Problems on RNA Secondary Structure Prediction and Design ⋆

Anne Condon

The Department of Computer Science
2366 Main Mall
University of British Columbia
Vancouver, B.C. V6R 2C8
condon@cs.ubc.ca

Abstract. We describe several computational problems on prediction and design of RNA molecules.

1   Introduction

Almost a decade ago, I ventured two blocks from my Computer Sciences department to a very unfamiliar world - the Chemistry Department. This short walk was the start of a rewarding ongoing journey. Along the way, I have made wonderful new friends - both the real sort and the technical sort that like to make their home in the heads of us theoreticians, there to remain indefinitely. In this article, I will describe some of the the latter.

The subjects are nucleic acids: DNA and RNA. From a *biological perspective*, the role of double-helical DNA in storing genetic information is well known. The central dogma of molecular biology posits that in living cells, this genetic information is translated into proteins, which do the real work. The traditional view of RNA is as a helper molecule in the translation process. That view has changed in recent years, with RNA getting star billing in regulation of genes and as a catalyst in many cellular processes [9]. Attention on RNA stems also from the many diseases caused by RNA viruses. Accordingly, significant effort is now expended in understanding the function of RNA molecules. The structure of RNA molecules is key to their function, and so algorithms for prediction of RNA structure are of great value.

While the biological roles of DNA and RNA molecules are clearly of great importance, they are only part of the story. From an *engineering perspective*, DNA and RNA molecules turn out to be quite versatile, capable of functions not seen in nature. These molecules can be synthesized and used as molecular bar-codes in libraries of polymers [24] and as probes on DNA chips for analysis

of gene expression data. RNA's with new regulatory properties are designed, with hopes of applications in therapeutics [25]. Tiny instances of combinatorial problems have been solved in a wet-lab, using DNA or RNA to represent a pool of solutions to a problem instance [4]. Novel topological and rigid three-dimensional structures have been built from DNA [22, 30], and a theory of programmable self-assembly of such structures is emerging [20]. Scientists are working to create catalytic RNA molecules that support the so-called "RNA world hypothesis": prior to our protein-dominated world, RNA molecules functioned as a complete biological system capable of the basic processes of life [26]. Naturally, advances in these areas also rely greatly on good understanding of function, and hence structure, of RNA and DNA molecules.

The problems described in this article are motivated more by the engineering, rather than the biological perspective of the potential roles of DNA and RNA. Even for the problem of predicting RNA structure, the two different perspectives suggest somewhat different approaches. In the biological setting, it is often possible to get sequences of homologous (i.e. evolutionarily and functionally related) molecules from several organisms. In this case, a comparative approach that uses clues about common structure from all molecules in the set are the most successful in structure prediction. However, in the engineering setting, this approach is typically not applicable. Moreover, the inverse to the prediction problem, namely design of a DNA or RNA molecule that has a particular structure, is of central importance when engineering novel molecules.

We focus on problems relating to RNA and DNA *secondary structure*, which we describe in Section 2. In Section 3, we describe problems on predicting the secondary structure of a given DNA or RNA molecule. Section 4 considers more general problems when the input is a set of molecules. Finally, in Section 5, we describe problems on the design of DNA and RNA molecules that fold to a given input secondary structure.

2 Basics on RNA secondary structure

To keep things simple, consider an RNA molecule to be a strand of four types of bases, with two chemically distinct ends, known as the $5'$ and $3'$ ends. In RNA the base types are Adenine (A), Cytosine (C), Guanine (G), and Uracil (U). DNA also has four types of bases, including A, C, G and replacing Uracil (U) with Thymine (T). We represent an RNA (DNA) molecule as a string over $\{A, C, G, U\}$ ($\{A, C, G, T\}$), with the left end corresponding to the $5'$ end of the molecule.

