doi:10.1016/S0022-2836(02)00308-X available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 319, 1059–1066

JMB

Secondary Structure Prediction for Aligned RNA Sequences

Ivo L. Hofacker[1], Martin Fekete[1] and Peter F. Stadler[1,2]*

[1]*Institut für Theoretische Chemie, Universität Wien Währingerstraße 17, A-1090 Wien, Austria*

[2]*The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA*

**Corresponding author*

Most functional RNA molecules have characteristic secondary structures that are highly conserved in evolution. Here we present a method for computing the consensus structure of a set aligned RNA sequences taking into account both thermodynamic stability and sequence covariation. Comparison with phylogenetic structures of rRNAs shows that a reliability of prediction of more than 80% is achieved for only five related sequences. As an application we show that the Early Noduline mRNA contains significant secondary structure that is supported by sequence covariation.

Keywords: RNA secondary structure prediction; conserved substructures; compensatory mutations

Introduction

Most functional RNA molecules exhibit a characteristic secondary structure that is highly conserved in evolution. Examples include tRNAs,[1] the rRNAs (5 S, 16 S, as well as 23 S),[2–6] RNAseP RNA,[7] the RNA component of signal recognition particles (srpRNA),[8] tmRNA,[9] and group I and group II introns†. This list can be extended by numerous families of artificially selected catalytic RNAs.

It is of considerable practical interest therefore to compute efficiently the consensus structure of a collection of such RNA molecules. Such an approach must combine the phylogenetic information contained in the sequence covariations as well as thermodynamic stability of molecules. Combinations of phylogenetic and thermodynamic methods for predicting RNA secondary structure fall into two broad groups: those starting from a multiple sequence alignment and algorithms that attempt to solve the alignment problem and the folding problem simultaneously. The main disadvantage of the latter class of methods[10–13] is their high computational cost, which makes them unsuitable for long sequences such as 16 S or 23 S RNAs. Most of the alignment based methods start from thermodynamics-based folding and use the analysis of sequence covariations or mutual information for post-processing.[14–19] The converse approach is taken in,[20] where ambiguities in the phylogenetic analysis are resolved based on thermodynamic considerations.

In this contribution we describe a combined approach that integrates the thermodynamic and phylogenetic information into a modified energy model. This has a number of advantages: (i) It is sufficient to run the folding algorithm only once for the entire alignment, which significantly reduces the computational effort, in particular for larger data sets. (ii) The reliability of prediction can be assessed fairly directly by computing the matrix of base-pairing probabilities instead of the minimum energy structure (or a small ensemble of sub-optimal folds). (iii) If the sequences do not admit a common fold, the method will not predict base-pairs.

Theory

From an algorithmic point of view, RNA secondary structure prediction can be viewed as a (complicated) variant of the maximum circular matching problem (MCMP).[21] We briefly outline the simplified model here to highlight the idea behind the alifold algorithm. The RNA folding problem, with a realistic energy model that is based on extensive thermodynamic measurements,[22] can be solved[23,24] using a similar

Abbreviations used: MCMP, maximum circular matching problem; ORF, open reading frame; srpRNA, RNA component of signal recognition particles.

E-mail address of the corresponding author: studla@tbi.univie.ac.at

† References given only to databases compiling sequence and structure information.

