Chapter 7

Energy-Based RNA Consensus Secondary Structure Prediction in Multiple Sequence Alignments

Stefan Washietl, Stephan H. Bernhart, and Manolis Kellis

Abstract

Many biologically important RNA structures are conserved in evolution leading to characteristic mutational patterns. RNAalifold is a widely used program to predict consensus secondary structures in multiple alignments by combining evolutionary information with traditional energy-based RNA folding algorithms. Here we describe the theory and applications of the RNAalifold algorithm. Consensus secondary structure prediction not only leads to significantly more accurate structure models, but it also allows to study structural conservation of functional RNAs.

Key words RNA structure, Consensus structure, Structure prediction, Functional RNA

1 Introduction

1.1 Conserved RNA Secondary Structures in Functional RNAs

As for most biological macromolecules, there is a close connection between the function of an RNA molecule and its structure. Therefore, accurate structure predictions can help to better understand RNA functions. This applies to the whole functional spectrum of RNAs, which reflects a diverse structural spectrum ranging from simple hairpin structures up to the intricate multi-loop and pseudoknot structures that build up complex molecular machines such as the ribosome.

Often, evolutionary processes lead to distinct and characteristic evolutionary signatures for different types of functional molecules. For structural RNA molecules, it is the structure and not necessarily their primary DNA sequence which is selectively maintained during evolution. Because of the simple rules that govern their secondary structure, structural RNAs provide exceptionally clear patterns of selection with base pairing patterns directly reflecting structural conservation. With the exception of nonstandard base pairs, RNA secondary structures are generally formed by the Watson/Crick base pairs A·U and G·C as well as G·U pairs.

Jan Gorodkin and Walter L. Ruzzo (eds.), *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, Methods in Molecular Biology, vol. 1097, DOI 10.1007/978-1-62703-709-9_7,

Two nucleotides that form a base pair may be changed by mutations but preserve the propensity to form a valid base pair. We distinguish consistent mutations and compensatory mutations. A consistent mutation changes one base (e.g., A·U ↔ G·U) while compensatory mutations change both bases in the base pair (e.g., A·U ↔ G·C or C·G ↔ G·C).

These mutational patterns have some concrete, practical implications. Interpreted as "evolutionary signatures" they can help to decide if some RNA forms a functional structure or if some region of genomic DNA is likely to be expressed into a functional RNA [1]. Another very useful aspect is the fact that the additional evolutionary information from homologous sequences greatly enhances our ability to predict accurate secondary structure models. The combination of this information with traditional thermodynamic folding algorithms will be the topic of this chapter.

1.2 Strategies to Predict Consensus Structures

Assuming several homologous RNA sequences form the same or very similar structures, our goal is to predict the *consensus structure* common to these sequences. Analyzing evolutionarily conserved structures generally requires a direct or indirect estimation of the underlying sequence alignment, which can itself depend on the structure or not.

There are three main strategies to consensus structure prediction, which differ in how structure prediction and sequence alignment are combined [2]: (1) The alignment is created first based on the sequence alone and the structure is predicted afterward, (2) The structure is predicted first and the structures are aligned afterwords, (3) sequence alignment and structure predictions are performed simultaneously.

In this chapter, we focus on the first approach which is conceptually the simplest. We are given a sequence alignment for which we want to predict a consensus structure. Clearly, this approach can only be successful if the alignment reflects the structural properties. Extensive benchmarks [3, 4] showed that if sequences are similar enough (more than ≈70% sequence identity) this requirement is usually met. It also does not matter which of the many sequence alignment programs are used because on high-sequence similarity data these programs agree well. So, for the purpose of getting high-quality alignments, high-sequence similarities are desirable. In contrast, for the purpose of structure prediction the opposite is true. It is only possible to efficiently exploit the evolutionary signal if there is enough evolutionary divergence in the sequences compared. Fortunately, in many practical applications we have access to data sets at reasonable range of sequence similarity, i.e. similar enough to be reliably aligned and diverse enough to infer secondary structures from mutation patterns.

Strategy (2) is rarely used in practice. One program of this class is `RNAforester`, which uses a tree alignment algorithm to align sequences based on their pre-calculated structure [5].

Strategy (3) solves the problem in the most rigorous way and is clearly the most appealing. Algorithms following this approach are usually variants of the well-known Sankoff algorithm [6]. The main drawback of these algorithms is their computational complexity in terms of both CPU time and memory. However, more recent variants of the Sankoff algorithm introduced several heuristics that improved the performance considerably. Some of these algorithms are discussed in Chapters 13–15.

Here, we concentrate on the case of a fixed alignment. In some rare cases there is enough evolutionary information that allows to predict an accurate structure exclusively from the analysis of the mutational patterns. Such covariation analyses typically use a mutual information score to find columns that show highly correlated mutation patterns. For example, this method led to surprisingly accurate structures for rRNAs already 30 years ago [7].