In a process called hybridization, pairs of bases in RNA and DNA form hydrogen bonds, with the complementary pairs C-G and A-U (or A-T in the case of DNA) being the strongest and others, particularly the "wobble" pair G-U also playing a role [29]. A folded molecule is largely held together by the resulting set of bonds. called its *secondary structure*. Knowledge of the secondary structure of a folded RNA molecule sheds valuable insight on its function [27]. We note that while the DNA that stores genetic information in living organisms

is formed from two complementary strands, single-stranded DNA folds and forms structures according to the same basic principles as does a single stand of RNA.

Figure 1 depicts the secondary structure of two DNA molecules. In the graphical depictions (top), dots indicate base pairs, and "stems" of paired bases and "loops" of unpaired bases can be identified. The graphical depictions do not convey the three-dimensional structure of the molecules. For example, stems twist to form double helices familiar in illustrations of DNA, and angles at which stems emanate from loops cannot be inferred from the diagrams. In the arc depiction (bottom), arcs connect paired bases. In the left structure, arcs are hierarchically nested, indicating that this is a *pseudoknot free* structure. In contrast, arcs cross in the arc depiction of the structure on the right, indicating that it is *pseudoknotted*.
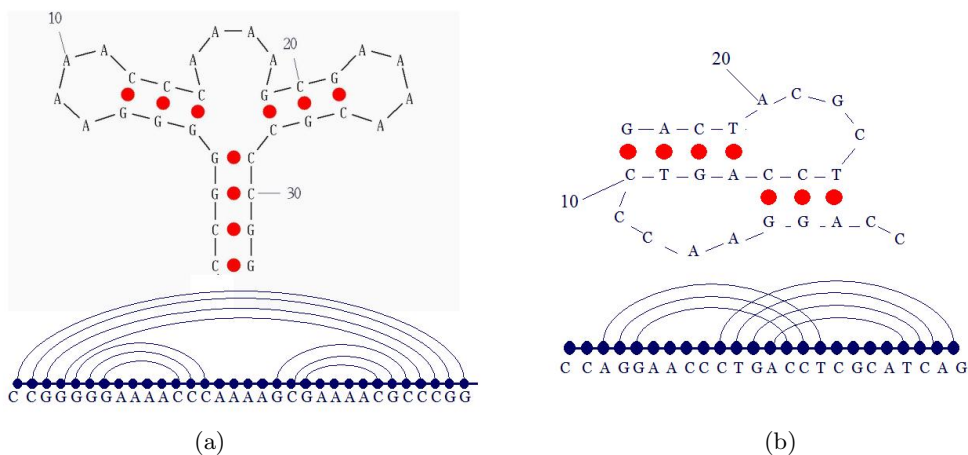

(a)   (b)

Fig. 1. (a) Pseudoknot free secondary structure. This structure contains 10 base pairs and three loops, two of which are hairpin loops (having one emanating stem) and one of which is a multi-loop (having three emanating stems). The numbers refer to base indices, in multiples of 10, starting at the $5'$ end (leftmost base in arc depiction). The substructure from index 19 to index 28 contains a stem with two stacked pairs, namely (G-C,C-G) and (C-G,G-C), and a hairpin loop with four unpaired bases (all A's) and closing base pair G-C. In set notation, this substructure is $\{(19, 28), (20, 27), (21, 26)\}$. The free energy contributions of the two stacked pairs and hairpin loop are -3.4 kcal/mol, -2.4 kcal/mol, and 4.5 kcal/mol, respectively, so the total free energy of the substructure from index 19 to 28 is -1.3 kcal/mol. (b) Pseudoknotted secondary structure.

Abstractly, we represent the secondary structure of a DNA or RNA molecule of length (i.e. number of bases) n as a set S of integer pairs $\{(i, j) \mid 1 \leq i < j \leq n\}$, where each i is contained in at most one pair of S. The pair (i, j) indicates

a bond between the bases at positions i and j of the corresponding strand. The secondary structure is pseudoknot free if and only if for all pairs (i, j) and (i', j'), it is not the case that $i' < i < j' < j$.

