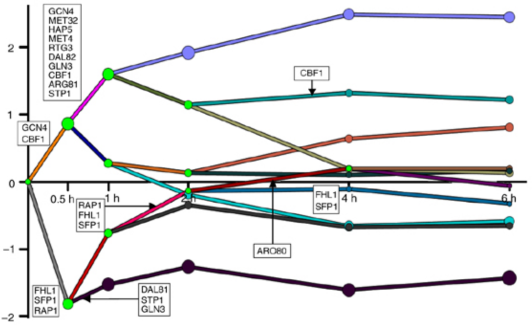
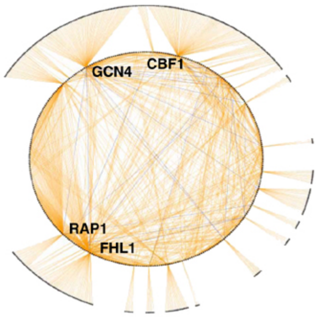
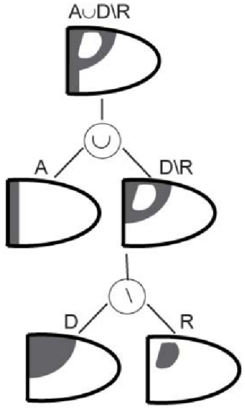
Abstracts of papers, posters and talks presented at the 2008 Joint RECOMB Satellite Conference on

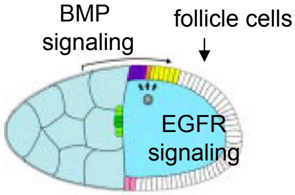
REGULATORY GENOMICS - SYSTEMS BIOLOGY - DREAM3

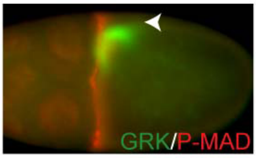
Oct 29-Nov 2, 2008
MIT / Broad Institute / CSAIL

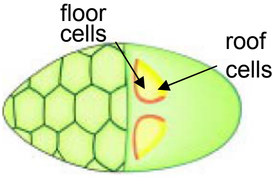


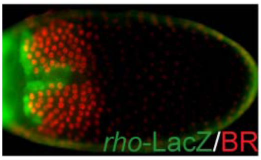












Organized by

Manolis Kellis, MIT
Andrea Califano, Columbia
Gustavo Stolovitzky, IBM

Abstracts of papers, posters and talks presented
at the 2008 Joint RECOMB Satellite Conference on

REGULATORY GENOMICS - SYSTEMS BIOLOGY - DREAM3

Oct 29-Nov 2, 2008
MIT / Broad Institute / CSAIL

Organized by

Manolis Kellis, MIT

Andrea Califano, Columbia

Gustavo Stolovitzky, IBM

Conference Chairs:

Manolis Kellis	Associate Professor, MIT
Andrea Califano	Professor, Columbia University
Gustavo Stolovitzky	Systems Biology Group, IBM

In partnership with:

Genome Research	editor: Hillary Sussman
Nature Molecular Systems Biology	editor: Thomas Lemberger
Journal of Computational Biology	editor: Sorin Istrail

Organizing committee:

Eleazar Eskin
Nir Friedman
Leroy Hood
Trey Ideker
Douglas Lauffenburger
Satoru Miyano
Eran Segal
Ron Shamir

Program Committee:

Regulatory Genomics:
Manolis Kellis, chair
Ziv Bar-Joseph
Panayiotis (Takis) Benos
Harmen Bussemaker
Ernest Fraenkel
Nir Friedman
Hui Ge
Mikhail Gelfand
Mark Gerstein
David Gifford
Sridhar Hannenhalli
Alex Hartemink
Uri Keich
Christina Leslie
Hao Li
Todd Lowe
Dana Pe'er
Yitzhak Pilpel
Mireille Regnier
Marie-France Sagot
Mona Singh
Eran Segal
Roded Sharan
Ron Shamir
Saurabh Sinha
John Stamatoyannopoulos
Alexander Stark
Gary Stormo
Saeed Tavazoie
Christopher Workman
Eric Xing
Michael Zhang

DREAM3:
Gustavo Stolovitzky, chair
M. Madan Babu
Gary Bader
Jim Collins
Joaquin Dopazo
Peicheng Du
Jean-Loup Faulon
Igor Jurisica
Winfried Just
Pascal Kahlem
Yannis Kevrekidis
Seungchan Kim
Andre Levchenko
Rune Linding
Kathleen Marchal
Adam Margolin
Satoru Miyano
Theodore Perkins
Raul Rabadan
Timothy Ravasi
John Rice
Michael Samoilov
Ilya Shmulevich
Eduardo Sontag
Brandilyn Stigler
Jesper Tegner
Yuhai Tu
Ioannis Xenarios

Systems Biology:
Andrea Califano, chair
Joel Bader
Harmen Bussemaker
Diego di Bernardo
Eric Davidson
Eleazar Eskin
Trey Ideker
Reinhard Laubenbacher
Andrew McCulloch
Ilya Nemenman
Dana Pe'er
Frederick Roth
Roded Sharan
Saeed Tavazoie
Ralf Zimmer

Local Organization and Student Volunteers:

Monica Concepcion, Broad
Sally Lee, CSAIL
Patrycja Missiuro
Julia Liu
Steve Hershman
Alexey Spiridonov
Deborah Ritter
Rachel Sealfon
Rogerio Candeias
Emily Capra
Subhajyoti De
Garrett Frampton
Eric Kelsic
Mei Lyn Ong
Sarah Pfau
Huangming Xie

Cover art credit:

Dynamic regulatory map reconstruction in yeast, Jason Ernst and Ziv Bar-Joseph.
Expression primitives in fly eggshell formation, Chris Bristow and Stas Shvartsman.

Industry and Government Sponsors:

The organizers wish to thank the following industrial and government sponsors for their generous support of this conference.

Merck Research Laboratories

The NIH National Centers for Biomedical Computing and the MAGNet Center at Columbia University

The New York Academy of Sciences

IBM Research

Scientific Partners

The organizers also wish to thank the following journals for their invaluable support in publishing conference manuscripts.

The Journal of Computational Biology

Genome Research

Nature / EMBO Molecular Systems Biology

Local Hosts and Organization

Many thanks also go to our local hosts for making this meeting possible.

Massachusetts Institute of Technology

Broad Institute of MIT and Harvard

Computer Science and Artificial Intelligence Lab

Wednesday, Oct 29, 2008

Systems Biology keynotes:

Aviv Bergman - A. Einstein Colleg.
David Botstein - Princeton
George Church - MIT/Harvard
Todd Golub - Broad Institute
Daphne Koller - Stanford University
Luis Serrano - CRG Spain
John Tyson - Virginia Tech

DREAM3 keynotes:

Doug Lauffenburger - MIT
Boris Kholodenko - T.Jefferson U

Regulatory Genomics Keynotes:

Uri Alon - Weizmann Institute
Chris Burge - MIT
Mark Gerstein - Yale
Tim Hughes - U. Toronto
Bing Ren - UC San Diego
Eddy Rubin - Berkeley/JGI
Leona Samson - MIT BE
Thomas Tuschl - Rockefeller

Wednesday, Oct 29, 2008		
5pm	Systems Biology: Registration / Poster SetUp	
5:15	**Pamela Silver: Designing Biological Systems**	**1**
5:45	Russo/diBernardo: Synchr. biol. clocks	2
6pm	Borenstein/Feldman: Metabolic topo sign	3
6:15	Rodriguez-Martinez/Furman: mRNA fluct	4
6:30	Mezes break - light snacks	
6:45	Burke/Sorger: Single-cell caspase amplif.	5
7pm	Fichtenholtz/Collins: Engin transc profiles	6
7:15	Szczurek/Vingron: Regul informative expts	7
7:30	**Daphne Koller: Activity patterns in biological networks**	**8**
8pm-11p	**Systems Biology: Welcome Reception Poster Session I** (hors-d'oeuvre, heavier snacks, cash bar)	

Thursday, Oct 30, 2008

8am	Breakfast	
9am	**Todd Gollub: Signatures for Small Molecule Discovery**	**9**
9:30	Karni/Sharan: Predicting Disease Genes	10
9:45	Iorio/di Bernardo: Network drug action	11
10am	Margolin/Carr: Drug target discovery	12
10:15	Coffee / Snacks / Fruit Break	
10:45	Veeramani/Bader: Metabol. flux in disease	13
11am	Chan/Song: CTCF confines Estrogen Rec	14
11:15	Wang/Schadt: Inter-Species Comparison of L	15
11:30	Miguez/Silver: Cancer bistability/drug resist.	16
11:45	Liu/Kasif: Type2 diabetes dysregulated nets	17
12pm	Lunch Break / Networking Opportunities	
1pm	**David Bostein: Coordination of growth, cell cycle, stress, metabolism in yeast**	**18**
1:30	Chechik/Koller: Gene expr. response timing	19
1:45	McDermott/Heffron: Virulence bottleneck nod	20
2pm	Qian/Yoon: Querying protein netw. pathways	21
2:15	Coffee Break	
2:45	Sahoo/Dill: Predicting B-Cell Develpt Genes	22
3pm	Beleva/Levchenko: Ecoli TCAcycle centrality	23
3:15	Eden/Alon: Dyn proteome response/cell fate	24
3:30	Degenhardt/Carlberg: Stochastic PPAR-depe	25
3:45-5:15p	**Systems Biology: Poster Session II** (hors-d'oeuvres, snacks, refreshments)	
5:15	**John Tyson: Reaction motifs and funct. modules in prot. networks**	**26**
5:45	Yosef/Sharan: Funct prot net reconstrction	27
6pm	Leung/Chin: Protein complexes from PPI data	28
6:15	Lee/Shakhnovich: Yeast complex Tx co-regul	29
6:30	Mezes break - light snacks	
6:45	Navlakha/Kingsford: Graph summarization	30
7pm	Shlomi/Ruppin: Human tissue-specific metab	31
7:15	Kuttykrishnan/Brent: Steady state to kinetics	32
7:30	**George Church: From personal genomes and enviromes to traits**	**33**
8pm-11pm	**Concert: Thousand Days**	

Friday, Oct 31, 2008

8am	Breakfast	
9am	**Aviv Bergman: Evolutionary Systems Biology in Health and Disease**	**34**
9:30	Hormozdiari/Cherkasov: Indels in prot netws	35
9:45	Mitchell/Pilpel: Adaptive envirmt conditioning	36
10am	Coffee / Snacks / Fruit Break	
10:15	DREAM introduction	
10:30	**Boris Kholodenko: Spatio-temp coding sign spec & netw reconstr**	**37**
11am	Popescu/D-Kumar: MAPK netw in plants	38
11:15	Sachs/Lauffenburger: Learning sign netw	39
11:30	Castelo/Roverato: Rev eng regulatory netw	40
11:45	Marucci/Bernardo: Rev eng yea synth net	41
12pm	Lunch Break / Networking Opportunities	
1pm	Altan-Bonnet: Challenge 1 intro	42
1:15	Saez-Rodriguez: Challenge 2 intro	43
1:30	Prill: Chllgs 1 and 2 overall results	
1:45	Guex: Challenge 2 best predictor talk	44
2pm	Bourque: Challenge 2 best predictor talk	45
2:15	Short Break	
2:30	Clarke: Challenge 3 intro	46
2:45	Marbach: Challenge 4 intro	47
3pm	Stolovitzky: Chllgs 3 and 4 overall results	
3:15	Gustafsson: Chlenge 3 best predictor talk	48
3:30	Ruan: Challenge 3 best predictor talk	49
3:45	Yip: Chllg 4 best predictor talk	50
4pm	Short Break	
4:15	**Doug Lauffenburger: Signaling netw in prim & transf hepatocytes**	**51**
4:30	DREAM3 - Future directions & adjourn	
5pm	Regulatory Genomics: Registration,Poster SetUp	
5:15	**Thomas Tuschl: Post-transcriptional networks**	**52**
5:45	Zhou/Zhang: anti-sense small RNAs	53
6pm	Khan/Marks: siRNAs vs. endogen. resp.	54
6:15	Gitter/Bar-Joseph: Euk code redundancy	55
6:30	Mezes break - light snacks	
6:45	**Mark Gerstein: Human intergenic annotation**	**56**
7:15	Erives: Ciona enhancer structure	57
7:30	Busser/Michelson: Muscle regulation	58
7:45	Megraw/Ohler,Hatzigeorgiou: TSS code	59
8pm-11p	**Regulatory Genomics: Halloween Poster Session III** (hors-d'oeuvre, snacks, refreshments, tricks, and treats)	

Saturday, Nov 1, 2008

8am	Breakfast	
9am	**Uri Alon: Design principles of biological systems**	**60**
9:30	Ahmed,Song/Xing: Temporal networks	61
9:45	Mezey/Logsdon: Bayesian networks	62
10am	Peterson/Vilo: Embryonic networks	63
10:15	Coffee / Snacks / Fruit Break	
10:45	Huttenhower/Coller: Integration networks	64
11am	Tchagang/Pan: Dynamic Factor Analysis	65
11:15	Madar/Bonneau: Coupled netw dynamics	66
11:30	Chen/Wang: Cancer knowledg integration	67
11:45	Zinman/Bar-Joseph: Cross spec expr clus	68
12pm	Lunch Break / Networking Opportunities	
1pm	**Bing Ren: Chromatin signatures of transcriptional enhancers**	**69**
1:30	Yuan: Histone mark recruitment prediction	70
1:45	Jothi/Zhao: Binding from ChIP-Seq	71
2pm	Ernst/Bar-Joseph: TF binding predition	72
2:15	Coffee / Snacks / Fruit Break	
2:45	Wasson/Hartemink: TF/nucl. Competition	73
3pm	Yeo/Clarke: TF targets from expr/binding	74
3:15	Halperin/Shamir: Modules from expr/seq	75
3:30	He/Sinha: Motif clustering/conservation	76
3:45-5:15p	**Regulatory Genomics: Poster Session II** (hors-d'oeuvres, snacks, refreshments)	
5:15	**Eddy Rubin: Human Enhancers**	**77**
5:45	DePace/Eisen: Fly embryo express. atlas	78
6pm	Frise/Celniker: Embryonic image mining	79
6:15	Li/Chuang,Guo: Brain express. enhancers	80
6:30	Mezes break - light snacks	
6:45	**Wang/Chris Burge: Tissue- and factor-specific RNA processing**	**81**
7:15	Foat/Stormo: Structural mRNA motifs	82
7:30	Reiss/Baliga: Reg. logic in coding regions	83
7:45	Waldman/Ruppin: Cancer transl. efficiency	84
8pm-11p	**Reception: A night at the MIT Museum**	

Sunday, Nov 2, 2008

8am	Breakfast	
9am	**Leona Samson: Predictors of interindividual diffs in DNA damage response**	**85**
9:30	Rosenfeld/Zhang: Single nucleo. repress.	86
9:45	Verlaan/Pastinen: Cis-regultory variation	87
10am	Lee/Bussemaker: Networks from linkage	88
10:15	Coffee / Snacks / Fruit Break	
10:45	Li/Leslie: Motifs from expr. trajectories	89
11:00	Li: EM spaced dyad motif discovery	90
11:15	Noyes/Wolfe: 1-hybrid TF characterization	91
11:30	**Tim Hughes: Protein-nucleic acid interaction mapping**	**92**
12pm	Adjourn	

RECOMB Systems Biology 2008

(★ Full length manuscript; ► Invited talk; Accepted abstract**)**

Wednesday, Oct 29, 2008

*Welcome / Registration / Poster Set Up – **5:00-5:15pm***

5:15► Pamela Silver: Designing Biological Systems 1
5:45★ Russo/diBernardo: How to synchronize biological clocks 2
6pm★ Borenstein/Feldman: Metabolic topological signatures 3
6:15 Rodriguez-Martinez/Furman: mRNA fluctuations / reg. RNAs 4

*Light snacks – **6:30-6:45pm***

6:45 Burke/Sorger: Single-cell caspase amplification 5
7pm Fichtenholtz/Collins: Engineering transcriptional profiles 6
7:15 Szczurek/Vingron: Regulation informative experiments 7
7:30► Daphne Koller: Activity patterns in biological networks 8

Systems Biology Poster Session I – 8pm-11pm
Authors of odd-numbered posters present 8pm - 9pm
Authors of even-numbered posters present 9pm - 10pm
Hors d'oeuvres, snacks, refreshments, cash bar

Thursday, Oct 30, 2008

Breakfast – 8am

9am► Todd Gollub: Signatures for Small Molecule Discovery 9
9:30★ Karni/Sharan: Predicting Disease Genes 10
9:45★ Iorio/diBernardo: Network drug action 11
10am★ Margolin/Carr: Drug target discovery... 12

*Coffee / Snacks / Fruit Break – **10:15-10:45am***

10:45★ Veeramani/Bader: Metabolic flux in disease 13
11am Chan/Song: CTCF confines Estrogen Receptor 14
11:15 Wang/Schadt: Inter-Species Comparison of Liver Networks..... 15
11:30 Miguez/Silver: Cancer bistability/drug resistance 16
11:45 Liu/Kasif: Type2 diabetes dysregulated networks 17

*Lunch Break / Networking Opportunities – **12-1pm***

1pm► David Bostein: Coord. growth, cell cycle, stress, metabolism 18
1:30★ Chechik/Koller: Gene expression response timing...................... 19
1:45★ McDermott/Heffron: Virulence bottleneck nodes 20
2pm★ Qian/Yoon: Querying protein network pathways 21

*Coffee / Snacks / Fruit Break – **2:15-2:45pm***

2:45 Sahoo/Dill: Predicting B-Cell Development Genes...................... 22
3pm Beleva/Levchenko: E. coli TCA cycle centrality........................... 23
3:15 Eden/Alon: Dyn proteome response/cell fate 24
3:30 Degenhardt/Carlberg: PPAR-dependent Tx cycling.................... 25

Systems Biology Poster Session II – 3:45pm-5:15pm
Authors of even-numbered posters present 3:45pm - 4:30pm
Authors of odd-numbered posters present 4:30pm - 5:15pm
Hors d'oeuvres, snacks, refreshments

5:15► John Tyson: Reaction motifs & funct modules in prot. networks. 26
5:45★ Yosef/Sharan: Functional prot net reconstrction 27
6pm★ Leung/Chin: Protein complexes from PPI data............................. 28
6:15★ Lee/Shakhnovich: Yeast complex Tx co-regul 29

*Light snacks – **6:30-6:45pm***

6:45★ Navlakha/Kingsford: Graph summarization 30
7pm Cabili/Ruppin: Human tissue-specific metab 31
7:15 Kuttykrishnan/Brent: Steady state to kinetics 32
7:30► George Church: From personal genomes, enviromes to traits ... 33

Concert – Thousand Days *8pm – 11pm*
Hors d'oeuvres, Cash Bar, Posters available for viewing

Friday, Oct 31, 2008

Breakfast – 8am

9am► Aviv Bergman: Evolutionary Sys Bio in Health and Disease....... 34
9:30★ Schoenhuth/Cherkasov: Indels in prot netws 35
9:45 Mitchell/Pilpel: Adaptive envirmt conditioning............................... 36
10am Systems Biology meeting adjourns

*Coffee Break – **10-10:15am***

DREAM3 Reverse Engineering Challenges

(★Full length manuscript; ► Invited talk; Accepted abstract**)**

Friday, Oct 31, 2008

10:15 DREAM introduction

10:30► Boris Kholodenko: Spatio-temp coding of signal specificity.......37

11am Popescu/D-Kumar: MAPK networks in plants............................38

11:15★ Sachs/Lauffenburger: Learning signaling networks...................39

11:30★ Castelo/Roverato: Rev eng regulatory networks 40

11:45 Marucci/diBernardo: Model & rev eng of a yeast synth netw.....41

*Lunch Break / Networking Opportunities – **12-1pm***

1pm Altan-Bonnet: Challenge 1 intro..42

1:15 Saez-Rodriguez: Challenge 2 intro...43

1:30 Prill: Challenges 1 and 2 overall results

1:45 Guex: Challenge 2 best predictor talk 44

2pm Bourque: Challenge 2 best predictor talk 45

*Short Break – **2:15-2:30pm***

2:30 Clarke: Challenge 3 intro 46

2:45 Marbach: Challenge 4 intro...47

3pm Stolovitzky: Challenges 3 and 4 overall results

3:15 Gustafsson: Challenge 3 best predictor talk................................48

3:30 Ruan: Challenge 3 best predictor talk 49

3:45 Yip: Challenge 4 best predictor talk...50

*Short Break – **4-4:15pm***

4:15► Doug Lauffenburger: Signaling networks in hepatocytes 51

4:30 DREAM3 - Future directions & adjourn

RECOMB Regulatory Genomics 2008

(★Full length manuscript; ► Invited talk; Accepted abstract**)**

Friday, Oct 31, 2008

5pm Welcome / Registration / Poster SetUp

5:15► Thomas Tuschl: post-transcriptional networks 52
5:45★ Zhou/Zhang: anti-sense small RNAs.. 53
6pm Khan/Marks: siRNAs vs. endogen. resp...................................... 54
6:15 Gitter/Bar-Joseph: Euk code redundancy..................................... 55

*Light snacks – **6:30-6:45pm***

6:45► Mark Gerstein: Human intergenic annotation 56
7:15★ Erives: Ciona enhancer structure 57
7:30 Busser/Michelson: Muscle regulation 58
7:45 Megraw/Ohler,Hatzigeorgiou: TSS code 59

Regulatory Genomics Poster Session I – 8pm-11pm

Authors of odd-numbered posters present 8pm - 9pm
Authors of even-numbered posters present 9pm - 10pm
*Hors d'oeuvres, snacks, refreshments, cash bar, **tricks and treats***

Saturday, Nov 1, 2008

Breakfast – 8am

9am► Uri Alon: Design principles of biological systems 60
9:30 Ahmed,Song/Xing: Temporal networks 61
9:45 Mezey/Logsdon: Bayesian networks 62
10am Peterson/Vilo: Embryonic networks... 63

*Coffee / Snacks / Fruit Break – **10:15-10:45am***

10:45 Huttenhower/Coller: Integration networks.................................. 64
11am Tchagang/Pan: Dynamic Factor Analysis 65
11:15 Madar/Bonneau: Coupled network dynamics 66
11:30★ Chen/Wang: Cancer knowledge integration 67
11:45 Zinman/Bar-Joseph: Cross species expression clustering 68

*Lunch Break / Networking Opportunities – **12-1pm***

1pm► Bing Ren: Chromatin signatures of transcriptional enhancers 69
1:30★ Yuan: Histone mark recruitment prediction 70
1:45 Jothi/Zhao: Binding from ChIP-Seq 71

2pm Ernst/Bar-Joseph: TF binding prediction 72

*Coffee / Snacks / Fruit Break – **2:15-2:45pm***

2:45 Wasson/Hartemink: TF/nucleosome competition model 73

3pm★ Yeo/Clarke: TF targets from expression/binding 74

3:15 Halperin/Shamir: Modules from expression/sequence................ 75

3:30 He/Sinha: Motif clustering/conservation 76

Regulatory Genomics Poster Session II – 3:45pm-5:15pm

Authors of even-numbered posters present 3:45pm - 4:30pm

Authors of odd-numbered posters present 4:30pm - 5:15pm

5:15► Eddy Rubin: Human Enhancers 77

5:45 DePace/Eisen: Fly embryo expression atlas 78

6pm Frise/Celniker: Embryonic image mining 79

6:15 Li/Chuang,Guo: Brain expression enhancers............................... 80

*Light snacks – **6:30-6:45pm***

6:45► Wang/Chris Burge: Tissue-/Factor- specific RNA processing..... 81

7:15 Foat/Stormo: Structural mRNA motifs 82

7:30 Reiss/Baliga: Regulatory logic in coding regions.......................... 83

7:45 Waldman/Ruppin: Cancer translation efficiency........................... 84

A night at the museum reception *8pm – 11pm*

MIT Museum within walking distance down Main St to 265 Mass Ave

Sunday, Nov 2, 2008

Breakfast – 8am

9am► Leona Samson: Interindividual diffs in DNA damage response .. 85

9:30 Rosenfeld/Zhang: Single nucleosome repression 86

9:45★ Verlaan/Pastinen: Screening human cis-regulatory variation...... 87

10am Lee/Bussemaker: Networks from inferred TF activity linkage 88

*Coffee / Snacks / Fruit Break – **10:15-10:45am***

10:45 Li/Leslie: Motifs from expression trajectories 89

11:00★ Li: EM spaced dyad motif discovery 90

11:15 Noyes/Wolfe: One-hybrid Drosophila TF characterization......... 91

11:30 Tim Hughes: Protein-nucleic acid interaction mapping.............. 92

*Closing remarks + adjourn – **12pm***

Designing Biological Systems

Pamela A. Silver[1]

[1]*Department of Systems Biology, Harvard Medical School, Boston, MA.*

Biology presents us with an array of design principles. From studies of both simple and more complex systems, we understand at least some of the fundamentals of how Nature works. We are interested in using the foundations of biology gleaned from Systems Biology to engineer cells in a logical way to perform certain functions. In doing so, we learn more about the fundamentals of biological design as well as engineer useful devices with a myriad of applications. For example, we are interested in building cells that can perform specific tasks, such as memory and timing of past events. Moreover, we design and construct proteins and cells with predictable biological properties that not only teach us about biology but also serve as potential therapeutics, cell-based sensors and factories for generating bio-energy.

How to Synchronize Biological Clocks

Giovanni Russo[1], Mario di Bernardo[1-2].

[1] *Department of Systems and Computer Engineering, University of Naples Federico II;*
[2] *Department of Engineering Mathematics, University of Bristol.*

Clocklike rhythms are found in every organism, from bacteria to humans and need to be perfectly synchronized in order to regulate the basic mechanisms for life. Biological oscillators can synchronize even in noisy environments and despite differences between the biochemical parameters of the clocks. Even if this intrinsic robustness has stimulated several attempts to explain the emergence synchronization phenomena, a systematic theoretical framework for its study is still lacking.

This paper is concerned with a novel algorithm to study networks of biological clocks. A new set of conditions is established that can be used to verify whether an existing network synchronizes or to give guidelines to construct a new synthetic network of biological oscillators that synchronize. The methodology uses the so-called contraction theory from dynamical system theory and Gershgorin disk theorem. The main features of our algorithm is that it explicitly takes into account the presence of noise and parameter mismatches and provides simple algebraic constraints on the parameters of the oscillator. The strategy is validated on two examples: a model of glycolisis in yeast cells and a synthetic network of Repressilators that synchronizes.

Topological Signatures of Species Interactions in Metabolic Networks

Elhanan Borenstein[1,2], Marcus W. Feldman[1]

[1]*Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA;*
[2]*Santa Fe Institute, Santa Fe, New Mexico 87501, USA.*

The topology of metabolic networks can provide insight not only into the metabolic processes that occur within each species, but also into interactions between different species. Here, we introduce a pair-wise, topology-based measure of biosynthetic support, reflecting the extent to which the nutritional requirements of one species could be satisfied by the biosynthetic capacity of another. To evaluate the biosynthetic support for a given pair of species, we use a graph-based algorithm to identify the set of exogenously acquired compounds in the metabolic network of the first species, and calculate the fraction of this set that occurs in the metabolic network of the second species. Reconstructing the metabolic network of 569 bacterial species and several eukaryotes, and calculating the biosynthetic support score for all bacterial-eukaryotic pairs, we show that this measure indeed reflects host-parasite interactions and facilitates a successful prediction of such interactions on a large-scale. Integrating this method with phylogenetic analysis and calculating the biosynthetic support of ancestral Firmicutes species further reveal a large-scale evolutionary trend of biosynthetic capacity loss in parasites. The inference of ecological features from genomic-based data presented here lays the foundations for an exciting 'reverse ecology' framework for studying the complex web of interactions characterizing various ecosystems.

messenger RNA fluctuations and regulatory RNAs shape the dynamics of a negative feedback loop

María Rodríguez Martínez[1], Yitzhak Pilpel[1,2], Tsvi Tlusty[1,] Jordi Soriano[1,3] and Itay Furman[1]

[1]Weizmann Institute of Science, Israel; [2]Harvard Medical School, USA; [3]Facultat de Física, Universitat de Barcelona, Spain.

The growing interest in biological noise has led to many efforts to measure gene expression at the single cell level, revealing a very distinct dynamics when compared to population cell experiments. An outstanding example is provided by the p53-mdm2 negative feedback loop, where single cell experiments of both p53 and mdm2 expression show a series of sustained anti-correlated pulses, while cell population experiments show damped oscillations.

In this work, we focus on the dynamics of the p53 and mdm2 regulatory network at the single cell level. We consider three alternative designs of the loop and model their stochastic behavior using Gillespie's algorithm.

We first investigate the role of RNA stochastic fluctuations as a leading force to produce a sustained excitatory behavior in single cells. We find that RNA random fluctuations can be amplified during translation and produce pulses of protein expression with a characteristic inter-spike time lag. Ignoring the stochastic nature of the mRNA leads to the disappearance of the pulses. Furthermore, we show that for a system such as this, deterministic or stochastic linearized models are insufficient to provide an accurate dynamical description of the system, and therefore a full stochastic treatment such as provided by Gillespie's algorithm is required. We also study the effect of changing reaction rates on the coefficient of variation and observe that contrary to previous beliefs different levels of noise can be observed at the same protein levels.

On the other hand, non-coding regulatory RNAs have been shown to have an important role in regulating p53 activity. To account for it, we include a regulatory RNA transcript, the antisense, that sequesters mRNA and slows down translation. We find that under sufficiently strong mRNA and antisense interaction the system is less excitable and the pulses are partially smoothed out. Fast mRNA and antisense interaction is required for the mechanism to be efficient. We show that in the presence of a regulatory RNA, the individual cell's dynamics is more robust to transcript fluctuations, hence leading to a reduced cell-to-cell variability. We also find that the best agreement with experimental data is obtained for models that include antisense regulation.

In conclusion, we have explored three architectural variations of a negative feed back loop: a base model that consists of p53, mdm2 mRNA and mdm2 protein; a reduced model where the mRNA stochastic dynamics has been averaged out, and an extended model that contains an additional regulatory step through non-coding RNA interaction with the mRNA transcript. We have observed remarkable differences among them that emerge from the RNA dynamical description. We have demonstrated that RNA fluctuations are able to induce a sustained excitable behavior and have lent support to prior suggestions that the antisense mRNA mechanism can buffer noise.

Mathematical and experimental single-cell analysis of caspase amplification in the death receptor network

John M. Burke[1,2,4], John G. Albeck[1,4], Sabrina L. Spencer[1,3,4], Suzanne Gaudet[1,4], Bree B. Aldridge[1,2,4], Mingsheng Zhang[1,4], Douglas A. Lauffenburger[2,3,4], Peter K. Sorger[1,2,4]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; [2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [3]Computational and Systems Biology Initiative, Massachusetts Institute of Technology, Cambridge MA 02139; [4]Center for Cell Decision Processes, Massachusetts Institute of Technology, Cambridge, MA 02139.

Apoptotic cell death is an essential physiological process misregulated in many diseases. Understanding aspects of quantitative apoptotic regulation is a central challenge. For example, all-or-none activation of executioner caspases at the single cell level potentially has major consequences in evolution, disease, and drug resistance. While numerous theoretical mathematical models propose to explain these consequences, few are validated by direct experimental evidence. To address this challenge, a system of mass-action ordinary differential equations describing apoptotic regulation is derived and compared with extensive experimentation at the single cell level. First, model analysis identifies, and experiments verify, signal transduction control mechanisms in which the graded upstream signal induced by the initial death stimulus is converted into a rapid all-or-none downstream response. Second, the model predicts conditions under which all-or-none caspase activation fails, yielding live single cells with stable, nonzero cleaved PARP levels (substrates of cleaved executioner caspases; measure of cell death), or “undead” cells; that is, single cells that exhibit sub-lethal partial cleaved PARP levels under wild type lethal ligand doses, abrogating the all-or-none death switch. The existence of “undead” cells is experimentally validated. These undead cells proliferate, suggesting a mechanism of creating and/or perpetuating DNA-damaged cells, possibly leading to Cancer. Applying the knowledge gained from the synergy of math modeling and biology identifies key mechanisms of cellular control that, when targeted therapeutically, may alter the apoptotic fate to a more desirous outcome. Thus, computational and experimental studies have combined to generate a comprehensive model describing the caspase regulatory network and cell-to-cell variability, which accurately predicts normal and pathological behavior, which may have long lasting and critical effects curing diseases such as Cancer and controlling T cell regulation.

Engineering a Global Transcriptional Profile via Modification of Regulation Parameters

Alex Fichtenholtz, Diogo Camacho, Jim Collins.

Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215

Many cellular engineering applications have a prerequisite requiring precise, yet widespread control of gene expression. This control needs to be on a large scale to coordinate the proper signaling that orchestrates many complex phenotypes. For instance, the field of metabolic engineering is limited by system wide bottlenecks which can only be removed through simultaneously modifying the expression of a large number of genes. Indeed, several techniques exist for massively perturbing the gene expression profile of *E. coli*, however, these techniques lack flexibility and predictivity, two requirements for robust transcriptional engineering. We present a combined computational/experimental technique for precisely tuning the transcription profile of *Escherichia coli* on a global scale in a programmable fashion. Our method relies on modification of parameters governing regulation between central transcription factors, which can be thought of as control parameters for the expression profile of the cell. We slide these parameters into a particular combination of values to achieve the precise profile we are interested in. With this method we are able to hit target gene expression on a global scale, allowing considerable flexibility when designing cellular phenotypes.

Elucidating Regulatory Mechanisms Downstream of a Signaling Pathway Using Informative Experiments

Ewa Szczurek[1,2,3], Irit Gat-Viks[1], Jerzy Tiuryn[3] and Martin Vingron[1]

[1]Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany, [2]International Max Planck Research School for Computational Biology and Scientific Computing, Berlin, Germany, and [3]Institute of Informatics, Warsaw University, Banacha 2, 02-097, Warsaw, Poland

Signaling cascades are triggered by extra-cellular stimulation and propagate the signal to regulate transcription. Systematic reconstruction of this regulation requires pathway-targeted, informative experimental data. However, experimental design is difficult since even highly informative experiments might be redundant with other experiments. In addition, experimental outcomes vary not only between different genetic perturbations but also between the combinations of environmental stimuli.

We have developed a practical algorithmic framework that iterates design of experiments and reconstruction of regulatory relationships downstream of a given pathway. The experimental design component of the framework, called MEED, proposes a set of experiments the can be performed in the lab and given as input to the reconstruction component. Both components take advantage of expert knowledge about the signaling system under study, formalized in a predictive logical model. The reconstruction component reconciles the model predictions with the data from the designed experiments to provide a set of identified target genes, their regulators in the pathway and their regulatory mechanisms. Reconstruction based on uninformative data may lead to ambiguous conclusions about the regulation. To avoid ambiguous reconstruction, MEED designs experiments so as to maximize diversity between the predicted expression profiles of genes regulated through different mechanisms.

MEED has several important benefits and advantages over extant experimental design approaches: First, it considers potential dependencies between the suggested experiments, making it possible to design and perform in parallel a set of informative, non-redundant experiments. Second, MEED optimizes not only the required genetic perturbations, but also the combination of environmental stimuli that should trigger the system. Finally, by using only the model predictions, MEED has the ability to choose experiments without access to high-throughput experimental data.

Our framework was extensively analyzed and applied on random models as well as the model of interconnected osmotic stress and pheromone pathways in *Saccharomyces cerevisiae*. In comparison to other approaches, MEED allowed to provide significantly less ambiguous conclusions about the regulation in this system.

Activity Patterns in Biological Networks: Statistical Analysis and Network Reconstruction

Gal Chechik[1], Eugene Oh[2], Oliver Rando[3], Jonathan Weissman[2], Aviv Regev[4], Daphne Koller[1]

[1]Stanford University; [2]UC San Francisco; [3]Univ. Mass. Medical School; [4]Broad Institute

Significant insight about biological networks arises from the study of network motifs: small wiring patterns that are overly abundant in the network. However, wiring patterns, like a street map, only reflect the set of potential routes within a cellular network, but not when and how they are used within different cellular processes. We study the *functional* behavior of networks by introducing the notion of “activity motifs”, which, like traffic flow, reflect dynamic and context-specific patterns in functional data that are abundant relative to the given network. Activity motifs can be used to analyze properties of biological networks, and to inform algorithms that reconstruct the network structure. The talk will describe the framework of activity motifs and some of its applications to different networks and different types of functional data.

As one example, we use this framework to study the fine-grained timing of transcriptional regulation in Saccharomyces cerevisiae metabolism. We find that metabolic pathways are enriched for a variety of patterns, such as ordered activation or repression in linear chains; these patterns allow for efficient production and degradation in response to environment changes. Protein time-course experiments show that timing patterns in mRNA are conserved at the protein level. To study the mechanism underlying fine-grained timing, we define activity motifs involving the binding strength of a single transcription factor to its targets. Motifs where binding affinity is ordered in a linear chain are also abundant, and overlap significantly with the timed motifs, suggesting that fine gradations in transcription factor affinity provide one mechanism for ordered transcription.

Signatures for Small Molecule Discovery

Todd R. Golub

The Broad Institute of Harvard and MIT, Dana-Farber Cancer Institute, Boston, MA USA

The application of genomic approaches to the study of cancer holds tremendous promise for improved diagnostic and prognostic tests, and for the elucidation of new therapeutic targets by building a molecular taxonomy of the disease. More recently, we have addressed the challenge of using gene expression data in the drug discovery setting. That is, having defined a gene expression signature of a biological state of interest (e.g. tumor subtype or state of pathway activation), could a small molecule library be screened to identify compounds capable of modulating the signature of interest – and by inference, modulate the biological state under study. We piloted this idea, termed Gene Expresion-based High Throughput Screening (GE-HTS), and applied it to the discovery of compounds capable of inducing the myeloid differentiation of acute myeloid leukemia cells. Importantly, the discovery of these compounds did not require a specialized phenotypic assay, nor did it require prior knowledge of the mechanism by which differentiation occurs. We have subsequently applied this GE-HTS concept to the discovery of compounds that inhibit the activity of the Ewing Sarcoma oncogene EWS/FLI and that abrogate androgen receptor signaling in prostate cancer. These experiments establish the feasibility of using a gene expression signature as the read-out of a primary small-molecule screen.

Extending on this concept of signature-based chemical screening, we have recently established the feasibility of using a database of gene expression profiles to systematically connect signatures of diseases to signatures of gene product function or signatures of drug action. We refer to this project as the Connectivity Map project. By querying a centrally generated database of gene expression profiles, users can find ‘connectivity’ between a query signature of interest and one or more treatments (perturbagens) in the database. The data and tools are available at www.broad.harvard.edu/cmap, and we have used the method to discover relevant connections in dexamethasone-resistant childhood leukemia, androgen response in prostate cancer, and connections to HDAC inhibition in various cell types. These experiments demonstrate the feasibility of the Connectivity Map approach, and suggest the value of creating a larger, more extensive, publicly accessible Connectivity Map database.

A Network-Based Method for Predicting Disease-Causing Genes

Shaul Karni[1], Hermona Soreq[2], Roded Sharan[1]

[1]Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, 69978, Israel.

[2]Interdisciplinary Center for Computational Neuroscience, The Hebrew University of Jerusalem, Jerusalem, 91904, Israel.

A fundamental problem in human health is the inference of disease-causing genes, with important applications to diagnosis and treatment. Previous work in this direction relied on knowledge of multiple loci associated with the disease, or causal genes for similar diseases, which limited its applicability. Here we present a new approach to causal gene prediction that is based on integrating protein-protein interaction network data with gene expression data under a condition of interest. The latter are used to derive a set of disease-related genes which is assumed to be in close proximity in the network to the causal genes. Our method applies a set-cover like heuristic to identify a small set of genes that best “cover” the disease-related genes. We perform comprehensive simulations to validate our method and test its robustness to noise. In addition, we validate our method on real gene expression data and on gene specific knockouts. We apply it to suggest possible genes that are involved in Myasthenia Gravis. Finally, we use a large collection of gene expression data sets for different diseases to study the topological characteristics of the causal genes as a function of disease type and stage.

Identifying Network of Drug Mode of Action by Gene Expression Profiling

Francesco Iorio[1,2], Roberto Tagliaferri[2], Diego di Bernardo[1,3]

[1] *Systems and Synthetic Biology Lab – TeleThon Institute of Genetics and Medicine (TIGEM), Naples (ITALY);*

[2] *Dept. of Mathematics and Computer Science, University of Salerno (ITALY)*

[3] *Dept. of Computer Science and Systems, University "Federico II" of Naples (ITALY)*

Drug Mode of Action has so far been predicted using phenotypic features or, more recently, comparing side effects similarities. Attempts to use gene expression data in mammalian systems so far have met limit successes.

Here, starting from a public reference dataset containing genome-wide expression profiles following treatments with more than a thousand compounds, we build a Drug Similarity Network providing a map of similarities in mode of action. In this network, drugs sharing a subset of molecular targets are connected or lie in the same community.

Our approach is based on a novel similarity metrics, combining an original rank-aggregation method with the analysis of the enrichment of a set of genes in ranked lists (Gene Set Enrichment Analysis). On the basis of this similarity, distances between drugs are defined and the network is constructed by connecting pairs of drugs whose distance is below a pre-defined threshold.

With our method, *molecular signatures*, characterizing the Mode of Action of the drugs in the reference dataset, are automatically obtained, without across-conditions normalization procedures. These signatures are used to compute similarities between drug modes of action and to map this information in a network form.

We start by merging lists obtained from treatments with the same drug, on different cell lines and for different concentrations, in order to obtain a *prototype ranked list* (PRL). The PRL is computed by means of a novel rank aggregation method, combining an algorithm for the computation of the *Minimum Spanning Tree* in a graph (a majority voting system to merge ranked lists) and a distance metric to compare the agreement between ranked lists. From each of these PRLs, an *optimal signature* of genes is computed for each drug by selection the top- and bottom-ranked 250 genes in the PRL. Similarities in mode of action are obtained by quantifying the enrichment of these optimal signatures in the other PRLs.

As the distance threshold (used to construct the network) increases, the network 'grows' in a coherent way, connecting drugs with similar mode of action.

The *Receiver Operating Characteristic* analysis on the community structures and the neighborhood of predefined test-drugs has been performed against the "Anatomical Therapeutic Chemical Classification System" codes (ATC-codes), showing very promising performances. Connections have been also confirmed by literature search.

We show that, despite the complexity and the variety of the experimental conditions, our approach, using gene expression data only, is able to identify similarities in drug mode of action thus allowing the classification of new drugs.

SysBio

Proteome-wide discovery of drug targets by empirical Bayes analysis of quantitative mass spectrometry data

Adam A. Margolin[1], Shao-En Ong[1], Monica Schenone[1], Robert Gould[1], Stuart L. Schreiber[1], Todd R. Golub[1*], Steven A. Carr[1*]

[1] *The Broad Institute of Harvard and MIT, Cambridge MA, 02142*

** These authors contributed equally to this work*

Rapid and reliable proteome-wide identification of compound targets is a major challenge in modern drug discovery. The cellular targets of many FDA approved drugs remain uncharacterized while many other drugs, thought to interact specifically with a particular target, are later shown to have substantial off-target effects. The ability to rapidly test thousands of chemical compounds in a phenotypic screen exacerbates the need for methods able to link these functional effects to the compound's mechanism of action. We have recently developed a high throughput assay to quantitatively measure compound-target interactions from tandem affinity pull-down experiments. Novel analytical tools are required to understand the capabilities and limitations of this technology, inform optimal experimental design, and provide statistical measures of predicted targets. Here we develop an empirical Bayes framework for analysis of quantitative proteomics experiments and apply it to the identification of the protein targets of six compounds. The method identified targets of well-characterized kinase inhibitors at exquisite precision, and detected novel targets of immunophilin ligands that were biochemically validated. We believe this method will assist in accelerating drug discovery by confidently identifying compound targets in proteome-wide assays.

Metabolic flux correlations, genetic interactions and disease

Balaji Veeramani[1,2], Joel S. Bader[1,2]

[1]*Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA*

[2]*High-Throughput Biology Center, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA*

Many diseases are caused by failures of metabolic enzymes. These enzymes exist in the context of networks defined by the static topology of enzyme-metabolite interactions and by the reaction fluxes that are feasible at steady state. We use the local topology and the flux correlations to identify how failures in the metabolic network may lead to disease. First, using yeast as a model, we show that flux correlations are a powerful predictor of pairwise mutations that lead to cell death – more powerful, in fact, than computational models that directly estimate the effects of mutations on cell fitness. These flux correlations, which can exist between enzymes far-separated in the metabolic network, add information to the structural correlations evident from shared metabolites. Second, we show that flux correlations in human align with similarities in Mendelian phenotypes ascribed to known genes. These methods will be useful in predicting genetic interactions in model organisms and understanding the combinatorial effects of genetic variations in human..

CTCF Confines the Distal Action of Estrogen Receptor

Chang S. Chan[1], Jun S. Song[1]

[1]*The Simons Center for Systems Biology, The Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540*

Distal enhancers have recently emerged as a common mode of gene regulation for several transcription factors (TFs), including estrogen and androgen receptors, the two key regulators of breast and prostate cancer major subtypes. Despite the rapid success in genome-wide annotation of estrogen receptor-α (ER) binding sites in cell lines, the precise mechanism governing the gene-to-enhancer association is still unknown, and no quantitative model that can predict the estrogen-responsiveness of genes has been hitherto proposed. This manuscript presents an integrative genomics approach to construct a predictive model that can explain over 70% of estrogen-induced expression profiles. The proposed method combines a recent map of the insulator protein CTCF with previous ER location studies and expression profiling in the presence of the translation inhibitor cycloheximide, providing evidence that CTCF partitions the human genome into distinct ER-regulatory blocks. It is shown that estrogen-responsive genes with a decreased transcription level (down-regulated genes) have a markedly different relative distribution of ER binding sites compared to those with an increased transcription level (up-regulated genes). Finally, Bayesian belief networks are constructed to quantify the effects of ER binding distance from genes as well as the insulating effects of CTCF on the estrogen-responsiveness of genes. This work thus represents a stride towards understanding and predicting the distal activities of steroid hormone nuclear receptors.

Inter-Species Comparison of Liver Co-expression Networks Elucidates Traits Associated with Common Human Diseases

Kai Wang[1], Manikandan Narayanan[1], Martin Tompa[1], Jun Zhu[1], Eric E. Schadt[1]

[1]*Department of Genetics, Rosetta Inpharmatics, Seattle, WA, USA*

Multiple genetic and environmental factors contribute to common human disease risk. Developing a more comprehensive view of common human diseases requires a systems level approach. Human and mouse transcriptional networks from multiple tissues have been used to study human diseases like obesity and diabetes. Comparing these types of networks across different species can lead to the identification of common links as well as to the removal of species-specific links that are due to guilt-by-association in a single dataset. Network comparisons can also help characterize network plasticity due to evolution. Liver is an important organ for glucose and lipid metabolism, as well as for metabolizing toxic compounds. Mouse and rat are commonly used animal models in studying human diseases. In this paper we focus on comparing the co-expression networks in the liver of human, mouse and rat. The goal is to understand mechanisms that are conserved between human and rodent species as a way to provide support for drug targets identified in mouse and rat having relevance in humans. In addition, by identifying mechanisms that differ between these species, we hope to improve the design and interpretation of toxicity studies using rodent models.

Toward this end, we developed a novel meta-analysis approach for combining and contrasting gene-gene co-expression relationships across diverse expression profile datasets generated in several species. Our method is semi-nonparametric in nature, incorporating advantages of previously reported parametric and nonparametric methods. When applied to a meta-analysis of liver co-expression networks in human, mouse and rat, our method out-performed all existing methods with respect to identifying gene pairs with coherent biological functions. We identified modules of co-expressed genes representing core biological processes that are essential to liver function in all three species. We also demonstrate that the conserved co-expression networks provide a better platform for annotating human lipid candidate genes identified in genome-wide association studies (GWAS), when compared to co-expression networks built from human expression profiles alone. Based on gene modules derived from the conserved co-expression network, we associated SORT1, a gene strongly linked to LDL cholesterol levels, with cell-cell signaling, establishing a link between cell-cell signaling and blood lipid regulation. Our results also revealed a small module consisting of lipid metabolism genes, including 3 GWAS candidates that affect human blood lipid concentration, PCSK9, HMGCR and MVK.

We further proposed the use of heterogeneity statistics as a way to assess differences in the correlation structures between multiple datasets. With this method we identified differences in co-expression networks that are human- or rodent- specific, and demonstrated that genes with species specific interactions tend to be under positive selection during evolution. We identified RXRG as a key regulator for genes with human specific interactions, and speculate that its differential regulation in human and rodent involves activating a downstream gene, CETP, which is an important cholesterol metabolism gene existing only in human. Taken together, our approach represents a novel step forward in integrating gene co-expression networks from multiple large scale datasets to leverage not only conserved information but also differences that are dataset specific.

SysBio

Bistability in the Akt-mTOR crosstalk increases resistance to anti-cancer drugs

David G. Míguez,[1], Pamela A. Silver

[1] *Department of Systems Biology, Harvard Medical School Boston MA*

The crosstalk between the Akt and the mTOR pathways results in several feedback regulations of the signaling. The combination of these positive and negative feedback loops results in a nonlinear response that influences the outcomes of cancer drug therapies. These nonlinearities often lead to, for instance, switch-like behavior or hysteresis and determine the optimal drug concentration, since its specificity is lost at high concentrations.

Using a reversible inhibitor (i.e., it does not form a covalent bond with its target and can be washed out) we perturbed the AKT/mTOR pathways to test for bistability. Two separate sets of cells, one of them pre-treated with the inhibitor, were further treated with several dilutions of the same inhibitor. Measurements of the pathway activity demonstrated the existence of bistable behavior. This “memory” of the pathway has unprecedented characteristics, with the pre-treated cells exhibiting increased resistance to the inhibitor (during the second inhibition, more drug is needed to produce the same anti-cancer effect in the cells). Experiments where the mTOR pathway is also inhibited exhibited no bistability, showing that the ability to “remember” arises from the crosstalk between the Akt and mTOR pathways in the form of feedback loops.

Dysregulated Gene Networks in Type 2 Diabetes

Manway Liu[1], Jonathan Dreyfuss[2], Eric Kolaczyk[3], Isaac Kohane[4], ME Patti[5], Alison Goldfine[5], Simon Kasif[1]

[1] Department of Biomedical Engineering, Boston University; [2] Harvard Partners Center for Bioinformatics and Genomics; [3] Department of Mathematics and Statistics, Boston University; [4] Center of Biomedical Informatics, Harvard Medical School; [5] Joslin Diabetes Center

Type 2 diabetes is a metabolic disorder affecting hundreds of millions of people worldwide. It is a complex disease with both genetic and environmental factors and arises from dual defects in insulin action and secretion. While the general picture of dysregulation in this disease is known, the mechanistic details remain poorly understood. Identifying relevant genes is particularly challenging in type 2 diabetes as the disease has mostly weak effects at the gene expression level and consequently few genes are strongly dysregulated in the disease state. This has motivated research aiming to identify gene networks, rather than individual genes, that are cumulatively dysregulated in disease. Ultimately, such work may lead to genetic biomarkers for metabolic staging as well as putative drug targets. Nonetheless, as a first step, our goal here is to use gene networks to identify those genes differentiating between disease and normal phenotypes and also correlating to progression of insulin resistance (a precursor condition to type 2 diabetes).

Towards that end, we have developed a novel and efficient analysis approach that identifies a dysregulated gene network in disease. In contrast with our previous methods, the new approach runs in linear (rather than exponential) time relative to the input, while yielding biologically comparable results. When applied to a dataset of type 2 diabetic vs. non-diabetic patients in skeletal muscle tissue, our approach identifies a large, dysregulated gene network that is significantly enriched in insulin signaling, oxidative stress, and inflammation-related genes (FWER $p \leq 0.05$, Fisher Exact Test). Insulin signaling is not detected by standard gene set enrichment analyses and few genes are individually differentially expressed in this dataset, in agreement with other studies. Besides well-known members such as GLUT4, PPARG, PGC1A, HNF4A, IRS1, AKT2, and GSK3, the network also contains a significant number ($p \leq 0.05$, Fisher Exact Test) of genes that are individually correlated to the clinical measurement of insulin sensitivity. Indeed, over 10% of the genes within the network are correlated as such, including TCF7L2, MAPK8 (JNK), LPL, and PIK3R1. TCF7L2 has been recently implicated in diabetes through multiple genome-wide association studies, though its causal role remains unclear. It is known to interact with β-catenin and believed to affect glucose homeostasis through the Wnt signaling pathway. The gene is not considered to be strongly expressed in skeletal muscle and our observation of its negative correlation ($R = -0.40$, $p = 0.004$) to insulin sensitivity is particularly intriguing in this context. Likewise, LPL has been previously implicated in multiple metabolic disorders and causally associated with obesity. We also compare our human gene network with one derived from a mouse model of diet-induced obesity and consider commonly dysregulated genes and interactions between them. Similarly to the human network, the mouse network is enriched in inflammation-related genes and includes the genes TCF7L2, MAPK8, and LPL. In addition, both networks share commonly dysregulated interactions related to insulin and interleukin-6 signaling pathways.

Coordination of Growth Rate, Cell Cycle, Stress Response and Metabolic Activity in Yeast

David Botstein

Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton NJ

This talk will describe the coordination of growth rate, cell cycle, stress response and metabolic activity in yeast.

Timing of gene expression responses to environmental changes

Gal Chechik[1,2], Daphne Koller[2]

[1]Google Research; [2]Computer science department, Stanford University

Cells respond to environmental perturbations with changes in their gene expression that are coordinated in magnitude and time. Timing information about individual genes, rather than clusters, provides a refined way to view and analyze responses, but is hard to estimate accurately.

To analyze response timing of individual genes, we developed a parametric model that captures the typical temporal responses: an abrupt early response followed by a second transition to a steady state. This *impulse* model explicitly represents natural temporal properties such as the onset and the offset time, and can be estimated robustly, as demonstrated by its superior ability to impute missing values in gene expression data.

Using response time of individual genes, we identify relations between gene function and their response timing, showing, for example, how cytosolic ribosomal genes are only repressed after mitochondrial ribosom is activated. We further demonstrate a strong relation between the binding affinity of a transcription factor and the activation timing of its targets, suggesting that graded binding affinities could be a widely used mechanism for controlling expression timing.

Bottlenecks and Hubs in Inferred Networks are Important for Virulence in Salmonella typhimurium

Jason E. McDermott[1], Ronald C. Taylor[1], Hyunjin Yoon[2], Fred Heffron[2]

[1] *Computational Biology & Bioinformatics Group, Pacific Northwest National Laboratory;*
[2] *Department of Molecular Microbiology and Immunology, Oregon Health & Science University*

Recent advances in experimental methods have provided sufficient data to consider systems as large networks of interconnected components. High-throughput determination of protein-protein interaction networks has led to the observation that topological hubs, highly connected proteins, and bottlenecks, proteins defined by high centrality in the network, are enriched in proteins with systems-level phenotypes such as essentiality. Global transcriptional profiling by microarray analysis has been used extensively to characterize systems, for example, examining cellular response to environmental conditions and effects of genetic mutations. These transcriptomic datasets have been used to infer regulatory and functional relationship networks based on co-regulation, but the topology of these networks and its relation to essentiality or other phenotypic changes has not been fully investigated

We use the context likelihood of relatedness (CLR) method to infer networks from two transcriptomic datasets gathered from the pathogen *Salmonella typhimurium*; one under a range of environmental culture conditions and the other from deletions of 15 regulators found to be essential in virulence. Bottleneck and hub genes were identified from these inferred networks and we show that these genes are significantly more likely to be essential for virulence than their non-bottleneck or non-hub counterparts. The observed enrichment of essential genes in bottleneck and hub nodes from inferred networks compared very favorably with previous observations from experimentally-derived networks. Networks generated using simple similarity metrics (correlation and mutual information) did not display this behavior. Overall this study demonstrates that topology of networks inferred from global transcriptional profiles provides information about the systems-level roles of bottleneck genes. Analysis of the differences between the two CLR-derived networks suggests that the bottleneck nodes are either mediators of transitions between system states or sentinels that reflect the dynamics of these transitions.

Querying pathways in protein interaction networks based on hidden Markov models

Xiaoning Qian[1,2], Sing-Hoi Sze[3,4], Byung-Jun Yoon[1]

[1]Dept. of Electrical & Computer Engineering; [2]Dept. of Statistics; [3]Dept. of Computer Science; [4]Dept. of Biochemistry & Biophysics.

High-throughput techniques for measuring protein interactions have enabled the systematic study of complex protein networks. Comparing the networks of different organisms and identifying their common substructures can lead to a better understanding of the regulatory mechanisms underlying various cellular functions. To facilitate such comparisons, we present an efficient framework based on hidden Markov models (HMMs) that can be used for finding homologous pathways in a network of interest. Given a query path, our method identifies the top k matching paths in the network, which may contain any number of consecutive insertions and deletions. We demonstrate that our method is able to identify biologically significant pathways in protein interaction networks obtained from the DIP database, and the retrieved paths are closer to the curated pathways in the KEGG database when compared to the results from previous approaches. Unlike most existing algorithms that suffer from exponential time complexity, our algorithm has a polynomial complexity that grows linearly with the query size. This enables the search for very long paths with more than 10 proteins within a few minutes on a desktop computer.

Computational Prediction of Genes Associated with B Cell Development

D. Sahoo[1], J. Seita[3], D. Bhattacharya[3], IL. Weissman[3], SK. Plevritis[4], DL. Dill[2,]

[1]Dept of Electrical Engineering, Stanford University, Stanford, CA, 94305, USA, [2]Dept of Computer Science, Stanford University, CA, 94305, US, [3]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, CA, 94305, USA, [4]Dept of Radiology, Stanford University, CA, 94305, USA

Precursor cells differentiate to their terminal progeny through a series of developmental intermediates and a network of gene expression changes that gradually establish lineage commitment and the identity of the mature cell type. The identification of genes that are involved in this process has largely been dependent upon the physical isolation and characterization of gene expression patterns within these developmental intermediates. Thus, when intermediate steps are unknown for a particular cellular differentiation pathway, the identification of genes that are developmentally regulated in that pathway can be difficult.

We present a new method to discover developmentally regulated genes using large repositories of publicly available microarray data. The method uses Boolean implications (if-then rules) mined from existing data and conservation across species to predict genes whose expression is established as precursor cells differentiate to their progeny in the absence of data from cells at intermediate steps of development, which have not been previously identified. The algorithm was validated by applying it to B cell development, for which most developmental stages between hematopoietic stem cells and fully differentiated B cells are known and can be isolated. The algorithm predicted 19 genes that are expressed after the c-kit+ progenitor cell stage and remain expressed through CD19+ mature B cells. These predictions were validated by sorting 12 mouse hematopoietic populations at different stages of development ranging from hematopoietic stem cells to germinal center B-cells using fluorescence activated cell sorting. Empirical measurement of the expression of 14 of these predicted genes using qRT-PCR confirmed that the expression of 12 of these genes is indeed stably established during B-cell differentiation.

Since CD19 expression begins rapidly after KIT expression turns off, there are relatively few developmentally regulated genes that can be identified in the intervening stages. In order to develop a more comprehensive list of B cell precursor genes, we used the combined expression of both CD19 and AICDA as a seed because high expression levels of both are specific to germinal center B cells[8], which are developmentally downstream of mature B cells. Through combined inclusion of CD19 and AICDA, we expanded our list of predicted developmentally regulated genes to 62. Thirty-three of these genes have been ablated in mice, and 18 of these mutants have been reported to have defects in B cell function or development. Of the remaining 15 knockout strains, it is possible that B cell defects exist but have not been extensively evaluated. The expression patterns of five more selected genes were confirmed using qRT-PCR experiments. This study validated our prediction algorithm on B cell development. Based on this outcome, the method appears to have the potential to predict novel genes that are expressed early in other developmental pathways that are less well characterized than B cells.

Centrality of TCA cycle in the E. coli metabolism: interplay between specialized and general use network components

Andre Levchenko[1,2], Violeta Beleva[1]

[1]*Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA;*

The structures of various metabolic networks have long been the foci of analysis related to regulation of cell function. In particular, it has been suggested that many large scale metabolic networks can be described as scale-free, modular and hierarchical in nature, following the expected properties for robust, fault-tolerant systems, including those designed by humans. The modules of the large scale metabolic networks appear to correspond to functional pathways responsible for various catabolic and anabolic processes. The connectedness of metabolites within a network implies a large degree of inter-convertibility of the parts of metabolites, down to specific atoms. This also implies that information, e.g., in the form of a radio-labeled carbon atoms, can propagate over much of the network, given enough time and expression of the appropriate enzymes. Are any components of metabolic networks central in some sense, e.g. in the sense of how many reaction steps might be needed to reach most other network components? Here we report on the results of an analysis suggesting that there indeed is a central component to the *E. coli* metabolic network and that investigation of centrality of this component can suggest important insights into the network organization.

In our analysis, we studied how closeness centrality of the nodes of the large scale metabolic network of *E. coli* scales with the distance from various nodes or pathways (the ‘root nodes of pathways’) within the network. The scaling invariably displayed a peaked distribution, unless the root nodes belonged to the TCA (also known as the tricarboxylic acid, cytric acid or Krebs) cycle. In the latter case, taking the directedness of the network into account, the distributions of In-closeness centrality as a function of the distance from the TCA cycle and Out-closeness centrality as a function of the distance to the TCA cycle were monotonic. This result suggested that the TCA cycle holds a special central place within the network in terms of the information propagation. Interestingly, we further found that the centrality dependencies on the distance from the TCA cycle or other network components could be broken down into well defined parts, each corresponding to distinct identifiable paths within the metabolic network. The centrality-network distance scaling of these parts carried a specific signature suggestive of two general types of components: One – a generic network component corresponding to the central well-interconnected network of reversible reactions, and the other – a collection of more specialized and frequently irreversible metabolic pathways. We suggest that this network organization provides metabolic processes with regulatory flexibility and robustness required for successful function in complex environments.

Linking the dynamic proteome response of individual cells to their fate

E. Eden*[1], A. A. Cohen*[1], N. Geva-Zatorsky*[1], M. Frenkel-Morgenstern[1], I. Issaeva[1], A. Sigal[2], R. Milo[3], C. Cohen-Saidon[1], Y. Liron[1], Z. Kam[1], L. Cohen[1], T. Danon[1], N. Perzov[1], U. Alon[1]

[1]Department of Molecular Cell Biology, Weizmann Institute of Science; [2]Division of Biology, California Institute of Technology; [3]Department of Systems Biology, Harvard Medical School; * ***These authors contributed equally to this work***

Individual cells do not all respond to a drug in the same way. This variability is important because it can lead to different timing of responses, and even to the escape of a small fraction of cells from the drug action. However, little is known about the molecular response to drugs at the individual cell level. Here, we study the dynamics and variability of the proteome response of human cancer cells to a chemotherapy, camptothecin, at high temporal and spatial resolution. For this, we present a novel experimental and computational approach that can automatically track individual cells and measure the levels and localizations of nearly 1000 different proteins expressed from their endogenous chromosomal position in space and time. Furthermore, our approach can couple the protein behavior of an individual cell to its fate (survival or death).

We observe, that for most proteins, nearly all cells in the population showed similarly shaped profiles of protein dynamics, rising and falling together. Diverging from this norm were 2-3% of the proteins which displayed a special behavior. At first, they showed the typical variability with similar dynamics in each cell. Then, at about 20h after drug addition, the cell population began to show dramatic cell-cell differences in the dynamics of these proteins. Some cells showed an increase in their protein levels, while other stayed constant or showed a decrease, demonstrating a bimodal dynamical behavior. We found that two of the bi-modal proteins, the RNA-helicase DDX5 and the replication factor RFC1, showed behavior that correlated with cell fate on the individual cell level. Furthermore, knock-down of DDX5 coupled with chemotherapy increased cell death by two fold further supporting its potential role in cell survival.

The ability to view the dynamic response of the proteome in individual cells over time provides a new window into human cell biology, and opens the way for understanding how seemingly identical cells show different responses to signals and drugs.

Stochastic population-based model captures PPARβ/δ-dependent transcriptional cycling

Tatjana Degenhardt[1], Katja N. Rybakova[2], Aleksandra Tomaszewska[1], Martijn J. Moné[2], Hans V. Westerhoff[3,4], Frank J. Bruggeman[2,4] and Carsten Carlberg[1,5]

1Department of Biochemistry, University of Kuopio, 70211 Kuopio, Finland,
2Molecular Cell Physiology, Institute of Molecular Cell Biology, BioCentrum Amsterdam,Vrije Universiteit, 1081 HV Amsterdam, The Netherlands,
3Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester, M1 7ND, UK,
4Netherlands Institute for Systems Biology&Multiscale Modelling and Nonlinear Dynamics, Centre for Mathematics and Computer Science, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands,
5Life Sciences Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg

A large number of proteins concertedly catalyze eukaryotic transcription initiation and monitor its progress. Nuclear receptor-mediated transcription initiation requires progressive chromatin remodeling, loop formation and polymerase preprocessing executed by a plethora of proteins. Large distinct protein complexes are formed both at the response element (RE) and the transcription start site (TSS) that besides other activities sequentially remodel chromatin, catalyze cofactor exchange, and pre-process RNA polymerase II. Elongation can commence after completion of loop formation between the distant chromatin regions of the pre-processed REs and TSSs. We study the progression dynamics of transcription initiation at the human *pyruvate dehydrogenase kinase* (*PDK) 4* gene, which has single peroxisome proliferator-activated receptor (PPAR) β/δ RE that can be selectively perturbed by the addition of synthetic ligand. We monitor the state of the RE and TSS (using chromatin immunoprecipitation (ChIP) for multiple cofactors and histone modifications), loop formation (chromatin conformation capture) and the synthesized *PDK4* transcript (qPCR). We observe periodicity in state changes and the level of *PDK4* transcript. Hierarchical clustering of the normalized ChIP time courses resulted in the identification of two protein complexes. RNA interference of a subset of co-factors allowed us to deduce that these complexes carry out distinct functions. Using computational modeling we show that the times scales of the dynamic measurements at the cellular population level are in accordance with the expected time scales occurring at the level of a single *PDK4* gene. We use stochastic theory and simulations of a realistic transcription-initiation model to explain that both the scaffold-nature of the RE and TSS and the sequential nature of transcription initiation allow cells in a population to display synchronized stochastic dynamics upon simultaneous stimulation by ligand. This passive single-time synchronization slowly desynchronizes as function of the number of cycles of transcription initiation with a rate that is inversely proportional to the number of sequential reaction steps in mechanism for transcription initiation. Interestingly, we determined that one promoter activity cycle caused multiple polymerases to commence elongation. We explain how the stochastic activity of a single gene can lead to periodicity discernible at the cell population level, showing that transcription initiation behaves as a reliable stochastic timer despite its inherent noise.

