Recent Developments in RNA Secondary Structure Prediction

Adelene Sim

December 5, 2008

Abstract

Here we discuss recent developments in predicting RNA secondary structure, giving an overview of methods used, including energy minimization and Boltzmann sampling, a probabilistic approach, and the use of fragment libraries generated from high resolution three-dimensional RNA crystal structures. An example for each of these three methods is discussed in detail (namely Sfold, CONTRAfold and MC-FOLD respectively), and compared using the P4-P6 domain of *Tetrahymena ribozyme* as the test case. In all instances, the predicted secondary structures were very similar to the native one, with the main helices being captured quite accurately. This mini-review is by no means exhaustive, but gives a guide to the current research in pushing the field of understanding and predicting RNA structure.

Introduction

Ribonucleic acids (RNAs), are typically regarded as the intermediate molecule between deoxyribonucleic acids (DNAs) and proteins. DNA is the genetic information carrier, while proteins carry out a myriad of biological functions - RNA was simply a messenger molecule to convert the coding genetic sequence from DNA to the corresponding protein.

In recent years, it has been found that RNA has functions beyond simply the messenger RNA (mRNA). In fact, RNA has important gene regulatory roles (1). RNA also adopts intricate three-dimensional structures, much like proteins, and the axiom that structure affects function likely is relevant also in the case of RNA. In fact, this is explicitly seen in riboswitches (2), which is a class of RNA that binds to small metabolites, and adopts different conformation in its ligand bound and unbound states. This conformational change then results in regulation of gene expression further upstream in the translated regions of the mRNA.

As a result, understanding the tertiary structure of RNA will provide invaluable insights to how it functions. Unfortunately, *ab inito* RNA folding is both highly intractable for large RNAs, and also heavily influenced by the force-fields used. However, the RNA folding problem can be split up into smaller, digestible portions. This is based on the assumption that RNA folds in a hierachical manner - the secondary structure of RNA is stable under physiological conditions, and forms prior to any tertiary interactions. This assumption, while a simplification, appears not to be entirely invalid. In particular, the Watson-Crick base pairing between guanine (G) and cytosine (C) , and adenine (A) and uracil (U), are more strongly dependent on the hydrophobic effect of the bases, and the hydrogen bonding between base pairs, while the tertiary interaction is more strongly dependent on electrostatic effects, due to the highly negatively charged phosphate backbone of RNAs. The different dominant influences to the secondary and tertiary structures of RNA allows us to solve the RNA folding problem sequentially.

Predicting the secondary structure of RNA accurately drastically reduces the phase space needed to be explored for tertiary structure prediction. Having accurate secondary structure models also guides experimentalists in interpreting their experiments, and facilitates them in making further predictions for testing in their respective RNA systems.

A very popular secondary structure prediction tool for nucleic acids is Mfold, which adopts a Physics-based approach (3). This has also been improved upon by Boltzmann sampling (4, 5) and sub-optimal searches (6–8). More recently, CONTRAfold uses a flexible probabilistic approach via conditional log-linear model (CLLM) (9). CLLM is a generalized version of stochastic context-free grammars (SCFCs) which have been rather unsuccessfully been atttempted in RNA secondary strucure prediction (10–12). In addition, with the growing database of high resolution three-dimensional x-ray crystallographic structures, fragment libraries generated from this database are gathering stronger attention, both for secondary and tertiary structure predictions. For instance, the secondary structure prediction tool MC-FOLD uses the fragment libary approach (13), as does Fragment Assembly of RNA (FARNA) (14). Some of these methods will be discussed in greater detail below.

Here I will focus this review on the secondary structure prediction of non pseudo-knotted structures, and will test and compare these programs using the P4-P6 domain of the *Tetrahymena ribozyme* as the standard. While this single test is not statistically significant, it still provides insight to the ease of use of these programs and their accuracies.

