Bridging the gap in RNA structure prediction

Bruce A Shapiro[1], Yaroslava G Yingling[1], Wojciech Kasprzak[2] and Eckart Bindewald[2]

The field of RNA structure prediction has experienced significant advances in the past several years, thanks to the availability of new experimental data and improved computational methodologies. These methods determine RNA secondary structures and pseudoknots from sequence alignments, thermodynamics-based dynamic programming algorithms, genetic algorithms and combined approaches. Computational RNA three-dimensional modeling uses this information in conjunction with manual manipulation, constraint satisfaction methods, molecular mechanics and molecular dynamics. The ultimate goal of automatically producing RNA three-dimensional models from given secondary and tertiary structure data, however, is still not fully realized. Recent developments in the computational prediction of RNA structure have helped bridge the gap between RNA secondary structure prediction, including pseudoknots, and three-dimensional modeling of RNA.

Addresses
[1] Center for Cancer Research Nanobiology Program, NCI-Frederick, Building 469, Room 150, Frederick, MD 21702, USA
[2] Basic Research Program, SAIC-Frederick Inc, NCI-Frederick, Building 469, Room 150, Frederick, MD 21702, USA

Corresponding author: Shapiro, Bruce A (bshapiro@ncifcrf.gov)

Current Opinion in Structural Biology 2007, **17**:157–165

This review comes from a themed issue on
Theory and simulation
Edited by Richard Lavery and Kim A Sharp

Available online 23rd March 2007

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2007.03.001

Introduction

Knowledge of the 3D structure and dynamics of RNA is important for understanding its function in the cell. Experimental techniques used to derive a 3D structure are time consuming and expensive, and include X-ray crystallography of single crystals of purified RNA molecules, NMR spectroscopy and cryo-electron microscopy. Moreover, the complexity and flexibility of RNA molecules makes the determination of 3D structures even more difficult. Thus, the disparity is increasing between known RNA 3D structures and known RNA sequences. This encourages the use of computational methods to obtain information on RNA 3D conformations. Much progress is being made in this research area, and the prediction of small and simple RNA structures is now a perfectly realistic goal. However, more complex structures with many helical stems and pseudoknots are much more difficult to predict. A simple pseudoknot (H-type) can be thought of as a secondary structure that forms from a hairpin loop that base pairs with a single-stranded sequence that is outside the loop, forming another helical stem. In most cases, these stems coaxially stack on one another (Figure 1 contains an example of a pseudoknot). More complex pseudoknot structures are also possible, as described later. The determination of RNA 3D structures is usually attempted by the use of a combination of theoretical data, algorithms and experimental observations.

The prediction of an RNA 3D structure directly from its sequence can be accomplished either by a detailed simulation of the folding process or by searching the entire conformational space for the correct fold. However, both approaches are well beyond current computational capabilities. Folding simulations of short sequences are possible via atomistic molecular dynamics, but might not be entirely accurate due to estimations in force-fields and the feasible duration of run times. Atomic-level simulations of large and complex structures are beyond current computational resources because of the enormous number of possible conformations. In addition, environmental factors, such as ion concentrations, solvent, interacting proteins and other RNAs, ultimately contribute to RNA folding pathways.

The RNA folding process is believed to be partly hierarchical, whereby helical domains fold first followed by compaction of the structure via tertiary interactions and associations between RNA domains and motifs. Thus, a more practical approach is to predict an RNA 3D structure using algorithms that are constrained by experimentally derived data. This experimental data might be obtained by new methods such as SHAPE [1••] or microarrays for chemical mapping [2•]. The difficulty of secondary structure prediction is exemplified by the fact that a sequence of n nucleotides can form on the order of 1.8^n possible secondary structures [3]. Therefore, numerous approaches to the problem have been used that combine the strengths of computational and experimental methods.

Computational secondary structure prediction falls into two general categories: one uses multiple sequence alignments to predict structures and the other predicts the structure of single sequences using free energy minimization. However, some programs combine these concepts. The accuracy of predictions is usually best for methods that consider secondary structures common to multiple

