Prediction of Locally Stable RNA Secondary Structures for Genome-Wide Surveys

I.L. Hofacker, B. Priwitzer and P.F. Stadler

Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstraße 17, Vienna, A-1090, Austria, Bioinformatik, Institut für Informatik, Universität Leipzig, Leipzig, D-04103, Germany and The Santa Fe Institute, Santa Fe, New Mexico, USA

ABSTRACT

Motivation: Recently novel classes of functional RNAs, most prominently the miRNAs have been discovered, strongly suggesting that further types of functional RNAs are still hidden in the recently completed genomic DNA sequences. Only few techniques are known, however, to survey genomes for such RNA genes. When sufficiently similar sequences are not available for comparative approaches the only known remedy is to search directly for structural features.

Results: We present here efficient algorithms for computing locally stable RNA structures at genome-wide scales. Both the minimum energy structure and the complete matrix of base pairing probabilities can be computed in $\mathcal{O}(N \times L^2)$ time and $\mathcal{O}(N+L^2)$ memory in terms of the length N of the genome and the size L of the largest secondary structure motifs of interest. In practice, the 100Mb of the complete genome of *Caenorhabditis elegans* can be folded within about half a day on a modern PC with a search depth of $L = 100$. This is sufficient e.g. for a survey for miRNAs.

Availability: The software described in this contribution will be available for download at http://www.tbi.univie.ac.at/~ivo/RNA/ as part of the `Vienna RNA Package`.

Contact: Ivo L. Hofacker,
Tel: ++43 1 4277 52738, Fax: ++43 1 4277 52793,
ivo@tbi.univie.ac.at

INTRODUCTION

Structural genomics, the systematic determination of all macro-molecular structures represented in a genome, is at present focused almost exclusively on proteins. Although it is common place to speak of *"genes and their encoded protein products"*, thousands of human genes (THE GENOME SEQUENCING CONSORTIUM, 2001) produce transcripts that exert their function without ever producing proteins. The list of functional non-coding RNAs (ncRNAs) includes well-known key players in the biochemistry of the cell, such as tRNAs, rRNAs, tmRNA, and the RNA components of RNAseP and signal recognition particles, as well as recently discovered functional RNAs such as the miRNAs (Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee & Ambros, 2001) that regulate gene expression by regulating mRNA expression. Many of these RNAs have characteristic secondary structures that are highly conserved in evolution.

Another level of RNA function is presented by functional motifs within protein-coding RNAs. A few of the best-understood examples of structurally conserved RNA motifs are found in viral RNAs such as the TAR and RRE structures in HIV and the IRES regions in Picornaviridae and many Flaviviridae. A textbook example of a functional RNA secondary structure is the *Rho*-independent termination in *E.coli*. The newly synthesized mRNA forms a hairpin in the 3'UTR that interacts with the RNA polymerase causing a change in conformation and the subsequent dissociation of the Enzyme-DNA-RNA complex.

It is not hard to argue therefore that **RNomics**, i.e., the understanding of functional RNAs (both ncRNA genes and functional motifs in protein-coding RNAs) and their interactions at a genomic level, is of utmost practical and theoretical importance in modern life sciences: The comprehensive understanding of the biology of a cell obviously requires the knowledge of identity of *all* encoded RNAs, the molecules with which they interact, and the molecular structures of these complexes (Doudna, 2000).

This ambitious goal requires first of all the development of versatile and reliable computational methods that can detect and classify functional RNAs, preferably within a single genome. An necessary prerequisite is the computation of locally stable secondary structures. This can be achieved by folding sub-sequences of length L in a window sliding along the genomic sequence nucleotide by nucleotide. In practice, however, the sequence windows have to be shifted by a substantial fraction of L in order to keep the CPU requirements manageable. As a consequence, a large number of relevant local structures are ignored. In this contribution we report a computationally efficient method for surveying *all* thermodynamically favorable local RNA structures at a genome-wide scale.