In practice, however, most data sets do not contain enough evolutionary information to predict reliable structures based on this information alone. In that case, thermodynamic folding algorithms might give better results. An obvious solution is to combine both strategies. This approach was first taken 2002 when `RNAalifold` [8] was presented as a tool to fold aligned sequences by extending Zuker's algorithm for folding single sequences. `RNAalifold` is now part of the widely used Vienna RNA packages and is routinely used to predict consensus structures by many researches.

2 The RNAalifold Algorithm

2.1 Averaged Energy Minimization for Multiple Alignments

`RNAalifold` extends the RNA structure prediction algorithms based on the nearest neighbor energy model in two ways: It averages the energy contributions over the sequences in the alignments and incorporates phylogenetic information as "pseudo-energies" into the energy model.

First, we want to recall the algorithm to predict the minimum free energy as introduced by Zuker and Stiegler [9] (refer to Chapter 4 for details). The recursions to calculate the minimum free energy F on a subsequence from i to j can be written as follows:

$$F(i,j) = \min\{F(i+1,j), \min_{i<u\leq j} C(i,u) + F(u+1,j)\} \quad (1)$$

$$C(i,j) = \min\{\mathscr{H}(i,j), \min_{i<k<l<j}\{C(k,l) + \mathscr{I}(ij,kl)\},$$

$$\min_{i<u<j} M(i+1,u)M^1(u+1,j-1) + a\}$$

$$M(i,j) = \min\{M(i+1,j) + c, \min_{i<u\leq j} C(i,u) + b + (j-u)c,$$
$$\min_{i<u<j} C(i,u) + b + M(u+1,j)\}$$
$$M^1(i,j) = \min\{M^1(i,j-1) + c, C(i,j)\}$$

$C(i,j)$ is the minimum free energy given i,j form a base pair. The technical details of the multi-loop matrices M and M^1 will not concern us here. Relevant to understand how `RNAalifold` extends the single sequence folding are the terms $\mathscr{H}(i,j)$ (the free energy of a hairpin between bases i and j) and $\mathscr{I}(ij,kl)$ (the free energy of an interior loop between base pairs i,j and k,l).

In `RNAalifold`, the free energy contributions from $\mathscr{H}(i,j)$ and $\mathscr{I}(ij,kl)$ are replaced by the mean contributions over all m sequences s of the alignment $\mathscr{A}$. The recursion for matrix C in the case of a multiple alignment thus reads:

$$C^{\mathscr{A}}(i,j) = \min \begin{cases} \frac{1}{m}\sum_{s\in\mathscr{A}} \mathscr{H}(i,j,s), \\ \min_{i<k<l<j}\{C^{\mathscr{A}}(k,l) + \frac{1}{m}\sum_{s\in\mathscr{A}} \mathscr{I}(ij,kl,s)\}, \\ \min_{i<u<j} M(i+1,u)M^1(u+1,j-1) + a \end{cases} \quad (2)$$

Here, the indices i, j, k, and l correspond to columns of the alignment $\mathscr{A}$. While this step is straightforward, some problems arise. One of them is the presence of gaps in the alignments. Clearly, a gap character is not considered in the standard energy model but needs to be dealt with during the energy evaluation step. It can cause problems, for example, when evaluating loop energies that depend on the loop length or assigning dangling energies to base pairs. As a solution, energy contributions are generally computed from the *gap free* sequences s. To this end, a map from the alignment column to sequence index is computed for every sequence s before the recursion is started.

Another problem may arise when columns i and j cannot pair for some sequences of the alignment—either because the bases do not form a valid base pair or there are gaps at one or both of these positions. All these cases are treated as "nonstandard" base pairs which are considered in the standard thermodynamic model. As a consequence, these positions are assigned unfavorable energies and do not require special treatment in the algorithm.

2.2 A Simple Covariance Score

The second extension to single sequence structure prediction is the introduction of covariance terms that take into account the phylogenetic signal of a conserved structure. `RNAalifold` uses two distinct approaches for this.

The simpler one calculates the sum of the hamming distances of all sequence pairs s_α and s_β at columns i and j if they can form a valid base pair between i and j. More formally, with h being the hamming distance, $\mathscr{B}$=AU,CG,GC,GU,UA,UG the set of allowed base pairs, the conservation score γ' between two columns i and j can be written as:

$$\gamma'(i,j) = \frac{1}{2} \sum_{s_\alpha, s_\beta \in \mathscr{A}} \begin{cases} h(s_\alpha(i), s_\beta(i)) + h(s_\alpha(j), s_\beta(j)) & \text{if } (s_\alpha(i), s_\alpha(j)) \in \mathscr{B} \\ & \wedge (s_\beta(i), s_\beta(j)) \in \mathscr{B} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This score only considers mutations that support a given base pair. Counterexamples, i.e. sequences where the base pair cannot be realized, are be penalized by another simple ad hoc score, leading to the total score γ:

$$\gamma(i,j) = \gamma'(i,j) + \delta \sum_{s \in \mathscr{A}} \begin{cases} 0 \text{ if } (s(i)s(j)) \in \mathscr{B} \\ 0.25 \text{ if } s(i) \wedge s(j) \text{are gaps} \\ 1 \text{ otherwise} \end{cases} \tag{4}$$

Here, the parameter δ balances between the covariation score supporting a base pair i, j and the penalty for counterexamples that cannot form a base pair.