Figure 1

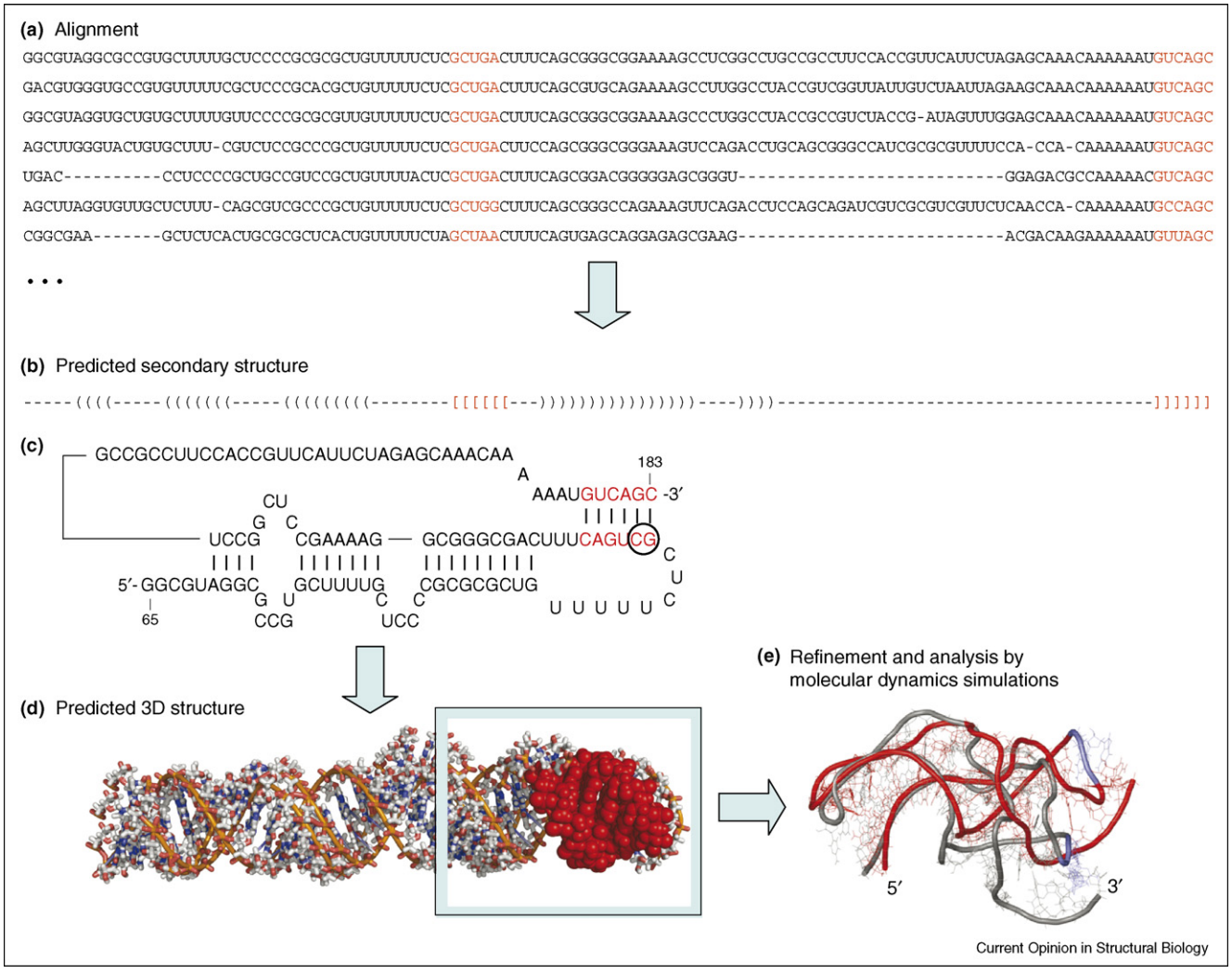

Diagram showing semi-automatic RNA 3D structure prediction. The workflow starts with a set of aligned sequences and ends with a refined 3D model of the wild-type human telomerase RNA pseudoknot region (pseudoknot bases are depicted in red). **(a)** The fraction of the 35 sequences used in the alignment (derived from the complete Rfam entry for vertebrate telomerase, RF00024) that corresponds to the telomerase pseudoknot region. Columns corresponding to gaps in the first sequence of the alignment (wild-type human telomerase) are not shown. Bases corresponding to positions 65, 67 and 68 are predicted to pair with positions outside the chosen pseudoknot region; these base pairs are not used to generate the 3D model. **(b)** KNetFold [6••] secondary structure prediction with the pseudoknot. **(c)** The secondary structure representation derived from KNetFold. The circled bases might be implicated in the genetic mutation causing dyskeratosis congenita (DKC). This structure is very similar to that shown in [72]. **(d)** The automatically generated 3D structure of the predicted wild-type region. The boxed region is believed to be critical for telomerase function and was further refined with manual manipulation and molecular dynamics simulations. **(e)** An overlay of the predicted optimized structure after molecular dynamics simulations of the wild-type telomerase RNA (red) and the DKC-mutated telomerase RNA (grey with the mutated bases in blue) [61••,70].

sequences [4,5,6••,7]. Gardner and Giegerich [8••] presented a comprehensive review and tests of RNA secondary structure prediction programs in their 2004 article. A review of free energy minimization methods, with an emphasis on dynamic programming algorithms (DPAs), was recently presented by Mathews and Turner [9••].

Secondary structure prediction methods are discussed in this review, with a strong emphasis on pseudoknot prediction. This is because pseudoknots add constraints that reduce flexibility, thereby simplifying somewhat the characterization of the complete 3D RNA conformation. This review is divided into three sections. The first describes methods for secondary structure and pseudoknot prediction given single-sequence input. We shall give a brief overview of DPA-based methods, and then add a discussion on genetic algorithms (GAs) and some other methodologies. The second section describes secondary structure and pseudoknot prediction using multiple

sequence alignments, and the third section describes programs that can use this data to determine 3D models of RNA.

RNA secondary structure and pseudoknot prediction using single sequences

Dynamic programming algorithms

The most familiar secondary structure prediction programs, such as Mfold [10], RNAfold [11] and RNAstructure [10,12•], are based on DPAs. They are deterministic in nature and guarantee returning the lowest free energy structure, within the accuracy limitations of the free energy rules employed. They can also enumerate a sample of energetically suboptimal structures requested by the user. The latest version of RNAstructure enables the user to include experimentally derived structure probing data as constraints in the folding algorithm [12•].

However, the biologically functional conformation of a given RNA molecule might not correspond to the minimum free energy structure. The problem then becomes one of searching for a relevant suboptimal structure or finding a subset of representative structures with the best positive predicted value, that is, the highest percentage of predicted base pairs that are correct [9••]. Predicting the probabilities of all base pairs in a structure is one way of determining such structures [13]. The programs Sfold [14,15] and RNAshapes [16••,17,18••] narrow the search for relevant solutions in the usually large solution spaces of DPA programs to a relatively few representative structures. In the case of Sfold, the predicted suboptimal structures are sampled based on the Boltzmann probability distribution. This statistical sampling can be used to produce a so-called centroid structure representing the whole solution space ensemble. Multiple clusters and their centroids are produced from the ensemble. None of the centroids has to represent a minimum free energy structure. The RNAshapes program extends the abstract shape analysis to the complete suboptimal solution space (not just a sample of it) within a requested energy range [11,19] and calculates cumulative probabilities of structures belonging to the identified shapes [18••].