Methods

Energy minimization

Mfold is a classic program that has commonly been used by biologists in finding both the lowest energy secondary structure of nucleic acids, as well as the thermodynamic stability of them. The energy calculations are based on what is known as the Turner rules (7, 15, 16), which was built up from a rigorous set of thermal melting experimental data of oligonucleotides. This covers the sequence dependent stability of oligonucleotides under different ionic strength conditions. Because of its experimental basis, methods which build on these Turner rules are often regarded as more "physical" than other methods, even though the extrapolation of knowledge gathered from studying small DNA and RNA fragments to larger RNA molecules might not be valid. Nonetheless, Mfold remains one of the "gold standard" programs that other up and coming algorithms often compare themselves to.

Mfold uses a dynamic programming algorithm to obtain a matrix of scores for possible secondary structures of RNA (6). A traceback algorithm is then able to find the optimal and suboptimal secondary structures. However, under physiological conditions, the Boltzmann energy distribution has a finite width, which means that conformations with energies close to the lowest energy state are still likely to be physically feasible structures. Hence sub-optimal results within a certain energy width are also often considered in the traceback algorithm. Unfortunately, this then heavily depends on how well the energy rules have described the true secondary structure energy landscape of the RNA.

Even considering suboptimal alignments, these optimal and suboptimal predicted models are not statistically weighted, and can therefore give thermodynamically skewed results. For instance, depending on the energy landscape of the RNA, the optimal (lowest) energy configuration could be less likely accessed, and therefore not representative of the ensemble averaged secondary structure. To account for this statistical weighting, Ding *et al.* (4, 17) introduced Sfold, which generates a Boltzmann weighted ensemble of structures. These structures are then clustered by base-pair distance (5). In some instances, the minimum energy structure is not found in the main cluster of the ensemble (see figure 1).

The Boltzmann equilibrium probability of a secondary structure I for a sequence S is given as (4)

$$P\left(I\right) = \exp\left[-E\left(S, I\right) / \left(RT\right)\right] / U$$

with $E(S, I)$ being the free energy of the structure for the sequence, and R is the gas constant, T the absolute temperature and U the partition function for all possible secondary structure of the RNA sequence. Therefore, $U = \sum_I \exp\left[-E\left(S, I\right) / \left(RT\right)\right]$. This probability $P(I)$ is determined recursively, by working inwards from the ends of the sequence. The energy $E(S, I)$ is evaluated each time based on the Turner energy rules for the particular subset of