Since γ is used as a "pseudo" energy during the energy minimization step, we need another parameter, β to balance between the thermodynamic and the phylogenetic score. The default value for both parameters δ and β is 1.

The conservation term γ is also used to decide whether a base pair is considered possible between two alignment columns. Base pairs are only evaluated if γ reaches a certain threshold. This reduces the number of possible base pairs. That is the reason why the speed of `RNAalifold` is usually not much slower than folding a single sequence although in theory the complexity of the algorithm is $\mathscr{O}(n^3 m)$ with n the length of the alignment and m the number of sequences.

2.3 An Improved Covariance Score Based on RIBOSUM Matrices

Many successful applications of `RNAalifold` over the past years demonstrate the power of this relatively simple scoring scheme. However, it suffers from one major drawback: Highly conserved base pairs not necessarily have a high covariance score and therefore only a few counterexamples can destroy a base pair in the consensus structure. In particular, this problem affects alignments with many sequences and high sequence conservation. In an alignment with 500 sequences, for example, three sequences that cannot form a

base pair would result in a prohibitive penalty even though all the other sequences have a totally conserved base pair in this position.

A new scoring scheme was introduced in 2008 to overcome this problem [10]. In this model, RIBOSUM like scoring matrices [11] are used as a replacement for the hamming distance-based scores. To obtain these scoring matrices, log-likelihood scores are calculated from alignments of ribosomal RNAs.

$$R(ab, cd) = \log \frac{f(ab, cd)}{f(ac)f(bd)} \tag{5}$$

Here, $f(ab, cd)$ is the frequency that base pair ab is aligned to base pair cd, and $f(ac)$ the frequency that base a is aligned with c. Following the example of the original RIBOSUM and BLOSUM matrices, these log-likelihood scores are computed for different subsets of the rRNA alignments, containing sequences of different sequence conservation. The RIBOSUM matrix to be used to score an alignment is the one that most closely reflects the minimum and maximum sequence distances in the alignment. `RNAalifold` compares the sequences of an input alignment and chooses the appropriate scoring matrix.

The new covariance score γ' can then be written as:

$$\gamma'(i,j) = \frac{1}{2} \sum_{\substack{s_\alpha, s_\beta \in \mathscr{A} \\ s_\alpha \neq s_\beta}} R\left(s_\alpha(i)s_\alpha(j), s_\beta(i)s_\beta(j)\right) . \tag{6}$$

The log-likelihood scores are scaled to approximately achieve the same absolute values as the hamming distance scores in Eq. 3. In spite of that, the relative impact of the covariation scoring increased from 5% to 50%, making it necessary to also adapt the factors β and δ. The new default values are $\beta = 0.6$ and $\delta = 0.5$. Overall, RIBOSUM scoring performs about 10% better than hamming distance scoring [10].

2.4 Extending Other Folding Algorithms to Multiple Alignments

It is important to note that all of the extension to the single sequence folding algorithm in `RNAalifold` only apply to the actual energy computations. This means that all other variants of folding algorithms that are available for single sequences can also be adapted to multiple alignments. In particular, partition function folding [12], stochastic backtracking [13], centroid structure [14], and local folding [15] are all implemented for the multiple alignment case in the Vienna RNA package.

3 Using RNAalifold

In the following sections, we briefly demonstrate how to use the `RNAalifold` program. We assume the reader has access to a UNIX like environment such as GNU/Linux or Mac OS X with the latest Vienna RNA package installed. Refer to **Notes 1** and **2** for details how to install and use the Vienna RNA package.

3.1 A Simple Example

To use `RNAalifold`, we need a multiple alignment of several homologous sequences. In principle, a pairwise alignment is sufficient although having more sequences clearly improves the results.

`RNAalifold` expects the alignment to be formatted in `CLUSTAL W` format (*see* **Note 3** on how to obtain and format alignments). As a short example, we use an alignment of 20 different isolates of Peach latent mosaic viroid. The short regions covers the so-called hammerhead ribozyme, a self-cleaving structural RNA. To predict the consensus secondary structure of this motif we run `RNAalifold` on the alignment file:

```
$ RNAalifold hammerhead.aln
```

which gives the following result:

```
20 sequences; length of alignment 55.
CAAAAGUCUGGGCUAA_GCCCACUGAUGAGUCGCUGAGAUGCGACGAAACUUUUG
(((((((..((((....)))).........((((.....))))).)))))))
 minimum free energy = -18.58 kcal/mol (-15.36 + -3.21)
```

The output shows the consensus sequence of the alignment and the predicted consensus structure in dot bracket notation. Every base pair is denoted by a pair of brackets "(" and ")" while unpaired positions are shown as dots ".". `RNAalifold` also reports a "minimum free energy" of the consensus structure of $-18.58 = -15.36 + (-3.21)$ kcal/mol. The value consists of two terms. The first term is the average free energy of the structures and the second part is the covariance term. A negative covariance term indicates the presence of many compensatory or consistent mutations supporting the structure. In contrast to the average free energy, the covariance term has no biophysical meaning and thus the `RNAalifold` score is more precisely referred to as a "pseudo-energy."