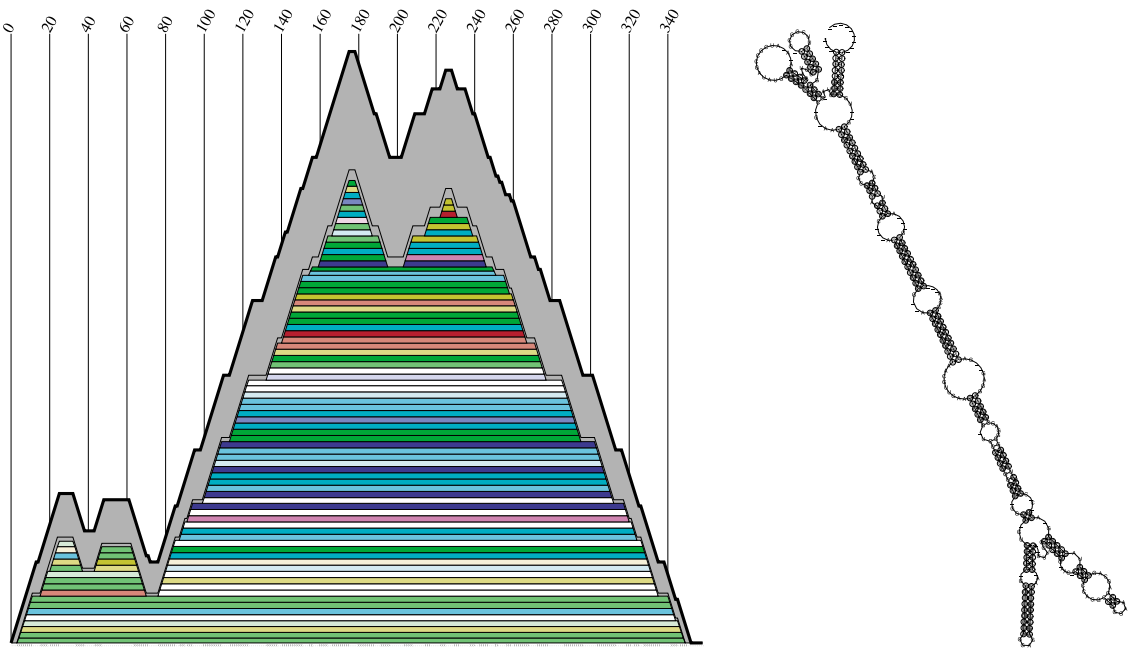

Figure 1. Consensus secondary structure of the 14 SRP RNA of Archea contained in SRPDB[8] (MET.JAN., MET.VOL., MET.FER., MET.THE., MET.ACE., HAL.HAL., ARC.FUL., PYR.ABY., PYR.HOR., THE.CEL., PYR.OCC., AER.PER., SUL.SO-A, SUL.SO-B). We use this example to explain the representation of the results: L.h.s: Mountain plot. A base pair (i,j) is represented by a slab ranging from i to j. The 5′ and 3′ sides of stems thus appear as up-hill and down-hill slopes, respectively, while plateaus indicated unpaired regions. Mountain plots[35] are equivalent to the conventional drawing (r.h.s.) but have the advantage that (1) they can be compared more easily, and (2) it is easier to display additional information about both sequence variation (color code) and thermodynamic likeliness of a base-pair (indicated by the height of the slab and the size of the dot, respectively). Colors in the order red, ocher, green, cyan, blue, violet indicate 1–6 different types of base-pairs. Pairs with one or two inconsistent mutation are shown in (two types of) pale colors. The shaded mountain in the background is the phylogenetic structure taken from the SRPDB. The close match is easily visible. It appear higher because the phylogenetic structure contains base-pairs that correspond to deletions in the majority of the structures and because the height of base-pair in the alifold structure is in general somewhat less than $p = 1.0$. R.h.s.: In the conventional secondary structure graph paired positions with consistent mutations are indicated by circles around the varying position. Compensatory mutations thus are shown by circles around both pairing partners. Inconsistent mutants are indicated by gray instead of black lettering.

dynamic programming scheme as for the MCMP (Figure 1).

We are given a sequence of nucleotides $x = (x_1, \ldots, x_n)$ of length n and energy parameters β_{ij} describing the stability of the base-pair (x_i, x_j). In the simplest case $\beta_{ij} = -1$ for every base-pair that is formed. RNA folding of course has to obey the logic of base-pairing, thus we introduce the pairing matrix Π of the sequence x with the entries $\Pi_{ij} = 1$ if sequence positions i and j can form a base-pair, i.e. if (x_i, x_j) is in the set of allowed base-pairs $B = \{\mathrm{GC}, \mathrm{CG}, \mathrm{AU}, \mathrm{UA}, \mathrm{GU}, \mathrm{UG}\}$, and $\Pi_{ij} = 0$ if x_i and x_j cannot pair. The second important restriction is that a base-pair must span at least $m = 3$ unpaired bases, i.e. if (i,j) is a pair then $j > i + m$. The RNA version of the MCMP thus consists of finding a secondary structure Ω on x that contains only allowed base-pairs ($\Pi_{ij} = 1$) and minimizes the total energy $E = \sum_{(i,j)\in\Omega} \beta_{ij}$.

The best energy on the subsequence from position i to j is denoted by E_{ij}. Because of the no-crossing rule a base-pair (i,k) separates the secondary structure into a secondary structure on the sub-sequence from $i+1$ to $k-1$ and a secondary structure from $k+1$ to j. The latter may be empty if $k = j$, of course. Therefore, E_{ij} satisfies the following recursion:

$$E_{i,j} = \min \left\{ E_{i,j-1};\ \min_{\substack{k:\ i+m<k\le j \\ \Pi_{ik}=1}} E_{i+1,k-1} + E_{k+1,j} + \beta_{ik} \right\} \tag{1}$$

The value $E_{1,n}$ is the minimal energy for a secondary structure of the sequence x. The (triangular) matrix $\mathbf{E}$ has $O(n^2)$ entries, and the computation of each entry requires a minimum over $O(n)$ terms, hence the total effort is $O(n^3)$. The structure itself, i.e. the list of base-pairs, can be recovered by standard back-tracking from the matrix $\mathbf{E}$.

While β_{ij} depends only on the type of the base-pair (x_i, x_j) in the usual ansatz there is nothing to prevent us to use a more sophisticated cost function that summarizes all our knowledge on the base-pair, not just its thermodynamic stability. Most importantly, we can use β_{ij} incorporate knowledge about sequence covariations into the folding procedure.