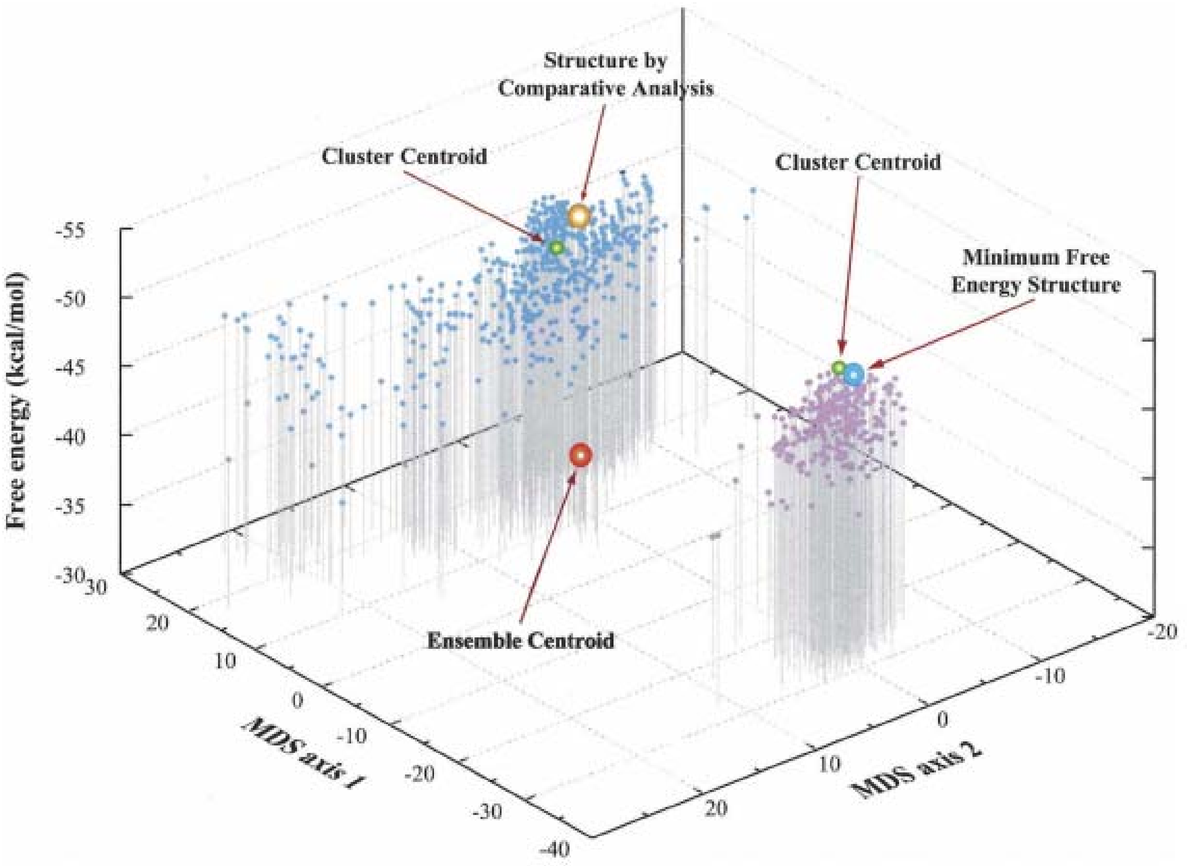

Figure 1: *The energy landscape of the sampled ensemble and representative structures for* Agrobacterium tumefaciens *5S rRNA (GenBank accession number X02627) of 120 nt. The structure determined by comparative sequence analysis is in the larger (blue color) cluster with a probability of 0.591 and the minimum free energy (MFE) structure is in the smaller cluster (purple color) with a probability of 0.409. The coordinates for a structure is (axis 1, axis 2, energy), where the horizontal axes are from multidimensional scaling (MDS; (18)) for presenting highdimensional objects in typically two dimensions, and the vertical axis is the free energy of a secondary structure. The base-pair distances between structures are used for MDS. The coordinates are (21.50, -5.73, -46.80) for the structure determined by comparative sequence analysis, (-27.92, -0.45, -50.50) for the MFE structure, (6.55, 3.15, -36.40) for the ensemble centroid, (20.14, -2.88, -45.80) for the larger cluster centroid, and (-25.95, -0.34, -50.50) for the smaller cluster centroid. Figure and caption from (5).*

the full sequence. Hence the energetic terms calculated using Sfold is identical to that in Mfold, only that instead of obtaining a lowest energy structure, one obtains a Boltzmann weighted distribution. The authors of Sfold show that this statistical sampling followed by clustering gives improved prediction results (5).

Probabilistic approach

Probabilistic methods score secondary structure predictions based on a series of conditional probabilities. These conditional probabilities are in turn determined from a known set of secondary structures, as opposed to experimentally determined energies.

For instance, in the SCFG representation, given a sequence $x = \mathrm{AGUCU}$ with secondary structure $y = ((.))$ (where a pair of matching parentheses indicates a base pair, and a period indicates an unpaired base), the unique parse σ corresponding to y is

$$S \rightarrow \mathrm{A}S\mathrm{U} \rightarrow \mathrm{AG}S\mathrm{CU} \rightarrow \mathrm{AGU}S\mathrm{CU} \rightarrow \mathrm{AGUCU}$$

and the corresponding joint probability is

$$P\left(x, \sigma\right) = p_{S \rightarrow \mathrm{A}S\mathrm{U}} \cdot p_{S \rightarrow \mathrm{G}S\mathrm{C}} \cdot p_{S \rightarrow \mathrm{U}S} \cdot p_{S \rightarrow \varepsilon}$$

where ε indicates the terminal transformation.