Reaction Motifs and Functional Modules in Protein Interaction Networks

John J. Tyson

Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg VA 24061 USA

The physiological responses of a cell to alterations of its internal or external state are governed by molecular regulatory networks involving genes, proteins and metabolites. These reaction networks are extremely complex, and it is difficult to comprehend how they process information, determine the cell's best course of action, and initiate appropriate responses. Understanding would come easier if we could, in principle, decompose a complex network into simple subsystems with identifiable roles. I will investigate this idea theoretically for an ideal protein interaction network with N nodes (proteins 1, 2, 3, … N) and an interaction matrix aij which specifies the effect of protein i on protein j (aij = +1 if i activates j, -1 if i inhibits j, 0 if no effect). For N=3, and with the proviso that aii = -1 (no direct autocatalysis), I will present a classification scheme for reaction 'motifs' (simple patterns of activation and inhibition among N proteins) and show that these motifs have identifiable dynamical functions within regulatory networks (i.e., they are functional 'modules'). Some characteristic modules are: toggle switch, oscillator, sniffer, cock-and-fire trigger, and hysteresis loop. I will illustrate all these motifs/modules operating in the molecular machinery governing cell growth, division and death.

References:

Novak & Tyson (2008) Nature Reviews Molecular Cell Biology, to appear.

Tyson & Novak (2008) Current Biology 18:R759-R768.

Tyson et al. (2003) Current Opinion in Cell Biology 15:221-231.

Toward Accurate Reconstruction of Functional Protein Networks

Nir Yosef[1], Lior Ungar[2] , Einat Zalckvar[3] , Adi Kimchi[3] , Martin Kupiec[2] , Eytan Ruppin[1,4] and Roded Sharan[1]

[1]*School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel.*

[2]*Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel-Aviv 69978, Israel*

[3]*Department of Molecular Genetics, Weizmann Institute of science, Rehovot 76100, Israel.*

[4] *School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel.*

Genome-scale screening studies are gradually accumulating a wealth of data on the putative involvement of hundreds of genes/proteins in a given cellular response or function. A fundamental challenge is to chart out the protein pathways that underlie the investigated system. Previous approaches to the problem have either employed a local optimization criterion, aiming to infer each pathway independently (typically via a shortest path computation), or a global criterion, searching for the most parsimonious subnetwork connecting the proteins involved. Here we study the tradeoff between the local and global approaches and present a new intermediary reconstruction scheme that provides explicit control over it. We demonstrate its utility in the analysis of the apoptosis network in human, and the telomere length maintenance system in yeast. Our results show that in the ma jority of real-life cases, the intermediary approach leads to the best description of the functioning of the system at hand. From a biological standpoint, the reconstructed apoptosis and TLM networks are comprehensively validated and analyzed, leading to new biological insights. In particular, we use a new set of perturbation experiments measuring the role of essential genes in telomere length regulation to study the TLM network. Surprisingly, we find that the proteasome plays an important role in telomere length regulation through its associations with transcription and DNA repair circuits. These findings demonstrate how network based descriptions can uncover complex and indirect functional relations.

Predicting Protein Complexes from PPI Data: A Core-Attachment Approach

Henry C.M. Leung[1] , Qian Xiang[2], S.M. Yiu[1], Francis Y.L. Chin[1]

[1] Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong;
[2] Department of Electronic & Communication Engineering, Sun Yat-Sen University, P.R. China

Protein complexes play a critical role in many biological processes. Identifying the component proteins in a protein complex is an impor-tant step in understanding the complex as well as the related bio-logical activities. This paper addresses the problem of predicting protein complexes from protein-protein interaction (PPI) network of one species using a computational approach. Most of the previous methods rely on the assumption that proteins within the same com-plex would have relatively more interactions. This translates into dense subgraphs in the PPI network. However, all existing software tools have limited success. Recently, [Gavin et al. 2006] provided a detailed study on the organization of protein complexes and sug-gested that a complex consists of two parts: a core and an attach-ment. Based on this core-attachment concept, we developed a novel approach to identify complexes from PPI network by identifying their cores and attachments separately. We evaluated the effectiveness of our proposed approach using three different datasets and com-pared the quality of our predicted complexes with three existing tools. The evaluation results show that we can predict many more com-plexes and with higher accuracy than these tools with an improvement of over 30%. To verify the cores we identified in each complex, we compared our cores with the mediators produced by [Andreopoulos et al. 2007], which were claimed to be the cores, based on the benchmark result produced by [Gavin et al. 2006]. We found that the cores we produced are of much higher quality ranging from 10-fold to 30-fold more correctly predicted cores and with better accuracy.

Availability: http://alse.cs.hku.hk/complexes/

Systems-level evidence of transcriptional co-regulation of yeast protein complexes

J. William Lee[1], Tomasz Zemojtel[2], Eugene Shakhnovich[1]

[1]Dept. of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA; [2]Dept. of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

Protein complexes are functional units in the cell whose formation is regulated by a number of mechanisms. Previous studies have suggested that transcriptional co-regulation of protein components play a minor role in assembly of individual complexes. Recent developments in experimental approaches and bioinformatic analyses offer an opportunity to investigate the cell at a global or systems level with better quality. Here we show that, using genome-/proteome-wide yeast data of protein complexes and protein-DNA interactions in normal growth conditions, a contribution of transcriptional co-regulation to global formation of protein complexes is statistically significant at a systems level. A significant fraction of proteins that form complexes shares common transcription factors on a global scale. We provide evidence for the significant contribution by quantifying transcriptional co-regulation of complexes as a global topological pattern of an integrated network. Our co-regulation measure for a complex takes into account both the absolute fraction of co-regulated protein components and their probabilistic scores from the hypergeometric distribution for all transcription factors. Biological relevance is examined by functional analysis of highly co-regulated complexes. Our finding indicates that there exists linear causality at a systems level between transcriptional co-regulation of genes and formation of multi-protein complexes in the unicellular organism.

References

[1] Gavin *et al.* (2006). **Proteome survey reveals modularity of the yeast cell machinery**, *Nature* **440**: 631 - 636

[2] Krogan *et al.* (2006). **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.**, *Nature* **440**: 637 - 643

[3] Reguly *et al.* (2006). **Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae**, *Journal of Biology* **5**(4): 11

[4] Harbison *et al.* (2004). **Transcriptional regulatory code of a eukaryotic genome**, *Nature* **431**(7004): 99-104

[5] Simonis *et al.* (2004). **Transcriptional regulation of protein complexes in yeast**, *Genome Biol* **5**: R33

[6] Yeger-Lotem *et al.* (2004). **Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction**, *Proc Natl Acad Sci USA* **101**: 5934 - 5939

[7] Zhang *et al.* (2005). **Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network**, *Journal of Biology* **4**(2): 6

Revealing Biological Modules via Graph Summarization

Saket Navlakha[1,2], Michael C. Schatz[1,2], Carl Kingsford[1-3]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park; [2]Department of Computer Science, University of Maryland, College Park; [3]Institute for Advanced Computer Studies, University of Maryland, College Park

The division of a protein interaction network into biologically meaningful modules can aid with automated complex detection and prediction of biological processes and can uncover the global organization of the cell. We propose a novel graph summarization (GS) technique, based on graph compression, to cluster protein interaction graphs into biologically relevant modules. The method is motivated by defining a biological module as a set of proteins that have similar sets of interaction partners. We show this definition, put into practice by a GS algorithm, reveals modules that are more biologically enriched than those found by other methods. We also apply GS to predict complex memberships, biological processes, and co-complexed pairs and show that in most settings GS is preferable over existing methods of protein interaction graph clustering.

Network Based Prediction of Human Tissue-Specific Metabolism

Tomer Shlomi[1†], Moran N. Cabili[1†], Markus J. Herrgård[3], Bernhard Ø. Palsson[3], Eytan Ruppin[1,2]

[1] School of Computer Science, Tel-Aviv University, [2] School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel; [3] Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093-0412, USA; [†] These authors contributed equally to this work

Research into human metabolism and its regulation has expanded rapidly due to the emergence of metabolic diseases such as diabetes and obesity as major sources of morbidity and mortality, with metabolic enzymes and their regulators increasingly emerging as viable drug targets. While much work has been done in the context of applying constraint-based modeling to study the metabolism of micro-organisms, large-scale modeling of human metabolism is still in its infancy. A fundamental step forward has been presented in recent studies by Duarte et al. and by Ma et al. that reconstructed the global human metabolic network based on an extensive evaluation of genomic and bibliomic data. The resulting network models are however non tissue-specific. The task of adapting constraint-based modeling methods from the realm of micro-organisms to that of multi-cellular organisms encounters two main hurdles: One major difficulty relates to the fact that different tissues have different metabolic objectives that are not well characterized and are largely unknown. Another major difficulty relates to the lack of information on tissue-specific metabolite uptake and secretion, which is essential for employing flux balance analysis. Here we present the first computational method that successfully obtains a large-scale, tissue-specific description of human metabolism. Our approach is based on integrating tissue-specific gene and protein expression data with an existing comprehensive reconstruction of the global metabolic network. Tissue-specific variations in enzyme expression levels are treated not as the final determiners of enzyme activity, but as cues for their likelihood of carrying metabolic flux. Network integration is then used to accumulate these cues into a global, consistent metabolic behavior, which reflects the outcome of putative post-transcriptional regulatory effects. Our method is first validated by predicting the metabolic state of the yeast S. cerevisiae under conditions for which reliable expression and flux data is readily available. Then, the method is applied to the genome-scale human metabolic network model of Duarte et al, integrated with tissue-specific enzyme expression data to predict tissue-specific metabolic behavior of 10 human tissues. The analysis reveals that post-transcriptional regulation plays a central role in shaping tissue-specific metabolic activity profiles, covering 18% of metabolic genes. In order to validate the predicted tissue-specific metabolic behavior, we relied on various data sources for tissue-specificity of genes, reactions and metabolites. In all cases, the predicted tissue-specificity was significantly correlated with these datasets, with the precision and recall varying between 0.36-0.7 and 0.37-0.55, respectively. Specifically, the prediction accuracy of disease-related gene-tissue associations is highly statistically significant, with precision of 0.49 and recall of 0.55. Overall, the method presented here lays the foundation for the rapid development of human tissue-specific metabolic models and is likely to advance the computational study of human metabolic disorders.

FROM STEADY STATE TO KINETICS: A DYNAMIC MODEL OF THE GLUCOSE TRANSPORT REGULATORY NETWORK IN S. cerevisiae.

Sooraj Kuttykrishnan[1], Jeff Sabina[2, 3], Mark Johnston[2, 3] Michael Brent[1-3]

[1]Department of Computer Science, Washington University in St. Louis; [2]Center for Genome Sciences, Washington University in St. Louis; [3]Department of Genetics, Washington University in St. Louis.

The topology of several small gene regulatory networks is fairly well understood. Using data on mRNA degradation rates, protein degradation rates, and a handful of measurements of steady state mRNA levels, it is now possible to move from topology to detailed kinetic models in which most parameters are directly measured.

Here we present the glucose transport regulatory network in *S. cerevisiae* as a case study for this approach. There is a fairly well established and mostly complete wiring diagram for this network. We model each regulatory interaction in the wiring diagram as a biochemical reaction and use an appropriate differential equation to express the kinetics of the reaction.

A typical transcription reaction can be written as:

$$\text{DNA} \xrightarrow{TF_1,\ TF_2} \text{mRNA}_1 + \text{DNA}$$

We modeled such reactions in the network using differential equations of the following functional form (1), where k is the degradation rate of mRNA$_1$.

$$\frac{d[mRNA_1]}{dt} = V_{max}\left(\frac{1}{1+\theta_1[TF_1]}\right)\left(\frac{1}{1+\theta_2[TF_2]}\right) - k[mRNA_1]$$

Degradation rates of mRNAs (and proteins) were estimated from published, high-throughput data. We then estimated V_{max} in these equations directly from steady state levels of mRNAs by inactivating all transcription factors in the model through a combination of gene knockouts and an inactivating external glucose condition. With ks and V_{max}s in hand, we could estimate Θs by restoring the activities of transcription factors one at a time. Finally, 10 (out of a total of 90) parameters were optimized to match steady state levels of mRNAs for glucose transporters in wild type yeast.

Thus, we have demonstrated the relative simplicity of estimating most parameters in a complex model by systematic application of quantitative PCR and gene knockouts. This approach promises much more realistic models than those obtained by fitting a large number of parameters to observations of the very kinetic behavior one hopes to predict. This approach has brought us to an integrated kinetic model of the glucose transport regulatory network that is ready to be validated by obtaining kinetic measurements not used to estimate its parameters.

1. *Computational and Experimental Approaches for Modeling Gene Regulatory Networks.* **Goutsias, J., Lee, N. H.** May 2007, Current Pharmaceutical Design, Vol. 13, pp. 1415-1436.

From personal genomes and enviromes to traits

George M. Church[1]

[1] *Harvard Medical School*

Relative to a reference human genome, a personal genome shows about 10,000 DNA variations which affect protein structure and 3 million which do not. While association studies of common DNA variants with diseases mostly yield, so far, weak predictive power and few causative alleles, most researchers expect that this will be soon remedied by genome-wide sequencing and aggregating alleles by system functions. Second-generation sequencing (e.g. Polonator.org -- open-source hardware, software, wetware) has brought costs down by over 10-fold per year for 4 years (from $100M to $5K), but this needs to improve further in cost and interpretability – e.g. by targeted sequencing of coding variants (~1% of the genome) plus analysis of regulatory variants via RNA quantitation by sequencing, and environmental components via microbiome and VDJ-ome. Haplotypes and allele specific expression should help establish causative links and improve association studies. PersonalGenomes.org is a unique effort to integrate personal genomes with comprehensive sets of medical and non-medical traits and environmental measures and share these in an open-access format. To assess personal variation in RNAs in a broad set of cell-types, we establish pluripotent stem cells from skin. We have IRB approval to expand our current cohort to 100,000 volunteers.

Evolutionary Systems Biology in Health and Disease: Aging and Cancer vis-à-vis Robustness and its Breakage

Aviv Bergman

Department of Systems & Computational Biology

Albert Einstein College of Medicine

Our long-term goal is a complete understanding of the functional properties of genes involved in biological processes that control key aspects of complex developmental, and pathophysiologic processes. We study these properties in their systems biological and evolutionary context, and utilize our findings to better understand the underlying systems level mechanisms of complex traits. In turn, we may be able to develop reliable diagnostic tools aimed at predicting the behavior of complex traits in response to perturbation. The evolutionary causes and consequences of robustness and its breakage will be explored in relation to health (healthy aging), and disease (head and neck squamous cell carcinoma).

Effects on insertions and deletions (indels) on wirings in protein-protein interaction networks: a large-scale study

Fereydoun Hormozdiari [1,*], Michael Hsing[2,*], Rahele Salari[1,*], Alexander Schoenhuth[1], Simon K. Chan[2], S. Cenk Sahinalp[1], Artem Cherkasov[2]

[1]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada; [2] Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada; [*]Joint first authorship

Although insertions and deletions (indels) are a common type of sequence variation, their origin and their functional consequences have not yet been fully understood. It has been known that indels preferably occur in the loop regions of the affected proteins. Moreover, it has recently been demonstrated that indels are significantly more strongly correlated with functional changes than substitutions. In sum, there is substantial evidence that indels, not substitutions, are the predominant evolutionary factor when it comes to structural changes in proteins.

As a consequence it comes natural to hypothesize that sizable indels can modify protein interaction interfaces, causing a gain or loss of protein-protein interactions, thereby significantly rewiring the interaction networks. In this paper, we have analyzed this relationship in a large-scale study. We have computed all paralogous protein pairs in S.cerevisiae (Yeast) and D.melanogaster (Fruit Fly) and sorted the respective alignments according to whether they contained indels of significant lengths as per a pair HMM based framework of a recent study. We subsequently computed well known centrality measures for proteins that participated in indel alignments indel proteins and those that did not. We found that indel proteins indeed showed greater variation in terms of these measures. This demonstrates that indels have a significant influence when it comes to rewiring of the interaction networks due to evolution, which confirms our hypothesis. In general, this study may yield relevant insights into the functional interplay of proteins and the evolutionary dynamics behind it.

Adaptive Environmental Conditioning: Preparation to Subsequent Environmental Stimuli in Eukaryotic and Prokaryotic Microorganisms.

Amir Mitchell[1], Gal H. Romano[2], Martin Kupiec[2], Orna Dahan[1], and Yitzhak Pilpel[1]

[1]*Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.*

[2]*Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Israel*

Living organisms constantly react to their surroundings in order to maintain their internal environment. In unicellular organisms external conditions can dramatically fluctuate, hence evolution selects cells that are well adapted to a wide range of environmental conditions. Although the adaptation to the various stimuli has been studied extensively over the years, little attention has been given to clarify the inter-dependencies between these stimuli. In reality, some ecological niches expose organisms to reoccurring sets of stimuli such that the appearance of an early stimulus may serve as a predictor for consequent ones. We propose that these habitats offer organisms the opportunity to prepare in advance for the next environmental change rather than merely respond to the change upon encounter. We term a regulatory strategy that captures the unidirectional temporal order of stimuli “Environmental Adaptive Conditioning”, in analogy to Pavlovian Classical Conditioning.

We develop a mathematical model to study the key forces that can select for Adaptive Conditioning. The model is used to predict the actual fitness gained from this regulation strategy under different environmental setups. The predictions of the model are experimentally tested and validated in a specific environment, exposure of *E. coli* to lactose with a preceding signal. We observe Environmental Adaptive Conditioning both in *S. cerevisiae* and in *E. coli* under conditions found in their natural habitats, the switch from fermentation to respiration and the passage through the digestive track, respectively. As expected by our hypothesis we observe that exposure to stresses that occur during yeast fermentation enhances their ability to cope with stresses typical of respiration. We find that this phenotype is accomplished by wiring of the transcription network. In *E. coli* we observe that a carbon source that is encountered early in the digestive track induces metabolic pathways that would be required only further down the track. Considering the great evolutionary distance between the two organisms, we expect that the ability to “predict” future environmental changes has evolved in additional Eukaryotic and Prokaryotic microorganisms.

Spatio-temporal coding of signal specificity and network reconstruction

Boris N. Kholodenko

Department of Pathology, Anatomy and Cell Biology, Thomas Jefferson University, Philadelphia, PA 19107

Cells respond to countless external cues using a limited repertoire of interconnected pathways activated by G protein-coupled receptors and receptor tyrosine kinases (RTKs). Recent discoveries changed our perception of the signal specificity, suggesting that this specificity is encoded by the spatial and temporal dynamics of downstream signaling networks. The ErbB receptors are the gatekeepers of a multilayered signal transduction network that converts external stimuli into specific gene expression responses and cell fate decisions. By combining computational modeling with experiments, we show how two ErbB receptor ligands, epidermal growth factor (EGF) and heregulin (HRG), induce distinct all-or-nothing responses of the transcription factor c-Fos by activating the extracellular regulated kinase (ERK) pathway. Although EGF and HRG induce transient versus sustained ERK activation in the cytoplasm, the nuclear ERK activity and the resulting *c-fos* mRNA expression are transient for both ligands owing to induced expression of nuclear dual-specificity phosphatases (DUSP). Our results demonstrate that the distinct c-Fos responses arise from ligand-dependent, spatiotemporal control of ERK activity emerging from DUSP-mediated negative feedback and cytoplasmic-signaling-to-protein-expression feedforward loops.

We also exploited experimental and computational approaches to reveal how a concordant interplay between the insulin and EGF signaling networks can potentiate mitogenic signaling. Computational modeling unveils that insulin endows the mitogenic EGFR pathway with robustness to perturbations of critical network nodes. Our results show the feasibility of using computational models to predict complex cellular responses and identify therapeutic targets.

In addition to bottom-up modeling, a top-down approach to inferring the structure of cellular regulatory networks is presented. Rapid advances in genomics and proteomics have enabled the acquisition of data on the expression of thousands of genes and the functional state of hundreds of proteins. However, there are no methods capable of providing quantitative interpretations of genomics and proteomics data sets in a manner that unravels the wiring of cellular machinery. Here, we propose a novel strategy of unraveling functional interactions in cellular signaling and gene networks. We demonstrate how dynamic connections leading to a particular module (e.g., an individual gene/protein or a cluster) can be retrieved from experimentally measured network responses to perturbations influencing other modules.

Large scale identification of MAPK networks in Arabidopsis thaliana

Sorina C. Popescu[1], George V. Popescu [2], Michael Snyder [3,4] , S. P. Dinesh-Kumar[3]

[1]Boyce Thompson Institute for Plant Research, Ithaca NY; [2]ACM; Departments of Molecular, Cellular and Developmental Biology[3], and Molecular Biophysics and Biochemistry[4], Yale University, New Haven, CT

Signaling through mitogen-activated protein kinase (MAPK) cascades is a conserved and fundamental process in all eukaryotes. However, only a limited repertoire of MAPK phosphorylation substrates has been revealed so far. We report here a large scale analysis of the *Arabidopsis thaliana* MAPK signaling pathways using a hierarchical approach that reconstructs multiple levels of the cellular signaling network. We have identified *in vitro* MAPKK and MAPK phosphorylation substrates, and we have constructed a MAPKK/MAPK/Effector phosphorylation network including previously known and novel signaling pathways. The first level of the network is represented by the MAPKK->MAPK tier of the phosphorylation cascades. We have identified 48 MAPKK/MAPK functional modules of nine MAPKKs and ten MAPKs combinations *in vitro,* using a combinatorial approach, and subsequently verified them *in vivo*. The second level of network contains the phosphorylation substrates identified by probing protein microarrays containing a significant fraction of transcription factors, protein kinases, and DNA and RNA binding proteins, with all ten *in vivo* activated *Arabidopsis* MAPKs. We have identified 570 putative MAPK substrates, spanning a variety of Arabidopsis protein families, including transcription and translation factors, protein kinases, metabolic enzymes, and protein with unknown function. Additionally, *in planta* reconstitution experiments validated selected MAPK pathways and effectors.

A complex image of signaling pathways emerges from our analysis. The structure of our reconstructed signaling network is significantly different from the signaling modules view, supporting the hypothesis of a combinatorial control of transcription through selective phosphorylation of a large number of transcription factors. According to our data, MAPKs phosphorylated network is denser than expected for a simple modular signal transduction network. Despite the fact that additional co-factors and scaffold proteins may regulate signal transduction, the number of phosphorylation events departs significantly from a modular signal transduction network. Functional analysis reveals that *Arabidopsis* MAPKs phosphorylate mostly transcriptional regulators and are associated with developmental and stress response cellular processes. Our *in vitro* reconstructed signaling pathways are significantly overrepresented in transcription activity, response to stress, response to biotic and abiotic factors and development processes.

Finally, our global analysis of MAPK signaling removes some of the bias in previous studies of signaling networks guided by genotype/phenotype association methods and reveal potential new roles of MAPK signaling in cellular control. From the reconstructed network perspective, transcription regulation appears to be the result of a superposition of several common pathways and subsets of highly specific signaling pathways activated by combinations on MAPKK and MAPK. The integrated analysis of signaling and control pathways is the starting point for further studies of transcription regulation.

Learning signaling network structures with sparsely distributed data

Karen Sachs[1,7], Solomon Itani[2,3], Jennifer Carlisle[2,3], Garry P. Nolan[1,7], Dana Pe'er[6] and Douglas A. Lauffenburger[2,4,5.]

[1]Stanford University School of Medicine; [2]Massachusetts Institute of Technology; [3]Electrical Engineering and Computer Science; [4]Biological Engineering; [5]Chemical Engineering; [6]Columbia University, Biology Department; [7]Baxter laboratory for Genetic Pharmacology.

Flow cytometric measurement of signaling protein abundances has proved particularly useful for elucidation of signaling pathway structure. The single cell nature of the data ensures a very large dataset size, providing a statistically robust dataset for structure learning. Moreover, the approach is easily scaled to many conditions in high throughput. However, the technology suffers from a dimensionality constraint: at the cutting edge, only about twelve protein species can be measured per cell, far from sufficient for most signaling pathways. Because the structure learning algorithm (in practice) requires that all variables be measured together simultaneously, this restricts structure learning to the number of variables that constitute the flow cytometer's upper dimensionality limit. To address this problem, we present here an algorithm that enables structure learning for sparsely distributed data, allowing structure learning beyond the measurement technology's upper dimensionality limit for simultaneously measurable variables. The algorithm assesses pairwise (or n-wise) dependencies, constructs 'Markov neighborhoods' for each variable based on these dependencies, measures each variable in the context of its neighborhood, and performs structure learning using a constrained search.

Reverse engineering molecular regulatory networks from microarray data with qp-graphs

Robert Castelo[1,2], Alberto Roverato[3]

[1]Pompeu Fabra University, Barcelona, Spain; [2]Municipal Institute for Medical Research, Barcelona, Spain; [3]University of Bologna, Bologna, Italy.

Reverse engineering bioinformatics procedures applied to high-throughput experimental data have become instrumental to generate new hypotheses about molecular regulatory mechanisms. This has been particularly the case for gene expression microarray data where a large number of statistical and computational methodologies have been developed in order to assist in building network models of transcriptional regulation.

A major challenge faced by every different procedure is that the number of available samples *n* for estimating the network model is much smaller than the number of genes *p* forming the system under study. This compromises many of the assumptions on which the statistics of the methods rely, often leading to unstable performance figures. In this work we apply a recently developed novel methodology based in the so-called *q*-order limited partial correlation graphs, qp-graphs, which is specifically tailored towards molecular network discovery from microarray expression data with $p >> n$.

Using experimental and functional annotation data from Escherichia coli here we show how qp-graphs yield more stable performance figures than other state-of-the-art methods when the ratio of genes to experiments exceeds one order of magnitude. More importantly, we also show that the better performance of the qp-graph method on such a gene-to-sample ratio has a decisive impact on the functional coherence of the reverse-engineered transcriptional regulatory modules and becomes crucial in such a challenging situation in order to enable the discovery of a network of reasonable confidence that includes a substantial number of genes relevant to the essayed conditions.

Derivation, identification and validation of the mathematical model of IRMA: a yeast synthetic network for In vivo Reverse-engineering and Modelling Assessment in Systems and Synthetic Biology.

Marucci L.[1,2], Cantone I.[1,3], Santini S.[2], di Bernardo M.[2], Cosma MP.[1], di Bernardo D.[1,2]

1 Telethon Institute of Genetics and Medicine (TIGEM), Naples 80131, Italy. 2 Department of Computer and Systems Engineering, Federico II University, Naples 80125, Italy; 3 European School of Molecular Medicine (SEMM), Naples, Italy

We derived, identified and validated a Delay Differential Equations mathematical model for IRMA, a novel synthetic network we built in *Saccharomyces cerevisiae* for 'benchmarking' modelling and reverse-engineering methods. The network, isolated from endogenous cell genes, is composed of five genes (*CBF1* [*x1*], *GAL4* [*x2*], *SWI5* [*x3*], *GAL80* [*x4*] and *ASH1* [*x5*]) and includes one positive and one negative feedback loop and a protein-protein interaction (Gal4-Gal80). IRMA can be 'switched' on or off by culturing cells in galactose or in glucose. We performed perturbation experiments by shifting cells from glucose to galactose and vice versa and measured time-series expression profiles of genes by q-PCR. We also performed genetic perturbations by over-expressing each of the five genes. Data show a rich dynamic behaviour with delays and the appearance of damped oscillations. We then performed experimental measurements of promoters' strength. The mathematical model consists of five nonlinear delay differential equations that describe production rates of the five mRNAs concentrations, assuming Hill kinetics and proportionality between protein and mRNA levels:

$$\frac{dx_1(t)}{dt}=\alpha_1+v_1\left(\frac{x_3(t-\tau)}{(k_1+x_3(t-\tau))\left(1+\frac{x_5(t)}{k_2}\right)}\right)-d_1x_1(t);\frac{dx_2(t)}{dt}=\alpha_2+v_2\left(\frac{x_1(t)}{k_3+x_1(t)}\right)-(d_2\ -\Delta(\beta_1))x_2(t)$$

$$\frac{dx_3(t)}{dt}=\alpha_3+v_3\left(\frac{x_2}{(k_4+x_2(t))\left(1+\frac{x_4(t)}{GAL+\gamma}\right)}\right)-d_3x_3(t);\frac{dx_4(t)}{dt}=\alpha_4+v_4\left(\frac{x_3^h(t)}{k_5^h+x_3^h(t)}\right)-(d_4-\Delta(\beta_2))x_4(t)$$

$$\frac{dx_5(t)}{dt}=\alpha_5+v_5\left(\frac{x_3^h(t)}{k_6^h+x_3^h(t)}\right)-d_5x_5(t)$$

We used a phenomenological rate law to model the galactose-induced activation of Gal4 driving the transcription of *SWI5* via the *GAL10* promoter, and included a delay in the activation of *CBF1* by Swi5 via the *HO* promoter. To estimate the 26 unknown parameters we used time-series data together with a Hybrid Genetic Algorithm and an empirical exploration of the parameters' space, taking into account the measured promoter strengths. To validate the predictive power of our mathematical approach we simulated the dynamics of IRMA relative to experiments not used for the parameter estimation: we performed *in silico* the over-expressions and compared them with the *in vivo* results, showing good quantitative agreement. The model can be used to link the observed dynamics to the topology of the network; both the positive feedback loop (Swi5-Cbf1-Gal4) and the time delay are essential for the damped in the "switch on" experiment. Here we show that mathematical modelling is essential to deeply understand the complex dynamics of synthetic Gene Regulatory Networks like IRMA. Currently we are using the model to understand how to modify the network topology to achieve sustained oscillations, thus transforming our synthetic network in a robust oscillator.

Challenge 1 introduction: Variability and Robustness in T Cell Activation from Regulated Heterogeneity in Protein Levels

Ofer Feinerman[1], Joël Veiga[1], Jeffrey Dorfman[1,2], Ronald N. Germain[3] & Grégoire Altan-Bonnet[1]

[1] ImmunoDynamics Group. Program in Computational Biology and Immunology. Memorial Sloan-Kettering Cancer Center. New York NY - USA; [2] Present address: Seattle Biomedical Research Institute. Seattle, WA - USA; [3]Lymphocyte Biology Section, Laboratory of Immunology. Program in Systems Immunology and Infectious Disease Modeling (PSIIM). National Institute of Allergy and Infectious Diseases. Bethesda, MD– USA

Background. Vertebrate organisms rely on the ability of their adaptive immune system to distinguish self- from non-self agents, to fight efficiently viral or bacterial infections without endangering their own viability. This discrimination between self and non-self agents is fundamental: in the immune system, antigen-presenting cells (APC) constantly process proteins to present on their surface as peptide Major Histocompatibility Complex (pMHC). During an infection, pMHCs' derived from the pathogen are presented by APC to T cells: the immune response relies on the specific activation of T cells upon detecting these non-self (pathogen-derived) pMHCs'. However, spurious activation of T cells by self-derived pMHC must be avoided to prevent auto-immune responses. A major contribution to this pMHC discrimination is the elimination during thymic development of many immature T cells possessing T cell receptors (TCRs') that are highly reactive with self pMHCs. However, this cellular selection itself depends on the capacity of the TCR to make fine distinctions between closely related pMHC structures when transducing signals that regulate cell survival and differentiation, distinctions that also must be made by mature, post-thymic T cells. Hence, at different stages of their lifespan, T cells endowed with a given TCR must be able to perform reliable yet flexible pMHC discrimination [1].

Results. We developed a single-cell assay to monitor how endogenous variation in the expression levels of signaling proteins influences response variability [2]. We have shown how two competing feedback loops control a high gain digital amplifier that sets a threshold in terms of the quality of ligand-receptor interaction and defines self/non-self discrimination. Combining this new methodology and computer modeling to study T cell activation, we identified and characterized two key regulators of antigen-induced signaling. The CD8 coreceptor functions as an analog regulator that tunes the activation threshold, while SHP-1 phosphatase acts as a digital regulator whose level determines whether a cell is either responsive or non-responsive. Stochastic variation in the levels of these two proteins generates substantial activation response diversity among cells in a clonal population, but co-regulation in the expression of these molecules limits the extent of this effect [2]. Together, these properties of the signaling network allow T cells to have functional flexibility without sacrificing accurate discrimination between self and foreign antigens.

[1] O. Feinerman, R. N. Germain, G. Altan-Bonnet (2008) *Molecular Immunology* **45**: 619

[2] Ofer Feinerman, Joël Veiga, Jeffrey R. Dorfman, Ronald N. Germain & Grégoire Altan-Bonnet (2008) Science **321**: 1081-4.

Challenge 2 introduction: Defining signaling differences between cancerous and healthy hepatocytes with cell-specific pathway models

Julio Saez-Rodriguez[1,2,3], Leonidas G. Alexopoulos[1,2,3,5], Ben Cosgrove[2,3] Regina Samaga[4], Steffen Klamt[4], Douglas A. Lauffenburger[2,3], Peter K. Sorger[1,2,3]

[1] Harvard Medical School, Department of Systems Biology, Boston, MA [2] Massachusetts Institute of Technology, Department of Biological Engineering, Cambridge, MA [3] Cell Decision Processes Center [4]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany [5]Current Address: Department of Mechanical Engineering, National Technical University of Athens, Athens, Greece

Liver cancer is the third most common cause of cancer death. Understanding the differences in the signaling networks between primary and cancerous hepatocytes can help to identify targets to affect selectively cancer cells. Towards this end, a data compendium of around 10,000 experimental points was generated in primary hepatocytes and the liver cancer cell line HepG2. Both cell types were stimulated with 10 different pro-growth and inflammatory cytokines in combination with 7 small-molecule inhibitors targeting key signaling mediators. For all these conditions, the phosphorylation levels of 17 essential readouts were measured at 0, 30 minutes and 3 hours, as well as the release of 20 cytokines at 0, 3 and 24hrs. This data set defines the DREAM3 Signaling-Response Prediction Challenge 2.

The data was processed with *DataRail*, an open source MATLAB toolbox for managing, transforming and visualizing varied high-throughput data (Saez-Rodriguez et al., *Bioinformatics*, 24(6): 840, 2008). Multiple linear regression was used to extract the correlations between stimuli/inhibitors and signals, and between signals and cytokine releases. To obtain mechanistical insight using the knowledge accumulated over decades of research in signaling pathways, we developed a methodology to identify the logic model that optimally describes a certain data set, based on prior knowledge casted as signaling maps. We chose a Boolean (logic) formalism as it encapsulates the topology and causality of the network without dealing with kinetic parameters. We also developed an efficient algorithm to find novel connections that uses the fit of the current topology to identify potential solutions. Furthermore, we are able to design new experiments to discriminate between model variants by systematically testing new conditions in silico. These methods are embedded in *CellNetOptimizer* (CNO), an open source MATLAB toolbox that uses *CellNetAnalyzer* as simulation engine, and works in concert with *DataRail*. Based on the phosphorylation data described above and canonical pathways retrieved from curated databases, we were able to delineate mechanistic models specific for both primary and cancer hepatocytes. By comparing the models we could uncover significant differences in the signaling networks. Furthermore, data that could not be reconciled with the *a priori* knowledge suggested gaps in our current knowledge that pointed at potential new connections involving crosstalk between inflammatory signals and survival pathways.

Challenge 2 Best predictor talk: Applying Multiple Imputation to the Phospho-Proteomics Challenge

Nicolas Guex[1], Eugenia Migliavacca[2], Ioannis Xenarios[1]

[1] *Vital-IT group Swiss Institute of Bioinformatics, Lausanne, Switzerland;*

[2] *Bioinformatic Core Facility, Swiss-Institute of bioinformatics, Lausanne, Switzerland.*

DREAM3 signaling response challenge consisted in predicting the concentration of 17 phospho-proteins at two time points for 7 combinations of stimuli and inhibitors applied to normal and cancer hepatocytes. For each of the 17 phosphoproteins, 42 distinct combinations of stimuli and inhibitors measurements were available, in addition to unstimulated and uninhibited controls.

Given the complexity of the underlying regulatory networks affected by the various stimuli and inhibitors, we decided to approach this challenge based solely on the existing measurements using a multiple imputation algorithm. We took advantage of the Vital-IT high-performance computing center (www.vital-it.ch) to determine through simulations the best possible imputation parameters to apply for our final prediction.

This approach could also be applied to experimental design, probing for combination of inhibitors without the necessity to measure all possible combinations.

Challenge 2 Best predictor talk: Thoughts on the DREAM3 signal transduction challenges, and methodology of our best performance algorithms

Guillaume Bourque[1] and Neil D. Clarke[1]

[1]*Genome Institute of Singapore*

Members of the Genome Institute of Singapore participated in three DREAM3 challenges, one as the poser of the challenge (Challenge 3) and two as predictors (Challenges 1 and 2). In Challenge 1, four proteins were assayed by FACS analysis, and we were asked to assign to each protein one of eight possible labels, corresponding to nodes in a signal transduction network. We were one of six groups tied for best, with two out of the four proteins labeled correctly. In Challenge 2, we were asked to predict certain phosphoprotein or cytokine levels under particular conditions, given a large set of data for other proteins under the same conditions, and for the same proteins under different conditions; our predictions were deemed best among those submitted for this Challenge. In this talk we will discuss our approaches to Challenges 1 and 2, and give some general thoughts on the nature of these challenges and on the DREAM experiment.

Challenge 3 introduction: The challenge of predicting gene expression

Xiao Wei Xue[1], Ai Li Yeo[1], Zhen Xuan Yeo[1] and Neil D. Clarke[1]

[1] *Genome Institute of Singapore*

GAT1, GCN4, and LEU3 are all transcription factors involved, in some way, in amino acid metabolism. Challenge 3 of the DREAM3 community-wide experiment required predictors to rank 50 yeast genes by fold-expression change following an environmental perturbation (3-aminotriazole treatment) in a gat1 deletion strain. Complete time course data was provided for all genes except the 50 to be predicted. In addition, analogous data was provided for all genes in a wild-type strain, in a gcn4 deletion strain, and in a leu3 deletion strain. Publicly available data could be used in the predictions, including chromatin immunoprecipitation data for all three of the transcription factors. A naïve prediction method, based on averaging the expression levels for each of the 50 genes across the three strains for which the data was provided, correlates reasonably well with the observed expression ranks. Two groups performed substantially better than this.

Challenge 4 introduction:
Generating Realistic in silico Gene Networks for Performance Assessment of Reverse Engineering Methods

Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Laboratory of Intelligent Systems
CH-1015 Lausanne

Reverse engineering methods are typically first tested on simulated data from *in silico* networks, for systematic and efficient performance assessment, before an application to real biological networks. In this paper we present a method for generating biologically plausible *in silico* networks, which allow realistic performance assessment of network inference algorithms. Instead of using random graph models, which are known to only partly capture the structural properties of biological networks, we generate network structures by extracting modules from known biological interaction networks.

Using the yeast transcriptional regulatory network as a test case, we show that extracted modules have a biologically plausible connectivity because they preserve functional and structural properties of the original network. We provide the gold standard networks for the gene network reverse engineering challenge of the third DREAM conference (Dialogue on Reverse Engineering Assessment and Methods 2008, Cambridge, MA). Here, we disclose the method used to generate these networks and their structure.

Our Java tool used to generate the DREAM3 *in silico* gold standards is available on our website (http://lis.epfl.ch/grn).

Keywords: reverse engineering, gene regulatory networks, DREAM Challenge, modularity, network motifs

Challenge 3 Best predictor talk:
Gene-expression prediction by the elastic net

Mika Gustafsson[1], Michael Hörnquist[1]

[1]*Dept. of Science and Technology, Linköping University, Sweden. [mikgu,micho]@itn.liu.se*

To predict the differential gene expressions for new experiments is an important endeavour within computational systems biology. It can both be a way to appreciate how drugs affect the system, as well as providing a framework for finding which genes are interrelated in a certain process. Here we predict the order of the levels for 50 genes out of 9335 Affymetrix probes for one time-series out of four, with eight experiments in each. This is the Gene Expression Prediction Challenge of DREAM3.

All Affymetrix probe names are mapped onto gene names, which enable incorporation of external data sources, such as other expression data sets and TF-binding data. This mapping reduces the number of units to 7805. We assume the level to be predicted is a linear combination of all other (known) levels, given as log-ratios, together with the estimated derivatives of those levels (thereby, we have 2 x (7805 – 50) = 15510 putative predictors for each gene). The parameter fit is performed as a minimization problem supported by some method to handle the lack of data. These methods cover lasso, weighted regularized least absolute deviation (wrlad), ordinary least squares (ols) with minimal L2 norm, ridge regression and the elastic net, where the model selection criterion is a three-fold cross-validation (CV) with respect to the three complete time-series. We also explore various forms of pre-transformations of the given log-ratios, including power-, exponential-, and arctangent-transformations.

The most promising method turns out to be the elastic net, avoiding derivatives and pre-transformations. This method enforces constraints both on the L1 and L2 norms of the coefficients, and we determine both the mixture and the magnitude of these norms by CV. To take into account public data, we download from the ncbi omnibus webpage (www.ncbi.nlm.nih.gov/geo/) a compendium of 256 gene expression profiles in time-series, and 515 profiles from the Rosetta database (Hughes et al, 2000, Bernardo et al 2005). These profiles are used independently in the weighted elastic net inference, but their importance turn by CV estimate out to be significantly less than the profiles given in the challenge. Finally, we integrate transcription factor (TF) binding data, downloaded from the public database Yeastract (www.yeastract.org). Here we utilize the full database, which consists of all documented TF-binding interactions, but with half the weight for bindings which are not experimentally confirmed. From this data, and for each gene, we count the number of identical upstream TFs. Based on the fraction of shared TFs, we promote those genes which share TFs as predictors for each other. Also this integration is performed in a soft, data-driven manner by CV, and the inclusion is only to the extent it actually lowers the prediction error.

Finally, we utilize all the parameter values for the model settled by the procedure above. For the test data, we obtain a Spearman rank correlation above 0.81, and hopefully the correlation for the time-series where the levels should be predicted is even higher.

Challenge 3 Best predictor talk:
A Nearest-Neighbor Co-expression Network for the Gene-Expression Prediction Challenge

Jianhua Ruan

Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249, USA, email: jruan@cs.utsa.edu

In this abstract, we present our experiments and results for the gene-expression prediction challenge in the Dream3 project. In this challenge, gene expression time course data is provided for four different strains of yeast (one wild type, and three mutants: Δ*gcn4*, Δ*leu3*, Δ*gat1*). The challenge is to predict the relative change of 50 selected target genes in the Δ*gat1* strain, given the other expression data and any other publicly available information about the genes.

The three mutant genes are probably the most useful biological information. All three genes encode transcription factors, and their interacting promoters have been identified using ChIP-chip assays. However, most of the 50 selected target genes do not interact with *gat1* with significant p-values, and therefore may have been affected by downstream regulators. Furthermore, ChIP-chip data does not provide time-series information. Therefore, we decided to base our prediction on expression data alone.

Our submitted prediction results were obtained using gene co-expression networks. Briefly, we first measured the similarity for every pair of genes based on the Euclidean distance between their expression profiles in the three strains other than Δ*gat1*. We then connected each gene to its k-nearest neighbors according to the similarity measure. Edges between target genes were prevented. The expression of g_i at time point t_j in Δ*gat1* was then predicted by the average expression at t_j of its immediate neighbors on the network. We estimated the prediction accuracy using reserved independent test cases: we randomly selected 50 genes that are not target genes, and computed the accuracy as the Spearman's rank correlation coefficient between the true expression values and the predicted expression values. This test was repeated 100 times and the average accuracy was computed. We tested different values of k for their accuracy, and found that the best accuracy, 0.75, was achieved at $k = 10$. Among the eight time points, the accuracy at time 0 was consistently the lowest.

We also tested several alternative strategies based on gene co-expression networks. First, instead of using a fixed k for all genes, we let the data to decide a different k for each gene. Briefly, we connected g_i to not only its top-5 neighbors, but also some of its top-30 neighbors if g_i was also listed by a neighbor as a top-30 gene. This improved the prediction accuracy by ~0.01 for all test cases (however, this was obtained after the submission deadline). In another strategy, we first selected the top-20 genes for each gene, and then identified among them the 10 genes that were highly connected to each other. This strategy did not improve the results significantly. In addition, using Pearson correlation coefficient as the similarity measure resulted in lower accuracy than using Euclidean distance. Finally, we also tested to predict the expression of gene g_i in Δ*gat1* from the expression of g_i in the other three strains using linear regression. The average accuracy for this strategy is 0.74. As the information used by the linear regression and the co-expression network based methods is orthogonal, it is likely that the combination of the two methods may further improve the results.

Challenge 4 Best predictor talk: Combining Multiple Models in Reconstructing In Silico Regulatory Networks

Kevin Y. Yip[1], Roger P. Alexander[2,3], Koon-Kiu Yan[2] and Mark Gerstein[1,2,4]

[1]Department of Computer Science; [2]Department of Molecular Biophysics and Biochemistry; [3]Department of Molecular, Cellular and Development Biology; [4]Program in Computational Biology and Bioinformatics, Yale University.

We performed computational reconstruction of the in silico gene regulatory networks in the DREAM3 Challenges. Our task was to learn the networks from two types of data, namely gene expression profiles in deletion strains (the ‘deletion data’) and time series trajectories of gene expression after some initial perturbation (the ‘perturbation data’). In the course of developing the prediction method, we observed that the two types of data contained different and complementary information about the underlying network. In particular, deletion data allow for the detection of direct regulatory activities with strong responses upon the deletion of the regulator while perturbation data provide richer information for the identification of weaker and more complex types of regulation.

We applied different techniques to learn the regulation from the two types of data. For deletion data, we learned a noise model to distinguish real signals from random fluctuations using an iterative method. The key idea is to first identify a conservative set of regulatory events that are unlikely to contain false positives. Wild-type expression levels and the width of the background Gaussian noise are then learned from the data outside this set of regulatory events. These learned values provide a probabilistic estimate of the chance that a given observed fluctuation is due solely to noise. The ones with very small probabilities then constitute a refined set of potential regulatory events. The whole process was repeated a number of times, which finally produced a probability for each gene A to be regulated by another gene B. Some ambiguous cases were resolved by checking the consistency between expression profiles in null mutants and heterozygous strains.

For perturbation data, we used differential equations to model the change of expression levels of a gene along the trajectories due to the regulation of other genes. We combined the predictions of various models (linear, sigmoidal and multiplicative). Model parameters were learned by using Newton’s method with multiple starting points and the differential equations solved numerically by Runge-Kutta method. If a set of potential regulators result in a small squared difference from the observed expression values, they are likely the real regulators of the target gene. Due to the high computational cost, and to avoid over-fitting, we started with models involving only one or two potential regulators. We then performed guided model fitting by using the obtained high-confidence set and the predictions from deletion data to construct more complex models.

The final predictions were obtained by combining the results from the two types of data. A comparison with the actual regulatory networks suggests that our approach is effective for networks with a range of different sizes.

To apply our method to real datasets, additional practical issues such as indirect regulation need to be confronted. Also, instead of performing model fitting in a purely unsupervised manner as described above, with the availability of some known examples of real regulatory networks, supervised or semi-supervised approaches could potentially lead to better performance.

Differences in Signaling Network Activities Between Primary and Transformed Human Hepatocytes Revealed by Cue-Signal-Response Pathway Maps

Leonidas G. Alexopoulos, Julio Saez-Rodriguez, Benjamin D. Cosgrove, Peter K. Sorger, and Douglas A. Lauffenburger

Center for Cell Decision Processes, MIT

Biological function is context-sensitive, so that the structures, dynamics and outputs of canonical signaling pathways vary from cell type to cell type and between normal and diseased cells. To date, data-driven pathway reconstruction has emphasized the inclusion of all possible data sources, producing “averaged” models that have some validity for pathways in general but do not accurately describe the actual operational network of any particular cell. Effective application of pathways maps to the study of disease and therapy requires context-specific information on specific cells and tissues in normal and abnormal states. Here we report the collection and analysis of a systematic dataset comprising the levels and states of modification of signaling proteins in primary hepatocytes and hepatocellular carcinoma (HCC) cells exposed to a range of cytokines and growth factors. We find significant and striking differences between the two cell types, notably significant alterations in pathways involved in innate immune responses and NF-κB function that cause profound changes in consequent cytokine secretion. Comparative pathway analysis should find wide application in the analysis of cell signaling for diagnostic and therapeutic biotechnology applications.

Small RNA sequencing resolves posttranscriptional regulatory networks

Thomas Tuschl

Howard Hughes Medical Institute, Laboratory of RNA Molecular Biology, The Rockefeller University, 1230 York Avenue, Box 186, New York, 10065, USA.

Over the past several years, our laboratory developed and applied small RNA library sequencing techniques to define RNA silencing mechanism-associated small RNAs in different animals. In mammals, the two predominant classes of endogenously expressed small RNAs are microRNAs (miRNAs) and piwi-interacting RNAs (piRNAs). miRNAs are associated with the ubiquitously expressed Argonaute (Ago) proteins, while piRNAs and Piwi proteins are restricted to the germline. Using this RNA library sequencing approach, we generated a large collection of small RNA expression profiles available at "www.mirz.unibas.ch/smiRNAdb/". To further analyze cell-type specific expression patterns in complex tissues, we have modified in situ hybridization protocols by enhancing miRNA crosslinking and using LNA-residue containing DNA probes.

More recently, and further exploiting sequencing approaches, we adapted and modified the CLIP (cross-linking and immunoprecipitation) technology to systematically clone and sequence millions of mRNA segments bound by miRNPs or mRNA-binding proteins. We have overcome two major limitations of standard CLIP assays. (1) By incorporation of photoreactive nucleosides (e.g. 4-thiouridine, 5-iodouridine, or 6-thioguanosine) in living cells, we were able to improve upon the yield of mRNA crosslinking by 2 to 3 orders of magnitude as compared to standard 260 nm crosslinking. (2) We are able to precisely locate the position of crosslink in CLIPed sequence segments by incorporation of 4-thiouridine. The crosslinking reaction alters the structure of the modified uridine, and reverse transcription/PCR returns a cytidine for the original uridine position. We applied 4-thioU-CLIP to isolate miRNA-targeted mRNA segments using IP of Argonaute proteins, or applied this to a series of mRNA-binding proteins to obtain a transcriptome-wide set of targets, from which consensus binding motifs and the context-dependence of binding sites can be evaluated. Our major goal is to systematically characterize the molecular network mediating post-transcriptional regulation of gene expression. These resent studies are carried out in collaboration with the computational biology group of Mihaela Zavolan (Biozentrum Basel).

Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in Oryza sativa

Xuefeng Zhou[1], Ramanjulu Sunkar[3], Hailing Jin[4], Jian-Kang Zhu[5], Weixiong Zhang[1,2]

[1] Department of Computer Science and Engineering; [2] Department of Genetics, Washington University in St.Louis; [3]Department of Biochemistry and Molecular Biology, Oklahoma State University; [4]Departments of Plant Pathology and Microbiology; [5]Department of Botany and Plant Sciences, University of California, Riverside.

A pair of natural antisense transcripts (NATs) are endogenous protein-coding or non-coding RNAs with sequence complementarity to each other. There are two classes of NATs, cis-NATs and trans-NATs.

NATs have been shown to play important roles in post-transcriptional regulation through the RNA interference pathway. We have combined the pyrophosphate-based high-throughput sequencing and computational analysis to identify and characterize, in genome-scale, cis- and trans-NAT small RNAs that are responsive to drought and salt stresses and under a normal condition in staple plant Oryza sativa. Out of a total of 714,202 sequence reads from high-throughput sequencing, we obtained 58,781, 43,003 and 80,990 unique small RNAs matching perfectly to the rice genome, from the control, salt and drought libraries, respectively. Computationally, we identified 344 cis-NATs and 7,142 trans-NATs that are formed by protein-coding genes. From the deep sequencing data, we found 108 cis-NATs and 7,141 trans-NATs that gave rise to small RNAs from their overlapping regions. Consistent with early findings, the majority of these 108 cis-NATs seem to associate with specific conditions or developmental stages. Furthermore, the overlapping regions of the cis- and trans-NATs appear to be more enriched with small RNA loci than non-overlapping regions. The small RNAs generated from cis- and trans-NATs have a length bias of 21-nt, even though their lengths spread over a large range. Our analyses also revealed several interesting results. More than 40% of the small RNAs from cis- and trans-NATs carry an 'A' as their 5' terminal nucleotides; a substantial portion of transcripts are involved in both cis- and trans-NATs; and many trans-NATs can form many-to-many relationships, indicating that NATs may form complex regulatory networks in O. sativa. Our work is the first genome-wide investigation of NAT-derived small RNAs in O. sativa. It revealed the importance of NATs in biogenesis of small RNAs, and broadened our understanding of the roles of NAT-derived small RNAs in gene regulation, particularly in response to environmental stimuli.

siRNA and miRNA transfections may reduce endogenous miRNA regulation

Aly Azeem Khan[1], Doron Betel[2], Christina Leslie[2*], Debora Marks[3*]

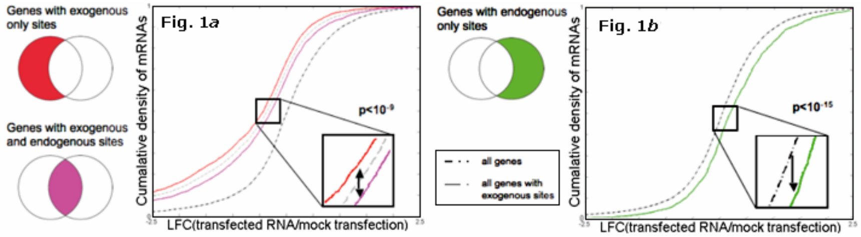
[1]Dept. Comp. Sci., Columbia University; [2]Comp. Bio. Prog., Memorial Sloan-Kettering Cancer Center; [3]Dept. Sys. Bio., Harvard Medical School; [*]Joint corresponding authors.

When a si/miRNA is transfected into a cell, hundreds of genes are down-regulated as observed through microarray analysis and mass spectrometry. Evidence suggests that this down-regulation is a direct effect of the transfected small RNA binding in a protein complex (RISC) to the cognate mRNA. However, some of the changes in gene expression are unexplained (e.g. up-regulation) and toxic effects after small RNA transfections suggest saturation of the protein machinery necessary for miRNA function. In addition, over 70% of genes that are predicted to be down-regulated remain unchanged after transfections.

We hypothesize that small RNA transfections have a global effect on the gene expression in a cell, beyond the more direct effects of down-regulation of target genes, due to this saturation of the small RNA pathway. To test this hypothesis we looked at changes in mRNA in a panel of more than 25 transfection experiments in 6 different cell types and changes in protein levels ascertained by mass spectrometry after small RNA transfections. We used a one-sided Kolmogorov-Smirnov (KS) statistic to test whether genes with conserved sites (in 3' UTR) for endogenous miRNAs were differentially affected by the transfection.

Our results show: **(i)** Genes that contain only transfected si/miRNA sites (and no endogenous miRNA sites) are significantly more down-regulated than those with transfected sites *and* endogenous sites, ($p<1.8E-09$ for pooled HeLa experiments, Figure 1a) **(ii)** Genes that contain only endogenous sites (and no transfected si/miRNA sites) are significantly up-regulated over background, ($p<1.3E-41$ for pooled HeLa experiments, Figure 1b) **(iii)** Genes that contain a single transfected si/miRNA site exhibit change in efficacy of down-regulation due to the endogenous profiles of differing cells; despite identical transfections ($p<.05$ for mir-17 transfection in HCT116 and HeLa).

Figure 1: a) Pooled data from 16 independent si/miRNA transfections into hela cells. b) Changes in gene expression after miR-17 transfection into HeLa cells showing effect on endogenous miRNA targets.



The implications of our findings are far ranging for siRNA and miRNA target gene prediction. They suggest a model that must quantitatively take into account the endogenous miRNA profile and the concentration of the protein machinery of the cells or tissues in question, in order to predict the outcome of small RNA perturbations.

Redundancy in the Eukaryotic regulatory code explains differences between binding and knockout experiments

Anthony Gitter[1], Oriol Fornés Crespo[2], Baldo Oliva[3], Itamar Simon[4], Ziv Bar-Joseph[1,5]

[1]Computer Science Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213 USA; [2]Institut Municipal d'Investigació Mèdica (IMIM), Doctor Aiguader, 88, Barcelona 08003, Catalonia, Spain; [3]Grup de Bioinformàtica Estructural (GRIB-IMIM), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra Barcelona 08003, Catalonia, Spain; [4] Department of Molecular Biology, Hebrew University Medical School, Jerusalem, Israel 91120; [5]Machine Learning Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213 USA

Recent studies in multiple species have indicated that most genes bound by a specific transcription factor (TF) are not affected when that TF is knocked out. The most comprehensive study of this phenomenon, in which all budding yeast TFs were knocked out one at a time [1], determined that only 3% of bound genes [2] were affected by the knockout, and similarly only 3% of knockout-affected genes were bound by the corresponding TF. This led some researchers to claim that most binding events do not affect expression levels, a process that is often termed 'non-functional binding'.

To investigate whether this was indeed the case or if there are other ways to explain the disagreement between binding and knockout experiments, we studied the dependence of this lack of agreement on the homology relationships and on the protein-protein and protein-DNA interactions of the TF. We first cleaned the expression and binding data by removing genes that are non-specifically affected by multiple knockouts and using a binding dataset that leverages conservation criteria to yield higher confidence interactions. Next, for each TF we identified its most similar homologous TF (based on either the entire sequence or the binding domain) and grouped TFs by this similarity score. For the set of TFs with the least similar homologs, the overlap between binding and knockout results was 4 times greater than the baseline overlap, and those TFs with the most similar homologs had 0% overlap. Within each similarity group, we found that TFs that shared many protein interactions with their homologs had lower overlap than those with few common interactions. To explain the low number of knockout-affected genes that are bound by the TF, we used a protein-protein interaction network. Even when focusing on nodes at a distance of at most two from the TF, we were able to increase the overlap more than 10-fold and to correspondingly increase the significance of the overlap. Analysis in other species, specifically human, led to similar results. Our findings indicate that redundancy in the Eukaryotic regulatory code can explain some of the disagreement between the binding and knockout experiments. Thus, rather than non-functional binding, we have redundant functional units that can compensate for the loss of other units in the cell leading to robust control mechanisms.

[1] Harbison, C.T. *et al*. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99-104 (2004).

[2] Hu, Z. and Killion, P.J. and Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genetics* 39, 683-687 (2007).

Human Genome Annotation, Focusing on Intergenic Regions

Mark Gerstein[1,2,3]

[1]Computational Biology & Bioinformatics; [2]Molecular Biophysics & Biochemistry; [3]Computer Science, Yale U.

A central problem for 21st century science will be the annotation and understanding of the human genome. My talk will be concerned with topics within this area, in particular annotating pseudogenes (protein fossils), binding sites, CNVs, and novel transcribed regions in the genome. Much of this work has been carried out in the framework of the ENCODE and modENCODE projects.

In particular, I will discuss how we identify regulatory regions and novel, non-genic transcribed regions in the genome based on processing of tiling array and next-generation sequencing experiments. I will further discuss how we cluster together groups of binding sites and novel transcribed regions.

Throughout I will try to introduce some of the computational algorithms and approaches that are required for genome annotation -- e.g., the construction of annotation pipelines, developing algorithms for optimal tiling, and refining approaches for scoring microarrays.

http://gersteinlab.org , http://pseudogene.org , http://tiling.gersteinlab.org

References:

Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification.
TE Royce, JS Rozowsky, MB Gerstein (2007) Nucleic Acids Res 35: e99.

Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions.
ZD Zhang, A Paccanaro, Y Fu, S Weissman, Z Weng, J Chang, M Snyder, MB Gerstein (2007) Genome Res 17: 787-97.

The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci.
JS Rozowsky, D Newburger, F Sayward, J Wu, G Jordan, JO Korbel, U Nagalakshmi, J Yang, D Zheng, R Guigo, TR Gingeras, S Weissman, P Miller, M Snyder, MB Gerstein (2007) Genome Res 17: 732-45.

What is a gene, post-ENCODE? History and updated definition.
MB Gerstein, C Bruce, JS Rozowsky, D Zheng, J Du, JO Korbel, O Emanuelsson, ZD Zhang, S Weissman, M Snyder (2007) Genome Res 17: 669-81.

Analysis of Copy Number Variants and Segmental Duplications in the Human Genome: Evidence for a Change in the Process of Formation Mechanism in Recent Evolutionary History
Philip M. Kim Hugo Y. K. Lam Alexander E. Urban, Jan Korbel, Xueying Chen, Michael Snyder and Mark B. Gerstein
Genome Res. (in press, 2008)

Modeling ChIP sequencing in silico with applications.
ZD Zhang, J Rozowsky, M Snyder, J Chang, M Gerstein (2008) PLoS Comput Biol 4: e1000158.

Non-homologous structured CRMs from the Ciona genome

Albert Erives[1]

[1] *Department of Biological Sciences, Dartmouth College, Hanover, NH, U.S.A.*

The internal functional organization of cis-regulatory modules (CRMs) is critically important for understanding the mode and tempo of gene regulatory evolution as well as for deciphering and annotating genomic sequences. In an open-ended search for loose clusters of known mesodermal enhancer motifs in the Ciona intestinalis genome, I discovered the existence of a class of highly organized CRMs in otherwise unrelated genes expressed early in development. Each such CRM is composed of distinct motifs located at specific positions along ~160 bp of DNA sequence, and is able to drive expression in distinct mesodermal compartments descended from the B4.1 blastomere. The majority of the loci bearing these B4.1-specific modules encode important early mesodermal transcription factors at the snail, paraxis, and tbx6 orthologous loci of this invertebrate chordate system. These unrelated genes encode members of the C2H2 zinc-finger, bHLH, and T-box transcription factor families, and likely serve as a chordate-specific trans-code for paraxial mesoderm. One other similarly organized enhancer was discovered in the TNC3 muscle structural locus. These results suggest that organization of binding sites over the length of the enhancer sequence is a critical aspect of gene regulatory biology. The extent to which this is a general principle will facilitate our ability to identify, decipher, and categorize the regulatory functions contained in whole genome assemblies.

Regulation of muscle gene expression by homeodomain transcription factors

Busser, B.W.[1], Jaeger, S.A.[2], Singhania, A.[1], Gisselbrecht, S.S.[2], Berger, M.F.[2,5], Bulyk, M.L.[2-5], and Michelson, A.M.[1]

[1]National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD. [2]Division of Genetics, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA. [3]Department of Pathology, Brigham & Women's Hospital and Harvard Medical School, Boston, MA. [4]Harvard/MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA. [5]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138.

Homeodomain (HD) proteins comprise a large family of highly conserved transcription factors (TFs) with diverse developmental functions. In *Drosophila*, a set of tissue-specific HD TFs has been proposed to control the unique gene expression programs of individual muscle founder cells (FCs). We have investigated this hypothesis using an integrated approach that combines genome-wide expression profiling, microarray-based TF binding site determination, computational prediction of *cis*-regulatory modules (CRMs), *in silico* evaluation of combinatorial TF codes, empirical testing of candidate CRMs, and both *cis* and *trans* tests of target gene regulation. We first showed that a small set of previously characterized FC genes is differentially responsive to over-expression of two muscle HD TFs, Slouch (Slou) and Muscle segment homeobox (Msh). FC gene expression was activated and/or repressed in a cell-specific manner by these TFs, and responsiveness correlated with TF co-expression in wild-type embryos. Next, we extended the identification of HD-responsive genes by genome-wide expression profiling of mesodermal cells purified from embryos in which an individual HD TF was over-expressed. These experiments revealed that different FC genes can be activated, repressed or remain unaffected by ectopic HD TFs. Computational analysis then showed that the inclusion of Slou or Msh binding sites in a previously delineated core transcriptional FC code defines a more specific combination of co-regulatory TFs for the corresponding HD-responsive gene sets. While many binding sites are shared by Slou and Msh ("HD-common"), each TF also binds a few unique sequences ("Slou- or Msh-specific"). Computational searches for combinations of binding motifs that are overrepresented in the noncoding regions of Slou-responsive FC genes showed a statistical enrichment for both HD-common and Slou-specific sites, again clustered with the core set of FC TFs. These results suggest that distinct binding preferences can mediate distinct biological effects. Indeed, mutagenesis of HD-common sites completely inactivated known FC CRMs, consistent with multiple cell-specific HD TFs functioning through these sites. Finally, gene expression profiling showed that HD TFs also activated genes uniquely expressed in fusion-competent myoblasts (FCMs), which do not normally contain these TFs. These results are consistent with HD TFs regulating two distinct temporal waves of myogenic gene expression, one in the developing muscle FC, and a second following fusion of the FC and FCM in the mature multinucleated myotube. Collectively, these studies provide new insights into the transcriptional codes that regulate muscle gene expression and into the roles of individual HD TFs in specifying cellular identity.

A Transcription Factor Affinity Based Code for Mammalian Transcription Initiation

Molly Megraw[1], Fernando Pereira[2], Shane T. Jensen[3], Uwe Ohler[1], Artemis G. Hatzigeorgiou[2,4]

[1]Institute for Genome Sciences and Policy, Duke University, Durham, NC, USA.
[2]Department of Computer and Information Science, School of Engineering, [3]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA.
[4]Institute of Molecular Oncology, BSRC "Alexander Fleming", Athens, Greece.

The recent arrival of large-scale Cap Analysis of Gene Expression (CAGE) datasets in mammals provides a wealth of quantitative information on coding and non-coding RNA polymerase-II transcription start sites (TSSs). Genome-wide CAGE studies reveal that a large fraction of TSSs exhibit peaks where the vast majority of associated tags map to a particular location (~45%), whereas other active regions contain a broader distribution of initiation events. The presence of a strong single peak suggests that transcription at these locations is mediated by position specific sequence features. Indeed, we demonstrate that a probabilistic model leads to near-perfect classification results in cross-validation (auROC = 0.98), enabling high-resolution genomic scans for single-peak TSSs. This new model is solely based on known transcription factors (TFs) and their respective regions of positional enrichment. The model combines the region-specific TF affinities by logistic regression, and includes a sparseness constraint which effectively leads to a selection of the most salient features. Inspection of model parameters confirms that canonical sequence features such as TATA box, Initiator, and GC content do play a significant role in transcription initiation, but that the positioning of many other TF binding sites adds significantly to the predictive power of de novo TSS classification. This interpretable model structure also leads to the discovery of multiple single-peak promoter subcategories with contrasting TF feature composition. Together these results offer a new perspective on single-peak promoters, and demonstrate that high accuracy computational TSS prediction is achievable for a large subgroup of mammalian promoters. The identification of single-peak promoters is of further interest for the analysis of microRNA genes, where the vast majority of full transcripts remain unknown.

Design principles of biological systems

Uri Alon

Weizmann Institute

This talk will describe design principles of biological circuits, including experimental results on biological circuits working in living cells at high temoporal resolution.

Reference: Alon U. An Introduction to Systesm Biology: Design principles of biological Circuits, CRC 2006**.**

Time-Varying Networks: Reconstructing Temporally Rewiring Gene Interactions

Amr Ahmed, Le Song, Eric Xing*

*School of Computer Science, Carnegie Mellon University; *To whom correspondence should be addressed*

Biological networks underlying temporal process, such as the development of *Drosophila melanogaster*, can exhibit significant topological changes to facilitate dynamic regulatory functions. Thus it is essential to develop methodologies that capture the temporal evolution of networks, which make it possible to study the driving forces underlying dynamic rewiring of gene regulation circuitry, and to predict future network structures.