Assume that we are given a multiple sequence alignment $\mathbb{A}$ of N sequences. By $\mathbb{A}_i$ we denote the ith column of the alignment, while a_i^α is the entry in the αth row of the ith column. The length of $\mathbb{A}$, i.e., the number of columns, is n. Furthermore, let $f_i(\mathrm{X})$ be the frequency of base X at aligned position i and let $f_{ij}(\mathrm{XY})$ be frequency of finding X in i and Y in j.

The most common way of quantifying sequence covariation for the purpose of RNA secondary determination is the mutual information score:[25–27]

$$M_{ij} = \sum_{\mathrm{X,Y}} f_{ij}(\mathrm{XY}) \log \frac{f_{ij}(\mathrm{XY})}{f_i(\mathrm{X}) f_j(\mathrm{Y})} \tag{2}$$

Usually, the mutual information score makes no use of RNA base-pairing rules. For large datasets this is desirable, since it allows identifying non-canonical base-pairs and tertiary interaction. For the small datasets considered here, neglecting base-pairing rules does more harm (by increasing noise) than good. In particular, mutual information does not account at all for consistent non-compensatory mutations, i.e. if we have, say, only GC and GU pairs at positions i and j then $M_{ij} = 0$. Thus sites with two different types of base-pairs are treated just like a pair of conserved positions. We argue, however, that the information contained in consistent mutations such as GC → GU should not be neglected when dealing with sparse datasets that contain too little sequence variation to use phylogenetic methods alone.

As a consequence we prefer a covariance-like measure distinguished between conserved pairs, pairs with consistent mutations, and pairs with compensatory mutations. It is convenient to use the abbreviation:

$$d_{ij}^{\alpha,\beta} = 2 - \delta\left(a_i^\alpha, a_i^\beta\right) - \delta\left(a_j^\alpha, a_j^\beta\right) \tag{3}$$

where $\delta(a', a'') = 1$, if $a' = a''$ and 0, otherwise. Thus $d_{ij}^{\alpha\beta} = 0$ if the sequences α and β coincide in both aligned positions i and j, equal to 1 if they differ in one position, and is 2 if they differ in both positions. In other words, $d_{ij}^{\alpha,\beta}$ is the Hamming distance of the restriction of the sequences α and β to the two aligned positions i and j.

A straight forward measure of covariation is then

$$\begin{aligned} C_{ij} &= \frac{1}{\binom{N}{2}} \sum_{\alpha<\beta} d_{ij}^{\alpha,\beta} \Pi_{ij}^\alpha \Pi_{ij}^\beta \\ &= \sum_{\mathrm{XY,X'Y'}} f_{ij}(\mathrm{XY}) \mathrm{D}_{\mathrm{XY,X'Y'}} f_{ij}(\mathrm{X'Y'}) \end{aligned} \tag{4}$$

where the 16×16 matrix $\mathbf{D}$ has entries $\mathbf{D}_{\mathrm{XY,X'Y'}} = d_{\mathrm{H}}(\mathrm{XY}, \mathrm{X'Y'})$ if both $\mathrm{XY} \in B$ and $\mathrm{X'Y'} \in B$ and $\mathbf{D}_{\mathrm{XY,X'Y'}} = \mathbf{0}$, otherwise. Here $d_{\mathrm{H}}(\mathrm{XY}, \mathrm{X'Y'})$ is again the Hamming distance of XY and X′Y′. In passing we note that the second form of equation (4) is a scalar product, $C_{ij} = \langle f_{ij} \mathbf{D} f_{ij} \rangle$, and hence can be evaluated efficiently.

Both the mutual information score and the covariance score give a bonus to compensatory mutation. Neither score deals with inconsistent sequences, i.e. with sequences that cannot form a base-pair between positions i, j. The simplest ansatz for this purpose is:

$$q_{ij} = 1 - \frac{1}{N} \sum_\alpha \left\{ \Pi_{ij}^\alpha + \delta(a_i^\alpha, \mathrm{gap}) \delta\left(a_j^\alpha, \mathrm{gap}\right) \right\} \tag{5}$$

This simply counts the number of inconsistent sequences, where combinations of a nucleotide and a gap are counted as inconsistent while gap–gap combinations (i.e. deletions of an entire base-pair) are ignored.