Genetic algorithms and other methodologies

Some secondary structure prediction algorithms use statistical sampling of known RNA secondary structures to create a model that can then be used to predict the secondary structure of a given sequence. The new algorithm CONTRAfold uses such a methodology [20••]. Its strength lies in training on a large set of experimentally verified structures from the Rfam database [21] and encoding thermodynamic knowledge in a feature-rich scoring scheme provided by a conditional log-linear model (CLLM).

However, an RNA structure is not necessarily static. A molecule might pass through several active and inactive conformations during its lifetime. These states might be related to the kinetics of full sequence folding or folding during transcription (sequence synthesis, i.e. elongation). They might also result from interactions with its environment. Capturing radically different but biologically functional states of some sequences might still require going beyond what the statistical sampling of the suboptimal solution spaces of DPA algorithms can show.

The program KINEfold [22,23••] implements stochastic simulations of folding kinetics using a Monte Carlo methodology. It permits the formation of pseudoknots using polymer theory, with constraints based on topological and geometric considerations, for sequences up to 300–400 nucleotides in length.

GAs are based on the concepts of biological evolution and the survival of the fittest individuals [24]. GAs can go beyond earlier pure Monte Carlo algorithm based approaches [25,26] in that they include results of evolving intermediate structural states, in which the folding kinetics might involve unfolding and refolding of domains within transitional structures. The essential GA operations, repeated in every step (generation) of the algorithm, are mutation, crossover and selection. They introduce, respectively, random changes to the solutions, exchange of parts of the evolving solutions and selection of 'fit' solutions for survival into the next generation. Because GAs are stochastic in nature, they are usually run multiple times to produce consensus structures.

One GA implementation, which relies on a relatively small number of co-evolving structures, is part of the STAR program and is aimed at personal computer platforms [27,28]. Our group has developed a massively parallel GA, MPGAfold, for the prediction of RNA secondary structure. Described in detail in [29–33], this GA implementation evolves, essentially in parallel, a population of thousands of structures, logically connected by a 2D toroidally wrapped mesh. Initially developed for special SIMD and MIMD parallel architectures, its current implementation makes it available on parallel Linux clusters. The stochastic parallel nature of MPGAfold is quite different from DPA paradigms for RNA structure prediction. New structural motifs, including multiple nucleating structures, can form at the same time within a particular generation in a given structure in the population. These motifs can persist for several generations before possibly changing to a new motif, thus suggesting a sense of the importance of a particular motif or structure. Both STAR and MPGAfold use the same free energy rules as the DPA-based programs for their fitness criterion, with the objective of converging to stable low energy states, which might not be the minimum free energy state. An important MPGAfold parameter is population size. Structures predicted in runs with different

population sizes have been shown to capture significant intermediate and final RNA secondary structure states, thus elucidating the dynamic folding process inherent to many RNA molecules [32,34•,35•,36••,37–39]. This program can also incorporate some experimental data and bias the maturation process toward known interactions via the use of so-called sticky stems [32,33,34•]. MPGAfold can be run in close conjunction with our StructureLab program and our recently developed visualizer program for interactive viewing and analysis of the fitness, trace and pseudoknot maps of the evolving solution spaces [36••,40].

MPGAfold, STAR and KINEfold are capable of simulating co-transcriptional folding (folding with sequence elongation), and can capture folding kinetics unique to RNA transcription.

Prediction of RNA secondary structures with pseudoknots

Pseudoknot-free secondary structures can be represented as tree topologies, which are naturally suited to free energy calculations based on the assumption that the total free energy of a structure (a tree) is the sum of the free energies of its independent elements (branches). The problem of including pseudoknot interactions in secondary structure prediction requires the evaluation of the free energy of a graph topology, instead of a tree. In addition, at the present time, no precise set of energy rules exists for the varied pseudoknot topologies. Rivas and Eddy [41] presented an empirical set of parameters to compute the free energy of secondary structures with pseudoknots. DPAs have been extended to incorporate several differently limited classes of pseudoknots at the cost of increasing their complexity. A reduced-complexity class can be represented by the simple H-type pseudoknot described above, whereas a more complex class can incorporate multiple nested interactions that might be recursive (i.e. contain pseudoknots within pseudoknots). Although DPAs scale in $O(n^3)$ time for secondary structure prediction of a sequence n nucleotides long, the scaling becomes $O(n^6)$ for the most general program, Pseudoknots [41], $O(n^5)$ for NUPACK [42] and $O(n^4)$ for pknotsRG [43]. Heuristic algorithms that are capable of computing structures with pseudoknots but do not guarantee finding the minimum energy solutions are less complex and, therefore, can be used to predict the structures of longer sequences. They are implemented in the STAR package [28], ILM [44,45•] and HotKnots [46]. HotKnots, the most recent of these, considers multiple partial solutions with multiple substructure additions for each of them. MPGAfold is also capable of predicting H-type pseudoknots [31] with minimal impact on speed.

Pseudobase [47,48], a web-retrievable database of RNA pseudoknots, is a good source of experimentally and computationally determined pseudoknot structures. The reliability of the pseudoknot structures presented is left to user discretion, with the help of the database field ''supported by'', which indicates how the pseudoknot was determined.