Here, we present a successful reverse engineering of a series of time-varying gene interaction networks from microarray measurements. Our study is based on a new machine learning algorithm called TEmporally Smoothed L1-regularized LOgistic Regression, or Tesla (stemmed from TESLLOR, the acronym of our algorithm). Tesla is based on a key assumption that temporally adjacent networks are unlikely to be dramatically different from each other in topology, and therefore are more likely to share common edges than temporally distal networks. An important property of this novel idea is that it fully integrates all available samples of the entire time series in a single inference procedure that recovers the wiring patterns between genes over a time series of arbitrary resolution, from a network for every single time point, to one network for every K time points where K can be very small.

We applied Tesla to the genome-wide reverse-engineering of the latent sequence of temporally rewiring gene networks over more than 4000 genes during the life cycle of *Drosophila melanogaster*. Our methods offer the first glimpse of time-specific snapshots and temporal evolution patterns of gene networks in a living organism during its full developmental course (several temporal snapshots of the time varying networks are shown in Figure 1). The recovered networks with this high temporal resolution chart the onset and duration of many gene interactions which are missed by typical static network analysis, and are suggestive of a wide array of other temporal behaviors of the gene network over time not noticed before.

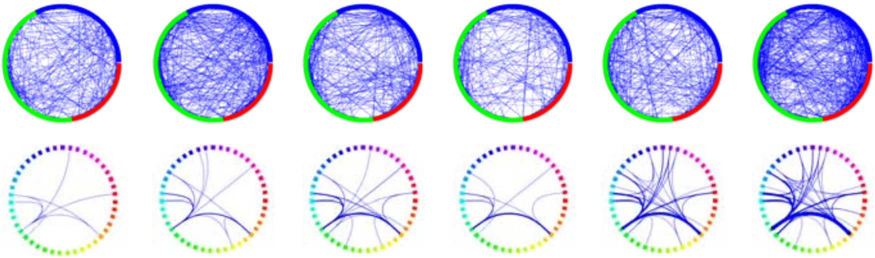


Figure 1. Temporal snapshots of the time rewiring gene networks during the life cycle of *Drosophila Melanogaster*. Top row: genes are ordered according to their functions (red: cellular component, green: molecular function, and blue: biological function) on the circle, and edges denote reverse engineered interactions. Bottom row: genes are further clustered according to 23 ontology groups. Each color patch on the circle represents an ontology group and the width of an edge is proportional to the number of between-group gene interactions.

Bayesian structural equation model (SEM) identification of regulatory networks.

Jason G. Mezey[1], Chuan Gao[1], Benjamin Logsdon[1]

[1]*Department of Biological Statistics and Computational Biology, Cornell University.*

Cellular transcription profiles are determined in part by underlying regulatory networks which control gene expression and transcript degradation. Within populations, segregating alleles can produce additional (genetic) variation in transcription profiles by acting through regulatory networks. In theory, population-level genotype and transcription profile data can therefore be used to infer regulatory network structure and to map loci where allelic variation affects transcription through a network.

We develop a general statistical framework for mapping loci affecting gene expression and for identifying network regulatory structure using genotype and transcription profile data from a population sample. For these purposes, we employ an appropriate form of Structural Equation Models (SEMs), a broad class of linear statistical models that include Bayesian Networks, path models, and multivariate regressions. These models have the appealing property of explicitly modeling the directional connections within networks and in the general form, can include both acyclic and cyclic structures. Our framework includes a set of novel computational approaches for dealing with the problems of model identification and equivalence. For genome-wide applications, we introduce Markov chain Monte Carlo methods to perform an efficient search over possible models. Our Bayesian framework makes use of mixture priors on genetic and regulatory parameters and we use these to incorporate assumptions which limit our inferences to cases where genetic effects and distinct network structures can be distinguished.

We demonstrate that our approach is far more powerful than individual marker testing approaches for mapping genetic loci with effects on gene expression and that our method can correctly identify complex network structure including cyclic regulation when analyzing simulated data. We also analyze data from several human population studies which included the collection of genome-wide genotype and transcription profile data. Some of the more interesting results of these analyses include the identification of multiple independent *cis*- genetic effects on gene expression in human lung cells and the identification of cyclic network structure responsible for regulation of transcription in cell lines derived from individuals in the international HapMap project.

Regulatory Networks Pertinent To Self-Renewal In Human Embryonic Stem Cells

Hedi Peterson[1,2], Marc Jung[3], Hans Lehrach[3], Pascal Kahlem[4], James Adjaye[3], Jaak Vilo[1,2]

[1]University of Tartu; [2]Quretec Ltd; [3]Max Planck Institute for Molecular Genetics; [4]EMBL-European Bioinformatics Institute

Understanding the complex nature of gene regulation in human embryonic stem cells (hESCs) is expected to provide a strong impact in both fundamental and clinical research. The pluripotent state of ES cells is known to be regulated by three transcription factors – Oct4, Sox2 and Nanog.

Current studies employing chromatin immunoprecipitation methods (ChIP-chip or ChIP-seq) have identified several direct targets of these three regulators and have proposed regulatory networks crucial for maintaining self-renewal in both human and mouse ES cells. Preliminary regulatory networks based on direct targets of Oct4, Sox2 and Nanog have been published.

We expanded the ESC regulatory network by combining both data integration and network reconstruction methods. We studied genes showing significant expression level changes in Oct4 pertubation experiments but lacking support in any Oct4 ChIP-chip experiments, and therefore being most probably indirectly regulated. More specifically we specialized our analysis on putative targets that play a role in early differentiation.

Based on Gene Ontology analysis, these putative indirect targets are enriched in development and chromatin structure related terms. De novo motif discovery and matching combined with conservation information resulted in enriched binding sites for transcription factors that are known to be connected to Oct4 and/or take part in distinct differentiation pathways. These factors seem to regulate different subsets of Oct4 indirect target genes either separately or in various combinations.

In our work we have identified candidate regulators that play an important role in human embryonic stem cell regulatory network by transducing the signal from Oct4 to its downstream targets.

This work has been supported by the EU FP6 Network of Excellence ENFIN (contract LSHG-CT-2005-518254).

Detailing regulatory networks through large scale data integration

Curtis Huttenhower[1,2,*], K. Tsheko Mutungu[1,*], Olga G. Troyanskaya[1,2], Hilary A. Coller[3]

[1]Computer Science; [2]Lewis-Sigler Institute for Integrative Genomics; [3]Molecular Biology, Princeton University, Princeton, NJ 08540; [*]Equal contribution

While the genome sequence of an organism describes its complement of potential proteins, it is the controlled expression, translation, and modification of these proteins that allows cells to survive and grow. At the level of transcription and mRNA stability, a complex regulatory network of transcription factors, RNA binding proteins, and microRNAs governs the interactions between components of a cell's internal state and its external environment. Understanding the elements of this regulatory network and the stimuli to which it responds in higher organisms has been of increasing recent interest[1,2,3] as a key to metazoan systems biology, particularly as genetic misregulation is a major cause of human disease.

Here, we describe a Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction (COALESCE) allowing the discovery of regulatory motifs and modules from large collections of genomic data. COALESCE takes advantage of Bayesian integration of multiple data types on a large scale to predict coregulated gene modules, the conditions under which they are coregulated, and the consensus binding motifs responsible for their regulation. Through a novel synthesis of gene expression biclustering, motif prediction, and data integration (including expression, DNA sequence, nucleosome positioning, and evolutionary conservation), COALESCE can successfully find coregulated modules for organisms ranging from *E. coli* to human beings and from data collections as large as 15,000 experimental conditions.

We present the results of applying COALESCE to data from a wide range of organisms, including *H. sapiens*, *M. musculus*, *C. elegans*, *S. cerevisiae*, *H. pylori*, and *E. coli*. Using ~2,200 yeast expression conditions, we recapitulate many known regulatory interactions (e.g. *AFT2* in iron transport, *STE12* activating mating genes) and highlight the importance of PUF family 3' UTR binding in a wide variety of targets, often ribosomal. In an analysis of ~15,000 human gene expression conditions, we extract a wide variety of putative upstream binding sites and potential 3' miRNA sites. On synthetic data comprising 5,000 genes and 100 conditions with 10 "activators" and "repressors" generated from a randomized model, COALESCE successfully recovered 60-90% of the affected genes, conditions, and binding motifs. In five sets of synthetic data containing no such regulators, COALESCE generated zero false positives. We are currently in the process of testing several novel transcriptional regulators of quiescence in human fibroblasts as predicted by COALESCE, as its ability to probabilistically leverage large collections of heterogeneous data is particularly suited to unraveling complex metazoan regulatory networks.

[1] Reiss DJ, Baliga NS, Bonneau R. "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks" BMC Bioinformatics 7:280, 2006
[2] Kloster M, Tang C, Wingreen NS. "Finding regulatory modules through large-scale gene-expression data analysis" Bioinformatics 21(7):1172-9, 2005
[3] Elemento O, Slonim N, Tavazoie S. "A universal framework for regulatory element discovery across all genomes and data types" Mol. Cell 28(2):337-50, 2007

Modeling Transcriptional Regulatory Modules Using Dynamic Factor Analysis

Alain Tchagang, Sieu Phan, Fazel Famili, and Youlian Pan

Knowledge Discovery Group, Institute for Information Technology, NRC 1200 Montreal Road, Ottawa ON, K1A 0R6 Canada

One of the major goals in computational biology is to identify transcriptional regulatory modules from gene expression datasets. Transcriptional modules are groups of transcription factors (TFs) and their target genes, such that genes in the same group tend to be similarly expressed and regulated by the same transcriptional regulatory machinery across a number of experimental conditions or time points. To analyze gene expression datasets containing large numbers of observed variables, many researchers have applied dimension reduction techniques such as principal component analysis (PCA), independent component analysis (ICA), singular value decomposition (SVD), network component analysis (NCA), factor analysis (FA), correspondence analysis (CA), or multidimensional scaling. In these techniques, a measure of similarity between the observed variables is defined and a low dimensional graphical representation of these similarities is presented. Although some of these techniques have been shown to reveal significant patterns in gene expression data, none of them, however, is designed to analyze time-series. It is possible to apply PCA to time-series gene expression data and connect consecutive time points with each other; interpretation of the results is most likely to be difficult because PCA method does not consider the sequential nature of time series data. If the order of time in the input gene expression data matrix is altered, the same results are obtained.

Dynamic factor analysis (DFA) is a dimension-reduction technique especially designed for time-series data. It has been intensively used in econometric and psychological fields for the past two decades. DFA can be used to model time-series in terms of common patterns and explanatory variables. This includes short time-series and non-stationary time-series, which may contain missing values. For example, DFA is able to reveal underlying common patterns in the multivariate time-series, whether they are related or independent to each other.

In this study, we developed a DFA approach and evaluated its performance on two criteria: gene clustering and TFs activity profiles reconstruction. The first tests its capability to uncover biologically meaningful groups of genes whereas the second assesses the reproducibility of TF activity profiles from the time-series gene expression data. We tested this method using synthetic data and applied it to well-defined *Arabidopsis thaliana*, and *Brassica napus* time-series gene expression datasets. Our results indicate that DFA outperforms PCA, ICA, SVD, NCA, and FA.

Markov Chain Monte Carlo (MCMC) optimization to learn coupled Gene Regulatory Networks: the Inferelator 2.0

Aviv Madar[1], Eric Vanden-Eijnden[2], Richard Bonneau[1,2]

[1]Center for Genomics & Systems Biology, New York University, New York, NY 10003, USA; [2]Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, New York, NY 10003, USA

Learning and characterizing Regulatory Networks (RNs)—responsible for the remarkable ability of organisms to adapt to changing environment—is a key problem in biology with applications spanning bioengineering, drug development and many other biological fields. Detailed transcriptional RNs can be modeled as a system of ordinary differential equations (ODEs), describing the rate of change in transcripts as a function of transcription factors. We have recently described a network inference algorithm, the Inferelator, which infers regulatory influences for genes and gene clusters. The typical input is: 1) a microarray compendium composed of time-series and equilibrium measurements, and 2) prior information such as a set of considered predictors (e.g. transcription factors). The output is a dynamical model for each gene, i.e. a differential equation describing the rate of change in mRNA concentration as a function of relevant predictors. At the core of the algorithm is a model shrinkage step (L1-shrinkage) that allows the Inferelator to learn sparse models. We have shown that the Inferelator is descriptive and predictive (up to the next time point in a time series) over a large test-set (with different conditions then train-set). The Inferelator, however, treats the predictors as constant throughout each time-interval, which becomes a crude estimation as the length of the time interval increases. Here, we extend the Inferelator procedure by learning dynamical models for all genes in the system, as a coupled system. In this coupled system all components (including predictors) are predicted and updated on a number of intermediate points in each time interval resulting in a more accurate description of the system's time evolution. To this end we have developed a MCMC optimization algorithm, which learns models that agree with the uncoupled as well as coupled dynamics of the system, and is constrained by the sparsity expected from biological systems. Preliminary results suggest improved predictive performance for long time intervals. This should allow us to model the dynamics of a cell's mRNA expression levels, over longer time scale, such as the cell cycle.

Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer

Xi Chen[1], Lily Wang[2]

[1]*Department of Quantitative Health Sciences, The Cleveland Clinic;*

[2]*Department of Biostatistics, Vanderbilt University*

Due to the large variability in survival times between cancer patients and plethora of genes on microarrays unrelated to outcome, building accurate prediction models that are easy to interpret remains to be a challenge. In this paper, we propose a general strategy for improving performance and interpretability of prediction models by integrating gene expression data with prior biological knowledge. First, we link gene identifiers in expression dataset with gene annotation databases such as Gene Ontology. Then we construct “supergenes” for each gene category by summarizing information from genes related to outcome using a modified Principal Component Analysis (PCA) method. Finally, instead of using genes as predictors, we use these supergenes representing information from each gene category as predictors to predict survival outcome. In addition to identifying gene categories associated with outcome, the proposed approach also carries out additional within category selection to select important genes within each gene set. We show using two real breast cancer microarray datasets the prediction models constructed based on gene sets (or pathway) information outperforms prediction models based on expression values of single genes, with improved prediction accuracy and interpretability.

Constrained clustering for cross species analysis of gene expression data

Guy E. Zinman[1], Dwight Kuo[2], Kai Tan[2], Trey Ideker[2], Ziv Bar-Joseph[1]

[1]Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213;

[2]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093.

Many biological systems operate in a similar manner across a large number of species or conditions. The differences in the dynamics of these systems point to evolutionary changes in their genetic programs. In this study we built a framework that allowed us to easily find and compare sets of orthologous genes that show conserved or differentiated expression patterns.

Microarrays are one of the few tools that we have today to measure dynamic condition-specific response of complex biological systems. However, unlike most other biological datasets which are represented by a few letters or binary edges (e.g., DNA sequences, protein interactions) microarray data values are continuous and often noisier, making it hard to compare results across species and distinct conditions. Another challenge arises from differences in expression analysis methods, which often occur when each of the species is analyzed independently.

To solve these problems we developed a computational method that relies on constrained clustering for combining experiments of the same biological system in multiple species. Our method modifies the target function of the clustering algorithm to encourage homologous genes to co-cluster. This helps to overcome the influence of small changes in expression, attributed to noise, which often lead to homologs being assigned to different clusters.

We have used our method, termed Gene-SoftClust to analyze time-series microarray data from two distinct biological systems; the response of three yeast species treated with the anti-fungal drug Fluconazole, and mammalian immune response experiments (obtained from the Pittsburgh Center for Modeling Pulmonary Immunity). Our method allowed us to identify both common and unique response patterns. In the yeast analysis we have identified some clusters with common responses for all species, while other clusters showed similar responses for only two of the three species, allowing us to focus on the molecular basis for the differences in phenotypic responses to the drug. In the immune response analysis we were able to find a 'core' set of genes that respond in similar way across all conditions and species as well as species and pathogen specific response genes. Both studies highlight the usefulness of cross species analysis of microarray data for explaining similarities and differences using common and unique response patterns.

Chromatin Signatures of Transcriptional Enhancers

Bing Ren

Ludwig Institute for Cancer Research and University of California, San Diego

The human body is composed of a large number of cell types, each defined by a unique gene expression profile. Cis-regulatory sequences, such as promoters, enhancers and insulator elements, control gene expression by interacting with specific transcription factors, many of which modulate the local chromatin modification structures. While each class of cis-regulatory elements may contribute to cell-type dependent gene expression, previous studies have mainly focused on the role of promoters as a driving force behind tissue-specific and differential expression, partly due to a scarcity in our knowledge of the long range regulatory elements. In order to better characterize the mechanisms of cell-type specific gene expression, we have performed experiments to localize the genomic binding sites of general transcription factors, active chromatin modifications, and insulator binding protein CTCF in the human genome in multiple cell types. We find that transcriptional promoters and enhancers are associated with distinct chromatin signatures. Further, we show that the chromatin signatures at promoters, as well as the localization pattern of the insulator-binding protein CTCF, remains largely invariant across cell types. By contrast, the majority of enhancers are marked by specific histone modifications in a cell-type dependent manner. We observe that cell-type specific gene expression correlates with not only changes in chromatin marks at promoters, but also changes at enhancers, and that the effects of multiple enhancers toward the adjacent genes tend to be synergistic. These results show at a large scale that enhancers play an important role in cell-type dependent gene expression, and highlight the necessity to identify these sequences for understanding mechanisms of cell-type specific gene expression.

Targeted recruitment of histone modifications in humans predicted by genomic sequences

Guo-Cheng Yuan[1,2]

[1]Department of Biostatistics, Harvard School of Public Health; [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute

Histone modifications are important epigenetic regulators and play a critical role in development. The targeting mechanism for histone modifications is complex and still incompletely understood. Here we applied a computational approach to predict genome-scale histone modification targets in humans by the genomic DNA sequences using a set of recent CHIP-seq data. We found that a number of histone modification marks can be predicted with high accuracy. On the other hand, the impact of DNA sequences for each mark is intrinsically different dependent upon the target- and tissue-specificity. The predicted average profiles at promoters and enhancers are in good agreement with experimental data. Diverse patterns are associated with different repetitive elements. Unexpectedly, we found that non-overlapping, functionally opposite histone modification marks may share similar sequence features. We propose that these marks may target a common set of loci but are mutually exclusive and that the competition may be important for developmental control. Taken together, we showed that our computational approach may provide insights not only into pattern detection but also into the targeting mechanism of histone modifications.

Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq Data

Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, Keji Zhao

Laboratory of Molecular Immunology, National Heart Lung and Blookd Institute (NHLBI), National Institutes of Health (NIH), Bethesda, MD 20892

ChIP-Seq, which combines chromatin immunoprecipitation (ChIP) with ultra high-throughput massively parallel sequencing, is increasingly being used for mapping protein-DNA interactions in-vivo on a genome scale. Typically, short sequence reads from ChIP-Seq are mapped to a reference genome for further analysis. Although genomic regions enriched with mapped reads could be inferred as approximate binding regions, short read lengths (approximately 25-50 bp) pose challenges for determining the exact binding sites within these regions.

Here, we present SISSRs (Site Identification from Short Sequence Reads), a novel algorithm for precise identification of binding sites from short reads generated from ChIP-Seq experiments [1, 2]. The sensitivity and specificity of SISSRs are demonstrated by applying it on ChIP-Seq data for three widely studied and well-characterized human transcription factors: CTCF (CCCTC-binding factor), NRSF (neuron-restrictive silencer factor; also known as REST, for repressor element-1 silencing transcription factor) and STAT1 (signal transducer and activator of transcription protein 1). We identified 26814, 5813 and 73956 binding sites for CTCF, NRSF and STAT1 proteins, respectively, which is 32, 299 and 78% more than that inferred previously for the respective proteins. Motif analysis revealed that an overwhelming majority of the identified binding sites contained the previously established consensus binding sequence for the respective proteins, thus attesting for SISSRs' accuracy. Binding sites identified by SISSRs are of high resolution, i.e. the identified sites are within few tens of base pairs from the center of the nearest canonical motif. For example, >90% of CTCF sites were within 32-bp from the motif center.

SISSRs' sensitivity and precision facilitated further analyses of ChIP-Seq data revealing interesting insights, which we believe will serve as guidance for designing ChIP-Seq experiments to map in vivo protein-DNA interactions. Our analysis reveals that tag densities at the binding sites are a good indicator of protein-DNA binding affinity, which could be used to distinguish and characterize strong and weak binding sites. Using tag density as an indicator of DNA-binding affinity, we have identified core residues within the NRSF and CTCF binding sites that are critical for a stronger DNA binding.

SISSRs is robust, yet flexible enough that it allows the user to control for elements such as antibody specificity and sequencing errors, which could affect the quality of generated data, and thus the accuracy and resolution of identified binding sites.

[1] Jothi, R. *et al*. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq Data. *Nucl. Acids Res.* 36(16):5221-31 (2008).

[2] http://sissrs.rajajothi.com

Integrating Multiple Evidence Sources to Predict Transcription Factor Binding Across the Human Genome

Jason Ernst[1,2], Ziv Bar-Joseph[1]

[1]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA

[2]Current affiliation: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA

Transcriptional gene regulation is a central biological process that is largely controlled by transcription factors binding to specific locations in an organism's genome. For many transcription factors information about their sequence binding preferences is known and characterized by a sequence binding motif. However computationally determining the locations that a transcription factor binds in the genome based on its motif is a very challenging problem particularly in species with larger genomes, such as human, since there are often many sequences in the genome which match the motif, but are not bound. Recognizing this challenge researchers have proposed specific rules based on sequence conservation or location relative to a transcription start site to help differentiate potentially true binding site locations from random ones. A number of other evidence sources have become available that are also informative of transcription factor binding such as high resolution genome wide histone modification data and DNaseI hypersensitivity data. Integrating multiple evidence sources has the potential to improve prediction of bound locations in the genome. The recent availability of experimental genome wide binding data in human cells for a number transcription factors allows supervised machine learning methods to be used to integrate the multiple evidence sources to predict the likely locations transcription factor binding across the human genome.

We present a method that first uses a logistic regression classifier to infer an empirical prior of transcription factor binding based on 29 general evidence features, but without using any motif specific information. Surprisingly, using cross-validation, we show that this prior can be highly predictive of true locations of transcription factor binding compared to randomly selected sites even when no binding motif is used. We are able to obtain an average area under the receiver operator curve (AUC) value of 0.77 across all test cases and a maximum AUC value of 0.90. We also show that it improves on any single feature we consider. We then combined this empirical prior with motif information specific to the transcription factor. This allows us to improve binding predictions compared to either of these information sources alone as we show by using cross validation on the set of transcription factors considered. We further tested our method on an independent set of 30 ChIP-chip experiments for the E2F family of transcription factors and on new E2F binding data. In both cases our method was able to accurately predict many of the true binding locations. Our results demonstrate that integrating multiple evidence sources informative of transcription factor binding in a principled manner can lead to improved prediction of transcription factor binding. These in turn could be used to study biological systems by integrating them with other types of data including gene expression data.

A Model of the Competitive Binding of DNA by Nucleosomes and Transcription Factors

Todd Wasson[1], Alexander J. Hartemink[1,2]

[1]Computational Biology and Bioinformatics; [2]Department of Computer Science, Duke University, Durham, NC 27708

Transcriptional behavior is determined in large part by the occupancy of a gene's promoter by a set of various DNA binding proteins and protein complexes, most notably transcription factors. However, as 80% of the genome is occupied by nucleosomes, chromatin structure plays an important role in the accessibility of DNA to transcription factors. Furthermore, as DNA occupancy is the dynamic result of thermodynamic competition amongst a set of DNA binders, a model of occupancy including transcription factors and nucleosomes and expressed in terms of probabilities or frequencies is likely to give a more complete view of the underlying biology. We consider a mechanistic view of DNA occupancy to be especially useful in understanding the nature of competition and cooperativity amongst DNA binders and the resulting landscape of occupancy of nearby genomic DNA in particular. That is, we model DNA binders explicitly and individually in thermodynamic competition with one another based on sequence affinity and concentration, as opposed to considering a more generalized model of the aggregate occupancy behavior.

We propose a statistical thermodynamic model to understand DNA occupancy in yeast from a mechanistic standpoint. While similar existing techniques may consider either multiple transcription factors and naked DNA [Sinha 2006] or nucleosomes and naked DNA [Segal 2006], our model allows for all of these at once. Given a sequence, a collection of generic DNA binders with defined sequence specificities, concentrations of each binder, and a system temperature, our model produces a posterior decoding under the Boltzmann distribution of the binding probability per sequence position of each binder along the sequence.

Our model is quite flexible and extensible. Any DNA binder whose binding preferences can be described by sequence affinity may be incorporated into the model. These preferences can be represented by arbitrary probability models: currently our nucleosome model assumes position-specific dinucleotide preferences while our transcription factor model assumes mononucleotide preferences (PSSMs). We may ask nuanced questions of the model, such as what the occupancy of a region looks like given that particular transcription factors are held in place in specific positions, or how often two transcription factors bind within a given distance of one another. We may allow refinement of the sequence preference of DNA binders, and aid in discovery of hitherto unknown roles of transcription factors. We may perform mutational analyses to both sequence and binders to predict novel binding behavior and guide biological experiments.

Importantly, our technique is computationally efficient. Entire yeast chromosomes with models on the order of 5000 states representing hundreds of transcription factors and nucleosomes are decoded in minutes.

This detailed view of DNA binding allows for future understanding of the interplay between the transcriptional control exerted by a gene product and the gene's own transcriptional rate; that is, the exploration of the larger transcriptional regulatory network in terms of the mechanistic interplay of its constitutive parts.

Inferring transcription factor targets from gene expression changes and predicted promoter occupancy

Zhen Xuan Yeo[1], Hock Chuan Yeo[1,2], Joan Keng Suan Yeo[1], Ai Li Yeo[1], Ye Li[1], Neil D. Clarke[1]

[1]Genome Institute of Singapore; [2]Bioprocessing Technology Institute

We have developed a method for inferring condition-specific targets of transcription factors based on ranking genes by gene expression change and ranking genes based on predicted transcription factor occupancy. The average of these two ranks, used as a test statistic, allows target genes to be inferred in a stringent manner. The method complements chromatin immunoprecipitation experiments by predicting targets under many conditions for which ChIP experiments have not been performed. We used the method to predict targets of 102 yeast transcription factors in ~1600 expression microarray experiments. The reliability of the method is suggested by the strong enrichment of genes previously shown to be bound, by the validation of binding to novel targets, by the way transcription factors with similar specificities can be functionally distinguished, and by the greater-than-expected number of regulatory network motifs, such as auto-regulatory interactions, that arise from new, predicted interactions. The combination of ChIP data and the targets inferred from this analysis results in a high-confidence regulatory network that includes many novel interactions. Interestingly, we find only a weak association between conditions in which we can infer the activity of a transcription factor and conditions in which the transcription gene itself is regulated. Thus, methods that rely on transcription factor regulation to help define regulatory interactions may miss regulatory relationships that are detected by the method reported.

Discovering transcriptional modules by combined analysis of expression profiles and regulatory sequences

Yonit Halperin[1]*, Chaim Linhart[1]*, Igor Ulitsky[1], Ron Shamir[1]

* These authors contributed equally to this work.

[1]*Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, ISRAEL*

A key goal of gene expression analysis is the characterization of transcription factors (TFs) and micro-RNAs (miRNAs) regulating specific transcriptional programs. The most common approach to address this task is a two-step methodology: In the first step, a clustering procedure is executed to partition the genes into groups that are believed to be co-regulated, based on expression profile similarity. In the second step, a motif discovery tool is applied to search for over-represented *cis*-regulatory motifs within each group. In an effort to obtain better results by simultaneously utilizing all available information, several studies have suggested computational schemes for a single-step combined analysis of expression and sequence data. Despite extensive research, reverse engineering complex regulatory networks from microarray measurements remains a difficult challenge with limited success, especially in metazoans.

We present Allegro, a new method for *de-novo* discovery of TF and miRNA binding sites through joint analysis of genome-wide expression data and promoter or 3' UTR sequences. In brief, Allegro enumerates a huge number of candidate motifs in a series of refinement phases to converge to high-scoring motifs. For each candidate motif, it executes a cross-validation-like procedure to learn an expression model that describes the shared expression profile of the genes, whose *cis*-regulatory sequence contains the motif. It then computes a *p*-value for the over-representation of the motif within the genes that best fit the expression profile. The output of Allegro is a non-redundant list of top-scoring motifs and the expression patterns they induce.

The expression model used by Allegro is a novel log likelihood-based, non-parametric model, analogous to the position weight matrix commonly used for representing TF binding sites. Unlike most extant methods, our approach does not assume that the expression values follow a pre-defined type of distribution, and can capture transcriptional modules whose expression profiles differ from the rest of the genome across a small fraction of the conditions. Furthermore, it successfully handles cases where the expression levels are correlated to the length and GC-content of the *cis*-regulatory sequences. Such correlations are quite common in practice, and often bias existing techniques, leading to false predictions and low sensitivity.

Allegro introduces several additional unique ideas and features, such as joint analysis of multiple datasets from several species, and is implemented in a graphical, user-friendly software tool. We apply it on several large datasets (>100 conditions), report on the transcriptional modules it uncovers, and show that it outperforms extant techniques. Our analysis reveals a novel motif over-represented in the promoters of genes highly expressed in murine oocytes, and several new motifs relevant to fly development. Finally, using a large stem-cell expression dataset, we identify three miRNA families with pivotal roles in human embryogenesis. Allegro is available at http://acgt.cs.tau.ac.il/allegro.

Combining Transcription Factor Binding Site Clustering and Evolutionary Conservation for Predicting cis-Regulatory Modules

Xin He[1], Xu Ling[1], Saurabh Sinha[1,2]

[1]*Department of Computer Science;* [2]*Institute of Genomic Biology; University of Illinois at Urbana-Champaign*

Two of the most common techniques of identifying cis-regulatory modules (CRMs) are: finding clusters of transcription factor binding sites (TFBSs) through the use of binding motifs, and finding evolutionarily conserved sequence elements. The most powerful strategy would be, naturally, a combination of the two in a single statistical framework. Two important problems with this strategy are: (a) modeling the evolution of TFBSs, in particular, the evolutionary loss and gain of binding sites; and (b) treating possible alignment errors.

We have developed a computational framework, EMMA, for finding CRMs in pair-wise genome comparison, that relies on a comprehensive probabilistic model of CRM evolution. A CRM is modeled as a mixture of background sequences and TFBSs. The evolution of background sequences is modeled by the HKY model of substitution and a stochastic model of indels, adapted from the area of statistical alignment. The evolution of TFBSs follows the Halpern-Bruno model that relates substitution rates to binding specificities. In addition, a site can switch from functional to non-functional state, and vice versa, in a manner that depends on the binding energy of the site. All possible alignments of the two sequences are summed over, with probabilistic weights. By applying our method to CRMs involved in *Drosophila* early development, we found that our method is able to overcome errors of general-purpose alignment tools, such as arbitrary gap placements; and significantly improve the performance of predicting regulatory target sequences from a given TF binding motif.

We also extended the EMMA model to the comparison of multiple genomes related by a phylogenetic tree. Direct generalization is computationally difficult though because: (i) multiple sequence alignment, even without TFBSs, is difficult to solve in a rigorous statistical fashion; (ii) computation of TFBS gain and loss in EMMA requires enumeration of “evolutionary trajectories”, which will become more expensive with multiple sequence comparison. We thus made some simplifications while preserving the spirit of the EMMA model: the evolution of functional and non-functional sites are modeled similarly, but binding site gain and loss is now modeled as a simple two-state Markov chain whose rates are independent of the sequences. The alignment of sequences is fixed. The effect of alignment errors on TFBSs is, however, reduced because such errors will create lineage-specific TFBSs, which are allowed in our model. The new model is applied to two important problems: (i) genome-wide search of target sequences of TFs with given binding motifs; (ii) search of the overrepresented TF motifs in the noncoding sequences of a given set of co-regulatory genes, and simultaneous identification of the regulatory sequences, which could be distant from the immediate promoter sequences.

With the growing number of genomes sequenced and TF motifs identified via protein-DNA binding assays, we believe our framework will provide a powerful means to integrate these data for regulatory sequence analysis.

Genome Wide Set of Human Enhancers

Eddy Rubin

Joint Genome Institute, Lawrence Berkeley National Laboratory

Tissue-specific enhancers are principal regulators of spatiotemporal gene expression and alterations in their activities contribute to many human disorders. Due to their location in non-coding genome regions and limited knowledge about their sequence features, human enhancers have been only minimally annotated. I will describe a large program focused on leveraging extreme evolutionary sequence conservation to identify putative regulatory sequences in the human genome and characterizing their/ in vivo/ enhancer activity in a transgenic mouse assay. To date, we have tested more than 500 such elements including all non-coding human-rodent ultraconserved elements in the human genome. More than 200 of them function as tissue-specific enhancers and reproducibly target gene expression to a broad range of anatomical structures. As a community resource, we have established a database to visualize and query the activity of these enhancer sequences at http://enhancer.lbl.gov/ and will be generating additional data for several thousand enhancers over the next several years.

A cellular resolution atlas of gene expression in Drosophila pseudoobscura reveals interspecies variation in embryonic patterning.

Angela H. DePace[1], Charless C. Fowlkes[4], Cris L. Luengo-Hendriks[6], Lisa Simirenko[2,3], Soile V.E. Keränen[2], Clara Henriquez[2], David W. Knowles[2], Mark D. Biggin[2], Jitendra Malik[5], Michael B. Eisen[2,3]

[1]Department of Systems Biology, Harvard Medical School, Boston MA; [2]Genome Sciences Department, Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA; [3]Department of Molecular and Cell Biology, UC Berkeley, CA; [4]Department of Computer Science, UC Irvine, CA; [5]Department of Computer Science, UC Berkeley, CA; [6]Centre for Image Analysis, Uppsala, Sweden

Understanding how gene regulatory networks evolve requires us to measure the functional consequences of even small changes in sequence. We have applied high-resolution microscopy and image processing methods to blastoderm embryos of *D. melanogaster* and *D. pseudoobscura* to determine the expression patterns of key transcriptional regulators and a subset of their targets in their native context at cellular resolution in 3D over the hour prior to gastrulation. These two species shared their last common ancestor nearly 30 million years ago, and comparative sequence analysis reveals a wide variety of changes in *cis*-regulatory elements. Our imaging techniques allow multiple types of statistically rigorous inter-species comparisons to be made, both between individual embryos and between composite multi-gene models, revealing widespread quantitative changes in expression patterns. We measure multiple types of gene-specific variation, including changes in spatial position, number of cells comprising a pattern, and the dynamics of expression. Our comparative analyses aim to put these differences in the context of complete developing embryos. Which changes are due to differences in the geometry of the embryos and which are due to genetic differences in the transcriptional networks? Furthermore, which are changes initiated by variation in the *trans*-network, and which are due to changes in how *cis*-regulatory sequences interpret that network? Differentiating these types of variation will allow us to interpret which specific sequence changes have functional consequences for gene expression, and provide insights into the functional constraints under which *cis*-regulatory elements evolve.

Mining embryonic expression images reveals novel developmental pathway components

Erwin Frise, Richard Weiszmann, Ann Hammonds and Susan E. Celniker.

Berkeley Drosophila Genome Project, Lawrence Berkeley Natl. Labs, Berkeley, CA 94720

Analyzing both temporal and spatial gene expression is essential for understanding the development and regulatory networks of multicellular organisms. Interacting genes are commonly expressed in overlapping or adjacent domains. Thus, gene expression patterns can be mined to infer candidates for networks.

We have generated a systematic 2D mRNA expression atlas profiling embryonic development of *Drosophila melanogaster*. To date, we have collected over 70,000 images for over 6,000 genes. To explore spatial relationships between gene expression patterns, we used a novel computational image processing approach by converting expression patterns from the images into virtual representations. Using a custom designed automated pipeline, for each image we segmented and aligned the embryo to an elliptically shaped mesh, comprised of 311 small triangular regions each defining a unique location within the embryo. By comparing corresponding triangles, we produced a distance score to identify similar patterns.

We used the virtualized expression patterns (VEP) to conduct a comprehensive analysis of the expression landscape. We filtered and automatically compacted and temporally sorted multiple images in each stage range using a novel approach combining clustering and graph theory. For developmental stage 4-6, we reduced the initial set of about 5,800 to 553 VEP containing 364 genes. To discover the range of unique patterns, we used affinity propagation clustering and identified 39 clusters each representing a distinct pattern class. We integrated the remaining genes into the 39 clusters and compared the clusters to genetic interactions, functional annotations and known protein interactions. While these data sets generally overlap, some surprising exceptions were found, such as genetic interactions anti-correlated between a ventral and dorsal class.

Clustered expression patterns were used to identify putative positive and negative regulatory interactions. The similar VEP in each cluster not only grouped already known genes from the same pathway, but previously undescribed genes. A comparative analysis identified subtle differences between the genes within each expression cluster. To investigate these differences we developed a novel Markov Random Field (MRF) segmentation algorithm to binarize the patterns. We then extended the MRF algorithm to detect shared expression boundaries, generate similarity measurements and discriminate even faint/uncertain patterns between two VEPs. This enabled us to identify more subtle partial expression pattern overlaps and adjacent non-overlapping patterns. For example, by conducting this analysis on the cluster containing the gene *snail*, we identified the previously known *huckebein*, which restricts *snail* expression, and *zfh1*, which interacts with *tinman*.

Representing expression patterns with geometric meshes facilitates the analysis of a complex process involving thousands of genes. We have demonstrated that it can be used for predicting relationships in regulatory and developmental pathways.

Comparative Analysis of Enhancers and Regulatory Motifs for Gene Expression in the Vertebrate Brain

Qiang Li[1] , Deborah Ritter[2], Nan Yang[1], Zhigiang Dong[1], Hao Li[3], Jeffrey H. Chuang[2], Su Guo[1]

[1]Department of Biopharmaceutical Sciences, UCSF; [2]Department of Biology, Boston College; [3]Department of Biochemistry and Biophysics, UCSF;

The vertebrate genome contains a vast amount of non-coding sequence with strong conservation across the vertebrate phylogeny. Deciphering the functions and mechanisms of action of these non-coding sequences is a challenging, but important, problem. Many Conserved Noncoding Elements (CNEs) are thought to regulate vertebrate gene function as transcriptional enhancers, capable of controlling gene expression either spatially and temporally. However, the tissue- or timing- specificity of CNE enhancers is not known *a priori*.

Here we report a combined computational/experimental strategy in which we used the developing vertebrate brain as an example and identified pattern-associated CNEs conserved between human and zebrafish. By selecting CNEs adjacent to genes of known developmental expression pattern, we were able to efficiently identify CNEs with brain enhancer activity in a zebrafish reporter system. These enhancer experiments comprise studies of more than 100 zebrafish CNEs. Application of *de novo* motif prediction algorithms on forebrain enhancers uncovered short sequences that were experimentally validated as critical codes for forebrain enhancer activity. In addition, we assessed cross-species expression of orthologous CNE sequences from zebrafish, human and mouse, via experiments in zebrafish and mouse embryos. A significant fraction of CNE sequences exhibited divergent tissue-specific enhancer activity across species, suggesting that positive selection and host-specific effects are common even for these highly conserved sequences.

We have built a database of experimental images and annotations on all CNEs we have tested, which we will continue to develop as a public resource for zebrafish CNE studies (cneBrowser). In addition, we have created an online tool (cneViewer – cneviewer.zebrafishcne.org) to simplify the process of selecting CNEs for experimental validation based on the anatomical and timing-specific expression of nearby genes. These websites provide user-friendly resources to facilitate information transfer and analysis for deciphering the noncoding functions in vertebrate genomes.

Global patterns in tissue- and factor-specific RNA processing

Eric Wang[1], Rickard Sandberg[1,2], Xinshu Xiao[1,3], Razvan Nutiu[1], Christin Mayr[4], Stephen Kingsmore[5], Gary Schroth[6], Chris Burge[1,7]

[1]Department of Biology, MIT, Cambridge, MA; [2]Department of Cell and Molecular Biology, Karolinska Institute, Stockholm, Sweden; [3]Department of Physiology, UCLA; [4]Whitehead Institute, Cambridge, MA; [5]National Center for Genome Resources, Santa Fe, NM; [6]Illumina, Inc., Hayward CA; [7]Department of Biological Engineering, MIT.

Through alternative processing of pre-mRNAs, individual mammalian genes often produce multiple mRNA and protein isoforms that may have related, distinct or even opposing functions. We have conducted an in-depth analysis of 15 diverse human tissue and cell line transcriptomes based on deep sequencing of cDNA fragments (mRNA-SEQ), yielding a digital inventory of gene and mRNA isoform expression[1]. Analysis of mappings of sequence reads to exon-exon junctions indicate that 92-94% of human genes undergo alternative splicing (AS), most with a minor isoform frequency of 15% or more. Differences in isoform-specific read densities indicated that a majority of AS and of alternative cleavage and polyadenylation (APA) events vary between tissues, while variation between individuals was ~2- to 3-fold less common. Extreme or 'switch-like' regulation of splicing between tissues was associated with increased sequence conservation in regulatory regions and with generation of full-length open reading frames. Patterns of AS and APA were strongly correlated across tissues, suggesting coordinated regulation of these processes, and sequence conservation of a subset of known regulatory motifs in both alternative introns and 3' UTRs suggested common involvement of specific factors in tissue-level regulation of both splicing and polyadenylation. We are also studying in depth the regulatory properties of the widely expressed splicing factor hnRNP H, using RNAi knockdown followed by deep sequencing of mRNAs (KD-SEQ) and other approaches. Our results indicate that hnRNP H activates splicing from intronic locations, and represses splicing from exonic locations in a manner that is highly dependent on the strength (but not precise sequence) of the adjacent 5' splice site. These observations have important implications, in particular suggesting that hnRNP H functions as an evolutionary capacitor of alternative splicing evolution.

[1] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge (2008) Alternative isoform regulation in human tissue transcriptomes. Nature (in press).

Discovering Structural Cis-Regulatory Elements by Modeling the Behaviors of mRNAs

Barrett C. Foat[1], Gary D. Stormo[1]

[1]*Department of Genetics, Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108*

Gene expression is regulated at each step from chromatin remodeling through translation and degradation. Yet, most efforts to understand the regulation of gene expression have focused on transcription and DNA-binding regulatory proteins. While regulatory RNAs have received appreciable attention, regulatory elements within mRNAs that are recognized by nucleic acid binding proteins have been largely ignored until recently. Several known RNA-binding regulatory proteins interact with specific RNA secondary structures in addition to specific nucleotides. Thus, to provide a more comprehensive understanding of the regulation of gene expression, we developed a novel, alignment-free computational approach that leverages functional genomics data and nucleotide sequences to discover RNA secondary structure-defined *cis*-regulatory elements (SCREs).

We applied our Structural *Cis*-Regulatory Element Detector (StructRED) to microarray and mRNA sequence data from *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Homo sapiens*. We recovered the known specificities of Vts1 in yeast and Smaug in flies, two sterile alpha motif-containing RNA-binding proteins. In addition, we discovered six more SCREs in flies and three in humans. We characterized the condition-specific regulatory activities of the factors that recognize each SCRE by following the strength of their regulatory effects across multiple microarray experiments. We characterized the SCRE-containing mRNAs based on available annotations, including Gene Ontology, phenotype, and *in situ* expression. We show that expression-predictive instances of many of the SCREs, including Smaug binding sites, are most often found in coding sequences. In addition, the regulatory effects of the *Drosophila* SCREs are conserved in other *Drosophila* species. Finally, we provide evidence that Smaug's mRNA stability and translation-regulating effects are independent. Overall, modeling functional genomics data in terms of combined RNA structure and sequence motifs is an effective method for discovering the binding specificities and regulatory roles of RNA-binding proteins.

High-resolution transcriptome analysis reveals prevalent regulatory logic interspersed within prokaryotic coding sequences

David J Reiss[1], Tie Koide[1], Nitin S Baliga[1]

[1] *Institute for Systems Biology*

Despite prior knowledge of complex prokaryotic transcription mechanisms, generalized rules, such as the organization of genes into operons with well-defined promoters and terminators, still play a significant role in systems analysis of regulatory logic in bacteria and archaea. Here, we have investigated the prevalence of alternate regulatory mechanisms through high-resolution, genome-wide characterization of transcript structures of 64% of all genes, plus putative non-coding RNAs, in the archaeon *H. salinarum NRC-1*. This study has revealed widespread environment-dependent modulation of operon architectures, transcription initiation and termination inside coding sequences, and extensive overlap in 3' ends of many convergently transcribed transcripts. A significant fraction of these alternate transcriptional events correlate with binding locations of 12 transcription factors (TFs) inside genes and operons – events usually considered spurious or non-functional. With experimental validation, we illustrate the prevalence of overlapping genomic signals in prokaryotic transcription, casting doubt on the general perception of rigid boundaries between coding and regulatory sequence.

We recently constructed a global gene regulatory network (GRN) model for *H. salinarum NRC-1* that accurately predicts transcriptional changes in 80% of all genes to new environmental and genetic perturbations [1]. Using biclustering and statistical inference algorithms that integrated diverse system-wide data types, we predicted the combinatorial and conditional regulation of genes by multiple TFs and environmental factors. Although several of the inferred regulatory influences in this network were demonstrated to be likely mediated through direct interactions with the promoters of regulated genes, a large number of influences are thought to be indirect. The logical next step is to make this quantitative and predictive network model also mechanistically accurate on a systems-scale.

Here, we report a significant step toward that goal by characterizing, at high resolution, the dynamic remodeling of the transcriptome structure of *H. salinarum NRC-1* during a complex cellular response, and correlating these changes to genome-wide binding locations of 10% of all predicted TFs. Specifically, we have (i) defined transcription start sites and termination sites for ~64% of the genes – including new and revised protein coding genes; (ii) discovered at least 61 novel ncRNA candidates; (iii) quantified 5' and 3' untranslated regions; (iv) identified functional promoters upstream and internal to coding regions; (v) discovered instances of transcription termination inside coding regions; (iv) uncovered the prevalence of mRNA populations with variable 3' end locations; (v) found extensive overlaps in the 3' ends of many convergently transcribed transcripts; and (vi) discovered that the lengths of a significant fraction of operon-encoding transcripts are variable and condition-dependent. These findings suggest that the construction of mechanistically accurate GRN models will require treating genes, operons, promoters, and terminators as dynamic entities.

[1] Bonneau, R., Facciotti, M.T., Reiss, D.J., Schmid, A.K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M.H., Bare, J.C., et al. (2007). A predictive model for transcriptional control of physiology in a free living cell. Cell 131, 1354-1365.

Gene Translation Efficiency in Healthy and Cancerous Human Tissues

Yedael Y. Waldman[1]*, Tamir Tuller[1-3]*, Tomer Shlomi[1], Shaul Karni[1], Roded Sharan[1] and Eytan Ruppin[1,3]

[1]School of Computer Sciences; [2]Department of Molecular Microbiology and Biotechnology; [3]School of Medicine, Tel Aviv University

Previous studies have left the question of gene translation efficiency (TE) in humans an intriguingly open one. We conduct a comprehensive analysis of TE in humans using data on tRNA copy numbers and tissue-specific gene expression measurements.

We find strong evidence for the efficiency of gene translation in humans, supporting the hypothesis that it has evolved under selection (vs. a neutral scenario). Different tissues have fairly different overall translation efficiency levels, with the heart, lung and liver having the largest scores. We find marked correlations between the translation efficiency of a gene, measured via its tRNA adaptation index (tAI) (a measure of its TE), and its functional importance, assessed via its expression breadth across tissues, its evolutionary rate, degree in protein interaction network and protein length. A (yet small scale) study indicates that adult tissues have higher translation efficiency scores than fetal tissues, suggesting that the genomic copy number is adapted to the adult period. Accordingly, we find that the tRNA pool is geared towards efficient translation.

Interestingly, studying two types of cancer with considerable chromosomal alterations in the genomic tRNA pool, we find that these alterations lead to a significant increase in translation efficiency compared with the TE levels that would have been obtained if the original pool would remain unaltered. This finding should motivate future genome-wide studies of translation efficiency alterations in cancer.

* Y.W. and T.T contributed equally to this work.

Genomic predictors of interindividual differences in response to DNA damaging agents.

Leona D. Samson

MIT Biological Engineering Department and Center for Environmental Health Sciences

Human lymphoblastoid cells derived from different healthy individuals display considerable variation in their transcription profiles. Here we show that such variation in gene expression underlies interindividual susceptibility to DNA damaging agents. The results demonstrate the massive differences in sensitivity across a diverse cell line panel exposed to an alkylating agent. Computational models identified 48 genes with basal expression that predicts susceptibility with 94% accuracy. Modulating transcript levels for two member genes, MYH and C21ORF56, confirmed that their expression does indeed influence alkylation sensitivity. Many proteins encoded by these genes are interconnected in cellular networks related to human cancer and tumorigenesis.

Repression at the single-nucleosome level

Jeffrey A. Rosenfeld[1,2,3], Zhenyu Xuan[1], Rob DeSalle[2,3], Michael Q. Zhang[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; [2]Department of Biology, New York University; [3]American Museum of Natural History, New York, NY

Eukaryotic DNA is wrapped around nucleosomes consisting of histone proteins. Many residues of histones can be post-translationally modified. In particular, methylation of lysine residues has been extensively investigated. Most of the research has concerned modifications that are found at actively transcribes genes including H3K4me and H3K36me. In contrast, repressive modification have received significantly less attention. For this reason we have focused on the analysis of H3K9me3 a modification that traditionally been considered repressive and associated with heterochromatin. A strong reason for this characterization is the finding of a strong affinity between H3K9me3 and heterochromatin protein 1 (HP1). More recently, evidence has been found for the presence of H3K9me3 at some actively transcribed genes.

We have used high-throughput Solexa sequencing data for H3K9me3 to investigate its locations throughout the entire human genome. We have looked at promoters, entire gene bodies and the ends of genes and determined that overall a small percentage of genes have significant levels of H3K9me3. Of the genes with H3K9me3 enriched, it is much more prevalent at the ends of genes than at promoters. We have also used GO analysis to determine that the functional categories of genes containing H3K9me3. The strongest represented categories are transcription factors with an enrichment for zinc-finger proteins.

For non-genic regions of the genome we have used our nucleosome level resolution to give a clear demarcation of locations containing and lacking H3K9me3. We have investigated the presence of barriers or insulators at the edges of H3K9me3 regions. In addition, we have compared our identified locations with known locations of different types of genomic repeats as well as different types of genomic staining bands to determine their level of overlap with H3K9me3.

Targeted screening of cis-regulatory variation in human haplotypes

D. J. Verlaan[1,2,3,5], B. Ge[2,5], E. Grundberg[1,2], R. Hoberman[1], K. Lam[2], V. Koka[2], J. Dias[2], S. Gurd[2], N. Martin[2], H. Mallmin[4], O. Nilsson[4], E. Harmsen[2], T. Kwan[2] and T. Pastinen[1,2]

[1]Department of Human Genetics, McGill University, Montréal, Canada; [2]McGill University and Genome Québec Innovation Centre, Montréal, Canada; [3]Hôpital Ste-Justine, Université de Montréal, Montréal, Canada; [4]Department of Surgical Sciences, Uppsala University, Uppsala, Sweden; [5]These authors have contributed equally to this work

Regulatory cis-acting variants account for a large proportion of gene expression variability in populations. Cis-acting differences can be specifically measured by comparing relative levels of allelic transcripts within a sample. Allelic expression (AE) mapping for cis-regulatory variant discovery has been hindered by the requirements of having informative or heterozygous SNPs within genes in order to assign the allelic origin of each transcript. In this study, we have developed an approach to systematically screen for heritable cis-variants in common human haplotypes across >1000 genes. In order to achieve the highest level of information per haplotype studied, we carried out allelic expression measurements by using exonic and intronic SNPs in primary transcripts. We employed a novel RNA pooling strategy in immortalized lymphoblastoid cell lines (LCLs) and primary human osteoblast cell lines (HObs) to allow for high-throughput AE. Screening hits from RNA pools were further validated by performing allelic expression mapping in individual samples. Our results indicate that >10% of expressed genes in human LCLs harbour common cis-acting variants. In addition, we have validated cis-acting variants in over 20 genes linked with common disease susceptibility in recent genome-wide studies. More generally, our results indicate that RNA pooling coupled with AE read-out by 2nd generation sequencing or by other methods provides a high-throughput tool for cataloguing the impact of common non-coding variants in the human genome.

Linkage analysis of inferred transcription factor activity reveals regulatory networks

Eunjee Lee[1] and Harmen J. Bussemaker[1,2]

[1]Department of Biological Sciences, Columbia University, New York, NY 10027, USA; [2]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA

The genomewide pattern of steady-state mRNA abundance depends on the regulatory activity of transcription factors (TFs). This activity is usually not well predicted by the mRNA expression level of the TF, as it is heavily modulated at the post-translational level through non-covalent modification by signaling proteins, changes in the subcellular localization of the TF protein, and the availability of co-factors. The fact that genetic variation is likely to perturb many of these processes provides a strategy for mapping the molecular networks between TFs and their upstream modulators. We analyzed the genetics of transcription factor activity in the yeast *S.cerevisiae* using genotype and gene expression data of segregants between BY4716 (BY; laboratory strain) and RM11-1a (RM; a natural isolate) strains from the study performed by Brem and Kruglyak [1]. First, we inferred the regulatory activity in each segregant for a large number of TFs, using the *MatrixREDUCE* software developed by our laboratory, which uses a biophysical model to relate the mRNA expression levels of target genes to their upstream regulatory sequence. Next, we identified chromosomal loci (referred to as activity quantitative trait loci, or “aQTLs”) whose genotype affects the activity of each TF. We found that about a third of the transcription factors surveyed have at least one significant aQTL. In the majority of the cases this regulatory relationship could not be detected using the mRNA expression level of the TF (*i.e.*, by an “eQTL” approach), underscoring the importance of accounting for post-translational modulation of TF activity. Finally, we revealed molecular mechanisms underlying the identified genetic associations by using interactome data (protein-protein, chromatin modifier-transcription factor, an kinase-substrate) to connect TFs to specific genes in their aQTLs. Our approach recovers known regulatory mechanisms for transcription factor activity modulation, and provides a rich set of novel predictions regarding the functional connectivity between signaling networks and transcriptional networks.

[1] Brem, R. B. and L. Kruglyak (2005). "The landscape of genetic complexity across 5,700 gene expression traits in yeast." Proc Natl Acad Sci U S A **102**(5): 1572-1577.

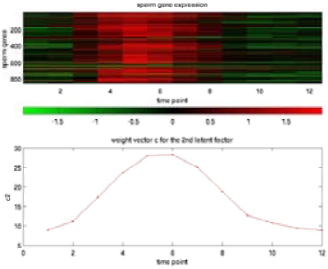
Learning regulatory motifs from gene expression trajectories using graph-regularized partial least squares regression

Xuejing Li[1], Chris H. Wiggins[2], Valerie Reinke[3], Christina Leslie[4]

[1]*Department of Physics, Columbia University;* [2]*Department of Applied Physics and Applied Mathematics, Columbia University;* [3]*Department of Genetics, Yale University;* [4]*Computational Biology Center, Memorial Sloan-Kettering Cancer Center*

A gene's mRNA expression level is determined by multiple input signals that are integrated by the cis regulatory logic encoded in its promoter sequence. Genes whose regulatory sequences contain similar DNA motifs are likely to have correlated expression profiles, and typical motif discovery approaches rely on first clustering genes by expression in search of potentially coregulated sets. Realistically, however, distinct regulatory programs may lead to similar patterns of differential expression. We present an algorithm that models the natural flow of information, from sequence to expression, to learn cis regulatory motifs and characterize gene expression patterns. We use a novel algorithm based on partial least squares (PLS) regression to learn a mapping from the set of *k*-mers in a promoter to the expression profile of the gene across experiments. We also introduce a graph-regularized version of the PLS algorithm to enable motif discovery by imposing two constraints: a lasso constraint for sparsity and a graph Laplacian constraint for smoothness over sequence-similar motifs.

We run PLS regression on a time series gene expression data for wild-type germ line development in *C. elegans* and demonstrate that the minimal mean squared prediction error is obtained with 4 latent factors. In particular, we study two gene sets consisting of sperm and oocyte genes. The first and second latent factor account for the largest reduction in loss for oocyte and sperm genes, respectively. PLS learns a correspondence between each latent factor and a weight vector ***c*** in the output space, which gives the weights over time points. Thus ***c*** can be interpreted as an expression pattern, and we show that oocyte and sperm gene expression profiles are strongly correlated with the $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$, respectively (Figure 1). Additionally, each input weight vector ***w*** in PLS gives weights over *k*-mers, and highly weighted *k*-mers should be relevant for the gene expression pattern. We plot *k*-mer graphs corresponding to the first two latent factors. From the second factor, we successfully find the ELT-1 motif GATAA and bHLH-1 motif ACGTG for the sperm genes (Figure 2). From the first factor, we learn CG-rich *k*-mers that are highly enriched in oocyte gene promoters. We also investigate motif conservation between *C. elegans* and *C. briggsae*, and find that these *k*-mers are highly conserved in their corresponding gene sets.



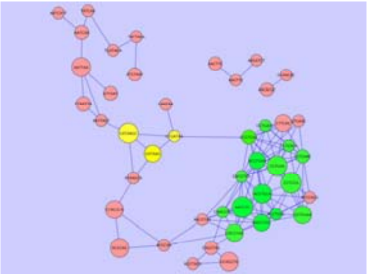


Figure 1: Sperm gene expression versus c_2. **Figure 2: *k*-mer graph for the second latent factor.**

GADEM: A genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery

Leping Li

Biostatistics Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, North Carolina 27709, USA

Genome-wide analyses of protein binding sites generate large amounts of data; a ChIP dataset might contain 10,000 sites. Unbiased motif discovery in such datasets is not generally feasible using current methods without restricting the initial motif profiles. We propose an efficient method, GADEM, which combines spaced dyads and an expectation-maximization (EM) algorithm. Candidate words (4-6 nucleotides) for constructing spaced dyads are prioritized by their degree of overrepresentation in the input sequence data. Spaced dyads are converted into starting position weight matrices (PWMs). GADEM then employs a genetic algorithm (GA), with an embedded EM algorithm to improve starting PWMs, to guide the evolution of a population spaced dyads toward one whose entropy scores are more statistically significant. Spaced dyads whose entropy scores reach a pre-specified significance threshold are declared motifs.

We applied GADEM to six genome-wide ChIP datasets. Approximately, 15 to 30 motifs of various lengths were identified in each dataset. Remarkably, without any prior motif information, the expected known motif (e.g., P53 in P53 data) was identified every time. Unlike any other *de novo* motif discovery tools, GADEM can also identify long (>40bp) motifs. For instance, in the P53 ChIP data, it identified several abundant long motifs (70-100bp) that happened to correspond to retroelements. GADEM discovered motifs of various lengths (6-100bp) and characteristics in these datasets containing from 0.5 to >13 millions nucleotides with run times of 3 to 72 hours. We believe that GADEM is an efficient tool for *de novo* motif discovery in large scale genome-wide data.

[Supplementary material and C code are available at:
http://www.niehs.nih.gov/research/resources/software/gadem]

A systematic characterization of Drosophila transcription factors via a bacterial one-hybrid system

Marcus B. Noyes[1,2], Ryan Christensen[4], Xiangdong Meng[1], Atsuya Wakabayashi[1], Gary D. Stormo[4], Saurabh Sinha[5], Michael H. Brodsky[1,3] and Scot A. Wolfe[1,2]

[1]Program in Gene Function and Expression, [2]Department of Biochemistry and Molecular Pharmacology, and [3]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA. [4]Department of Genetics, Washington University, School of Medicine, St. Louis, MO 63110, USA. [5]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Complementary experimental and computational approaches will be required to deconvolute the regulatory networks that control the spatial and temporal patterns of transcription in complex genomes. DNA-binding specificity data for a transcription factor (TF) or set of TFs can be used to computationally identify evolutionarily-conserved binding sites throughout a genome via comparisons with syntenic regions of the genomes of related species. One critical limitation for this type analysis is the absence of specificity data for most TFs within a genome. Using a recently described bacterial one-hybrid system, we have characterized the specificities of over 100 *Drosophila* TFs. These factors are focused on two sets: TFs that are involved in early regulation of anterior-posterior (A-P) patterning and 84 independent members of the homeodomain family. As a complement to this dataset, we have developed web-based tools (biotools.umassmed.edu/genomesurveyor) to identify cis-regulatory modules (CRMs) throughout the *Drosophila* genome using these TF specificities coupled with phylogenetic comparisons. These tools allow a user to either explore individual genomic regions (using Gbrowse) or perform genome-wide searches for clusters of over-represented sites using any TF combination. Using these tools known target genes and *cis*-regulatory modules for many TFs in this dataset can be predicted. The comprehensive analysis of a family of DNA-binding domains, such as the homeodomains, allows a unique, specificity-based analysis of this family. Correlation of DNA-binding specificities with residues at the protein-DNA interface reveals new recognition principles that allow more elaborate reengineering of homeodomain specificities. Importantly, this dataset can serve as a template for the prediction of specificities of the majority of homeodomains in eukaryotic genomes. Utilizing our bacterial one-hybrid system it should be possible to characterize the specificity of most TFs in the fly genome with applications for broadly predicting CRMs in the fly and for predicting TF specificities in other eukaryotic species for studies on their transcriptional regulatory networks.

Protein-nucleic acid interaction mapping

Timothy R. Hughes

Banting and Best Department of Medical Research, Department of Molecular Genetics, and Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON CANADA

One of the major impediments to our understanding of genome function is lack of accurate and complete knowledge of how proteins interact with nucleic acid. Recent technical advances now make it possible to assay relative preference of proteins for tens of thousands of distinct sequences *in vitro*, thus removing confounding variables at work in vivo. Current work in my laboratory is aimed at cataloguing descriptions of protein interactions with DNA and RNA *in vitro* on a genomic scale. We anticipate that these data will be invaluable in formulating models that explain how cells recognize genomic features, how gene regulatory networks are organized, and how genomes evolve.

System-Identification of Gene Regulatory Networks by Systematic Gene Perturbation Analysis

Vida Abedi[1], Dawn Fraser[3], Mila Tepliakova[1], Simon St-Pierre[1], Mads Kaern[1,2,3]

[1]Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Canada; [2]Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Canada; [3]Department of Physics, University of Ottawa, Ottawa, Canada

A key goal of systems biology is to develop improved methods to map complex biological regulatory systems providing insight into how biomolecular interaction networks propagate and process information. Systems Biology studies typically aims to generate new knowledge by integrating experimental inquiry, computational data analysis and model-based hypothesis generation. While methods developed in engineering could in principle be utilized to analyze biological data, the high dimensionality and often qualitative nature of experimental datasets, as well as the predominance of non-linear co-dependencies among biomolecules, make the task of adapting these methods to a biological context very challenging. The development of novel approaches for systems-level characterization thus requires not only new computational methodologies to analyze complex datasets, but also the design of new experimental assays that can provide the appropriate types of data. Here, we have developed a system-identification framework for the extraction of quantitative and mechanistic information about causal relationship among genes using the canonical galactose utilization pathway in *Saccharomyces cerevisiae*. The developed methodology, referred to as systematic gene perturbation analysis (SGPA), is based on the effects of systematic pair-wise gene deletions. This methodology is inspired by a classical engineering approach to gain insight into the properties of a "controller" by measurements of input/output relationships under different conditions. In essence, the method establishes dynamical models of the regulatory network from single-cell measurements of steady-state input-output relationships, in systematically perturbed networks without *a priori* knowledge. SGPA framework succeeds in identifying, with no prior knowledge, the main regulatory interactions among genes in the galactose utilization pathway. This strategy leads the way to a better application of available resources and provides a scalable framework for system-identification and reverse-engineering of biological networks based on in vivo systematic data generation. The approach could potentially be applied to other networks in yeast as well as higher organisms.

Evaluating regularization methods for inferring Gene Networks by a linear perturbation strategy

Animesh Acharjee[1], Alberto Rezola Urquia[2], Alberto de la Fuente[3], Chris Maliepaard[4]

[1,4]Wageningen University and But the Research Center, Laboratory of Plant Breeding, The Netherlands; [2]University of Navarra, Spain; [3]CRS4 Bioinformatica in Sardinia, Italy

In the post-genomic era, biology has undergone a transition from focusing on simple, small components of cells such as DNA, RNA and proteins in isolation, to the analysis of relationships and interactions between various parts of a biological system. One of the fundamental tasks in the post-genomic era is to understand the Gene Networks [1], controlling cellular functions and uncovering the underlying mechanisms.

We have applied a linear dynamical modeling framework to simulated heterozygous knock-out data [2, 3]. This approach had a winning performance in the DREAM2 competition.

In this research, we improve upon the algorithm by are employing 'regularization' methods and evaluate the algorithms using simulated data sets [7]. Different 'regularization' methods like ridge [4], lasso [5], and elastic net [6] were incorporated and tested on different sizes of the networks consisting of 10, 50, 100 genes. Regularization parameters in three methods were optimized to improve the results. We evaluated the methods in terms of the areas of the precision and recall (PvR) and receiver operator (ROC) curves. Results show that elastic net method outperforms the other methods.

[1] Brazhnik, P., A. de la Fuente & P. Mendes. 2002. Gene networks: how to put the function in genomics*. Trends Biotechnol*. 20: 467-72.
[2] de la Fuente, A., P. Brazhnik & P. Mendes. 2002. Linking the genes: inferring quantitative gene networks from microarray data*. Trends Genet.* 18: 395-398.
[3] de la Fuente, A. & D.P. Makhecha. 2006. Unravelling gene networks from noisy under-determined experimental perturbation data*. Syst Biol (Stevenage).* 153: 257-62.
[4] Tikhonov, A.N. (1963) Dokl. Akad. Nauk SSSR, 153(1963) 49-52, MR 28# 5577.
[5] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B, 58(1):267–288
[6] Zou, H. and Trevor Hastie,T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B,* 67(2), 301-320.
[7] Soranzo, N (2007). Comparing association network algorithms for reverse engineering of large scale gene regulatory networks: synthetic vs real data. *Bioinforamtics.* Vol.00 no. 00 :1-7

Phosphoproteomic-derived pathway maps using Integer Linear Programming (ILP)

Alexander Mitsos[1], Julio Saez-Rodriguez[2,4], Paraskevas S Siminelakis[3], Leonidas G Alexopoulos[3]

[1]Aachen Institute for Advanced Study in Computational Engineering Science, RWTH Aachen, Aachen, Germany; [2] Harvard Medical School, Boston, MA; [3]Department of Mechanical Engineering, National Technical University of Athens, Athens, Greece; [4]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA

Signaling pathways maps, assembled either from manual literature search or from automated text mining algorithms, are the cornerstone for understanding molecular systems for applications in the medical or pharmaceutical arena. Several signaling databases (i.e. Reactome, Pathway Interaction Database, Pathway Commons, KEGG, STKE, Pathway Studio, Ingenuity, etc) have been developed and coupled to gene expression profiling which provided the experimental front for correlating biological functions and/or diseases to the genes in the network (i.e. Ingenuity, Pathway studio). On the other side, advances on high-throughput protein assays (i.e. Luminex platform, mass spectrometry-based protein phosphorylation) gave the opportunity for broad exploration of signaling networks using phosphoproteomic datasets (see also DREAM3 - Challenge 2: The Signaling-Response Prediction Challenge). Despite their limited coverage of the whole signaling network, phosphoprotein measurements are the ultimate reporters of protein function, and thus can carry significant biological insight. In an effort to simulate the information flow on this dataset, Klamt et al. have developed Boolean (logical) descriptions and Saez-Rodriguez et al. have developed a methodology for automatic identification of optimal pathway models based on high-throughput phosphoproteomic data and prior knowledge. Here, we present an alternative method of optimal pathway identification based on an Integer Linear Program (ILP) formulation. Three types of variables are introduced: the first type models the presence of reactions (connectivity existence); the second type models whether reactions take place in each experiment; the third type models the occurrence of species (signaling molecules) in each experiment. The constraints of the optimization problem are formulated based on two logical rules: 1) a reaction takes place if and only if all of its signaling molecules are present and its inhibitor is absent and 2) if a reaction takes place, all its products are formed. The objective of the optimization problem is to minimize the discrepancy between prediction and measurements. A secondary objective is to obtain the smallest possible pathway without a sacrifice of goodness of predictions. The ILP is solved with state-of-the-art commercial codes to guaranteed global optimality; this ensures that the best possible fit is obtained. The computational requirements are manageable, in the order of minutes. Thus, ILP might represent an alternative method to determine optimal Boolean networks with a much higher efficiency which makes it scalable to much larger network topologies. If that proves true, it can provide scientists with a powerful tool for functional pathway analysis to better understand the complexity of biological systems.

