Theoretical Computer Science 410 (2009) 294–301

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Computational prediction of nucleic acid secondary structure: Methods, applications, and challenges

Anne Condon, Hosna Jabbari *

Department of Computer Science, University of British Columbia, Canada

ARTICLE INFO

Keywords:
RNA
Secondary structure prediction
Pseudoknot
Interacting RNAs
RNAi

ABSTRACT

RNA molecules are crucial in different levels of cellular function, ranging from translation and regulating genes to coding for proteins. Additionally, nucleic acids (RNA and DNA molecules) are designed for novel applications in biotechnology. Understanding the structure of a molecule is important in inferring its function, and computational methods for structure prediction have captured the interest of many researchers.

Some functions of RNA molecules in cells, such as gene regulation, result from the binding of one RNA molecule to another, so-called target RNA molecule. This has led to recent interest in prediction of the secondary structure formed from interacting molecules. In this paper, we provide a brief overview of methods, applications, and challenges in computational prediction of nucleic acid secondary structure, both for single strands and for interacting strands.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

In this article, we review algorithms for predicting the secondary structure of nucleic acids – RNA and DNA molecules – from their base sequence. The challenge of nucleic acid structure prediction presents interesting combinatorial problems, solutions to which are of real practical value in genomics and biotechnology.

The varied natural cellular roles, as well as designed uses, of nucleic acids arise from the diverse structural forms that these molecules can assume. Very familiar is the double helix, formed by base pairing of a DNA strand with its complement: A pairs with T, and C pairs with G. Unlike DNA, RNA strands in the cell typically are not paired with a complementary partner; yet the bases of an RNA strand also energetically favor pairing with complementary bases. As a result, a long single strand of RNA tends to fold on itself, in a way that causes pairing of complementary bases within the strand. The resulting paired complementary regions form helical stems; these stems may form branched or even more complex (pseudoknotted) patterns, as explained in Section 2.

With several genomes now sequenced, there is growing recognition of the degree to which RNA molecules are essential to the operation of the cell. The central dogma of molecular biology is that genes are transcribed into RNA molecules called messenger RNA's (mRNA's), which are then translated into proteins. However, many RNA molecules which are obtained via transcription from genes are *not* intermediates in the process of creating proteins from DNA. Rather, these so-called non-coding RNA's (ncRNA's) perform cellular functions in their own right, largely derived from their structures. As Couzin stated, "RNA, long upstaged by its more glamorous sibling, is turning out to have star qualities of its own" [8]. Some cellular roles of RNA's include the following.

- RNA molecules catalyze cellular processes, such as splicing of RNA strands; this is one mechanism that cells use to degrade foreign matter such as RNA viruses [15].
- Naturally-occurring short RNA molecules can help regulate the expression of genes. One mechanism for achieving this is RNA interference, whereby a short interfering RNA molecule (siRNA) binds to a perfect complement on the target

* Corresponding author.
E-mail addresses: condon@cs.ubc.ca (A. Condon), hjabbari@cs.ubc.ca (H. Jabbari).

0304-3975/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2008.09.042

strand. In this way, the siRNA guides an attached protein complex, called the RISC (RNA-induced silencing complex) to the target, and the target is cleaved by the RISC. Other RNA's, called microRNA's, inhibit translation of their target mRNA's by (imperfectly) binding to the target so as to prevent the cell's translation machinery from accessing the target.

- Yet another mechanism for gene regulation is the riboswitch. The structure of a riboswitch can change, upon sensing the presence of a particular small molecule; if the riboswitch is embedded in an mRNA then the structural change can inhibit or enhance translation efficiency for that mRNA [40,41].
- Designed siRNA's can be synthesized in the lab and used to "knock down" the regulation of certain genes in cell cultures [16,17]. This mechanism is extremely useful in genome studies, since it can help scientists infer the functions of certain genes and their products.
- Mechanisms of RNA interference are also being explored as a treatment for HIV and viral hepatitis [17].

Apart from naturally-occurring RNA's, other RNA molecules have been designed which have novel catalytic functions not found in nature [20]. One motivation for design of such molecules is to lend support to the hypothesis that a primitive form of life – the RNA world – preceded the modern DNA-protein world.

Often the catalytic and regulatory roles of RNA involve the interaction of two RNA strands. For example, a ribozyme needs to bind to a target RNA strand in order to splice it, and similarly a miRNA or siRNA must bind to its target in order to regulate gene expression. As a result, prediction of the structure formed by two interacting RNA strands is valuable.