In a multiple alignment of a larger number of sequences we have to expect one or the other sequencing error and of course there will be alignment errors. Thus, we cannot simply mark a pair of positions as non-pairing if a single sequence is inconsistent. Furthermore, there is the possibility of a non-standard base-pair.[27] Thus we define a threshold value B^* for the combined score $B_{ij} = C_{ij} - \phi_1 q_{ij}$ and set:

$$\Pi_{ij}^{\mathbb{A}} = \begin{cases} 0 & \text{if } B_{ij} < B^* \\ 1 & \text{if } B_{ij} \geq B^* \end{cases} \tag{6}$$

for the pairing matrix of the alignment. The energy model for the MCMP is then obtained as a linear combination of the average pairing energy and the combined covariation score B_{ij}:

$$\beta_{ij}^{\mathbb{A}} = \frac{1}{N} \sum_\alpha \epsilon\left(a_i^\alpha, a_j^\alpha\right) - \phi_2 B_{ij} \tag{7}$$

where $\epsilon(a_i^\alpha, a_j^\alpha)$ is the pairing energy contribution for a (a_i^α, a_j^α) pair in sequence α. In practice, "loop-based" energy models perform much better. The secondary structure is decomposed into its loops (faces of the planar drawing) and each loop is assigned an energy dependent on loop-type (stacked pairs, hairpin loops, interior loops, multi-branched loops), size, and sequence. Up-to-date parameters for this model are tabulated.[22] We set the total energy of an alignment-folding as the average of loop-based energies of all sequences plus the covariance contribution.

In addition to the standard energy model for RNA folding we have therefore only the threshold value B^* and the two scaling factors ϕ_1 and ϕ_2. Their default values are listed in Table 1. With these values the effect of a compensatory mutation

Table 1. Additional "energy" parameters for alignment folding

Parameter	Default
Threshold for pairing (B^*)	−1.00
Relative weight of inconsistent sequences (ϕ_1)	1.00
Weight of sequence covariation (ϕ_2) (kcal/mol)	1.00

is comparable to the energy gained by extending a helix by one base-pair. In a few tests using three 23 S rRNA sequences, these default values were indeed optimal, while variations within a factor 2 of the default did not have strong effects. In addition, non-standard base-pairs (including gap–gap combinations) can occur in the alignment folding for which no measured energy parameters are available. We substitute the default stacking energy of 0.0 kcal/mol in this case (1 cal = 4.184 J).

The values for B^* and the linear combination coefficients ϕ_1 and ϕ_2 have to be estimated with the expected values of the covariance term C_{ij} and the non-bonding term q_{ij} for uncorrelated random