A MODIFIED FOLDING ALGORITHM

The RNA folding problem is complicated considerably by the details of the modern energy model (Mathews *et al.*, 1999) which is based upon interactions of adjacent base pairs and loop contributions. Dynamic programming solutions were described by Zuker & Stiegler (1981); McCaskill (1990). Efficient implementations are available e.g. in the `Vienna RNA Package` (Hofacker *et al.*, 1994) and `mfold` (Zuker, 1989).

Maximum Circular Matching

While in practice all computations are performed using the full loop based energy model, the logic of the folding problem and its solution is much easier to explain in terms of a simplified model, the so-called Maximum Circular Matching Problem (MCMP), that considers only base pairing strength. We therefore use this simplified version to explain the modifications to the folding algorithm that are necessary to find locally optimal structures.

Given a sequence x we define the matrix $\mathbf{\Pi}$ with entries $\Pi_{ij} = 1$ when x_i and x_j can form a base pair and $\Pi_{ij} = 0$ otherwise. In the MCMP we arrange the sequence x along a circle and ask for the maximum matching $\mathfrak{M}$ such that (1) $\{i, j\} \in \mathfrak{M}$ implies $\Pi_{ij} = 1$ and such that (2) two chords do not cross. This *no-crossing condition* is equivalent to the "no-pseudoknots" condition in nucleic acid folding. In fact, MCMP can be interpreted as the problem of finding the secondary structure that maximizes the number of base pairs. In order to stay closer to the folding algorithm we define E_{ij}, the energy of the most stable structure on the subsequence from i to j (inclusive) as the negative of the maximal number of base-pairs that can be formed on this subsequence.

The MCMP is then solved by the dynamic programming recursion (Nussinov *et al.*, 1978)

$$E_{ij} = \min\left\{ E_{i,j-1},\ \min_{\substack{k=i\ldots j-m \\ \Pi_{kj}=1}} E_{i,k-1} + E_{k+1,j-1} + \varepsilon(k,j) \right\} \tag{1}$$

with the initial conditions $E_{i,i+d} = 0$ for $0 \leq d \leq m$, where m denotes the minimum unpaired segment in a hairpin loop, usually $m = 3$. Here $\varepsilon(i, j)$ is the energy contribution for forming the base pair (i, j), in the simplest case $\varepsilon(i, j) = -1$ iff $\Pi_{ij} = 1$. The secondary structure graph can be retrieved by straightforward backtracking from the (E_{ij}) array.

Forward Recursion

Restricting the maximum span of a base-pair to $L < n$ poses no problem. For the optimal energy subject to this restriction, we have

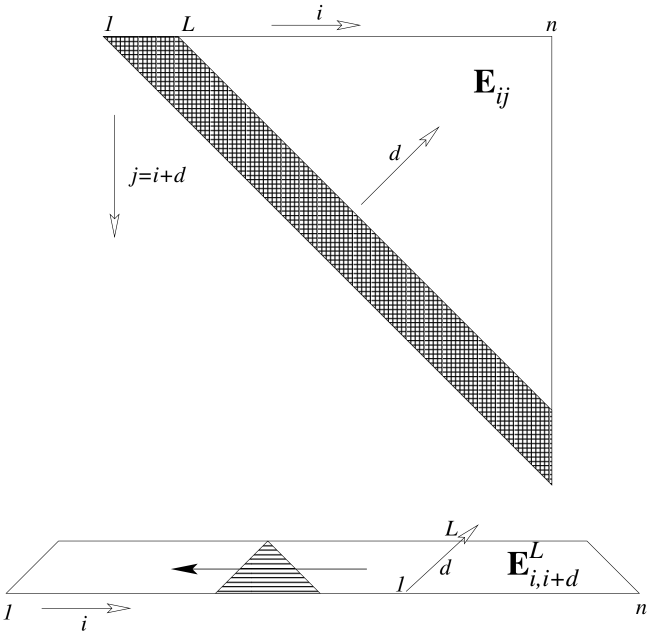

Fig. 1. The array E^L_{ij} is needed only close to the diagonal. [...]