Then, we have the probability of the secondary structure y given sequence x as:

$$P\left(y|x\right) = \sum_{\sigma \in y} P\left(\sigma|x\right) = \frac{\sum_{\sigma \in y} P\left(x, \sigma\right)}{\sum_{\sigma' \in \Omega(x)} P\left(x, \sigma'\right)}$$

where $\Omega\left(x\right)$ is the space of all possible parses of x. (This example was adapted from reference (9), and the reader is referred to the paper for more detailed information.)

CLLMs are more generalized versions of SCFGs. This flexibility allows the inclusion of complex scoring schemes into the CLLM framework. CONTRAfold combines the probabilistic approach of CLLM with the energy-based model of Mfold, which, while still not physical, is a more sophisticated development compared to the standard SCFG, and gives the user more flexibility in controlling the sensitivity and specificity of the algorithm than in Mfold (9).

The weighting on each probability (or equivalently, the energy for each interaction) is parameterized using a training set of known secondary structures obtained from the Rfam database (9). Hence unlike Mfold which obtains thermodynamic energies experimentally, CONTRAfold uses a bioinformatic approach to parameterize the probabilities used in their model.

The authors compared their algorithm to some other available softwares, and found that CONTRAfold gave the highest sensitivity for a given specificity (see figure 2).

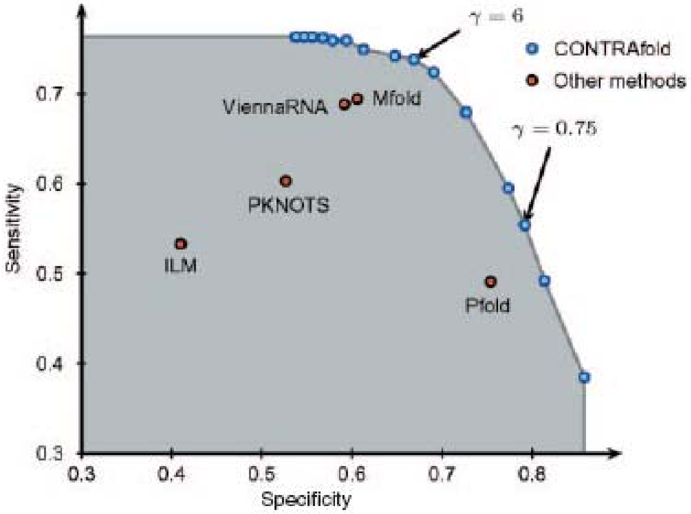

Figure 2: *CONTRAfold does better than Mfold, ViennaRNA, PKNOTS, ILM and Pfold. γ is the parameter that controls the tradeoff between the sensitivity and specificity of CONTRAfold. Figure was taken from reference (9).*

Use of fragments library

Nucleotide cyclic motifs (NCMs) are fragments of RNA structures obtained from the high resolution x-ray crystallographic structure of the 23S ribosomal RNA of *Haloarcula marismortui* (19). The particular database used in MC-FOLD (13) contains lone-pair loops up to six nucleotides, and double-stranded NCMs up to eight nucleotides. For instance, there can be loop NCMs of sequence AAA, AAAA and so on, while the double-stranded version can incorporate bulges by having uneven lengths of both ends of the NCM. An example is a 4_2-UGGUAA, which has a GG bulge in the 5' strand (first integer of NCM name represents number of nucleotides in the 5' strand, and the second integer is the same for the 3' strand). This comprehensive fragments library allows for the calculation of probabilities of secondary structure models. Hence this approach is somewhat similar to CONTRAfold conceptually, but the library used for the probability calculation is based on NCMs.

MC-FOLD proceeds recursively. A list of initiation sites is first determined, and this is assigned lone pair NCMs (i.e. single stranded regions), and the rest of the sequence is matched to double-stranded NCMs. The scoring used in MC-FOLD assumes a Boltzmann distribution (13):

$$\Phi\left(\text{structure}|\text{sequence}\right) = -RT \ln \Psi\left(\text{structure}|\text{sequence}\right)$$

with

$$\begin{aligned}\Psi\left(\text{structure}|\text{sequence}\right) = {} & \Psi\left(\text{NCMs}|\text{sequence}\right) \times \Psi\left(\text{junctions}|\text{NCMs}\right) \\ & \times\Psi\left(\text{hinges}|\text{junctions}\right) \times \Psi\left(\text{pairs}|\text{hinges}\right)\end{aligned}$$

$\Psi\left(x|y\right)$ represents the probability of x given y, as determined in the NCM library.