The thermodynamic model for RNA structure formation posits that, out of the exponentially many possibilities, an RNA molecule folds into that structure with the *minimum free energy* (mfe). Free energy models typically assume that the total free energy of a given secondary structure for a molecule is the sum of independent contributions of adjacent, or stacked, base pairs in stems (which tend to stabilize the structure) and of loops (which tend to destabilize the structure). These contributions depend on temperature, the concentration of the molecule in solution, and the ionic concentration of the solution. *Standard models* additionally assume that the free energy contribution of a loop depends only on (i) the bases closing the stem and those unpaired bases in the loop adjacent to the stem, for each stem, (ii) the number of stems emanating from the loop, and (iii) the number of unpaired bases between consecutive stems. For loops with more than two stems, (ii) and (iii) are further simplified to be of the form $a + bs + cu$, where b, c are constants, s is the number of stems emanating from the loop, and u is the total number of unpaired bases in the loop.

Significant effort has been expended to determine many of these energy contributions experimentally [21, 23]. Other contributions are estimated based on extrapolations from known data or existing databases of naturally occurring structures [17]. More sophisticated models also associate energy contributions with coaxially stacked pairs and other structural features, but we will ignore these here for the sake of simplicity.

3   RNA secondary structure prediction

"If 10% of protein fold researchers switched to RNA, the problem could be solved in one or two years." - I. Tinoco Jr. and C. Bustamente

The best known algorithms for predicting the secondary structure of a single input RNA or DNA molecule work by finding the minimum free energy (mfe) secondary structure of the given input RNA molecule, with respect to a given standard thermodynamic model. Lyngsø and Pedersen [15] have shown that the task is NP-hard. However, the problem is not as intractable as this might suggest, because in practice the range of structures into which a molecule will actually fold is somewhat limited.

Zuker and Steigler [32] describe a dynamic programming algorithm for finding the mfe pseudoknot free secondary structure of a given molecule. (In practice, the algorithm can be used to gain insight on secondary structure even for molecules with pseudoknotted structures, because there is some evidence that molecules fold to form a pseudoknot free secondary structure first, and pseudoknotted features are added only at the end of the folding process.) Conceptually the algorithm is quite simple, exploiting the following fact. Let the input strand be $b_1 b_2 \ldots b_n$. Suppose that $W(i, j)$ is the energy of the mfe pseudoknot free

secondary structure for strand $b_i \ldots b_j$, and $V(i,j)$ be the energy of the mfe pseudoknot free secondary structure for strand $b_i \ldots b_j$, among those structures containing base pair (i,j). Then, W satisfies the following recurrence (base cases excluded):

$$W(i,j) = \min[V(i,j), \min_{k:i<k<j}\{W(i,k) + W(k+1,j)\}].$$

$V(i,j)$ also satisfies a recurrence that is expressed in terms of the different types of loops (omitted here). A refinement of the original Zuker-Steigler algorithm, due to Lyngsø et al. [16], has running time $O(n^3)$. We note that the algorithm exploits the simplified loop energy contributions of the standard thermodynamic model mentioned earlier. Implementations of this algorithm are available on the world wide web as part of the mfold [17] and the Vienna [13] packages.

Mathews et al. [17] report that on a large data set of RNA molecules of length up to 700, the algorithm reports 73% of known base pairs. On longer molecules, the prediction accuracy is poorer. Thus, there is certainly room for improvement in the current mfe approach to secondary structure prediction. Perhaps the most important problem listed in this article is to find algorithms for pseudoknot free secondary structure prediction that have improved accuracy. We expect that significant progress will only come through a greater understanding of the underlying biological forces that determine folding, perhaps by refining the currently used thermodynamic model or by considering the folding pathway of molecules. In light of this and the subtle interplays between algorithmic and modeling considerations, we believe that the best progress can be made only through productive collaborations between algorithm designers and experts on nucleic acids.

So far, we have focused on the problem of finding the mfe secondary structure (with respect to some thermodynamic model) of a DNA or RNA molecule. Other information on the stability of the molecule's structure can also be very useful. A better view is that each possible secondary structure S for molecule M occurs with a probability that is proportional to $e^{-\Delta G(S)/RT}$ where $\Delta G(S)$ is the free energy associated with structure S, R is the Boltzmann constant, and T is temperature. Associated with each possible base pair of the molecule is a weight, defined to be the sum of the probabilities of the structures in which it occurs. McCaskill [18] gave an $O(n^3)$ dynamic for calculating the set of base pair weights of a molecule. This algorithm is incorporated into standard folding packages [17, 13], significantly enhancing their utility. Another useful enhancement to the Zuker-Steigler algorithm outputs not just the mfe structure, but all structures with energy below a user-supplied threshold [31, 33].