$$E^L_{ij} = \min\left\{ E_{i,j-1},\ \min_{\substack{k=j-L\ldots j-m \\ \Pi_{kj}=1}} E^L_{i,k-1} + E^L_{k+1,j-1} + \varepsilon(k,j) \right\} \tag{2}$$

Define $f_k = E^L_{k,n}$ to be the minimal energy on the tail of the sequence starting at position k. It is clear that

$$f_k = \min\left\{ f_{k+1},\ \min_{\substack{d=m+1\ldots L \\ \Pi_{k,k+d}=1}} E^L_{k+1,k+d-1} + \varepsilon(k,k+d) + f_{k+d+1} \right\} \tag{3}$$

since the structures beginning with base pairs at position k can be decomposed into the optimal structure with a base pair from k to $k + d$ and the tail beyond this pair. The span d of the pair is of course constrained by $d \leq L$. The optimal folding energy is $E_{\text{opt}} = f_1$. It follows from equ.(3) that f_1 can be computed if all f_k and $E^L_{k+1,k+d-1}$ are known. Since the computation of f_k only requires f_l with $l > k$ and the part of the (E_{ij})-array in the triangle between $E_{k,k+L}$ and the diagonal we need to store only $\mathcal{O}(L^2)$ entries of the (E_{ij})-array and the (f_k)-array, see Fig. 1.

We shall see below that the backtracking step can be partitioned such that no further information has to be stored. The forward step of the algorithm therefore requires $\mathcal{O}(nL^2)$ operations and $\mathcal{O}(n + L^2)$ memory, see Fig. 2.

Backtracking

The array (f_k) contains the energies of locally optimal components that *begin* at position k. Since there are no energy contributions in the "out-side loop", i.e., of the joints connecting structural components, we know that a locally optimal component begins at position k if and only if $f_k < f_{k+1}$. The pairing partner can now be obtained by backtracking within the E^L array. This backtracking step works on the sub-sequence $x[k \dots k+L]$ as in the standard MCMP (Nussinov *et al.*, 1978). As a result we obtain a list of locally optimal components $\mathcal{C}(k)$ together with their position in the full sequence and the energy of the optimal "tail structure" on $x[k \dots N]$. Frequently, a component $C(l)$ consists simply of a smaller (previously detected) locally optimal component $C(k)$ enclosed by one additional base pair. The size of the output can be removed considerably if we store only the locally optimal components that are also maximal w.r.t. inclusion.

If desired, the optimal structure of the complete sequence can be reconstructed from this list of components $\mathcal{C}(k)$ starting now at the 5' end.

Performance

The `lfold` algorithm has been implemented in `C` as variant of the `fold` routine of the `Vienna RNA package`. To assess the performance we applied the algorithm to several viral and bacterial genome, as well as the complete genome of *Caenorhabditis elegans*. Figure 1 shows the `lfold` performance as a function of sequence length n and maximum pair span L. Typical bacterial genomes can be handled with moderate computer requirements even when using a span of $L = 1000$. Extrapolating from the data shown even the human genome $n \approx 3 \cdot 10^9$ should be doable with $L = 100$ and a week's computer time.

As another test case we have predicted secondary structures with $L = 100$ for all six chromosomes *Caenorhabditis elegans*, total size about 100Mb. The span size was chosen so that it should be possible to search the predicted structures for small temporal RNAs (stRNA), the precursors of miRNAs. *C. elegans* chromosomes consist of 14 to 21 million bases and folding took between 1.5 and 2.5 hours. The resulting list of locally optimal components contained between 700000 and 1 million structures per chromosome. The results of the rather tedious analysis of these data will be reported elsewhere.