This procedure can be slow, and the server version of MC-FOLD is limited to 150 nucleotides (20).

Discussion

The four algorithms (Mfold, Sfold, CONTRAfold and MC-FOLD) discussed above are tested using the P4-P6 domain of *Tetrahymena ribozyme* as the test case. In all cases, default parameters were used. This test is not expected to be statistically representative, but rather to provide some insight to the efficacy of the different RNA secondary structure prediction algorithms. P4-P6 was chosen since it is relatively long (159 bases), with several loops and branches which provides complexity for testing (21).

Mfold and Sfold are expected to be more similar, since they adopt the same energy rules, with the latter including statistical sampling to generate an ensemble of structures that are then clustered. The lowest energy result from Mfold is compared to the native state structure (as determined from crystal structure (21)), while in the case of Sfold, the centroid of the largest cluster is taken as the representative one. In general, the largest cluster might not correspond to the one that the lowest energy structure resides in, but it appears to be the case for our particular test sequence. However, since the centroid is determined from the mean of the main cluster (5), the result from Sfold is different from Mfold.

The Mfold and Sfold servers both ran very quickly, and gave thermodynamic parameters along with their structures. On the other hand, since CONTRAfold works with a probabilistic model, no energies were given with the secondary structure structure, and only the structure with the best score is given (server version (22)). CONTRAfold also returned results very efficiently.

MC-FOLD was a lot slower, and since the server version was limited to 150 bases (P4-P6 has 159), the package had to be downloaded and run on the Bio-x^2 cluster (23). The job took almost an hour of CPU time on a Intel E5345 (Clovertown) quad-core 2.33GHz processor. Again, the best scoring structure was taken as the representive one.

The representative secondary structures are shown in figure 3, and were rendered using the server version of plot_rna from the CONTRAfold package made available in the MC-FOLD website (24). To compare these structures, two parameters were used, namely:

SN = Number of correctly predicted base-pairs/Total number of base-pairs in native structure

and

PPV = Number of correctly predicted base-pairs/Total number of predicted base-pairs

which are standards commonly used to compare quality of predicted secondary structures (5, 25). Ideally, the predicted structures should have SN and PPV closest to 1 as possible. Figure 4 indicates that CONTRAfold does best using this test.

Clearly all the aforementioned algorithms have flaws of their own, and it is likely that how well each algorithm works is somewhat system specific. Mfold and Sfold depend strongly on the experimental thermodynamic data, which might not be accurate when extrapolated to larger RNA. MC-FOLD, on the other hand, is slow, and the server version is limited by the number of nucleotides allowed. While MC-FOLD appears to do badly in this particular test case, it is likely that it will work better for smaller RNA, as illustrated by the authors (13). It is also possible that the best scored RNA secondary structure might not be the closest to the native one, which likely indicates the scoring is not ideal.

Even though CONTRAfold does well for the test, the server version only provides the best scoring secondary structure (22). Also, the score cannot be directly converted to a thermodynamic property, unlike in Mfold or Sfold. Having the thermodynamic information