RNA secondary structure and pseudoknot prediction using multiple sequence alignments

Several programs predict RNA secondary structures, including pseudoknots, using a set of aligned sequences. All reviewed methods use the following paradigm: in the first stage, a matrix with scores corresponding to each base pair is computed. These scores typically incorporate both thermodynamic and covariation information. In the second stage, this matrix is mapped to one unique secondary structure. Both the computation of the scoring matrix and the mapping of the matrix to the secondary structure differs among the various programs. One mapping approach is called maximum weighted matching (MWM) [49]. The idea is to find a set of non-overlapping edges in a graph (each edge, in this case, corresponds to a potential base pair), such that the sum of the weights of the edges (in this case, the sum of the base pair scores) is maximal. This is an attractive approach; however, in practice it has the problem of sometimes predicting spurious base pairs [50•]. The programs ILM and HXMATCH address this problem differently (see below).

A commonly used measure for covariation is mutual information [6••,51•]. However, other measures that take RNA base pairing preferences into account are also used. The covariance measure of Hofacker *et al.* [7] has the advantage of being able to detect consistent but non-compensatory mutations (e.g. the mutation of base pair GC to GU) [7,52].

Programs using multiple sequence alignments

The program ILM (iterative loop matching) uses as a base pair score a linear combination of a thermodynamic term and a covariation term [45•]. Mapping to a secondary structure starts with a pseudoknot-free structure using maximum circular matching, sometimes referred to as the Nussinov algorithm [53]. Iteratively, the helix with the highest score is chosen to be part of the predicted structure until there are no remaining base pairs to be found.

The program KNetFold computes, for each pair of alignment columns, a thermodynamic score, the fraction of complementary base pairs and a covariance measure (the mutual information) [6••,51•]. The resulting base pair scoring matrix is not a linear combination of these three scores, but is instead based on a k-nearest neighbor machine learning approach that classifies small 5×5 neighborhoods of a base pair. Mapping to a unique secondary structure with possible pseudoknots is performed by iteratively choosing the highest-ranking base pairs until all non-overlapping base pairs above a cutoff

score are found. An example of the application of KNetFold is illustrated in Figure 1, in which the prediction of the human telomerase RNA pseudoknot is depicted.

The program HXMATCH uses a linear combination of a thermodynamic score and a covariation score [7,50•]. The thermodynamic score is computed by using the longest helix that a base pair is assigned to. This is averaged over the different sequences of the alignment. HXMATCH then uses the MWM algorithm with a post-processing step restricting the allowed solutions to bi-secondary structures, that is to say, the superimposition of two predicted pseudoknot-free secondary structures, thus producing a secondary structure with possible pseudoknots.

The program MIfold is a MATLAB® package that uses the mutual information measure or a covariance measure as a scoring matrix [54•]. It then uses this matrix in conjunction with the maximum circular matching algorithm to compute the highest-scoring pseudoknot-free structure [53]. The scoring matrix elements corresponding to the bases that are paired in this structure are, in the next stage, set to zero and a second pseudoknot-free structure is computed. The final structure is determined from bi-secondary structures.

RNA three-dimensional structure prediction

To assist in RNA 3D structure prediction, several programs have been developed and successfully applied. Most use data derived from experiments and programs for secondary structure and pseudoknot prediction. RNA 3D structure prediction programs include YAMMP [55], NAB [56], ERNA-3D [57], MANIP [58], S2S [59], MC-Sym [60] and RNA2D3D [61••] (see Table 1). The increasing number of known RNA 3D structures that have been organized into databases (such as PDB [62], SCOR [63], RNABase [64] and NCIR [65]) makes it possible to integrate and refine the prediction of 3D structures. However, no automated process can successfully generate the entire structure without problem-specific input and user knowledge of RNA structure.

Programs for three-dimensional structure prediction

ERNA-3D can produce a 3D representation of an RNA from a known secondary structure [57,66••]. It automatically generates representations of A-form helices directly from the specified base-paired regions of the secondary structure. In the first version of the program, single strands were derived from iterated rotations along the backbone. In the most recent version, single-stranded regions and motifs are extracted from known high-resolution structures of other RNAs using the SCOR database and incorporated into the models. If high-resolution data are available for a particular motif, then these nucleotides can be positioned manually [66••]. Comparative sequence analysis and the ERNA-3D program have been used to determine the 3D backbone arrangement of a small domain of signal recognition particle RNA that includes a pseudoknot [57]. Recently, they were used to build a high-resolution 3D structure of the transfer-messenger RNAs (tmRNAs) of *Escherichia coli*, *Bacillus anthracis* and *Caulobacter crescentus* [66••]. The resulting models show functionally significant features, such as single-stranded regions and the close proximity between the tRNA-like domain (TLD) and the resume codon (the codon in the mRNA portion of the tmRNA where translation is continued), thus significantly advancing our understanding of *trans*-translation.

Table 1

Comparison table of programs for building all-atom high-resolution RNA 3D structures.

Software/method	Input	Output	User's input	Comment
MANIP	Database of known fragments and secondary structure	Complex 3D architecture	Rotation, translation of fragments; interactive manipulation.	
S2S	3D structure	Multiple alignments	Interactive manipulation	Need 3D structure
NAB	Secondary structure and distance constraints	3D structure	Interactive manipulation	Possible use of known tertiary RNA structure fragments. Built-in energy minimization and molecular dynamics optimization.
ERNA-3D	Secondary structure	3D structures	Interactive manipulation	Possible use of known tertiary RNA structure fragments
MC-Sym	Secondary structure; distance, torsion and other structural constraints; database of known fragments.	Series of 3D structures		Counterions can be implicitly represented; no interactive manipulation
RNA2D3D	Secondary structure; can also use known fragments.	3D structure	Interactive manipulation if needed	Possible automatic stacking of helices; compactification, kissing loops and pseudoknots. Built-in molecular mechanics and dynamics.
YAMMP (YUP)	Reduced model representations and secondary structure	3D structure	Interactive and batch mode	

MANIP is a program that can assemble known RNA fragments into a complex 3D architecture [58]. It requires a database containing RNA 3D fragments with a specified sequence. It automatically recognizes and displays allowed hydrogen bonds between residues. The program is interfaced with the online refinement tool NUCLIN-NUCLSQ; the refinement is based on various constraints, such as canonical and non-canonical base pairing, covalent geometry, stereochemistry and van der Waals contacts. MANIP has been used to predict 3D models of 377-nucleotide RNase P RNAs of different structural subtypes at atomic resolution [67,68••].