As mentioned earlier, `RNAalifold` not only implements Zuker's algorithm to find the minimum free energy, but it also implements McCaskill's algorithm [12] to calculate base pair probabilities. To run both variants of the algorithm we use the command line switch "`-p`":

```
$ RNAalifold -p hammerhead.aln
```

The base pair probabilities can be visualized in a so-called dotplot. This and other ways to visualize the structure predictions are described in the next section.

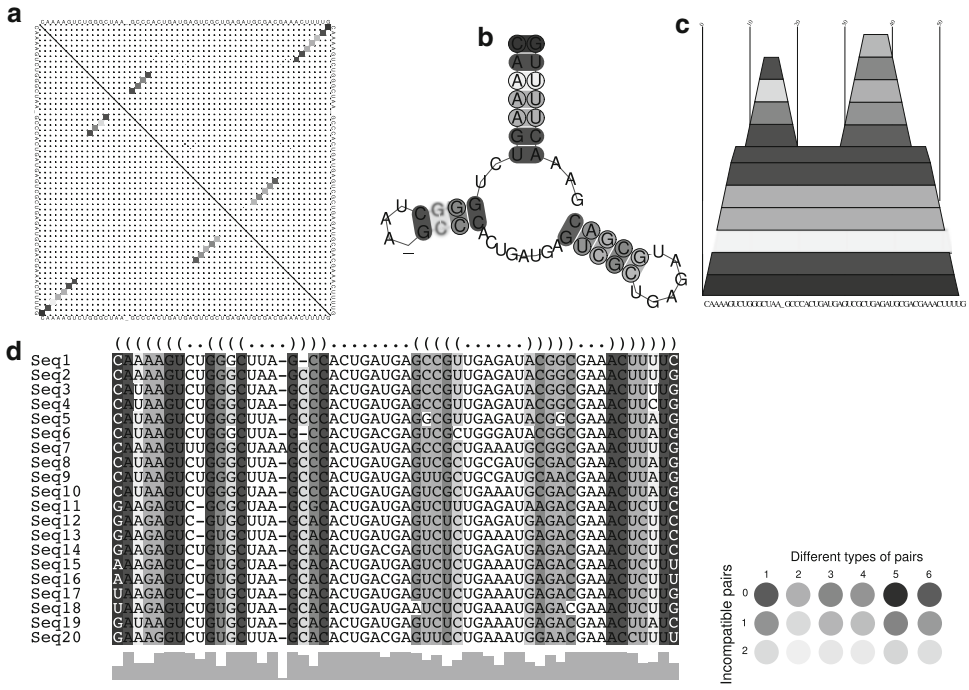

Fig. 1 Visualizations of consensus secondary structures. A color version of this figure can be found online at http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/MiMB/ (**a**) Dotplot of base pair probabilities (*upper right*) and pairing matrix of the optimal consensus structure (*lower left*). (**b**) Traditional RNA representation (**c**) Mountain plot (**d**) Sequence alignment with dot bracket notation of consensus structure

3.2 Visualizing the Results

By default `RNAalifold` generates two representations of the consensus structure (`alirna.ps`) and if run with "`-p`" a dotplot of pair probabilities (`alidot.ps`). However, with additional tools it is possible to create a variety of other visualizations (Fig. 1). In all of these representations the mutational patterns are highlighted using a color code. That way mutations that support a predicted structure (consistent/compensatory mutations) or disrupt a predicted structure (incompatible mutations) can be easily identified. Before we explain this code in more detail, we demonstrate how to generate these pictures on the hammerhead example (*see* **Note 2** for how to view EPS formatted pictures on Linux and OS X).

The dotplot representation (Fig. 1a) is created by default. Every element (i,j) in the matrix represents a base pair formed between i and j. The upper right half of the matrix represents the pair probabilities; the size of every dot is drawn proportional to the probability of this base pair. The lower left half of the matrix represents the optimal structure, i.e. the structure with minimum free (pseudo-)energy. Comparing both parts of the dotplot can be useful to find out whether the optimal predicted structure is well defined or whether there are many alternative

structures possible. In the hammerhead example both dotplots are nearly identical because the prediction is supported by many compensatory mutations which are not compatible with alternative structures.