Computational Cis-Regulatory Transcriptional Grammar and Developing Predictive Mathematical Models.

Ahmet Ay[1], Walid D. Fakhouri[2], Chichia Chiu[1], David N. Arnosti[2]

[1]Department of Mathematics and [2]Department of Biochemistry&Molecular Biology

Michigan State University, East Lansing, MI, 48824, USA

Cis regulatory information comprises a key portion of genetic coding, yet despite the abundance of genomic sequences now available, identifying and characterizing this information remains a major challenge. We are pursuing a unique “bottom up” approach to understand the mechanistic processing of regulatory elements (input codes) by the transcriptional machinery, using a well defined and characterized set of repressors and activators in Drosophila blastoderm embryos. We are identifying quantitative values for parameters affecting transcriptional regulation in vivo, and these parameters are used to build and test mathematical models that predict the outputs of novel cis-regulatory elements.

Giant, Kruppel, Knirps are short-range transcriptional repressors involved in the developmental patterning of Drosophila blastoderm embryo. Using defined regulatory modules tested in germ line transformed embryos, we are measuring quantitative parameters describing the effects of spacing, stoichiometry, arrangement and binding site affinities of these repressors on cassettes driven by endogenous activators. To develop predictive models we employ a hybrid thermodynamic gene regulation model in which we included variables for the spacing effect between activators and repressors. This model is being used to predict the output of novel permutations of binding sites, which will allow us test and refine parameters used. In one line of investigation, fluorescence quantization of lacZ reporter gene expression was used to measure the effect of moving Giant repressor binding sites from a position adjacent to Twist/Dorsal activator sites to a distal site 125bp away. Our mathematical model successfully predicted the distance effect of intermediate positions, such as 25, 50, 75 and 100bp compared to experimental results. Further tests will illustrate the effects of site permutation, transcription factor concentration and activator/repressor stoichiometry and arrangement. Extension of these predictive models to endogenous cis elements will provide novel insights on regulatory element design and evolution, and should provide a bioinformatics method for predicting quantitative output of novel regulatory elements.

Power and Study Design Considerations for Multivariate Classification in Systems Biology Experiments

Raji Balasubramanian[1], Yu Guo[2], Armin Graber[2]

[1]School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA;
[2]BG Medicine, Inc., Waltham, MA

Data from high throughput Data from high throughput systems biology experiments profiling transcripts, metabolites and proteins are characterized by a large number of measurements (p) made per sample (n), where p >> n. In addition, due to the discovery nature of the experiments, these data typically contain only a small proportion of all measured features that are associated with the outcome of interest – thus, computational methods employed in the analysis of such experiments often involve procedures to effectively identify the subset of biomolecular entities that are significantly associated with the outcome (phenotype). Design of such high throughput experiments must take into account the high dimensional nature of the resulting data and the subsequent procedures employed to reduce dimensionality. Previous work by Dobbin et al. (2007) describes an algorithm for determining sample size and power that incorporates both the feature reduction and predictor development procedures. In this work, the authors assume that the data are normally distributed and that feature reduction is based on prioritization of features resulting from a t test. In our work, we present simulation results that compare the performance of three commonly used classifiers (Random Forests, PAM and Support Vector Machines) to the algorithm described in Dobbin et al. (2007), where feature reduction is based on a recursive feature elimination procedure. We evaluate the effects of (a) varying proportion of true signal in the data and of (b) violations of the normality assumptions on the resulting power of the study. To support our simulations and to guide study design of future systems biology experiments, we also present a summary of the characteristics of metabolomics and proteomics data profiled across 6 animal and 2 human studies using a variety of mass spectrometry based techniques. As expected, our simulations reveal that statistical power is adversely affected with decreasing proportion of true signal in the data. Moreover, our simulation results reveal that when the data distributions are non-normal, all three multivariate techniques resulted in decreased power - however, this effect was observed to be most severe in PAM.

Inference of the human and mouse interactome from massive data using a parallel-computing inference algorithm.

V. Belcastro[1,3], L. Cutillo[1], F. Gregoretti[4], G. Oliva[4], D. di Bernardo[1,2]

[1]Telethon Institute of Genetics and Medicine, Via P. Castellino, Naples, Italy; [2]Dipartimento di Informatica e sistemistica. Universitá degli studi di Napoli Federico II; [3]The Open University, PO Box 197, MK7 6BJ, Milton Keynes, United Kindom; [4]Institute of High Performance Computing and Networking ICAR-CNR, Naples, Italy.

The advent of microarray technologies to measure gene expression, as well as, the ease in obtaining large amount of data via public repositories, makes it possible to infer a genome-wide interactome via the analysis of thousands of microarrays. An interactome is a map of all direct and indirect regulatory interactions among genes in specific species. Such a network is extremely useful in discovering gene function and in elucidating the molecular basis of genetic disease; genes involved in the same molecular pathways are likely to be close in the recovered interactome.

Here we analyzed a collection of 20255 HG-U133A microarrays coming from 703 experiments measuring 22283 human transcripts. As well as, a collection of 8895 Mouse403_2 microarrays coming from 613 experiments measuring 23000 mouse transcripts. Transcripts are the nodes of the species-specific interactome; a connection between two nodes is weighted according to their statistical dependence computed across all the expression values. The higher the value the stronger the connection is.

In order to analyze this massive dataset, we implemented a normalization step via the application of a novel discretization procedure that yields a single dataset containing comparable expression values. Then, for each pair of transcripts (over 200 million pairs), we assigned a score that weights their statistical dependence by computing their Mutual Information (MI). Since MI needs to be computed for about 200 million pairs of transcripts on a massive dataset, using a standard algorithm, would take months of computational time. Therefore we designed, implemented and tested an efficient parallel version of an algorithm that normalizes data and computes MI, thus dramatically reducing the computational time to about 8 and 4 hours to recover the Human and the Mouse Interactome respectively.

The human interactome has been validated by comparing it against a published protein-protein interaction network including 51486 interactions among 6404 genes. By sorting the interactions of the inferred network according to their weights, we obtain a ROC curve with a maximum PPV of 91% and 1% sensitivity (Sens); A random prediction would have yielded a PPV of only 0.0025%. To validate the mouse interactome, we chose a tissue specific transcription factor (p63) responsible for the keratinocytes self-renewal, involeved in human malformation syndromes, for which a list of 785 target genes have been recently validated. The interactome PPV peaks at 20% with a Sens=3%.

We show that massive expression datasets, including a variety of tissues and cell types, if properly analysed are able to shed light on gene regulation, even when tissue-specific. The interactomes we generated are a valuable resource to help in understanding gene function and can be accessed via a web interface from http://dibernardo.tigem.it.

microRNA transcript and promoter features emerge from polymerase II chromatin immunoprecipitation data

David L. Corcoran[1], Kusum V. Pandit[1], Sabah Kadri[2], Ben Gordon[3], Veronica Hinman[2], Arindam Bhattacharjee[3], Naftali Kaminski[4], Panayiotis V. Benos[5]

[1]Department of Human Genetics, University of Pittsburgh, PA, USA; [2]Department of Biological Sciences, Carnegie Mellon University, PA, USA; [3] Genomics, Agilent Technologies, Inc., Santa Clara, CA, USA; [4] Allergy and Critical Care Medicine, University of Pittsburgh, PA, USA; [5] Department of Computational Biology, University of Pittsburgh, PA, USA.

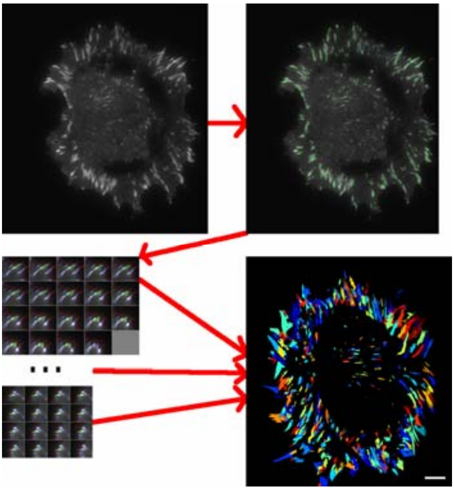
MicroRNAs (miRNAs) are short, non-coding RNA regulators of protein coding genes. They bind to their target mRNAs and degrade them or inhibit protein translation. miRNAs play an important role in diverge biological processes and they constitute important markers for cancer and other diseases. A number of algorithms have been developed for predicting miRNA genes and their targets, but surprisingly little is known about their transcriptional regulation. miRNA genes located in introns of protein coding genes (intronic miRNAs) are assumed to be co-transcribed with their host genes, while intergenic miRNAs are believed to be transcribed by their own RNA polymerase II promoters. However, the exact length of the miRNA transcripts and their promoter organization is currently unknown. In an attempt to understand better the transcription of miRNA genes, we performed RNA polymerase II (Pol II) chromatin immunoprecipitation (ChIP)-chip using a custom designed location array surrounding regions of known human miRNA genes. To identify the exact position of the transcription start sites (TSSs) of the miRNA genes and to compare the promoter features of protein coding and miRNA genes we developed a new modeling tool called Core Promoter Prediction Program (CPPP). CPPP uses the Support Vector Machine (SVM) framework to identify the TSS in the miRNA surrounding region identified by the ChIP-chip probes. Various promoter features (*n*-mers, PSSMs), backgrounds and kernel functions were compared on known promoters from protein coding genes and the best performing model was used for the miRNA promoter analysis. The results indicate that miRNA genes can be transcribed from promoters that are located several kilobases away. Analysis of the miRNA promoter features also showed that the promoters of intergenic miRNA genes share the same features as those of protein coding genes. Finally, we found evidence that about 25% of the intronic miRNAs may be transcribed from their own promoter.

Focal Adhesion Dynamics Analysis Through Quantitative Image Processing

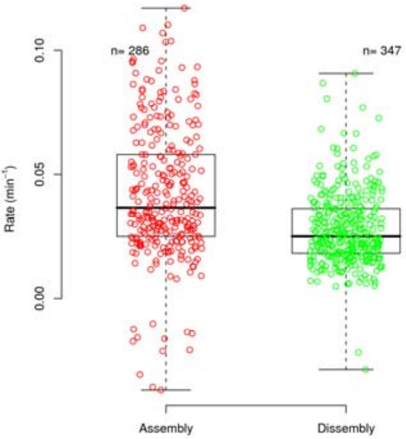
Matthew E. Berginski[1], Eric Vitriol[2], Klaus Hahn[2], Shawn M. Gomez[1]

University of North Carolina at Chapel Hill, [1]Department of Biomedical Engineering, [2]Department of Pharmacology

Focal adhesions (FAs) are highly dynamic protein complexes responsible for mediating the cell's interaction with the outside environment; acting as a point of integration for both mechanical and chemical signaling. The biochemical components of FAs have been well characterized and cataloged, but comprehensive, automated methods to analyze the dynamic properties of these protein complexes have not been developed. To better understand the regulation and spatio-temporal dynamics of FAs, we have developed a set of computational tools to extract relevant information from time-lapse TIRF microscopy images. We applied this set of tools to monitoring paxillin, a well-studied protein that is recruited early in FA assembly and remains associated until disassembly of the FA complex. We then monitored GFP-paxillin localization within 3T3 fibroblasts at 1-minute time intervals.



The computational tools implement a workflow of automated image and data analysis (Top Figure, adhesions outlined in color). This software identifies each individual FA, tracks their movement through time and collects a set of properties concerning the location, shape, size and intensity of each FA. All detectable adhesions are included in this analysis, so a thorough picture of global adhesion behavior is captured. Many properties of FA have been extracted using this system and the analysis of FA assembly and disassembly rates has shown that the system is capable of collecting data beyond the typical range found using manual analysis methods (Bottom Figure). Such quantitative measurements of FA dynamics will be crucial for studying the results of perturbing the system of signaling networks that affect FA development.



Global identification and characterization of Hox response enhancers in Drosophila melanogaster

Daniela Bezdan[1], Martin H. Schaefer[1], Andreas Hunczek[1], Stephan Ossowski[1], Ingrid Lohmann[1,2]

[1] Max-Planck-Institute for Developmental Biology, AG I. Lohmann, Spemannstrasse 37-39, D-72076 Tuebingen, Germany, [2]BIOQUANT Center, NWG Developmental Biology / AG Lohmann, 69120 Heidelberg, Germany

Hox genes control morphogenesis along the anterior-posterior (A/P) body axis of animals. As transcription factors, they regulate a battery of downstream genes by direct binding to Hox response enhancers (HREs). Despite their specific effects *in vivo*, Hox proteins bind indiscriminately to only one distinct core sequences *in vitro*, a main limitation for identification of Hox target genes and associated HREs. Therefore it is necessary to globally identify and characterize HREs in the genome to understand the regulatory code and environmental conditions by which Hox proteins achieve specificity.

To identify unknown HREs we established a pipeline that identifies accumulations of Hox consensus binding sites and collaborating transcription factors (cTFs). These binding site clusters are further tested for conservation across Drosophila species considering evolution branch factor and core conservation.

Features like alignment-free similarity measures, motif search tools, motif distance preferences, gene function and gene pathway information and expression data from two microarray approaches were applied to refine the prediction of functional HRE clusters.

To validate the most promising HREs, candidate motifs were tested by high throughput EMSA, cell culture assays and reporter constructs. In addition we preformed microarray based expression analysis for six Hox genes and detection of direct target HREs/cTFs using CHIP/^{2}. To improve resolution of binding site detection we designed a ChIP-Seq experiment using Illumina sequencing technology.

In summary, we identified and validated a considerable number of unknown HREs *in vitro* and *in vivo*. We could show that a data mining approach using multiple strategies is essential to identify HREs. The similar arrangements of motifs correspond to similar expression patterns and functional annotation (e.g. GO terms). Interestingly the simple presence of a consensus-binding site is not sufficient for Deformed/Ultrabithorax binding even *in vitro*. We conclude that Hox gene regulation depends on a combination, arrangement and distances of collaborators and binding sites in Hox Response Enhancers.

Simultaneous learning of the topology and dynamics of global regulatory networks using the Inferelator

Richard Bonneau'[1,2,3], Aviv Madar[1-3]

[1]NYU Center for Genomics and Systems Biolgy; [2]Courant Institute for Mathematical Sciences; [3]New York UniversBut it ity, Dept. of Biology.

Our system for network inference and modeling consists of two major components: cMonkey (a method for learning co-regulated biclusters and pathways), and the Inferelator (regulatory network inference, regulatory networks modeled as large systems of ODEs). The Inferelator infers regulatory influences for genes and/or gene clusters from mRNA and/or protein expression levels. The procedure models transcriptional networks as large sets of coupled ODEs and can simultaneously model equilibrium and time-course expression levels, such that both kinetic and equilibrium expression levels may be predicted by the resulting models (or used to learn/parameterize regulatory network models). Through the explicit inclusion of time, and gene-knockout information, the method is capable of learning causal relationships. It also includes a novel solution to the problem of encoding interactions between predictors. We discuss the results from application of this method to organisms spanning the tree of life including prokaryotic functional genomics, mouse T-cell differentiation, etc. We focus on an initial application of this method to the halophilic archaeon, Halobacterium NRC-1. We have found the network to be predictive of 150 newly collected microarray datasets and have also validated parts of the network using ChIP-chip. This network offers a means of deciphering how this intensely tough organism maintains homeostasis and responds to wide varieties of metabolic, genetic and environmental states. We will also discuss recent improvements including improvements to the Inferelator method that remove several of the assumptions inherent in our original method and methods for learning networks from multiple-species data compedium.

Gene-expression Interpretation Tool for toxicogenomics

André Boorsma[1,2] and Rob Stierum[1]

[1]TNO quality of Life P.O. box 360,3700 AJ Zeist, The Netherlands. [2]Department of Risk and Health Analysis and Toxicology Maastricht University, Maastricht, The Netherlands.

One of the aims of toxicogenomics is to classify compounds based on their toxicity. Another important issue is the dose-response assessment of the toxicity of compounds using gene expression profiles. In order to compare the toxic response of compounds in different species, tissues and in vivo and in vitro, cross comparison of gene expression profiles is of great importance. The majority of the microarray analysis methods used within toxicogenomics are based on the single gene level. Using gene groups, can improve sensitivity of analysis and has the advantage that interpretation of the data is more straightforward.

We addressed aforementioned issues by creating a web-application based on T-profiler. T-profiler is a method that quantifies the expression of a gene group within a gene expression profile. This has the benefit of a better biological interpretation and allows for cross comparison of gene expression profiles. Our application is used for the analysis of gene expression profiles and the comparison of gene groups of interest to those derived from public toxicogenomics-related gene expression data. A possibility to upload custom gene sets is included together with a database of public available toxicogenomics experiments that can be used as reference material.

We illustrate the utility of our application by showing gene specific response for compound classes and dose dependent responses for the HMG-CoA reductase inhibitor Atorvastatin. We also performed correlation analysis between t-values of gene groups and physiological parameters; this revealed mostly significant results for immunology related gene groups. Finally, we used cluster analysis to visualize the relation between the gene groups over all experiments. Altogether, we think that our web-application is a valuable contribution to the field of toxicogenomics.

Evolution of the mammalian transcription factor binding repertoire via transposable elements

Guillaume Bourque[1*], Bernard Leong[1], Vinsensius B. Vega[1], Xi Chen[2], Yen Ling Lee[3], Kandhadayar G. Srinivasan[3], Joon-Lin Chew[2], Yijun Ruan[3], Chia-Lin Wei[3], Huck Hui Ng[2] and Edison T. Liu[4]

[1]Computational and Mathematical Biology, [2]Stem Cell and Developmental Biology, [3]Genome Technology and Biology, [4]Cancer Biology and Pharmacology, Genome Institute of Singapore, 138672, Singapore

Identification of lineage-specific innovations in genomic control elements is critical for understanding transcriptional regulatory networks and phenotypic heterogeneity. We analyzed, from an evolutionary perspective, the binding regions of seven mammalian transcription factors (ESR1, TP53, MYC, RELA, POU5F1-SOX2 and CTCF) identified on a genome-wide scale by different chromatin immunoprecipitation approaches and found that only a minority of sites appear to be conserved at the sequence level. Instead, we uncovered a pervasive association with genomic repeats by showing that a large fraction of the bona fide binding sites for five of the seven transcription factors (ESR1, TP53, POU5F1-SOX2 and CTCF) are embedded in distinctive families of transposable elements. Using the age of the repeats, we established that these Repeat-Associated Binding Sites (RABS) have been associated with significant regulatory expansions throughout the mammalian phylogeny. We validated the functional significance of these RABS by showing that they are over-represented in proximity of regulated genes and that the binding motifs within these repeats have undergone evolutionary selection. Our results demonstrate that transcriptional regulatory networks are highly dynamic in eukaryotic genomes and that transposable elements play an important role in expanding the repertoire of binding sites.

Fast Statistical Alignment of genes and genomes

Robert K. Bradley[1,2], Ariel Schwartz[3], Nicolas Bray[1], Colin N. Dewey[4], Lior Pachter[1,2]

[1]Department of Mathematics, University of California, Berkeley, CA 94720, USA; [2]Department of Molecular & Cellular Biology, University of California, Berkeley, CA 94720, USA; [3]Department of Bioengineering, University of California, San Diego, CA 92093, USA; [4]Department of Biostatistics & Medical Informatics, and the Department of Computer Sciences, University of Wisconsin, Madison, WI 53706, USA.

Accurate alignment is essential for both functional and evolutionary genomics studies, ranging from assessing conservation of transcription factor binding sites to interpreting data from comparative ChIP-Chip or ChIP-Seq experiments. Focusing on the tractability of the alignment problem, we argue that many recent negative results emphasizing uncertainty in alignment are misleading in that they confound uncertainty in the choice of model, uncertainty in alignment given a model and error due to heuristics used for inference. We explain how hidden Markov models for pairwise alignment can be extended to provide effective models for multiple alignment, and show that these models indicate little uncertainty in alignment of both unrelated sequences and of orthologous sequences from related species.

Guided by our analyses of the sources of uncertainty in alignment, we present an algorithm which provides an efficient approach to finding the alignments with highest expected accuracy. We show that this algorithm, implemented in the program FSA (Fast Statistical Alignment), accurately aligns proteins, RNAs and orthologous genomic regions. We perform large-scale tests to demonstrate that most current programs give near-complete alignments of random sequences, implying that presence of alignment does not necessarily imply homology. In contrast, our method leaves random sequence largely unaligned. FSA's statistical model can automatically learn different parameterizations for each alignment problem, allowing it to accurately align sequences under very different evolutionary constraints. Furthermore, FSA estimates the accuracy with which each character is aligned, allowing biologists to incorporate alignment reliability measures into downstream analyses. Crucially for large-scale analyses, FSA is fast; we used it to align 1,502 genes in *Saccharomyces cerevisiae* with their orthologs in six other species of yeast in 10 minutes on a cluster with 36 nodes.

While there are many sources of uncertainty in alignment, they can be efficiently controlled and quantified with an appropriate statistical model. Taken together, these results provide a path to the removal of lingering doubts about the accuracy of multiple alignments.

Systems-level analysis of transcription factor spatial patterns and targets in drosophila embryogenesis

Bristow, C.A. [1,2], Kheradpour, P.[1], Frise, E.[3], Celniker, S.[3]. Kellis, M.[1,2]

[1]CSAIL, MIT, Cambridge, MA; [2]Broad Institute of MIT and Harvard, Cambridge, MA; [3]Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, CA.

Drosophila embryogenesis is well established experimental system for the elucidation and analysis of gene regulation. The implementation of high-throughput *in situ* hybridization protocols has rapidly increased the amount of spatial and temporal gene expression data. Over 4000 genes have detectible embryonic expression patterns and are annotated using a controlled vocabulary of anatomical terms. Spatial expression patterns for ~50% of all transcription factors are available, and the remaining factors are currently being imaged and annotated. The combination of a large set of patterns for both factors and potential targets allows for the systematic assembly and analysis of gene regulation across a range of developmental systems. We are developing a series of approaches to associate transcription factors with their respective targets.

The first approach involves tracking the correlations between spatial patterns of transcription factors and potential targets using anatomical terms. For ~60 of the 300 factors with known spatial profiles, binding site data is available and is incorporated into the analysis. In parallel, we are developing methods to extract correlations between factors and targets directly from images. The images of embryos are segmented and converted into a geometric mesh. For the representation of the spatial pattern, we are developing geometric models that assemble complex patterns from simple building blocks using Boolean operations.

A final aspect of this project is the development of an annotation pipeline to assign controlled vocabulary terms directly to specific anatomical features and expression patterns. At present, controlled vocabulary terms are provided as captions describing spatial expression patterns. The assignment of labels to sets of pixels will greatly increase the utility of this dataset for automatic identification and classification of expression patterns.

Temporal Dynamics of the Drosophila Developmental Network

Rogerio Candeias[1,2,3], Manolis Kellis[2,3]

[1]PhD Program in Computacional Biology of Instituto Gulbenkian de Ciência, sponsored by Fundação Calouste Gulbenkian, Siemens SA, and Fundação para a Ciência e a Tecnologia, Portugal; [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA

Gene expression induction by a transcription factor is a complex stochastic interplay of a number of mechanisms. A delay exists between activation and binding of a transcription factor and the full change in the expression of the target gene. Extensive dynamic models of regulation exist for small network motifs assuming constant or null delay. These models will fail to predict the expression of highly temporal dependent processes as in the case of embryonic development.

Meanwhile, recent developments in comparative genomics have led to the reliable prediction of regulatory networks from sequence, producing the most complete and accurate regulatory network of Drosophila melanogaster.

Currently we are developing methods to compute the overall delay of the expression of two genes from time-course microarray data. As a preliminary method, we used the cross-correlation coefficient. This function outputs a measurement of the similarity between the two signals and the delay that maximizes the correlation. We are also using fitted multi-step functions in order to ascertain state changes in the expression of a given gene, effectively determining the transition time for each state.

We have searched for enriched network motifs and found that each edge on a particular motif has a unique distribution of delays associated with it. Moreover, for a given motif, repressing edges have different dynamics then activating ones. Furthermore, we discovered that transcription factors act has strong activators of expression with short delays in early embryonic development and that in later stages they will mildly repress gene expression with increasing delay.

Applying this novel approach to the fly regulatory network has led to a promising framework for the study of temporal developmental patterns and temporal information flow across the network, fundamentally changing the way we think about networks.

Genome-wide integration of transcriptional and post-transcriptional regulatory networks: looking at mixed TF/microRNA feed-forward circuits in the human genome.

Angela Re[1], *, Davide Corà[2], *, Daniela Taverna[3] and Michele Caselle[2]

* Equal contribution

[1] *CIBIO - Centre for Integrative Biology, University of Trento, Italy;*

[2] *Department of Theoretical Physics, University of Torino and INFN, Italy*

[3] *Department of Oncological Sciences, University of Torino, Italy*

Promoters and 3'-UTRs are thought to control the expression of coding genes mainly in response to transcription factors (TF) and microRNAs. Several methods exist to elucidate TF-related and microRNA-related regulatory networks, but comparable information is lacking to explicitly connect them. Moreover, a basic notion of modern system biology is that biological functions are usually performed by groups of genes which act in an interdependent and synergic way. A prominent role in this approach is played by the notion of ``network motif": a complex network can be divided into simpler, distinct, regulatory patterns, acting as the smallest functional modules of the whole ensemble of interactions [1,2].

Stemming from these considerations, we report here a systematic integration of a genome-wide transcriptional and post-transcriptional regulatory network, in the human genome, based on a bioinformatic sequence-analysis work. In particular, we focused in the study of functional properties of a special class of network motifs, the mixed TF/microRNA feed-forward circuits, recently proved to be of significant importance [1].

A human genome-wide catalogue of this type of circuits was assembled with a two step procedure. We first constructed a transcriptional regulatory network and, separately, a list of post-transcriptionally regulated genes for human by looking for conserved overrepresented motifs in human and mouse promoters and 3'-UTRs [3]. Second we combined the two datasets looking for mixed feed-forward regulatory loops, i.e. all the possible instances in which a master transcription factor regulates a microRNA and together with it a set of joint target coding genes. We obtained a catalogue composed by a total of a few hundreds of such regulatory circuits in human. To investigate the biological relevance of these interactions, the proposed circuits were prioritized based on the existence of enriched functional (Gene Ontology) annotations and comparing our results with external different and independent evidences, both of experimental and computational type. Several biologically relevant circuits are discussed, describing in particular aspects of organism development and differentiation. Among the various loops that we found we finally describe with particular attention circuits with potential involvement in cancer.

Our results indicate that mixed feed-forward regulatory circuits can play an important role in the global regulatory network of a complex organism.

[1] Hornstein E and Shomron N; Nat Genet 38 Suppl:S20–4 (2006).

[2] Shalgi R, Lieber D, Oren M, Pilpel Y; PLoS Comput Biology 3:e131 (2007).

[3] Cora D, Di Cunto F, Caselle M, Provero P; BMC Bioinformatics 8:174 (2007).

Towards a Quantitative Representation of the Cell Signaling Mechanisms of Hallucinogens: Measurement and Mathematical Modeling of 5-HT1A and 5-HT2A receptor-mediated ERK1/2 Activation

Chiung-wen Chang[1], Ethan Poteet [3], John A. Schetz[3], Zeynep H. Gümüş[1-2], Harel Weinstein[1-2]

[1]Department of Physiology and Biophysics; [2]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Medical College of Cornell University, 1300 York Ave, New York, NY 10021 USA; [3]Department of Pharmacology and Neuroscience, University of North Texas Health Science Center, 3500 Camp Bowie Blvd. Fort Worth, TX 76107.

Through a multidisciplinary approach involving experimental and computational studies, we address quantitative aspects of signaling mechanisms triggered in the cell by the receptor targets of hallucinogenic drugs, the serotonin 5-HT2A receptors. To reveal the properties of the signaling pathways, and the way in which responses elicited through these receptors alone and in combination with other serotonin receptors subtypes (the 5-HT1AR), we developed a detailed mathematical model of receptor-mediated ERK1/2 activation in cells expressing the 5-HT1A and 5-HT2A subt pes individually, and together. In parallel, we measured experimentally the activation of ERK1/2 by the action of selective agonists on these receptors expressed in HEK293 cells. We show here that the 5-HT1AR agonist Xaliproden HCl elicited transient activation of ERK1/2 by phosphorylation, whereas 5-HT2AR activation by TCB-2 led to higher, and more sustained responses. The 5-HT2AR response dominated the MAPK signaling pathway when co-expressed with 5-HT1AR, and diminution of the response by the 5-HT2AR antagonist Ketanserin could not be rescued by the 5-HT1AR agonist. Computational simulations produced qualitative results in good agreement with these experimental data, and parameter optimization made this agreement quantitative. *In silico* simulation experiments suggest that the deletion of the positive regulators PKC in the 5-HT2AR pathway, or PLA2 in the combined 5-HT1A/2AR model greatly decreased the basal level of active ERK1/2. Deletion of negative regulators of MKP and PP2A in 5-HT1AR and 5-HT2AR models was found to have even stronger effects. Under various parameter sets, simulation results implied that the extent of constitutive activity in a particular tissue and the specific drug efficacy properties may determine the distinct dynamics of the 5-HT receptor-mediated ERK1/2 activation pathways. Thus, the mathematical models are useful exploratory tools in the ongoing efforts to establish a mechanistic understanding and an experimentally testable representation of hallucinogen-specific signaling in the cellular machinery, and can be refined with quantitative, function-related information.

An Alignment Independent Algorithm for de novo Motif Discovery using Phylogenetic Conservation

Donavan T.S. Cheng[1], Michael A. Beer[1]

[1]*Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore MD.*

We present a novel Gibbs sampling based algorithm, Flipper, for transcription factor (TF) binding site discovery, which uses sequence conservation to inform its search for over-represented sequence motifs. Because TF binding sites can frequently be short and of varied composition, our method does not rely on an alignment pre-preprocessing step, unlike previous approaches [1,2,3] which may be overly restrictive for the detection of such sites. Flipper joins similar methods in advocating an alignment free approach for assessing conservation [6], with a difference that Flipper is capable of *de novo* motif discovery without prior knowledge of TF specificities.

We compared the performance of our algorithm against previous alignment based approaches [1,2,3] using synthetic sequence datasets. Our analysis shows that Flipper performs up to 25% better against competing methods, especially in simulations where the compared species are sufficiently diverged to the point that alignment quality is adversely affected.

In validation studies, we tested the ability of Flipper to recover known sequence motifs from *in vivo* binding data of yeast, worm and fly TFs [4,5]. For yeast, Flipper succeeded in recovering the motif for 41 out of 44 TFs where previous methods were also successful, but more significantly, Flipper recovered the known motif for another 12 of 37 TFs where previous methods had failed. For worm and fly, Flipper was able to recover binding motifs for worm TFs, ELT-2, NFI-1 and HSF-1 and fly developmental TFs involved in patterning of the AP axis.

We present motif predictions generated from sets of coexpressed genes from *C.elegans* and the related sequenced nematode species *C.briggsae, C.remanei* and *C.brenneri*. In particular, we have validated three novel functional motifs in promoters of ribosomal protein genes, by generating transgenic *C.elegans* carrying GFP reporters fused to *C.elegans* and *C.briggsae* promoters, either containing an intact or mutated copy of the motif.

In summary, we have developed a general algorithm that is not limited by constraints on sequence alignability and can be extended to motif discovery in larger genomes.

References:

[1] Siddharthan R, Siggia ED, van Nimwegen E, Plos Comp Biol, 1, e67 (2005)
[2] Sinha S, Blanchette M, Tompa M. BMC Bioinformatics. 5, 170 (2004)
[3] Wang T, Stormo GD. Bioinformatics (2003)
[4] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ et al. Nature 431, 99-104 (2004)
[5] Li XY et al, Plos Biol, 2008 Feb;6(2):e27.
[6] Ward D, Bussemaker HJ. Bioinformatics, 2008

Pathgen: Gene Regulatory Network Tool

Kendell Clement[1], Nathaniel Gustafson[1], Chris Merris[1], Steven Flygare[1],Kenneth Sundberg[1], Mark Clement[1], Quinn Snell[1], Jared Allen[2], Randall J. Roper[2]

[1]Brigham Young University, [2]Indiana University-Purdue University Indianapolis

As sequencing projects and microarray experiments continue to produce large amounts of genetic information, the task of understanding these sequences and the proteins they encode becomes increasingly important. Public access of protein interaction databases has allowed researchers to relate genetic sequences to function. Several protein interaction databases and analysis programs have been created to address the demand for interaction data, but several problems still exist.

First, accessibility of information has been limited because protein interactions databases are not connected. In order to find interactions for a specific protein, several databases with different input and output formats must be queried, and then the results are analyzed. This is often tedious and inefficient.

Another problem is that many pathway investigation tools allow researchers to see interactions within a single pathway, but it is impossible to search for interactions between genes across multiple pathways. Researchers may have a set of suspected source genes, and a set of target genes that are associated with a phenotype, and they want to find connections between the sets of genes. Many programs allow users to specify a set of search genes, but they don’t allow the searching of interactions from source to target genes.

The biggest problem with current pathway investigation tools is that biological researchers may be overwhelmed when they look for candidate genes in a whole-genome interaction display. Researchers are looking for novel connections between candidate genes and target genes known to be associated with a phenotype, and want to investigate genes that lie in the possible interaction pathway between the two genes without being overwhelmed by unrelated genes.

The PathGen database contains regulatory information from several existing interaction databases, interactions inferred from PubMed abstracts, as well as interactions provided by the user. Users input start and target genes, and interactions between these two genes are presented in a way that is understandable and useful. The only intermediate genes that are displayed are those connecting the start and target genes, so the user is not overwhelmed by the output.

We used PathGen to connect genes associated with Down syndrome and genes linked with developmental phenotypes disrupted by trisomy 21. PathGen produced a large network showing interactions between the genes of interest that were not previously identified in existing protein interaction databases. The genes connected *in silico* by PathGen networks can be studied *in vitro* and *in vivo* for possible therapeutic intervention to correct Down syndrome phenotypes.

In addressing these problems, we have created a useful pathway analysis tool that is accessible, flexible, and useful. These features make PathGen a valuable tool for *in silico* discovery of novel gene interaction pathways, which can be experimentally tested and verified.

This resource can be accessed online at *http://dna.cs.byu.edu/pathgen/*.

A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence

Hilary A. Coller[1], Joshua J. Forman[1], Aster Legesse-Miller[1]

[1]*Department of Molecular Biology, Princeton University, Princeton, NJ 08544*

Recognition sites for microRNAs (miRNAs) have been reported to be located in the 3' untranslated regions of transcripts. In a computational screen for highly-conserved motifs within coding regions, we found an excess of sequences conserved at the nucleotide level within coding regions in the human genome, the highest scoring of which are enriched for miRNA target sequences. To validate our results, we experimentally demonstrated that the *let-7* microRNA directly targets the miRNA-processing enzyme Dicer within its coding sequence, thus establishing a mechanism for a miRNA/Dicer autoregulatory negative feedback loop. We also found computational evidence to suggest that miRNA target sites in coding regions and 3'UTRs may differ in mechanism. This work demonstrates that miRNAs can directly target transcripts within their coding region in animals, and suggests that a complete search for the regulatory targets of miRNAs should be expanded to include genes with recognition sites within their coding regions. As more genomes are sequenced, the methodological approach that we employed of identifying motifs with high sequence conservation will be increasingly valuable for detecting functional sequence motifs within coding regions.

A web interface to extract protein motifs by constrained co-clustering

Francesca Cordero [1, 2], Alessia Visconti [2], Marco Botta[2]

[1]Molecular Biotechnology Center, Department of Clinical and Biological Science, University of Turin, Via Nizza 52, 10126 Torino, Italy and [2]Department Computer Science, University of Turin, Corso Svizzera 185, 10149 Torino, Italy

Pattern discovery in biological sequences is a fundamental problem in both computer science and molecular biology. We propose a framework based on a constrained co-clustering technique that simultaneously groups protein sequences and their associated patterns. The main goal of our approach is to split a set of (possibly unknown function) protein sequences in groups characterized by some common motifs. The novelty of our approach is the protein motif discovery methodology, which relies on the following two main ideas: exhaustive search, based on prefix tree, and automatic association between motifs and protein sub-families using a constrained co-clustering algorithm.

Motif discovery is performed on a complete *ab-initio* technique where any biological knowledge is considered *a-priori*. We use a prefix tree as data structure, since it allows to store all possible motifs of a given length extracted from the protein sequences. In this way, we maintain *all* those patterns present there along with both the protein name and their frequencies.

A constrained co-clustering technique finds both protein motif classes and protein groups: in such a way, it is possible to correlate every protein group with one or more motif classes. This technique is applied to the frequency matrix built on those values extracted from prefix tree. Therefore, the co-clustering association is given by a statistical measure based on cluster cardinality. In test on experimentally determined protein datasets, the presented framework is able to identify the correct pairs of protein family and motifs that are very similar to PROSITE's patterns.

We have built a web interface in which the best identified motifs are displayed in association with a list of related protein names. In particular, the patterns in each motif cluster have been multiple aligned with ClustaW. In such way the information content is graphically pictured with sequence logos. Our approach is able to group together protein sequences belonging to the same families and, at the same time, provides a set of characterizing motifs.

Plasticity of the Yeast Transcription Network After the Whole-genome Duplication.

Luigi Grassi[1,*], Diana Fusco[2,3,*],Davide Cora'[1,5], Bruno Bassetti[2,4], Michele Caselle[1,5], Marco Cosentino Lagomarsino[2,4]

[1]Dip. Fisica Teorica, Univ. di Torino, Italy; [2]Gruppo Teorico, Dip. Fisica, Univ. di Milano, Italy; [3]Duke University Durham NC, USA; [4]I.N.F.N. Milano, Italy; [4]I.N.F.N. Torino, Italy; [*]Equal Contribution

Gene duplication, a major evolutionary drive for functional innovation, can occur at the scale of an entire genome. One of such "whole-genome duplication" (WGD) events has been proven among the Ascomycota fungi, in the branch of *S.cerevisiae*. The genes duplicated during this event are shown to have different functional and interaction properties than conventional (local) duplicates, indicating that the WGD is free of some of the constraints affecting local duplications. We studied the evolution of transcriptional interactions triggered by the yeast WGD with a combined methodology including (i) transcription network data analysis and graph-growth modeling, and (ii) cis-regulatory sequence analysis comparing pre- and post- WGD yeasts.

Our results uncover the WGD as a major source for the evolution of complex pathways in the transcriptional regulation of yeast. The inheritance of interactions among WGD duplicates does not occur independently, but follows elementary "duplication subgraphs", that relate ancestral interactions with newly formed ones. Moreover, duplication subgraphs are strongly dependent on their network neighborhood and give rise to higher order circuits with two elementary properties: newly formed transcriptional pathways are not broken and preferentially cross-connected with older ones. In contrast, local duplications involve mainly target genes (and not transcription factors). Growth by local duplication has created a few cross-connected interaction subgraphs by successive duplications, but only over hundred million years. Based on these observations, we develop a quantitative "one-shot" evolutionary graph duplication model for the WGD with predictive power, and use it on the data to estimate quantities that are not observable directly, such as the probability of loosing transcriptional interactions. At the cis-regulatory sequence level, we observe large plasticity in the promoters of WGD duplicates. A comparative analysis of regulatory motifs shows a marked tendency for regulatory sub-functionalization, i.e. motifs that were present beforethe WGD are divided between the promoters of WGD duplicates, and tend not to be present on both. Finally, the evolutionary rate of the promoters of duplicates decreases smoothly with duplication date and does not show a distinct trend for WGD paralogs.

Conservation of two SVD modes of gene expression in the rat brain following stroke and seizures permits splitting analysis of gene co-regulation into tractable sub-problems

Michal Dabrowski[1], Norbert Dojer[2], Malgorzata Zawadzka[1], Bozena Kaminska[1]

[1]Nencki Institute, Laboratory of Transcription Regulation, Warsaw; [2]Institute of Informatics, University of Warsaw, Poland

We previously reported identification of cis-regulatory features associated with a simple pattern of expression (down-regulation) in a subspace of SVD mode of gene expression conserved between in vitro and in vivo neural development. In the present work, we performed SVD of data from global gene profiling in the brain following a transient cerebral ischemia – an animal model of stroke (M); and of the published data from the rat hippocampus following kainate-induced seizures (A). The two datasets, obtained on different microarray platforms, comprised of profiles for 2786 and 2392 genes, with 737 common genes. We report that the loadings of two corresponding SVD modes (M2 – A2, M3 – A3) across the common 737 genes are highly conserved between stroke and seizures. Functional annotation of the genes ranked separately on loadings of each of the conserved modes (M2, A2, M3, A3) revealed that in both systems the second modes (M2, A2) reflect apoptosis and/or inflammatory response, while the third modes (M3, A3) reflect increased neuronal activity. Separately for each mode (M2, A2, M3, A3), we looked for cis-regulatory features associated with either sign of its loading (up- or down-regulation in the subspace of this mode). Applying Bayesian networks learning algorithm, with the BDE scoring function, to a database of putative cis-regulatory regions and known transcription factor binding motifs, we looked for cis-regulatory features (motifs, modules, and sets of modules) associated with the sign of the loading of a particular mode. The significance (p-value) of the BDE scores was obtained from analysis of 1000 datasets with randomly permuted gene labels. We identified AP1F motif – binding the transcription factor AP1 as the highest-ranking feature, significantly associated with the sign of the second modes (A2, M2) in both systems. Several motifs, among the highest-ranking features in one or the other system, were identified as highly significantly associated with the sign of the third modes (M3, A3) in both systems. The motifs associated with the third modes bind transcription factors involved in neuronal plasticity (Egr1-3, Creb, Zfp161), differentiation (Lhx1-3), oxidative stress/response to hypoxia (Nfe2l1-2, Arnt). The results obtained on one dataset were confirmed on the other dataset, which validated the identified features and, at the same time, revealed mechanisms of gene co-regulation conserved between stroke and seizures.

Supported by grant PBZ-MNiI-2/1/2005.

Models from patients: reverse-engineering glioblastoma

Teresia Dahl[,1], Weiqing Wang[2], Nikolaus Schultz[2], Björn Nilsson[3], Peter Gennemark[4], Chris Sander[2], Sven Nelander[1,4]

[1]University of Gothenburg, Wallenberg laboratory; [2]Memorial Sloan Kettering Cancer Center, Computational Biology Center; [3]Brigham and Women's hospital, [4]Gothenburg Mathematical Modeling Center

A key aim of cancer systems biology is to relate targeted therapy to phenotypic consequences, To achieve this aim, computational models of cellular processes are extremely useful, if not essential.

We have developed a particular approach to constructing, optimizing and applying computational models of cellular processes, which we call Combinatorial Perturbation-based Interaction Analysis (CoPIA) [1]. The key purpose of CoPIA is to predict and explain the action of drug combinations on cancer cells.

We now aim to use an extension of CoPIA to derive predictive models for transcriptional regulation in glioblastoma. The key idea is to view mutational profiles as combinatorial perturbations, and the corresponding transcriptional profiles as the phenotype response. Model construction is based on detailed genotypic, epigenetic and transcriptional profiles across 200 primary glioblastoma tumors, generated by the Cancer Genome Atlas consortium.

As far as we know, this is the first attempt to reverse engineer oncogenic pathways based on paired genotype/phenotype measurements across a set of patients.

[1] Nelander S, Wang W, Nilsson B, She QB, Pratilas C, Rosen N, Gennemark P, Sander C. Models from experiments: combinatorial drug perturbations of cancer cells. *Mol Syst Biol.* 2008;4:216.

A webserver for the prediction of kinase-specific phosphorylation sites using Conditional Random Fields

Thanh Hai Dang[1], Alain Verschoren[1], Kris Laukens[1]

[1]*Intelligent Systems Laboratory, University of Antwerp, Belgium*

Phosphorylation is an important regulatory post-translational protein modification, catalyzed by kinases, and consisting of the addition of a phosphate group to S, T or Y residues of subtrate proteins. Experimentally verified phosphorylation sites have been compiled into dedicated databases, such as “Phospho.ELM” [1]. To complement the not trivial experimental phospho-site identification, computational phosphorylation site prediction models have been developed using a range of machine learning techniques. A number of these models are implemented in public web servers.

We developed a new model for kinase-specific phosphorylation site prediction based on Conditional Random Fields (CRFs) [2]. A CRF is a discriminative probabilistic model often used for labeling sequential data. It has several advantages compared to Hidden Markov Models and Maximum Entropy Markov Models. Given that X and Y are a set of data observations and label sequences, respectively, HMMs require enumerating all possible data observation sequences to maximize the probability P(X,Y). The CRF model does not have to model the observation sequences explicitly. It, maximizes the conditional probability P(Y|X) from the training datasets. Furthermore, it is still valid if dependencies between arbitrary features exist in observation sequences. Moreover, the CRF model overcomes a weakness called the label bias problem of which HMMs and MEMMs fall victim.

In contrast to existing models, the proposed CRF model is trained only from the positive (“golden”) dataset, whereas the negative data are used to specify a decision treshold by using Chebychev’s statistical rule of inequality. The flanking sequence surrounding phosphorylated and non-phosphorylated sites on the same substrate protein sequence are extracted as positive data and negative data respectively. Apart from the amino acids themselves, we also incorporated additional useful physicochemical information of each amino acid (e.g. hydrophobicity, accessible surface area, bulkiness,. ..) into the model. The chemical values and any combinations thereof are complemented to the feature set and learned in globally fashion from training dataset. The model performance was validated using *k*-fold cross and Jackknife validation procedures on a benchmark dataset (Phospho.ELM). Receiver Operating characteristic Curves (ROCs) and Area Under these ROCs (or AUC) were generated for comparing with existing methods. Compared to the other ones, this CRF based model worked well and in most cases it yielded better performances.

We are now deploying this CRF based model as a web server program. With this web server the end-users can receive a kinase-specific prediction of phosphorylated residues on any entered substrate protein sequences. Moreover, it also has the capability of yielding a prediction with a user-defined allowed False Positive Rate. This gives the end-users extra flexibility, especially in situations where either complete detection is required, or false positives are undesired.

[1] Diella,F. et al. (2008) Phospho.ELM: a database of phosphorylation sites update 2008. *Nucleic Acids Res.* **36**, D240-D244.

[2] Lafferty,J.D. et al. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*., 282 – 289.

Genomic rewiring that causes a severe regulatory challenge exposes flexibility in gene regulation and points out a novel adaptation mechanism

Lior David[1], Erez Braun[2]

[1] *Department of Animal Sciences, Hebrew University of Jerusalem, Rehovot 76100, Israel;*
[2] *Department of Physics & Network Biology Research Laboratories, Technion- Israel Institute of Technology, Haifa 32000, Israel;*

Instances of genomic rewiring, where expression of an existing gene is regulated by novel elements, are key evolutionary occurrences that contributed greatly to the remarkable diversity of life. So far most of what we understand on the role of genomic rewiring in evolution comes from comparisons among contemporary organisms. Such comparisons however, cannot reveal the adaptation process that is necessary for establishment of a new regulatory system. Thus, understanding how cells can overcome unforeseen regulatory challenges will teach us a lot about the properties of the regulatory network and will address one of the basic open questions in biology. Commonly, we assume that in accordance with the Darwinian Theory, populations comprise random genetic variability from which advantageous genotypes could be selected in certain environments. Since the DNA provides a memory mechanism, a population might more readily adapt to environments that were already encountered in its past evolution as compared to situations that confront the population with a novel challenge. The multitude of novel challenges that could emerge due to genome rewiring events requires flexibility in regulatory responses and adaptation capacity well beyond the given spectrum of genetic variability and thus suggests alternative adaptation mechanisms might exist. To study if alternative mechanisms for adaptation to novel challenges exist, we have confronted yeast cells carrying a rewired regulatory circuit, with a severe and unforeseen challenge. The essential *HIS3* gene from the histidine biosynthesis pathway was placed under the sole regulation of the *GAL1* promoter from the galactose utilization system. Histidine biosynthesis and galactose utilization are two separated yet evolutionary conserved pathways so connecting these two distinct modules most likely confronted the cells with an unforeseen regulatory challenge. Glucose containing medium strongly represses utilization of galactose including the essential *HIS3* and rewired cells are severely challenged. We show that this challenge was successfully met by a large fraction (>50%) of the population within only a few generations. The dynamics of the population in response to glucose medium leads to stabilization of a metabolic state that could be memorized and propagated for many generations and thus signifies a genuine adaptation process. By following the dynamics of the process we show this adaptation was due to the response of many cells to the challenge and thus was induced by the environment rather than a result of a common selection process. No fitness advantage was measured for adapted cells. Therefore, we provide evidence for a fast adaptation mechanism, accessible simultaneously to many cells and challenges the common view of selection as a sole adaptation mechanism. Such adaptation capability demonstrates the plasticity of regulatory networks and supports the significant potential of genomic rewiring in evolution.

An excess of human-oriented non-coding mutations are linked to expression divergence of genes in the ENCODE regions

Subhajyoti De and Sarah A. Teichmann

MRC Laboratory of Molecular Biology, Cambridge CB2 0QH. UK
Email: sde@mrc-lmb.cam.ac.uk, sat@mrc-lmb.cam.ac.uk

Evolutionary analyses of protein-coding gene-sequences are yet to explain many of the distinct phenotypic differences like cognition, bipedalism, complex behavior and diet between human and chimpanzees. On the other hand, findings of several recent studies (1-3) suggest that evolution of non-coding regions, and in particular those associated with regulation of gene expression may hold important clues.

Here we develop a novel method called Revolvar and scan the human ENCODE regions for windows enriched in mutations accumulated in humans at positions that are (i) strongly conserved (HOE) or (ii) strongly divergent (MAE) in other mammals. We report that presence of HOE elements in proximity of protein-coding genes is significantly linked to divergence of gene expression between human and chimpanzee, suggesting that many such HOE elements may overlap with cis-regulatory elements. Several of the ENCODE genes that contained HOE elements in their promoter regions are known to be involved in functions associated with cognition (e.g ACSL6, SYN3, NRX2), nutrition (e.g. EHD1) and skeletal development (e.g. OSTM1). Using Revolvar method, we also find subtle evidence of widespread mutations shared among all mammals in CpG-rich regions, and accumulation of lineage-specific mutations after loss of function in pseudogenes.

Extending our analysis to the complete human genome may provide us with many interesting insights into the human evolution.

1. Haygood, R *et al.* (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*. 39:1140-4.

2. Prabhakar, S *et al.* (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science.* 314:786.

3. Pollard, KS *et al.* (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature.* 443:167-72.

Abbreviation used:

HOE: Human oriented evolution

MAE: Mammalian accelerated evolution

The use of gene network modeling to analyze common and specific mechanisms of liver toxicity induced by the ORCs TNT, 2,6-DNT, 2,4-DNT, 4A-DNT, and 2A-DNT

Youping Deng[1, 2], Lynn L. Escalon[1], Xin Guan[1], Mitchell S. Wilbanks[3], Sharon A. Meyer[3], Natalia G. Reyero[4], Edward J. Perkins[5]

[1]SpecPro, Vicksburg, MS, USA; [2]University of Southern Mississippi, Hattiesburg, MS, USA; [3]University of Louisiana at Monroe, Monroe, LA, USA; [4]University of Florida, Gainesville, FL, USA; [5]US Army Engineer Research and Development Center, 3909 Halls Ferry Road, Vicksburg, MS, USA

The ordnance constituents (ORCs) 2,4-dinitrotoluene (2,4-DNT), 2,6-dinitrotoluene (2,6-DNT), 2,4,6-trinitrotoluene (TNT) and its microbial degradation products, 4-amino-2,6-nitrotoulene (4A-DNT) and 2-amino-4,6-dinitrotoluene (2A-DNT) are often found as pollutants in the environment. They have been reported to be toxic to humans and other animals. However, their molecular mechanism of toxicity is still unclear. Here we describe the use of gene expression profiles and network modeling to try to understand the mechanisms of toxicity of these compounds. We used a total of 200 microarrays to analyze liver samples from rats that had been exposed to one of 5 doses for each one of the five compounds (TNT, 2,6-DNT, 2,4-DNT, 4A-DNT, or 2A-DNT) at either 24h or 48 h. We found that 2,6-DNT was the strongest regulator to affect the expression of the highest number of genes , followed by 2,4-DNT, TNT, 4A-DNT and 2A-DNT. While each chemical induced a distinctive expression pattern, there was a common set of genes that were significantly regulated by all chemicals. We used this set of common genes to build a gene network, using both data from the literature and our microarray data. The resulting network was mainly involved in the metabolic pathway of xenobiotics by Cytochrome P450; androgen and estrogen metabolic pathways; and the PXR/RXR activation pathway. The analysis of the biological function has shown that these genes are strongly associated with hepatic system disease and liver cholestasis. We also constructed specific gene networks for each one of the specific chemicals. Moreover, we found that all the chemicals with the lowest doses (close to background) could modulate the expression of a number of genes that were connected to some diseases and physiological functions. Overall, our results indicate that the genomics and systems biology strategy can help to understand the common and specific mechanisms of toxicity of DNT, TNT and their derivatives. The genomics approach has been proven to be a very sensitive method to detect specific gene expression changes due to chemical exposure.

Evaluation of local reliability of gene networks inferred from time series expression data.

Francesco Sambo[1], Barbara Di Camillo[1], Marco Falda[1], Gianna Toffolo[1], Silvana Badaloni[1]

[1]*Department of Information Engineering, University of Padova, Italy*

Reverse Engineering algorithms are commonly used to infer gene networks from microarray data. However, it is not clear how reliable this methods are and what determines the reliability of these methods. To this purpose, we analyze the behaviour of two Reverse Engineering algorithms, Dynamic Bayesian Networks and Causal Networks, on a set of realistically simulated networks and dynamic expression data. The two algorithms have been chosen because both reconstruct directed graphs from time series data and are representative of two different approaches to reverse engineering: a model based method applied to continuous expression data and a Mutual Information based method applied to quantized data.

Since at present no biological network is known with sufficient precision to serve as a standard, quantitative assessment of reverse engineering algorithms is accomplished using simulation studies. This allows the assessment of average performance on a large number of networks and experimental conditions. In particular, in-silico data consist of 60 networks of 10 genes and 60 networks of 20 genes, generated with a simulator, which: 1) generates network topology with a scale-free distribution of the connectivity and a constant clustering coefficient; 2) creates gene expression profiles by using differential equations, accounting for saturation effects and transcription activation thresholds. For each network, 4 different time series of 50 samples are generated: the first is obtained observing free evolution from random initial conditions; the other three are obtained stimulating the hub, i.e. the node with the highest out degree, with a sinusoid, a ramp and a step signal respectively, so to study the effect of the propagation of the stimulus through the network.

Performances of the algorithms are always higher when the system is externally stimulated. In particular, performance is significantly higher than the average on edges directly outgoing from the stimulated node and unrecognized nodes concentrate in regions far from the stimulus.

In the intermediate portions of the networks, performance is significantly higher than the average on edges whose destination nodes have lower in-degree and clustering coefficient, using both externally stimulated and not externally stimulated data. Consistently with this observation, specific motif structures, such as feedback loops involving more than three nodes, are recognised with significantly higher performance than others. Although the two algorithms are able to identify portions of the network with similar topological properties, the sets of true positives they identify seldom shows an intersection greater than 50%. On the opposite, false negatives produced by the two algorithms overlaps for the 75% on average, showing that there are portions of the networks inherently difficult to be detected by both algorithms.

Understanding the limitations of Reverse Engineering problems is fundamental for the design of both experiments and algorithms: we intend to explore further this direction, both theoretically and experimentally.

Annotating the Zebrafish Transcriptome

Anthony DiBiase, Anhua Song, Fabrizio Ferre, Gerhard Weber, David Langenau, Len Zon, and Yi Zhou

Stem Cell Program and Division of Hematology and Oncology, Children's Hospital Boston and Howard Hughes Medical Institutes, Harvard Medical School, Boston, MA 02115

Our informatics research focuses on enabling the translation of knowledge of biology and disease from the zebrafish model system to humans. Two ongoing projects are presented here: Zebrafish Gene Transcript Collection (ZGTC) and Zebrafish Gene and Microarray Annotation Project (ZGMAP). ZGTC is a relatively stable, genome-wide and non-redundant collection of zebrafish gene transcripts and aims to facilitate gene annotation and expression profile studies. 31,052 different transcripts, encoded by 25,494 genes and 9,063 splice variants are represented on a new NimbleGen Roche microarray. ZGMAP is in its second phase and annotates genes and microarray probes by assigning zebrafish orthologs to known human genes using both manual and automated procedures. ZGMAP takes a conservative approach and trades accuracy for genome coverage. The first phase the project has resulted in 7,680 high-quality orthologous pairs. Our initial analyses of first phase results estimated the ZGMAP ortholog accuracy at over 91 percent. The second phase annotates the entire ZGTC and is nearing completion. The continuation of genome sequence annotation and its associated projects in the post genome sequencing era is critical for advancing the use of the zebrafish system for studying vertebrate genetics and developmental biology, and for modeling of human diseases.

Integrating knowledge on transcriptional regulation of prokaryotic genes in the RegTransBase database and the semi-automatic analysis system.

Michael J. Cipriano[1], Pavel Novichkov[1,4], Alexei E. Kazakov[2], Dmitry Ravcheev[2], Adam P. Arkin[1,3,4], Mikhail S. Gelfand[2], Inna Dubchak[1,4,5]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [2]Institute for Information Transmission Problems, RAS. Moscow, 127994, Russia; [3]Department of Bioengineering, University of California, Berkeley, CA, 94710; [4]Virtual Institute of Microbial Stress and Survival http://vimss.lbl.gov/, Berkeley, CA, 94710; [5]Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

RegTransBase database has been developed as a component of the MicrobesOnline/RegTransBase framework. RegTransBase contains experimental data on the regulatory interactions with known elements in a wide range of bacteria and archaea. It is manually curated and based on published scientific literature. Current database content is derived from close to 4500 relevant articles describing over 10000 experiments in relation to 180 bacteria and archaea. It contains data on the regulation of ~18000 genes and evidence for ~8800 interactions with ~970 regulators. It is available at http://regtransbase.lbl.gov.

RegTransBase allows searching for experimentally proven facts of regulation based on the different types of classifications, such as the type of experimental results, taxonomy, phenotype, habitat, genome relevance. More sophisticated classifications will include metabolic pathways, effectors and types of experimental techniques.

RegTransBase additionally provides an expertly curated library of alignments of known transcription factor binding sites for many bacterial and archaeal species. Each alignment contains information on the transcription factor which binds the DNA sequence, the exact location of the binding site on a published genome, and links to published articles.

RegTransBase builds upon these alignments by providing a set of computational modules for the comparative analysis of regulons among related organisms. These modules guide a user through the appropriate steps of transferring known or high confidence regulatory binding site results to other microbial organisms, allowing them to study many organisms at the same time, while warning of analysis possibly producing low confidence results, and providing them with sound default parameters.

There is an increasingly tight coupling of RegTransBase with MicrobesOnline in reporting cis-regulatory sites and regulatory interactions, and integrating RegTransBase searches into MicrobesOnLine cart functions.

SLIPPER: An Iterative Mapping Pipeline for Short DNA Sequences

Lee Edsall[1,2], Bing Ren[1], Terry Gaasterland[2]

[1] Laboratory of Gene Regulation, Ludwig Institute for Cancer Research, Department of Cellular and Molecular Medicine, University of California, San Diego School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA; [2] Laboratory of Computational Genomics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Next generation sequencing technologies can generate millions of short DNA sequences for mapping to a reference genome. While Illumina's Genome Analyzer Pipeline accurately and efficiently maps some sequences, there are some sequences it cannot map. We developed an iterative mapping pipeline called SLIPPER that increases the number of sequences mapped, providing a larger dataset that can be used to gain greater insight into biological processes such as transcription and gene regulation.

Ambiguous base calling, incorrect base calling, and single nucleotide polymorphisms between the cell line used in the experiment and the reference genome contribute to the number of mismatches between a read and the reference genome. Illumina's software limits the number of mismatches to 2 and may result in mappable reads being unmapped.

Illumina's Genome Analyzer Pipeline version 1 generates sequences from experimental data and performs an initial mapping to the human genome. Sequences are mapped to hg18 with at most 2 mismatches.

Sequences that do not map are extracted and used as input to SLIPPER, which proceeds in two phases.

In Phase 1, 25-base subsequences of unmapped sequences are converted to FASTA format and used as input to Illumina's Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) mapping software. The subsequences are mapped to hg18 with at most 2 mismatches.

In Phase 2, Active Motif's TimeLogic DeCypher System is used to align the remaining unmapped sequences against the human genome using the BLASTN algorithm. The unmapped sequences from Phase 1 are converted to FASTA format and aligned against hg18. Sequences must align with an E-value of $1x10^{-3}$ or better.

Combining Illumina's Genome Analyzer Pipeline with SLIPPER increases the proportion of mapped sequences compared to using the Illumina software alone. A proof of principle analysis of 40 lanes taken from six different experiments was performed. The total number of reads mapping to the genome increased by an average of 3.73%. An average of 83.50% of those mapped uniquely to the genome.

Silencing by microRNAs is strongly modular in human

Gabor Boross [1,2], Katalin Orosz [1,2], Illes J. Farkas [1,2]

[1]Dept. of Biological Physics, Eotvos Lorand Univ.; [2] Statistical and Biological Physics Res. Group of the Hung. Acad. of Sci., Pazmany P. stny. 1A, H-1117 Budapest, Hungary

Short regulating RNAs are essential for guiding a large number of cellular processes. Their main role is to reduce or fully inhibit the production of proteins, the type of biomolecule with the broadest spectrum of cellular locations, biochemical functions and interactor partners. Compared to transcription factor proteins short RNAs appear to provide more specialized control over cellular processes, for example, single deletions less frequently cause observable phenotype changes.

After comparing four computational databases of microRNA targets in human we find that despite large differences the four lists become increasingly similar with growing sample size and we use one of the lists instead of data integration. Based on microRNA - target gene silencing scores we define the co-regulation score of two microRNAs and in the resulting weighted network we locate modules of tightly co-regulating microRNAs. The well-separated co-regulation modules we identify quantitatively confirm on the system level previous, mostly qualitative, results about the specialized control RNAs provide. By comparing the roles of individual microRNAs within the modules we define and compute for each human microRNA its relative essentiality readily available for experimental testing.

Support:

Hungarian National Science Fund (OTKA)

Hung. Natl. Office for Res. and Techn. (NKTH) in CellCom RET

A Systematic Study of The Murine Dendritic Cell Response to Lipopolysaccharide Through Time-Series Analysis of Gene Expression and Multiplex Cytokine Assays

Feng Feng[1], Ana Paula Sales[1], Thomas B Kepler [1,2]

[1]*Duke University Laboratory of Computational Immunology, Durham, NC 27705;*
[2]*Department of Biostatistics, Duke University, Durham, NC 27705;*

Vaccine adjuvants are known to exert their influence on adaptive immunity primarily through dendritic cells, the most potent of the antigen presenting cells (APC). The rational design of novel vaccine adjuvants to elicit strong memory responses to pathogen components while avoiding excessive inflammation therefore necessitates an understanding of dendritic cell biology. For the current study, we generated gene expression profiles on both primary murine dendritic cells and a murine APC cell line *in vitro* at 8 different time points through two days in response to lipopolysaccharide (LPS). In addition, the culture supernatants were analyzed for the concentrations of 23 cytokines at each time point. Statistical analysis using Gaussian Process/Dirichlet Process modeling revealed a multi-phase response to LPS treatment, with the first-phase peaking at about three hours, and subsequent phases appearing sequentially thereafter. Furthermore, we identified over 2000 genes that participate in these response waves, under stimulation by LPS. Some of these genes were confirmed by quantitative PCR. Gene ontology and pathway analyses were also applied to these data to capture the underlying regulatory pathways mediating the LPS effect at a higher organizational level. Through such profiling studies and analyses, interesting genes, molecules and pathways—and their interactions—have been identified. This study with LPS is the first of several such studies that will be conducted using other potential adjuvant compounds under the similar experimental conditions to move us toward the rational design of vaccine adjuvants.

Most Mammalian mRNAs are Conserved miRNA Targets

Robin C. Friedman[,1,2,3], Kyle Kai-How Farh[1,2,4], Christopher B. Burge[1], David P. Bartel[1,2]

[1]Department of Biology, Massachusetts Institute of Technology; [2]Whitehead Institute for Biomedical Research and Howard Hughes Medical Institute; [3]Computational and Systems Biology Program, Massachusetts Institute of Technology; Division of Health Sciences and Technology, Massachusetts Institute of Technology;

MicroRNAs (miRNAs) are small endogenous RNAs that bind to short, seed-matched sites in the 3'UTRs of mRNAs to direct the posttranscriptional repression of these messages. Comparative sequence analyses have shown that, on the whole, a sizable number of these sites are preferentially conserved, but previous analyses were not suitable for measuring the probability of individual sites being selectively maintained and may have substantially underestimated the true scope of conserved targeting. We have overhauled our comparative genomics tool for finding and analyzing the conservation of 3' UTR motifs and applied it to the analysis of human miRNA sites. The new tool more completely controls for background conservation by accounting for mutational biases, dinucleotide conservation rates, and the conservation rates of individual UTRs. These improvements allow statistically powerful analysis of the conservation of individual miRNA target sites — an important criterion for assessing which of the many sites are most promising for experimental follow-up. Moreover, substantial increases in sensitivity were achieved by more efficiently incorporating new genomes and different seed-match types. When considering sites matching the 87 mammalian miRNA families conserved to chicken or beyond, over 45,000 sites within human 3'UTRs are conserved above background levels. This number represents a ~3-fold increase over the most sensitive lower-bound reported previously and indicates that more than 60% of human protein-coding genes have been under selective pressure to maintain pairing to miRNAs. Mammalian-specific miRNAs have far fewer conserved targets than do the more broadly conserved ones, even when considering only more recently emerged targets. Several seed-match types contribute similar numbers of targets, indicating that purifying selection acts on weaker but more common target sites roughly as often as on stronger but rarer sites. Although pairing to the 3' end of miRNAs can compensate for seed mismatches, this class of sites appears to constitute less than 2% of all conserved sites. Our greatly expanded set of target predictions, which now include conserved 3'-compensatory sites, are available at the TargetScan web site, which displays the conserved targeting probability for each site and each predicted target.