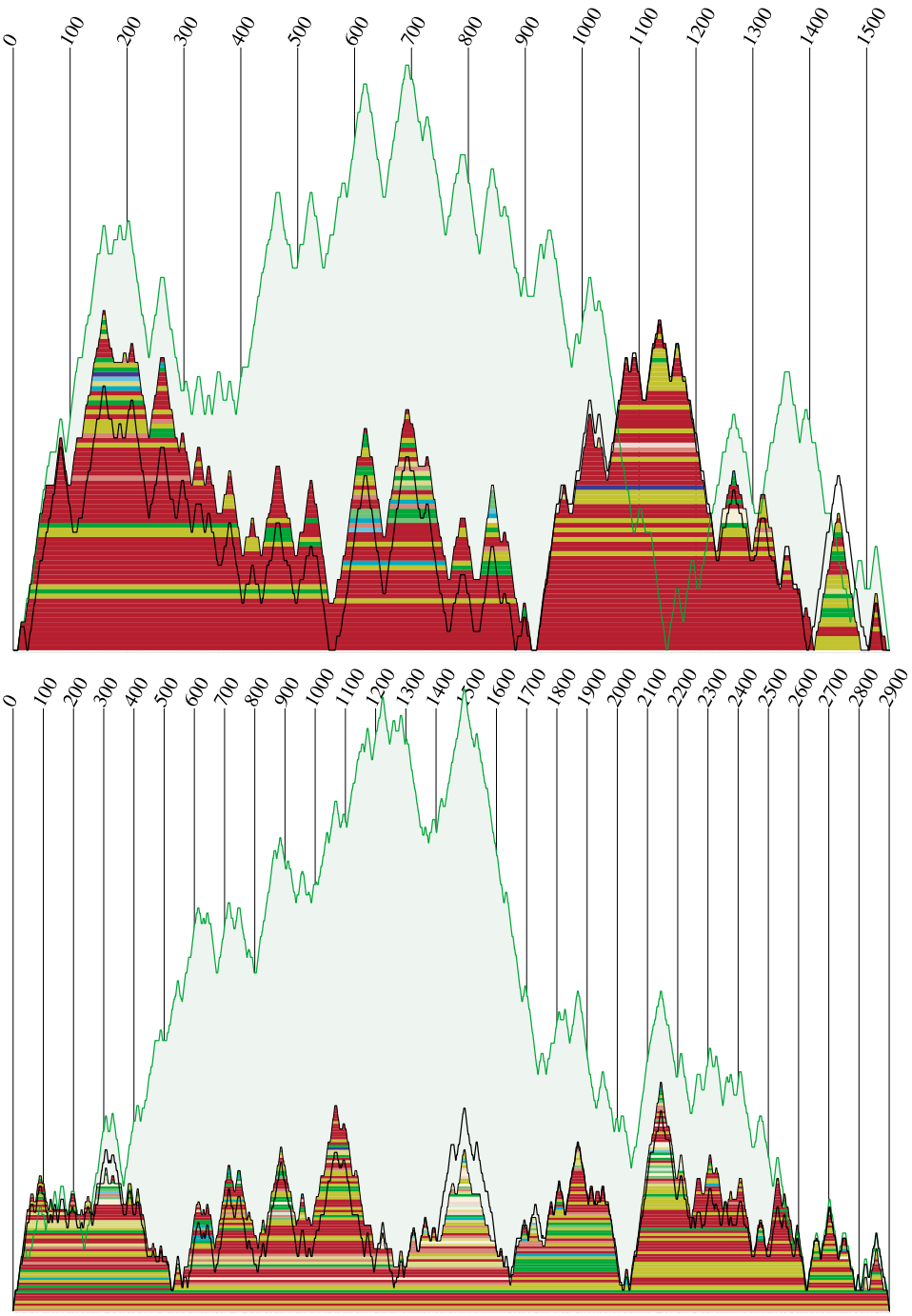

Figure 2. Mountain representation of the secondary structure of *E. coli* rRNAs. Upper panel: 16 S RNA (*A. globiformis*, *Anabaena* sp., *A. tumefaciens*, *B. japonicum*, *E. coli*), lower panel 23 S RNA (*B. subtilis*, *T. thermoph*, *Pir. marina*, *Rb. sphaero*, *E. coli*). Green line: predicted mfe structure; black line: phylogenetic structure; solid colored area: RNAalifold prediction for *E. coli* from alignment of five sequences.

Table 2. Quality of predictions

	E. coli 16 S RNA				*E. coli* 23 S RNA			
	ClustalW		RDB		ClustalW		RDB	
N	Raw	Filled	Raw	Filled	Raw	Filled	Raw	Filled
1	47.2	N/A	47.2	N/A	52.2	N/A	52.2	N/A
2	64.7	67.1	73.8	73.4	71.0	69.4	83.7	82.6
3	74.1	77.2	78.1	79.9	71.2	73.7	85.3	84.9
5	74.5	81.2	85.2	86.6	76.2	82.4	86.6	86.8
9	74.1	82.1	85.9	88.6	74.6	82.6	86.1	86.2

We list the percentage of the base-pairs of the phylogenetically reconstructed structure for *E. coli* rRNA that are correctly predicted. Data are compared for two alignments and different number *N* of aligned sequences, both for the raw RNAalifold prediction and the filled-in structure (refer to text)

sequences in mind:

$$\langle C_{ij}\rangle = \frac{6\times 0 + 8\times 1 + 22\times 2}{16^2} = \frac{13}{64} \approx 0.203$$
$$\langle q_{ij}\rangle = 1 - \frac{6}{36} \approx 0.833 \qquad (8)$$

Here the expectation of C_{ij} is computed for a sample of independent random RNA sequences.

The reliability of thermodynamics-based RNA folding is increased substantially by taking suboptimal structures into account. This can be achieved either by explicitly generating a list of suboptimal structures (as in Zuker's mfold[28] or the program RNAsubopt from the Vienna group[29]) or by directly computing the pairing probabilities for all possible base-pairs. McCaskill's partition function algorithm[24] produces the complete matrix **P** of pairing probabilities with time and memory requirements comparable to the simpler minimum energy folding. The partition function algorithm is easily extended to work with the modified energy functions in the same way as the minimum energy folding algorithm.

The covariance term (equation (4)) can be biased if the sequences are strongly clustered. A more accurate approach to quantifying the sequence covariations should therefore explicitly account for the phylogenetic relationships of the aligned sequences. A maximum likelihood approach for this task is outlined.[30] We have experimented with a parsimony-based approach in which covariations are not counted between all pairs of sequences but only along the edges of an inferred phylogenetic tree. It appears, however, that at least for data sets considered in this study the simple covariance term yields equally good results.

Results and Discussion

Purely phylogenetic methods can be used to derive conserved elements or a consensus structure only when a sufficiently large number of sequences is available, while the accuracy of purely thermodynamic structure prediction is often not satisfactory. In contrast, the alignment folding procedure introduced in this contribution predicts over 80% of the base-pairs correctly from a dataset of only five sequences with an automatically generated alignment, as the examples in Figure 2 show, see also Table 2.

The consensus structure of a set of RNA sequences has to be distinguished from the collection of structural features that are conserved. Whenever there are reasons to assume that the structure of the whole molecule is conserved one may attempt to compute a consensus structure. On the other hand, consensus structures are unsuitable when a significant part of the molecule has no conserved structures. RNA virus genomes, for instance contain only local structural patterns (such as the IRES in pircorna viruses or the TAR hairpin in HIV). Such features can be identified with a related approach that is implemented in the algorithms alidot and pfrali[17,18] and requires structure prediction for each individual sequence. The automatic search for conserved structures should not return false positives and hence has been designed not to predict secondary structures at all unless structure is unambiguously preserved among the sequences. For small sets of sequences pfrali therefore predicts only about half of the base-pairs of the phylogenctie structure and leaves out regions with little sequence variation and ambiguous thermodynamic structure predictions (data not shown).