To obtain a color coded representation of the RNA structure as shown in Fig. 1b, we use the Perl script `colorrna.pl`. This script takes the (black and white) structure plot in `alirna.ps` and the dotplot in `alidot.ps` and writes a colored structure plot which is saved to the file `colored_rna.ps`:

```
$ colorrna.pl alirna.ps alidot.ps > colored_rna.ps
```

To produce a mountain plot (Fig. 1c), we simply run the script `cmount.pl` on the dotplot file `alidot.ps` and save the results in a new postscript file `mountain.ps`:

```
$ cmount.pl alidot.ps > mountain.ps
```

In a mountain plot each base pair (i, j) is represented by a trapeze with a baseline from i to j and a height proportional to the probability of base pair (i, j). Unpaired bases form plateaus in this representation.

To show the consensus structure as a dot bracket string on the sequence alignment itself (Fig. 1d) and to color the alignment columns, we use the script `coloraln.pl`. This script reads the original alignment file and the structure file `alirna.ps` and generates a postscript file which is saved in `color_aln.ps`:

```
coloraln.pl -s alirna.ps hammerhead.aln
    > colored_aln.ps
```

All the different representations use the same color code (a colored version of Fig. 1 is available at: http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/MiMB/). If a paired position (i, j) is formed by the same base pair in every sequence of the alignment, the base pair is shown in red. If i, j is supported by 2, 3, 4, 5, 6 different base pairs, they are colored in brown, green, turquoise, blue, and violet, respectively. In this particular example we see positions that have 2, 3, and 5 different base pairs. For example, the outermost base pair is colored in blue because we observe C·G, G·C, A·U,G·U, U·G providing compelling evidence for selective pressure on this base pair. On the other hand, if one or more sequences cannot form a base pair present in the consensus structure, these positions are shown in pale versions of the respective color. If there are more than two inconsistent base pairs, the position is not colored and appears white.

The colored alignment (Fig. 1d) is particularly useful to quickly assess the structure conservation. Many columns in saturated colors other than red indicate a conserved structure well supported by compensatory mutations. This representation also allows to identify sequences that do not fold into the consensus structure.

In practice, some sequences might be misaligned, have changed their structure, form nonstandard base pairs, or contain sequence errors. All of which leads to incompatible mutations that appear without coloring. Also the hammerhead example contains a few positions in sequences 5, 6, and 18 that are incompatible with the consensus although the general structure is very well conserved in all sequences.

4 Advanced Usage

4.1 Refining Structure Prediction of Individual Sequences

`RNAalifold` predicts the consensus structure of a set of aligned RNA sequences. However, there may be parts of the molecules that are not conserved, e.g. because they are not essential for the function or because they have been the subject to recent structural changes. Furthermore, it is not uncommon that the length of conserved stems and loops varies somewhat between the sequences.

To get a potentially more accurate secondary structure of a single RNA sequence in an alignment, one can use a two-step process. First the consensus structure is predicted using `RNAalifold`. As a second step, the consensus structure is used as a *structural constraint* for single structure prediction using `RNAfold`. The ViennaRNA package provides a Perl script `refold.pl` that uses the output of `RNAalifold` and the input alignment to generate the constraints for `RNAfold` for every sequence in the alignment.

The `refold.pl` approach is especially useful for large alignments with extensive unconserved regions. A good example is the very diverse RNase P RNA class A. We used `RNAalifold` to predict a structure for *Vibrio cholerae* from an alignment of 305 bacterial species and compared it to an experimentally well-established reference structure. The consensus prediction is relatively poor and only achieves a sensitivity and specificity of 38.2% and 68.1%, respectively, of correctly predicted base pairs. Refolding the sequence with constraints from the consensus structure leads to a much improved sensitivity of 59.3% at the cost of a slightly lower specificity of 62.4%.

```
$ RNAalifold < rnasep.aln > rnasep.alifold
$ refold.pl rnasep.aln rnasep.alifold | RNAfold -C
```

Note that consensus base pairs that lead to a positive energy contribution in a particular sequence will not be part of the new prediction for this sequence. Also base pairs that are mutually exclusive to those of the consensus structure will not be included. This can lead to a largely unfolded structure if many consensus base pairs are affected. Consequently, sequences that are largely unstructured where the consensus structure contains base pairs are

most likely wrongly aligned or do not share the same consensus structure as the rest of the alignment.

4.1.1 Local Structure Prediction in Long Alignments

The prediction of secondary structures for long (>500nt) RNA molecules has two major drawbacks: As the algorithm complexity scales cubic in time, the computation time becomes prohibitive for very long sequences. Furthermore, the quality of the predictions decreases rapidly with the length of the molecules. In practice this is not necessarily a problem. For many applications there is no biological reason to assume that a long RNA molecule (e.g., an mRNA transcript of several kilobases in length) has a defined global structure. Often local structural features such as regulatory elements in an mRNA are of interest. In such cases it is useful to predict *local structural components* only.