A Novel Sequence Based Model for Nucleosome Positioning in Yeast

Mahmoud Ghandi, Michael Beer

Department of Biomedical Engineering and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University

Nucleosome positioning and chromatin remodeling is thought to play a significant role in transcriptional regulation by controlling the accessibility of DNA binding proteins to their DNA binding sites. It has been shown *in vitro* that some DNA sequences have lower affinity to bind histones and/or are less flexible to form nucleosomes [1]. Data obtained from isolating and sequencing nucleosomal bound DNA fragments in *C. elegans* suggest that nucleosomal bound DNA is more GC rich than genomic DNA [2]. Microarray experiments have provided nucleosome positioning data in yeast [3], and we are collaborating to analyze more recent experiments which extend this approach using whole genome tiling arrays. Current mathematical and computational models have low accuracy to predict nucleosome positioning [1][4]. We are developing a sequence based model to predict nucleosome positioning in yeast and *C. elegans*. We employ a context-based arithmetic coding scheme to build two probability models for DNA sequences, one for the nucleosome bound regions and one for the nucleosome free regions. In this scheme the probability of each base pair occurring in nucleosomal bound vs. unbound regions is estimated from its local sequence context (neighboring base pairs in both directions). The entropy of a given DNA segment is computed using each of the probability models, and each DNA fragment is classified as nucleosome bound or unbound based on which model resulted in lower entropy. The distributions for each class are inferred from whole genome nucleosome positioning data available from *S. cerevisiae* and *C. elegans*. This method should be general enough to capture sequence affinity for nucleosomes due to either flexibility or electrostatic DNA-histone interactions, but further improvements would incorporate the geometry of the DNA-histone complex. Preliminary results indicate that this approach can successfully identify more than 60% of the nucleosome bound sequences by their DNA sequence alone, while only 20% of unbound sequences are misclassified as bound. In future work, we plan to incorporate our models of nucleosome positioning into more accurate predictive models of combinatorial gene expression.

References

[1] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J. Z. Wang, J. Widom, "A genomic code for nucleosome positioning", Nature, Vol. 442, No. 7104. (19 July 2006), pp. 772-778.

[2] S. M. Johnson, F. J. Tan, H. L. McCullough, D. P. Riordan, A. Z. Fire, "Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin", Genome Res. 2006 Dec;16(12):1505-16.

[3] G. Yuan, Y. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, O. J. Rando, "Genome-Scale Identification of Nucleosome Positions in S. cerevisiae", Science 22 July 2005, Vol. 309. no. 5734, pp. 626 – 630.

[4] G. Yuan, J. Liu, "Genomic Sequence is Highly Predictive of Local Nucleosome Depletion ", PLoS Computational Biology, January 2008, Vol. 4, Issue 1, pp. 164 – 174.

Genetic algorithm suggests novel role for codons in the N-terminal region of some protein-coding transcripts

Sean Connolly[1], Elizabeth Grayhack[2], and Shuba Gopal[1]

[1] Dept. of Biological Sciences, Rochester Inst. Of Tech., Rochester, NY 14623, USA
[2] Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA

We used a genetic algorithm (GA) to better understand the role of codon usage in protein-coding transcripts that must be expressed at low levels. Preferential codon usage and its correlation with high expression is well documented, but little is known about whether certain codons might be selected to preferentially reduce the expression of a transcript. To study this, we utilized a novel dataset on protein expression in *Saccharomyces cerevisiae* from the movable ORF (MORF) collection. Each MORF is an epitope-tagged vector containing a gene's coding region with a common extragenic promoter and terminator. As a result, the sole determinants of expression are properties of the open reading frame. Expression within the MORF collection correlates well with native expression.[1] Using data from the MORF collection, we studied codon usage and expression levels in a set of genes that were cytotoxic if over-expressed. We hypothesized that these genes' transcripts might provide insights into codon usage in transcripts that must be expressed at low levels. Codon adaptation index (CAI)[2] has an accuracy of only 45% for genes in this class (35% of known high expressors and 56% of known low expressors). In contrast, we have used a GA to optimize a codon scoring scheme that has an overall accuracy of 61% (65% of high expressors and 57% of low expressors correctly predicted). Interestingly, the GA optimized scoring scheme was able to correctly predict the expression level of 71% of low expressors by considering just the 50 codons at the N-terminus of each gene transcript. The GA scoring scheme has an accuracy of 64% in this region (56% of high expressors and 71% of low expressors correctly predicted). In contrast, CAI has an accuracy of only 40% (24% of high expressors and 56% of low expressors). This suggests that transcripts expressing at low levels may in fact "set" their expression level very early in the translation process by manipulating codon composition at the N-terminus of the transcript. The mechanisms that might underlie such preferential codon organization will be explored, along with detailed descriptions of the GA and results obtained.

1. Gelperin DM, White MA, Wilkinson ML, Kon Y, Kung LA, Wise KJ, Lopez-Hoyo N, Jiang L, Piccirillo S, Yu H, Gerstein M, Dumont ME, Phizicky EM, Snyder M, Grayhack EJ. (2005). Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes Dev.* 19:2816-2826.

2. Sharp P and Li W. (1987). The codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.* 15:1281-1295.

Sequence determinants of origin localization in metazoa

Raluca Gordan[1], David MacAlpine[2], Alexander J. Hartemink[1]

[1]Department of Computer Science; [2]Department of Pharmacology and Cancer Biology;
Duke University, Durham, NC 27708

DNA replication, one of the fundamental steps in the cell cycle, is initiated at specific sites across the genome, called origins of replication. These sites direct the assembly of multi-protein complexes that eventually lead to the formation of replication forks. The selection of replication origins is carefully regulated by the cell, as this process is essential for genomic stability. Each origin is marked by the formation of the pre-replicative complex, which starts with binding of the origin recognition complex (ORC) at specific DNA sites. The sequence specificity of ORC has been well characterized in *S. cerevisiae*, where a consensus binding motif has been identified. In higher eukaryotes, however, little is known about the sequence determinants that guide ORC toward specific DNA sites, and it is believed that ORC does not have strong sequence specificity. Although the proteins that comprise ORC are evolutionarily conserved in eukaryotes, no ORC binding motif has been identified to date in organisms other than *S. cerevisiae*.

Here, we analyze genomic regions bound by *D. melanogaster* ORC in chromatin immunoprecipitation experiments. Our goal is to discover what sequence determinants, if any, are important for ORC binding and origin localization in metazoa.

We first searched for over-represented motifs in the set of sequences bound by ORC in *D. melanogaster* using PRIORITY with informative priors based on DNA stability profiles and evolutionary conservation. We chose PRIORITY because, unlike other motif finders, it successfully identified the known ORC binding motif in yeast. When applied to the fly sequences, the algorithm reports mostly repetitive sequences (*e.g.,* (AC)n, (AG)n) and low complexity AT-rich motifs. These sequences may interact directly with ORC or alternatively, they may have an indirect effect on ORC binding or subsequent replication initiation by preventing nucleosome formation (*e.g.,* in the case of the (AC)n repeat) or by facilitating DNA unwinding (in the case of AT-rich motifs).

Next, we trained an SVM based on sequence features to distinguish between DNA regions bound versus not bound by ORC. For each training sequence the classifier uses 2772 features that describe the frequency of all possible words of size 1 to 6 in that sequence. After feature selection (yielding ~200 essential features), the classifier achieves remarkable performance: 92% accuracy in a 10-fold cross-validation test, and an AUC-ROC of 0.96. This suggests that the DNA sequence itself contains plenty of information regarding ORC binding and origin localization. What features are most important for the performance of the classifier? Not surprisingly, some of the best features correspond to oligomers that are part of motifs reported by PRIORITY (ACACAC, AGAGAG), while other discriminating oligomers (AATA, AAAACA, CGAAA, GAGCGA, etc.) have been shown to be correlated with nucleosome depletion.

Our analysis is consistent with the hypothesis that the DNA-binding specificity of metazoan ORC may not be well described by a single motif. Nevertheless, we find that complex sequence determinants are strongly predictive of ORC association, possibly due to an indirect effect, at metazoan origins of replication.

Unraveling the “backbone” of vertebrate gene regulation

Valer Gotea[1], Ivan Ovcharenko[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894

The understanding of complex gene regulation in higher metazoans has advanced significantly during recent years when more and more complete genome sequences have become available. It has even become possible to identify regulatory elements with shared function between distant genomes beyond the level permitted by sequence similarity.

Here we attempt to identify vertebrate regulatory elements in the form of homotypic clusters of binding sites, defined as short regions with multiple copies of binding sites of the same type. This type of clusters has been described mostly in Drosophila, but little attention has been paid to them in vertebrates, where most efforts concentrated to identify heteroptypic clusters. The homotypic clusters offer the advantage of providing the redundancy needed for key regulatory events, allowing low-abundance transcription factors to find their targets, or even providing a way to respond to various concentrations of a particular transcription factor. Multiple copies of binding sites of the same type are known to exist even in heterotypic clusters, where redundancy and/or dosage are important.

In our approach we used a hidden Markov model to identify homotypic clusters of binding sites corresponding to every matrix available from TRASNFAC and JASPAR3 databases for vertebrate species. Clusters were identified independently in the genomes of human, mouse, chicken, and fugu. We find that such clusters are present throughout all these genomes, and for 272 non-redundant matrices we could identify clusters that are conserved at least beween human and mouse, indicating that vertebrate species can also employ this type of simple regulation. The conservation with mouse is defined here by whether an independently identified cluster in mouse is present in the orthologous region indicated by an evolutionary conserved region (ECR) if one is located within the human cluster, or otherwise by the two flanking non-repetitive ECRs. Using ChIP-chip and ChIP-seq data available for STAT1, NRSF, YY1, E2F1, E2F4, and c-MYC, we verified that our cluster predictions overlap significantly with experimental data ($p\text{-val} < 0.001$ for each of these transcription factors), indicating that homotypic clusters are indeed bound by their corresponding transcription factors. For 119 matrices, we found that their clusters are preferentially found near promoter regions, indicating a direct contribution to gene regulation. Interestingly, a significant class of genes targeted by such clusters is that of transcription factors themselves. By further restricting our dataset to 453 clusters that are conserved in human, mouse, chicken, and fugu, we found that among the 378 target loci the only significant class of genes targeted by such deeply conserved clusters is that of transcription factors (1/3 of the target loci), indicating that these might be a set of genes found at the top of the regulatory network shared by all vertebrate species, around which additional complexity has developed.

In conclusion, we present evidence that homotypic clusters of binding sites are present throughout the vertebrate lineage, and their deep conservations indicates an important role in vertebrate gene regulation.

Ethanol related gene resource and candidate gene analysis

An-Yuan Guo[1], Bradley T. Webb[2], Michael F. Miles[3], and Zhongming Zhao[1,4]

[1]Department of Psychiatry and Virginia Institute for Psychiatric and Behavior Genetics, Virginia Commonwealth University, Richmond, VA, USA

[2]Department of Pharmacy, Virginia Commonwealth University, Richmond, VA 23298, USA

[3]Department of Pharmacology/Toxicology, Virginia Commonwealth University, Richmond, VA 23298, USA

[4]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA

Over the last decade rapid progress has been made in the study of ethanol-related traits including alcohol dependence and behavioral responses to ethanol in both humans and animal models. To collect, curate, integrate these results so as to make them easily accessible and interpretable for researchers, we developed ERGR, a comprehensive ethanol-related gene resource. We collected and curated more than 30 large-scale data sets including linkage, association, and microarray gene expression from the literature and 21 mouse QTLs from public databases. At present, the ERGR deposits ethanol-related information of ~7000 genes from 5 organisms: human (3311), mouse (2129), rat (679), fly (614) and worm (228). ERGR provides gene annotations and orthologs, detailed gene study information (e.g., fold changes of gene expression, *P* values), and both the text and BLAST searches. Moreover, ERGR has data integration tools such as for data union and intersection and candidate gene selection based on evidence in multiple datasets or organisms. ERGR has a user-friendly web interface with browse and search functions at multiple levels.

Using the data we collected and curated, we prioritized 42 ethanol related candidate genes based on the evidence in multiple datasets and in multiple organisms. These candidate genes include not only some well studied ethanol related genes (e.g., *ADH*, *ALDH*, *GABA* receptors, and *NPY*) but also some rarely studied genes (e.g., *CPE*, *GFAP*, *CRYAB*, *GAD1* and *NTRK2*) with multiple lines of evidence. We found a motif 'TACGAAGG' in the promoter regions of most of these candidate genes by performing motif search on Weeder web server. Gene ontology analysis indicated that these genes are enriched in GABA signaling pathway, cell-cell signaling, neurophysiological process, synaptic transmission, transmembrane receptor activity, and ion transporter activity. The pathway analysis revealed that 10 of the 42 genes are in neuroactive ligand-receptor interaction pathway. Using Ingenuity Pathway Analysis system, we found that MAP kinase, Akt, AGT, retinoic acid, PTK2, and calmodulin are the central nodes linking these candidate genes.

Gene-expression prediction by the elastic net

Mika Gustafsson[1], Michael Hörnquist[1]

[1]*Dept. of Science and Technology, Linköping University, Sweden. [mikgu,micho]@itn.liu.se*

To predict the differential gene expressions for new experiments is an important endeavour within computational systems biology. It can both be a way to appreciate how drugs affect the system, as well as providing a framework for finding which genes are interrelated in a certain process. Here we predict the order of the levels for 50 genes out of 9335 Affymetrix probes for one time-series out of four, with eight experiments in each. This is the Gene Expression Prediction Challenge of DREAM3.

All Affymetrix probe names are mapped onto gene names, which enable incorporation of external data sources, such as other expression data sets and TF-binding data. This mapping reduces the number of units to 7805. We assume the level to be predicted is a linear combination of all other (known) levels, given as log-ratios, together with the estimated derivatives of those levels (thereby, we have 2 x (7805 – 50) = 15510 putative predictors for each gene). The parameter fit is performed as a minimization problem supported by some method to handle the lack of data. These methods cover lasso, weighted regularized least absolute deviation (wrlad), ordinary least squares (ols) with minimal L2 norm, ridge regression and the elastic net, where the model selection criterion is a three-fold cross-validation (CV) with respect to the three complete time-series. We also explore various forms of pre-transformations of the given log-ratios, including power-, exponential-, and arctangent-transformations.

The most promising method turns out to be the elastic net, avoiding derivatives and pre-transformations. This method enforces constraints both on the L1 and L2 norms of the coefficients, and we determine both the mixture and the magnitude of these norms by CV. To take into account public data, we download from the ncbi omnibus webpage (www.ncbi.nlm.nih.gov/geo/) a compendium of 256 gene expression profiles in time-series, and 515 profiles from the Rosetta database (Hughes et al, 2000, Bernardo et al 2005). These profiles are used independently in the weighted elastic net inference, but their importance turn by CV estimate out to be significantly less than the profiles given in the challenge. Finally, we integrate transcription factor (TF) binding data, downloaded from the public database Yeastract (www.yeastract.org). Here we utilize the full database, which consists of all documented TF-binding interactions, but with half the weight for bindings which are not experimentally confirmed. From this data, and for each gene, we count the number of identical upstream TFs. Based on the fraction of shared TFs, we promote those genes which share TFs as predictors for each other. Also this integration is performed in a soft, data-driven manner by CV, and the inclusion is only to the extent it actually lowers the prediction error.

Finally, we utilize all the parameter values for the model settled by the procedure above. For the test data, we obtain a Spearman rank correlation above 0.81, and hopefully the correlation for the time-series where the levels should be predicted is even higher.

A phylum-wide catalog of cis-regulatory elements and their evolution

Naomi Habib[1,2], Ilan Wapinski[3,4], Hanah Margalit[2], Aviv Regev[4], Nir Friedman[1]

[1] School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel; [2] Department of Molecular Genetics and Biotechnology, Faculty of Medicine, The Hebrew University; [3] School of Engineering and Applied Science, Harvard University; [4] Broad Institute of MIT and Harvard

Divergence of gene regulation is likely a major driving force in species evolution, from bacteria to mammals. This hypothesis relies however on a few specific examples and very general observations. Large-scale comparative analyses across species can reveal important new insights into the evolution of transcriptional regulation. While comprehensive maps between factors and their downstream targets are unknown for most non-model organisms, much can be inferred by considering the repertoire of cis-regulatory motifs in their genomes, the specific genes in which they occur, and their evolutionary trajectories. The motifs can be viewed as compact and informative representation of the regulatory signals.

Here we present a novel computational approach to define phylum-wide catalogs of functional DNA elements (motifs) that captures the regulatory code for dozens of Ascomycota species, spanning over 300 million years of evolution.

By tracing the evolutionary history of these regulatory elements we are able to capture an accurate regulatory map for each species and deduce how the mechanisms of transcription regulation have evolved in this phylogeny. We identify global trends of cis-regulatory conservation, turnover and divergence, as well as investigate these patterns between the targets of each regulatory element individually, suggesting testable hypotheses. Taken together, we are able to identify the significant events in the history of the yeast regulatory code.

GRENDEL: An in-silico Benchmarking Platform for Gene Regulatory Network Inference Algorithms

Brian C. Haynes[1], Michael R. Brent[1]

[1]*Department of Computer Science and Engineering, Washington University, St Louis, MO*

High throughput assays for mRNA expression have paved the way for computational methods that aim to reverse engineer the control architecture of gene regulation, but objectively evaluating these methods remains a challenge that has impeded their progress and adoption as practical tools. Evaluating gene network inference methods is difficult primarily due to the paucity of fully understood biological networks to use as gold standards. In-silico gene regulatory networks hold promise as an alternative where the ground truth is known, but the available network generation algorithms do not take full advantage of what we know about the kinetic and topological parameters of networks in model organisms such as E. *coli* and S. *cerevisiae*.

To generate and simulate more realistic synthetic networks, we developed an open and extensible software toolkit: Gene Regulatory Network Decoding Evaluations tooL (GRENDEL). GRENDEL generates random gene regulatory networks according to user defined constraints on the network topology and kinetics. The available topological parameters allow for compact in-degree distributions that fit those observed in S. *cerevisiae* better than the standard power-law distributions. The kinetic parameterization allows for sampling protein and mRNA half-lives, translation rates, and transcription rates, from the actual parameters estimated for S. *cerevisiae* from genome-wide measurements. We took advantage of both these features to increase the realism of the experiments reported below.

We evaluated the utility of synthetic benchmarks for two applications: assessing the relative performance of inference methods and supporting cost-benefit analysis of designs for gene expression experiments. Our results show that using realistic degree distributions can change the relative accuracy estimates for three well-known algorithms: ARACNE, CLR, and Symmetric-N. The design of experiments for measuring gene expression under a variety of conditions also had a huge effect on their relative accuracies. Surprisingly, simulated technical noise in gene expression measurements had a small effect on inference accuracy. We also evaluated two methods that can make use of gene expression time courses: DBmcmc and Symmetric-N. Networks using realistic kinetic parameters showed a greater accuracy improvement from increased sampling frequency than simulations using arbitrary parameters from the A-BIOCHEM benchmark, however, even with realistic kinetic parameters, sampling gene expression every 2 minutes improved inference accuracy by less than 15% relative to sampling hourly.

Annotating functional regions of the human genome with frequently occurring chromatin signatures

Gary Hon[1,2], Bing Ren[1-3], Wei Wang[1,4]

[1]Bioinformatics Program; [2]Ludwig Institute for Cancer Research; [3]Department of Cellular and Molecular Medicine and Moores Cancer Center; [3]Department of Chemistry and Biochemistry, University of California at San Diego.

Transcriptional regulation in human cells is a complex process involving a multitude of genomic regulatory elements. One of the central problems confronting genome-scale understanding of human transcription is an incomplete catalog of these regulatory elements and an even more incomplete understanding of how they affect transcription. Recent studies have shown that distinct chromatin signatures mark a variety of functional genomic elements, and that variants of chromatin signatures confer different functional specificities. To aid in the identification and annotation of novel regulatory elements in a systematic and unbiased way, we apply a computational clustering method called ChromaSig on recently published maps of 21 histone modifications (Barski et al 2007). We find distinct chromatin signatures at known classes of regulatory elements including promoters and enhancers that correlate with gene activity. To identify novel chromatin signatures spanning lesser-studied loci, we also apply ChromaSig to regions of the genome with strong enrichment of histone modifications. We observe unexpected chromatin signatures marking exons that are preferentially included in mature mRNAs and other signatures marking excluded exons, suggesting that co-transcriptional splicing is a widespread phenomenon that may be directed by chromatin structure.

Tip growth with molecular details

Seungpyo Hong[1], Dongsup Kim[1]

[1]Department of Bio and Brain Engineering, KAIST, Daejeon, Republic of Korea

Polarization of a cell takes important roles in motility, division, and morphogenesis. Here we are focusing on the root hair cells which have long hairs protruding several hundred micro-meters. What causes this polarization and how it is sustained are the main focus in the root hair cell research. Experimentalists found that a group of proteins and molecules - RHD2, Rho, RhoGDI, calcium ion and reactive oxygen species(ROS) - are involved in this process. Cellular structures including actin filament and vesicles are also important for the persistent tip growth. Several mathematical models have been suggested on this growth. Some models assume continuous incorporation of vesicles and observe expansion pattern of the membrane. Others are focused on how stochastic growth of actin filament changes the shape of the cell. However, none of them explains where the vesicles or actins come from. Here, we propose a model which explains the root hair cell growth in terms of molecular species involving tip growth. Especially, we modeled vesicle transport via the actin filament. This model shows the polarization of the cell and the growth of the tip and suggests more detailed mechanism of polarization and the hair growth.

We constructed a 3D cell model based on the experimental results. It is constituted of 1 μm^3 cubes with dimension of 15 μm x 6 μm x 24 μm as an initial structure. Three topologies, cytosol, Golgi complex, and membrane, are considered. Five molecular species involved in root hair growth are gathered from the previous literature: RHD2, Rho, RhoGDI, ROS and calcium ions. Their interactions are represented by the reaction-diffusion equations. The formation of actin filament and the transport of vesicle from Golgi complex to the growing tip are explicitly considered in our model. The concentration of actin filaments determines the intensity of the direction, and the concentration gradient of actin filaments determines the direction of the diffusion. Based on this model, partial differential equations (PDE) are generated and solved using an explicit finite difference method, and resulting ordinary differential equations (ODE) are solved by ‘ode15s’ module in MATLAB7.

Our model suggests that the initial polarization of the cell starts from the random fusion of vesicles onto the membrane. Then, positive feedback loops are working to sustain the localization of the signal. We can observe the pillar like growth of actin filament and the accumulation of vesicle at the growth tip reproducing the root hair cell growth. The dynamic equilibrium between vesicle synthesis at the Golgi complex and the consumption at the growth tip is important in sustaining the tip growth. The concentration of vesicle at the tip decreases with the tip growth. This explains the short root hair in Rho defect mutant; the growth is terminated earlier because spontaneous membrane fusion requires high concentration of vesicles at the tip.

In this research, we made a mathematical model describing the growth of the root hair cell. Unlike the previous mathematical models, we linked the root hair cell growth phenomenon with the several crucial molecular species. Thus, the function of each molecular species and the effect of their interactions can be studied more deeply.

In-silico network predictions by ODE and lasso

Mika Gustafsson[1], Michael Hörnquist[1]

[1] *Dept. of Science and Technology, Linköping University, Sweden. [mikgu,micho]@itn.liu.se*

To predict the edges in biological networks is an important endeavour within computational systems biology. It can both be a way to appreciate how drugs affect the system, as well as providing a framework for finding which genes are interrelated in a certain process. Here we explore one reverse engineering algorithm based on ordinary differential equations and regression by the lasso for non-linear transfer functions, as a part of the In-Silico-Network Challenges of DREAM3.

Data come as both time-series and steady state for three sizes of networks, number of nodes *N*=10, 50 and 100, respectively, and for each size from five networks of various topologies. There is considerably more data than is normally accessible for real systems, but otherwise the model is told to incorporate both transcription and translation as well as noise. The output from our reverse engineering algorithm is, for each possible edge in each network, a score indicating how likely the existence of the edge in the corresponding generating in-silico network is. Here we describe the algorithm in brief.

A visual inspection of the presented time-series yields that there is a considerable amount of noise in the data. Our first step is therefore to apply smoothing splines to obtain values of the derivatives, which are necessary components for our model. The basic assumption we utilize is that both the time-courses and the steady-state data can be described by a set of *N* ordinary differential equations. Since all data-points can be considered gene by gene, the network inference factorizes into *N* independent problems. Hence each equation represents the dynamics of a single gene in the network. We make the following ansatz:

$$\frac{dx_i}{dt}(t) = \sum_{j=1}^{N}\sum_{a} w_{ija} f_a[x_j(t)] - \lambda_i x_i(t), \quad i = 1,\dots,N.$$

Here $x_i(t)$ denotes the expression level of gene *i* at time *t* and w_{ija} is the net effect of gene *j* on gene *i,* mediated by the transfer-function f_a. These functions are picked from a pre-defined list with functions of “typical” behaviour. It might be beneficial to use more than one transfer-function for each gene, and therefore we index them by the letter *a*. The last term corresponds to degradation, which means that λ_i has to be non-negative. The left hand side is the effective dynamics of the mRNA concentration for gene *i*, that is, the transcription rate minus the degradation rate. The actual choice of transfer functions and coefficients (including in-degree) is performed by a linear regression combined with the lasso.

When a set of transfer functions eventually is picked, we disturb all the presented data with Gaussian noise, and repeat the algorithm outlined above (including the model selection). This process of disturbance is repeated many times, and by keeping track of the number of times a certain edge is picked (corresponding to the matrix element w_{ij} being non-zero), we obtain a score for each possible edge in the network.

Sequence-dependent Nucleosome Positioning

Ho-Ryun Chung[1], Martin Vingron[1]

[1]Max-Planck-Institute for molecular Genetics, Department of Computational Molecular Biology, Ihnestr. 73, 14195 Berlin, Germany

Eukaryotic DNA is organized by a macromolecular structure called chromatin. The basic repeating unit of chromatin is the nucleosome, which consists of two copies of each of the four core histones and DNA. The nucleosomal organization and the positions of nucleosomes have profound effects on all DNA-dependent processes. Understanding the factors that influence nucleosome positioning is therefore of general interest.

Among the many determinants of nucleosome positioning the DNA sequence has been proposed to play a major role. Here, we analyzed over 860,000 nucleosomal DNA sequences to identify sequence features that guide the formation of nucleosomes *in vivo*. We found that both a periodic enrichment of AT base pairs and an out of phase oscillating enrichment of GC base pairs as well as the overall preference for GC base pairs are determinants of nucleosome positioning.

The preference of GC pairs can be related to a lower energetic cost required for deformation of the DNA to wrap around the histones. In line with this idea, we found that only incorporation of both signal components into a sequence model for nucleosome formation results in maximal predictive performance on a genome wide scale. In this manner, one achieves higher predictive power than published approaches. Our results confirm the hypothesis that the DNA sequence plays a major role in nucleosome positioning *in vivo*.

Mapping Potential Regulatory SNPs in Human Genome

Grace Tzu-Wei Huang[1,2], Panayiotis (Takis) Benos[1,2]

[1]Joint CMU-Pitt PhD Program in Computational Biology; [2]Department of Computational Biology, School of Medicine, University of Pittsburgh

Gene expression is known to be influenced by genetic variances among different individuals. Genetic variances may partially account for expression differences, possibly through non-coding SNPs (Single Nucleotide Polymorphism) that assume regulatory roles on the nearby genes. With the availability of the high-density HapMap SNP data, as well as the efforts to profile the expression phenotypes of the individuals from whom SNP genotyping was performed, it is now possible to conduct a fine-scale analysis of the location and properties of genetic variants associated with gene expression. In this study, we investigate the extent of influence that cis-acting SNPs may have on the expression differences among individuals. We first perform a small-scale association study in the ENCODE region to determine, for a given gene, how much of the expression differences can be explained by the presence of particular SNP genotype acting *cis-* to the gene. We then constructed a non-linear decision tree predictor to test for the expression differences among genotype subgroups. The predictor takes into account the regulatory potential of a given SNP and successfully identified potential regulatory SNPs. These results could help elucidate the role of fine-scale genetic variances on expression phenotypes. The analysis is being extended to the entire genomic region to construct a detailed map of potential cis-acting non-coding SNPs in the human genome.

Modeling thyroid hormone binding sites from homologous promoters

Vidhya Jagannathan[1,2], Marc Robinson-Rechavi[1,2]

[1]Univeristy of Lausanne, CH-1015, Lausanne, Switzerland; [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Nuclear receptors are transcription factors that control developmental processes, homeostasis maintenance and medically important diseases and conditions. Moreover many hormonal actions are mediated by nuclear receptors. Yet the binding sites of most nuclear receptors remain poorly characterized, with the exception of estrogen receptor binding sites(EREs). Thyroid hormone receptors (THRα and THRβ) are a major class of nuclear receptors, whose complex patterns of binding are not described in available models. THR dimers bind to regulatory sequences composed of two half sites (AGGTCA) that are separated by variable spacing and can occur in different orientations. An additional obstacle to modeling is the absence of a good list of known binding sites. Here we start with the assumption that DNA segments under evolutionary constraint may be good candidates for functional elements.

We have constructed a flexible hidden Markov model (HMM) framework capable of predicting THR binding sites (thyroid hormone response elements-TREs). TREs matching exactly the consensus sequence, with 1 mismatch to the consensus and located within 25kb of an annotated cap site were listed for the human and mouse genomes. Using the human co-ordinates (hg18) we obtained 17-way vertebrate multiple alignments using the phastcons scores. The regions with an average score of above 0.4 were used to build a flexible TRE HMM. The model allows for variable spacing and orientation of half sites. The model was used to scan the Eukaryotic promoter database (EPD) for human promoters with high scoring TRE binding sites. The high scoring list includes promoters which were not used for training the HMM, but have already shown in the literature to be likely direct targets of THR. These include the promoters of thymosin, oligodendrocyte myelin glycoprotein, ADP-ribosylation factor 1, ATPase, H+ transporting lysosomal 38kDa V0 subunit d isoform.

We are now addressing the possibility that genes containing TRE promoters could be direct THR genes by monitoring expression databases. We are also scanning the human and mouse genome with the HMM model for response elements more loosely related to the consensus, but conserved among vertebrates, as this procedure may allow to identify other THR target genes.

Network properties of schizophrenia candidate genes in the human interactome

Jingchun Sun[1] and Zhongming Zhao[1,2]

[1] Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA

[2] Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA, USA

Schizophrenia is a heritable complex disease whose pathogenesis likely involves multiple genes interacting themselves or with environment. Though hundreds of reports declared that linkage peaks or genes are associated with schizophrenia in the last decade, its genetic mechanisms have not been well understood. In order to further understanding the pathogenesis of schizophrenia, in this study we first prioritized 161 schizophrenia candidate genes (SZGenes) from multiple evidence in linkage scans, association studies, and gene expression. We then analyzed their network properties in the whole human protein-protein interaction (PPI) networks. Our results demonstrate that SZGenes tend to have intermediate connectivity and intermediate efficiency with which a perturbation can spread throughout the network relative to essential genes and non-essential genes. We compared schizophrenia-specific subnetworks and cancer-specific subnetworks, both were extracted from the human interactome, and found that schizophrenia genes do not have a strong trend on interacting with each other or clustering compared to cancer genes. This difference suggests that schizophrenia and cancer might have different pathological mechanisms even though both have been considered as complex disease. Our results are helpful for better understanding the genetic mechanisms of schizophrenia. The extracted schizophrenia specific networks also provide an opportunity for identifying novel candidate genes for further verification.

Convex Optimization for Identification of Minimal Models

A. Agung Julius[1,2], George J. Pappas[1]

[1]GRASP Lab. School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia PA 19104, USA.

Progress in computational and systems biology in past decade has given rise to an understanding that entities involved in the regulation, signaling, and metabolism of the cell form large scale complex networks. Nevertheless, finding out the topology of such networks, in most cases, is still an open research problem. The availability of high-throughput measurement data provides an essential ingredient for solving the problem. Despite the availability of the data, identification of a very large scale biochemical network is still a hard problem. This is because (i) the data is usually expensive to obtain and therefore only available in small quantity relative to the size of the network, (ii) the data is corrupted by noise, (iii) the scale of the network is very large.

A straightforward approach in network identification is to simply 'invert' the data. While this approach is computationally simple and intuitive, it can be shown that when the data is noisy the approach leads to overfitting. In this case, any sparse structure in the real network is lost in the model. One way to overcome this drawback is to include some minimality/sparsity constraints while identifying the model. However, including minimality/sparsity constraints in the network identification procedure typically leads to NP-Hard computational problems. This kind of problems is practically impossible to solve in large scale.

In this talk, we will present some convex optimization relaxation techniques for the problem of identification of sparse/minimal models. Convex optimization is a branch of mathematical optimization theory that deals with specially structured optimization problems. Because of the special structure, certain mathematical properties of the solutions can be guaranteed, and efficient solver algorithms have been developed. Convex relaxation techniques do not provide exact solution to the NP-Hard problems. Instead, they can provide high quality approximations to the solution through much more tractable computation. The talk will be motivated by a few examples on the application of the techniques.

HHMMiR: Efficient de novo Prediction of MicroRNAs using Hierarchical Hidden Markov Models

Sabah Kadri[1], David L. Corcoran[2], Kusum Pandit[2], Naftali Kaminski[3], Veronica Hinman[4], and Panayiotis V. Benos[5]

[1]Lane Center for Computational Biology, Carnegie Mellon University, PA, USA; [2] Department of Human Genetics, University of Pittsburgh, PA, USA; [3] Allergy and Critical Care Medicine, University of Pittsburgh, PA, USA; [4] Department of Biological Sciences, Carnegie Mellon University, PA, USA; [5] Department of Computational Biology, University of Pittsburgh, PA, USA

MicroRNAs (miRNAs) are small non-coding single-stranded RNAs (20-23 nts) that are known to act as post-transcriptional regulators of their target genes. The role of miRNAs in many important biological processes, such as development, cell differentiation, and cancer has been established in recent years. In spite of their biological significance, the identification of miRNA genes in newly sequenced organisms is still based, to a large degree, on evolutionary conservation. We have developed **HHMMiR**, a novel algorithm for *de novo* miRNA hairpin prediction in the absence of evolutionary conservation. Our method implements a Hierarchical Hidden Markov Model (HHMM) that utilizes regional structural and sequence information of miRNA precursors. The structure of a typical miRNA hairpin was derived from summarized characteristics of miRNA hairpins in publicly available databases. A typical precursor is divided into four distinct regions, each with different characteristic features. The underlying HHMM provides an intuitive model for these regions. Our algorithm achieved average sensitivity of 84% and specificity of 88%, on 10-fold cross-validation of human datasets. We also showed that, the model trained on human sequences performs very well in predicting hairpins from other vertebrate, invertebrate and plant species. In particular, the human-trained model was able to correctly classify ~97% of known plant miRNA precursors. Preliminary results showed that three of five newly predicted mouse miRNA genes are expressed in lung cells. These are very encouraging results showing that HHMMiR may be very useful in predicting miRNA genes across long evolutionary distances without the requirement for evolutionary conservation of sequences. This is the first time HHMMs have been used to predict miRNAs. The importance of HHMMiR is expected to be high in predicting novel miRNA genes in any organism, including organisms with no closely related sequenced species (e.g., sea urchin, platypus, etc). Further developments of the HHMMiR algorithm include the extension of the model to predict the miRNA genes in the pre-miRNA stem-loops.

A genomic approach to map cell cycle transcription pathways in Saccharomyces cerevisiae

Pinay Kainth[1-3], Holly Sassi[2,3], Lourdes Pena-Castillo[2,3], Jeffrey Fillingham[2,3], Alex Kostic[1], Harm van Bakel[2,3], Kyle Tsui[2,4], Corey Nislow[1-3], Jack Greenblatt[1-3], Brendan Frey[5], Timothy Hughes[1-3] and Brenda Andrews[1-3]

[1]Department of Molecular Genetics; [2]Terrence Donnelly Centre for Cellular & Biomolecular Research; [3]Banting & Best Department of Medical Research; [4]Department of Pharmacy [5]Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada.

We are using tools of yeast functional genomics to systematically discover new regulators controlling gene expression patterns in *Saccharomyces cerevisiae.* To identify such proteins, we combined a promoter-reporter system with our capacity for high-throughput genetics to determine the effect of gene deletion on a promoter of interest. Our general strategy utilizes the synthetic genetic array (SGA) platform to systematically introduce promoter-GFP reporter constructs along with a control promoter-RFP gene into the array of ~5000 yeast deletion mutants. Fluorescence intensities from each reporter are assayed from individual colonies arrayed on solid agar plates using a scanning fluorimager and the ratio of GFP to RFP intensity reveals deletion mutants that cause differential GFP expression. We expect deletion of putative activators to result in decreased GFP intensity while deletion of putative repressors to result in increased GFP expression.

We are using this approach to systematically delineate pathways of cell cycle transcription. The cell cycle of all eukaryotes is hallmarked by successive waves of gene expression. In budding yeast, ~15-20% of all genes show this pattern of expression but a complete map of pathways that impinge upon the regulation of these genes remains unknown. We generated a panel of 27 cell cycle regulated promoter-GFP constructs that represent all cell cycle phases and used our reporter system to discover new regulators of these genes. By doing so, we created a data matrix of quantitative gene expression measurements representing 27 reporter genes by ~5000 yeast deletion mutants. We are using computational approaches to sort this information into biologically relevant modules where particular deletion mutants explain the regulation of groups of cell cycle promoters we screened.

In our screening effort, we identified both known and novel candidate regulators of cell cycle genes and are using real-time PCR and other assays to further understand the regulatory roles of these genes. For instance, we screened the promoter that normally drives transcription of histone H2A (*HTA1*) to identify the regulators responsible for its S-phase specific expression. We discovered that Rtt109, which is a histone acetyltransferase that acetylates H3 lysine 56, is required for proper activation of *HTA1* gene expression. Our screen also revealed that the histone chaperone, Rtt106, is required for proper repression of *HTA1* gene expression. We have shown by chromatin immunoprecipitation that Rtt106 localizes to a negative regulatory region of the histone promoter and that its presence there is dependent on the previously characterized repressors, Hir1 and Asf1. These results highlight the utility of our approach to pinpoint the regulators that control a promoter of interest.

We hope to exploit this genome-wide screening approach to construct a detailed map describing the interplay of regulators controlling the eukaryotic cell cycle and are also applying this strategy more generally to understand other transcription factor pathways.

Detecting the Presence and Absence of Causal Relationships Between Yeast Gene Expressions with Very Few Samples

Eun Yong Kang[1], Hyun Min Kang[2], Ilya Shpitser[4], Chun Ye[3], Eleazar Eskin[1]

[1]Department of Computer Science, University of California Los Angeles; Department of Computer Science and Engineering and [3]Bioinformatics Program, University of California San Diego; [4]Department of Epidemiology, Harvard School of Public Health.

The availability of high throughput microarray data has encouraged several genetical genomics studies where researchers have identified genomic locations correlated with expression quantitative traits in the form of linkages or associations. Various statistical approaches have been applied to these results to tease out the underlying biological networks that govern how genes regulate and interact with each other. Extracting causal relationships from these networks is a challenging but important step to understanding the complex diseases and phenotypes under perturbation.

In this paper, we discover the presence and absence of causal relationships between gene expressions in yeast. Causal discovery is challenging in our case because our domain has many thousands of variables, while our number of samples is very limited. In particular, most conventional conditional independence tests are not reliable in the small sample case, since conditioning severely reduces the power of the test, and as a result we cannot infer independence with high confidence.

Our approach is to rely on basic properties of graphical models along with prior knowledge of biology to infer or exclude directionality of edges based on results of marginal independence tests, and limited model selection. To perform marginal independence tests we calculate the likelihood ratio statistics, use a permutation test to obtain null statistics, and obtain our p-values from these statistics. We handle the multiple testing problem in the standard way by using False Discovery Rate(FDR). Finally, we settle edge orientation in ``promising'' three node regions of the graph with a certain pattern of marginal (in)dependence among the three nodes by performing model selection using a likelihood ratio test. While our data has so few samples that no method can recover the complete causal graph, our method is able to identify small causal subgraphs which have particularly strong signals. Nevertheless, the predictions we do make are surprisingly consistent with previous experimentally validated knowledge of yeast regulation.

We applied our method to an expression dataset of 5534 genes and a genotyping dataset of 2956 SNPs collected over 112 genetic segregants of yeast. After applying our method, we found 24620 (SNP, expression) pairs where the SNP is causal to the expression at a FDR of $q < 0.05$ and found 1148 causal relationships consisting of 108 causal regulator genes and 809 affected target genes. The list of 108 causal regulators we found includes many genes identified in previous analyses[1], and is remarkably consistent with experimental findings[2].For example, for CNS1, we found that its affected genes are enriched for the ribosome biogenesis pathway similar to previous results. However, we also found that genes unaffected by CNS1 are enriched for the metabolism pathway.

[1] Lin Chen, Frank Emmert-Streib, and John Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biology, 8:R219, 2007.
[2] J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. Bumgarner, and E. E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.

Identification of regulatory motifs and their targets using comparative genomics

Pouya Kheradpour[1], Alexander Stark[1,2], Manolis Kellis[1,2]

[1] Computer Science and Artificial Intelligence Laboratory of MIT; [2]Broad Institute of MIT and Harvard.

The short length and many weakly-specified positions of regulators make the chance matching of their motifs to a single genome very frequent. Comparative genomics provides a powerful methodology for increasing discovery power in order to identify the functional binding sites with both high sensitivity and high specificity.

We have developed a statistical measure that leverages many genomes in order to assign a confidence score to individual motif instances for both transcriptional regulators and microRNAs. The method uses a phylogenetic measure of the overall neutral branch length over which a motif instance is conserved, and compares it to the expected conservation for random motifs of similar properties in order to obtain a confidence value (1 - false discovery rate) for every motif instance.

We have used ChIP-chip and ChIP-Seq experiments for several transcription factors to evaluate our predicted binding sites and have found that in both flies and mammals the instances our method selects as having high confidence strongly correlate with those bound in vivo. For example, at 50% confidence our predicted CTCF instances are more than 10-fold enriched in ChIP-Seq CTCF human bound regions. We find similar enrichments for several factors including Stat1, HNF1A, HNF4A, HNF6, ERalpha, p53, RELA, and Myc in the human, and for Twist, Snail, Dorsal and CrebA in the fly.

We have also used the confidence score to discover novel regulatory motifs in the fly and human genomes, both genome-wide, as well as associated with regions bound by in ChIP-chip and ChIP-Seq experiments. This allows us to discover motifs associated with transcription factor binding and changing chromatin states, thus suggesting candidate regulators for enhancer and promoters regions of interest.

Conservation of Transcriptional Autoregulatory Loops in Vertebrate Evolution

Szymon M. Kiełbasa[1], Martin Vingron[1]

[1]Max Planck Institute for Molecular Genetics, Ihnestr. 73, D-14195 Berlin, Germany

An autoregulatory feedback loop is an elementary motif observed in transcriptional regulatory networks. The cells utilize direct autoregulation in many fundamental processes, for example to guarantee fast transcriptional responses or to provide multiple stable states of gene expression. Therefore, the conservation of autoregulatory feedback loops in the course of evolution is of prime interest.

We study the enrichment of the number of autoregulatory feedback loops among higher organisms [1]. First, the number of loops constructed out of predicted transcription factor binding sites is counted. Next, we compare the result to an estimate obtained by assuming that each (conserved) gene has the same chance to be a target of a given transcription factor or by assuming that each conserved promoter position has an equal chance to become a binding site of the transcription factor.

Our analysis shows that the numbers of putative autoregulatory feedback loops conserved between human and fugu, danio or chicken are significantly higher than the estimations. Additionally, we demonstrate that the conserved autoregulatory binding sites tend to concentrate around genomic locations corresponding to the transcription start sites of the respective transcription factors. We conclude, that transcriptional autoregulatory feedback loops constitute a core transcriptional network motif and their conservation has been maintained in evolution of higher vertebrate organisms.

[1] Kiełbasa SM, Vingron M (2008) Transcriptional Autoregulatory Loops Are Highly Conserved in Vertebrate Evolution. PLoS ONE 3(9): e3210. doi:10.1371/journal.pone.0003210

Using SNPs for Measuring MicroRNA-mediated mRNA Destabilization

Jinkuk Kim[1,2], David P. Bartel[2,3]

[1]Harvard-MIT Division of Health Sciences and Technology, E25-519, 77 Massachusetts Avenue, Cambridge, MA 02139, USA; [2]Howard Hughes Medical Institute and Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA; [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

MicroRNAs (miRNAs) are ~23 nt endogenous RNAs that interact with mRNAs and repress the expression of the mRNAs by destabilization or translational inhibition. Crucial for target recognition is the pairing between a 7-8mer target site of the mRNA and the seed region of the miRNA. We examined heterozygous SNPs that create or destroy the target sites. Such a SNP will allow one allele of a gene to have a target site but the other allele not, and may therefore contribute to allelic imbalance at mRNA level. Using pyrosequencing, we measured the allele-specific expression of 67 such SNPs of mice. We observed that, for a significantly large fraction of the SNPs, target alleles are indeed downregulated compared to the matching non-target alleles, and the efficacy of target sites is dependent on site type and context features. Our study not only confirmed the validity of seed-based target model and the utility of site type and context as predictors of target site efficacy, but also demonstrated an approach for measuring miRNA-mediated mRNA destabilization in vivo. In addition, we noticed a potential of the approach to be used for testing the in vivo functionality of miRNA target sites or other cis-elements that may influence mRNA level.

Modeling exopeptidase activity from LC-MS data

Bogusław Kluge[1], Anna Gambin[1], Wojciech Niemiro[2]

[1] Institute of Informatics, University of Warsaw, Warsaw, Poland, {bogklug,aniag}@mimuw.edu.pl; [2] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland, wniemiro@gmail.com

Recent studies demonstrate that the peptides in the serum of cancer patients that are generated (ex-vivo) as a result of tumor protease activity can be used for the detection and classification of cancer. In this paper we propose the first formal approach to modeling exopeptidase activity from liquid chromatography mass spectrometry samples. We design a statistical model of peptidome degradation and a Metropolis-Hastings algorithm for Bayesian inference of model parameters. The model is successfully validated on a real LC-MS dataset. Our findings support the hypotheses about disease-specific exopeptidase activity, which can lead to new diagnostic approach in clinical proteomics.

Genome-wide identification of distal 5’ and 3’ transcriptional start and stop sites using exon-tiling microarrays

Tony Kwan[1], Daniel Gaffney[1], Christel Dias[1], Tomi Pastinen[1], Jacek Majewski[1]

[1]Department of Human Genetics, McGill University and Genome Québec Innovation Centre, Montréal, Canada.

Regulatory regions within the genome harbor specific sequences that control gene expression, and are often located within the 5’ and 3’ untranslated regions. Recent studies have identified novel, distally located exons that extend the annotated 5’ transcription start sites (TSS) of most ENCODE genes and that contain potential regulatory motifs for the control of gene expression. Here, we use the Affymetrix Exon Array to derive expression information from the HapMap CEU panel and identify potential exons extending the gene past the annotated 5’ and 3’ boundaries, and reveal possible new regulatory regions. Our strategy involves identifying genes showing significant cis-association to a SNP, and then looking for expressed flanking exons that also show concordant association to the same SNP, and with the same direction of effect for the over-expressed allele. All significant gene-SNP associations (P<1e-4) from the Exon Array, as well as significant cis-eQTLs from the Sanger CEU LCL dataset were selected. Subsequently, we performed an association analysis of the same SNP to all expressed non-refSeq (non-core) probesets within 250kb flanking either side of the gene, with significance cutoff at P<0.05. We identified 266 genes with expressed, concordant probesets (1039 probesets total) flanking the 5’ and/or 3’ ends. There were 179 and 189 transcripts with extended 5’ and 3’ ends, respectively, as well as some transcripts extending at both ends. Consistent with recent data, ~20% of the exons are located in the adjacent intergenic regions. Another 65% of probesets are located within flanking genes, suggesting either transcript extensions or co-regulated overlapping transcripts. The overlap of these probesets with the Affymetrix Transcriptome Phase 3 data indicates that >60% of these probesets are expressed and correspond to an mRNA transcript, further supporting these non-core probesets as ‘real’ exons. We observed ~30% of the distal 5’ probesets overlapping with a genome-wide dataset of DNAse I hypersensitive sites, which are typically found near the TSS and is a hallmark of transcriptional activity. The concordant probesets also show enrichment above expectations with overlapping regulatory regions, such as those identified from ChIP-chip data, cis-regulatory modules (PreMod), and the 7x Regulatory Potential sites, suggesting that these probesets contain sequences that are regulatory in nature. Although this strategy is limited to transcripts showing genetic associations, this study nonetheless shows the value of leveraging expression microarray data for the identification of novel TSS and potentially new sites of regulatory action.

Global Network Analysis of Protein and Gene List Sets

Alexander Lachmann[1], Avi Ma'ayan[1]

[1]*Mount Sinai School of Medicine, New York, NY*

Biological experiments such as gene expression microarrays and mass-spectrometry proteomics produce results in a form of lists of proteins or genes. To be able to comprehend the most information that is contained within those lists a multitude of algorithms and tools were developed in the past. For example, Gene Ontology enrichment analysis is used to find common functions enriched within the proteins or genes from a list. Another example is protein-protein interaction databases used to construct graphs that visualize a subnetwork connecting proteins from a list. There are numerous other types of background knowledge datasets that can be projected onto proteins/gene lists. Most commonly existing tools only look at a single type of background knowledge examining only one list. We are implementing a solution that combines several of the current resources and approaches to analyze protein/gene lists. Our aim is to provide a utility for a broader insight into the function and attributes of selected protein lists by comparing them. To achieve this, we designed a framework that provides users with the ability to input a number of protein/gene lists. From those lists we generate additional lists, such as the intersection, union, and complementation between pairs of lists. Additional related proteins/genes identified based on protein-protein interactions, or other methods, can also be added to the lists. We then analyze each list with Gene Ontology using Ontologizer, Protein-Protein Interaction using Genes2Networks , structural domains using data from Protosite, phoshoporylation targets and kinases responsible for the phosphorylation using NetworKIN, protein/DNA interactions using ARACNE, and disease-gene association using data from OMIM . Each analysis tool with its reference data is wrapped as a predefined framework object, so that new components can be plugged into the system without the need of changing it. Each framework object contains the following APIs: FuzzyCluster, AnalyzeEnrichment, FindAndRankIntermediats, and Visualize. The FuzzyCluser API clusters the genes/proteins within each list creating sublists, each sublist is fed into a FuzzyClusterIntegrator module. The Visualize API formats output web-pages with images of the results from the FuzzyCluster, AnalyzeEnrichment, and FuzzyClusterIntegrator modules (Fig. 1). To demonstrate the capabilities of this global network analysis system we applied it to analyze lists of proteins from Mass-Spectrometry data reporting lists of proteins pulled-down with bait proteins in a beta-arrestin-GPCR pathway.

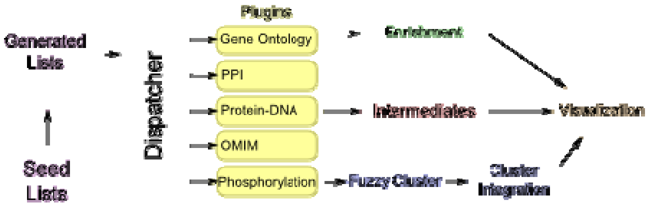


Figure 1: Global network analysis framework for protein/gene-list attributes comparison

Posters

Inferring transcription factor targets from multiple heterogeneous data sources

Xiaofeng Dai[1] and Harri Lähdesmäki[1]

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland

Transcriptional regulation is a central control mechanism for many biological processes, such as development and cell cycle. Transcriptional regulation is largely controlled by transcription factors (TF) that bind gene promoters in a sequence specific manner. Thus, revealing genome-wide protein-DNA interactions is one of the key problems in understanding transcriptional regulation at mechanistic level.

Computational transcription factor binding site (TFBS) predictions rely on sequence specificities that are taken from a database (JASPAR, TRANSFAC), obtained as an output from a motif discovery method or, more recently, measured using high-throughput techniques. Sequence specificities alone, however, are not sufficiently informative to accurately predict TF targets. A natural way to improve TF target predictions is to incorporate additional data into statistical inference of TFBSs. We have recently developed a probabilistic TFBS prediction method that is able to make use of practically any additional genome-level information source [1]. Statistical data fusion becomes more challenging when several information sources need to be combined in a meaningful way.

Here we extend our previously published method [1] by incorporating novel data sources into TFBS prediction and by developing a new method for multiple data fusion. In particular, we use evolutionary conservation, nucleosome positioning data from a recently published method, regulatory potential and DNA duplex stability to improve TFBS predictions. These data sources are informative of binding sites because functional binding sites are typically conserved and free of stable nucleosomes, regulatory DNA sites have different characteristics than neutral sites, and different TFs can bind DNA in a single or double strand manner, Some of these individual data sources have already been shown to improve *de novo* motif discovery, but we demonstrate how these multiple data sources can be combined to make joint statistical inference of TF targets. Integration of those data sources that have a probabilistic interpretation is relatively straightforward [1]. For other cases, we convert the raw data into probabilities, or priors, by applying a previously proposed Bayesian transformation method. In addition, for efficient use of DNA duplex stability data, we develop a simple heuristic that can assess the binding preference (single or double stranded DNA) for a TF from a set of known binding sites.

Results on a carefully constructed test set of verified binding sites in the mouse genome (ABS, ORegAnno) demonstrate that principled data fusion can significantly improve the performance of TF target prediction methods. Our statistical data fusion method can gain valuable new insights into genome-wide models of transcriptional regulatory networks.

[1] Lähdesmäki H, Rust AG and Shmulevich I. (2008) Probabilistic inference of transcription factor binding from multiple data sources. *PLoS ONE*, Vol. 3, No. 3, e1820.

New insights into gene and protein regulatory networks during the initial stages of human T helper 2 differentiation through LC-MS/MS and stable isotope labeling.

Robert Moulder[1], Tapio Lönnberg[1.2],Jan-Jonas Filén[1,2], Sanna Filén[1,3], Riitta Lahesmaa[1]

[1]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University; [2]The National Graduate School in Informational and Structural Biology, Finland.; [3]The Drug Discovery Graduate School, Finland

T cells play a key role in orchestrating the immune response and recognizing the foreign structures e.g. invading pathogens or cancer from the healthy cells and tissues of the host. Lymphocytes can differentiate into functionally distinct subsets that play important roles in host immune responses against pathogens as well as several autoimmune and inflammatory diseases. Dissecting the molecular mechanisms of T helper cell (Th) differentiation to functional effector cells is hence important for understanding the pathogenesis of immune mediated diseases. While Th cell differentiation has been extensively studied in the mouse system very little is known about the molecular basis of Th differentiation in human. We have exploited genome wide transcriptomics and proteomics approaches to create a holistic view on the human Th differentiation. The data provides an excellent starting point for generating novel hypotheses, to be experimentally tested, on the genes and pathways involved in the process.

To identify novel factors directing the early human Th2 cell differentiation we have studied the detailed kinetics of genes regulated during the process. Besides a detailed transcriptomics profiling [1] we have used LC-MS/MS with stable isotope labeling technology to investigate proteomic changes in the microsomal fraction [2] and the nucleus. Using the cysteine specific cICAT labeling to identify IL-4 regulated proteins from the microsomal fractions of human naive CD4+ cells 557 proteins were identified, of which 304 were also quantified. Combined with the genome wide transcriptomics analysis our results indicated that the entire GIMAP family of proteins is differentially regulated during the early Th cell differentiation.

To study further the quantitative changes in nuclear proteins during the early stages of human Th2 differentiation we used 4-plex iTRAQ regents. We determined changes in the proteomic profiles of the nuclear fraction of activated naïve human CD4+ cells in association with IL-4 stimulation at the time points of 6 and 24 h. Altogether 834 proteins were detected and quantified, and a number of statistically significant changes were determined in nuclear localization/expression with several proteins, both known and novel in this context. These results suggest a role for the nuclear localization of the nuclease sensitive element-binding protein 1 (YB-1) in the resistance to activation-induced cell death characteristic of Th2 cells. We have established a siRNA based screen to further characterize the functional role of the newly identified candidates in the Th2 differentiation process. The newly discovered candidates with the pathways they participate in are potential targets for developing therapeutics to modulate human Th2 cell responses.

[1] Lund R et al. J. Immunol, 2007:15;178(6):3648-60.

[2] Filén J-J & Filén S et al. Molecular Cellular Proteomics, 2008 Aug 12. [Epub ahead of print].

MODELLING AND DOCKING STUDIES OF AMYLOID PRECURSOR PROTEIN IN ALZHEIMER’S DISEASE

Lalith krishna*, Hemaswaroop*, Febin Prabhu Dass*

*School of Biotechnology, Chemical and Biomedical Engineering

Vellore Institute Of Technology University, Vellore-632014, India.

ABSTRACT: Alzheimer’s disease is a fatal Neuro-degenerative disorder where neuronal cell die at an accelerated rate. This disease has characteristic pathological findings of senile plaques and neurofibrillary tangles along with less production of neurotransmitter acetylcholine. Currently available medication aims at affecting the acetylcholine cycle there by preventing the breakdown of acetylcholine. This offer relatively small symptomatic benefit for some patients but do not slow disease progression. A new approach utilizing the genomic and chemo informatics tools such as Modeling software– INSIGHT II, docking programs – AUTODOCK, ARGUSLAB and 3D- Pharmacophore designing server are used to identify a better lead compound which can possibly slowdown the progression of neuronal death by targeting plaque proteins. Beta-amyloid peptides (forming the senile plaques) are formed by action of beta-secretase and gamma-secretase enzyme on amyloid precursor protein (APP). The lead compound designed in such a way it specifically binds to the active site of APP where beta-secretase binds. This inhibits the binding of beta secretase to the active site of APP thereby preventing the cleavage of APP to beta-amyloid peptide, eventually preventing the formation of plaque. The lead molecule determined by this approach is intended to overcome the shortcomings of the existing drugs against Alzheimer’s disease progress.

KEYWORDS: Alzheimer, Amyloid plaques, Beta-secretase, Amyloid precursor protein (APP),INSIGHT II, AUTODOCK, 3D Pharmacophore.

Geometric Interpretation of Gene Co-Expression Network Analysis

Steve Horvath[1,2], Peter Langfelder, Jun Dong[1]

[1]Human Genetics, Medical School, University of California, Los Angeles; [2]Biostatistics, School of Public Health UCLA

Similar to natural languages, network language is ever evolving. While some network terms (concepts) are widely used in gene coexpression network analysis, others still need to be developed to meet the ever increasing demand for describing the system of gene transcripts. There is a need to provide an intuitive geometric explanation of network concepts and to study their relationships. For example, we show that certain seemingly disparate network concepts turn out to be synonyms in the context of coexpression modules.

We take advantage of the relationship between network theory and the field of microarray data analysis to clarify the meaning of and the relationship among network concepts in gene coexpression networks [1]. Network theory offers a wealth of intuitive concepts for describing the pairwise relationships among genes, which are depicted in cluster trees and heat maps. Conversely, microarray data analysis techniques (singular value decomposition, tests of differential expression) can also be used to address difficult problems in network theory. We describe conditions when a close relationship exists between network analysis and microarray data analysis techniques, and provide a rough dictionary for translating between the two fields. Using the angular interpretation of correlations, we provide a geometric interpretation of network theoretic concepts and derive unexpected relationships among them. We use the singular value decomposition of module expression data to characterize approximately factorizable gene coexpression networks, i.e., adjacency matrices that factor into node specific contributions. High and low level views of coexpression networks allow us to study the relationships among modules and among module genes, respectively. We characterize coexpression networks where hub genes are significant with respect to a microarray sample trait and show that the network concept of intramodular connectivity can be interpreted as a fuzzy measure of module membership. We illustrate our results using human, mouse, and yeast microarray gene expression data. The unification of coexpression network methods with traditional data mining methods can inform the application and development of systems biologic methods.

[1] Horvath S, Dong J (2008) Geometric Interpretation of Gene Coexpression Network Analysis. PLoS Comput Biol 4(8): e1000117

WGCNA: R package for Weighted Gene Co-expression Networks

Peter Langfelder[1], Steve Horvath[1,2]

[1]*Dept. Of Human Genetics;* [2]*Dept. Of Biostatistics, University of California, Los Angeles, CA 90095*

Weighted gene co-expression network analysis (WGCNA) is a systems biologic method for analyzing microarray data, gene information data, and microarray sample traits (e.g., case-control status or clinical outcomes). WGCNA can be used for constructing a weighted gene co-expression network, for finding co-expression modules, and for calculating module membership measures (related to intra-modular connectivity). WGCNA facilitates a network based gene screening method that can be used to identify candidate biomarkers or therapeutic targets. These methods have been successfully applied in various biological contexts, e.g., cancer, mouse genetics, yeast cell cycle, and complex disease gene identification.

We present the WGCNA R software package, a comprehensive collection of R functions for performing a weighted gene co-expression network analysis of microarray data. It includes functions for network construction, module detection, gene screening, calculations of topological properties, data simulation, and visualization. Along with the R package we also present an R software tutorial that illustrates important analysis tasks addressed by the package including network construction, module detection and different gene selection methods.

As an illustration of the utility of the WGCNA package, we present examples of network analyses of gene expression data. The first example studies the preservation of eigengene networks across different human brain areas. Eigengene networks represent relationships among entire gene co-expression modules and can be thought of as a bridge between individual genes and a systems-level view of an organism. Our results suggest that gene co-expression modules and their relationships are highly robust and preserved between different data sets.

In the second example, we study gene co-expression networks in an F2 mouse intercross. Building networks from over 23000 surveyed micorarray probes in adipose and liver tissues, we identify gene modules that are related to physiological traits and study their inter-module relationships. We illustrate incorporation of genetic data into the analysis that allows the study of causal relationships among co-expression modules and clinical traits.

The R package along with a tutorial is freely available and can be dowloaded from http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA.

Acknowledgments: Preservation of eigengene networks across brain areas was studied in collaboration with Michael Oldham and Daniel Geschwid; mouse data were analyzed in collaboration with Margarete Mehrabian and Aldons J. Lusis.

Seeking higher-order chromatin domains in humans via hidden Markov models

Jessica Larson[1], Guocheng Yuan [1,2]

[1] Department of Biostatistics, Harvard School of Public Health; [2] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute

Eukaryotic chromatin is not uniformly packaged but organized into higher-order regulatory domains. There is evidence that neighboring genes, although not involved in the same pathways, are still similarly regulated at the level of transcription via various histone modifications. We seek to develop a genome wide map of these higher-order (i.e., multi-gene) domains through a novel analysis of ChIP-seq histone modification data in human CD4+ T cells. As a starting point, we have described 30 genes on chromosome 2 with five modifications (H3K4me2, H3K4me3, H4K12ac, H3K27me3, H3K9me3). To determine each domain we first collapsed our five sets of ChIP-seq data into a more general characterization using hierarchical clustering and heat-map visualization. Then, with hidden Markov models and corresponding algorithms, we were able to determine the most probable domain status of each gene. There is evidence that our 30 genes can be described as three neighboring higher-order domains. To verify our findings, we found that our domain boundaries are in agreement with nuclear lamina associated domains (LADs) as described by Guelen, *et al*. (2008). We also noted less gene expression variability within each of our domains when compared to randomly selected boundaries (p-value=0.059). We thus have evidence of multi-gene domains in our region, which are characterized by similar patterns in five histone modifications. As we expand our region and modifications, we will provide important insight into the general structure, organization, and regulation of the human genome.

[1] Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, and B van Steensel. 2008. "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." *Nature*. 453(7197): 948-51.

A plugin to connect Cytoscape to local relational databases

Kris Laukens[1], Jens Hollunder[2], Thanh Hai Dang[1], Alain Verschoren[1]

[1]*Intelligent Systems Laboratory, University of Antwerp, Belgium,*

[2]*Department of Plant Systems Biology, VIB, Gent, Belgium*

Network visualization software such as Cytoscape is widely used to study molecular interaction networks. The relevant features that need to be visualized are often coming from a number of different sources such as public databases, local experiments or computational tools. It is common practice to collect relevant information from selected model organisms into local relational databases. Such a local database may serve as a common repository for all researchers in a group are working with the same organism. It unifies data format conversion efforts, avoids duplicate work and its data can be retrieved using a simple common query syntax.

Thus far, Cytoscape database connectivity is offered by a number of plugins. Most of these plugins are either dedicated to a certain public database or subset of databases, or to a specific data type. No general solution exists for connecting Cytoscape to custom-built local databases. We therefore developed a new plugin that provides a generic approach to establish a connection between Cytoscape and any relational database. The Cytoscape user can launch SQL ('Structured Query Language') queries from within Cytoscape, with the option to use existing Cytoscape node or edge features as SQL arguments. The plugin then converts the retrieved data to Cytoscape network components according to user-defined mappings, thereby allowing the user to specify how each column of the result table should be imported into Cytoscape (e.g. node versus edge attributes).

Using real data we demonstrate how this plugin allows to answer complex research questions, even if the data is spread over multiple data sources, directly from within Cytoscape. It can for example be used to enrich existing networks with potentially interesting extra node or edge attributes (e.g. expression data, mass spectrum features, biological functions, ...), using attributes from existing nodes or edges as query criteria. The tool also supports expansion of an existing (sub)network with other interactions (e.g. from known interactions).

This plugin augments Cytoscape's functionality with generic database connectivity. The ability to import data from any user accessible database simplifies otherwise extensive processing tasks for network generation, network analysis, and network attribute manipulation, to one or more simple SQL statements. For example, looking up whether interactions equivalent to an open Cytoscape network exist in a database becomes now very straightforward.

Overall the extreme flexibility of this tool makes it it particularly useful for the integration of local experimental data into rich molecular interaction networks. Since it rapidly connects to virtually any database it is optimally suited to integrate data from a diverse range of sources.

The implications of human metabolic network topology for disease comorbidity

Deok-Sun Lee[1,2], Juyong Park[1,2], Krin A. Kay[3], Nicholas A. Christakis[4], Zoltan N. Oltvai[3], Albert-László Barabási[1,2]

[1]Center for Complex Network Research and Department of Physics, Biology, and Computer Science, Northeastern University, Boston, MA 02115; [2]Center for Cancer Systems Biology, Dana–Farber Cancer Institute, Boston, MA 02115; [3]Department of Pathology, University of Pittsburgh, Pittsburgh, PA 15261; [4]Department of Health Care Policy, Harvard Medical School, Boston, MA 02115

Given that most diseases are the result of the breakdown of some cellular processes, a key aim of modern medicine is to establish the relationship between disease phenotypes and the various disruptions in the underlying cellular networks. Here we show that our current understanding of the structure of the human metabolic network can provide insight into potential relationships among often distinct disease phenotypes. Here we constructed a human disease network based on metabolism, in which nodes are diseases and two diseases are linked if mutated enzymes associated with them catalyze adjacent metabolic reactions. Adjacent metabolic reactions have their flux rates highly correlated, as shown by the co-expression of corresponding enzyme-encoding genes, and connected disease pairs display higher comorbidity than those that have no metabolic link between them. Furthermore, the more connected diseases a disease has, the higher is its prevalence and associated mortality rate. The network topology-based approach also helps to uncover potential mechanisms that contribute to their shared pathophysiology. Thus, the structure and modeled function of the human metabolic network can provide insights into disease comorbidity, with potentially important consequences for disease diagnosis and prevention[1].