Another reason why prediction of nucleic acid structure is important is because of the emerging uses of designed DNA and RNA molecules in nanotechnology. Designed DNA molecules can be self-assembled into rigid scaffolds for organization of matter at the nanoscale, or designed to serve as biosensors that can signal the presence of tiny concentrations of a chemical in solution [14]. DNA is suitable for these applications because, when single-stranded, it forms branched or pseudoknotted secondary structures in a similar manner to RNA; also DNA tends to be quite stable, unlike RNA which degrades quite quickly under standard conditions. In order to design DNA molecules with novel uses, it is necessary to first have good tools for prediction of DNA structure. Many designed nucleic acid structures incorporate pairing interactions between two (or more) strands, and so predicting the structure of interacting strands is again important for these applications.

Following some useful background in Section 2, we describe some algorithms for prediction of the secondary structure of single-stranded nucleic acids in Section 3, and for interacting nucleic acids in Section 4. We conclude with some areas for future work in Section 5. We consider only algorithms that aim to predict structure from the base sequence: that is, a single DNA or RNA sequence is given as input, and the task is to predict the structure into which this molecule folds, under fixed conditions. This paper is not intended to be a comprehensive review, but aims to highlight some of the key algorithmic advances and challenges pertaining to nucleic acid secondary structure prediction from the base sequence. We note that, from a computational standpoint, there are other very interesting challenges pertaining to nucleic acid structure. One is prediction of secondary structure common to two or more homologous sequences, that is, sequences which are evolutionarily related and form similar structures [7]. Another is discovery of new non-coding RNA's in sequenced genomes [40].

2. Background

Throughout this paper, we focus on RNA molecules, but the principles of structure prediction are essentially the same for DNA molecules. An RNA molecule is a sequence of nucleotides, or bases, of which there are four types: Adenine (A), Guanine (G), Cytosine (C), and Uracil (U). The molecule has chemically distinct ends, called the 5′ and 3′ ends. We model an RNA molecule as a sequence over the alphabet {A, C, G, U}, with the left end of the sequence being the 5′ end. Throughout, n denotes the length of an RNA sequence. We index the bases consecutively from the 5′ end starting from 1, and refer to a base by its index.

When an RNA molecule folds, bonds may form between certain pairs of bases, where each base may pair with at most one other base. The resulting structure depends on environmental conditions, such as temperature and salt concentration of the solution in which the molecule resides. A *secondary structure* R is a set of pairs $i.j$, $1 \leq i < j \leq n$, such that no index occurs in more than one pair. The pair $i.j$ denotes that the base indexed i is paired with the base indexed j. The canonical base pairs, which form the secondary structure, are the Watson–Crick pairs A–U and C–G, as well as the wobble pair G–U.

Pair $i.j$ is *pseudoknotted* if it *crosses* some base pair $i'.j'$, that is, exactly one of i' and j' is in the set $\{i, \ldots, j\}$. A structure R is considered pseudoknotted if it has at least one pseudoknotted base pair. Similarly R is pseudoknot free if it has no pseudoknotted base pair.

We can illustrate the secondary structure of an RNA molecule on a circular graph called a polymer graph, in which the bases of the RNA molecule are presented on the circumference of the circle and the base pairs are presented by straight lines (undirected edges) connecting the two pairing bases. A pseudoknotted secondary structure has crossing lines in the polymer graph. See Fig. 1.

As noted in the introduction, the RNA structure forms because it is thermodynamically (energetically) favourable for bases to form paired helices. Accordingly, computational secondary structure prediction from the base sequence is typically done by finding the thermodynamically most stable (minimum free energy) secondary structure. To explain this, we need to describe loops and loop energies.