BASE PAIRING PROBABILITIES

At physiological temperatures an RNA molecule may exhibit an ensemble of structures with similar, near optimal, energy. Therefore, as well as because of the unavoidable inaccuracies of predicted structures, it is often insufficient to describe an RNA molecule by a single optimal secondary structure. An elegant way to

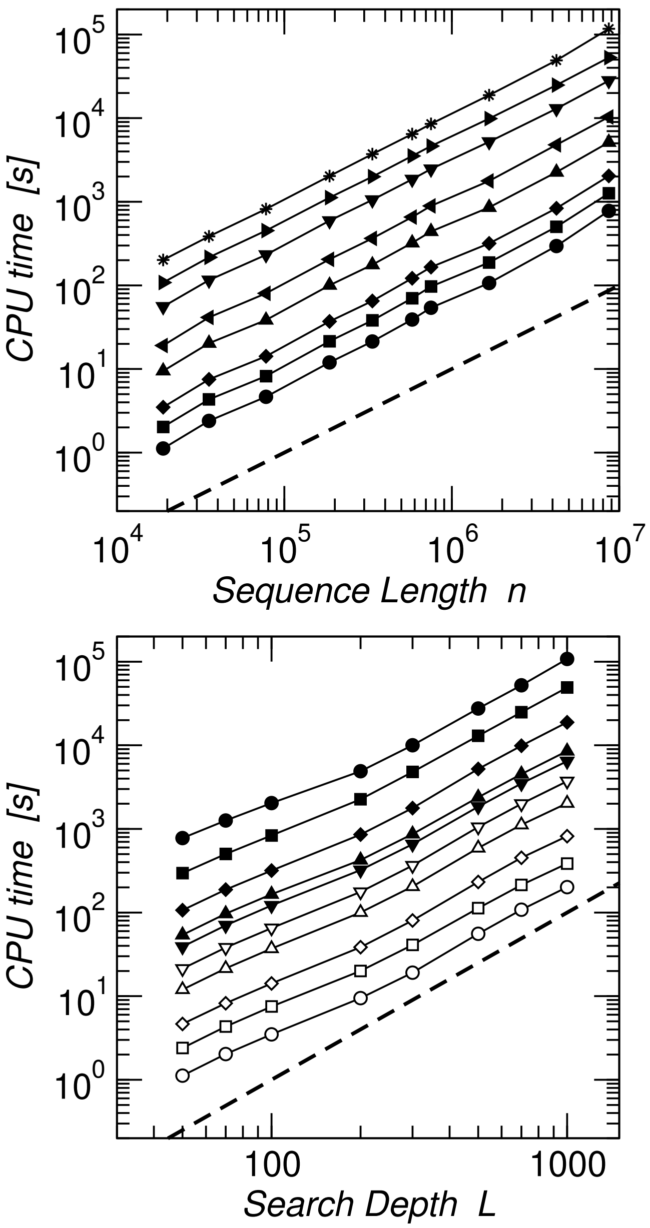

Fig. 2. Performance of `RNALfold`. Upper panel: search depths L are 50, 70, 100, 200, 300, 500, 1000, and 2000, from bottom to top. The dashed line is $t = 10^{-5}n$ sec. Data are with the `-noLP` option, i.e., excluding isolated base pairs, timings are for a LINUX PC with 2.2GHz pentium4 processor.
Lower panel: Folding times for test genomes: Ebola virus $n = 18890$, ○; Sulfolobus virus $n = 35450$, □; Halovirus HF2 $n = 77670$, ◇; Variola major $n = 186103$, △; Ectocarpus siliculosus virus $n = 335593$, ▽; Mycoplasma genitalium G37 $n = 580074$, ▼; Ureaplasma urealyticum $n = 751739$, ▲; Aeropyrum pernix K1 $n = 1669695$, ◆; Bacillus subtilis $n = 4214814$, ■; Streptomyces coelicolor $n = 8667507$, ●. The dashed line is $t = 10^{-4}L^2$sec.

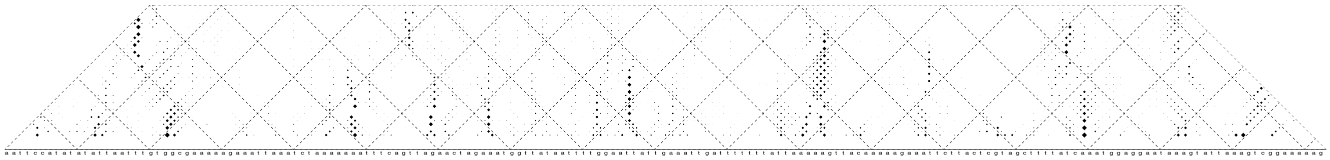

Fig. 3. Example of a base pairing probability matrix obtained from solving the partition function version of the MCMP with $\varepsilon = -3$ and $L = 40$ for a sequence from the aster yellows phytoplasma (AY191296). The dot plot depicts each possible base pair by a square with an area proportional to the pairing probability. Helical regions therefore appear as vertical "lines", structural alternatives as fuzzy clouds of points.