The program `RNALalifold` is included in the ViennaRNA package since version 2.0. It predicts local structures with a maximum base pair span L (usually L is between 100 and 200 nt). It uses the algorithm by Hofacker et al. [15] to calculate "locally stable closed structures." "Closed" means that there is always an outer base pair i,j if a substructure for the sequence interval i,j is reported (Exceptions are unpaired bases that contribute favorable "dangling energies" for the structure). "Local stability" means that a substructure starting at base i is only reported if there are no substructures with better or equal energy starting at bases $i-1$ or $i+1$.

While `RNALalifold` is not guaranteed to get all important substructures, it usually helps to get a reasonable list of candidates. As an example we calculate local structures for a 1.1-kb human region on chromosome X that contains a cluster of miRNAs (miR-106a, miR-20b, miR-19b2, miR-92a, and miR-363). We use a 12-way alignment of this region with human aligned to 11 other mammalian species.

For a maximum base pair span of 120, the example can be computed like this:

```
$ RNALalifold -L 120 mirsequences.aln
```

The output consists of a list of locally stable structures with the pseudo free energy of the local structure and the start and stop columns in the alignment:

```
12 sequences; length of alignment 1171.

...

(((.(((.....))).))). ( -2.10)  995 - 1014
(((........))). ( -0.28)  976 -  990
((..((((..............))))..)). ( -1.87)  946 -  977

...
```

Because of space constraints in this book, we have only shown three of the predicted local structures. In total `RNALalifold` predicts 33 structures in this region including all 5 known miRNA precursors. As a comparison, the single sequence variant `RNALfold` predicts 125 local structures. This clearly shows that using consensus structure prediction increases the specificity in structure prediction. A single sequence can fold easily into some random structure even though it has no biological meaning. In contrast, it is difficult to find a consensus structure by chance in an alignment of sequences that have accumulated many random mutations. However, the fact that `RNAalifold` or `RNALalifold` predicts a consensus structure does not necessarily mean that it is a biologically relevant structure. To answer this particular question other algorithms and metrics have been developed [16–18]. As a ready-to-use implementation to detect biologically relevant structures the program `RNAz` [19, 20] can be used.

5 The RNAalifold Web-Server

An alternative way to use `RNAalifold` is the public web-server at http://rna.tbi.univie.ac.at presenting and explaining all available options and allows to predict consensus structures and all the various visualization without installing software locally (Fig. 2).

6 Alternative Methods and New Approaches

An alternative to thermodynamic folding algorithms are pure probabilistic methods. Stochastic context free grammars (SCFG) are widely used to model RNA structures (*see* Chapter 5). They have also been used to address the consensus folding problem. `Pfold` [21] explicitly models the evolutionary relationship of the sequences in the alignment using a phylogenetic tree and the RNA structures using a simple grammar. Since both components are probabilistic in nature, `Pfold` can easily combine them in one framework to calculate the probability distribution of all secondary structures for a given alignment and phylogenetic tree.

Seemann et al. extended `Pfold` by incorporating pair probabilities from thermodynamic folding [22]. Their program `PETFold` thus unifies evolutionary and thermodynamic information in this attractive framework.

Despite the differences between these algorithms in how RNA structures are modeled and how evolutionary information is incorporated, all need to solve the same problem of predicting an optimal structure from their model. In the original implementation