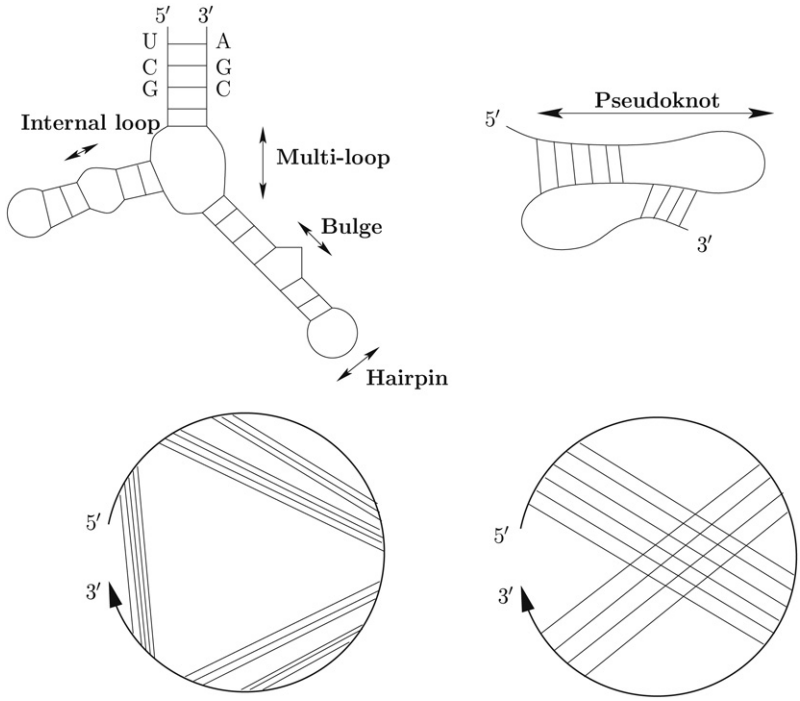

Fig. 1. From left to right and top to bottom, a pseudoknot free and a pseudoknotted structure in their base pair structure and polymer graphs. In the figure at the top left, bases are arranged along the backbone; three bases at each of the 5′ and 3′ ends are illustrated. In the top left structure, loops are annotated by their type: loops with one emanating branch are hairpin loops, with two emanating branches are either internal or bulge loops; and with three or more emanating branches are multi-loops.

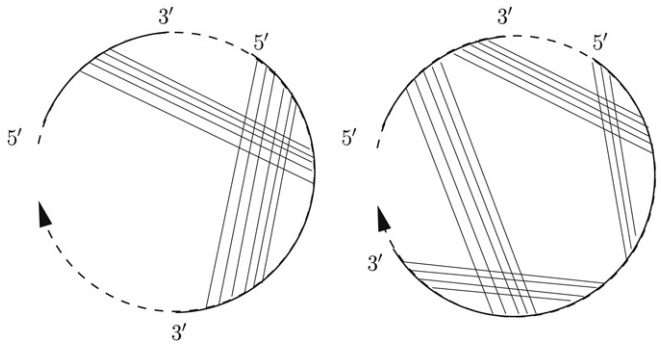

Fig. 2. Two pseudoknotted structures of multiple RNA molecules.

Loosely speaking, the region enclosed by one or more base pairs forms a *loop*. With respect to fixed environmental conditions, each loop has an associated energy value. Biochemists have developed tables and rules for calculating the energy of the loop, based on its closing base pairs and the unpaired bases around the loop [26]. (For large loops, the number of unpaired bases is used, rather than the actual bases.) The free energy of a secondary structure is calculated by summing its loop energies. The lower the free energy of the structure, the more stable is the structure.

So far, we have considered only structures formed from a single RNA molecule. When two or more molecules interact, the situation is similar. We can represent the secondary structure of interacting RNAs by representing each participating RNA strand by an arc, or segment, of the circle enclosing a polymer graph, such that there is a gap between two consecutive segments and no segments overlap. The base pairings are then represented just as in the single-stranded case, by straight lines connecting the two pairing bases. A multi-RNA structure is *connected* if the corresponding polymer graph is connected. There are $(L-1)!$ different ways (orderings) to place L RNA molecules on a circle. A connected multi-RNA structure is considered pseudoknotted if all of these orderings result in a structure with crossing base pairs (see Fig. 2). Thus, a secondary structure is considered pseudoknot free if at least one of its corresponding polymer graphs does not have crossing base pairs. Dirks et al. [12] showed an elegant *representation theorem*, that if a connected multi-RNA structure is pseudoknot free, there is exactly one ordering of the RNA molecule segments around the circle that results in a structure with non-crossing base pairings. See Fig. 3.

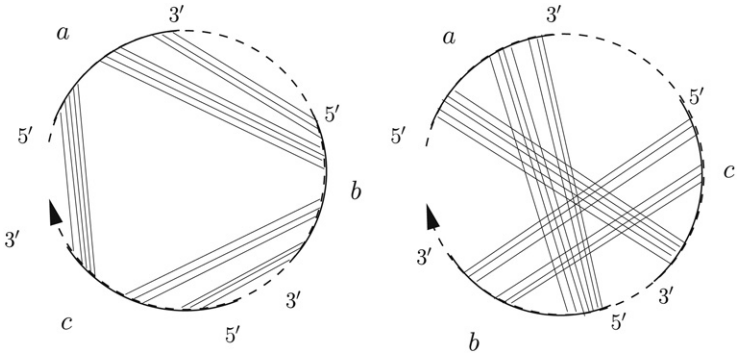

Fig. 3. Each connected pseudoknot free structure involving interacting strands has exactly one polymer graph with no crossing lines. For the example shown, involving three strands (a, b, and c), the graph on the left is that with no crossing lines. In the polymer graph obtained by arranging the strands in the order a, c, b, there are crossing lines, as illustrated on the right.