In Figure 3 we compare the RNAalifold consensus structure with the conserved parts of the structure as predicted by pfrali for the mRNAs of the early nodulin gene enod40 from nine plant species. Enod40, which is coding for an RNA of about 700 nt, is expressed in the nodule primordium developing in the root cortex of leguminous plants after infection by symbiotic bacteria. Translation of two sORFs (I and II, 13 and 27 amino acid residues, respectively) present in the conserved 5′ and 3′ regions of enod40 was required for this biological activity.[31]

A stem-loop structure located just after the first ORF is proposed.[31] Its location, indicated by a narrow bar in Figure 3, coincides with a signal in the pfrali prediction but does not appear in the RNAalifold consensus structure. A comparison of

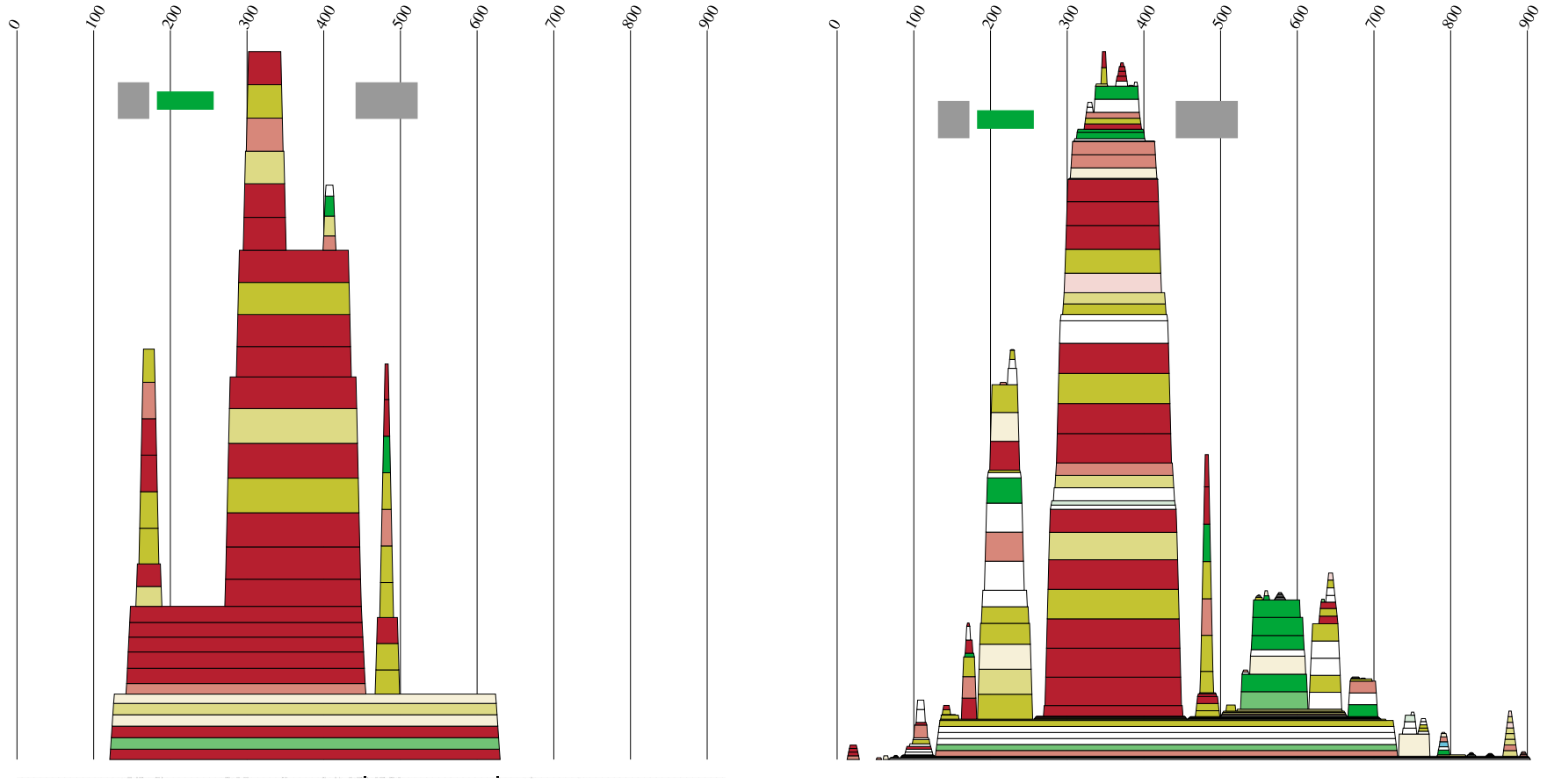

Figure 3. Mountain plots for nine enod40 sequences (PSENOD40, TRJ00268, MSENOD40, MTENOD40R, MSENOD40R, AF013594, PVENOD40, GMENOD401, SRENOD40) taken from the database.[33] The short ORF is marked by a gray bar. L.h.s.: consensus structures from RNAalifold; r.h.s: pfrali prediction. Both methods unambiguously detect a stem-loop structure (alignment positions 272–450), arid the hairpin structure (468–500) which is located within the longer ORE II. The structure (156–190) partially overlaps with ORF I; it is not well predicted by pfrali. The location of the putative RNA secondary structure described[31] (in Figure 7) is marked by the narrow bar.

this element between different enod40 transcripts (Figure 7 of Ref. 31) shows that there is a thermodynamically exceptionally stable stem-loop structure that exhibits so much structural variation that only a few base-pairs are conserved among all sequences. Hence, there is no (thermodynamically reasonable) consensus structure which explains the absence of a signal in the RNAalidot computation. The pfrali program, on the other hand, picks up the few conserved pairs and reports a structural element with many "holes".

Both methods agree on a number of other conserved secondary structure elements in enod40 RNAs that are supported by a significant number of sequence covariations. Whether some or all of these structural features are functional is unknown at present. One likely possibility is that they might take part in localization of mRNA translation.[32]

Materials and Methods