describe the ensemble of plausible structures is given by McCaskill's partition function algorithm, which allows to compute the probabilities of all possible base pairs in thermodynamic equilibrium. Again restricting the span of base pairs yields an $\mathcal{O}(n \times L^2)$ algorithm, as shown below.

Let Z_{ij} be the partition function of the substructures from i to j, and denote by Z^B_{ij} the partition function of the substructures from i to j that have a base pair from i to j. We have

$$\begin{aligned} Z_{ij} &= Z_{i+1,j} + \sum_{k=i+m}^{j} Z^B_{ik} Z_{k+1,j} \\ &= Z_{i+1,j} + \sum_{k=i+m}^{j} Z_{i+1,k-1} Z_{k+1,j} e^{-\varepsilon(i,k)} \end{aligned} \tag{4}$$

In order to incorporate the restriction of range d of a base pair we first note that $Z^B_{i,i+d} = 0$ for $d > L$. This yields

$$Z_{ij} = Z_{i+1,j} + \sum_{d=m+1}^{L} Z_{i+1,i+d-1} Z_{i+d+1,j} e^{-\varepsilon(i,i+d)} \tag{5}$$

which can be evaluated in $\mathcal{O}(L)$ time. Furthermore we introduce the abbreviation $Z^*_k = Z_{kn}$ and observe

$$Z^*_k = Z^*_{k+1} + \sum_{d=m+1}^{L} Z^B_{k,k+d} Z^*_{d+1} \,. \tag{6}$$

Again, only on the most recent triangular part of the matrix Z^B needs to be stored. The partition function $Z = Z_{1n} = Z^*_1$ can therefore be evaluated in $\mathcal{O}(nL^2)$ time and $\mathcal{O}(L^2)$ memory.

The probability P_{kl} that the bases k and l are pairing in thermodynamic equilibrium can be computed from the partition function $\widetilde{Z}_{kl}$ of structures *outside* the sequence interval $[k, l]$ and Z^B_{kl} as the ratio

$$P_{kl} = \widetilde{Z}_{kl} Z^B_{kl} / Z \tag{7}$$

The exterior partition functions $\widetilde{Z}_{kl}$ satisfy the recursion

$$\begin{aligned} \widetilde{Z}_{kl} = {} & Z_{1,k-1} Z_{l+1,n} + \\ & \sum_{i<k;l<j} \widetilde{Z}_{ij} Z_{i+1,k-1} Z_{l+1,j-1} e^{-\varepsilon(i,j)} \end{aligned} \tag{8}$$

Note that the sum in equ.(8) vanishes if $l \geq k + L$ which defines the initial values $\widetilde{Z}_{k,k+L} = Z_{1,k-1} Z_{k+L+1,n}$. Recall that we do not consider base pairs spanning more than L bases, i.e., $e^{-\varepsilon(i,j)} = 0$ for $j > i + L$. We can reduce the computational complexity by introducing the auxiliary $L \times L$ field

$$Z^M_{il} = \sum_{j=l+1}^{\min\{i+L,n\}} \widetilde{Z}_{ij} Z_{l+1,j-1} e^{-\varepsilon(i,j)} \tag{9}$$

Equ.(8) can now be rewritten in the form

$$\widetilde{Z}_{kl} = \bar{Z}_{k-1} Z^*_{l+1} + \sum_{i=l-L}^{k-1} Z^M_{il} Z_{i+1,k-1} \tag{10}$$

where $\bar{Z}_j$ is the partition function of the initial subsequence, which satisfies the recursion

$$\bar{Z}_j = \bar{Z}_{j-1} + \sum_{d=m+1}^{L} Z^B_{j,j-d} \bar{Z}_{j-d-1} \,. \tag{11}$$