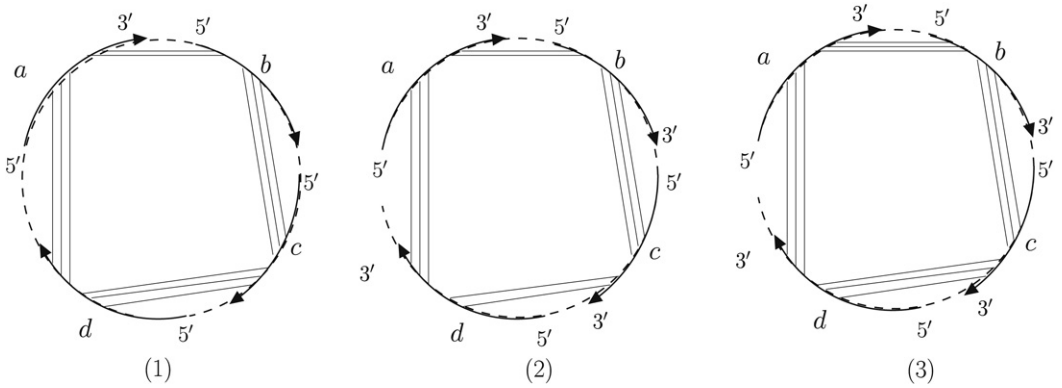

Fig. 4. Polymer graphs of four interacting strands. (1) The structure has 1-fold rotational symmetry (i.e. does not have rotational symmetry). (2) If strands a and c, and b and d are respectively indistinguishable, then this structure has 2-fold rotational symmetry. (3) If all strands are indistinguishable, then this structure has 4-fold symmetry.

In what follows, when we refer to a multi-RNA structure, we mean a connected multi-RNA secondary structure.

The free energy of a multi-RNA structure is also defined as the sum of energy contributions of loops. But, when two (or more) of the participating strands are *indistinguishable* (have identical sequences), an additional term must also be added to the sum. A (connected) structure formed from indistinguishable strands may have *rotational symmetries*. Informally, rotational symmetries arise in a structure when the polymer graph can be rotated (less than 360 degrees) to a configuration that is identical to its original configuration, both in sequence composition and in structure. Fig. 4 illustrates 1-fold, 2-fold, and 4-fold rotational symmetries; see Dirks et al. [12] for details. The additional term is $kT \log R$ for a structure with R-fold symmetry, where k is the Boltzmann constant and T is the temperature.

3. Predicting the secondary structure of single stranded RNA molecules

In this section, we first give an overview of methods for predicting the minimum free energy secondary structure for an RNA sequence. Then, we briefly discuss a related problem, that of calculating the partition function. Finally, we describe an alternative approach to prediction of pseudoknotted structures, based on the hierarchical folding hypothesis.

3.1. MFE secondary structure prediction

There has been significant success in prediction of *pseudoknot free* secondary structures, that is, structure which have no crossing base pairs. State-of-the-art prediction algorithms, such as Mfold [26] or RNAfold [18] find the structure with *minimum free energy* (MFE) from the set of all possible pseudoknot free secondary structures. The energy of a structure is estimated as the sum of energies of loops that form when the molecule folds. Algorithms for MFE secondary structure prediction of pseudoknot free secondary structures exploit two useful properties. First, a pseudoknot free secondary structure for a sequence is either closed by a base pair connecting the first and last base in the sequence, or can be broken down into two independent substructures on a prefix and suffix of the sequence. Second, the total energy of a structure which is composed of two independent substructures is the sum of the energies of the loops of the substructures. As a result, a divide-and-conquer approach, based on dynamic programming, can be used to find the MFE pseudoknot free secondary structure; the runtime of such algorithms is $\Theta(n^3)$ for standard energy loop models [30].

While many small RNA secondary structures are pseudoknot free, pseudoknots do arise frequently in biologically-important RNA molecules, both in the cell [36,39], and in viral RNA [9]. Examples include simple H-type pseudoknots, with two interleaved stems, which are essential for certain catalytic functions and for ribosomal frameshifting [2], as well as kissing hairpins, which are essential for replication in the coxsackie B virus [28].