The program Nucleic Acid Builder (NAB) describes nucleic acid structures in a hierarchical fashion and can be used to construct helical and non-helical nucleic acids up to a few hundred nucleotides in size ([56]; http://www.scripps.edu/mb/case/). Some of the features of this software include specified base transformations relative to other bases or base layouts along arbitrary curves in space. Sets of distance constraints can be applied to structures (especially non-helical regions, hairpins, loops and pseudoknots) based on specific hydrogen bonds and cross-linking or footprinting results, or on derived data from known 3D structures. Constructed models can be optimized or modified using energy minimization or molecular dynamics simulations.

MC-Sym uses symbolic and numerical computations to build 3D RNA structures using structural data that are represented as a series of application domain symbols and RNA structure expert terms. Numerical computation is used to refine the symbolic representations [60]. It builds RNA 3D structures using coordinates and relationships between residues extracted from known 3D structures. Structural constraints can be interactively applied to the building procedure. These constraints are taken from cross-linking assays, and topographic and crystallographic data. Model RNA structures are further refined with molecular mechanics calculations. MC-Sym has been used to generate several models. Most recently, the structure of the hairpin ribozyme catalytic core was determined [69•]. The methodology provided an alternative conformation of the active ribozyme, which is supported by crystallographic data and by cross-linking experiments.

Recently, we developed RNA2D3D — a program that can generate, view and compare 3D RNA structures. Helical stems are generated from the reference triad of any of its nucleotides using helical coordinates. Unpaired nucleotides, bulges, hairpin loops, branching loops and other non-helical motifs are generated using the coordinates of their reference triad relative to the 5′ neighboring nucleotide and are placed in 3D space using a special 3D embedding procedure. This procedure equally spaces atomic models of nucleotides along the fixed backbone, which initially resides in a planar representation of the secondary structure of the RNA before 3D helical winding takes place. Ultimately, a first-order approximation of the actual 3D molecule is established. Structure refinement involves interactive editing via rotation and translation of a nucleotide or a group of nucleotides to remove structural clashes, thus enforcing tertiary interactions, and modification of mutual stacking. Specified pseudoknot stems can be moved relative to each other to modify their mutual stacking or the nucleotides in the loop can be rotated to remove tertiary interactions. Stems can be 'compactified' (extended into single-stranded regions to form non-canonical base pairs) or interactively stacked. Database motifs can also be included. The 3D model can be further refined with built-in molecular mechanics and molecular dynamics simulations. We have recently published a study in which we used the RNA2D3D software and molecular dynamics to predict the 3D structure of the pseudoknot domain of wild-type human telomerase RNA [61••,70]. In addition, the implications of genetic mutations for its structure were explained. An application of RNA2D3D is illustrated in Figure 1. RNA2D3D was given the pseudoknot prediction from KNetFold for the full wild-type human telomerase pseudoknot. A compactified rendition of the pseudoknot was automatically generated and later refined.

The software S2S (sequence to structure) has been designed to construct multiple alignments of RNA molecules for which 3D structures are known. It enables the display, manipulation and interconnection of RNA data, such as multiple sequence alignments and secondary and tertiary interactions [59].

Another methodology is based on the assumption that similar secondary structures possess similar tertiary interactions. An algorithm was developed that maps RNA 3D structures found in the PDB onto RNA 2D structures. In addition, algorithms were developed that measure secondary structure similarity and pick 3D structural motifs from a specially constructed database. The resultant product is a set of fragments from which an entire RNA can be built [71].

In general, all predicted 3D structures modeled using the methods discussed above need to be manually refined and adjusted to achieve meaningful structures. The initial model can be built using the programs discussed above, assisted by the use of standard RNA fragments. Molecular mechanics and molecular dynamics can be used to refine the resulting structures.

Conclusions

The ability to computationally determine RNA structure and function from sequence data is still quite limited. The issue becomes even more complex when one considers that the final structure of an RNA does not necessarily represent the full functionality of the RNA in

question. An RNA might fold into intermediate or alternative states that permit the molecule to partake in more than one function. A basic assumption that applies to all methodologies is that the 3D structure of an RNA can ultimately be determined from its given sequence and the environment in which the sequence finds itself. Solvent, ions, proteins and other RNAs are significant environmental factors determining the structure and function of an RNA. A few programs have the ability to incorporate information about known constraints and attempt to use this information in secondary structure prediction and molecular dynamics. The validity and accuracy of the predicted structures also depends on the nature of collected experimental observations, as well as on the starting secondary and tertiary structures. Even though the correctness of atomistic details depends on the limitations of the force-field used in molecular mechanics and molecular dynamics, the above-discussed methods can delineate the overall shape and spatial relationships inherent to an RNA molecule. The use of experimental techniques, such as X-ray crystallography, NMR or cryo-electron microscopy, is still necessary to determine precise atomic details of the structure.

To improve the accuracy of 3D structure prediction from secondary structures, the non-atomistic free energy rules that are commonly used in secondary structure prediction algorithms need to be further refined. The inclusion of more context sensitivity for loop structures would be helpful. Current rule sets calculate the energy of loops mostly based on their size rather than on their base composition. Also, the existing rules for pseudoknots rely mostly on non-thermodynamic empirical parameters. More experiments are required to determine the thermodynamic properties of simple pseudoknots and complex pseudoknots that may contain, for example, recursive structures. Finally, free energy rules for tertiary interactions are basically non-existent. This, however, is a difficult problem because thermodynamic experiments would have to be done on a variety of such interactions in a variety of contexts.