[1] This presentation is based on D.-S. Lee *et al.*, Proc. Nat'l. Acad. Sci. USA **105**, 9880 (2008).

Testing the Limits of Bacterial Robustness

Andrew Krueger[2], Joseph Lehár[1-2]

[1]CombinatoRx Inc; [2]Boston University BME;

Biological systems are robust, in that they can maintain stable phenotypes under varying conditions or attacks, and are also complex, being organized into many functional modules that communicate via interlocking pathways and feedbacks. In these systems, robustness and complexity are linked because both qualities arise from the same underlying mechanisms. When perturbed by multiple attacks, such complex systems become fragile in both theoretical and experimental studies, and this fragility depends on the number of agents applied (Fig. 1). We discuss how this relationship can be used to study the functional robustness of a biological system using both simulations and systematic high-order combination experiments.

Simulated target perturbations in genomic-scale models of *Escherichia coli* metabolism and systematic experiments on *E. coli* cultures with high-order combinations of targeted chemical probes demonstrate that fragility can be achieved at accessibly low orders of combination. Detailed analyses of the simulations and combination experiments show that topologically complex subsystems can tolerate more simultaneous perturbations than can simpler subsystems. We also explore the graph-theoretic properties of different target sets and how they relate to the observed high order combination responses.

High order combinations represent a promising approach towards many biomedical and bioengineering challenges. Systematic high-order experiments could determine the point of fragility for pathogenic bacteria and help identify optimal treatments against multi-drug resistance. This contributes to the growing realization that precise control of biological systems may best be achieved by targeting not single proteins, but well-chosen sets of nodes.

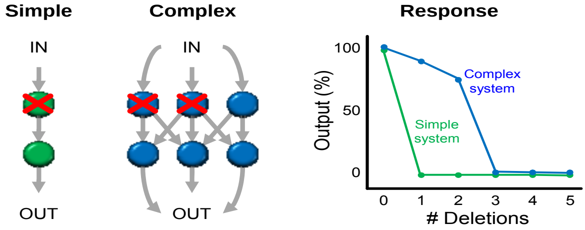


Figure 1: How complex systems are less fragile to high-order perturbations. Biological processes modeled as networks of interacting components generate output responses based on input conditions. For a trivially simple system with a single pathway, removing a single node will eliminate all output, so applying additional deletions will yield no synergistic responses. However, networks that have more redundancy will resist such attacks until all available alternative pathways have been blocked. Figure from Lehár *et al.* 2008, *Mol Sys Biol* **4**:215.

Quantum Physical Representation of Biological Systems

Azi Lipshtat[1], Ravi Iyengar[1]

[1]*Mount Sinai School of Medicine*

Quantum physics is a generalization of classical mechanics which applies to atomic and subatomic scales. Its mathematical formulation provides a framework for analysis of systems composed of discrete “particles”. This representation can be used to describe subcellular systems, in which cellular component present in low copy number and discreteness plays a role.

To demonstrate this approach, we present the biological equivalents of boson and fermion gases. We show that polymerization process can be compared to boson gas. Using stochastic calculations we analyze the biological system to show that the length distribution could be predicted simply by using Bose-Einstein distribution. Similarly, regulation of GTPase activation is analogous to Fermi gas. Here again, direct calculation yields the expected result of Fermi-Dirac function.

Predicting genetic interactions based on patterns of genetic and physical interactome networks

Kesheng Liu[1], Sira Sriswasdi[1], Thomas Martinez[1], Pengyu Hong[2], and Hui Ge[1]

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142;
[2]Department of Computer Science, Brandeis University, Waltham, MA 02139.

Identification of genetic interaction is an important approach to probe gene functions and to understand structures of signaling pathways. The amount of currently known genetic interactions is very limited for metazoans. Exhaustive experimentation is expensive and time-consuming, given the enormous combinations of genes in a genome. Currently, physical interactome networks are available for human and a number of model organisms, thanks to the systematic mapping of physical interactions between proteins. We develop a computational approach based on network motifs to predict novel genetic interactions from combined physical and genetic interactome networks for C. elegans. Our approach achieves a high specificity and a reasonable sensitivity. We predict potential interactors for components of a TGF-beta signaling pathway in C. elegans. By systematic RNAi interference (RNAi) perturbation and image analysis, we confirm 20 genes that genetically interact with components of the TGF-β signaling pathway. Taken together, our approach greatly facilitates the identification of genetic interactions and knowledge expansion of biological pathways based on interactome networks.

Maximum Entropy Reconstructions of Phosphotyrosine Signaling Networks from Quantitative Proteomic Data

Jason W. Locasale, Alejandro Wolf-Yadlin, Lewis C. Cantley

Harvard Medical School Department of Systems Biology/BIDMC

Advances in mass spectrometry based phosphoproteomics have allowed for quantitative, reproducible, measurements of levels of tyrosine phosphorylation as signals propagate through complex networks in response to external stimuli under different conditions. However, computational approaches to infer elements of the signaling network strictly from the quantitative aspects of mass spectrometry data are not well established. We considered a method using the principle of maximum entropy to infer a network of interacting phosphotyrosine sites from pairwise correlations in a mass spectrometry data set. We first investigated the applicability of this approach by using a simulation of a model biochemical signaling network whose dynamics are governed by a large set of coupled differential equations. We found that in a simulated signaling system, the method detects interactions with significant accuracy. We then analyzed a growth factor signaling network in a human mammary epithelial cell line that we inferred from mass spectrometry data and observe a biologically interpretable, small-world structure of signaling nodes, as well as a catalog of predictions regarding the interactions among previously uncharacterized phosphotyrosine sites. Our findings suggest that maximum entropy network models are an important tool for interpreting quantitative phosphoproteomics data.

Calculating Local Optima in the Turner energy model for RNA secondary structure

Andy Lorenz[1], Peter Clote[1],

[1] *Boston College.*

An RNA secondary structure S is locally optimal if the free energy of the structure obtained by removing or adding a base pair is not lower than that of S. We present a novel technique to compute the partition function with respect to the Turner nearest neighbor energy model of all locally optimal secondary structures of a given RNA sequence. A secondary structure S is saturated if no base pair can be added; i.e. saturated structures are locally optimal with respect to the simple Nussinov energy model.

Using our new algorithm, we show that there are far fewer (Turner) locally optimal secondary structures than saturated structures. This strongly suggests the existence of a funnel in the folding landscape of RNA, thus explaining rapid formation of secondary structure. Using our partition function algorithm, we sample locally optimal secondary structure (kinetic traps) and study the kinetics of folding.

Noise tolerance of micro RNA expression quantification by next-generation sequencing

Jonas Maaskola[1], Nikolaus Rajewsky[1]

[1] *Max-Delbrück-Centrum für molekulare Medizin, Berlin-Buch, Germany*

The typically desired DNA sequencing error rate of below 1 in 1000 is historically founded on the necessity for very high accuracy in conventional, low-sequence-number cloning approaches. Progress in the development of new sequencing technologies mainly aims to increase data volume while at the same time sustaining sequencing accuracy. However, current high-throughput techniques are still prohibitively expensive for mainstream applications, and it is thus interesting to explore possible avenues to lower cost sequencing machines.

We studied the impact of lower sequencing accuracy on micro RNA quantification by next-generation sequencing. We used real *C. elegans* and HeLa small RNA samples that were sequenced using current high-throughput technology. Before mapping to the genome, sequence reads were subjected to increasing levels of white noise to mimic higher sequencing error rates. Subsequently, miRNAs were quantitated. We establish that miRNA quantitation specificity is practically unaffected by even high levels of white noise (specificity at 20% added noise in *C. elegans:* 0.999, HeLa: 0.991). Sensitivity is affected to a degree determined by noise level and error tolerance of the mapping procedure, and found to be only mildly deteriorating at high noise by using a noise tolerant mapping procedure (sensitivity at 20% noise in *C. elegans*: 0.458, HeLa: 0.281). The results suggest that for the studied sequencing application the trade-off between read abundance, sequencing accuracy, and costs could be resolved in favor of cheaper, “sloppier” sequencing machines with higher throughput.

The Limits of Subfunctionalization in Gene Regulatory Networks

Thomas MacCarthy, Aviv Bergman

Dept of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, NY, 10461, USA

The duplication-degeneration-complementation (DDC) model has been proposed as an explanation for the unexpectedly high retention of duplicate genes. The hypothesis proposes that, following gene duplication, the two gene copies degenerate to perform complementary functions that jointly match that of the single ancestral gene, a process also known as subfunctionalization. Previous genomic studies investigating the prevalence of subfunctionalization have suggested an important additional role for neofunctionalization (recruitment of novel interactions) following duplication. Here, we distinguish between subfunctionalization at the regulatory level and at the product level (e.g within temporal or spatial expression domains).

In contrast to what is expected under the DDC model, we use in silico modeling to show that regulatory subfunctionalization is expected to peak and then decrease significantly. At the same time, neofunctionalization increases monotonically, eventually affecting the regulatory elements of the majority of genes. Furthermore, since this process occurs under conditions of stabilizing selection, there is no need to invoke positive selection. At the product level, the frequency of temporal subfunctionalization is no higher than would be expected by chance, a finding that was corroborated using yeast microarray time-course data. We also find that product subfunctionalization is not necessarily caused by regulatory subfunctionalization.

Our results suggest a more complex picture of post-duplication evolution in which subfunctionalization plays only a partial role in conjunction with redundancy and neofunctionalization. We argue that this behavior is a consequence of the high evolutionary plasticity in gene networks.

Discovering transcriptional modules involved in A. thaliana growth and development

Steven Maere[1,2], Eric Bonnet[1,2], Yves Van de Peer[1,2]

[1]Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB), 9052 Gent, Belgium ; [2]Department of Molecular Genetics, Ghent University, 9052 Gent, Belgium

One of the major goals in plant molecular biology is to understand the molecular interactions and regulation networks that underlie plant growth and development. A better understanding of these processes will have important implications for fundamental and applied plant research. The wealth of plant transcriptome data that has become available in the past few years now allows us to reverse engineer plant transcriptional networks on a larger scale than before. Recently, we developed the ENIGMA method to extract expression modules from perturbational microarray experiments [1]. ENIGMA leverages differential expression analysis results to extract potentially overlapping groups of co-differentially expressed genes from the data. The core parameters of the ENIGMA clustering procedure are automatically optimized to reduce the redundancy between modules. The expression modules are further characterized by incorporating other data types such as GO annotation, protein and regulatory interactions, and by suggesting potential regulators that govern the modules' expression behavior. One of the main strengths of ENIGMA is its ability to uncover subtle crosstalk between processes. Here, we apply ENIGMA to a comprehensive set of Affymetrix ATH1 microarray experiments probing the cell cycle, growth and development in *Arabidopsis thaliana*. We identified close to 500 expression modules, many of which are related to plant growth and development. Among the most intriguing finds is a potential link between auxin-mediated developmental processes and stress responses. In addition, we identify several novel candidate cell cycle and developmental genes and regulators.

[1] Maere, S, Van Dijck, P, Kuiper M (2008) Extracting expression modules from perturbational gene expression compendia. BMC Systems Biology 2:33

A Combinatorial Model for Synthetic Cell-Cell Communication

Amin R. Mazloom[1], Anastasia Sagalovitch[2], Avi Ma'ayan[1]

[1]Departmet of Pharmacology & System Therapeutics, Mt. Sinai School of Medicine;
[2]Baruch College, City University of New York

Synthetic biology aims to design engineered biological systems that can perform predictable computation using living cells and their integral biochemical molecular parts. Such efforts are challenged with scarcity of molecular parts and methods for controlling cellular behavior. Therefore, computational modeling and simulations could be used to construct realistic complex synthetically engineered biological systems to identify potential obstacles and propose optimized designs before such circuits are implemented with real natural parts.

Here we designed a synthetic network of cell colonies communicating by source cells secreting molecules that induce the production of other molecules in target cells. These cell colonies reside on a microfluidic device organized as an instance of a Road Coloring Problem (RCP) graph where cell colonies represent vertices and directional fluidic tubes represent the arcs. The out-degree of an RCP graph is constant, where in our case each node has two outgoing arcs. Arcs are colored-coded distinctly with a finite a set of colors. This allows for every walk on the graph that follows a specific sequence to converge onto the same target node. Each cell colony contains cells transfected with genetic constructs consisting of an artificial promoter coupled to a gene that can give rise to an extracellular ligand. We utilized the Graph Coloring Problem (GCP) algorithm to find an optimal coloring scheme for real construct design for a realistic biological implementation. To simulate the network, we employed NetLogo which is a powerful *agent-based* simulation platform. The NetLogo simulation environment is composed of '*turtles'* (agents) autonomously moving on '*patches*' (gird environment). We were able to simulate diffusion of extracellular signals (ligands), as well as gene activation and protein translation that resulted in realistic cell-cell communication. The simulation helped us identify several potential barriers for real biological implementation.

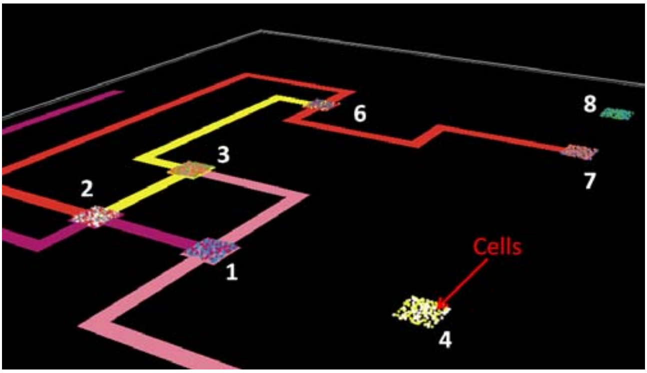


Figure 1 NetLogo simulation snapshot depicts seven (out of eight) cell colonies along with different transfected cells within each colony. The colored connections among the colonies show ligand diffusion, with the color specifying the type of ligand within the micro fluidic tube.

Motif based module identification in integrated networks

Tom Michoel[1,2] and Yves Van de Peer[1,2]

[1]*Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium;*
[2]*Department of Molecular Genetics, UGent, Technologiepark 927, B-9052 Gent, Belgium;*

Cellular functions are carried out by modules, defined as sets of interacting molecules functioning more or less independently. The automatic identification of functional modules from various types of high-throughput data is one of the foremost challenges in the top-down approach to systems biology. In undirected networks, modules can be defined unambiguously as sets of densely (ideally, completely) connected nodes with few ties to the rest of the network. However, such a simple definition is too limited for integrated networks where one is interested in identifying mixed modules with edges between nodes coming from various types of directed and undirected networks.

Here we present a general framework for the identification of modules in integrated networks based on the theory of network motifs. Network motifs are small subgraphs (2-4 nodes) which occur significantly more often than expected by random chance in real networks, including integrated networks. In transcription regulatory networks, network motifs have been shown to aggregate into larger topological units which form prototypical examples of functional modules which receive, process and output signals as independent units. Motivated by this example we define for each network motif a module score which counts for a given subset of nodes the number of motif occurences relative to the number of nodes. The module score is maximized by perfect topological generalizations of the motif without missing edges, similar to the maximization of modularity scores in undirected networks by completely connected sets. This score is used to partition all motif occurences in the integrated network. Since subgraphs are partitioned instead of nodes, it follows automatically that nodes, and even edges, may belong to different modules. By repeating the algorithm for all overrepresented network motifs a complete module atlas for a given integrated network can be derived. Mathematically, the maximization of the module score is carried out by repeated sparse eigen- and singular vector computations which generalize the computation of the PageRank or hub and authority weights for adjacency matrices to higher dimensional adjacency tensors.

We have applied this approach to an integrated network for *S. cerevisiae* consisting of 6306 nodes, 37299 protein-protein, 9328 transcription factor binding and 4624 phosphorylation interactions, using a set of previously identified mixed motifs. As an example, we will highlight the structural differences between modules involving transcriptional and phosphorylation interactions, despite their similarity at the motif level.

Biophysical model of interactions between transcription factors and chromatin

Leonid A. Mirny

Harvard-MIT Division of Health Science and Technology, Massachusetts Institute of Technology, Cambridge, MA

Binding of transcription factors (TFs) to DNA is critical for triggering a cascade of events that lead to gene expression. The role of chromatin in this process is not considered by traditional biochemical models of protein-DNA interaction, or is limited to the passive DNA sequestration by the nucleosomes. Taking into account dynamic structure of chromatin is important for understanding transcription regulation in eukaryotes.

Here we present a biophysical model of interactions between TFs and chromatinized DNA. The model takes into account dynamics of nucleosomes as well as other important features of eukaryotic regulatory regions such as the clustering TF binding sites, nucleosome-positioning DNA signals etc. Our model demonstrates that a wide range of biological phenomena can be explained by interactions between TFs and chromatin, and provides a quantitative description of the following processes:

- cooperative binding and synergistic action of non-interacting TFs;
- access of TFs to chromatinized DNA;
- displacement of nucleosomes from regulatory regions;
- rapid evolutionary changes in arrangement and membership of TF-binding sites in eukaryotic regulatory regions.

Strikingly, we found that cooperative binding of TFs to chromatinized DNA is identical to the Monod-Wyman-Changeux model of allosteric cooperativity in hemoglobin, pointing at a general mechanism of cooperativity employed in a range of biological systems. This parallel allowed us to use classical results in biochemistry to gain deep insights into the mechanisms of gene regulation.

Information flow analysis of protein networks in *C.elegans* and *S. cerevisiae* predicts functionally important genes

Patrycja V. Missiuro[1,2], Kesheng Liu[1], Lihua Zou[3], Brian C. Ross[1], Guoyan Zhao[5], Jun S. Liu[4], and Hui Ge[1]

[1]*Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142;*

[2]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139;*

[3]*Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115;*

[4]*Department of Statistics, Harvard University, Cambridge, MA 02138;*

[5]*Department of Genetics, Washington University, St. Louis, MO 63108.*

Recent studies of cellular networks have revealed modular organizations of genes and proteins. For example, in interactome networks, a module refers to a group of interacting proteins that form molecular complexes and/or biochemical pathways and together mediate a biological process. However, it is still poorly understood how biological information is transmitted between different modules. We have developed “information flow analysis,” a new computational approach that identifies proteins central to the transmission of information throughout the network. We hypothesize that the proteins scoring high in information flow are in central positions of an interactome network and connect various functional modules. In the information flow analysis, we represent an interactome network as an electrical circuit, where interactions are modeled as resistors and proteins as interconnecting junctions. Construing the propagation of biological signals as flow of electrical current, our method calculates an “information flow score” for every protein. Unlike previous metrics of network centrality such as degree or betweenness that only consider topological features, our approach incorporates the confidence scores of protein-protein interactions and automatically considers all possible paths in a network when evaluating the importance of each protein.

We apply our method to protein-protein interaction networks in *S. cerevisiae* and *C. elegans*. We find that the likelihood of observing lethality and pleiotropy when a protein is eliminated is positively correlated with its information flow score. Even among proteins of low degree or low betweenness, high information scores serve as a strong predictor of loss-of-function lethality or pleiotropy. The correlation between information flow scores and phenotypes supports our hypothesis that the proteins of high information flow play important roles in connecting different modules of an interactome network. Using gene expression pattern data, we extract tissue specific networks from *C. elegans* interactome and find that information flow can detect genes locally important to proper functional development of a specific tissue. We show that the information flow scoring of proteins is more consistent than betweenness when a large amount of noise is added to an interactome. We conclude that the information flow analysis is an effective method for identifying genes that reside in central positions in interactome networks.

A Novel Genome-Wide Analysis of Comparative Genomic Hybridization (CGH) Microarray to Reveal Phylogenetic Relationships in Bovines

Apratim Mitra[1], George Liu[2], Jiuzhou Song[1]

[1]Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742; [2]Bovine Functional Genomics Lab, ANRI, ARS, USDA, Beltsville, MD 20705-2350

Array-based comparative genomics hybridization (CGH) has gained prevalence as a technique of choice for the detection of structural variations in the genome. Although analysis tools for CGH data are abundant, they have found major applications predominantly in the study of diseases, such as cancer, that are characterized by widespread copy number variations (CNV). However, such variations are not limited to diseased subjects. Indeed, normal individuals also exhibit CNVs on a smaller scale. Such variations have been annotated to some extent in humans, but attempts to find characteristic CNVs for distinct populations have been largely unsuccessful.

In this study, we propose a novel genome-wide method of classification using CGH data, in order to reveal putative phylogenetic relationships between bovines. We analyze data from 32 individuals from 5 breeds. The data is first denoised using Hu's algorithm [1] before we perform feature extraction with the Haar wavelet. The wavelet power spectrum is then calculated and the spectrum profiles are classified using Ward's hierarchical clustering. Our approach correctly classified 4 out of the 5 breeds. Pair-wise comparisons of the clusters using the exact F-test [2] showed that they were significantly different ($p<0.05$) in all except two cases. Thus, our results suggest that CGH data can be used for uncovering potential phylogenetic relationships between closely related individuals. This can lead to increased efforts towards characterization of CNVs in distinct populations and in conjunction with single nucleotide polymorphism information, might lead to improved understanding of various genetic phenomena.

[1] Hu J, Gao JB, Cao Y, Bottinger E, Zhang W (2007) Exploiting noise in array CGH data to improve detection of DNA copy number change. Nucleic Acids Res 35(5): e35.

[2] Lauter J (1996) Exact t and F Tests for Analyzing Studies with Multiple Endpoints. Biometrics 52(3): 964-970

Determining the extent of epistasis among whole-genome duplicates in yeast

Gabriel Musso[1,2], Michael Costanzo[1,2], Charles Boone[1,2], Andrew Emili[1,2] & Zhaolei Zhang[1,2]

[1] Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College St., Toronto ON, Canada M5S 3E1; [2] Department of Molecular Genetics, University of Toronto, 1 Kings College Circle, Toronto ON, Canada M5S 1A8;

Immediately following a gene duplication event there is a period of enhanced selection acting on either one or both duplicates (paralogs) which results in some extent of functional divergence. However, as redundancy among extant duplicates is thought to confer genomic robustness, a consequent question is: how much functional overlap exists between the subset of duplicates that are retained over long periods of time? To examine this issue we recently determined the extent of synthetic genetic interactions between paralogous gene pairs resulting from an ancient Whole Genome Duplication (WGD) event in yeast. Logically, if paralogs have retained substantial functional overlap the cell should be able to compensate for loss of either individual paralog without observable phenotypic consequence, but should exhibit morbidity upon double-deletion of sister paralogs (i.e. epistasis being observable as an aggravating synthetic genetic interaction).

Our findings indicated a complex relationship between the functional redundancy of paralog pairs, gauged mainly as the overlap in protein interaction partners, and the propensity to exhibit epistasis. One potential explanation for this discrepancy is the fact that some functional categories of paralogs were markedly under-represented in their frequencies of aggravating genetic interactions. Further, we noted that some of these same non-epistatic paralog pairs, notably those involved in cell signaling, showed an increased likelihood to exhibit genetic interactions only under cellular duress, suggesting that epistasis may be far more pervasive among WGD paralogs than had been observed in the limited conditions tested. In order to further identify the full extent of epistasis among this group, Support Vector Machine analysis conducted using an expansive set of functional attributes was used to identify a small number of pairs with high circumstantial evidence for functional redundancy but which did not exhibit epistasis under conditions tested. We next seek to quantify the extent of redundancy between these selected pairs through a novel application of the Synthetic Genetic Array (SGA) platform, that of comparing the SGA profiles of strains carrying double-deletions of paralog pairs to that of the constituent single mutant strains. By so doing, sensitivities of the double-mutant strains which were not obvious using either constituent single mutant strain will be determined, and ultimately an atlas of redundancy and epistasis can be derived.

Uncovering the Organization of Physical Interactomes via Network Schemas

Elena Nabieva[1,2,3], Eric Banks[1,2], Bernard Chazelle[1], Mona Singh[1,2]

[1]Princeton University, Department of Computer Science; [2]Princeton University, Lewis-Sigler Institute for Integrative Genomics; [3]present address: Whitehead Institute for Biomedical Research

Understanding the ways in which proteins come together to perform various biological processes and thus create the life of a cell is one of the key challenges of biology. Large-scale determination of protein-protein interactions is an important step towards addressing this fundamental question. Commonly represented as networks or graph, protein interactions create new opportunities for understanding cellular organization and functioning, but simultaneously pose the challenge of interpreting these data to obtain biological knowledge. Here, we focus on the problem of identifying shared mechanisms within interactomes, and introduce *network schemas* to describe patterns of interaction among distinct types of proteins. Network schemas specify descriptions of proteins and the topology of interactions among them. We develop a novel computational procedure for systematically uncovering recurring, over-represented schemas in interaction networks. We apply our methods to the *S. cerevisiae* physical interactome, focusing on schemas consisting of proteins described via sequence motifs and molecular function annotations and interacting with one another in one of four basic network topologies. We identify hundreds of recurring and over-represented network schemas of various complexities, and demonstrate via graph-theoretic representations how more complex schemas are organized in terms of their lower-order constituents. The uncovered schemas span a wide-range of cellular activities, with many signaling and transport related higher-order schemas. We establish the functional importance of the schemas by showing that they correspond to functionally cohesive sets of proteins, are enriched in the frequency with which they have instances in the *H. sapiens* interactome, and are useful for predicting protein function. In addition, we touch upon the use of schema analysis for comparative interactomic studies by examining the simplest network schemas in the *H. sapiens* interactome. Our findings suggest that network schemas are a powerful paradigm for organizing, interrogating, and annotating cellular networks

An unbiased approach for inferring sparse multivariate autoregressive models from time series measurements

H. Koeppl[1], F. Parisi[2], F. Naef[2]

[1]*School of Computer and Communication Sciences, Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland;* [2] *School of Life Sciences, Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland*

The ambitious goal of reverse-engineering is to learn causal relationships between proteins or genes from systematic time series data. For example we would like to learn from data, which are the direct targets of a transcription factor or a protein kinase. The parametric approach to this inference problem is to postulate a mechanistic model that underlies the recorded time series. A parsimonious model in such contexts is a linear model and should be able to capture the data well if the underlying system is locally perturbed around one of its operating points. Models relating dynamics of proteins with dynamics of gene or the dynamics of different genes to each other should account for memory as interaction between these players are known to be not instantaneous.

In this work we propose to infer causal structures by linear convolutional models, in particular linear multivariate autoregressive (AR) models. Such models can integrate events from several past time steps and can thus account for the delays due to transcription and translation imprinting in the transient dynamics. We distinguish between process noise and measurement noise and assume that the dominating noise in population measurements is the latter. Accounting for measurement noise leads to a non-quadratic estimation problem as opposed to the classical linear AR estimates for process noise. We develop a multivariate extension of the Steiglitz-McBride estimation scheme resulting in a convergent sequence of linear estimation steps. Augmenting the vector of unknown model parameters with the initial conditions of the model, the proposed scheme yields unbiased estimates and we numerically show that those are of minimum variance and thus optimal in the sense of the Cramer-Rao bound.

In the sequence of quadratic objective functions we incorporate a quadratic regularizer with individual regularization parameters enforcing sparsity of the solution by employing an expectation-maximization (EM) procedure to adjust the regularization strength. We discuss the issue of reconstructing a continuous-time dynamical system from equidistant sampled data and detail its implication for the form of the regularizer. For linear systems conditions are given under which the resulting exact non-quadratic regularizer can well be approximated by a quadratic one.

The proposed scheme is evaluated using a large population of random linear systems, one well known 4-gene *in silico* network and the real qPCR data of the 5-gene network supplied by the DREAM2 challenge. Different performance metrics such as the trajectory L2-norm, the parameter L2-norm and the Hamming distance after thresholding are deployed. The approach is put in context by comparison to the classical AR estimation scheme and to a non-regularized version of the presented Steiglitz-McBride scheme.

Inferring signal transduction cascades from synthetic and experimental data: a comparison of statistical approaches

Xiaoyu Jiang[1], Peter V. Henstock[1], and Madhusudan Natarajan[1]

[1] *Pfizer Research Technology Center, 620 Memorial Dr., Cambridge MA 02139, USA*

The cornerstone of quantitative approaches to understand biology is the ability to identify causal relationships between biological entities from sparse experimental data. Several recent studies have involved systematically perturbing cellular systems and assaying responses downstream of these perturbations. The reconstruction of the network from these assayed responses remains an open research question.

To address the important issues in this context including efficiently using time course data for each protein, quantifying the confidence of predicted edges, and inferring causality (i.e. directions) of the associations, we have explored three different statistical approaches that utilize conditional independence and partial correlations to infer the network topology. The techniques include (1) recursive structural learning of directed acyclic graphs (DAGs) where the structural learning problem is divided into small subsets of vertexes and edges linked based on conditional independence; (2) inferring dependencies from dynamic partial correlations that capture the functional nature of the observations; and (3) a dynamic Bayesian network approach that approximates the full order conditional dependencies with two lower order dependency relationships.

A comparison of these methods in their ability to infer the correct edges and directions for a medium-sized network of ~20 nodes is first performed on various synthetic networks with known structures and specified assumptions to understand the advantages and disadvantages of each method. The methods are then applied to analyze experimental measurements of both proximal signaling of protein phosphorylation, and downstream signaling of measured cytokine secretion in response to specific perturbations in U937 cells. By examining the consistencies of the results from different approaches, we extract plausible hypotheses for further studies.

Estimation of Statistical Significance for Motif Scans by Hidden Markov Models

Lee A. Newberg[1,2]

[1]Center for Bioinformatics, New York State Department of Health Wadsworth Center, P.O. Box 509, Albany NY 12201-0509, USA; [2]Department of Computer Science, Rensselaer Polytechnic Institute, 110 8th St., Troy NY 12180-3590, USA.

Scanning for regulatory elements with a known motif model is an important part of regulatory genomics. Approaches exist for accurately evaluating the statistical significance of finds, for mononucleotide models / position weight matrices [1], multi-nucleotide models [2], and phylogenetic mononucleotide models [3, 4], but the best approaches for the most general hidden Markov models, powerful enough to exploit modules of regulatory elements, are more approximate (e.g., [5] relies on a Gumbel distribution approximation). Especially with multi-genome scans, non-negligible false positive rates will cause large numbers of false hits and needless subsequent experiments, thus there is an important need for accurate evaluations of statistical significance for located regulatory elements.

We have developed a technique that accurately estimates the statistical significance (also termed *p*-value or type I error or false positive rate) for a scan by any hidden Markov model. The technique works whether regulatory element finds are evaluated by a "maximum score" approach, where the scan score is from the best placement of regulatory elements in the scanned sequence, or are evaluated by a "forward sum" approach, where all possible placements of regulatory elements contribute to the final score.

The technique employs a thermodynamic Boltzmann probability distribution to randomly generate sequences to be scanned. By controlling a temperature parameter, the sampling can be biased towards sequences yielding a scan score of interest. A calculation involving the scan scores of such sampled sequences quickly gives the accurate *p*-value estimate. We have used the technique to accurately estimate *p*-values as low as 1e-4000.

As scans of databases of trillions of nucleotides becomes the norm, we will need the ability to quickly estimate *p*-values at the 1e-12 level and smaller. This technique thus fills an essential niche.

[1] Staden, R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci*. http://dx.doi.org/10.1093/bioinformatics/5.2.89.

[2] Huang, H., Kao, M.-C. J., Zhou, X., Liu, J. S. & Wong, W. H. (2004) Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *J Comput Biol*. http://dx.doi.org/10.1089/106652704773416858.

[3] Moses, A. M., Chiang, D. Y., Pollard, D. A., Iyer, V. N. & Eisen, M. B. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*. http://dx.doi.org/10.1186/gb-2004-5-12-r98.

[4] Carmack, C. S., McCue, L. A., Newberg, L. A. & Lawrence, C. E. (2007) PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol Biol.* http://dx.doi.org/10.1186/1748-7188-2-1.

[5] Eddy, S. R. (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol*. http://dx.doi.org/10.1371/journal.pcbi.1000069.

CATCH those ChIP profiles!

Fiona G Nielsen[1,2], Kasper Markus[3], Henk Stunnenberg[2], Lene Monrad Favrholdt[3], Martijn Huynen[1]

[1]Department of Molecular Biology, Nijmegen Centre for Molecular Life Sciences, the Netherlands; [2]University of Nijmegen Medical Centre, the Netherlands; [3]University of Southern Denmark, Odense, Denmark.

In the field of epigenetics the regulation of genes is examined not at the level of DNA (genomics) but at the level of proteins associated to and interacting with the DNA. The distinct patterns of bound proteins are indicators of gene activity, as well as clues to decyphering the complex combinatorial regulatory machinery of the cell. Characterising epigenetic profiles from ChIP profiling experiments in a genome-wide scale is an overwhelming task that calls for automated data-mining techniques to aid analysis.

We present the program CATCH to perform unsupervised clustering of ChIP profiles. CATCH implements a hierachical clustering algorithm with an exhaustive all-pairs comparison at each clustering step. As part of the comparison CATCH simultaneously aligns the ChIP signal profiles, so the profiles are repeatedly clustered by their most informative alignment.

With CATCH we have analysed the ENCODE ChIP-on-chip data set from Heintzman et al [1]. Next to recovering the average promoter and enhancer profiles as described in [1], we discovered several other distinct epigenetic patterns. Our results include specific profile patterns for bidirectional promoters, as well as a pattern specific for promoters of the heavily repressed genes from the olfactory receptor gene family.

CATCH automates the clustering of ChIP profiles and enables easy browsing and export of results. This ease of analysis is paramount to manage the increasing volume of experimental ChIP profiling data.

CATCH is open-source and freely available from the authors homepage: www.cmbi.ru.nl/~fnielsen/CATCH

[1] Heintzman et al, Nature Genetics vol 39, 3, March 2007.

Top-Down Approaches to the Reconstruction of Adaptive and Oscillatory Networks

Evgeni V. Nikolaev and Boris N. Kholodenko

Department of Pathology, Anatomy and Cell Biology, Thomas Jefferson University, Philadelphia, PA 19107

Recent advances in high-throughput experimental technologies have facilitated the acquisition of large genomics and proteomics datasets. Although high-throughput measurements are now widely available, the detailed biochemical knowledge of networks of interacting molecular species is still lacking. Therefore, further development of powerful "reverse engineering" concepts and efficient computational procedures to infer network's structure is increasingly appreciated. To this end, Modular Response Analysis (MRA) has been recently developed to unravel connections between network modules from experimental system-wide responses to small perturbations at steady sates. MRA exploits the stationary conditions to dramatically decrease the number of measurements. To infer the modular structure of a network, all what is needed is to measure responses of so –called "communicating" species, and not all network species.

Because biological networks are dynamic, there is an unmet need to develop the inference methods that will utilize time-resolved measurements. To meet this challenge, we have generalized MRA to cover adaptive and oscillatory networks. Specifically, we extended the unraveling MRA method to

- *adaptive* systems displaying peaks before stationary levels are reached;
- *oscillatory* systems that exhibit limit cycle periodic behavior.

We note that as the time derivatives vanish both at stationary and extremum time points, the connection architecture can be inferred from the responses of each individual node determined at the corresponding extremum point. However, for large-scale networks, it is impractical (and often impossible) to measure the time-resolved responses of all species, and a modular approach must be employed. An additional reason of using MRA is that while the material fluxes between modules are not allowed, these fluxes can occur within single modules.

Application of MRA inference methods is challenged by the fact that the time moments of attaining extrema are generally different for different nodes. Moreover, MRA algorithms cannot be directly applied to the cell cycle or circadian rhythm networks, because time-dependent responses generally diverge as the time progresses due to changes in the phase between the corresponding reference and perturbed oscillations. Remarkably, a solution to these problems lies in the averaging of the time trajectories for intramodular species, while determining the responses to perturbations at the time of extrema for those nodes that provide communication between different modules. The averaging procedure leads to the formulation of a linear integral equation encompassing the following three important additive terms, (*a*) the term accounting for response coefficients at extrema, (*b*) the integral term accounting for the averaged modular behavior before the extremum is reached or over the minimal oscillatory period, and (*c*) the term describing the sensitivity of the extremum moment or the oscillatory period to small perturbations, respectively. In this work, we present and benchmark the developed integral MRA method by using several *in silico* models of classical MAPK and GTPase cascades under certain noise levels.

Modeling the Cell Decision Process in Response to the DNA Lesion O6Methylguanine

Ericka Noonan[1,2], Douglas Lauffenburger[1-3], Leona Samson[1-3]

[1]Department of Biological Engineering, MIT; [2]Center for Environmental Health Sciences, MIT; [3]Department of Biology, MIT.

The cellular outcome resulting from DNA damage is determined by the coordinated activity of signaling molecules from a number of pathways including stress response pathways, survival pathways, checkpoint signaling pathways, DNA damage repair pathways, and apoptotic pathways. How a cell deciphers this complex signaling network to yield a particular phenotypic response is unclear. The objective of this project is to construct an experimentally-based computational model capable of interpreting and predicting the cell responses to a specific DNA base lesion commonly induced by chemotherapeutic agents, based on measurements of key intracellular signaling pathways.

Alkylating agents are a class of DNA damaging agents which can be found endogenously, exogenously, and are commonly used clinically as cancer chemotherapeutic agents due to their cytotoxicity. They produce a diversity of DNA lesions which if left unrepaired could result in mutations leading to human diseases such as cancer. The O^6-methylguanine lesion is a critical base lesion which can be cytotoxic as well as mutagenic. This lesion has been shown to induce apoptosis in a manner that is both O^6MeG DNA methyltransferase (MGMT) repair- and mismatch repair (MMR)-dependent. Through manipulation of the repair of this adduct, it is possible to examine the cellular response resulting specifically from O^6MeG.

The signal transduction network downstream of this damage is largely regulated by dynamic protein phosphorylation events. Signaling events were monitored across multiple pathways by measuring kinase activities or phosphorylation states over time after treatment with the DNA methylating agent MNNG. In parallel, the phenotypic responses cell cycle arrest and cell death were measured. IP kinase activity assays revealed that both the stress activated protein kinase JNK and Chk2 kinase are activated in response to MNNG. In addition, the proteins ATM(S1981), Chk2(T68), Chk1(S317), p53(S15 and S20), and H2AX(S139) were found to be phosphorylated at sites which are linked to their activation.

Using the data-driven modeling approach Partial Least Squares Regression (PLSR) we hope to gain biological insight into the DNA damage response by revealing combinations of signals which are predictive of different cellular outcomes (e.g. death, cell cycle arrest). To test model predictions we will perturb the network by overexpressing or knocking down key signaling network proteins. Comparison between *in silico* and experimental perturbations of the network will permit us to optimize the predictive power of our computational model.

RegPrecise: database of manually curated inferences of regulons in bacteria.

Pavel S. Novichkov[1], Mikhail S. Gelfand[2], Inna Dubchak[1,4], Dmitry A. Rodionov[2,3]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia; [3]Burnham Institute for Medical Research, La Jolla, CA 92037,USA; [4]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA.

Identification and reconstruction of various transcriptional regulons in bacteria using a computational comparative genomics approach is coming of age. During the past decade a large number of manually-curated high quality inferences of transcriptional regulatory interactions were accumulated for diverse taxonomic groups of bacteria. These data provide a good foundation for understanding molecular mechanisms of transcriptional regulation, identification of regulatory circuits, and interconnections among circuits within the cell. Traditional experimental methods for regulon analysis have certain limitations both in terms of productivity and feasibility. While the development of high-throughput transcriptome approaches allow to obtain genome-scale gene expression patterns, in many cases the complexity of the interactions between regulons makes it difficult to distinguish between direct and indirect effects on transcription. The availability of a large number of closely related genomes allows one to apply comparative genomics to accurately expand already known regulons to yet uncharacterized organisms, and to predict and describe new regulons. Due to fast accumulation of such valuable data, there is a need for a specialized database and associated analysis tools that will compile and present the growing collection of high quality predicted bacterial regulons.

The RegPrecise database was developed for capturing, visualization and analysis of transcription factor regulons that were reconstructed by the comparative genomic approach. The primary object of the database is a single regulon in a particular genome, which is described by the identified transcription factor, its DNA binding site model (a profile), as well as the set of regulated genes, operons and associated operator sites. Regulons for orthologous transcription factors from closely related genomes are combined into the collections that provide an overview of the conserved and variable components of the regulon. A higher level representation of the regulatory interactions is also provided for orthologous regulons described in several bacterial taxonomic groups enabling comparison and evolutionary analysis of the transcription factor binding motifs. Another view of complex data in the database is a general overview of multiple regulons inferred in a set of closely related group of genomes.

The current version of database covers more than 250 genomes and 180 profiles. Among others, it represents the results of our recent comparative genomic reconstruction of metabolic regulons in 13 Shewanella species that included near 70 transcription factors, approximately 400 binding sites and more than 1000 target genes per each genome. The database gives access to large regulatory networks reconstructed for certain metabolic pathways, e.g. degradation of fatty acids, branch chain amino acids, and aminosugars, homeostasis of biometals, and biosynthesis of NAD cofactor. In the near future we are planning to add a large collection of regulons for the LacI family transcription factors.

Coregulation Mapping Based on Individual Phenotypic Variation in Response to Virus Infection

German Nudelman[1†], Jianzhong Hu[2†], Yongchao Ge[1†], Madhu Kumar[2], Jeremy Seto[1], Jamie L. Duke[3], Steven H. Kleinstein[3], Fernand Hayot[1], Stuart C. Sealfon[1], James G. Wetmur[2]

[1]*Center for Translational Systems Biology and Department of Neurology, Mount Sinai School of Medicine, New York, NY 10029, USA*

[2]*Department of Microbiology, Mount Sinai School of Medicine, New York, NY 10029, USA*

[3]*Interdepartmental Program in Computational Biology and Bioinformatics and Department of Pathology, Yale University School of Medicine, New Haven, Connecticut 06510, USA*

Dendritic cells (DCs) are specialized antigen presenting cells that are important mediators of the early responses to virus infection. Infection of DCs induces secretion of cytokines and chemokines that influence the efficiency of the immune response. Variation in the responses of individual hosts to virus infection, for example due to genetic heterogeneity, is widely acknowledged but difficult to study. In order to develop and test a computational approach for investigating individual variations in responses to infection, we generated a dataset consisting of cytokine profiles induced by infection in cells obtained from 200 different human donors. The levels of eight secreted immune signaling proteins induced by Newcastle Disease Virus infection of cultured primary DCs showed a large variation across individuals. We generated a map of the relative coregulation of the proteins using pairwise correlation and multidimensional scaling (MDS) methodology. The analysis of these data showed that among the eight secreted proteins, CCL5, TNF and IL6 showed the most significant coregulation. We then applied phylogenically-constrained binding site analysis to the promoter regions of the genes encoding these proteins in order to identify the possible transcription factors driving the observed coregulation patterns. Through this analysis, we identified E2F and DP families of TFs as common and restricted to the promoters of clustered genes in the coregulation map. In order to facilitate the use of our methodology by other researchers, we are developing an interactive coregulation explorer web-based tool called CorEx. We propose therefore, a novel approach to study the individual response variation data based on two-dimensional mapping of the pairwise correlations, combined with computational prediction of TFs uniquely present in the promoters of genes clustered in the coregulation map.

Supported by NIAID Contract HHSN2662000500021C and Grant U19 AI06231

Identification and characterization of trans-factors that regulate the splicing / alternative splicing specificity using cell-based assays and RNAi

Razvan Nutiu, Christopher Burge

Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.

Pre-mRNA splicing is a fundamental, precisely regulated post-transcriptional step in gene expression in which specific pre-mRNA sequences are removed and the remaining segments are joined to obtain mature mRNAs that direct protein synthesis. In mammals constitutive and alternative splicing events are regulated by trans-acting proteins that are recruited by cis-acting elements along the RNA sequence such as exonic and intronic splicing enhancers (ESE and ISE) as well as exonic and intronic splicing silencers (ESS and ISS). Since a good deal of progress has been made recently on identification and characterization of cis-acting elements that regulate splicing, we set our goal to identify and study the corresponding trans-acting proteins in mammalian cells. We have designed an ESS responsive three exon minigene reporter that enables identification of functional relationships between cis elements and trans factors by means of an RNAi screen. The first exon of the minigene is a constitutively spliced exon and the third exon is a reporter gene (such as GFP or luciferase) that only lacks the start codon. The second exon of the minigene (test exon) contains an ESS and has a start codon at its 3' end. Skipping of the test exon due to an active ESS leads to a mRNA lacking an in-frame start codon for the reporter gene. Conversely, if the ESS is inactivated by knockdown of the trans factor, the test exon is included to allow proper translation of the reporter gene. Using this minigene reporter we performed a pilot screen where we tested six different ESSs against eighteen different proteins from the hnRNP family. The screen correctly identified our positive control, a G-rich ESS bound by hnRNP H1, as well as a new relationship between a pyrimidine rich ESS and PTB1.

An Evolutionary Approach to Alignment of Multiple Transcription Factor Binding Motifs

Pilib Ó Broin[1], Terry Smith[1], Aaron Golden[1]

[1] *National Centre for Biomedical Engineering Science, NUI, Galway.*

Structurally related Transcription Factors are known to share similar DNA binding preferences. Binding motifs from individual Transcription Factors (represented as Position Weight Matrices) can be aligned to produce a Familial Binding Profile (FBP) or 'average' binding specificity for a group of related proteins. FBPs can be used not only to provide information about the evolution of a particular binding site, but also as a means to bias motif-finding tools in order to identify binding sites for proteins from a particular Transcription Factor family. FBPs can also help to classify novel binding proteins.

In this work we present a Genetic Algorithm for the alignment of multiple Position Weight Matrices. We discuss (i) the scoring metric used to assess the quality of alignments, (ii) the evolutionary mechanisms used to explore the search space, and (iii) the advantages of GAs over other multiple alignment strategies.

Gene regulatory network reconstruction using discrete dynamical system modeling with exact p-values

Zhengyu Ouyang and Mingzhou (Joe) Song

Department of Computer Science, New Mexico State University

DREAM3 *In-Silico*-Network Challenges (Challenges 4) provides 15 trajectory data sets to reconstruct 15 partial gene regulatory networks in *E. coli* and yeast. The trajectories represent time course expressions of selected subsets of genes. In conjunction with the transient states in the trajectories, steady states under "Heterozygous knock-down" and "Null-mutant" experimental conditions are also provided. The steady states can be used to form trajectories by repeating themselves of each experimental condition.

We enter DREAM3 challenges using the discrete dynamical system (DDS) model (discrete time, but continues value), composed of difference equations. Currently, the model contains linear difference equations. A difference equation would specify the regulators for each gene and the coefficients indicating the nature of gene regulation. DDS can be considered discrete-time analogy of the ordinary differential equation (ODE) model.

The least-squares method from linear regression is used DDS model reconstruction by enumerating all possible combinations of regulators for specified maximum number of regulators. This maximum number was determined by the available computational capability. One best set of regulators for each gene from all possible combinations was chosen to optimize the *F* test statistic, often used in linear regression. The null hypothesis in DDS modeling is that no variables interact with each other. However, enumeration of regulators presents a major difference in computing the statistical significance in DDS modeling from the *F*-distribution typically used in linear regression. To enumerate all possible combinations of independent variables in DDS modeling, versus to involve all independent variables without any enumeration in a typical linear regression, increases the probability of getting a greater *F* value under the null hypothesis. Thus the *F*-statistic might not follow the *F*-distribution in DDS modeling.

We randomly permute each original trajectory 400 times to form 400 trajectories in order to obtain realistic null distributions of the *F*-statistic under the experimental condition of the original trajectory. The permutation was achieved by randomly and independently re-ordering the time points on the time course of each gene. Statistically permuted trajectories destroy the original interactions among gene nodes and can thus serve as a sample from the null distribution. Applying DDS reconstruction on 15x400 permuted trajectory files of the original 15 data sets in Challenges 4, we recorded the largest *F*-statistics in DDS modeling for each gene on given numbers of regulators and sample sizes, in 15 three-dimensional permutation statistic tables. Using each table and the corresponding original trajectory file, we can determine the *p*-values of each gene by the percentage of null *F*-statistics that are greater than the observed *F*-statistic of the original trajectory file in DDS modeling. Such computed *p*-values are called exact *p*-values. A regulator set with the least *p*-value was assigned to each gene as its regulators. We finally reconstructed 15 DDS models by obtaining the regulators for all genes from each of the 15 data sets.

Asynchronous Inference of Regulatory Networks

David Oviatt[1], Kenneth Sundberg[1], Mark Clement[1], Quinn Snell[1], Randall Roper[2], Jared Allen[2]

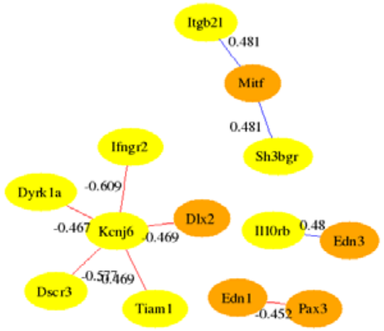
[1]Brigham Young University, [2]Indiana University Purdue University Indianapolis

Modern approaches to treating genetic disorders, cancers and even epidemics rely on a detailed understanding of underlying gene signaling network. Previous work has used time series microarray data to infer gene signaling networks given a large number of accurately timed samples. Microarray data available for many biological experiments is limited to a small number of arrays with little or no time series guarantees. Asynchronous Inference of Regulatory Networks (AIRnet) provides gene signaling network reverse engineering using more practical assumptions about the microarray data. By learning correlation patterns from all pairs of microarray samples, accurate network reconstructions can be performed with data that is normally available in microarray experiments.

The first step in analyzing microarray data is discretizing the data. Data for each gene is separated into clusters using a K-means clustering algorithm. Gene transitions between these clusters are then compared for all pairs of genes on the microarray to produce an influence score indicating the confidence that two genes are related. A filter is then applied to allow the user to filter out network connections that are below a specified level. AIRnet allows the researcher to specify a group of genes of interest and then only displays genes that connect to this interest group with influence scores above the user specified level.

AIRnet has been tested with simulated microarray data as well as real microarray data for the Ts1Cje mouse model and for patients being treated for breast cancer. The Ts1Cje mouse model data revealed some promising information related to Down Syndrome that will be tested in wet-lab experiments. The breast cancer microarrays produced a network, for which many of the relations have been previously verified. On average, over 60% of the gene relationships inferred by AIRnet can be verified using KEGG, Pubmed abstracts or other biological databases included in the PATHGEN search engine (http://dna.cs.byu.edu/pathgen).

In our Down Syndrome analysis, little is known about the connections between Chromosome 21 genes and genes associate with jaw development phenotype genes. Through the AIRnet analysis, several interesting connections have been determined that warrant further experimentation. AIRnet can also be used to determine the difference between the network inferred from trisomic microarrays and euploid samples.



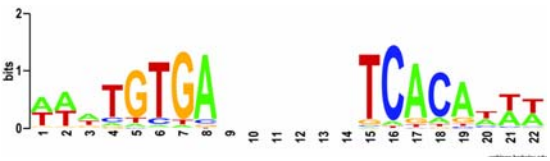
Further development of AIRnet promises assist researchers in understanding gene relationships, which will lead to improved drug development and gene therapies. AIRnet information can be found at http://dna.cs.byu.edu/AIRnet.

PhyloScan II Web Server: Scanning for *Cis*-Regulatory Elements in Mixed Aligned and Unaligned Sequence Data.

Michael J. Palumbo[1], Lee A. Newberg[1,2]

[1]Center for Bioinformatics, New York State Department of Health Wadsworth Center, P.O. Box 509, Albany NY 12201-0509, USA; [2]Department of Computer Science, Rensselaer Polytechnic Institute, 110 8th St., Troy NY 12180-3590, USA

The PhyloScan II Web Server scans for novel *cis*-regulatory elements matching a known motif model <http://bayesweb.wadsworth.org/cgi-bin/phyloscanV2.pl>. It scans the multi-genomic intergenic sequence data you provide (in MAF format) using the model you supply: a phylogenetic tree (in Newick format) and a set of known *cis*-regulatory elements. You may designate which positions within the model are relevant to binding and whether the motif is palindromic. Additionally you specify a species of primary interest and the extent to which you wish to permit weak signals in the primary species to be "rescued" by stronger signals from orthologous regions. PhyloScan II reports discoveries to you with an associated *q*-value (also termed false discovery rate) that estimates the fraction of reported discoveries of this quality or better that are likely to be spurious.



To better detect regulons when scanning for *cis*-regulatory elements matching the motif model, PhyloScan [1] combines evidence from both orthologous sites found in other species and from multiple sites within an intergenic region. PhyloScan garners statistical sensitivity without compromising specificity by accepting both multiply aligned and unaligned orthologous sequence data; analyzing the aligned data by adapting Staden's approach [2] to Felsenstein's tree-likelihood algorithm [3,4]. PhyloScan's ability to combine evidence from multiple sites within an intergenic region further bolsters its statistical sensitivity, allowing it to find genes regulated by multiple weak sites. The statistical significance of PhyloScan's gene predictions is calculated directly, without employing training sets. The previous web server, PhyloScan I, provides direct access to intergenic region sequence data for *E. coli* and several related gamma-proteobacteria.

[1] Carmack, C. S., McCue, L. A., Newberg, L. A. & Lawrence, C. E. (2007) PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol Biol.* http://dx.doi.org/10.1186/1748-7188-2-1.

[2] Staden, R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci*. http://dx.doi.org/10.1093/bioinformatics/5.2.89.

[3] Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. http://dx.doi.org/10.1007/BF01734359.

[4] Moses, A. M., Chiang, D. Y., Pollard, D. A., Iyer, V. N. & Eisen, M. B. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* http://dx.doi.org/10.1186/gb-2004-5-12-r98

Signal propagation in mutually communicating networks

Keunwan Park[1], Dongsup Kim[1]

[1]*Department of Bio and Brain Engineering, KAIST, Daejeon, Republic of Korea*

Many biological systems can be modeled as some type of interaction network. Networks need to respond to external perturbations adequately for performing necessary functions. Thus, the key feature of interaction networks is the network's ability to communicate necessary information between nodes for efficient signal transduction. In addition, the communication process can vary depending on the specific type of networks, which depends on the types of signal flows from node to node. Therefore, understanding how the signals in a network propagate and which nodes are crucial for mediating communications is important for understanding the properties of interaction networks.

In this study, we propose a simple method for modeling signal propagation in mutually communication networks, relying solely on its structure. In our method, weighted undirected network was represented by a transition matrix whose sum of each rows is one (The signal flows from row to column). Using the transition matrix, we obtained the expected visiting time (EVT) profile from a starting node to a destination node at which the signal arrives. The EVT profile could be interpreted as information of affected intermediate nodes while a starting signal go around until it reaches an end node. In addition, the EVT profiles for the other nodes were averaged under the assumption that one node communicates with every other nodes. Finally, we propose that the averaged EVT profile from a starting node would represent the propagated effect from a starting node.

To validate our method, the protein structure network of PDZ domain was analyzed. PDZ domain is a small structural domain found in many signaling proteins and one of the popular examples studying intramolecular signaling pathway. One of its structures (1GM1) was transformed into an interaction network based on distance relationships between amino acids. That is, closely located amino acids (network nodes) were assumed to interact with each other. The transition matrix representing the transition probability between nodes was constructed. In the matrix, edges between closer nodes had larger weights, or higher transition probability. We chose a starting node 71H (peptide binding site) as the previous studies, and applied the average EVT profile from 71H. The result showed that the signal from 71H affects distal nodes as well as close nodes. The 35I was the highest affected node by 71H, and a mutation I35V was experimentally verified to have a significantly decreased binding affinity for peptide (K_d~100μM). Moreover, the dominant signaling pathway was proposed by maximizing EVT from 71H to 40I, and the resulting traverse order was 71H→35I→22V→37V→40I. The second dominant signaling pathway was also proposed with the same procedure after eliminating the nodes belonged to the first dominant pathway: 71H→74A→78L→20I→40I. The previous study using anisotripic thermal diffusion suggested the similar pathway, 71H→22V→20I→40I.

If a signal affects neighborhoods simultaneously, it is unlikely that it propagates only through the single optimal path. Accordingly, the previous studies describing the signaling pathway as a single shortest path may lose some information. Our method models the signaling pathway in a mutually communicating network using transition probability and EVT profile, so it does not take any sort of ideal path, and represents the multiple-path effects (EVT) well.

Hierarchical feedback control and reverse engineering of transcriptional networks involved in sex hormone synthesis in ovaries

Edward J Perkins, Natàlia Garcia-Reyero[2], Sandy Brasfield[1], Daniel L Villeneuve[3], Xin Guan[4], Dalma Martinovic[3], Nancy Denslow[2], Youping Deng [4,5], Ying Li[5], Tanwir Habib[5], Jason Shoemaker[6], and Francis Doyle III[6]

[1]US Army Engineer Research and Development Center, Vicksburg, MS, USA;
[2] University of Florida, Gainesville, FL, USA;
[3]US EPA Mid-Continent Ecology Division, Duluth, MN, USA;
[4] SpecPro, US Army ERDC, Vicksburg, MS, USA;
[5] University of Southern Mississippi, MS, USA;
[6] University of California at Santa Barbara, CA, USA;

Sex steroid synthesis is controlled through a highly conserved and complex interplay within the endocrine pathway comprised of the hypothalamus-pituitary-gonad (HPG) axis. We investigated regulation of steroid synthesis in ovaries of the experimental fish model *Pimephales promelas* upon exposure to chemicals using a combination of microarrays, PCR, hormone monitoring, dynamic simulation and reverse engineering of transcriptional interactions. The HPG axis could compensate for chemical inhibition of steroid synthesis in exposed whole fish, however ovary tissue exposed in vitro could not. In the absence of hormonal feedback control from the HPG axis, local responses in gene expression and enzyme inhibition dominated in ovary tissues resulting in limited hormone production. Sensitivity analysis of an ovary metabolic model suggested local regulatory events would occur rapidly after stimulation. Consistent with this prediction, oscillatory behaviors in expression of key metabolic genes, StAR and CYP 11, were revealed within 60 min of substrate addition characteristic of feedback control systems. The transcriptional interactions in ovaries were further investigated by inference of a transcriptional interaction network in ovaries. Microarray data was generated for ovaries functioning in over 180 different conditions including time series exposure to various chemicals, in vitro vs in vivo exposures, and stages of ovary maturation. Differentially expressed genes were identified within each condition. The union of all differentially expressed genes across all conditions was used to infer transcriptional interaction networks in ovaries. 2000 genes plus 55 genes known to be involved in steroid metabolism were used in network inference. The inferred network revealed novel interactions within the ovary and provide the basis for understanding local control strategies. Overall, these experiments reveal distributed control of steroid synthesis where global control compensates for local regulatory inhibition of metabolic gene expression to optimize hormone production and prevent adverse impacts on reproduction.

Interactive diagrams and ontology: tools to capture and explore pathway data at Rat Genome Database

Victoria Petri, Mary Shimoyama, Jeff dePons, Simon Twigger, Melinda Dwinell, Howard Jacob

Rat Genome Database, Human and Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA

Essential to understanding how complex biological systems work and why they sometimes malfunction is to delineate the role gene products play within a network or pathway and the ways in which these circuits connect to and impact upon each other. In order to address these issues we have developed a Pathway Ontology (PW) and started providing interactive pathway diagrams.

The ontology has been designed to provide: a) a means to standardized annotations of genes to pathway terms, including altered and disease pathways; b) a platform to illustrate the relationships and dependencies between pathway types. The four major categories or nodes are: metabolic, regulatory, signaling and disease pathway; the ontology also accommodates altered version(s) of pathways. Pathway data from multiple sources (KEGG, Reactome, Metacyc, GenMapp) have been integrated into the ontology. Pathway data as described in the scientific literature and captured in reviews have contributed to its continuous development in terms of both informational content and structural design. The literature is used to identify the components of a selected pathway and the rat and homologous human and mouse genes are annotated to the term. As applicable, certain genes may also be annotated to 'altered' pathway terms. In addition to providing for the standardized annotations of genes, the structured nature of the ontology allows for easy data mining, presentation and integration with various tools. The Ontology Browser can be used to explore the ontology. The Ontology Report for a selected pathway provides the pathway description, list of all genes annotated to the term with links to the individual gene reports, a Genome-wide View (GViewer) of the genes in the pathway across chromosomes with data download option, and if available, a link to the interactive diagram.

The interactive diagrams provide another venue for representing as well as accessing and exploring pathway data. The diagrams are created using the Ariadne Genomics Pathway Studio software tool. The emphasis is on representing various signaling pathway. Diagrams offer an instant 'snapshot' of the molecules involved and the relationships among them. They also capture the roadmap of the conduits through which the signals may travel, the pathways that may be triggered. This adds a new dimension to the pathway data offered, beyond what the relationships in the ontology can provide. But, the one-to-one relationship between names in the roadmap and terms in the ontology allows for easy link(s) between graphics and reports. Very importantly, it offers a unique ability to 'walk' from one diagram to another and to explore pathway connectivity. The diagrams provide a description of the pathway and acronyms, links to references, legends for figure elements, links to gene reports.

Future goals include: continue the development of the ontology, particularly with respect to cross-products as well as to altered and disease pathways; continue to build interactive diagrams, particularly with respect to expanding pathway connectivity to deepen and enrich the unique 'walk' through the pathway landscape feature; address the issue of representing altered pathways.

Metamotif - a novel probabilistic framework for describing patterns in sequence motif families

Matias Piipari[1], Thomas Down[2], Tim Hubbard[1]

[1]Wellcome Trust Sanger Institute, Cambridge, Hinxton, UK; [2]Gurdon Institute, University of Cambridge, UK

Understanding genome regulation is one of the most important goals of modern molecular biology. This goal is being made increasingly more tractable through the availability of numerous model organism genome sequences, associated large scale gene expression and protein–DNA interaction data, and advances in computational methodologies. Two key steps involved are 1) identifying complete sets of transcription factors and 2) finding their specific DNA binding site motifs. We have previously shown that large scale computational motif discovery is possible [1]. Our aim with this work is to improve the discovery and categorisation of DNA binding site motifs of regulatory proteins through a novel probabilistic framework that describes sequence motif families through a generative model.

Development of motif comparison methods is important if computationally discovered motifs are to be used in e.g. automated genome annotation or regulatory module detection. Analysis of motif discovery experiments introduces many unsolved problems: e.g. finding duplicate motifs from large motif sets is difficult, or finding ones that are closely associated together in sequences of interest. Finding such links between motifs is important not only for linking motif data with putative function, but also for guiding further motif discovery experiments. Both unsupervised and supervised machine learning methods have been applied for the problem of comparison and clustering of motif, as well as for informing motif discovery algorithms of known motifs.

We describe here a Markov chain Monte Carlo (MCMC) method for discovering over-represented patterns in sequence motif sets represented as weight matrices, a model which we call a 'metamotif'. In short, a metamotif is an alignment model for motifs where each independent column is represented by a Dirichlet distribution, and models are estimated using an MCMC method called nested sampling. Such a model yields naturally a probabilistic distance metric for scoring motifs for clustering and classification purposes. The metamotif distance metric is applied to the classification problem of assigning sequence motifs to their corresponding DNA binding domain families and we find that our method outperforms previously described classification methods [2]. We have also extended the NestedMICA sequence motif discovery algorithm [1] to allow the use of metamotif models as a motif prior in large motif discovery experiments. Results of simulations showing promising increase in sensitivity in detecting low-frequency sequence motifs from large sequence sets are presented.

[1] ***Down et al.*** Large-Scale Discovery of Promoter Motifs in Drosophila melanogaster. PLoS Comput Biol (2007) vol. 3 (1) pp. e7

[2] ***Narlikar et al.*** Sequence features of DNA binding sites reveal structural class of associated transcription factor. Bioinformatics (2006) vol. 22 (2) pp. 157-163

Identifying regulatory genetic variants underlying dominance expression patterns in Arabidopsis thaliana by module-based probabilistic graphical models

Nathalie Pochet[1], Hilde Van den Daele[1], Steven Maere[1], Heidi Wouters[1], Fanghong Zhang[1], Tom Michoel[1], Eric Bonnet[1], Anagha Joshi[1], Yves Van de Peer[1], Marnik Vuylsteke[1]

[1]Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.

Integrated analysis of DNA sequence variants and phenotypic differences at the level of gene expression allows for unraveling molecular pathways underlying phenotypic variation across individuals. The intermediate position of gene expression between genotype and organismal phenotype makes it ideally suited to serve as a bridge between heterozygosity and heterosis. Ultimately, it will be necessary to include variation in proteins and metabolites in the analysis as well. At present, however, genetic dissection of transcript abundance is the emerging approach that sheds light on the genetic complexity of transcript levels. Methods such as Geronemo and LeMoNe aim for deciphering both the cell's regulatory network and perturbations to it resulting from sequence variability. Variations in both sequence and expression of regulators are considered to act on sets of co-regulated genes (modules). Modularity of genetic regulatory systems allows recovering complex combinatorial regulation. Incorporating both expression and genotype of regulators enables capturing cases where the effect of sequence variation on its targets is indirect.

Our aim is to identify regulatory mechanisms underlying heterosis in *Arabidopsis thaliana*. For that purpose, ten (homozygous) accessions were used as parents and crossed in a full diallel design to produce 45 F1 hybrids. For all 55 individuals we measured gene expression using the 44K Agilent microarray. Genetic marker information (SNP) for the ten accessions was acquired through public databases and F1 hybrid genotypes were derived from parental genotypes. Using the original formulation, however, both Geronemo and LeMoNe were incapable of adequately capturing information acquired from a diallel design. We therefore extended LeMoNe to also incorporate diallel information comprising more than two genotype states, and subsequently applied this method within the context of heterosis in *Arabidopsis thaliana*.

Modeling and construction of an epigenetic synthetic 'toggle switch' in mammalian cells based on RNA interference.

Cuccato Giulia[1], Siciliano Velia[1], Polynikis Athanasios [2], Marucci Lucia [1,3], Hogan S. John [2], di Bernardo Mario[2,3], di Bernardo Diego[1,3]

[1]Telethon Institute of Genetics and Medicine, Via P. Castellino 111, 80131 Naples, Italy; [2] Department of Engineering Mathematics, University of Bristol, University Walk, Bristol BS8 1TR, United Kingdom; [3] Department of Systems and Computer Science, University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy.

This project is aimed at building a synthetic circuit in mammalian cells acting as a "toggle switch" for in vivo delivery of mRNA/protein. This synthetic network enables to "switch on" and "off" expression of a protein at will, without the need of external inducer molecules, but transiently.