Unfortunately, MFE pseudoknotted secondary structure prediction is NP-hard [1,22,23], even for a simple energy model that depends on base pairs but not on unpaired bases and a single RNA. Polynomial-time MFE-based approaches to pseudoknotted structure prediction have been proposed [1,13,32,34,38], with respect to various sum-of-loops energy models for pseudoknotted structures, which find the MFE structure for a given input sequence, from a restricted class of structures. A class of structures can be defined by specifying allowable patterns of interleaving among base pairs. For example, Mfold and RNAfold handle the class of pseudoknot free secondary structures; we provide more examples later. We say that a structure *R can be handled* by a given algorithm if R is in the class of structures over which the algorithm optimises.

Algorithms for MFE pseudoknotted secondary structure prediction trade off run-time complexity and *generality* — the class of structures handled, that is, the class of structures over which the algorithms optimise. For example, kissing hairpins are not in the class of structures handled by the $\Theta(n^5)$ algorithms of Akutsu [1] and Dirks and Pierce [13] but are in the class handled by the $\Theta(n^6)$ algorithm of Rivas and Eddy [34]. (We note that, even when the true structure R for a sequence is handled by an algorithm, the algorithm still may not correctly predict R, because correctness depends not only on the generality of the algorithm but also on the energy model and energy parameters.)

3.2. Calculating the partition function

The MFE approach to structure prediction assumes that a molecule folds into one stable structure, namely that with the lowest energy. A more accurate viewpoint is that the folding process is probabilistic, with the MFE structure being the most likely to form. It is assumed that the probability of a sequence being in a given structure depends on the energy of that structure, following a Boltzmann distribution. Thus, the probability of a given structure for a sequence is proportional to an exponent in the free energy of that structure. More precisely, for a fixed sequence, the probability that the sequence folds into structure S is $e^{\Delta G(S)/RT}/Q$. Here, $\Delta G(S)$ is the free energy of secondary structure S, R is the universal gas constant, and T is the absolute temperature. The quantity Q is called the partition function, and is given by

$$Q = \sum_S e^{\Delta G(S)/RT},$$

where the sum is taken over all secondary structures S for the fixed sequence.

Knowing the base pair probabilities is very valuable; they can indicate, for example, whether a sequence has two quite different low-energy structures. Furthermore, predictions of base pairs that occur with high probability tend to be more accurate than predictions of base pairs that occur with low probability [24].

In 1990, McCaskill [27] presented a dynamic programming algorithm for computing the partition function Q of a given sequence, where the sum is taken over all possible pseudoknot free secondary structures. This algorithm determines the partition function for a strand using a divide-and-conquer dynamic programming approach that is quite similar to that used for MFE structure prediction. Care must be taken that to avoid over-counting: each structure should make exactly one contribution to Q. (In contrast, when finding the MFE energy, it is OK to take the minimum over a set of structures containing duplicate structures.) McCaskill's algorithm has run time complexity of $O(n^4)$ which can be further reduced to $O(n^3)$ using standard methods.

In 2003, Dirks and Pierce [13] extended McCaskill's $O(n^3)$ algorithm to handle a class of pseudoknotted structures. Their algorithm has run time complexity of $O(n^5)$.

3.3. Prediction of pseudoknots based on hierarchical folding hypothesis

Recently, we introduced a new approach, called HFold [19], motivated by the two limitations of MFE-based algorithms for pseudoknotted secondary structure prediction: they have high time complexity, and ignore the folding pathway from unfolded sequence to stable structure. HFold is based on the *hierarchical folding hypothesis*, which is succinctly stated by Tinoco and Bustamante as follows: "An RNA molecule [has] a hierarchical structure in which the primary sequence determines the secondary structure which, in turn, determines its tertiary folding, whose formation alters only minimally the secondary structure" [37]. (These and other authors consider the initially-formed secondary structure to be pseudoknot free, and refer to base pairs that form pseudoknots as part of the tertiary structure. However, here we refer to all canonical base pairs, namely A–U, C–G, and G–U, as secondary structure.)

HFold predicts RNA secondary structures in a manner consistent with a natural formalisation of the hierarchical folding hypothesis. Consider the problem of hierarchically predicting secondary structure as follows: given a sequence S and a pseudoknot free secondary structure G (a set of base pairings), find a pseudoknot free secondary structure G' (a set of base pairings disjoint from G) for S, such that the free energy of $G \cup G'$ is less than or equal to the free energy of $G \cup G''$ for all pseudoknot free structures $G'' \neq G'$.

As with algorithms for MFE pseudoknotted secondary structure prediction, algorithms for Hierarchical-MFE secondary structure prediction may handle a restricted class of structures. That is, the type of structure formed by $G \cup G'$ may have restricted patterns of interleaving among base pairs. Since both G and G' are pseudoknot free, the most general class of structures that could be handled by an algorithm for hierarchical-MFE secondary structure prediction would be the *bi-secondary* structures of Witwer et al. [42] — those structures which can be partitioned into two pseudoknot free secondary structures G and G'. There is no known way to solve the hierarchical-MFE prediction for the class of bi-secondary structures. Instead, HFold solves the problem with respect to a subclass of the bi-secondary structures, which we called *density-2* structures.