Sequence data were retrieved from publicly accessible RNA databases: the SRPDB†,[8] the non-coding RNA database‡,[33] and the Ribosomal Database Project§.[3] The *E. coli* rRNA reference secondary structures were retrieved from Robin Gutell's Comparative RNA Web Site‖. These structures are generated from covariation data alone and manually refined; the *E. coli* structure represents the standard model of rRNA structure. Non-standard base-pairs and pseudo-knots pairs were removed for comparison with predicted structures. The database names of the sequences used here are listed in the corresponding Figure captions.

Alignments were generated either automatically using ClustalW[34] or taken from the website of the Ribosomal Database Project.

The consensus structure for a set of aligned RNA sequences was computed using the program RNAalifold as described in detail in Theory. Both optimal consensus structures and base-pair probabilities were computed using the simple covariance scoring scheme (equation (7)) and the standard nearest neighbor energy model as compiled[22] and the additional parameters listed in Table 1.

For the test of prediction accuracy (Table 2), 16 S and 23 S rRNAs from *E. coli* were aligned with 1–8 sequences from other prokaryotic species (16 S rRNA: *A. tumefaciens*, *A. globiformis*, *B. japonicum*, *Anabaena* sp., *B. burgdorferi*, *B. melitensis*, *B. suis*; for 23 S rRNA: *B. subtilis*, *Pir. marina*, *Rb. sphaero*, *T. thermoph*, *Ps. cepacia*, Syn. 6301, *Tt. maritim*, *Myb. leprae*). The predicted optimal consensus structure was then compared to the phylogenetically reconstructed *E. coli* structure, by counting the percentage of the base-pairs of reference structure present in the predicted structure. Since the *E. coli* structure may contain additional non-conserved base-pairs, we also compared the "filled-in" structure obtained by computing the thermodynamically most favorable structure consistent with the consensus prediction (using RNAfold –C).

† http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html
‡ http://biobases.ibch.poznan.pl/ncRNA/
§ http://rdp.cme.msu.edu/html/
‖ http://www.rna.icmb.utexas.edu/

Availability

Source code implementing the method described here will be distributed with the next release of the Vienna RNA Package, a corresponding fold server can be found at http://rna.tbi.univie.ac.at/

Acknowledgments

This work was supported in part by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung, Project No. P-13545-MAT.