From a purely algorithmic standpoint, the problem of predicting RNA and DNA secondary structure becomes more interesting when one considers pseudoknotted structures. The thermodynamic model for pseudoknot free secondary structures has been extended to include contributions of pseudoknotted stems and loops. Several algorithms have been proposed for predicting the mfe secondary structure from a class of secondary structures that allows limited types of pseudoknots [1, 15, 19, 28]. Other algorithms are heuristic in nature, such as the

genetic algorithm of Gultyaev et al. [12]. The dynamic programming algorithm of Rivas and Eddy [19] is the most general in terms of the class of structures handled. The authors claim that all known natural structures can be handled by the algorithm, although they do not provide evidence for this claim. However, the authors state that "we lack a systematic *a priori* characterization of the class of configurations that this algorithm can solve". Another limitation of the algorithm is its high running time of $\Theta(n^6)$. An algorithm of Akutsu [1] runs in $O(n^4)$ time and $O(n^2)$ space, but there are natural pseudoknotted structures that cannot be handled by this algorithm.

An interesting goal for further research is to precisely classify pseudoknotted structures, refining the current partition into pseudoknot free and pseudoknotted structures. As a first step in this direction, we have developed a characterization of the class of secondary structures that can be handled by the Rivas and Eddy algorithm. Roughly, a secondary structure can be handled by that algorithm if and only if in the arc depiction of that structure (see Figure 1), all arcs can be reduced to one arc by repeatedly applying a *collapse* operation. In a collapse operation, two arcs can be replaced by one arc if one can colour at most two line segments along the baseline of the depiction, and touch all four end points of the two arcs but no other arc. (We note that a natural approach to classification of secondary structures, which does not seem to be particularly fruitful, is to consider the crossing number of the arc depiction of the secondary structure.)

With a good classification of secondary structures in hand, one can then hope to clarify the trade-offs between the class of structures that can be handled, and the time or space requirements of algorithms for predicting mfe pseudoknotted structures. Perhaps the classification would provide a hierarchy of structure classes, parameterized by some measure k, and a fixed-parameter tractability result for this classification is possible, as in the work of Downey et al. [10].

It would be very useful to calculate the partition function for pseudoknotted structures. An extension of the Rivas and Eddy algorithm along the lines of McCaskill [18] should be possible, but would be computationally expensive and limited by the range of structures handled by the Rivas and Eddy algorithm. It may be possible to approximate the partition function via the Markov chain monte carlo method of Jerrum and Sinclair [14].

Finally, we note that secondary structures can also form between two or more RNA or DNA molecules in solution, so a natural generalization of the problem discussed so far is to predict the mfe secondary structure formed by two or more input molecules. Conceptually, the thermodynamic model for a secondary structure formed from multiple strands is very similar to that for a single strand, but an initiation penalty is added to the total free energy. An algorithm for predicting the secondary structure of a pair of molecules is publically available [2]. Some interesting algorithmic questions arise in design of algorithms for handling multiple strands. For example, what does it mean for a structure with multiple strands to be pseudoknot free?

4 Prediction for combinatorial sets of strands

The problems in this section are motivated by the use of *combinatorial sets* of strands in various contexts. In the first context, described by Brenner et al. [7], the goal is to sequence millions of short DNA fragments (these fragments could be in a gene expression sample). DNA sequencing machines handle one sequence at a time, and it would be infeasible to separate out the millions of short fragments and sequence each separately. Instead, Brenner described an ingenious "biomolecular algorithm" to sequence the molecules in a massively parallel fashion. One step of this algorithm attaches a unique DNA "tag" molecule to each of the DNA fragments. The tags are used to help to organize the DNA fragments in further steps of the algorithm. Let

$$S = \{\text{TTAC, AATC, TACT, ATCA, ACAT, TCTA, CTTT, CAAA}\}. \quad (1)$$

The tags constructed by Brenner et al. [8] are all of the 8^8 strands in the combinatorial set S^8. The strands in S were carefully designed so that each contains no G's, exactly one C, and differs from the other strands of S in three of the four bases. The reason for this design is to ensure that the tags do not fold on themselves (that is, have no secondary structure), in which case they would not be useful as tag molecules in the sequencing scheme.