HFold handles a general class of structures, including H-type pseudoknots and kissing hairpins, as well as structures containing nested instances of these structural motifs. The only known algorithm for predicting MFE nested kissing hairpins, that of Rivas and Eddy, requires $\Omega(n^6)$ time. Rastegari and Condon [31] showed that, out of a set of over 1,100 biological structures, all but nine are density-2 (when isolated base pairs are removed), and six of these nine are also not in the class handled by Rivas and Eddy's algorithm.

4. Interacting RNA molecules

Because of the variety of ways in which interacting RNA and DNA molecules contribute to cellular function, are useful in gene knockdown studies (that is, techniques in which an organism is genetically modified to have reduced expression of one or more genes), and have potential in therapeutics and in nanotechnology, methods for computational prediction of the structure of interacting RNA molecules are very useful. A key challenge is to predict the MFE structure formed from two or more strands. A closely related problem is to calculate the partition function for a complex of interacting strands. We describe methods for these purposes in Section 4.1.

A related problem is to calculate how tightly one molecule binds to another. For example, in order to design siRNA's (small interfering RNA) for a target mRNA, we need to assess which short strand that is complementary to a stretch of the target has strongest "binding affinity" to the target. Several different measures of binding affinity have been proposed, and algorithms to calculate these measures have been developed. We review these in Section 4.2.

4.1. MFE and partition function calculations for interacting strands

MFE pseudoknot free joint secondary structure. Mathews et al. [25] noted that the dynamic programming method for predicting the secondary structure of single-stranded nucleic acids can easily be adapted to two strands. Briefly, the method is run on the concatenation of the strands, but keeps track of the "break point" at which one strand starts and the other ends. The recurrences are adjusted to account for the fact that an unpaired region surrounding the break point is not a loop, and to include a penalty for associating the two strands. Andronescu et al. [4,6] extended the same idea to handle multiple strands. However, neither of these methods account for symmetry correction, and so may not compute the true MFE structure if two or more of the interacting strands are indistinguishable.

Dirks et al. [12] explained one method for calculating the MFE pseudoknot free secondary structure formed by two strands, which does correct for symmetry. First, calculate the MFE secondary structure, say S, without symmetry correction. If this secondary structure has no symmetries, it is the true MFE structure, since the symmetry correction term is always positive. Otherwise, if S has R-fold rotational symmetry, then enumerate all secondary structures whose energies without symmetry correction are within $kT \log R$ of the free energy of S without symmetry correction. Finally, re-evaluate the energy of these structures, this time correcting for symmetry, and choose the structure with the lowest energy. Unfortunately, this method requires exponential time in the worst case. No efficient algorithm is known for calculating the MFE pseudoknot free secondary structure formed by two strands, which corrects for symmetry.

MFE pseudoknotted joint secondary structure. Algorithms for predicting the MFE pseudoknotted secondary structure formed by multiple stands, without symmetry correction, can be easily obtained by adapting pseudoknotted algorithms for single strands, in the same manner as described in the pseudoknot free case. Another dynamic programming algorithm was proposed by Alkan et al. [3]. This method takes $\Theta(n^6)$ time on two strands of length n. The authors note that their method can predict structures which could not be predicted by an adaptation of the Rivas and Eddy algorithm.

Partition function. Dirks et al. [12] presented the first algorithm for calculating the partition function of a pseudoknot free complex of multiple interacting nucleic acid strands. Their method also uses dynamic programming. In addition to the need to correct for symmetry as discussed above, the partition function calculation (unlike the MFE calculation) should ensure that each possible structure that could be formed from the interacting strands is counted exactly once. While correcting for symmetry or properly counting appear to be hard to do on their own, Dirks et al. showed that both could be done simultaneously, using a correction factor which is derived from group-theoretic insights. The time and space complexity of the algorithm of Dirks et al. are $O(n^3)$ and $O(n^2)$ respectively when n is the total length of input strands.

4.2. Measuring binding affinity

Now, consider the case when one of two interacting strands is a short oligo (short sequence of nucleotides, also called oligonucleotide), which binds with a much longer, so-called target strand. The free energy of the MFE joint structure formed

by the two strands is an inappropriate measure of binding affinity, because it includes terms pertaining to structure formed by parts of the target alone.