References

1. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.* **26**, 148–153.
2. Gutell, R. R., Cannone, J. J., Shang, Z., Du, Y. & Serra, M. J. (2000). A story: unpaired adenosine bases in ribosomal RNA. *J. Mol. Biol.* **304**, 335–354.
3. Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Jr, Saxman, P. R., Farris, R. J. *et al.* (2001). The RDP-II (ribosomal database project). *Nucl. Acids Res.* **29**, 173–174.
4. Van de Peer, Y., De Rijk, J., Wuyts, J., Winkelmans, T. & De Wachter, R. (2000). The European small subunit ribosomal RNA database. *Nucl. Acids Res.* **28**, 175–176.
5. Szymanski, M., Barciszewska, M. Z., Barciszewski, J. & Erdmann, V. A. (2000). 5 S ribosomal RNA database Y2K. *Nucl. Acids Res.* **28**, 166–167.
6. Wuyts, J. P., De Rijk, P., Van de Peer, Y., Winkelmans, T. & De Wachter, R. (2001). The European large subunit ribosomal RNA database. *Nucl. Acids Res.* **29**, 175–177.
7. Brown, J. W. (1999). The ribonuclease P database. *Nucl. Acids Res.* **27**, 314.
8. Gorodkin, J., Knudsen, B., Zwieb, C. & Samuelsson, T. (2001). SRPDB (signal recognition particle database). *Nucl. Acids Res.* **29**, 169–170.
9. Williams, K. P. (2002). The tmRNA website: invasion by an intron. *Nucl. Acids Res.* **30**, 179–182.
10. Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.* **45**, 810–825.
11. Tabaska, J. E. & Stormo, G. D. (1997). Automated alignment of RNA sequences to pseudo-knotted structures. In *Proceedings of the ISMB-97* (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. & Valencia, A., eds), pp. 311–318, AAAI Press, Menlo Park, CA.
12. Gorodkin, J., Heyer, L. J. & Stormo, G. D. (1997). Finding common sequences and structure motifs in a set of RNA molecules. In *Proceedings of the ISMB-97* (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. & Valencia, A., eds), pp. 120–123, AAAI Press, Menlo Park, CA.
13. Gorodkin, J., Heyer, L. J. & Stormo, G. D. (1997). Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucl. Acids Res.* **25**, 3724–3732.
14. Le, S.-Y. & Zuker, M. (1991). Predicting common foldings of homologous RNAs. *J. Biomol. Struct. Dyn.* **8**, 1027–1044.

15. Lück, R., Steger, G. & Riesner, D. (1996). Thermodynamic prediction of conserved secondary structure: application to the RRE element of HIV, the tRNA-like element of CMV, and the mRNA of prion protein. *J. Mol. Biol.* **258**, 813–826.
16. Lück, R., Gräf, S. & Steger, G. (1999). ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucl. Acids Res.* **27**, 4208–4217.
17. Hofacker, I. L., Fekete, M., Flamm, C., Huynen, M. A., Rauscher, S., Stolorz, P. E. & Stadler, P. F. (1998). Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.* **26**, 3825–3836.
18. Hofacker, I. L. & Stadler, P. F. (1999). Automatic detection of conserved base-pairing patterns in RNA virus genomes. *Comp. Chem.* **23**, 401–414.
19. Juan, V. & Wilson, C. (1999). RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.* **289**, 935–947.
20. Han, K. & Kim, H.-J. (1993). Prediction of common folding structures of homologous RNAs. *Nucl. Acids Res.* **21**, 1251–1257.
21. Nussinov, R., Piecznik, G., Griggs, J. R. & Kleitman, D. J. (1978). Algorithms for loop matching. *SIAM J. Appl. Math.* **35**, 68–82.
22. Mathews, D., Sabina, J., Zucker, M. & Turner, H. (1999). Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940.
23. Zuker, M. & Stiegler, P. (1981). Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* **9**, 133–148.
24. McCaskill, J. S. (1990). The equilibrium partition function and base-pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
25. Chiu, D. K. & Kolodziejczak, T. (1991). Inferring consensus structure from nucleic acid sequences. *CABIOS*, **7**, 347–352.
26. Gutell, R. R. & Woese, C. R. (1990). Higher order structural elements in ribosomal RNAs: pseudoknots and the use of noncanonical pairs. *Proc. Natl Acad. Sci. USA*, **87**, 663–667.
27. Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J. & Stormo, G. D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.* **20**, 5785–5795.
28. Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
29. Wuchty, S., Fontana, W., Hofacker, I. L. & Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
30. Gulko, B. & Haussler, D. (1996). Using multiple alignments and phylogenetic trees to detect RNA secondary structure. In *Proceedings of the Pacific Symposium on Biocomputing* (Hunter, L. & Klein, T., eds), pp. 350–367, World Scientific, Singapore.
31. Sousa, C., Johansson, C., Charon, C., Manyani, H., Sautter, C., Kondorosi, A. & Crespi, M. (2001). Translational and structural requirements of the early nodulin gene enod40, a short-open reading frame-containing RNA, for elicitation of a cell-specific growth response in the alfalfa root cortex. *Mol. Cell. Biol.* **21**, 354–366.
32. Oleynikov, V. & Singer, R. H. (1998). RNA localization: different zipcodes, same postman? *Trends Cell Biol.* **8**, 381–383.
33. Erdmann, V. A., Szymanski, M., Hochberg, A., de Groot, N. & Barciszewski, J. (2000). Non-coding, mRNA-like RNAs database Y2K. *Nucl. Acids Res.* **28**, 197–200.
34. Thompson, J. D., Higgs, D. G. & Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
35. Hogeweg, P. & Hesper, B. (1984). Energy directed folding of RNA sequences. *Nucl. Acids Res.* **12**, 67–74.

Edited by J. Karn

(Received 2 November 2001; received in revised form 3 April 2002; accepted 4 April 2002)