Acknowledgements

We wish to thank Hugo Martinez for his work on RNA2D3D. This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health (NIH), under contract N01-CO-12400. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. •• Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM: **RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE)**. *J Am Chem Soc* 2005, **127**:4223-4231.
This paper introduces a novel approach to experimental high-throughput determination of RNA secondary structures.

2. • Duan S, Mathews DH, Turner DH: **Interpreting oligonucleotide microarray data to determine RNA secondary structure: application to the 3′ end of *Bombyx mori* R2 RNA**. *Biochemistry* 2006, **45**:9819-9832.
This important paper shows how microarrays of 2′-*O*-methyl RNA 9-mers can be used to obtain RNA secondary structure constraints. The added experimental information is shown to significantly improve the prediction accuracy of RNA secondary structure prediction programs.

3. Zuker M, Sankoff D: **RNA secondary structures and their prediction**. *Bull Math Bio* 1984, **46**:591-621.

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

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

6. •• Bindewald E, Shapiro BA: **RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers**. *RNA* 2006, **12**:342-352.
A novel machine learning algorithm is presented that integrates mutual information, thermodynamic predictions and the fraction of complementary base pairs in an alignment to predict a consensus RNA secondary structure with possible pseudoknots. The method is shown to outperform RNAalifold and PFOLD for a test set of 49 Rfam alignments.

7. Hofacker IL, Fekete M, Stadler PF: **Secondary structure prediction for aligned RNA sequences**. *J Mol Biol* 2002, **319**:1059-1066.

8. •• Gardner PP, Giegerich R: **A comprehensive comparison of comparative RNA structure prediction approaches**. *BMC Bioinformatics* 2004, **5**:140-157.
This paper introduces a very useful scheme for classifying secondary structure prediction methods. It also provides the prediction accuracies of several methods with respect to a test set of RNA sequences.

9. •• Mathews DH, Turner DH: **Prediction of RNA secondary structure by free energy minimization**. *Curr Opin Struct Biol* 2006, **16**:270-278.
A comprehensive review of methods and programs that use the DPA algorithm for RNA structure prediction by free energy minimization. Thermodynamics, pseudoknot prediction, fidelity of structure prediction, solution space sampling, inclusion of experimental data and use of homologous sequences to find a common structure are discussed.

10. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure**. *J Mol Biol* 1999, **288**:911-940.

11. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures**. *Monatsh Chem* 1994, **125**:167-188.

12. • Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH: **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure**. *Proc Natl Acad Sci USA* 2004, **101**:7287-7292.
This paper describes how the DPA algorithm is improved by the addition of coaxial stacking energy to the free energy calculations, updates to the free energy rules and the incorporation of user-provided base pairing constraints that are derived from experimental probing data.

13. Mathews DH: **Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization**. *RNA* 2004, **10**:1178-1190.

14. Ding Y, Chan CY, Lawrence CE: **RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble**. *RNA* 2005, **11**:1157-1166.

15. Chan CY, Lawrence CE, Ding Y: **Structure clustering features on the Sfold web server**. *Bioinformatics* 2005, **21**:3926-3928.

16. •• Giegerich R, Voss B, Rehmsmeier M: **Abstract shapes of RNA**. *Nucleic Acids Res* 2004, **32**:4843-4851.

This paper presents a formalized approach to abstract shape analysis of RNA secondary structures. Each shape is treated as a distinct class represented by the so-called *shrep*, a structure of minimum energy within the class. *Shreps* are computed directly, bypassing the costly enumeration of suboptimal structures.

17. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R: **RNAshapes: an integrated RNA analysis package based on abstract shapes**. *Bioinformatics* 2006, **22**:500-503.

18. •• Voss B, Giegerich R, Rehmsmeier M: **Complete probabilistic analysis of RNA shapes**. *BMC Biol* 2006, **4**:5-27.
This paper builds upon the abstract shape analysis formalized in [16••] and adds complete probability computations for all identified distinct shapes. The approach is now available in the RNAshapes program.

19. Wuchty S, Fontana W, Hofacker IL, Schuster P: **Complete suboptimal folding of RNA and the stability of secondary structures**. *Biopolymers* 1999, **49**:145-165.

20. •• Do CB, Woods DA, Batzoglou S: **CONTRAfold: RNA secondary structure prediction without physics-based models**. *Bioinformatics* 2006, **22**:e90-e98.
This paper presents a single-sequence secondary structure prediction model based on multiple sequence/structure training. It builds on earlier stochastic context-free grammar training algorithms, but it uses a discriminative training approach to create a feature-rich CLLM from free energy data extracted from the training set. Comparison tests against several widely used single-sequence structure prediction programs are presented.

21. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic Acids Res* 2005, **33**:D121-D124.

22. Xayaphoummine A, Bucher T, Thalmann F, Isambert H: **Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations**. *Proc Natl Acad Sci USA* 2003, **100**:15310-15315.

23. •• Xayaphoummine A, Bucher T, Isambert H: **Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots**. *Nucleic Acids Res* 2005, **33**:W605-W610.
This paper describes the Kinefold server and the underlying stochastic algorithm, which performs RNA/DNA folding simulations on molecular timescales. Both full sequence folding and folding with elongation (co-transcriptional) are simulated.

24. Holland JH: *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications in Biology, Control, and Artificial Intelligence*. Cambridge, MA: MIT Press; 1992.

25. Martinez HM: **An RNA folding rule**. *Nucleic Acids Res* 1984, **12**:323-334.

26. Gultyaev AP: **The computer simulation of RNA folding involving pseudoknot formation**. *Nucleic Acids Res* 1991, **19**:2489-2494.