A MFE measure of binding affinity. One alternative is to assume that the MFE joint structure, in the region where the oligo binds to the target, involves no intra-molecular base pairings and that loop lengths are short – bounded by a constant. With these assumptions, the goal is to predict the MFE joint structure formed by oligo and target. Rehmsmeier et al. [33] noted that the dynamic programming recurrences for MFE structure prediction can be simplified to yield an algorithm with running time $O(mn)$, where m and n are the lengths of the oligo and target, respectively. They developed a method, RNAhybrid, to do this.

When the short oligo is a microRNA or siRNA, m is typically small (at most 25 nucleotides) and can be treated as a constant compared with n. Then the time complexity of RNAhybrid is linear to the length of the target. RNAhybrid also predicts optimal and additional suboptimal, non overlapping hits. RNAhybrid does not handle pseudoknotted structures.

Accounting for target structure. Several authors have stressed the need to account for secondary structure of the target, when measuring binding affinity of an oligo to the target [21,26,29,35]. This is because, if the target has structure, part of the structure may need to be disrupted, in order to create or expose an unpaired region which can then bind to the oligo.

Mathews et al. [25] introduce a measure of oligo-target binding affinity which they call equilibrium affinity. They assume that the target structure is known (perhaps obtained via MFE prediction). Roughly speaking, the equilibrium affinity is the overall free energy change, at equilibrium, of a system that includes targets which have structure and are not bound to an oligo, targets with disrupted structure which are bound to an oligo, and oligos which are self-folded or bound to another oligo. Given a target strand and its structure, the OligoWalk method calculates the equilibrium affinity for each oligo of a given length which is complementary to a stretch of the target. The run time for OligoWalk is $O(n^2)$, where n is the length of the target sequence, if only local rearrangements in target structure are allowed upon binding of oligos, and the run time is $O(n^4)$ if global rearrangements in target structure are allowed.

Mückstein et al. [29] described an extension to McCaskill's partition function algorithm for single strands, which calculates the probability that a region is unpaired in a given structure. They then used this probability value in predicting where the oligo would bind to the target strand. The authors showed that when the target strand is large, this algorithm has time complexity $O(n^3)$. This algorithm, RNAup, handles a restricted class of pseudoknotted RNA–RNA structures in which, for example, the oligo binds to the unpaired bases of a hairpin loop of the target structure. This class of structures is not handled by the concatenation based RNA–RNA structure prediction algorithms that do not consider pseudoknots. However, this algorithm cannot predict complex pseudoknotted structures resulting, for example, from binding of an oligo to unpaired bases of two hairpin loops of the target structure (see Fig. 2).

Shao et al. [35] note that long targets may form one of several stable structures. This, together with the fact that MFE secondary structure predictions are inaccurate, supports the consideration of multiple target structures when measuring oligo binding affinity. They found that, rather than using suboptimal structures, it is better to obtain statistically representative sample of structures, from the Boltzmann weighted ensemble of secondary structures. The Sfold program [10,11] efficiently provides such a sample. They defined a quantitative measure called "target disruption energy", which is the energy change (increase) needed to locally disrupt the target structure at the point where the oligo binds, averaged over 1000 target structures predicted by Sfold, and showed that accounting for target disruption energy is important, when predicting binding affinity of an oligo to its target.

In an investigation of microRNA target recognition, Kertesz et al. [21] also account for target structure. They proposed an energy-based score equal to the difference between the free energy gained by the binding of the microRNA to the target and the free energy cost of unpairing the target site nucleotides. To compute the above energy values, they used MFE secondary structure prediction methods.

Other considerations. When predicting microRNA or siRNA efficiency (that is, the likelihood that they are effective in gene regulation), factors other than thermodynamic measures of binding affinity are also important. For example, microRNA's are held in a protein complex called the RISC (RNA-induced silencing complex); Kertesz et al. [21] found that accounting also for the cost of unpairing bases flanking the region to which microRNA is bound resulted in better predictions. Tertiary structure may also play a role [26]. We note that none of the methods above handle pseudoknotted structures.

Shao et al. [35] provide a nice summary of several important factors that influence the efficiency of potential siRNA's. For example, the 5′ end of the oligo should form a stem with its complement on the target that is less stable than the stem formed by the 3′ end. Also, in some positions, certain nucleotides are preferred.

Further discussion of these considerations is beyond the scope of this review, but we note that computational methods for predicting binding efficiency can be fruitfully combined with methods that assess other factors. For example, RNAhybrid allows the user to force hybridisations to contain perfect helices in the 5′-part of the microRNA.