The set S of tags given in (1) above is an example of a *complete combinatorial set*, defined as a set of strings (strands) in $S(1) \times S(2) \ldots \times S(t)$, where for each $i, 1 \leq i \leq t$, $S(i)$ is a set of strings, all having the same length l_i. The l_i are not required to be equal. Complete combinatorial sets are also used to represent solution spaces in biocomputation that find a satisfying assignment to an instance of the Satisfiability problem [6, 11]. Again, for this use, all strands in the complete combinatorial sets should form no secondary structure.

These applications motivate the **structure freeness problem for combinatorial sets**: given the description of a complete combinatorial set S, determine whether all of the 2^t strands in S are structure free. Here, we consider a strand to be structure free if its mfe pseudoknot free secondary structure is the empty set. We limit our definition to pseudoknot free secondary structures here because in the case of predicting the mfe secondary structure of a single molecule, the pseudoknot free case is already well understood, as discussed in the last section of this article.

Given sets of strings $S(1), S(2), \ldots, S(t)$, one can test that all strands in $S = S(1) \times S(2) \ldots \times S(t)$ are structure free by running the Zuker-Steigler algorithm on each strand of S. This would take time proportional to $|S|n^3$, where $n = l_1 + l_2 + \ldots + l_t$ is the total length of strands in S. In general, this running time is exponential in the input size. Andronescu et al. [3] describe a simple generalization of the Zuker-Steigler algorithm, which has running time $O(max_i|S(i)|^2n^3)$.

The algorithm of Andronescu et al. handles only complete combinatorial sets. More general combinatorial sets can be defined via an acyclic graph G with a special start node and end node. Suppose that each node i in the graph is labeled

with a set of strands S_i. Then, each path $n_1, n_2, \ldots, n_t$ in the graph from the start node to the end node corresponds to the set of strands $S(n_1) \times S(n_2) \ldots \times S(n_t)$. The combinatorial set of strands $S(G)$ associated with the graph is the union of the set of strands for each path of G from the start node to the end node. (Since G is acyclic, there are a finite number of such paths). Such a combinatorial set of strands was used by Adleman [4] in his biomolecular computation for a small instance of the Hamiltonian Path problem. It is open whether there is an efficient algorithm to test if all strands $S(G)$ are structure free, where the input is the graph G and the set $S(i)$ of strands for each node i of G. The case where all strands in $S(i)$ have the same length, for any node i of G, is also open. By adding cycles to G, the problem becomes even more general, and its complexity remains open even for the simplest case that the nodes and edges of G form a simple cycle.

5   Secondary structure design

"... rather than examining in detail what occurs in nature (biological organisms), we take the engineering approach of asking, what can we build?" - Erik Winfree.

The simplest version of the RNA design problem is as follows: given a secondary structure S (that is, set of desired base pairings), design a strand whose mfe secondary structure is S, according to the standard thermodynamic model. There has been relatively little previous theoretical work on algorithms for design of DNA or RNA molecules that have certain structural properties. Indeed, it is open whether the problem is NP-hard, although we conjecture that this is the case. Even if the range of secondary structures is restricted to be the pseudoknot free secondary structures, the complexity of the problem is open.

However, as with RNA secondary structure prediction, we expect that the range of structures one may wish to design in practice will be somewhat limited. Thus, it would certainly be useful to provide characterizations of secondary structure classes for which the design problem is efficiently solvable. More useful versions of the RNA design problem may pose additional requirements, perhaps on the stability of the mfe structure or on the base composition of the RNA molecule.