Both Z^M_{il} and $\widetilde{Z}_{kl}$ can be obtained in $\mathcal{O}(L)$ time because the sums span at most L index values. Furthermore, we only need matrix entries Z^M_{kl} and $\widetilde{Z}_{kl}$ with $l - k \leq L$, i.e., $\mathcal{O}(n \times L)$ matrix entries. The algorithm therefore requires $\mathcal{O}(n \times L^2)$ steps and $\mathcal{O}(L^2 + n)$ storage, where the $\mathcal{O}(n)$ contribution is used to store the input and the arrays $\bar{Z}$ and Z^*, respectively.

Tools for the analysis of very large base pairing probability matrices are not *yet* available. We have therefore refrained from implementing the complete energy model at this time and use the partition function version of the MCMP to demonstrate the feasibility of the approach.

Fig. 3 gives a small example. It is clear that this type of data is not amenable to manual analysis; the design of corresponding data-mining tools hence is ongoing research.

DISCUSSION

We have presented here an efficient algorithm for surveying local RNA secondary structures at genome-wide scales. At least for the minimum free energy problem, we also describe a versatile implementation that makes use of the full RNA energy model.

The use of structural information appears to be necessary. Various groups have tried to detect functional RNA structures based on local thermodynamical stability alone, see e.g. (Le *et al.*, 1988; Huynen *et al.*, 1996). While such procedures are capable of detecting some particularly stable features, a recent study of Rivas & Eddy (2000) concludes that "although a distinct, stable secondary structure is undoubtedly important in most noncoding RNAs, the stability of most noncoding RNA secondary structures is not sufficiently different from the predicted stability of a random sequence to be useful as a general genefinding approach." Thus, the explicit usage of structural information is indispensable.

The list of locally optimal components produced by `lfold` is therefore a necessary first step for approaches to search for both known and novel functional RNA structures. This is most obvious when searching for a class of functional RNAs for which information on conserved structural features is already known, such as stRNAs. In these cases one can obtain a structural model from known instances of the RNA in question, and simply search the list for the reference structure, possibly using a local structural alignment algorithm (Höchsmann *et al.*, 2002).

In principle one can also hope to identify novel functional RNAs based on predicted structures. To this end the frequencies of structural motifs are correlated with their genome context. Such an approach could detect both potential regulatory features in mRNAs and new functional RNAs depending on whether ones searches near or far away from protein-coding genes. The computational methods for such comparisons, however, go beyond the scope of this contribution.

Acknowledgments

This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893.

REFERENCES

Doudna, J. A. (2000). Structural genomics of RNA. *Nature Struct. Biol.*, **7**, 954–956.

Höchsmann, M., Thomas, T., Giegerich, R. & Kurtz, S. (2002). A new algorithm for local similarity of RNA secondary structures. Technical report, University of Bielefeld, Germany. Preprint.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, **125**, 167–188.

Huynen, M. A., Perelson, A. S., Viera, W. A. & Stadler, P. F. (1996). Base pairing probabilities in a complete HIV-1 RNA. *J. Comp. Biol.*, **3**, 253–274.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–857.

Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *caenorhabditis elegans*. *Science*, **294**, 858–862.

Le, S.-Y., Chen, J.-H., Currey, K. & Maizel, J. (1988). A program for predicting significant RNA secondary structures. *CABIOS*, **4**, 153–159.

Lee, R. C. & Ambros, V. (2001). An extensive class of small RNAs in *caenorhabditis elegans*. *Science*, **294**, 862–864.

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.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Nussinov, R., Piecznik, G., Griggs, J. R. & Kleitman, D. J. (1978). Algorithms for loop matching. *SIAM J. Appl. Math.*, **35**, 68–82.

Rivas, E. & Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.

THE GENOME SEQUENCING CONSORTIUM (2001). Gene content of the human genome. *Nature*, **409**, 860–921.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.

Zuker, M. & Stiegler, P. (1981). Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, **9**, 133–148.