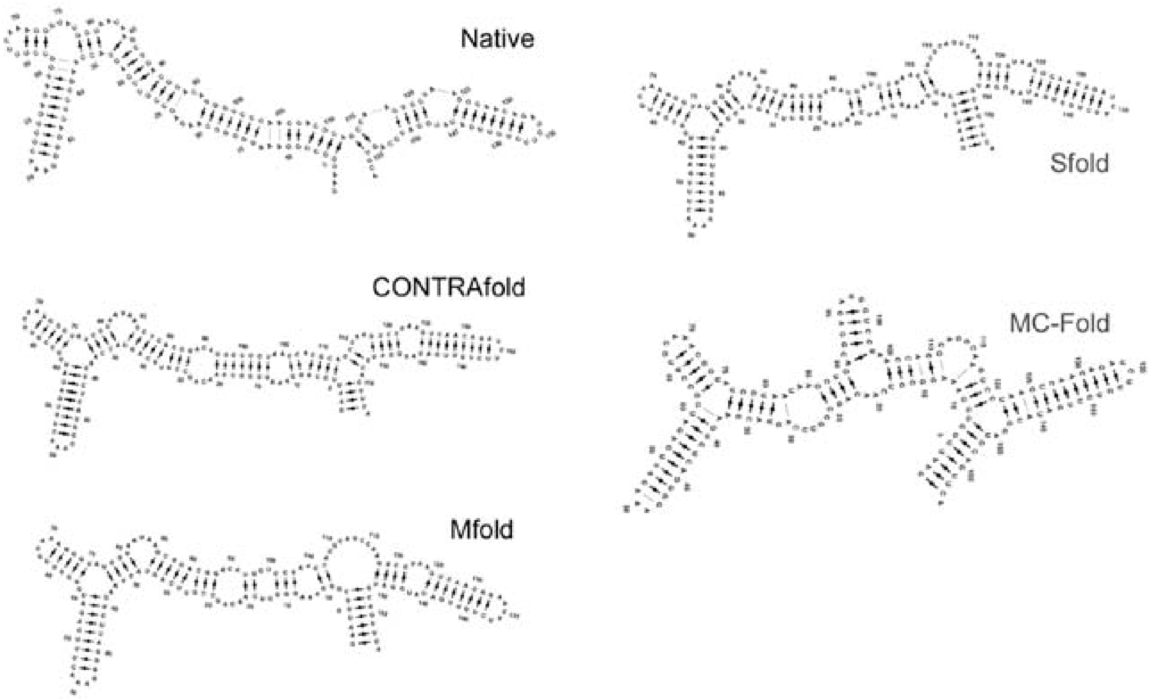

Figure 3: *Secondary structures obtained using the various algorithms. The native secondary structure was determined from the crystal structure (21). The secondary structures were rendered using the server version of plot_rna from the CONTRAfold package made available in the MC-FOLD website (24).*

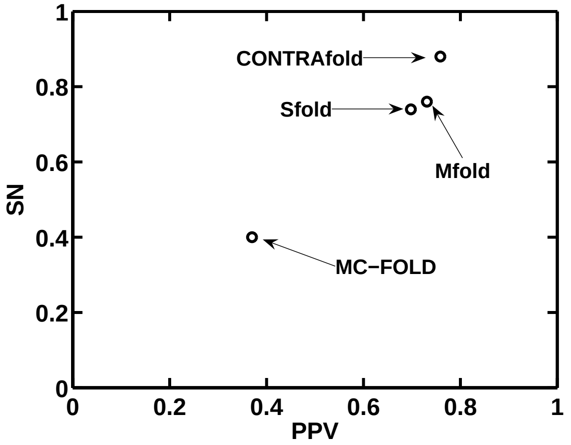

Figure 4: *SV (a measure of true positives) against PPV (an inverse measure of false positives) plot for the four algorithms tested. The ideal case would have both SV and PPV as 1. The results suggest that for the case of P4-P6, CONTRAfold does best, while MC-fold does the worst.*

from Mfold or Sfold can be very valuable if the user would like to do a comparison of energies with, for instance, different RNA sequences, or to use the energies for other thermodynamic calculations. Even though these values are dependent on how well the experimental data can be extrapolated, hairpin pulling experiments suggest that for small nucleic acids, the results can be very predictive (26).

In all these cases, the secondary structure prediction was conducted independent (at least explicitly) of tertiary interactions. However, *in vivo*, it is likely that tertiary structure affects secondary structure as well (i.e. the hierachical assumption in RNA folding might not be valid in all cases). This is somewhat included indirectly in CONTRAfold and MC-FOLD, since the probabilistic model of the former, and the fragment library of the latter were generated from existing RNA structural databases. The Turner rules used in Mfold and Sfold however, do no account for tertiary structure effects.