A generalization of the RNA secondary structure design problem above arises when the desired structure is composed of more than one strand. Many of the applications of RNA secondary structure design that we are familiar with involve multiple strands. For example, Seeman has designed several multi-strand structural motifs, and has developed an interactive software tool to help design the component strands [22]. Winfree et al. [30] proposed a method for self-assembly of DNA "tile" molecules in a programmable fashion, and has shown that programmable self-assembly is in principle capable of universal computation. The component tile molecules used in these self-assembly processes involve four component strands, which form a rigid two-dimensional structure with protruding

short single strands, called sticky ends, that are available for hybridization with the sticky ends of other tile molecules. RNA molecules are designed as molecular switches, biosensors, and even for therapeutic uses. For example, it is possible to inhibit the action of certain pathogenic RNA molecules (such as viruses) using carefully-designed short RNA molecules, called *trans*-cleaving ribozymes, that can bind to the pathogenic RNA and cleave it [25]. The *trans*-cleaving ribozymes are currently developed via *in-vitro* evolution, in which a large library of RNA molecules is screened to select for those that exhibit some tendency towards the desired function and the screened molecules are then randomly mutated, in order to diversify the pool. The screening and diversification steps are repeated until a molecule with the desired function is obtained. Computational methods for design of RNA molecules could help provide good starting points for *in-vitro* evolution processes. As with the RNA secondary structure design problem for a single strand, while ad-hoc techniques are in use by researchers in Chemistry, there is little theoretical knowledge of good algorithmic design principles.

Finally, a design problem that has received significant attention is that of designing combinatorial sets of molecules that have no secondary structure. This is the inverse of the prediction problem mentioned in Section 4. BenDor et al. [5] describe a combinatorial design scheme with provably good properties that address one version of this problem. Other approaches, such as the simple design of Brenner described in Section 4, construct strands in the component sets $(S(i))$ of the combinatorial sets to be over a three-letter alphabet and have certain coding-theoretic properties. In light of the wide uses of these designs, further insights as to good design strategies would be useful.

6   Conclusions

This article has described several problems of a combinatorial flavour relating to RNA secondary structure prediction and design. These problems are motivated by work in design of RNA and DNA strands for diverse applications with both biological and computational motivations. The prediction and design problems are inter-related, with good algorithms for prediction being a prerequisite to tackling the secondary structure design problems. In light of the importance of these problems in both the biological and engineering settings, and the relatively little attention they have received to date from the computer science community, they represent a fruitful direction for algorithms research.

Inevitably, the problems reflect my own interests and biases. Many other theoretically interesting problems, motivated by three-dimensional RNA structure prediction, visualization of secondary structures, and more are not covered here, but raise interesting questions in computational geometry and graph drawing.