A plasmid containing the network has been constructed. It comprises a transcription factor (TF) that inhibits transcription of a shRNA silencing the TF itself. This reciprocal inhibition enables the toggle to be in one of two possible states, high TF and low shRNA, or *vice-versa*. The switch between the two possible states is obtained by the transient introduction of an external inducer of the currently active repressor. Once the network has switched, the new epigenetic state is maintained indefinitely or until the application of the other inducer. We are using a lentiviral backbone, pLVUTH, containing the doxycycline inducible Tet-repressor-KRAB (TET-ON) under the control of the Ubiquitin promoter. The vector also includes a TRE-H1 promoter upstream of an aptamer-fused shRNA (silencing the TTR mRNA) controlled by the TET system. The switch transition is provoked by theophylline that inhibits the aptamer-fused shRNAs or doxycycline that inhibits the TTR protein.

The use of an aptamer fused shRNA responsive to theophyllline to silence transcription is a major novelty which also allows for the whole circuit to be contained in a single lentiviral vector. The same vector is utilized throughout the study. It allows testing the circuit on stably integrated primary cell and later animal models. So far, both the inducible systems have been tested with positive results and the virus expressing the circuit has been produced in Hek293T cells. The bistability of the network is currently being tested.

In order to investigate design constraints and parameters, using the preliminary experimental results, we derived a semi-quantitative model based on ordinary differential equations. The mathematical variables of this model are the concentrations of the TTR-KRAB fusion protein, the TTR-KRAB mRNA and the shRNA. Transcription was modeled using Hill kinetics and translation was modeled as a first order process. The degradation of TTR-KRAB mRNA by the shRNA has been modeled in the literature with two different functional forms, a Michaelis-Menten like term (Khanin et al, 2008) and a linear one (Malphettes et al 2006), but with limited experimental validation. Interestingly, using only one of the two forms, the circuit is predicted to behave as a bistable switch. To clarify the issue, we treated TTR-KRAB expressing Hek293 cells with varying concentrations of siRNA oligomers directed against the TTR-KRAB mRNA. We measured mRNA levels by qPCR. We then fitted the two functional forms to the data via a generalized nonlinear regression with bootstrapping, and found the functional form that best explains the data. The resulting mathematical model shows that the circuit can behave as a bistable switch. Additionally, the mathematical model captures well the qualitative behavior observed in the experiments so far.

A biophysical model inferred from ChIP-seq data accurately predicts TF-DNA binding

Wenjie Sun[1], Michael Lim[1], Linus Lee[1], Shyam Prabhakar[1]

[1] *Computational & Mathematical Biology, Genome Instt. of Singapore, Singapore 138672*

Knowledge of the strengths and locations of transcription factor (TF) binding sites in the genome is central to our understanding of transcriptional regulation. There has therefore been great interest in computational prediction of TF-DNA binding. Traditionally however, computational predictions have relied upon approximate, non-physical models that arbitrarily partition genomic regions into “bound” and “unbound” sets. Moreover, the set of bound sites used to construct motif models is often small (~10-30 sites compiled from locus-specific studies) or of low spatial resolution (ChIP-chip assay). In order to obtain a more precise and physically realistic model of TF-DNA binding affinity, we developed an algorithm that exploits the abundance (~1,000-10,000 sites) and spatial resolution of ChIP-seq data to directly estimate the position-specific binding energy matrix 👎👎G_{ij}, as well as the intranuclear TF concentration in units of the dissociation constant K_d of the highest-affinity binding site. The algorithm, known as COSMIC (**C**hIP-seq-based **O**ptimization of position-**S**pecific affinity **M**atrix and **I**ntranuclear **C**oncentration), uses nonlinear regression to fit the observed distribution of sequence tags in a ChIP-seq assay and yields explicit estimates of TF occupancy at each site. In cross-validation tests on ChIP-seq data for 7 mouse embryonic stem cell TFs COSMIC predicted quantitative TF-DNA binding with accuracy comparable to or greater than that of four widely used motif discovery algorithms, explaining 60% of the power of the tag distribution signal for Esrrb and Klf4 and 51% for STAT3. Moreover, the efficacy of tag distribution analysis was highlighted by identification of a TAAT[G/T][A/G] binding motif for the homeobox TF Nanog that had eluded previous computational analyses of the same data. Intranuclear TF concentrations estimated by COSMIC, though difficult to directly validate, appeared biologically plausible in that they lay within an order of magnitude of K_d. Our results suggest that physically realistic modeling, when combined with the rich information content of ChIP-seq data, yields increased accuracy in predicting TF-DNA binding. Methods such as COSMIC that explicitly predict binding site occupancy can potentially quantify the transcriptional regulatory consequences of human SNPs and assist in identifying disease-causing variants in the noncoding genome.

Posters

An Analysis of the Temporal Regulation of the Vaccinia virus Transcriptome

Anjan Purkayastha[1], Ron Golan[1], Judy Yen[1], Kathleen Rubins[1]

1 Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142 US

The Poxviridae family consists of double-stranded DNA viruses that replicate exclusively in the host cell cytoplasm. They coordinate their DNA replication and assembly by regulating the timing of expression of their genes. Poxvirus gene expression takes place in three transcriptional waves- Early, Intermediate and Late. However, a number of issues remain unresolved. A systematic study that characterizes the transcriptional profile of each Poxvirus gene needs to be done. It is entirely possible that the temporal regulation that has been reported in a handful of Poxvirus genes provides a biased view of the total regulatory process. The role transcriptional regulation may play in determining virulence and host tropism has not been investigated. Finally, the viral and host factors involved in this regulatory process remain unidentified. We have used a Vaccinia virus whole genome tiling array that provides an unbiased, nucleotide-resolution view of transcription to address these issues. Our initial experiments have provided the transcription profiles for many Poxvirus genes. They also suggest that while the regulatory process may be divided into Early, Intermediate and Late phases, the genes are very heterogeneous, both in terms of the strength and the exact timing of their transcription. Follow up experiments will be done to provide a more refined view of the kinetics of transcription; to identify possible regulatory motifs in the promoters of co-regulated genes and to identify the viral and host factors that control the transcription of the Poxvirus genes.

Bayesian Co-clustering of Heterogeneous Data Types

Saumyadipta Pyne[1], Roee Gutman[2] and Nicholas A. Heard[3]

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, USA; [2]Department of Statistics, Harvard University, Cambridge, USA; [3]Department of Mathematics, Imperial College, London, UK.

Clustering algorithms help us to identify groups of features, such as genes, with similar expression in high-throughput data. Often data are generated from different platforms, and therefore measurements for the same feature across independent experiments may not be comparable directly. Further, the differences in reproducibility and conditions of experiments might require that such data be analyzed separately.

However, given data of heterogeneous types, most clustering algorithms, which were originally designed to work with data from a single experiment, either merge the datasets undermining their heterogeneity or rely on post hoc meta-analysis of clustering results from individual experiments. Limitations in either approach can have serious impact on the overall clustering results especially since multi-platform investigation is becoming increasingly common.

We built upon earlier work [1] to develop an algorithm for Bayesian co-clustering of experiments with heterogeneous data types. Based on a fully probabilistic linear regression model, it allows us to integrate diverse types of data: instead of coerced concatenation, the experiments are represented individually by suitable basis functions. Consensus across types is modeled to allow for clustering of features that are similar in many but not necessarily all of the experiments. The consensus also allows for statistically more powerful and nuanced clustering of large number of features, in particular for identification of features with weak but generally consistent signals across experiments.

Regardless of the nature of the basis functions used for individual data types (e.g. orthonormalized Fourier bases for periodic time course data), a common mean expression profile specific to each cluster is determined. The algorithm has the computational advantage of using conjugate priors for the regression parameters in the linear model framework which makes it efficient. By extending the regression model to be a mixture of regressions based on the cluster-experiment error variance, the algorithm can allow some experiments to be more influential than others in each cluster, which is often useful for multi-platform data analysis.

We ran the algorithm on several genome-wide datasets across different types and platforms, and identified functionally homogeneous biological clusters (in fact, it performed better than many well known clustering algorithms) leading to classification of a large number of previously uncharacterized genes and proteins, thereby allowing genome-wide discovery of regulatory motifs, modules and networks in fungal species.

[1] Heard N.A. et al. *Proc. Nat. Acad. Sci.* USA 102, 16939-16944, 2005

Reducing the Computational Complexity of Information Theoretic Approaches for Reconstructing Gene Regulatory Networks

Peng Qiu, Andrew J. Gentles, and Sylvia K. Plevritis

Department of Radiology, Stanford University, CA, USA

Information theoretic approaches are increasingly being used for reconstructing gene regulatory networks from gene expression microarray data. Examples include the relevance network (RelNet), ARACNE, and the maximum relevance minimum redundance network (MRNet). These specific approaches and others start by computing the pairwise mutual information between all possible pairs of genes, resulting in a mutual information matrix. The resulting mutual information matrix is then manipulated to identify regulatory relationships; for example, thresholding in RelNet, the data processing inequality in ARACNE, and the maximum relevance minimum redundance criterion in MRNet. These approaches have been successfully applied to simulated data and real microarray data for identifying regulatory targets and pathways. The most time-consuming step in these approaches is computing the mutual information matrix. For example, in a recently analyzed B-cell data set consisting of 336 samples and 9563 genes per sample (Basso et al, *Nature Genetics* 2005), ARACNE takes about 142 hours to compute the mutual information matrix. In this work, we present two independent methods to reduce the computation time needed to compute the mutual information matrix.

We use spectral analysis to re-order the genes, such that genes that share regulatory relationship are more likely to be close to each other. We then apply a sliding window to examine subsets of the re-ordered genes, where the number of subsets depends on the total number of genes and the size of the sliding window. Since we compute the mutual information only among genes within the sliding window, only part of the mutual information matrix is computed. Depending on the window size, the computational complexity can be significantly reduced. Although only part of the mutual information matrix is computed, as long as connected genes are close to each other after the re-ordering, the reconstruction performance will not decrease much. With simulated data, we observed that our approach does not incur performance loss in operating regions of high precision and low recall. Using the above mentioned B-Cell data set, we show that with a small amount of reconstruction loss (<10%), our method reduces the computation time by 84%, which is from 142 hours to 23 hours.

We also developed a fast software implementation to calculate of the pairwise mutual information matrix based on kernel estimation. We noticed that the calculation of pairwise mutual information involves a few nested loops, which contain a lot of repeated operations. The essential idea of our fast implementation is to switch the order of the nested loops, so that the repeated operations can be pushed out of some of them. For the above example, where ARACNE takes 142 hours to compute the mutual information between all gene-pairs, our fast implementation requires only 1.6 hours.

Computational prediction of RNA structural motifs involved in post transcriptional regulatory processes

Michal Rabani, Michael Kertesz, and Eran Segal

Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel

mRNA molecules are tightly regulated, mostly through interactions with proteins and other RNAs, but the mechanisms that confer the specificity of such interactions are poorly understood. It is clear, however, that this specificity is determined by both the nucleotide sequence and secondary structure of the mRNA. Here, we develop RNApromo, an efficient computational tool for identifying structural elements within mRNAs that are involved in specifying post-transcriptional regulations. By analyzing experimental data on mRNA decay rates, we identify common structural elements in fast-decaying and slow-decaying mRNAs, and link them with binding preferences of several RNA binding proteins. We also predict structural elements in sets of mRNAs with common sub-cellular localization in mouse neurons and fly embryos. Finally, by analyzing pre-microRNA stem-loops, we identify structural differences between pre-microRNAs of animals and plants, which provide insights into the mechanism of microRNA biogenesis. Together, our results reveal unexplored layers of post-transcriptional regulations in groups of RNAs, and are therefore an important step towards a better understanding of the regulatory information conveyed within RNA molecules. Our new RNA motif discovery tool is available online.

Analysis of Co-evolution in Drosophila regulatory genome

Pradipta Ray[1] and Eric P. Xing[1]

[1]*School of Computer Science, Carnegie Mellon University*

Many biological processes such as embryogenesis and development are surprisingly robust to genetic perturbations, such as DNA mutations in the genome. For example, the genetic circuitry underlying the establishment of morphogenic landmarks in an early embryo, e.g., the polarity, gradient, and segmental patterns of key morphogens that determine body plan and organ formation during ontogenesis, appears to have inherent properties to dampen genetic noise in the relevant regulatory elements up to some threshold. A possible mechanism that explains the molecular and evolutionary basis of biological robustness is the *co-evolution* of potentially functionally compensating elements in the regulatory system in a way that ensures phenotypes to be buffered against these genetic perturbations. These concealed variations may provide the chance for further molecular evolution and selection via enabling complex phenotypic changes only possible under the synergetic effect of accumulated concealed genetic changes.Little progress has been made in unraveling and understanding the dynamics and mechanisms of phenotypically buffered co-evolution of regulatory elements. In this work, we report an analysis of co-evolution of such elements in the cis-regulatory modules (CRMs) of Drosophila early developmental genes across 12 sequenced *Drosophila* species.

Using a new algorithm based on spectrum clustering of site-specific or region-specific molecular phylogenies, we have identified multiple co-evolving regions within the CRMs of Drosophila developmental genes that display similar special expression patterns across species, and regulatory elements possibly working in tandem as in the case of adaptive or compensatory evolution; we have also detected possible co-evolution between transcription factors (Tfs) and their binding sites (TFBSs); and we have found long-range dependencies of evolution in regulatory regions and functional units of different regulatory genes in the developmental network. Interestingly, we find that most TFBSs are tightly co-evolving, even when they are far apart spatially on the regulatory sequence, based the highly correlated evolutionary rates and common topology of their respective phylogenetic trees. We also found rate correlations with strong statistical significance ($p <=0.05$) when DNA binding domains of the TFs are analyzed along with TFBSs, but we do not find such strong signals when the entire TF is analyzed along with the TFBSs in question.

Our work is important in three major respects: (1) we model co-evolution in the regulatory genome at the sequence level of various functional granularities rather than expression or other levels as previously attempted [1], (2) we present a holistic view of evolution by analyzing other evolutionary phenomena like selection and duplication in parallel, and (3) we introduce a novel tree-spectral-clustering (TSC) algorithm for co-evolution analysis, which builds on a similarity function using the Felsenstein-Kuhner metric on phylogenies. TSC can naturally employ other co-evolution metrics derived from alignments or other sources, and therefore is easily generalizable to other data types and modeling assumptions/methods.

[1] Hunter Fraser, Aaron Hirsh, Dennis Wall and Michael Eisen. (2004). `Coevolution of gene expression among interacting proteins.' *Proc Natl Acad Sci* **101**(24):9033-9038.

Global structure inference of the transcription regulatory network in *S.cerevisiae*

Jüri Reimand[1,2], Jaak Vilo[2], Nicholas Luscombe[1]

[1] EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD Cambridge, United Kingdom; [2] University of Tartu, Institute of Computer Science, Liivi 2, 50409 Tartu, Estonia

Dissecting transcription regulatory networks is a great challenge of contemporary systems biology. High-throughput measurements from gene perturbation experiments provide useful clues for the above task, as resulting data reveal downstream targets that are certainly affected by the gene in question. However, naïve reconstruction of regulatory relations from gene perturbations provides only a unilateral and noisy view of the network. Advanced computational techniques of probabilistic modeling may aid in tackling largely unanswered questions such as cascades of direct and indirect regulatory targets, combinatorial regulation via multiple transcription factors, and transcriptional compensation mechanisms.

We have carefully reprocessed a recently published transcription factor (TF) knockout microarray collection for *S.cerevisiae* and are combining the data with putative TF binding sites (TFBS), ChIP-chip binding, protein-protein interactions and chromatin state measurements to construct a high-confidence global regulatory network. We have developed a Bayesian network structure inference engine that treats differentially expressed genes in knockout mutants as independent modules and applies Markov Chain Monte Carlo (MCMC) sampling with epistatic constraints to compose a consensus network from many high-scoring candidates. The network reconstruction process takes advantage of structural priors that increase confidence in regulatory edges that are supported in previously published experimental data. Densely connected substructures of the resulting network reflect underlying biological functions as the genes in such modules are well enriched in various Gene Ontology categories and KEGG and Reactome pathways. In addition to global functional profiling of network modules, we predict transcriptional regulators for less-studied processes such as quiescence (G0), DNA damage response and stress response.

Prediction of Gene Phenotypes in Humans Using Phenotype-Specific Transcription Modules

Michael R. Mehan[1], Juan Nunez Iglesias[1], Chao Dai[1], Xianghong Jasmine Zhou[1]

[1]Program in Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles CA 90089, USA

Human phenotypes such as complex diseases, are often caused by multiple mutations, each of which contributes only a minor effect to the phenotype. Traditional association studies are known to have difficulties in discovering gene-gene interactions contributing to complex phenotypes. In this project, we developed a novel network-based approach, based on Multiple Objective Simulated Annealing, to systematically map transcription modules to diverse phenotypes. Specifically, we integrated 338 microarray datasets, covering 178 phenotype classes in the Unified Medical Language System (UMLS). We identified approximately 200,000 gene modules coexpressed only in datasets studying specific phenotypes. The study of these modules can provide a deeper understanding of the genetic bases for complex phenotypes not possible with traditional disease mapping techniques. We used these modules to make novel gene phenotype predictions, discover pleiotropic genes, and construct phenotype-specific regulatory networks. We detail these three applications below.

Novel gene phenotype prediction: Given a phenotype-specific co-expression module, we predicted whether previously unannotated genes are associated with the phenotype by considering two aspects of the module: the strength of the phenotype-specificity, and the phenotype annotations of other genes in the module. Integrating this information with protein-protein interaction data and annotations from Gene Ontology, we predicted 3,642 novel gene-phenotype associations covering 1,952 genes. These predictions could be extremely useful in a clinical setting to further our understanding of disease pathways, as well as serve as potential drug targets.

Pleiotropic gene discovery: From our phenotype predictions, we can determine if a gene is associated with two or more distinct phenotypes, and therefore constitutes an example of pleiotropy. We discovered that 47% of the genes in our study are known to affect at least two unrelated phenotypes. The contribution of our 3,642 gene-phenotype predictions resulted in 308 novel pleiotropic genes that previously had at most one associated phenotype. An example is the gene GMFG, which is known to be associated with “Dysmyelopoietic Syndromes.” We predicted that it is also involved in “Head and Neck Neoplasms,” which is consistent with its function of being a nerve growth factor. This analysis indicates that pleiotropy may be a more widespread phenomenon than previously believed.

Phenotype specific regulatory network reconstruction: ChIP-chip experiments provide binding data derived under particular conditions, representing a static snapshot of a dynamic, phenotype-dependent network. In conjunction with our phenotype-specific coexpression modules, we begin to reconstruct these dynamic regulatory networks. Our results show that bound genes are more likely to form coexpression modules in datasets whose phenotype annotations match those of the ChIP-chip experiment. Out of 67 matching phenotypes, 38 exhibited a statistically significant preferential binding. This validates the accuracy of our phenotype specific-modules, as well as highlights the importance of considering ChIP-chip data as a phenotype-specific resource.

Comparative Analysis of Enhancers and Regulatory Motifs for Gene Expression in the Vertebrate Brain

Qiang Li[1] , Deborah Ritter[2], Nan Yang[1], Zhigiang Dong[1], Hao Li[3], Jeffrey H. Chuang[2], Su Guo[1]

[1]Department of Biopharmaceutical Sciences, UCSF; [2]Department of Biology, Boston College; [3]Department of Biochemistry and Biophysics, UCSF;

The vertebrate genome contains a vast amount of non-coding sequence with strong conservation across the vertebrate phylogeny. Deciphering the functions and mechanisms of action of these non-coding sequences is a challenging, but important, problem. Many Conserved Noncoding Elements (CNEs) are thought to regulate vertebrate gene function as transcriptional enhancers, capable of controlling gene expression either spatially and temporally. However, the tissue- or timing- specificity of CNE enhancers is not known *a priori*.

Here we report a combined computational/experimental strategy in which we used the developing vertebrate brain as an example and identified pattern-associated CNEs conserved between human and zebrafish. By selecting CNEs adjacent to genes of known developmental expression pattern, we were able to efficiently identify CNEs with brain enhancer activity in a zebrafish reporter system. These enhancer experiments comprise studies of more than 100 zebrafish CNEs. Application of *de novo* motif prediction algorithms on forebrain enhancers uncovered short sequences that were experimentally validated as critical codes for forebrain enhancer activity. In addition, we assessed cross-species expression of orthologous CNE sequences from zebrafish, human and mouse, via experiments in zebrafish and mouse embryos. A significant fraction of CNE sequences exhibited divergent tissue-specific enhancer activity across species, suggesting that positive selection and host-specific effects are common even for these highly conserved sequences.

We have built a database of experimental images and annotations on all CNEs we have tested, which we will continue to develop as a public resource for zebrafish CNE studies (cneBrowser). In addition, we have created an online tool (cneViewer – cneviewer.zebrafishcne.org) to simplify the process of selecting CNEs for experimental validation based on the anatomical and timing-specific expression of nearby genes. These websites provide user-friendly resources to facilitate information transfer and analysis for deciphering the noncoding functions in vertebrate genomes.

Comparative genomic reconstruction of transcriptional regulons in the Shewanella genus

Dmitry A. Rodionov[1,2], Pavel S. Novichkov[3], Mikhail S. Gelfand[2], Andrei L. Osterman[1]

[1]Burnham Institute for Medical Research, La Jolla, California; [2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; [3] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California.

Integrative comparative genomics approaches were used to infer transcriptional regulatory networks (TRNs) in 13 *Shewanella* species and a set of other γ-proteobacteria with sequenced genomes. To accomplish this goal, we combined the identification of transcription factors (TFs), TF-binding sites (TFBSs) and cross-genome comparison of regulons with the analysis of the genomic and functional context inferred by metabolic reconstruction. The reconstructed TRNs for the key pathways involved in central metabolism, production of energy and biomass, metal ion homeostasis and stress response provide a framework for the interpretation of gene expression data. This analysis also helps to improve functional annotations and identify previously uncharacterized genes in metabolic pathways. Finally, we attempted to reconstruct possible evolutionary scenarios of these TRNs.

Using comparative genomics approach we identified candidate TFBSs for near eighty TFs from *Shewanella* group. For thirty described regulons, the TF was conserved between *E. coli* and *Shewanella.* These include global regulators (Crp, Fnr, ArcA, Fur, LexA) and specialized regulators of the metabolism of nitrogen (NarP, NsrR, DNR, NorR, NtrC), amino acids (ArgR, MetJ, TrpR, TyrR, HutC, IlvY, MetR), fatty acids (FadR, FabR), carbohydrates (SdaR, PdhR, HexR, GntR), and cofactors (BirA, IscR, ModE). Another fifty regulons described in *Shewanella* spp. are operated by TF that do not have orthologs in *E. coli*. In particularly, we characterized novel regulons that control gene involved in the degradation of branch chain amino acids (named LiuR), degradation of fatty acids (PsrA), and catabolism of various sugars (e.g. NagR, ScrR, AraR, and BglR tentatively implicated in the control of utilization of N-acetylglucosamine, sucrose, arabinose and β-glucosides, respectively).

Although some diversity of the predicted regulons is observed within the collection of *Shewanella* spp., the most striking difference in the overall regulatory strategy is revealed by comparison with *E. coli* and other γ-proteobacteria. Multiple interesting trends in diversification and adaptive evolution of TRNs between lineages were detected including regulon “shrinking”, “expansion”, “mergers”, and “split-ups”, as well as multiple cases of using nonorthologous regulators to control equivalent pathways or orthologous regulators to control distinct pathways.

Within the *Shewanella* lineage, the two major diversification strategies are: constrained (“all or none”), when the regulon is either present or absent in its entirety with tightly conserved regulation of all genes (e.g. for local regulons), and permissive (“loose”), when most genes of a regulon are conserved between genomes, whereas the conservation of respective regulatory sites is much weaker and sometimes not mandatory (e.g. for global regulons). Many aspects of metabolic regulation in *Shewanella* species are substantially different from TRN models that were largely derived from studies in *E. coli*. Among the most notable are the differences in TRNs for the central carbohydrate pathways. In enterobacteria the central carbon metabolism is controlled by catabolic regulators FruR and Crp, whereas *Shewanella* species use two other TFs, HexR and PdhR, for this control. The content and functional role of the Crp regulon is significantly different in these two lineages: the catabolism of carbohydrates and amino acids in enterobacteria, and the anaerobic respiration in *Shewanella* species.

Tissue-specific Sequence Signals in Core Promoters: CpG-islands vs. Transcription Factors

Helge G. Roider, Stefan A. Haas, Thomas Manke, Aditi Kanhere, and Martin Vingron

Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

One major challenge in regulatory genomics is the identification of transcription factors which control the expression of tissue-specific genes. Recent experiments have also highlighted the importance of chromatin modifications and methylation of CpG-islands for tissue-specific gene regulation.

Here we use a novel method (PASTAA) for detecting transcription factors associated with functional categories, which utilizes the prediction of binding affinities of a transcription factor to promoters. This binding strength information is compared to the likelihood of membership of the corresponding genes in the functional category under study. Coherence between the two ranked data sets is seen as an indicator of association between a transcription factor and the category. PASTAA is applied primarily to the determination of transcription factors driving tissue specific expression. We show that PASTAA is capable of recovering many known tissue specifically acting transcription factors and provides novel associations so far not detected by alternative methods.

We used this framework to investigate a large class of different tissues and find that most of the tissue-specific signals from transcription factors are associated with CpG-poor promoters and tend to cluster near the transcription start site. Moreover, those genes which are specifically expressed in certain tissues are also found to have predominantly CpG-poor promoters. In contrast, specifically expressed genes from neuronal tissues and many stem cells tend to have CpG-rich promoters. For these genes we find general transcription factors which are strongly associated with proximal promoter regions.

These findings suggest a fundamental separation of the regulatory mechanisms at CpG-poor and CpG-rich promoters - through either transcription factor binding or chromatin modifications.

A relational framework for predicting tissues and links in the Drosophila regulatory network.

Sushmita Roy[1], Alexander Stark[2,4], Pouya Kheradpour[2], Manolis Kellis[2], Margaret Werner-Washburne[3], Terran Lane[1]

[1]UNM Computer Science; [2]MIT Computer Science & Artificial Intelligence Lab; [3]UNM Biology; [4]Institute of Molecular Pathology.

Background: A systematic understanding of gene regulation requires a dynamic view of regulatory networks, which includes both regulatory links between transcription factors (TFs) and their targets, and condition-specific expression information for each gene. Unfortunately however, both types of information are incomplete in practice, due to methodological limitations and biological noise. Classification approaches that exploit partial information from each domain to make predictions about the other can provide an effective and efficient way to jointly predict both regulatory links as well as tissue-specific gene expression.
Approach: We describe a joint prediction framework for predicting tissue-specific gene expression and regulatory links by integrating partial data from a computationally-defined regulatory network (Kheradpour *et al*., Genome Res. 2007), and tissue-specific gene expression (Tomancak *et al.,* Genome Biol. 2002). Our “tissue predictor” for tissue-specific gene expression is a relational classifier that incorporates information from neighboring TF and target genes (Lu & Getoor, ICML 2003), in contrast to a non-relational classifier where tissues of one gene are predicted independent of other genes. Our ”link predictor” for regulatory links between TFs and target genes is a non-relational classifier, using tissue expression of the TF and all its putative target genes. We evaluate performance on a previously unseen test set, based on our ability to predict the tissue-specific gene expression and TF-target regulatory links, using an iterative approach alternating between tissue and link prediction.
Results: We compared the performance of our relational tissue classifier against a non-relational classifier, using the *Drosophila* Image ontology tissues, while assuming regulatory links are fixed. The relational classifier outperformed the non-relational classifier on 73 of 187 tissues, but was outperformed in 25 tissues. It also had significantly higher average scores ($P<10^{-7}$, t-test). We obtained similar results using a variety of classifiers (Maximum entropy, Naïve Bayes), indicating that our results were not an artifact of the classifier type, and reflected the importance of incorporating gene neighborhood information in tissue prediction.

Tissues with the best prediction rate included head, trunk mesoderm and pharynx. These relied on both generic TFs associated with many tissues (*Kr, dri*, *en*), and specific TFs associated with few tissues (*ems* with trachea). These TFs have known phenotypes in nervous system and pharynx tissues, suggesting that our predictor was correctly exploiting tissue-specific information.

Overall, application of our joint predictor to infer both tissues and links produced better link prediction results than a link predictor with fixed known tissues, when the number of tissues was small (*n*<50). Thus, when prior tissue information is scarce, a joint predictor can provide greater coverage by using neighborhood information, and improving link prediction by simultaneously improving upon the tissue assignment. This represents a significant improvement over methods that require known tissues or links.

Clustering of temporal gene expression data with a Dirichlet and Gaussian processes mixture model

Ana Paula Sales[1], Feng Feng[1], Thomas B Kepler[1,2]

[1]Duke University Laboratory of Computational Immunology, Duke University, Durham, NC 27705; [2]Department of Biostatistics & Bioinformatics and Immunology, Duke University, Durham, NC 27705;

The identification of co-regulation and interaction among genes is a first step toward a comprehensive understanding of cell signaling pathways and consequently of prediction of how cells will respond to novel stimuli, such as vaccine adjuvants. Clustering of microarray data aims exactly at identifying such patterns and has been used up to now with varying degrees of success. There are, however, many difficulties associated with these approaches, including determining the number of clusters and, in the case of longitudinal datasets, being able to account for the correlations among time points and in general having to impose restrictions on the shape of the temporal trajectories.

We have developed an integrated approach to dynamically model and cluster time-course gene expression data across multiple treatments using a Bayesian nonparametric mixture model. The mRNA abundance for any given gene is assumed to follow a continuous trajectory, which is represented as a Gaussian process. We assume that genes that share regulatory elements follow the same underlying trajectories (up to additive constants) and that the number of underlying trajectories is unknown and must be inferred from the data. By modeling the data as a mixture of components where the data follows a Gaussian process and the mixture proportions follow a Dirichlet process, we are able to identify clusters of genes with similar expression patterns while benefiting from several advantages. First, by using a countably infinite number of clusters we bypass the need to pre-specify the number of clusters, which can introduce a great deal of bias to the process. In addition, because we obtain posterior samples from the model specified by this mixture we are able to provide confidence measures for each cluster individually as well as to the number of clusters, which is typically not accounted for when the number of clusters is determined in a stage separate from the clustering. Finally, the modeling of the trajectories as Gaussian process provides a dynamic representation of the trajectories since it defines a probability distribution over functions, such that one is not forced to impose any particular parametric form for the time trajectory.

We demonstrate the utility of this method by applying it to a four-treatment eight-point time-course gene expression microarray dataset generated by stimulating dendritic cell cultures with vaccine adjuvant compounds. Using our model, we are able to recover known and novel aspects of the signaling pathway triggered by these stimuli. This approach is general and can be applied both short and long- time course data, and to any number of treatments.

Dissecting the dynamics of yeast transcription factors

Raja Jothi[1,¶, *], S Balaji[2,¶], Arthur Wuster[3], Joshua A. Grochow[4], Jorg Gsponer[3], Teresa M. Przytycka[2], L Aravind[2], M. Madan Babu[3,*]

[1]Laboratory of Molecular Immunology, National Heart Lung and Blookd Institute (NHLBI); [2]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20892, USA; [3]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH,UK; [4]Department of Computer Science, University of Chicago, Chicago, IL 60637, USA; *[¶]Equal contribution; *Correspondence: jothi@mail.nih.gov, madanm@mrc-lmb.cam.ac.uk*

There has been a surge of interest in uncovering the organizing principles underlying biological networks. While advances in this direction have largely focused on understanding how topological properties and network organization characterize complex biological systems, the genome-scale dynamics of the underlying entities (genes, mRNA and proteins) and their role in systems behavior remain unclear.

Here, we first classify transcription factors (TFs) in the yeast transcription network into three mutually exclusive hierarchical layers (top, core, and bottom) using a novel *vertex sort* algorithm. Second, by integrating diverse molecular datasets with the inferred hierarchical structure, we show that the top-, core-, and bottom-layer TFs have strikingly distinct dynamic properties that are characteristic of their position in the hierarchy. Our analysis reveals that the top-layer TFs are present in higher abundance and have a much longer protein half-life although their transcript abundance and degradation rates are similar to those of core- and bottom-layer TFs. We also find that the top-layer TFs display more noise (variability) in protein levels compared to core- and bottom-layer TFs. These results suggest that the core- and the bottom-layer TFs are more tightly regulated at the post-transcriptional level rather than at the transcriptional level itself. The high variability in the expression of top-level TFs in a population of cells may confer a selective advantage as this permits at least some members in a population to respond to changing conditions. Analogously, low noise in the other two layers may minimize noise propagation and ensure fidelity in regulation of relevant target genes.

Taken together, our findings suggest that the interplay between the inherent hierarchy of the network and the dynamics of the TFs are critical to making the transcription network both robust and adaptable, and permits differential utilization of the same underlying network in distinct members of a population. The results presented here have implications in synthetic biology experiments aimed at engineering regulatory networks.

HDAC Inhibitors Reverse CpG Methylation and Gene Silencing Through Regulation of ERK Phosphorylation and DNMT1 Expression

Sibaji Sarkar, Ana Abujamra, Lora Forman and Douglas V. Faller

Cancer Research Center, Department of Medicine, Boston University School of Medicine

Boston 20018

Methylation of CpG repeats in the upstream/promoter regions of genes is an established mechanism of gene silencing and is correlated with the silencing of critical genes, including tumor suppressors, in many types of cancers. Gene silencing by methylation is also prevalent in stem and neural cells. Though this phenomenon appears widely operative, the regulation of this mechanism of gene silencing is not understood. DNA methylation results in the recruitment of HDACs (histone deacetylases) to promoter regions, eventually inhibiting the binding of RNA polymerase II and thereby repressing expression of genes and their products, leading to the paradigm that DNA methylation lies upstream of histone acetylation in this pathway of transcriptional repression. General inhibitors of class I and II HDACs (HDACi), such as sodium butyrate and SAHA, suppress the growth of prostate cancer cells *in vitro* and *in vivo*. Higher concentrations and longer exposures to HDACi cause cell cycle arrest and apoptosis. We report that exposure to HDACi reverse promoter methylation of three silenced tumor suppressor genes retinoic acid receptor beta2 [RARbeta2)], and the cyclin-dependent kinase inhibitors p16 and p21, in prostate cancer cells. The epigenomic silencing of RARbeta2 was also reversed by exposure to these HDACi, as indicated by the induction of its transcript. DNA methylation is maintained by DNA N-methyl transferases (DNMTs). No changes in the expression of any DNMT transcripts were observed after HDACi exposure, but DNMT1 protein levels were markedly down-regulated post-translationally, possibly by increased degradation through a proteosomal pathway. To determine the mechanism of regulation of DNA methylation by HDAC inhibitors, we investigated the activity of mitogenic kinases, which may regulate DNMT1 levels. HDAC inhibitors suppressed basal ERK activity (phosphorylation) in LNCaP cells, but it did not alter Akt and p38 MAP kinase phosphorylation. Silencing of MEK 1 and 2, the upstream activating kinases of ERK, by SiRNA expression abolished ERK phosphorylation and decreased DNMT1 protein levels. Furthermore, the basal hypermethylation of the promoters of the RARbeta2, p21 and p16 tumor suppressors was reversed under these conditions, suggesting that ERK activation is necessary for their methylation, through regulation of DNMT1. Exposure to the MEK inhibitor PD 98059 similar suppressed DNMT1 expression and promoter methylation. Suppression of DNMT1 levels by SiRNA also inhibited methylation of these tumor suppressor genes, without inhibition ERK phosphorylation and expression, demonstrating that repression of DNMT1 is sufficient to reverse their methylated state. Collectively, these data suggests that ERK activity regulates DNMT1 levels and that HDACi, by inhibiting ERK phosphorylation, regulate DNMT1 stability and ultimately DNA methylation. The reversal of DNA methylation of gene promoters and gene silencing by HDAC inhibitors demonstrated here thus identifies a new level of control over genome methylation and transcriptional regulation and silencing, and suggests that promoter acetylation may influence methylation, in addition to the established and reciprocal ability of CpG methylation to influence HDAC recruitment and transcriptional repression.

Transcription factor binding within repetitive sequence elements

Christoph D. Schmid[1], Philipp Bucher[1,2]

[1]*Swiss Institute of Bioinformatics and* [2]*Swiss Institute for Experimental Cancer Research, Swiss Federal Institute of Technology Lausanne (EPFL), 1066 Epalinges, Switzerland*

The binding of transcription factor proteins to specific loci on genomic sequences is involved in the regulation of gene expression. A recently introduced combination of Chromatin immunoprecipitation and novel mass-sequencing methods (ChIP-seq) enables the precise and quantitative evaluation of DNA-protein interactions *in vivo* in a genome-wide scale. The method relies on short (~30bp) sequence reads with a unique occurrence in the genome.

Here we report the results of a re-analysis of ChIP-seq data (Robertson et al., 2007) including the refinement of a previous motif description of STAT1 binding sites based on this novel kind of data. The close similarity of the corresponding motifs derived from the present approach or from an *in vitro* SELEX assay (Ehret et al., 2001) indicates very comparable STAT1 binding specificities *in vivo* and *in vitro*. Accordingly the majority of genomic loci with a strong occupation in the human HeLa cell line as derived from ChIP-seq signals also feature the STAT1 binding motif. However the inverse holds not true with large numbers of genomic sequences matching or closely resembling a STAT1 binding motif being devoid of significant ChIP-seq signal. This putatively generalized property of DNA-binding proteins suggests that the nucleotide sequence represents a required, but not sufficient factor determining DNA-protein interactions and their functional consequences in gene regulation.

Furthermore we describe an IFN-γ induced binding of the transcription factor STAT1 to repetitive sequence elements of the MER41 family. Repetitive sequences represent almost 50% of the human genome, however functions of these sequence elements remain poorly characterized. Highly repeated occurrences of these elements considerably complicate the functional characterization or even disable the assessment of repetitive genomic regions using hybridization-based approaches. Millions of years of evolution have introduced sufficient diversity in the nucleotide sequence of individual copies of repetitive elements to create a number of unique 30bp sequences within repetitive sequences. Therefore in contrast to ChIP-chip the ChIP-seq readout allows the monitoring of binding of transcription factors at specific sites also within repetitive sequences. The observed regulation of binding of a transcription factor to repetitive sequence elements strongly suggests a role of at least subsets of repeats in the regulation of gene expression.

References:

Ehret, G. B., Reichenbach, P., Schindler, U., Horvath, C. M., Fritz, S., Nabholz, M., and Bucher, P. (2001). DNA binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites. J Biol Chem 276, 6675-6688.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4, 651-657.

The effects of parameter deviations on reactant concentrations in metabolic networks

Eric Kramer[1], Moritz Schütte[2], Oliver Ebenhöh[2,3]

[1]Northwestern University, 633 Clark Street, Evanston, IL 6020; [2]Max-Planck Institute of Molecular Plant Physiology, Potsdam-Golm, 14476, Germany; [3]Potsdam University, Institute of Biochemistry and Biology, Karl-Liebknecht-Straße 24-25, Potsdam-Golm, 14476, Germany

Recent advances in high-throughput techniques allow for simultaneously measuring a large number of metabolites, so called metabolic profiles. Handling the tremendous mass of data produced by these methods currently poses a major challenge in the field of theoretical biology. In particular, it is extremely difficult to interpret the data within a biological context and to derive molecular mechanisms which are responsible for the observed behavior. A straightforward approach is the construction of a metabolic network from correlation measurements between the metabolites. However, it is not a priori clear whether correlations reflect the underlying reaction system.

Repetition of an experiment is hardly possible under exactly the same conditions. Changes in external parameters such as temperature, pressure or humidity are not entirely excludable. Moreover, enzyme concentrations may vary even among different cells. Thus the parameters in a metabolic network experience slight changes for every repetition. These perturbations produce less apparently significant correlations. To give promising hints on improvements for the experimental setup it is of interest to investigate which parameters result in decisive deviations.

We present a theoretical approach to reversely determine the standard deviation of every particular parameter from covariance measurements of the concentrations of the reactants. Our theory relies on concentration control coefficients. Thus it is restricted by the existence of a mathematical model of the investigated metabolism and knowledge of the mean parameter values. To overcome the latter restriction, we developed a fitting tool using the concept of simulated annealing that yields parameter values for any common kinetics, e.g. Michaelis-Menten, from measured steady-state metabolite concentrations. These assumptions are necessary to calculate the concentration control coefficients which connect the standard deviations with the covariances.

Our theory is best applicable if the inequality $P \leq (R +1)*R/2$ holds true, where P labels the number of parameters and R the number of metabolites. The quadratic dependence makes it especially suitable for data sets comprising a large number of metabolites.

Often there are competing models describing the same metabolism, for example for Glycolysis or the Calvin cycle several models exist. As a future application, we will try to use our theoretic approaches for the validation of different models for which sufficient data is available.

A Probabilistic Approach for Detection and Evaluation of Dependencies in Regulator-Target Gene Pairs

Akdes Serin[1], Martin Vingron[1]

[1] Computational Biology, Max Planck Institute f. Molecular Genetics, Ihnestr. 73, 14195 Berlin

Over the past decade, there has been a great development in high throughput molecular technologies for measuring mRNA levels genome wide, which enable us to obtain gene expression profiles under different experimental conditions. Gene regulatory networks can be inferred from gene expression datasets using computational and statistical methods. However, there are few methods to identify condition specific regulator-target gene pairs and to predict their regulatory behaviors.

Here, we propose a probabilistic approach for detecting and evaluating regulator-target gene pair dependencies under different experimental conditions. Since the most commonly used method of correlation is not able to detect non-linear dependencies between genes, we applied density classification based on possible logical models of a regulator and its target gene. These logical models are defined according to regulators functional behavior (activator/repressor) and experimental condition. We assume that condition does not influence the functional behavior of a regulator but it influences its activity.

Our method takes as input *Saccharomyces cerevisiae* gene expression data set in several different conditions and candidate target genes for a given regulator. With the given input, classification was done according to regulators function and experimental condition. After the comparison of our method to mutual information, which also provides a general measure of non-linear dependencies between variables, we showed that our method provides more information about the functions of the regulators in certain conditions.

Systems Biology Approach To Map DNA Damage Networks In Response to 3-Methyladenine DNA Lesion

Dharini Shah[1,2,3,4], Neil Tritman[2,4], Georgi Marinov[4], Michael Yaffe[1,2,3,4],Douglas Lauffenburger [1,2,3,4] and Leona D. Samson[1,2,3,4]

[1]Department of Biological Engineering; [2]Center for Environmental Health Sciences, [3]Koch Institute for Integrated Cancer Research and [4]Massachusetts Institute of Technology.

3-Methyladenine (3MeA) lesions are highly cytotoxic and relatively non-mutagenic making them clinically relevant for cancer chemotherapy. Tumors that have become resistant to other alkylated bases like O6-alkylguanine remain sensitive to DNA damaging agents that generate 3MeA lesions. In response to DNA damage, the PI3KK like kinases ATM and ATR are activated to initiate a cascade of signaling events that can lead to either cell cycle arrest and/or cell death. To explore the dynamics of the DNA damage signaling network in response to 3MeA DNA lesions we have employed the systems biology approach. The biological system involves damaging WT, ATM-/- and Seckel (ATR hypomorphic) human lymphoblastoid cell lines with methyl methanesulfonate (MMS), a non-specific DNA alkylating agent (generates 10% 3MeA) or Me-lex that specifically generates 3MeA (~ 95%). In response to damage by MMS or Me-lex, the activation of various network proteins will be measured and correlated to cell cycle arrest and/or cell death. 3MeA is known to be a replication blocking lesion. Hence it was expected that ATR would be a central player in response to Me-lex (3MeA). Surprisingly, WT cells treated with Me-lex show robust activation of ATM that is similar to double strand breaks induced by ionizing radiation (IR). In contrast, the cellular response is opposite. ATM-/- cells are resistant to Me-lex but sensitive to IR. Detailed mechanistic ODE and statistical correlative PCA/PLSR model will be implemented to shed light on the complex dynamic signaling network in response to both MMS and Me-lex. This project will enhance our understanding of DNA alkylation chemotherapeutics and allow us to predict additional target for combination chemotherapy.

Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation

Ophir Shalem[1,2], Orna Dahan[1], Michal Levo[1], Maria Rodriguez Martinez[1,3] Itay Furman[1] Eran Segal[2]*, and Yitzhak Pilpel[1]*

[1]Department of Molecular Genetics
[2]Department of Applied Mathematics and Computer Science
[3]Department of Physics of Complex Systems

Weizmann Institute of Science,
Rehovot 76100, Israel

*corresponding authors

Abstract

The state of the transcriptome reflects a balance between mRNA production and degradation. Yet how these two regulatory arms interact in shaping the kinetics of the transcriptome in response to environmental changes is not known. We subjected yeast to two stresses, one that induces a fast and transient response, and another that triggers a slow enduring response. In parallel to conventional microarrays measurements of mRNA abundance, we also used microarrays following transcriptional arrest to measure genome-wide decay profiles under each condition. We found condition-specific changes in mRNA decay rates and coordination between mRNA production and degradation. In the transient response, most induced genes were surprisingly de-stabilized, while repressed genes were somewhat stabilized, exhibiting counter-action between production and degradation. This strategy can reconcile high steady-state level with short response time among induced genes. In contrast, the stress that induces the slow response displays the more expected behavior, whereby most induced genes are stabilized, and repressed genes de-stabilized. We show genome-wide interplay between mRNA production and degradation, and that alternative modes of such interplay determine the kinetics of the transcriptome in response to stress.

Posters

Towards a multi-scale cell specific knowledgebase of the adaptive immune system

Shai S. Shen-Orr[1,2], Ofir Goldberger[2], Yael Garten[3], Yael Rosenberg-Hasson[4], Patricia A. Lovelace[4,5], David L. Hirschberg[4], Russ B. Altman[6], Mark M. Davis[2,7], Atul J. Butte[1,8]

[1] Stanford Biomedical Informatics Research, [2] Department of Microbiology & Immunology,[3] Stanford Biomedical Informatics Training Program,[4] Stanford Human Immune Monitoring Center, [5] Stem Cell Biology and Regenerative Medicine,[6] Departments of Bioengineering and Genetics,[7] The Howard Hughes Medical Institute,[8] Department of Pediatrics, Stanford University, School of Medicine, Stanford, CA 94305-5479, USA

The immune system of higher organisms is, by any standard, complex. To date, using reductionist techniques, immunologists have elucidated many of the basic principles of how the immune system functions, yet our understanding is still far from complete. In an era of high throughput measurements, it is already clear that the scientific knowledge we have accumulated has itself grown larger than our ability to cope with it, and thus it is increasingly important to develop bioinformatics tools with which to navigate the complexity of the information that is available to us. Here, we describe ImmuneXpresso, an information extraction system, tailored for parsing the primary literature of immunology and relating it to experimental data. The immune system is very much dependent on the interactions of various white blood cells with each other, either in synaptic contacts, at a distance using cytokines or chemokines, or both. Therefore, as a first approximation, we used ImmuneXpresso to create a literature derived network of interactions between cells and cytokines. Integration of cell-specific gene expression data facilitates cross-validation of cytokine mediated cell-cell interactions and suggests novel interactions. We evaluate the performance of our automatically generated multi-scale model against existing manually curated data, and show how this system can be used to guide experimentalists in interpreting multi-scale, experimental data. Our methodology is scalable and can be generalized to other systems.

Beyond the position weight matrix

Rahul Siddharthan[1]

[1]*The Institute of Mathematical Sciences, Chennai 600113, India*

Transcription factors (TFs) bind preferentially to short, inexactly-conserved DNA sequences or “motifs”, commonly represented by “position weight matrices” (PWMs) that measure the relative frequencies of individual nucleotides at each binding-site position. Most motif-finders, including our recent PhyloGibbs / PhyloGibbs-MP (RS et al, 2005/RS, 2008, PLoS Comput Biol), assume this model for binding sites. A drawback is that PWMs assume independence of nucleotides at different positions, though we know that generic DNA sequence contains significant dinucleotide correlations. It has been pointed out (eg, Djordjevic et al, 2003, Genome Res.) that, even if one assumes independence of binding *energies* for individual nucleotides to the TF, independence of weight-matrix columns does not follow except in a low-concentration limit. Attempts have been made (eg, Djordjevic et al, ibid; Berger et al, Cell, 2008) to make biophysically-motivated models for binding sites. The main drawback is that estimating dinucleotide correlations requires many more known binding sites than a simple PWM does, since there are 4 nucleotides but 16 dinucleotides. Since PWMs seem “good enough” in most applications, and are easily visualised via “sequence logos”, they remain widely used.

We present a dinucleotide model for binding sites that can be seen as a second-order approximation to the underlying biophysical binding-energy model, just as a PWM is a first-order approximation. Unlike previous dinucleotide models (eg, Segal et al, 2006, Nature describing nucleosomes), we consider all possible spacings, not just nearest-neighbours. We show that where sufficient data exist, such gapped nucleotide pairs can be significantly correlated. Since dinucleotide frequencies are not independent, posteriors are calculated approximately.

We observe a spectacular improvement in specificity of predictions with the dinucleotide model (that is, a sharp reduction in false positives), compared to PWMs made from the same binding-sequence data. Further, we observe that flanking sequence (about 20bp on either side) significantly improves the dinucleotide model's predictions, though it has no effect on the PWMs. As a benchmark, we use known binding sites from the RedFly2 TFBS database, excluding sites near a chosen gene (eg, *eve*); use those sites to construct a PWM and a dinucleotide model; and make predictions for the chosen gene. While PWM predictions are scattered over the entire region of interest, most dinucleotide model predictions occur within known enhancers. Results are particularly good for TFs (eg, *hunchback*) for which many known binding sequences exist. Notably, such improved specificity is also observed with dinucleotide models constructed from 28bp B1H-derived binding sequences selected from randomised libraries (Noyes et al, NAR, 2008). This suggests that the dinucleotide correlations, and the importance of flanking sequence, are not entirely explained by the chromatin structure in eukaryotes, or by quirks of DNA statistics, but are a consequence of the DNA-protein-binding mechanism.

Work is on progress on quantifying dinucleotide models for various TFs in fly and yeast, using publicly-available binding data. We argue that PWMs have severe inadequacies that reflect in false-positives, and our approach is significantly superior when prior binding data exist. *Ab initio* prediction is a future topic.

Positive and negative selective pressures preserve microRNA secondary structure

Alexey Spiridonov[1]

[1]*Massachusetts Institute of Technology*

The function of many different RNA molecules depends critically on the formation of a specific, stable secondary structure. MicroRNAs must fold to single-stem hairpins, and are thus an especially simple model for studying the evolutionary effects of such structural constraints. Here, we quantify the amount, and type of selective pressure resulting from the need to maintain a functional RNA secondary structure.

Starting with a set of 143 miRNAs orthologs from *Homo sapiens* and *Danio rerio*, we obtained orthologous miRNAs from 10 other mammalian and non-mammalian species (100-140 sequences in most). We constructed 3-fold alignments of the form human - "other" - zebrafish, and the consensus secondary structures from these alignments. The resulting mutation counts, compared to randomized controls and a theoretical model, showed very clear evidence of a *structural* selective constraint. Our most conservative randomized control preserved the original alignment almost exactly --- including nucleotide content, dinucleotide content, number of mismatches, and the kinds of mismatches (e.g. human A -> other T). Only the positions of the mismatches was shuffled, with the constraint that miRNA stem mismatches stay in the stem, and the rest stay out of the stem. We also compared our results with a simple theoretical control, parametrized by the mutation rate, the transition/transversion ratio, and the nucleotide contents.

The mutation statistics on the original data differed markedly and consistently from the controls. These differences indicate that both positive and negative selective pressure was involved in the evolutionary maintenance of miRNA secondary structure. A particularly striking example of positive selection is the 3-fold excess of substitutions that affected both ends of a secondary structure pairing, but preserved the pairing. Thus, substitution events in RNA sequences are not independent, and in fact exhibit long-range (40nt+) dependencies caused by secondary structure constraints. In the 10 species, we found that 1%-5% of the pairings are under structural selective constraint relative to Homo sapiens. This number appears to reflect evolutionary distance from the human.

Our methods readily extend to other RNA sequences, and should permit the calculation of structural selective constraint in any RNA molecule with cross-species orthologs. Our signals of positive and negative selection could also potentially be used for detecting novel regions of evolutionarily stable structures.

A Computational Method For Identifying Novel Genes Prognostic of Breast Cancer Development in High-Risk Populations

Ying Sun[1], Raffaella Soldi[1], Darren Walker[1], Thomas Conner[2], Saundra Buys[2], W. Evan Johnson[3], Andrea Bild[1]

[1]Department of Pharmacology and Toxicology, College of Pharmacy, University of Utah
[2]Huntsman Cancer Institute, University of Utah [3]Department of Statistics, Brigham Young University.

Family history is an important factor contributing to a woman's risk of breast cancer development. This increased risk reflects the participation of inherited genetic components such as breast cancer susceptibility genes. However, many of the genetic components contributing to breast cancer remain unknown, and certain women with familial histories of breast cancer, or a mutation in the BRCA genes, live disease-free lives despite their high-risk status, while others develop breast cancer. Thus, it is clear that we lack crucial pieces of information to help define the true risk of getting breast cancer. We hypothesize that there are many undiscovered germline genetic aberrations predisposing patients to cancer that will significantly contribute to an accurate assessment of a women's true risk.

We have developed a computational model for breast cancer risk that is sensitive and specific at determining an individual's true risk for development of breast cancer by using 120 samples from exon-level genome-wide expression profiling of peripheral blood mononuclear cells (PBMCs). These samples include women with strong family histories of breast cancer, as well as women with no family history. Half of the women had developed breast cancer, and half have never developed breast cancer. Our goal was to develop a genomic model capable of predicting which high-risk women will actually get breast cancer. We applied three different statistical approaches for predicting cancer status in 120 samples hybridized on exon arrays. These methods include logistic regression (Logistic), support vector machine (SVM) and Bayesian classification tree (Bintree) approaches. Overall, our model is over 80% accurate in predicting cancer development in high-risk women on an internal independent test set.

In addition to providing a predictive model of a woman's true risk for cancer development, by using the exon array data, we have developed a novel computational method based on a Hierarchical Bayesian approach to identify novel splicing variants that correlate to cancer development. The preliminary results highlight the effectiveness and accuracy of a novel normalization and splice variant detection approaches. Our top candidates all have a significantly high probability (greater than 0.95) for loss of a specific exon's expression ("exon skipping") in high-risk women who develop breast cancer. Interestingly, these genes include DNA damage repair, growth factor signaling, epigenetic, RNA interference, and steroid response pathways.

Overall, these results highlight the ability of our predictive models to more accurately assess a woman's true risk of breast cancer occurrence. Further, our studies identify potential splice variants correlated to breast cancer development in high-risk women. We expect these approaches and experiments to identify the genetic changes that underlie cancer predisposition, and assist clinicians and patients in determining the appropriate preventative measures.

A language for transcriptional regulation

Leila Taher[1], Ivan Ovcharenko[1]

[1] *CBB, NCBI, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894*

Divergence in gene expression arises primarily during transcription. The molecular basis for much of the transcriptional control is the specific binding of proteins called transcription factors (TFs) to clusters of binding sites (TFBSs) that are hardwired in the genome. A generally accepted criterion for identifying such regulatory clusters is the evolutionary conservation between two or more nucleotide sequences. However, regulatory clusters often exhibit great variability, both in terms of the spatial arrangement of the TFBSs and the sequence of the TFBSs, and hence, sequence similarity approaches will often fail to detect them.

In our approach, we model DNA sequences based on an alphabet of hypothetical binding specificities. Our alphabet captures general properties observed for known TFBSs, but is not constrained by the motifs in databases. The alphabet consists of a set of 8-mers, in which each nucleotide is associated with a probability; these probabilities serve to define a similarity function. Combinations of smoothed 8-mers and relative positions can be used to describe actual TFBSs. After representing a set of orthologous sequences with this alphabet, we can compute pairwise alignments to obtain clusters of binding specificities. The alignment algorithm that we employ has been adapted from a dynamic programming algorithm initially developed to align restriction enzyme maps, and encompasses operations such as deletions, insertions, inversions, and reshuffling of our regulatory units. The output of the alignment is postprocessed using a Hidden Markov Model.

We designed a strategy directed towards the discovery of functional noncoding elements that have diverged to an extent that prevents their identification by standard sequence comparison methods. First, we trained the model parameters by verifying the recapitulation of evolutionary conserved regions (ECRs), which are computed with classical alignment methods; our results show a very satisfactory AUC (area under the ROC curve) value of 0.85 +/- 0.04. Next, we applied our method to align the sequences, and looked for predictions of noncoding elements that were identified by standard alignment methods (VISTA/AVID, VISTA/LAGAN and BLASTZ) as conserved in human and mouse, but not in either human and frog or human and zebrafish. We tested our approach on a small but relatively well curated set of ten 50 kb orthologous human, mouse, frog and zebrafish sequences. In this manner, we identified eight putative noncoding regulatory elements in the loci of FLOT2, SSH, DLL4, and C10orf114. We are currently in the process of constructing a larger dataset to apply our method and experimentally evaluate the enhancer functionality of our predictions. Furthermore, the analysis of the predicted clusters provides an insight into the structure of the regulatory sequences in relation to the binding sites they comprise.

In conclusion, our novel representation of the regulatory code provides an original means to indirectly compare orthologous sequences and identify extensively diverged regulatory elements. More importantly, our results illustrate the importance of looking for alternative languages to understand the different instructions coded in biological sequences.

The CYRENE Project: A cis-Lexicon and a cis-Browser for Gene Regulatory Systems and Networks

Ryan Tarpine[1], Eric Davidson[2], Sorin Istrail[1,3]

[1]Computer Science Department and Center for Computational Molecular Biology, Brown University; [2]Division of Biology, California Institute of Technology

The CYRENE Project seeks to address the fundamental problem of determining de novo the function of regulatory sequence by developing the cis-Lexicon, a database of known cis-regulatory modules, the cis-Browser, a next-generation regulatory genome browser, and a library of tools for assisting in the annotation pipeline. The cis-Lexicon will be a comprehensive catalog of experimentally-validated gene regulatory knowledge, designed to be a foundation and benchmark for future prediction algorithms. Presently the cis-Lexicon contains cis-regulatory modules of over 100 transcription factors. The cis-Browser is a high-speed integrative genomics environment for viewing and annotating a variety of types of genomic information in full genome context. It is based on the Celera Genome Browser, redesigned in the past two years into the Regulatory Genome Browser. It is capable of displaying data from the cis-Lexicon, public online databases, and comparative genomics analyses. To aid annotators' entry of information into the cis-Lexicon, we are developing high-throughput tools for finding relevant literature and assisting in the extraction of correct information.

[3]Corresponding author: sorin_istrail@brown.edu, Computer Science Department and Center for Computational Molecular Biology, Box 1910, 115 Waterman Street, CIT 523, Providence, RI 02912, Brown University

Systems biology platform established for studying KRAB-ZNF cell biology in human tissues as well as human neoplasias

Peter Lorenz[1]; Michael Kreutzer[1]; Larisa Kiseleva[2]; Ziliang Qian[3]; Kurt Frischmuth[4]; Yixue Li[3]; Hans-Jürgen Thiesen[1]

[1] Institute of Immunology, University of Rostock, Germany; [2]AIST, Computational Biology Research Center, Tokyo, Japan; [3]Shanghai Institute for Biological Sciences, CAS, China, Institute of Mathematics[4], University of Rostock, Germany;

The human genome encodes 350 KRAB zinc finger genes (www.SysZNF.org). The KRAB domain initially described as a heptad repeat of leucines (Thiesen, 1990, New Biol) represents one of the strongest repression domains found in mammalian organisms (Margolin et al., PNAS 1994; Deuschle et al., MCB 1995; Lorenz et al, Biol Chem 2001). In a long term effort in strong collaboration with the Swedish Human Proteome Resource (HPR) program (www.proteinatlas.org) antisera against numerous KRAB-ZNF proteins have been generated to dissect their biological functions in respect to expression patterns in normal and cancer tissues finally enabling systems biology oriented modeling approaches.

Novelty of the approach: Tissue expression profiles of 31 KRAB-ZNF proteins established by immuno-histochemistry on a panel of 48 normal human tissues and 20 different cancer types originally categorized by 4 expression levels (high, medium, low and negative) have been converted into numeric values such as 4, 3, 2 and 1 in order to establish cell distance networks between tumour tissues, normal tissues and the combination thereof.

Importance of the results: I. A comprehensive heat map displaying the semi-quantitative tissue- and cell type specific signals of 38 KRAB-ZNF proteins appears to be more heterogeneous in normal than in cancer tissues. II. Principal component analysis (PCA) of the semi-quantitative patterns shows classification within a decision boundary with an accuracy of 92 % distinguishing normal from cancer tissue based on KRAB-ZNF protein expression. III. A support-vector machine (SVM) initially trained with 92 ZNF expression data sets i. describing 20 human tumor types, ii. derived from 48 normal tissues including 24 sets representing cellular subtypes demonstrates that ZNF genes are differentially expressed in normal tissues versus various cancer tissues. IV. A minimum spanning tree of a full graph taking Pearson correlation coefficients translated into correlation distances (Kruskal's algorithm) is computed between every pair of samples using the KRAB-ZNF protein expression data of each normal tissue. V. Pearson correlations based on the KRAB-ZNF protein expression data of cancer tissue are visualized. Since signals are quite heterogenous between tumours, the highest and lowest expression values for particular ZNF genes presented by one tumour type has been used as input to establish cell distance networks.

Perspectives: The KRAB ZNF protein distribution in human tissues and related cancer types provides a platform (www.toponostics.org; http://proteinatlas.biosino.org/) suitable for data mining of KRAB-ZNF-related target gene expressions in time and space as well as for modelling dysregulated disease pathways found in human cancer tissues. Numerous of the more than 5000 proteins (Proteinatlas) that are inversely correlated in their expression to KRAB-ZNF proteins could be putative target genes of KRAB-ZNF proteins.

Comparative analysis of nucleosome sequence organization in human and yeast gnomes

Michael Y. Tolstorukov[1], Peter V. Kharchenko[1,2], Joseph A. Goldman[3], Robert E. Kingston[3] and Peter J. Park[1,2]

[1]Harvard Partners Center for Genetics and Genomics, New Research Building, 77 Avenue Louis Pasteur, Boston, MA 02115 USA; [2]Children's Hospital Informatics Program, 300 Longwood Ave., Boston, MA 02115 USA; [3]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA.

Nucleosomes comprise 147-bp of DNA tightly wrapped around a histone protein core, constituting the fundamental repeating units of eukaryotic chromatin. Placement of nucleosomes at specific positions in the genome may regulate gene function by altering accessibility of transcription factor binding sites and facilitating formation of higher-order chromatin structures. The affinity of the histone core for DNA depends on the nucleotide sequence; however, it is unclear to what extent DNA sequence determines nucleosome positioning *in vivo*, and if the same rules of sequence-directed positioning apply to genomes of varying complexity.

High-throughput DNA sequencing technology in combination with chromatin immunoprecipitation allows genome-wide mapping of histone variants and covalent modifications. We have developed computational methods to detect stable nucleosome positions from such data, and used them to determine positions of nucleosomes containing the H2A.Z histone variant and histone H3 tri-methylated at lysine 4 (H3K4me3) in human $CD4^+$ cells.

Our results show that DNA sequences associated with human nucleosomes lack the pronounced 10-bp periodicity in dinucleotide distribution that is characteristic for nucleosome-positioning sequences in yeast and other organisms. We also find that compared with the H3K4me3-enriched nucleosomes, human H2A.Z nucleosomes protect shorter DNA fragments from MNase digestion and exhibit different sequence preferences, suggesting a novel mechanism of nucleosome organization for the H2A.Z variant.

Functional Variation of Degenerate Positions in Transcription Factor Binding Sites

Chung-Yi Cho[1,2], Krishna. B. S. Swamy [1,3], Huai-Kuang Tsai [1,2,3]

[1]Institute of Information Science, [2]Research Centre for Information Technology Innovation, [3]Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan

Transcription factors regulate the gene expression by binding to specific cis-regulatory elements (transcription factor binding sites (TFBSs)) in gene promoters that directs the initiation of transcription and expression rate. Some TFBSs contain one or more highly degenerate positions. Although the prevalent assumption is that these degenerate positions are functionally equivalent, there is increasing evidence showing that the gene expression is dependent on TFBS degenerate positions. In this study, we proposed a method to study the relationship between the gene expression and TFBS degenerate positions. The results showed that more than one-third degenerate positions were significantly related to the expression of target genes. Besides, the influence of degenerate positions on expression is condition-specific. Further, we extended the proposed method to investigate combinatorial effects of degenerate positions. About 20% of the pair of degenerate positions were found depending on each other and highly associated with the expression of their target genes. Surprisingly, some degenerate positions when considered together are functionally significant in several microarray datasets but inactive when considered alone. Our result should be helpful to predict the outcome considering the effect of more than one degenerate position, giving a deeper understanding of the regulatory mechanism. In addition, this method can be easily extended to study such transcriptional functional relationship in other eukaryote species.

A novel probability-based pattern of transcription factor activity and approach to identifying transcriptional regulatory relationship

Tsung-Yeh Tsai [1,2], and Chung-Ming Chen [2]

[1]Taiwan International Graduate Program, Academia Sinica, 128, Section 2, Academia Road, Nankang, Taipei 115, Taiwan; [2]Institute of Biomedical Engineering, National Taiwan University, 1, Section 1, Jen-Ai Road, Taipei 100, Taiwan.

Transcriptional regulation is a keystone in the analysis of biological pathway and genetic therapy. However, there are little explicit genome-wide information due to the complexity of gene regulation and the lack of effective and efficient experimental technique. Also the prevalent computational models for prediction of gene regulation are not accurate enough. Thus there is a requirement of looking at the gene regulation problem from a new perspective. Consequently, to depict real phenomenon of transcriptional regulation and improve the accuracy of identifying regulatory relationship, a novel serial pattern called Transcription Factor Activity Probability Pattern (TFAPP), a TF-gene regulatory relationship identification method has been developed in this study.

The main idea of TFAPP is to take into account change in the probability of TF activity based on the co-expression of potential target genes. Here the activity of a TF is probability based instead of taking quantitative measure of its activity, which is hard to measure. For each TF, the potential target genes are identified initially by ChIP-chip analysis. Then the expression patterns of these target genes from the microarray data are processed by wavelet de-noise and a binary threshold protocol. Further TFAPP is estimated from this by using the learning algorithm, which is based on binary factor analysis and random sampling process. Finally, the regulatory relationships between TF and genes can be identified according to the correlation between the TFAPP of a TF and the significant expression patterns of genes.

In this work, the TFAPP has been proved meaningful for 37 yeast TFs in cell cycle condition, according to the function and phase-specific information. The robustness of the learning algorithm has also been confirmed by 10-fold cross validation. The accuracy of the relationship identification method is validated by three high-confident targeted genes lists, including biology experimental results (80.77%), TRANSFAC database (79.81%), and reliable data collected from literature (88.24%). Our method can be valuable for studies on TF interactions prediction and regulatory network reconstruction. Further TFAPP can be extended to regulatory genomics and system biology analysis of higher eukaryotes.