Conclusion

Here I've given brief overviews on three main methods of RNA secondary structure prediction, namely: energy minimization, probabilistic approach and the use of fragments library. Mfold (energy minimization), Sfold (energy minimization and statistical sampling), CON-

TRAfold (probabilistic model) and MC-FOLD (fragments library) were then tested using the sequence of the P4-P6 domain of the *Tetrahymena ribozyme*. The native secondary structure used to validate the models was determined from the crystal structure of the RNA (21). Two evaluation criteria (SN and PPV) were used, which indirectly measure the number of true and false positive base-pairs were introduced in each secondary structure model. It was found that for this particular test system, based on the mentioned grading criteria, CONTRAfold did the best.

References

[1] Raymond F. Gesteland, J. F. A., Thomas R. Cech, editor, 2005. The RNA World. Cold Spring Harbor Laboratory Press.

[2] Mandal, M., and R. R. Breaker, 2004. Gene regulation by riboswitches. *Nat Rev Mol Cell Biol* 5:451–463. `http://dx.doi.org/10.1038/nrm1403`.

[3] M. Zuker, D. H. T., D. H. Mathews, 1999. RNA Biochemistry and Biotechnology, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, NL, chapter 2, 11–43.

[4] Ding, Y., and C. E. Lawrence, 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* 31:7280–7301.

[5] Ding, Y., C. Y. Chan, and C. E. Lawrence, 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* 11:1157–1166. `http://dx.doi.org/10.1261/rna.2500605`.

[6] Zuker, M., 1989. On finding all suboptimal foldings of an RNA molecule. *Science* 244:48–52.

[7] Mathews, D. H., J. Sabina, M. Zuker, and D. H. Turner, 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288:911–940. `http://dx.doi.org/10.1006/jmbi.1999.2700`.

[8] Mathews, D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* 101:7287–7292. `http://dx.doi.org/10.1073/pnas.0401799101`.

[9] Do, C. B., D. A. Woods, and S. Batzoglou, 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 22:e90–e98. http://dx.doi.org/10.1093/bioinformatics/btl246.

[10] Richard Durbin, A. K. G. M., Sean R. Eddy, 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

[11] Knudsen, B., and J. Hein, 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* 15:446–454.

[12] Knudsen, B., and J. Hein, 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* 31:3423–3428.

[13] Parisien, M., and F. Major, 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452:51–55. http://dx.doi.org/10.1038/nature06684.

[14] Das, R., and D. Baker, 2007. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A* 104:14664–14669. http://dx.doi.org/10.1073/pnas.0703836104.

[15] Freier, S. M., R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner, 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A* 83:9373–9377.

[16] Walter, A. E., D. H. Turner, J. Kim, M. H. Lyttle, P. Mller, D. H. Mathews, and M. Zuker, 1994. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci U S A* 91:9218–9222.

[17] Ding, Y., C. Y. Chan, and C. E. Lawrence, 2004. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* 32:W135–W141. http://dx.doi.org/10.1093/nar/gkh449.

[18] J. B. Kruskal, M. W., 1977. Multidimensional scaling. Sage Publications, Beverly Hills, CA.

[19] Lemieux, S., and F. Major, 2006. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res* 34:2340–2346. http://dx.doi.org/10.1093/nar/gkl120.

[20] http://www.major.iric.ca/MC-Fold/.

[21] Cate, J. H., A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, C. E. Kundrot, T. R. Cech, and J. A. Doudna, 1996. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 273:1678–1685.

[22] `http://contra.stanford.edu/contrafold/server.html`.

[23] The Bio-X2 computing cluster is supported by National Science Foundation Award CNS-0619926.

[24] `http://www.major.iric.ca/MC-Pipeline/`.

[25] `http://www.ece.tamu.edu/ bjyoon/ecen689-612-spring08/ecen689-612-final-project.pdf`.

[26] Woodside, M. T., W. M. Behnke-Parks, K. Larizadeh, K. Travers, D. Herschlag, and S. M. Block, 2006. Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins. *Proc Natl Acad Sci U S A* 103:6190–6195. `http://dx.doi.org/10.1073/pnas.0511048103`.