7   Acknowledgements

References

1. T. Akutsu, "Dynamic programming algorithms for RNA secondary prediction with pseudoknots", Discrete Applied Mathematics, 104, 2000, 45–62.
2. M. Andronescu, R. Aquirrez-Hernandez, H. Hoos, and A. Condon, "RNAsoft: a suite of RNA secondary structure prediction and design software tools", Nucleic Acids Research, In press.
3. M. Andronescu, D. Dees, L. Slaybaugh, Y. Zhao, A. Condon, B. Cohen, and S. Skiena, "Algorithms for testing that sets of DNA words concatenate without secondary structure", Proc. Eighth International Workshop on DNA Based Computers, Hokkaido, Japan, June 2002. To appear in LNCS.
4. L.M. Adleman, "Molecular computation of solutions to combinatorial problems," Science, Vol 266, 1994, 1021–1024.
5. A. Ben-Dor, R. Karp, B. Schwikowski, and Z. Yakhini, "Universal DNA tag systems: a combinatorial design scheme," Proc. Fourth Annual International Conference on Computational Molecular Biology (RECOMB) 2000, ACM, 65–75.
6. Braich, R. S., Chelyapov, N., Johnson, C., Rothemund, P. W.K. and Adleman, L. "Solution of a 20-variable 3-SAT Problem on a DNA computer", Science 296, 2002, 499–502.
7. S. Brenner, M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M, Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S.R. Williams, K. Moon, T. Burcham, M. Pallas, R.B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran, "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays," Nature Biotechnology, 18, 2000, 630–634.
8. S. Brenner, "Methods for sorting polynucleotides using oligonucleotide tags," U.S. Patent Number 5,604,097, 1997.
9. C. Dennis, "The brave new world of RNA", Nature, 418, 11 2002, 122–124.
10. R. G. Downey and M. R. Fellows, "Fixed-Parameter Tractability and Completeness I: Basic Results", SIAM J. Comput. 24(4), 1995, 873–921.
11. D. Faulhammer, A.R. Cukras, R.J. Lipton, and L. F. Landweber, "Molecular computation: RNA solutions to chess problems," Proc. Natl. Acad. Sci. USA, 97, 2000, 1385–1389.
12. P.Gultyaev, F.H.D.van Batenburg, and C.W.A.Pleij, "The computer simulation of RNA folding pathways using a genetic algorithm", J. Mol. Biol., 250, 1995, 37–51.
13. I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast Folding and Comparison of RNA Secondary Structures", Monatsh.Chem. 125, 1994, 167–188.
14. M. Jerrum and A. Sinclair, "Approximating the permanent", SIAM Journal on Computing 18, 1989, 1149–1178.
15. R.B. Lyngsø and C.N.S. Pedersen, "Pseudoknot prediction in energy based models", Journal of Computational Biology 7(3), 2000, 409–427.
16. R. B. Lyngso, M. Zuker, and C. N. S. Pedersen, "Internal Loops in RNA Secondary Structure Prediction", Proc. Third International Conference in Computational Molecular Biology (RECOMB), April 1999, 260–267.

17. D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure", J. Molecular Biology, 288, 1999, 911–940.

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

19. E. Rivas and S. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots," Journal of Molecular Biology, 285, 1999, 2053–2068.

20. P.W.K. Rothemund and E. Winfree, "The program-size complexity of self-assembled squares", Symposium on Theory of Computing, 2000.

21. J. SantaLucia, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics", Proc. Natl Acad Sci USA 95:4, 1998, 1460–1465.

22. N.C. Seeman, "De novo design of sequences for nucleic acid structural engineering," Journal of Biomolecular Structure and Dynamics, 8:3, 1990, 573–581.

23. M.J. Serra, D.H. Turner, and S.M. Freier, "Predicting thermodynamic properties of RNA", Meth. Enzymol., 259, 1995, 243–261.

24. D. D. Shoemaker, D. A. Lashkari, D. Morris, M. Mittman, and R. W. Davis, "Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy," Nature Genetics, 16, 1996, 450–456.

25. B.A. Sullenger and E. Gilboa. "Emerging clinical applications of RNA", Nature, 418, 2002, 252–258.

26. J.W. Szostak, D.P. Bartel, and L. Luisi. "Synthesizing life", Nature 409, 2001, 387-389.

27. I. Tinoco Jr. and C. Bustamante, "How RNA folds", J. Mol. Biol. 293, 1999, 271–281.

28. Y. Uemura, A. Hasegawa, Y. Kobayashi, and T. Yokomori, "Tree adjoining grammars for RNA structure prediction", Theoretical Computer Science, 210, 1999, 277–303.

29. E. Westhof and V. Fritsch, "RNA folding: beyond Watson-Crick pairs", Structure 2000, 8:R55-R65, 2000.

30. E. Winfree, F. Liu, L. Wenzler, and N. Seeman, "Design and self-assembly of 2D DNA crystals," Nature, 394, 1998, 539–544.

31. S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster, "Complete Suboptimal Folding of RNA and the Stability of Secondary Structures", Biopolymers, 1998, Vol. 49, 145–165.

32. M. Zuker and P. Steigler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," Nucleic Acids Res 9, 1981, 133–148.

33. M. Zuker, "On Finding All Suboptimal Foldings of an RNA Molecule", Science, 244, 1989, 48–52.