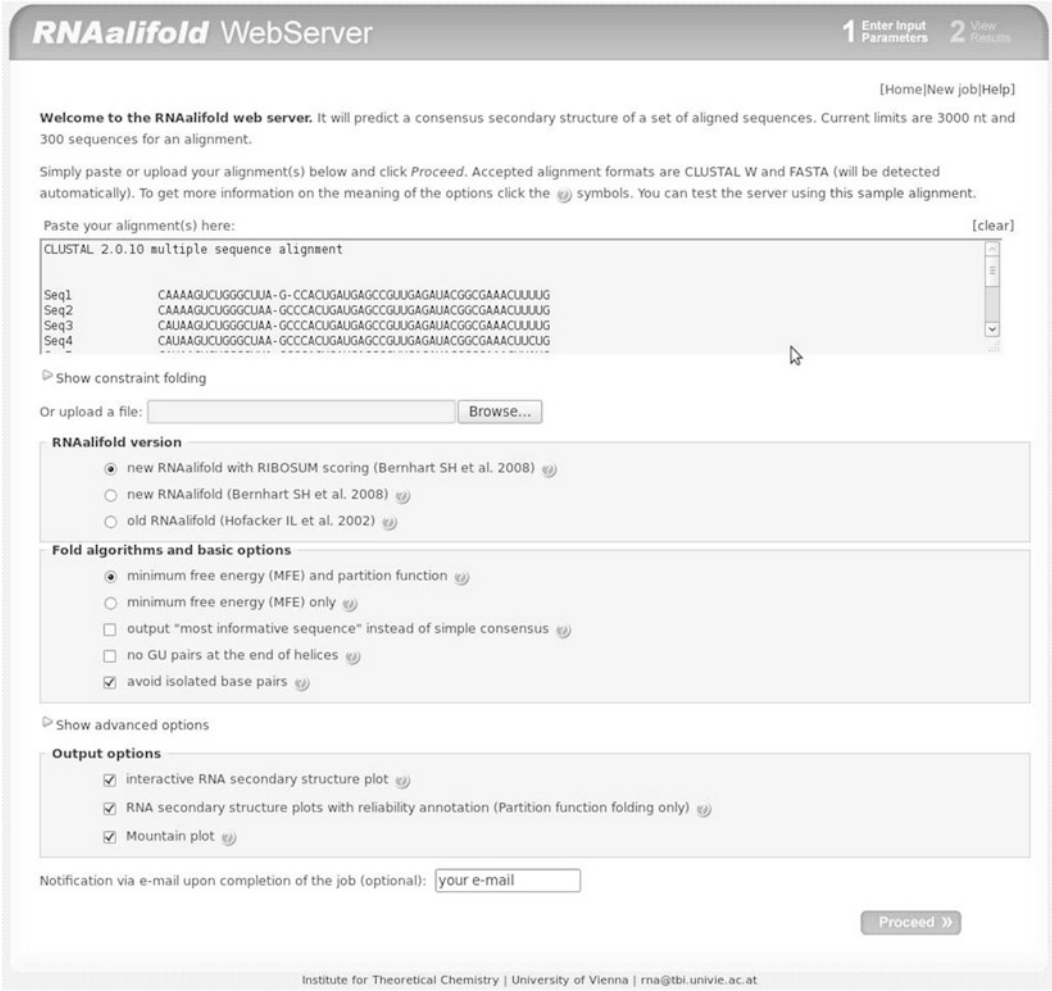

Fig. 2 Screenshot of the `RNAalifold` web-server available at: rna.tbi.univie.ac.at

of `Pfold` this was solved by finding the maximum likelihood solution which corresponds to the minimum free energy solution of `RNAalifold`. A later version of `Pfold` introduced a maximum expected accuracy (MEA) approach. The idea of MEA is to find a structure that when compared to the pair probabilities of all possible structures maximizes the expected base pair accuracy [23, 24]. Also `PETFold` uses this approach and a few other variants of this idea in the context of consensus structure prediction have been proposed [25–27]. In its latest version, also `RNAalifold` predicts MEA structures from the pair probabilities.

7 Notes

Note 1: Setting Up the Environment

To use `RNAalifold` and other programs of the Vienna RNA package, we recommend using a UNIX like environment such as Linux or OS X. All the examples in this chapter assume that you have access to a command prompt and that all relevant programs are installed. To install the Vienna RNA package download the latest `tar.gz` file from http://www.tbi.univie.ac.at/~ivo/RNA/. To install the package run the following on your command prompt:

```
$ tar -xzf ViennaRNA-2.0.0.tar.gz
$ cd ViennaRNA-2.0.0
$ ./configure
$ make
$ sudo make install
```

This will install all the programs under `/usr/local` on you computer. You can test if the installation was successful by running:

```
$ RNAalifold --version
```

If the program cannot be found, you need to make sure that `/usr/local/bin` is within executable path. This is the case by default on Linux. The perl scripts used to generate the various representations are installed by default in `/usr/local/share/ViennaRNA/bin`. Since this path is usually not in the PATH you can either add it in your shell configuration file or explicitly call the programs by their full name.

Note 2: Practical Tips to Use Programs from the Vienna RNA Package

All programs presented in this chapter come with detailed online information. To get help for any programs of the Vienna RNA package you can run:

```
$ RNAalifold --help
$ man RNAalifold
```

The perl scripts contain embedded documentation which can be viewed using `perldoc`:

```
$ perldoc -F `which refold.pl`
```

All the graphical output is formatted as EPS. There are various programs you can use to view an EPS image. Typically the following command will display an eps file on Linux:

```
$ gv image.ps
```

On OS X you can use the `open` command:

```
$ open image.ps
```

EPS images usually can be imported and edited by common graphics programs. Moreover, all EPS images produced by the Vienna RNA package are pure vector graphics that can be scaled and converted to other formats without loss of quality.