27. van Batenburg FH, Gultyaev AP, Pleij CW: **An APL-programmed genetic algorithm for the prediction of RNA secondary structure**. *J Theor Biol* 1995, **174**:269-280.

28. Gultyaev AP, van Batenburg FH, Pleij CW: **The computer simulation of RNA folding pathways using a genetic algorithm**. *J Mol Biol* 1995, **250**:37-51.

29. Shapiro BA, Navetta J: **A massively parallel genetic algorithm for RNA secondary structure prediction**. *J Supercomputing* 1994, **8**:195-207.

30. Shapiro BA, Wu JC: **An annealing mutation operator in the genetic algorithms for RNA folding**. *Comput Appl Biosci* 1996, **12**:171-180.

31. Shapiro BA, Wu JC: **Predicting RNA H-type pseudoknots with the massively parallel genetic algorithm**. *Comput Appl Biosci* 1997, **13**:459-471.

32. Shapiro BA, Bengali D, Kasprzak W, Wu JC: **RNA folding pathway functional intermediates: their prediction and analysis**. *J Mol Biol* 2001, **312**:27-44.

33. Shapiro BA, Wu JC, Bengali D, Potts MJ: **The massively parallel genetic algorithm for RNA folding: MIMD implementation and population variation**. *Bioinformatics* 2001, **17**:137-148.

34. • Kasprzak W, Bindewald E, Shapiro BA: **Structural polymorphism of the HIV-1 leader region explored by computational methods**. *Nucleic Acids Res* 2005, **33**:7151-7163.
This study presents a structural exploration of the HIV-1 5′-untranslated leader region. Full domain folding and co-transcriptional folding MPGAfold predictions are presented for 13 HIV-1/SIV strains; a KNetFold variant is used to generalize the results for 155 strains and StructureLab tools were used to analyze the results. The results are compared with *in vitro* and *ex vivo* data.

35. • Linnstaedt SD, Kasprzak WK, Shapiro BA, Casey JL: **The role of a metastable RNA secondary structure in hepatitis delta virus genotype III RNA editing**. *RNA* 2006, **12**:1521-1533.
This paper describes the computational and experimental analysis of hepatitis delta virus genotype III RNA using MPGAfold and native PAGE. The computational and experimental results indicate that the RNA has a tendency to form two functional conformations, one that is conducive to editing and the other that is conducive to replication.

36. •• Shapiro BA, Kasprzak W, Grunewald C, Aman J: **Graphical exploratory data analysis of RNA secondary structure dynamics predicted by the massively parallel genetic algorithm**. *J Mol Graph Model* 2006, **25**:514-531.
This paper presents, step by step, an interactive analysis process for the solution spaces produced by MPGAfold (or any DPA program) using the tools available in the StructureLab analysis program. Examples are based on the structures of two different states of the 5′-untranslated region of HIV-1.

37. Gee AH, Kasprzak W, Shapiro BA: **Structural differentiation of the HIV-1 polyA signals**. *J Biomol Struct Dyn* 2006, **23**:417-428.

38. Tortorici MA, Shapiro BA, Patton JT: **A base-specific recognition signal in the 5′ consensus sequence of rotavirus plus-strand RNAs promotes replication of the double-stranded RNA genome segments**. *RNA* 2006, **12**:133-146.

39. Zhang J, Zhang G, Guo R, Shapiro BA, Simon AE: **A pseudoknot in a preactive form of a viral RNA is part of a structural switch activating minus-strand synthesis**. *J Virol* 2006, **80**:9181-9191.

40. Kasprzak W, Shapiro B: **Stem Trace: an interactive visual tool for comparative RNA structure analysis**. *Bioinformatics* 1999, **15**:16-31.

41. Rivas E, Eddy SR: **A dynamic programming algorithm for RNA structure prediction including pseudoknots**. *J Mol Biol* 1999, **285**:2053-2068.

42. Dirks RM, Pierce NA: **A partition function algorithm for nucleic acid secondary structure including pseudoknots**. *J Comput Chem* 2003, **24**:1664-1677.

43. Reeder J, Giegerich R: **Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics**. *BMC Bioinformatics* 2004, **5**:104-115.

44. Ruan J, Stormo GD, Zhang W: **ILM: a web server for predicting RNA secondary structures with pseudoknots**. *Nucleic Acids Res* 2004, **32**:W146-W149.

45. • Ruan J, Stormo GD, Zhang W: **An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots**. *Bioinformatics* 2004, **20**:58-66.
This paper presents the ILM algorithm, which employs heuristics to find secondary structures with pseudoknots. It can use thermodynamic and/or comparative information, which makes it suitable for application to single and multiple aligned sequences.

46. Ren J, Rastegari B, Condon A, Hoos HH: **HotKnots: heuristic prediction of RNA secondary structures including pseudoknots**. *RNA* 2005, **11**:1494-1504.

47. van Batenburg FH, Gultyaev AP, Pleij CW: **PseudoBase: structural information on RNA pseudoknots**. *Nucleic Acids Res* 2001, **29**:194-195.

48. van Batenburg FH, Gultyaev AP, Pleij CW, Ng J, Oliehoek J: **PseudoBase: a database with RNA pseudoknots**. *Nucleic Acids Res* 2000, **28**:201-204.

49. Tabaska JE, Cary RB, Gabow HN, Stormo GD: **An RNA folding method capable of identifying pseudoknots and base triples**. *Bioinformatics* 1998, **14**:691-699.