Higher-order Genomic Organization of Cellular Functions in Yeast

Tamir Tuller[1-4], Udi Rubinstein[3], Dani Bar[3], Michael Gurevitch[3], Eytan Ruppin[1,4] and Martin Kupiec[2]

[1]School of Computer Sciences, Tel-Aviv University; [2]Department of Molecular Microbiology and Biotechnology, Tel-Aviv University; [3]Multiple Sclerosis Center, Sheba Medical Center, Tel-Hashomer; [4]School of Medicine, Tel Aviv University

Previous studies have shown that the distribution of genes in prokaryotes and eukaryotic genomes is not random. Using the thousands of cellular functions that appear in the Gene Ontology (GO) project, we exhaustively study the relation between functionality and genomic localization of genes across 16 organisms with rich GO ontologies (one prokaryote and 15 eukaryotes).

Overall, we find that the genomic distribution of cellular functions tends to be more similar in organisms that have higher evolutionary proximity. At the primary level, which measures localization of functionally related genes, the prokaryote *E. coli* exhibits the highest level of organization, as one would expect given its operon-based genomic organization.

However, examining a higher level of genomic organization by analyzing the co-localization of pairs of different functional gene groups, we surprisingly find that the eukaryote yeast *S. cerevisiae* is markedly more organized than *E. coli*. A network-based analysis further supports this notion and suggests that the eukaryotic genomic architecture is more organized than previously thought.

Computational methods for detecting immune selection and understanding affinity maturation

Mohamed Uduman[1], Uri Hershberg[2], Steven H. Kleinstein[1, 2]

[1] Interdepartmental Program in Computational Biology and Bioinformatics; [2] Department of Pathology, Yale University School of Medicine, New Haven, CT

Affinity maturation is the dynamic process by which the immune system produces B lymphocytes capable of binding to foreign pathogens with high affinity. It is a critical component of adaptive immunity, which protects the host from recurring and ever-evolving pathogens. This process is driven by somatic hypermutation (SHM) of B cell receptor genes and selection of the mutated B lymphocytes based on their affinity for the invading pathogen. Detecting the forces of selection from experimentally-derived B cell receptor sequences is critical in understanding affinity maturation, and can provide insights into antigen-driven selection in health (e.g., vaccination and immunity) and disease (e.g., autoimmunity and lymphomas).

Due to the short time scale and small region targeted for mutation, standard evolutionary biology methods cannot be directly applied to test for immune selection. Moreover, coming up with novel methods for detecting selection and validating these methods is difficult given the limited availability of controlled experimental data. To overcome this problem we have employed novel transgenic mouse models, and also developed a stochastic simulation of B cell clonal expansion in which underlying biological parameters (such as selection pressure, number of divisions, etc.) can be set and their individual influences studied in depth. The synthetic data generated by our simulation allows us to extensively validate previously proposed methods and gain valuable insights into their performances.

The most common methods to test for selection compare the expected and observed frequencies of replacement and silent mutations (R:S). While these methods have been criticized for high false-positive rates, we found this low specificity was due to incomplete accounting for intrinsic sequence-specific biases of SHM. We have corrected this shortcoming and show the resulting method has the expected specificity on a large set of non-productively rearranged, presumably unselected, antibody receptor genes. Our method was also able to detect positive selection in transgenic mice where selection was expected to be a major force. However the sensitivity of R:S-based methods is relatively low, leaving room for further improvement.

In order to improve the sensitivity for detecting antigen-driven selection, we have been investigating the use of graph theoretical properties of B cell lineage trees. Using our simulation-based validation approach we have shown that none of the previously proposed properties (such as the overall degree of "bushiness") are reliable surrogates for in vivo selection since they are confounded by methodological differences in experiments, such as sampling time and pick size. However, in both simulated and in vivo data, we found that the signal for selection is strongest near the root of the tree. Taking advantage of this observation, we propose new tree properties that are not affected by methodological factors, and show how they can be used to detect selection.

Gene Experession Signatures Characteristic of Cell Sensitivity to DNA Damaging Agents

Chandni Valiathan[1,2] and Leona Samson[1-4]

[1]Computational and Systems Biology, MIT; [2]Center for Environmental Health Sciences, MIT; [3]Department of Biological Engineering, MIT; [4]Department of Biology, MIT

We have evolved DNA damage response mechanisms to deal with the constant damage to our DNA by endogenous and exogenous agents. These mechanisms repair or eliminate damaged DNA and maintain our genomic integrity. Elimination of cells that have severely damaged DNA by cell death prevents loss of genomic integrity, aberrant proliferation and development of diseases like cancer. The process by which a cell decides to die or survive after DNA damage is yet unknown. However, the regulation of transcriptional response to DNA damage has been implicated in these death/survival decisions. In fact, a majority of tumors contain mutant forms of transcription factors that are activated after DNA damage.

Initial studies from our lab have shown that twenty-four genetically diverse human lymphoblastoid cell lines show a broad range of sensitivities to MNNG, a DNA alkylating agent[1]. The range of sensitivities observed in this panel suggests an intrinsic difference between the cell lines, in the transcriptional regulatory mechanisms that control a cell's decision to survive or die after DNA damage. To better understand this transcriptional control of cell survival or death decisions after DNA damage, we are screening the panel for sensitivity to various DNA damaging agents. The DNA damaging agents used here are representative of clinically prescribed cancer therapeutics. To facilitate rapid screening of the panel to various DNA damaging agents, we have developed a sensitive, high-throughput, 96-well plate assay which measures the proliferative ability of cells after treatment with a DNA damage agent.

From these screens, the two most resistant and two most sensitive cell lines for each agent will be used to make temporal gene expression measurements and transcription factor activation measurements induced after DNA damage. Comparing these measurements in resistant cell lines where most cells decide to survive after DNA damage, to those in sensitive cell lines where most cells decide to die after DNA damage, will reveal cell survival or death decision control at the transcriptional level. Computational tools will be used to extract both agent-specific expression patterns and common (agent non-specific) gene expression signatures for cell survival or cell death after DNA damage. Additionally, these algorithms will be used to learn the temporal transcriptional regulatory mechanisms in sensitive and resistant cells. This study will reveal general transcriptional regulatory mechanisms that govern a cell's decision to die or survive after DNA damage by the agents used here. This will improve our knowledge on development of drug resistance in tumors, and ways to reverse this resistance.

[1] Fry, R.C., et al., *Genomic predictors of interindividual differences in response to DNA damaging agents.* Genes and Development, Oct 1, 2008. **22**(19)

Efficient Query-Driven and Global Biclustering of Gene Expression Data Using Probabilistic Relational Models

Tim Van den Bulcke[1], Hui Zhao[2],Kristof Engelen[2], Tom Michoel[3,4], Bart De Moor[1], Kathleen Marchal[1,2]

[1]Department of Electrical Engineering(ESAT-SCD), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium;

[2]Department of Microbial and Molecular Systems(CMPG), Katholieke Universiteit Leuven, Kasteelpark Arenberg 20, 3001 Leuven, Belgium;

[3]Department of Plant System Biology, VIB, Technologiepark 927, 9025 Gent, Belgium;

[4]Department of Molecular Genetics, UGent, Technologiepark 927, 9025 Gent, Belgium;

The reconstruction of the cellular signaling pathways is one of the foremost challenges in systems biology. While a large amount of high throughput 'omics data are available, the reconstruction of this signaling network still remains a highly underdetermined problem. Currently, insufficient data is available to uniquely identify all the interactions and their parameters in the model. However, regulatory networks exhibit a modular organization, an aspect that is successfully exploited in a number of biclustering and module inference methods. Biclustering algorithms for gene expression data perform simultaneous clustering in both the gene and condition dimensions. The result is a subset of genes that are coexpressed under a subset of conditions.

The *Pro*Bic model uses a hybrid query-driven and model-based approach. This allows researchers to both identify biclusters using a global approach and to incorporate prior knowledge by performing directed queries around genes of interest. Although some other algorithms also allow for query-driven searches, they do not combine the advantages of the query-driven search with a model based approach for identifying overlapping biclusters. We present an alternative approach to identify overlapping biclusters in gene expression data using Probabilistic Relational Models. *Pro*Bic can be applied in both a query-driven and a global setting. Moreover, *Pro*Bic was designed such that it is easily extensible towards additional data types.

An extensive evaluation of the algorithm was performed on synthetic data to investigate the behavior of the algorithm under various parameter settings and input data. The results on a wide range of synthetic datasets show that *Pro*Bic successfully identifies biclusters under various levels of noise, overlap and missing values and this in both the query-driven and global setting. We show that prior knowledge in the form of a set of query genes guides the algorithm towards biclusters of interest to a biologist. Moreover, the bicluster identification using query genes is quite robust as the set of query genes can contain several 'noisy' genes that are not part of the bicluster of interest, a situation that often occurs in practice.

In conclusion, *Pro*Bic is an efficient biclustering algorithm that simultaneously identifies a set of overlapping biclusters in a gene expression dataset. It can be used in both a query-based and a global mode. Experiments on synthetic data showed that biclusters are successfully identified under various settings, both in the query-driven and the global setting.

Conserved noncoding elements are evolving rapidly in teleost fishes

Alison P. Lee, Sydney Brenner and B. Venkatesh

Institute of Molecular and Cell Biology, A*STAR (Agency for Science, Technology and Research), Biopolis, Singapore 138673

Vertebrate genomes contain several thousand conserved noncoding elements (CNEs), which are likely to be functional elements that are under purifying selection. Although the functions and significance of CNEs have not been fully understood, *in vivo* functional assays have indicated that many of them may function as transcriptional regulatory elements directing tissue- and developmental stage-specific expression of target genes. Cartilaginous fishes are the oldest living group of jawed vertebrates. We have identified about 5000 CNEs (more than 70% identical and longer than 100 bp) between the human genome and the genome of a cartilaginous fish, the elephant shark (*Callorhinchus milii*). These elements, conserved over 450 million years of evolution, represent ancient functional noncoding elements that are likely to be shared by all jawed vertebrates. Interestingly, although almost all of them are conserved in dog and chicken genomes, a significant number of them have been lost or diverged beyond recognition in teleost fishes such as fugu, zebrafish, medaka and stickleback. Furthermore, there is very little overlap between CNEs retained among different teleosts. Estimation of the substitution rates in CNEs shared by teleost fishes, human and elephant shark indicate that the CNEs in teleost fishes have accumulated substitutions at an increased rate compared to human. These analyses indicate that the ancient noncoding elements that are under selective constraint in cartilaginous fishes and tetrapods have experienced an accelerated rate of evolution solely in the teleost fish lineage. Such fast evolving CNEs may be associated with adaptive evolution of gene regulation in teleost fishes.

The role of whole gene duplication in the evolution of the yeast regulatory network

B. Vernot [1], R. D. Sedgewick [1,2], S. Tavazoie [3], D. Durand [1,4]

[1]Department of Biological Sciences, Carnegie Mellon University; [2]Currently at Google; [3]Lewis-Sigler Institute for Integrative Genomics and Department of Molecular Biology, Princeton University; [4]Department of Computer Science, Carnegie Mellon University.

Duplication is the primary force driving the evolution of novel genes. Current theory proposes that the evolution of novel function following duplication is less selectively constrained because of the availability of a duplicate copy that can perform the pre-duplication function. However, functional innovation is still constrained because paralogs must function in the context of existing networks. Some have proposed that whole genome duplication (WGD) relaxes these constraints because all interaction partners are duplicated simultaneously.

In this study, we investigate the fates of duplicated regulatory interactions when transcription factor (TF) and target are duplicated simultaneously. Genes that arose by WGD (ohnologs) can be identified by sequence comparison and synteny[1]. ChIP-chip data is used to infer regulatory interactions[2]. A number of studies have considered gene duplication in the context of networks. Our work differs from prior results in that we consider a very specific question: does simultaneous duplication of transcription factors and their targets offer different opportunities for innovation than duplication of genes in the same network at different times. By focusing on very simple network motifs, we are able to enumerate all possible outcomes following a duplicated interaction when both TF and target are retained in duplicate and tabulate how often each outcome occurs. We use randomization to identify significant enrichment (or depletion) of network motifs in the observed data compared with a null model. To assess the importance of simultaneous duplication, we performed the same analysis on regulatory interactions involving non-ohnolog paralogs; i.e., network motifs in which TF and target were duplicated at different times.

Our results include some surprising observations. First, our data refutes the common expectation that novel pathways arise through duplication and functional separation of entire pathways simultaneously. Second, simultaneous duplication of TF and target does not appear to be a strong predictor of the types of regulatory interactions that are retained following duplication. Rather, the type of duplication by which the TFs arose appears to be the determining factor. In other words, we see enrichment of the same network motifs whenever the TFs arose through WGD, regardless of whether the targets are were duplicated simultaneously or in a separate event. Our results suggest that factors indirectly related to WGD – such as the mechanism of duplication, the age of the duplicates and the tendency to form homo- or hetero-dimers – most strongly influence functional differentiation following duplication in the context of regulatory networks.

[1] Byrne and Wolfe, Genome Research, 2005
[2] MacIsaac *et al*, BMC Bioinformatics, 2006

This work was supported by NIH grants 1 K22 HG 02451-01 and GM 071508, and a David and Lucille Packard Foundation fellowship.

Genome-wide binding and regulatory targets of hnRNP H

Eric T. Wang[1,2], Razvan Nutiu[2], Xinshu Xiao[2], Gary Schroth[3], David Housman[3,4], Christopher Burge[2]

[1]Harvard-MIT Division of Health Sciences and Technology; [2]Department of Biology , Massachusetts Institute of Technology; [3]Illumina, Inc.; [4]Koch Institute for Integrative Cancer Research at MIT

RNA-protein interactions are critical for pre-mRNA processing, regulation of mRNA export, and ultimately, gene expression. The RNA binding protein hnRNP H, which binds G-rich RNA, can act as either a splicing repressor or a splicing activator depending on context, but the global targets and roles of this factor remain unknown. We have utilized two complementary high-throughput sequencing approaches to elucidate both the binding and regulatory targets of hnRNP H in human cells. First, Illumina mRNA-SEQ was performed using both control and hnRNP H RNAi-treated HEK 293T cells, yielding over 50 million 32-bp long reads, which we have mapped to the genome and transcriptome to identify differentially regulated mRNA processing events. HnRNP H was found to regulate the expression of both constitutive and alternative exons, in a manner dependent on both the density of G-runs and 5' splice site strength. Second, UV cross-linking and immunoprecipitation (CLIP) of hnRNP H in HEK 293T cells followed by Illumina sequencing (CLIP-SEQ) was performed, yielding over 1 million 32-bp long reads that could be mapped in a strand-specific manner to the transcriptome. Our analyses of CLIP-SEQ data confirmed that a large proportion of hnRNP H binding sites contain runs of G, with locations in exons, introns, and UTRs. We also observed an enrichment of hnRNP H CLIP tags in introns flanking exons whose inclusion levels decreased following H knockdown, and an enrichment of tags within exons whose inclusion levels increased after knockdown, consistent with exon-specific repressor and intron-specific activator functions of hnRNP H. These results broaden our understanding of hnRNP H and provide a general procedure for analysis of RNA binding protein function.

A Model of the Competitive Binding of DNA by Nucleosomes and Transcription Factors

Todd Wasson[1], Alexander J. Hartemink[1,2]

[1]Computational Biology and Bioinformatics; [2]Department of Computer Science, Duke University, Durham, NC 27708

Transcriptional behavior is determined in large part by the occupancy of a gene's promoter by a set of various DNA binding proteins and protein complexes, most notably transcription factors. However, as 80% of the genome is occupied by nucleosomes, chromatin structure plays an important role in the accessibility of DNA to transcription factors. Furthermore, as DNA occupancy is the dynamic result of thermodynamic competition amongst a set of DNA binders, a model of occupancy including transcription factors and nucleosomes and expressed in terms of probabilities or frequencies is likely to give a more complete view of the underlying biology. We consider a mechanistic view of DNA occupancy to be especially useful in understanding the nature of competition and cooperativity amongst DNA binders and the resulting landscape of occupancy of nearby genomic DNA in particular. That is, we model DNA binders explicitly and individually in thermodynamic competition with one another based on sequence affinity and concentration, as opposed to considering a more generalized model of the aggregate occupancy behavior.

We propose a statistical thermodynamic model to understand DNA occupancy in yeast from a mechanistic standpoint. While similar existing techniques may consider either multiple transcription factors and naked DNA [Sinha 2006] or nucleosomes and naked DNA [Segal 2006], our model allows for all of these at once. Given a sequence, a collection of generic DNA binders with defined sequence specificities, concentrations of each binder, and a system temperature, our model produces a posterior decoding under the Boltzmann distribution of the binding probability per sequence position of each binder along the sequence.

Our model is quite flexible and extensible. Any DNA binder whose binding preferences can be described by sequence affinity may be incorporated into the model. These preferences can be represented by arbitrary probability models: currently our nucleosome model assumes position-specific dinucleotide preferences while our transcription factor model assumes mononucleotide preferences (PSSMs). We may ask nuanced questions of the model, such as what the occupancy of a region looks like given that particular transcription factors are held in place in specific positions, or how often two transcription factors bind within a given distance of one another. We may allow refinement of the sequence preference of DNA binders, and aid in discovery of hitherto unknown roles of transcription factors. We may perform mutational analyses to both sequence and binders to predict novel binding behavior and guide biological experiments.

Importantly, our technique is computationally efficient. Entire yeast chromosomes with models on the order of 5000 states representing hundreds of transcription factors and nucleosomes are decoded in minutes.

This detailed view of DNA binding allows for future understanding of the interplay between the transcriptional control exerted by a gene product and the gene's own transcriptional rate; that is, the exploration of the larger transcriptional regulatory network in terms of the mechanistic interplay of its constitutive parts.

The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome

Yutao Fu[1], Manisha Sinha[2,3], Craig L. Peterson[3], Zhiping Weng[1,4,5]

1 Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America, 2 Interdisciplinary Graduate Program, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America, 3 Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America, 4 Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America, 5 Department of Biomedical Engineering, Boston University, Boston, Massachusetts, United States of America

Abstract
Chromatin structure plays an important role in modulating the accessibility of genomic DNA to regulatory proteins in eukaryotic cells. We performed an integrative analysis on dozens of recent datasets generated by deep-sequencing and high-density tiling arrays, and we discovered an array of well-positioned nucleosomes flanking sites occupied by the insulator binding protein CTCF across the human genome. These nucleosomes are highly enriched for the histone variant H2A.Z and 11 histone modifications. The distances between the center positions of the neighboring nucleosomes are largely invariant, and we estimate them to be 185 bp on average. Surprisingly, subsets of nucleosomes that are enriched in different histone modifications vary greatly in the lengths of DNA protected from micrococcal nuclease cleavage (106.164). The nucleosomes enriched in those histone modifications previously implicated to be correlated with active transcription tend to contain less protected DNA, indicating that these modifications are correlated with greater DNA accessibility. Another striking result obtained from our analysis is that nucleosomes flanking CTCF sites are much better positioned than those downstream of transcription start sites, the only genomic feature previously known to position nucleosomes genome-wide. This nucleosome-positioning phenomenon is not observed for other transcriptional factors for which we had genome-wide binding data. We suggest that binding of CTCF provides an anchor point for positioning nucleosomes, and chromatin remodeling is an important component of CTCF function.

Inference of Logical Regulatory Networks from Coarse Time Series

Roy Wilds[1], Leon Glass[2]

[1]Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada
[2]Department of Physiology, McGill University, Montreal, QC, Canada

We have developed a method for inferring robust regulatory networks from coarse expression data that is based on a logical control system. Provided the expression data can be qualitatively described as a sequence of HIGH/LOW states, we are able to infer candidate regulatory networks that are based on logical control which reproduce the dynamics. The identified networks are piece-wise linear ordinary differential equations in which the regulatory structure is specified by Boolean functions. Our approach produces networks that are highly robust in that the sequence of HIGH/LOW states persist over a wide range of kinetic parameters in the differential equations.

A powerful feature of our approach is the ability to include a priori information concerning regulatory interactions. We are able to identify the logical control structure necessary to yield robust dynamics, given an initial set of interacting components. Additionally, it is also possible to identify missing regulatory components under the hypothesis that the network dynamics must be highly robust.

We demonstrate the applicability of our approach by analyzing several models from the literature. Furthermore, we compare our approach with discrete time logical methods for identifying regulation structure (such as in [1]). Conditions under which the two models infer the same logical functions are proven.

[1] Global control of cell-cycle transcription by coupled CDK and network oscillators. Orland, D.A. et al., Nature (453), pp. 944-947 (2008).

Overcoming Cell Cycle Addiction: Conditional Module Discovery and Application to Breast Cancer Biomarkers

Chang-Jiun Wu[1], Tianxi Cai[5], Isaac Kohane[3,4], and Simon Kasif[1,2,3]

[1]Bioinformatics Program; [2]Department of Biomedical Engineering, Boston University; [3]Children's Hospital Informatics Program; [4]Harvard Medical School; [5]Department of Biostatistics, Harvard University, Boston.

Background

A large number of gene expression-based prognostic biomarkers have been evaluated by the research community for the prediction of the clinical prognosis of breast cancer patients. Several recent studies suggest that while these biomarkers are informative in predicting survival, they appear to primarily track the relationship of cell-cycle activity to outcome. However, these cell-cycle-related biomarkers show a declined association to outcome in specific breast cancer subpopulations, such as estrogen receptor (ER)-negative tumors. Thus, biomarkers aimed to discover tumor properties that are independent of cell cycle have the potential to improve the prediction. The challenge stems from the fact that no single gene is significantly correlated to breast cancer prognosis after controlling for the effect of cell-cycle activity. Many traditional module discovery algorithms based on existing pathways produce signatures strongly correlated to cell cycle and are not adding a significant predictive value.

Methods. We designed a Pheno-Genomic Module Profiler algorithm that integrates microarray expression profiles, systematic pathway annotations, transcription-factor binding information, and protein-interaction network data to search for functionally linked gene modules predictive of phenotype. The search engine uses a novel stepwise correlation approach to detect functional modules with significant additive predictive value (over cell cycle) with respect to breast cancer outcome.

Results. We identified three gene modules independently predictive of breast cancer outcomes. The first module is associated with cell-cycle-related pathways, and is correlated to commonly used prognostic signatures. The second module is enriched with extracellular matrix interaction pathway genes. Overexpression of this module is associated with poor prognosis (shorter survival) conditional on the cell-cycle activity. The third module is enriched with genes from multiple immune-related pathways. Overexpression of genes in this module is indicative of good prognosis after controlling for the effects of the other modules. These three modules are robustly associated with survival on independent validation datasets. A model combining these modules predicts outcome 2–8% better than signatures based solely on cell-cycle activities in terms of AUC. In the ER-negative cohort, an integrative model targeting this specific breast cancer subpopulation improves the predictive performance by 10–20% compared to existing biomarkers.

Conclusion. This study provides a new direction for biomarker design that improves both biological transparency and clinical utility. The robust predictive capability of the three modules suggests that these cancer mechanisms are consistently predictive across different patient cohorts.

Integrative inference of gene networks from expression data

Xuebing Wu, Feng Zeng, Rui Jiang*

MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China (corresponding author)*

It is common in bioinformatics that numerous methods have been developed for the same problem, using different models and data sources. The integration of these methods can typically yield approaches with higher performance and is therefore of great importance. We propose HHMI (Hierarchical Heterogeneous Model Integration), an unsupervised framework for integrating multiple methods, and we demonstrate the advantage of this framework in the network inference problem. We compile a wide spectrum of computational approaches for reconstructing networks from gene expression data, including methods that are inherently linear or nonlinear, local or global, aimed at inferring directed or undirected networks, and using steady-state or time-series data. A total of 16 predictive modules are obtained by applying 13 inference methods (3 are novel) to 4 types of expression data that are generated by a detailed kinetic model. Weight matrices are predicted by each module and then integrated using 3 strategies and 4 fusion statistics (resulting in 12 integrated modules). All 28 modules are validated using 20 synthetic 100-gene networks with scale-free topologies and evaluated in terms of their average AUC scores. As shown in Fig. 1, the novel HHMI strategy that recursively fuses the most dissimilar predictions is superior to other integration strategies, and HHMI combined with harmonic mean (module HeH) outperforms all other single and integrated modules. These observations are even more obvious when we use experimental expression data to reconstruct the E.coli transcriptional network which contains 1,426 genes. Moreover, our novel integration framework is model and problem independent, flexible to include other methods and data types, and ready to be applied to many other problems.

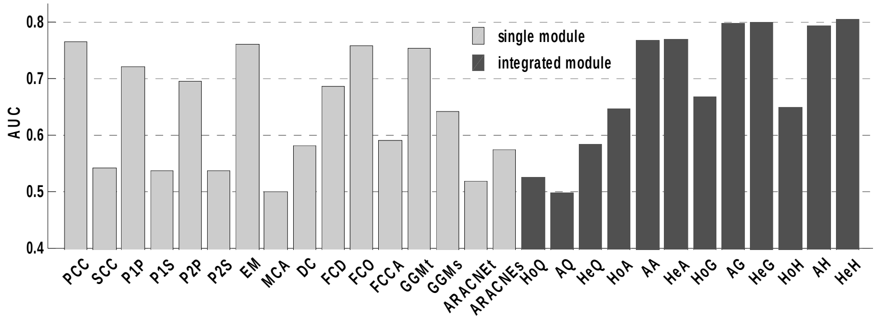


Fig. 1. Comparison of AUC. PCC/SCC: Pearson/Spearman Correlation Coefficient; P1P/P1S: 1st-order Partial PCC/SCC; P2P/P2S: 2nd-order Partial PCC/SCC; EM: Entropy Maximization; MCA: Metabolic Control Analysis; DC: Directed Correlation; FCD/FCO: Fold Change between wild-type and knock-Down/knockOut data; FCCA: Functional Canonical Correlation Analysis; GGM: Graphical Gaussian Model; ARACNE: Algorithm for the Reconstruction of Accurate Cellular Networks ('-s/-t' means the method is applied to steady-state / time-series data). Three integration strategies: A- (All-at-once integration), Ho- (hierarchical Homogeneous model integration) and He- (hierarchical Heterogeneous model integration, i.e. HHMI). Four fusion statistics: -Q (order statistics), -A (arithmetic mean), -G (geometric mean) and -H (harmonic mean).

Genomic profiling of adipocyte and preadipocyte microRNAs reveals deregulated microRNA expression in obese mice

Huangming Xie[1,2,3], Bing Lim[2,3,4], Harvey F. Lodish[1,2,5]

*[1] Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA; [2] Computation and Systems Biology, Singapore-MIT Alliance, National University of Singapore, 4 Engineering Drive 3, Singapore 117576; [3] Stem Cell and Developmental Biology, Genome Institute of Singapore, A*STAR (Agency for Science, Technology and Research), 60 Biopolis Street, Singapore 138672; [4] Beth Israel Deaconess Medical Center, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02215, USA; [5] Department of Biology and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.*

MicroRNAs (miRNAs) are important post-transcriptional regulators and affect diverse biological processes and many diseases. We profiled the expression of more than 370 miRNAs during adipogenesis of the preadipocyte 3T3-L1 cells using miRNA microarrays and validated by RT-PCR eight miRNAs that are significantly upregulated and four that are downregulated. Similar changes in miRNA expression during adipogenesis were observed by comparison of mature primary adipocytes and enriched primary preadipocytes. We also profiled miRNA expression in purified epididymal adipocytes from normal and leptin deficient or diet-induced obese mice; miRNAs were regulated similarly in these two different models of obesity. Importantly, miRNAs that were induced during adipogenesis were downregulated in adipocytes from both types of obese mice. Conversely miRNAs that were decreased during adipogenesis were elevated in adipocytes from obese mice. These changes are likely associated with the chronic inflammatory environment in obese adipose tissue as they were mimicked by TNFα treatment of differentiated adipocytes. Our results provide an important first step towards construction of the entire RNA regulatory network underlying fat cell development and adipocyte dysfunction in obesity and may lead to novel RNA-based therapies that complement current anti-obesity treatments.

DNA Methylation Profiling of Schizophrenia

Xin, Y.[1], Chanrion, B.[1], O'Donnell, A.[2], Ge, Y.[3], Haghighi, F. [1]

[1]Department of Psychiatry, Columbia University, New York, NY; [2]Department of Genetics and Development; [3] Department of Neurology, Mt. Sinai School of Medicine, New York, NY

Emerging evidence suggests that DNA methylation plays an expansive role in the central nervous system, linking this epigenetic process to neurogenesis and neuronal plasticity. Epigenetic processes may also play an important role in the etiology of neuropsychiatric disorders, perhaps as equally important as genetics and the environment. However, until recently epigenetics has received very little attention, yet there are likely pathological abnormalities in genomic methylation patterns that regulate genes involved in the development or physiology of the brain. Our aim is to explore the epigenetic profile of schizophrenia. We have limited our study to the dorsal lateral prefrontal cortex (dlPFC) because converging evidence from structural and functional imaging implicates this region in schizophrenia.

Using experimentally and computationally validated methods, genomic DNA from brain tissue from the dlPFC of schizophrenic and normal subjects was fractionated into methylated and unmethylated compartments and paired-end library construction and subsequent sequencing of both compartments performed via the ultra high-throughput SOLiD sequencing platform. These sequences were then mapped to the draft human genome, delineating the methylated and unmethylated compartments. In mammalian cells, DNA methylation occurs at the 5′-position of cytosine within CpG dinucleotides, and so with this method we characterized the methylation state of ~80% of the CpGs genome-wide for 4 brain specimens. These included two schizophrenic cases and two normal controls.

Posters

To further analyze and annotate these data, we developed an in-house database along with a suite of analytic tools to detect potential disease-specific DNA methylation profiles. In analysis of Next Gen sequencing data, sampling is a key consideration. The genomic coverage varies for each sequenced genomic library, reflecting different sampling of the genome for each individual subject examined. Thus in analyzing genomic features as DNA methylation signatures across multiple individuals, it is critical to consider potential sampling bias. We have developed a novel probabilistic score that accounts for intra-individual sampling bias. For each CpG, a methylation probability is estimated based on the magnitude and confidence associated with this probability score. Then the CpG is assigned a discrete methylation state (methylated, unmethylated, mixed, or unknown). In this way, the DNA methylation profile of each individual subject is determined for subsequent analysis. Using this data we have identified differentially methylated regions between our schizophrenic vs. normal subjects. Although we have observed numerous differentially methylated regions across the genome, our initial efforts are focused on gene promoters where extant DNA methylation is shown to be associated with gene expression. We have evaluated and prioritized these candidate genes for indications of involvement in central nervous system function (e.g., nervous system development, synaptic transmission, axon guidance). Application of these methods will improve our understanding of the potential causes of schizophrenia, a debilitating neuropsychiatric disorder.

A bistable Rb–E2F switch underlies the heterogeneous cell cycle entry at the restriction point

Guang Yao[1,2], Tae Jun Lee[3], Seiichi Mori[1,2], Joseph R. Nevins[1,2*], and Lingchong You[1,3*]

[1]Institute for Genome Sciences and Policy, [2]Department of Molecular Genetics and Microbiology, [3]Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA.

**Address correspondence to: L.Y. (you@duke.edu) or J.R.N. (j.nevins@duke.edu)*

The restriction point (R-point) marks the critical event when a mammalian cell commits to proliferation and becomes independent of growth stimulation. It is fundamental for normal differentiation and tissue homeostasis, and seems to be dysregulated in virtually all cancers. Although the R-point has been linked to various activities involved in the regulation of G1–S transition of the mammalian cell cycle, the underlying mechanism remains unclear. Using single-cell quantifications coupled with mathematical modeling, here we show that the Rb–E2F pathway functions as a bistable switch to convert graded serum inputs into all-or-none E2F responses. Once turned ON by sufficient serum stimulation, E2F can memorize and maintain this ON state independently of continuous serum stimulation. We further show that, at critical concentration and duration of serum stimulation, bistable E2F activation correlates directly with the ability of a cell to traverse the R-point. Lastly, we show that the degree of heterogeneity at cell cycle entry follows a simple mathematical rule, which can be modified by altering the positive feedback strength with Cdk inhibitors.

Ab initio construction of a Eukaryotic Transcriptome by Massively Parallel mRNA Sequencing

Tommy Kaplan[1,2,*], Moran Yassour[1,*], Hunter B.Fraser[3], Dawn-Anne Thompson[3], Joshua Z. Levine[3], Jenna Pfiffner[3], Xian Adiconis[3], Gary Schroth[4], Shujun Luo[4], Irina Khrebtukova[4], Andreas Gnirke[3], Chad Nusbaum[3], Nir Friedman[1], and Aviv Regev[3,5]

1. School of Computer Science and Engineering, The Hebrew University, Jerusalem, 91904, Israel

2. Department of Molecular Genetics and Biotechnology, Faculty of Medicine, The Hebrew University, Jerusalem 91120, Israel

3. Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA

4. Illumina, Inc., 25861 Industrial Boulevard, Hayward, CA 94545

5. Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02142

Defining the transcriptome, the repertoire of transcribed regions encoded in the genome, is a challenging experimental task. Current approaches, relying on sequencing of expressed sequence tags (ESTs) or cDNA libraries, are expensive and labor-intensive. Consequently, we know little about the transcriptome of most sequenced species. Advances in massively parallel sequencing can revolutionize the study of transcriptomes. Here, we present a novel approach for ab initio discovery of the complete transcriptome of the budding yeast, based only on the (unannotated) genome sequence and millions of short reads from a single sequencing run. Using novel algorithms, we automatically construct a highly accurate transcript catalogue, including most known transcripts, and adding 160 novel transcripts and 25 introns. Our results demonstrate that massive parallel sequencing provides accurate definition of a eukaryotic transcriptome without any prior knowledge. This framework can be applied to poorly understood organisms, for which only the genomic sequence is known.

Inferring stable and causal gene interactions through reverse engineering algorithms

M. Zampieri[1], N. Soranzo[1] and C. Altafini[1]

SISSA-ISAS, International School for Advanced Studies, via Beirut 2-4, 34014 Trieste, Italy

The possibility of using the information provided by high-throughput measurements in order to infer interactions between genes represents a first step towards a comprehensive understanding of a biological system in terms of gene functions, 'partner genes', conditions for activation and dynamical behavior.

In the field of systems biology, genome-wide reverse engineering methods try to infer putative gene-gene interactions from the analysis of large compendia of microarray data. Needless to say, the reconstruction of a gene network is a very challenging problem since the number of biological perturbations/experiments is typically comparable to the number of dynamical variables composing the system, while the possible connections scale as an higher power of the number of components. We focus here on two classes of reverse engineering algorithms, called relevance networks and graphical models. They are computationally more tractable than most of the other methods that have been recently proposed (e.g. Bayesian networks, Boolean networks and linear) and can therefore be applied in a truly genome-wide context. They consist essentially in computing a two-point similarity measure between gene pairs which is then used to weight the edges of a graph. We consider five similarity measures: two direct (Pearson correlation and mutual information) and three conditional (partial Pearson correlation, conditional mutual information and graphical Gaussian model). The first two metrics are based on gene-gene co-expression, while the remaining three perform a conditioning operation on the two-point measure.

Our approach confirms the several studies suggesting that gene co-expression is mainly associated to static and stable relationships, like belonging to the same protein complex, rather than other types of interactions, more of a "causal" and transient nature (e.g. transcription factor-binding site interactions). Based on our current knowledge, the interaction networks representing protein complexes (PCs) and transcription factor–binding site (TF–BS) are roughly characterizable by means of different 'recurrent' regulatory motifs. The motifs representing PCs are characterized by undirected subgraphs in which all nodes are mutually connected, while TF–BS modules are better represented by directed subgraphs with a 'scale-free like' connectivity.

Posters

The aim of this work is to verify if different network inference algorithms are indeed tailored to the detection of one of these two types of regulatory motifs. We apply these algorithms to an artificial dataset (generated by a system of non-linear ODEs associated to a network with a characteristic topology) and two microarray collections (for *E.coli* and *S.cerevisiae*). We compare the ability of the different similarity metrics to infer the above "physical networks" (PC and TF-BS collected from databases). In particular we investigate how the predictive power of the algorithms changes increasing the dimension of the motifs (e.g. number of genes forming a PC or the number of TFs regulating the same BS). The comparisons show that the static gene interactions associated to protein complexes are better detected by the direct methods (especially for large PCs), while conditional metrics are more suitable to detect the causal (even if combinatorial) transcriptional regulation.

miROrtho: the computational survey of microRNA genes

Daniel Gerlach, Evgenia V. Kriventseva, Evgeny M. Zdobnov

Department of Genetic Medicine and Development, University of Geneva Medical School, rue Michel-Servet 1, 1211 Geneva 14, Switzerland

MicroRNAs (miRNAs) are short, non-protein coding RNAs that direct the widespread phenomenon of post-transcriptional regulation of metazoan genes. The mature ~22nt long RNA molecules are processed from genome-encoded stem-loop structured precursor genes. Hundreds of such genes have been experimentally validated in vertebrate genomes, yet their discovery remains challenging, and substantially higher numbers have been estimated.

The miROrtho database (http://cegg.unige.ch/mirortho) presents the results of a comprehensive computational survey of miRNA gene candidates across the majority of sequenced metazoan genomes. We designed and applied a three-tier analysis pipeline: 1) an SVM-based ab initio screen for potent hairpins, plus homologs of known miRNAs, 2) an orthology delineation procedure, and 3) an SVM-based classifier of the ortholog multiple sequence alignments.

The miROrtho data are conceptually complementary to the miRBase catalogue of experimentally verified miRNA sequences, providing a consistent comparative genomics perspective as well as identifying many novel miRNA genes with strong evolutionary support.

Integrative Multi-Network Approach to Predict Synthetic Lethal Interactions

Gaurav Pandey[1*], Bin Zhang[2*], Aaron Chang[2], Jun Zhu[2], Vipin Kumar[1] & Eric Schadt[2]

[1] CSE Dept., University of Minnesota, Twin Cities, Minneapolis, MN 55414; [2]Rosetta Inpharmatics, LLC, wholly owned subsidiary of Merck & Co., Inc., Seattle, WA 98109

Genetic interactions characterize the relationships between two genes in terms of fitness when both are deleted and thus are important for predicting gene function, dissecting protein complexes into functional pathways and exploring the sources underlying complex inherited human diseases. In yeast, systematic deletion of all 6000 genes has been instrumental in delineating the non-essential genes, which may in combination with other gene mutations lead to loss of viability. However, testing all pairwise combinations is still prohibitive in terms of time and materials. Synthetic sickness and lethality are the most studied genetic interactions in yeast, however only 4% of gene pairs have been tested due to the huge number of combinations.

To expand the set of known synthetic lethal (SL) interactions, we have devised an integrative, multi-network approach for predicting synthetic lethal interactions that extends previous work by Wong et al. (PNAS 101**:** 15682-15687, 2004). First, we defined 152 features for characterizing the relationships between gene pairs from various sources such as gene expression studies, protein-protein interactions, transcription binding site information, functional annotations by gene ontology categories, gene network modules and communities, to name just a few. In particular, these features are independent of the known SL interactions (in contrast to the approach by Wong et al). To predict the SL or non-SL class for a given gene pair based on these features, we developed a non-parametric multi-classifier system that enabled the simultaneous use of multiple classification procedures, such as SVM, neural networks and decision trees. To access the performance of our approach, we collected 9,994 SL interactions and 125,509 non-SL interactions from the SGD database (www.yeastgenome.org). Results from the cross-validation show that: a) the multiple classifier system consistently performs better than any of the six classifiers on their own; b) the prediction precision can be as high as 90% at a coverage rate of 16.6%; c) these 152 SL-independent features lead to a 20% increase in true positive rate when compared with the performance reported by Wong et al. using their SL-independent features. Moreover, testing on a SGD subset consisting of 337 least connected SL interactions in the SGD SL set and an independent set of 506 SL interactions from E-MAP (Collins et al., *Nature* **446:** 806-810, 2007) which share no gene with the SGD SL set shows SL-independent features are consistently more robust than SL-dependent features.

These comprehensive experiments demonstrate that the SL-independent features in junction with the advanced classification scheme lead to an improved performance when compared to the current state of the art method and they are expected to be more effective and robust in uncovering new genetic interactions from the tens of millions of unknown gene pairs in yeast and from the hundreds of millions of gene pairs in higher organisms like mouse and human in which very few genetic interactions have been identified to date.

Incorporating modular information improves protein identification and data interpretation in shotgun proteomics

Jing Li[1], Lisa J. Zimmerman[2], Byung-Hoon Park[3], David L. Tabb[1,2], Daniel C. Liebler[2], Bing Zhang[1]

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232; [2]Departments of Biochemistry, Vanderbilt University, Nashville, TN 37232; [3]Oak Ridge National Laboratory, Oak Ridge, TN 37831.

Shotgun proteomics has emerged as a powerful technology for protein identification in complex samples with remarkable applications in elucidating cellular and subcellular proteomes and discovering disease biomarkers. Owing to the peptide-centric nature of shotgun proteomics, assembling peptides identified from mass spectra into proteins is a critical step in data analysis. In order to ensure the reliability of protein identification, existing protein assembly pipelines usually eliminate a large number of proteins with limited experimental evidence, including those supported by single peptide and those without unique peptide evidence. However, some of the eliminated proteins may exist in the original samples and play important biological roles. Unfortunately, distinguishing eliminated proteins that are present or absent in a sample is extremely challenging or even impossible based solely on the observed spectral data.

Here we pursue a protein interaction network-assisted approach to address this challenge. We first enumerate proteins complexes in a network. As protein complexes are key modules that integrate multiple proteins to perform cellular functions, we hypothesize that proteins eliminated as a result of insufficient experimental evidence are more likely to be present in the original sample if they exist in a complex enriched with confidently identified proteins. Using a yeast cell culture data set, a mouse organ data set, and a mouse breast cancer data set, we show that the complex enrichment approach (CEA) significantly improves protein identification and biological interpretation in shotgun proteomics data.

First, we demonstrated the accuracy and robustness of CEA through cross-validation studies. CEA achieved an accuracy of 0.85 with a sensitivity of 0.55 in the yeast cell culture data set. It outperformed other network-based methods in both accuracy and robustness. Secondly, applying CEA on the mouse organ data set increased the number of identified proteins by around 15% in individual organs. 92% of the rescued proteins were supported by existing transcriptome profiling studies or publications on corresponding organs. Finally, in the mouse breast cancer data set, CEA increased protein identification by 10% and 30% in the tumor and normal tissues, respectively. Among the 102 rescued proteins in the tumor tissue, 72% and 31% had been reported in cancer- and breast cancer-related publications, including products from some well-known breast cancer genes such as *Ctnnb1* and *Top1*. Moreover, CEA organizes identified proteins into complexes and thus provides a natural way to compare and interpret proteomics data at the modular level. Comparison of the normal and tumor tissues identified cancer-specific modules that corresponded to important biological processes involved in tumor biogenesis and progression, such as “programmed cell death”, “cell-matrix adhesion”, and “Wnt receptor signaling pathway” etc.

In conclusion, CEA is an accurate and robust approach that can be easily incorporated into routine shotgun proteomics protein assembly pipelines to improve protein identification and to gain a modular view of the identified proteins.

Cross-Species Comparison of Nucleosome Positioning Signals

Lu Zhang[1], Randy Wu[2], Hao Li[2], Jeffrey H. Chuang[1]

[1]Boston College Department of Biology; [2]UCSF Department of Biochemistry and Biophysics

How nucleosomal organization is specified by genomic sequence is unclear, although DNA affinities to nucleosomes can vary over a 1000-fold range. Current efforts on deciphering nucleosome positioning code focus on two paradigms. One employs dinucleotide periodicity, which describes locations which nucleosomes favor. However, existing methods can localize nucleosomes only 15% better than random guessing. Another approach is to look for signatures that define nucleosome free regions. One major type of nucleosome-free region in eukaryotic genomes is ~200 bp upstream of the transcription start site. Previous studies in S. cerevisiae reveal poly (dA:dT) tracts in these regions. We extend this analysis into the malaria species: *Plasmodium falciparum, Plasmodium yoelii, and Plasmodium vivax*, as well as the additional eukaryotic parasite *Cryptosporidium parvum*. Although these species have unusually high AT% in their promoter regions, there is still a significant poly (dA:dT) signal peak ~100 bp upstream of the transcription start site (TSS). This signal differs in order and position from yeast results, with stretches of 6+ poly A/T most significant and poly (dT) proximal to the TSS. These results indicate that poly (dA:dT) signal is also prevalent in malaria genomes and should play a functional role. This signal is likely to exclude nucleosomes from the area by encoding a non-favorable binding signal. This feature may shed light on nucleosomal organization and will improve the prediction of TSS.

Boosting The Performance Of Inference Algorithms For Transcriptional Regulatory Networks Using A Phylogenetic Approach

Xiuwei Zhang[1,2], Bernard M.E. Moret[1,2]

[1]Laboratory for Computational Biology and Bioinformatics, EPFL (Ecole Polytechnique Fédérale de Lausanne), Lausanne, Switzerland; [2]Swiss Institute of Bioinformatics.

Inferring transcriptional regulatory networks from gene-expression data remains a challenging problem, in part because of the noisy nature of the data and the lack of strong network models. Time-series expression data have shown promise and recent work by Babu *et al* on the evolution of regulatory networks in *E. coli* and *S. cerevisiae* opened another avenue of investigation. In this paper we take the evolutionary approach one step further, by developing ML-based refinement algorithms that take advantage of established phylogenetic relationships among a group of related organisms to improve the inference of these networks for these organisms from expression data gathered under similar conditions.

Our approach places the predicted networks for a collection of organisms at the leaves of the phylogenetic tree of these organisms, uses a standard maximum-likelihood approach to infer the ancestral networks, and then uses the information in the ancestral networks to refine the networks at the leaves. Our current implementation uses a simple evolutionary model for the regulatory networks where the networks are represented by binary adjacency matrices and the evolutionary events are gains and losses of edges.

We use simulations with different methods for generating gene-expression data, different phylogenies, and different evolutionary rates, and use different network inference algorithms, to study the performance of our algorithmic boosters. The results of simulations (including various tests to exclude confounding factors) demonstrate clear and significant improvements (in both specificity and sensitivity) on the performance of current inference algorithms. Thus gene-expression studies across a range of related organisms could yield significantly more accurate regulatory networks than single-organism studies.

Finally, our refinement algorithms can work on a set of regulatory networks from different sources which are to be refined, not only the networks inferred from gene-expression data.

Detecting Interactions among Genes using Generalized Logical Network Modeling with Permutation Tests

Mingzhou (Joe) Song, Yang Zhang, and Samuel D. Palmer
Department of Computer Science, New Mexico State University

Comparative gene expression profiles of *in-silico* gold standard networks from DREAM3 Challenge 4 were quantized into discrete variables prior to modeling the interactions. Using the generalized logical network modeling algorithm, significant interactions among genes were identified using a permutation test.

Quantizing continuous variables by sequential dynamic programming
When the quantization of continuous random variables into discrete random variables is performed, it is important to preserve interactions between variables and avoid creating interactions that are not present in the original data. We attempt to maintain the original distribution of data by using a sequential dynamic programming algorithm to maximize the log likelihood of the entire quantization grid one dimension at a time. This approach is greedy but it allows us to quantize variables based on all dimensions unlike a single dimensional quantization using *k*-means or the mutual information based pair-wise approach.

Reconstructing generalized logical networks based on chi-square statistics
The generalized logical network modeling algorithm determines significant interactions among genes. Using inference of the *K*-th order the temporal information in the data sets provides a basis to examine interactions. In a generalized logical network, a generalized truth table, associated with each gene, describes its behavior dictated by some other influential genes. The optimal logics at each node in the network will be searched so that they best explain the observed data. Determination of an optimal logic will involve parent node selection and generalized truth-table generation. The maximum number of parents is set to a given number. If the current node shows consistent behavior during transition from one state to another given the parent nodes, then the parent nodes will be kept. The actual goodness of the transition will be calculated using the chi-square statistics in conjunction with a permutation test to determine its statistical significance.

Using permutation test to obtain exact *p*-values for the chi-square statistics
In the permutation test, a statistical significance test, the test statistics are calculated under random rearrangements of observed time point values. In this case, we permuted the order of time points in trajectories separately for each gene. For each permutation we calculated chi-squares for each node in the generalized logical network model under various combinations of the number of parents and the Markovian order. Tabulating the statistics, we obtain the null distributions of the chi-square statistics, with which as a reference, we perform permutation tests in generalized logical network modeling when selecting the best parents and getting exact *p*-values for each candidate interaction.

Generalized logical network modeling can readily incorporate both the steady states (heterozygous knock-down and null-mutants experimental data) and transient states to the optimal estimation of gene interactions. Using a zero-th Markovian order enabled generalized logical network, we can capture dynamic interactions not only between different time points but also within the same time slices, which makes the data-driven method much flexible.

Identification of Antioxidant Control Models in Cell Survival / Death Decision Systems

He Zhao[1], Bahrad Sokhansanj[1]

[1]*School of Biomedical Engineering, Science, and Health Systems, Drexel University, USA*

Molecular biology experiments have conventionally emphasized descriptive, qualitative outcomes. A data-driven model of the changes in phenotypic variables upon perturbations is important to characterize the relative role of system components. Immune cells are particularly susceptible to damage generated by reactive oxygen species (ROS) in the course of inflammation. Apoptosis of irreparably damaged cells provides a means of preventing dysfunction and potential necrotic release of DNA and other cell products, which prevents inflammation from resolving. Studies show that ROS plays a complex role in regulating cell death and survival. We described and implemented methodology to generate quantitative information for cell phenotype in response to stress based on quantitative experimental data. In addition, our stochastic individual cell-based model that included ROS and major components of the cell survival/death decision system evaluated the effect of antioxidants.

Towards this goal, we developed and tested a cell population-based differential equation modeling to analyze flow cytometry data on a specific test case: the role of intracellular ROS and mitochondrial membrane potential (MMP) in Jurkat T lymphocyte death induced by CCCP (carbonyl cyanide m-chlorophenylhydrazone), a strong mitochondrial uncoupler and the effect of antioxidant N-acetylcysteine (NAC) on cell death. We used both morphological and histogram-based metrics to define cell population fractions instead of the traditional identification using the fluorescence intensity arbitrary values. This approach reduced the noise caused by the potential artifactual oxidation of dyes through dye-based methods in flow cytometry and provided the valuable data to better understand the mechanisms of system perturbations.

We applied nonlinear parameter estimation methods to relate the model to experimental data and analyze robustness of model fitting. Modeling allows for quantitative definition of the complex time-dependent impact of CCCP on cell phenotype. Two independent “training” and “test” time series were obtained for ROS, MMP, and cell viability through flow cytometry experiments. Nonlinear parameter estimation was performed on each data set, and parameters and predictions were compared between them to evaluate model robustness. CCCP was found to directly lead to Jurkat cell death, accompanied by increased ROS and mitochondrial membrane permeability. Model estimation on experimental data was quantitatively robust for MMP and cell death. ROS measurements were less reproducible, though consistently models predicting that ROS mediates the lethal effect of CCCP fit data better than those assuming direct mediation by MMP depolarization. This is consistent with additional experimental results showing that antioxidant N-acetylcysteine (NAC) inhibited the effects of CCCP on ROS and cell death. This “top down” system model is a necessary step for individual patient-based design of antioxidant and anti-inflammatory therapies. Indeed, our basic approach can be generalized to quantitatively test hypotheses on other kinds of measurements of multiple intracellular fluorescent markers within practical limits on data set size and quality.

A formal model for analyzing and predicting combination effects for biosystems based on Bliss independence method

Han Yan[1], Shao Li[1], Qianchuan Zhao[1]

[1]Department of Automation and TNList, Tsinghua University, Beijing 100084, China

Drug combination therapy is commonly used in clinical practice. Drug targets (most commonly proteins) design is therefore one of the hot research topics. As the foundation of drug targets design, numerous models to evaluate combinational effects (mostly synergism and antagonism) between drugs have been well developed. Combinational effects could be considered as system behaviors under perturbations on multi-targets and are relative to system structures (that is biochemical reactions network). Many methods based on evaluation models have been generally used in simulations and experiments for drug targets design by analyzing the relationship between system structures and combination effects. Some empirical results about the relationship have been obtained. However, these are not sufficient for drawing theoretical conclusions. Here, we present a novel method to analyze the relationship between structures and effects for combinational drug targets design from a mathematical aspect. Based on Bliss independence method which is one of the primary evaluation models, mathematical expressions of basic concepts like survival ratio in the basis method are given. And a new index named synergism assessment factor derivative is induced from the Bliss independence method, with which the traditional ODEs model of the biological system could be connected to combinational effects. Besides for specific simplified models it has been proved that synergism assessment factor derivative of the simplified model is in direct proportion with that of the original model. Using this index the generation of combination effects of some classic structure types has been analyzed. It showed that some types of structure (like OR relation between the combination targets, that means either branch of a biosystem where combination targets is could activate downstream reactions) must generate synergistic effect under any combinational perturbations, some types like AND relation must generate antagonistic effect or keep Bliss independent, and others, like phosphorylation cascade (e.g. IKKK-IKK cascade in NFκB pathway), may generate synergistic or antagonistic effects according to the variation ranges of parameters. These results supported some former simulation or experiments results. With the help of these conclusions about structures, it is more convenient to identify combinational drug targets in signal transduction pathways like NFκB pathway.

Combining Gene Expression and Cross-species conservation to discover TFBS motifs

Xiang Zhou[1,*], Pavel Sumazin[2,*] and Andrea Califano[1,2,3,#]

[1] Department of Biomedical Informatics (DBMI), Columbia University, New York, NY 10032, USA

[2] Center for Computational Biology and Bioinformatics (C2B2), Columbia University, New York, NY, USA.

[3] Institute for Cancer Genetics and Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA

* These authors contributed equally.

Gene expression and cross-species conservation data have been widely used to guide computational *cis*-regulatory motif discovery algorithms, and the integration of the two data promises to further improve discovery accuracy. We constructed an extensive human-promoter platform for estimating motif discovery accuracy, and used this platform to compare discovery approaches that use gene expression and cross-species conservation. We addressed fundamental problems for the use of each of the data individually, and compared integration methods for these orthogonal data. We showed that a regulatory network reverse-engineering algorithm that is designed to identify direct transcription factor targets identifies binding-site-enriched promoters with significantly better accuracy than traditional co-expression. Combining motif discovery in target promoters, alignment-based cross-species conserved regions in these promoters, and pattern-discovery-based cross-species conserved regions in these promoters significantly improved accuracy over methods that use gene expression alone. Finally, we showed that a careful greedy integration of the two data outperformed a recent Gibbs sampling strategy that integrates the two data simultaneously. We validated predicted binding sites derived from known motifs, *de novo* discovered motifs, and motifs associated with predicted co-factors. Our results demonstrate that gene expression and cross-species conservation data can be used to dramatically improve motif discovery accuracy. Our discovery approach uses these data to improve discovery accuracy on multiple levels, and our unbiased test platform is a general resource for estimating motif discovery accuracy in future studies.

Hidden Markov Model Clustering of Chromatin Marks Correlates with Functional and Conserved Elements

Jason Ernst[1,2], Manolis Kellis[1,2]

[1] *Computer Science and Artificial Intelligence Laboratory of MIT*; [2] *Broad Institute of MIT and Harvard*

The high resolution mapping of a diverse set of histone modifications across the human genome provides the opportunity to identify histone modification signatures associated with known and novel classes of genomic elements. We investigated using a set of 38 histone modifications for de-novo segmentation of the human genome and assignment of these segments into different clusters based only on its histone modification signature. For this we used a hidden markov model where for each hidden state there was an emission probability associated with each histone modification. The model learned these emissions probabilities as well as the transition probabilities between each hidden states from the data. Using the viterbi algorithm each location in the genome was assigned to its most likely hidden state under the model providing a clustering of the genome based on hidden state assignments. When analyzing these clusters we found a number of them to be significantly enriched for specific genomic elements such as transcription start sites, transcribed regions, exons, 5’ UTRs, and 3’ UTRs. A number of the clusters were also significantly enriched for highly conserved elements. Further analysis of the histone modification and conservation signatures associated with these clusters has the potential to yield insights into novel or underappreciated classes of genomic elements.

Novel microRNA Predictions in Human ES Cells and their Differentiated Progeny

Goff, L.A, Swerdel, M., Moore, J., Adams, C., Hart, R.

Keck Center for Collaborative Neuroscience, Rutgers University

MicroRNAs are often tissue-specific, so we predict that specific stages of human embryonic stem cell (hESC) differentiation will express novel miRNAs that are not observed in mature tissues. Many of these are likely to be important regulators of early differentiation. To identify novel sequences, we turned to deep sequencing, genomic alignment, and computational prediction of miRNA precursors. Using the SOLiD platform (ABI), we obtained >2.5x10^8 small RNA sequences from various undifferentiated and neural-precursor stage hESC lines. Accumulated counts of known miRNA sequences correlate with both qPCR and microarray results from previous studies. We demonstrate, however, that a broader dynamic range and sensitivity are obtained from sequence counts as compared to previous technologies. Using perfect alignments to known miRNA, we are able to clearly discriminate between various hESC lines and differentiation status, agreeing with earlier studies. Since deep sequencing technologies require no a priori knowledge of small RNA sequence, we are able to explore the data for putative novel miRNA sequences. Using a modified Smith-Waterman alignment (SHRiMP) we are able to determine the location of >99% of the small RNA reads within the human genome. We employed a probabilistic model (miRDeep, Friedlander, 2008) to discover novel miRNA sequences from the small RNA alignments as well as from a sample of human heart small RNA alignments obtained using the Solexa platform (Illumina). Several high-scoring predictions are validated using qRT-PCR and represent a selection of novel miRNA sequences within the human genome. These novel miRNAs in most cases demonstrate significant regulation during neural differentiation of hESCs. It is clear that the prevalence of small RNAs within hESC has been underestimated and that knowledge of the full gamut of miRNAs regulated during neural differentiation is required before understanding the cellular mechanisms regulated during this process.

Systems Biology Oral Presentations

(★ Full length manuscript; ► Invited talk; Accepted abstract)

DREAM3 Oral Presentations

Regulatory Genomics Oral Presentations

All Poster Presentations

Index

Author Index

Wednesday, Oct 29, 2008

Systems Biology keynotes:

Aviv Bergman - A. Einstein Colleg.
David Botstein - Princeton
George Church - MIT/Harvard
Todd Golub - Broad Institute
Daphne Koller - Stanford University
Luis Serrano - CRG Spain
John Tyson - Virginia Tech

DREAM3 keynotes:

Doug Lauffenburger - MIT
Boris Kholodenko - T.Jefferson U

Regulatory Genomics Keynotes:

Uri Alon - Weizmann Institute
Chris Burge - MIT
Mark Gerstein - Yale
Tim Hughes - U. Toronto
Bing Ren - UC San Diego
Eddy Rubin - Berkeley/JGI
Leona Samson - MIT BE
Thomas Tuschl - Rockefeller

Wednesday, Oct 29, 2008

5pm	Systems Biology: Registration / Poster SetUp	
5:15	**Pamela Silver: Designing Biological Systems**	1
5:45	Russo/diBernardo: Synchr. biol. clocks	2
6pm	Borenstein/Feldman: Metabolic topo sign	3
6:15	Rodriguez-Martinez/Furman: mRNA fluct	4
6:30	Mezes break - light snacks	
6:45	Burke/Sorger: Single-cell caspase amplif.	5
7pm	Fichtenholtz/Collins: Engin transc profiles	6
7:15	Szczurek/Vingron: Regul informative expts	7
7:30	**Daphne Koller: Activity patterns in biological networks**	8
8pm-11p	**Systems Biology: Welcome Reception Poster Session I** (hors-d'oeuvre, heavier snacks, cash bar)	

Thursday, Oct 30, 2008

8am	Breakfast	
9am	**Todd Gollub: Signatures for Small Molecule Discovery**	9
9:30	Karni/Sharan: Predicting Disease Genes	10
9:45	Iorio/di Bernardo: Network drug action	11
10am	Margolin/Carr: Drug target discovery	12
10:15	Coffee / Snacks / Fruit Break	
10:45	Veeramani/Bader: Metabol. flux in disease	13
11am	Chan/Song: CTCF confines Estrogen Rec	14
11:15	Wang/Schadt: Inter-Species Comparison of L	15
11:30	Miguez/Silver: Cancer bistability/drug resist.	16
11:45	Liu/Kasif: Type2 diabetes dysregulated nets	17
12pm	Lunch Break / Networking Opportunities	
1pm	**David Bostein: Coordination of growth, cell cycle, stress, metabolism in yeast**	18
1:30	Chechik/Koller: Gene expr. response timing	19
1:45	McDermott/Heffron: Virulence bottleneck nod	20
2pm	Qian/Yoon: Querying protein netw. pathways	21
2:15	Coffee Break	
2:45	Sahoo/Dill: Predicting B-Cell Develpt Genes	22
3pm	Beleva/Levchenko: Ecoli TCAcycle centrality	23
3:15	Eden/Alon: Dyn proteome response/cell fate	24
3:30	Degenhardt/Carlberg: Stochastic PPAR-depe	25
3:45-5:15p	**Systems Biology: Poster Session II** (hors-d'oeuvres, snacks, refreshments)	
5:15	**John Tyson: Reaction motifs and funct. modules in prot. networks**	26
5:45	Yosef/Sharan: Funct prot net reconstrction	27
6pm	Leung/Chin: Protein complexes from PPI data	28
6:15	Lee/Shakhnovich: Yeast complex Tx co-regul	29
6:30	Mezes break - light snacks	
6:45	Navlakha/Kingsford: Graph summarization	30
7pm	Shlomi/Ruppin: Human tissue-specific metab	31
7:15	Kuttykrishnan/Brent: Steady state to kinetics	32
7:30	**George Church: From personal genomes and enviromes to traits**	33
8pm-11pm	**Concert: Thousand Days**	

Friday, Oct 31, 2008

8am	Breakfast	
9am	**Aviv Bergman: Evolutionary Systems Biology in Health and Disease**	34
9:30	Hormozdiari/Cherkasov: Indels in prot netws	35
9:45	Mitchell/Pilpel: Adaptive envirmt conditioning	36
10am	Coffee / Snacks / Fruit Break	
10:15	DREAM introduction	
10:30	**Boris Kholodenko: Spatio-temp coding sign spec & netw reconstr**	37
11am	Popescu/D-Kumar: MAPK netw in plants	38
11:15	Sachs/Lauffenburger: Learning sign netw	39
11:30	Castelo/Roverato: Rev eng regulatory netw	40
11:45	Marucci/Bernardo: Rev eng yea synth net	41
12pm	Lunch Break / Networking Opportunities	
1pm	Altan-Bonnet: Challenge 1 intro	42
1:15	Saez-Rodriguez: Challenge 2 intro	43
1:30	Prill: Chllgs 1 and 2 overall results	
1:45	Guex: Challenge 2 best predictor talk	44
2pm	Bourque: Challenge 2 best predictor talk	45
2:15	Short Break	
2:30	Clarke: Challenge 3 intro	46
2:45	Marbach: Challenge 4 intro	47
3pm	Stolovitzky: Chllgs 3 and 4 overall results	
3:15	Gustafsson: Chlenge 3 best predictor talk	48
3:30	Ruan: Challenge 3 best predictor talk	49
3:45	Yip: Chllg 4 best predictor talk	50
4pm	Short Break	
4:15	**Doug Lauffenburger: Signaling netw in prim & transf hepatocytes**	51
4:30	DREAM3 - Future directions & adjourn	
5pm	Regulatory Genomics: Registration,Poster SetUp	
5:15	**Thomas Tuschl: Post-transcriptional networks**	52
5:45	Zhou/Zhang: anti-sense small RNAs	53
6pm	Khan/Marks: siRNAs vs. endogen. resp.	54
6:15	Gitter/Bar-Joseph: Euk code redundancy	55
6:30	Mezes break - light snacks	
6:45	**Mark Gerstein: Human intergenic annotation**	56
7:15	Erives: Ciona enhancer structure	57
7:30	Busser/Michelson: Muscle regulation	58
7:45	Megraw/Ohler,Hatzigeorgiou: TSS code	59
8pm-11p	**Regulatory Genomics: Halloween Poster Session III** (hors-d'oeuvre, snacks, refreshments, tricks, and treats)	

Saturday, Nov 1, 2008

8am	Breakfast	
9am	**Uri Alon: Design principles of biological systems**	60
9:30	Ahmed,Song/Xing: Temporal networks	61
9:45	Mezey/Logsdon: Bayesian networks	62
10am	Peterson/Vilo: Embryonic networks	63
10:15	Coffee / Snacks / Fruit Break	
10:45	Huttenhower/Coller: Integration networks	64
11am	Tchagang/Pan: Dynamic Factor Analysis	65
11:15	Madar/Bonneau: Coupled netw dynamics	66
11:30	Chen/Wang: Cancer knowledg integration	67
11:45	Zinman/Bar-Joseph: Cross spec expr clus	68
12pm	Lunch Break / Networking Opportunities	
1pm	**Bing Ren: Chromatin signatures of transcriptional enhancers**	69
1:30	Yuan: Histone mark recruitment prediction	70
1:45	Jothi/Zhao: Binding from ChIP-Seq	71
2pm	Ernst/Bar-Joseph: TF binding predition	72
2:15	Coffee / Snacks / Fruit Break	
2:45	Wasson/Hartemink: TF/nucl. Competition	73
3pm	Yeo/Clarke: TF targets from expr/binding	74
3:15	Halperin/Shamir: Modules from expr/seq	75
3:30	He/Sinha: Motif clustering/conservation	76
3:45-5:15p	**Regulatory Genomics: Poster Session II** (hors-d'oeuvres, snacks, refreshments)	
5:15	**Eddy Rubin: Human Enhancers**	77
5:45	DePace/Eisen: Fly embryo express. atlas	78
6pm	Frise/Celniker: Embryonic image mining	79
6:15	Li/Chuang,Guo: Brain express. enhancers	80
6:30	Mezes break - light snacks	
6:45	**Wang/Chris Burge: Tissue- and factor-specific RNA processing**	81
7:15	Foat/Stormo: Structural mRNA motifs	82
7:30	Reiss/Baliga: Reg. logic in coding regions	83
7:45	Waldman/Ruppin: Cancer transl. efficiency	84
8pm-11p	**Reception: A night at the MIT Museum**	

Sunday, Nov 2, 2008

8am	Breakfast	
9am	**Leona Samson: Predictors of interindividual diffs in DNA damage response**	85
9:30	Rosenfeld/Zhang: Single nucleo. repress.	86
9:45	Verlaan/Pastinen: Cis-regultory variation	87
10am	Lee/Bussemaker: Networks from linkage	88
10:15	Coffee / Snacks / Fruit Break	
10:45	Li/Leslie: Motifs from expr. trajectories	89
11:00	Li: EM spaced dyad motif discovery	90
11:15	Noyes/Wolfe: 1-hybrid TF characterization	91
11:30	**Tim Hughes: Protein-nucleic acid interaction mapping**	92
12pm	Adjourn	

Partner Scientific Journals

Journal of Computational Biology

Genome Research

Nature/EMBO Molecular Systems Biology

Local Host Institutions

Massachusetts Institute of Technology

BROAD INSTITUTE of MIT and Harvard

Computer Science and Artificial Intelligence Lab CSAIL

REGULATORY GENOMICS - SYSTEMS BIOLOGY - DREAM3   2008