Note 3: Creating and Formatting Sequence Alignments

`RNAalifold` expects a multiple sequence alignment in `CLUSTAL W` format. You can use the program `CLUSTAL W` itself which natively produces that output. But also more modern alignment programs such as MUSCLE [28] can create `CLUSTAL W` formatted output suitable for `RNAalifold`. If your alignment program does not support the `CLUSTAL W` format or you are using prebuilt alignments provided in a different format, you might need to convert the format. A typical `CLUSTAL W` alignment looks like this example:

```
CLUSTAL 2.0.10 multiple sequence alignment

Seq1   CAAAAGUCUGGGCUUA-G-CCACUGAUGAGCCGUUGAGAUACGGCGAAA
       CUUUUG
Seq2   CAAAAGUCUGGGCUAA-GCCCACUGAUGAGCCGUUGAGAUACGGCGAAA
       CUUUUG
Seq3   CAUAAGUCUGGGCUAA-GCCCACUGAUGAGCCGUUGAGAUACGGCGAAA
       CUUUUG
Seq4   CAUAAGUCUGGGCUAA-GCCCACUGAUGAGCCGUUGAGAUACGGCGAAA
       CUUCUG
```

There is no formal specification for the program but the first line should start with the word "CLUSTAL." After two empty lines the actual alignment data starts. Every entry has a unique name (no white spaces) and separated by a variable number of spaces a sequence string. For `RNAalifold` it can contain the letters A,U,T,G,C, and a dash ("-") for gaps. For long alignments, the data can be broken up into blocks. This is optional but if the data is given in multiple blocks, each block must contain the same list of names in the same order.

Acknowledgements

Stefan Washietl was supported by an Erwin Schrödinger Fellowship of the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung. Stephan H. Bernhart was funded by the Austrian GEN-AU project "Noncoding RNA." We thank Ivo Hofacker and Benjamin Holmes for comments on the manuscript.

References

1. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF (2007) Structured RNAs in the ENCODE selected regions of the human genome. Genome Res 17:852–864. doi:10.1101/gr.5650707
2. Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 5: 140. doi:10.1186/1471-2105-5-140
3. Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res 33:2433–2439. doi:10.1093/nar/gki541
4. Wilm A, Mainz I, Steger G (2006) An enhanced RNA alignment benchmark for sequence alignment programs. Algorithms Mol Biol 1:19. doi:10.1186/1748-7188-1-19
5. Höchsmann M, Voss B, Giegerich R (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. IEEE/ACM Trans Comput Biol Bioinform 1: 53–62. doi:10.1109/TCBB.2004.11
6. Sankoff D (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J Appl Math 45:810–825
7. Noller HF, Kop J, Wheaton V, Brosius J, Gutell RR, Kopylov AM, Dohme F, Herr W, Stahl DA, Gupta R, Woese CR (1981) Secondary structure model for 23s ribosomal RNA. Nucleic Acids Res 9(22):6167–6189. doi:10.1093/nar/9.22.6167
8. Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. J Mol Biol 319:1059–1066. doi:10.1016/S0022-2836(02)00308-X
9. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–148.
10. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474. doi:10.1186/1471-2105-9-474
11. Klein RJ, Eddy SR (2003) RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 4: 44. doi:10.1186/1471-2105-4-44
12. McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29: 1105–1119. doi:10.1002/bip.360290621
13. Ding Y, Lawrence CE (1999) A bayesian statistical algorithm for RNA secondary structure prediction. Comput Chem 23(3–4): 387–400.
14. Ding Y, Chan CY, Lawrence CE (2005) RNA secondary structure prediction by centroids in a boltzmann weighted ensemble. RNA 11(8): 1157–1166. doi:10.1261/rna.2500605
15. Hofacker IL, Priwitzer B, Stadler PF (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. Bioinformatics 20:186–190. doi:10.1093/bioinformatics/btg388
16. Washietl S, Hofacker IL (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol 342: 19–30. doi:10.1016/j.jmb.2004.07.018
17. Gruber AR, Bernhart SH, Hofacker IL, Washietl S (2008) Strategies for measuring evolutionary conservation of RNA secondary structures. BMC Bioinformatics 9:122. doi:10.1186/1471-2105-9-122
18. Gesell T, Washietl S (2008) Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinformatics 9:248. doi:10.1186/1471-2105-9-248
19. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF (2010) RNAz 2.0: improved noncoding RNA detection. Pac Symp Biocomput 15:69–79
20. Washietl S (2007) Prediction of structural noncoding RNAs with RNAz. Methods Mol Biol 395:503–526
21. Knudsen B, Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res 31: 3423–3428
22. Seemann SE, Gorodkin J, Backofen R (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Res 36:6355–6362. doi:10.1093/nar/gkn544
23. Do CB, Woods DA, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics 22:e90–e98. doi:10.1093/ bioinformatics/btl246
24. Lu ZJ, Gloor JW, Mathews DH (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA 15:1805–1813. doi:10.1261/rna.1643609
25. Kiryu H, Kin T, Asai K (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. Bioinformatics 23:434–441. doi:10.1093/bioinformatics/btl636

26. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K (2009) Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics 25:465–473. doi:10.1093/bioinformatics/btn601
27. Hamada M, Sato K, Asai K (2011) Improving the accuracy of predicting secondarystructure for aligned RNA sequences. Nucleic Acids Res 39:393–402. doi:10.1093/ nar/ gkq792
28. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797. doi:10.1093/nar/gkh340