50. Witwer C, Hofacker IL, Stadler PF: **Prediction of consensus RNA**
• **secondary structures including pseudoknots**. *IEEE/ACM Trans Comput Biol Bioinform* 2004, **1**:66-77.
This paper introduces the HXMATCH program. In this approach, a matrix of base pair scores (consisting of a covariation and a thermodynamic term) is used as input to an MWM algorithm. The result is post-processed to obtain bisecondary structures (structures that are a superposition of not more than two pseudoknot-free structures). The method is successfully applied to several pseudoknot-containing RNA structures.

51. Bindewald E, Schneider TD, Shapiro BA: **CorreLogo: an online**
• **server for 3D sequence logos of RNA and DNA alignments**. *Nucleic Acids Res* 2006, **34**:W405-W411.
This paper describes a web server that uses mutual information derived from a sequence alignment to generate a 3D sequence logo representing correlated bases and other useful properties. It also introduces a novel small sample correction methodology that enables the determination of statistically significant correlations.

52. Lindgreen S, Gardner PP, Krogh A: **Measuring covariation in RNA alignments: physical realism improves information measures**. *Bioinformatics* 2006, **22**:2988-2995.

53. Nussinov R, Piecznik G, Griggs JR, Kleitman DJ: **Algorithms for loop matching**. *SIAM J Appl Math* 1978, **35**:68-82.

54. Freyhult E, Moulton V, Gardner P: **Predicting RNA structure**
• **using mutual information**. *Appl Bioinformatics* 2005, **4**:53-59.
This paper presents the program MIfold, an RNA secondary structure prediction program based on mutual information. It is shown that MIfold can predict simple pseudoknots. For a set of sequences, MIfold turns out to be more sensitive but less selective than RNAalifold.

55. Wang R, Alexander RW, VanLoock M, Vladimirov S, Bukhtiyarov Y, Harvey SC, Cooperman BS: **Three-dimensional placement of the conserved 530 loop of 16 S rRNA and of its neighboring components in the 30 S subunit**. *J Mol Biol* 1999, **286**:521-540.

56. Macke T, Case D: **Modeling unusual nucleic acid structures**. In *Molecular Modeling of Nucleic Acids.* Edited by Leontes N, SantaLucia JJ. Washington, DC: American Chemical Society; 1998:379-393.

57. Zwieb C, Muller F: **Three-dimensional comparative modeling of RNA**. *Nucleic Acids Symp Ser* 1997, **36**:69-71.

58. Massire C, Westhof E: **MANIP: an interactive tool for modelling RNA**. *J Mol Graph Model* 1998, **16**:197-205, 255–257.

59. Jossinet F, Westhof E: **Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure**. *Bioinformatics* 2005, **21**:3320-3321.

60. Major F: **Building three-dimensional ribonucleic acid structures**. *Comput Sci Eng* 2003, **5**:44-53.

61. Yingling YG, Shapiro BA: **The prediction of the wild-type**
•• **telomerase RNA pseudoknot structure and the pivotal role of the bulge in its formation**. *J Mol Graph Model* 2006, **25**:261-274.
This paper describes the prediction of the 3D structure of the wild-type telomerase pseudoknot domain. RNA2D3D was used for the prediction, which was followed by 100 ns of implicit and explicit solvent molecular dynamics simulations. The result predicts novel tertiary interactions, possibly explaining the function.

62. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank. A computer-based archival file for macromolecular structures**. *Eur J Biochem* 1977, **80**:319-324.

63. Tamura M, Hendrix DK, Klosterman PS, Schimmelman NR, Brenner SE, Holbrook SR: **SCOR: Structural Classification of RNA, version 2.0**. *Nucleic Acids Res* 2004, **32**:D182-D184.

64. Murthy VL, Rose GD: **RNABase: an annotated database of RNA structures**. *Nucleic Acids Res* 2003, **31**:502-504.

65. Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q, Fox GE: **NCIR: a database of non-canonical interactions in known RNA structures**. *Nucleic Acids Res* 2002, **30**:395-397.

66. Burks J, Zwieb C, Muller F, Wower I, Wower J: **Comparative 3-D**
•• **modeling of tmRNA**. *BMC Mol Biol* 2005, **6**:14-31.
This paper describes a procedure for building 3D models of large tmRNAs from *E. coli*, *B. anthracis* and *C. crescentus* using the software ERNA-3D. *E. coli* tmRNA is 362 bases long and includes four pseudoknots.

67. Massire C, Jaeger L, Westhof E: **Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis**. *J Mol Biol* 1998, **279**:773-793.

68. Tsai HY, Masquida B, Biswas R, Westhof E, Gopalan V: **Molecular**
•• **modeling of the three-dimensional structure of the bacterial RNase P holoenzyme**. *J Mol Biol* 2003, **325**:661-675.
This paper discusses 3D models of large RNase RNAs from *E. coli* and *B. subtilis*. The models were interactively built from the bacterial RNase P holoenzyme either uncomplexed or bound to a ptRNA (precursor tRNA) using the software MANIP.

69. Lambert D, Heckman JE, Burke JM: **Cation-specific structural**
• **accommodation within a catalytic RNA**. *Biochemistry* 2006, **45**:829-838.
This paper describes the recent use of MC-Sym to propose a model of the $[Co(NH_3)_6]^{3+}$-mediated catalytic core of the hairpin ribozyme.

70. Yingling YG, Shapiro BA: **The impact of dyskeratosis congenita mutations on the structure and dynamics of the human telomerase RNA pseudoknot domain**. *J Biomol Struct Dyn* 2007, **24**:303-320.

71. Barreda DCJE, Shigenobu Y, Ichiishi E, Del Carpio MC: **RNA 3D structure prediction: (1) assessing RNA 3D structure similarity from 2D structure similarity**. *Genome Inform* 2004, **15**:112-120.

72. Chen JL, Blasco MA, Greider CW: **Secondary structure of vertebrate telomerase RNA**. *Cell* 2000, **100**:503-514.