5. Conclusion and future work

Much attention has been recently directed to RNA molecules in general and RNA secondary structure prediction in particular. There has been significant progress in the past decade, in development of algorithms for prediction of pseudoknotted structures, and of secondary structures formed from interacting strands, based on thermodynamic principles. However, the computational complexity of many of these methods still remains high. As discussed before, most of the

secondary structure prediction algorithms either exclude pseudoknotted structures in general or are very costly and inefficient in predicting pseudoknots.

We have not discussed the prediction accuracy of the methods presented here. Accuracy can be measured, for example, as the percentage of bases that are correctly predicted (as unpaired, or as paired with the proper partner). Mathews et al. [26] did an extensive study of the standard energy model on a large set of structures, and reported 72% accuracy for MFE pseudoknot free prediction. The accuracy of the methods hinges greatly on the accuracy of the underlying energy model. Andronescu et al. [5] have developed methods for improving the energy values of features (loop contributions) of the current energy model, resulting in improved parameters. For significantly better accuracy, particularly for pseudoknotted structures, changes in the feature set may be needed. Energy model improvement is also necessary for predicting the secondary structure of interacting RNA molecules. Also, in the case of interacting RNA molecules, much effort still needs to be put in the development of algorithms that accurately handle pseudoknotted structures. As the amount of data on the characteristics of the structure of interacting RNA's increases, it should be possible to improve our thermodynamic models of structure formation, and thus to improve the accuracy of structure prediction.

References

[1] T. Akutsu, Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots, Discrete Appl. Math. 104 (1–3) (2000) 45–62.
[2] S.L. Alam, J.F. Atkins, R.F. Gesteland, Programmed ribosomal frameshifting: Much ado about knotting! PNAS 96 (25) (1999) 14177–14179.
[3] C. Alkan, E. Karako, J.H. Nadeau, S.C. Sahinalp, K. Zhang, RNA–RNA interaction prediction and antisense RNA target search, J. Comput. Biol. 13 (2) (2006) 267–282.
[4] M. Andronescu, R. Aguirre-Hernàndez, A. Condon, H.H. Hoos, RNAsoft: A suite of RNA secondary structure prediction and design software tools, Nucleic Acids Res. 31 (13) (2003) 3416–3422.
[5] M. Andronescu, A. Condon, H. Hoos, D. Mathews, K. Murphy, Efficient parameter estimation for RNA secondary structure prediction, Bioinformatics 23 (13) (2007) i19–i28.
[6] M. Andronescu, Z.C. Zhang, A. Condon, Secondary structure prediction of interacting RNA molecules, J. Mol. Biol. 345 (5) (2005) 987–1001.
[7] M.F. Browner, C.B. Lawrence, Comparative sequence analysis as a tool for studying the secondary structure of mRNAs, Nucleic Acids Res. 13 (23) (1985) 8645–8660.
[8] J. Couzin, Breakthrough of the year: Small RNAs make big splash, Science 298 (5602) (2002) 2296–2297.
[9] B.A.L.M. Deiman, C.W.A. Pleij, Pseudoknots: A vital feature in viral RNA, Sem. Virol. 8 (3) (1997) 166–175.
[10] Y. Ding, C.Y. Chan, C.E. Lawrence, Sfold web server for statistical folding and rational design of nucleic acids., Nucleic Acids Res. 32 (Web Server issue).
[11] Y. Ding, C.E. Lawrence, A statistical sampling algorithm for RNA secondary structure prediction, Nucleic Acids Res. 31 (24) (2003) 7280–7301.
[12] R.M. Dirks, J.S. Bois, J.M. Schaeffer, E. Winfree, N.A. Pierce, Thermodynamic analysis of interacting nucleic acid strands, SIAM Rev. 49 (1) (2007) 65–88.
[13] R.M. Dirks, N.A. Pierce, A partition function algorithm for nucleic acid secondary structure including pseudoknots, J. Comput. Chem. 24 (13) (2003) 1664–1677.
[14] R.M. Dirks, N.A. Pierce, From the cover: Triggered amplification by hybridization chain reaction, PNAS 101 (43) (2004) 15275–15278.
[15] J.A. Doudna, T.R. Cech, The chemical repertoire of natural ribozymes, Nature 418 (6894) (2002) 222–228.
[16] E. Gerhart, H. Wagner, K. Flardh, Antisense RNAs everywhere? Trends in Genetics 18 (5) (2002) 223–226.
[17] G.J. Hannon, J.J. Rossi, Unlocking the potential of the human genome with RNA interference, Nature 431 (7006) (2004) 371–378.
[18] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster, Fast folding and comparison of RNA secondary structures, Monatshefte für Chemie / Chemical Monthly 125 (2) (1994) 167–188.
[19] H. Jabbari, A. Condon, A. Pop, C. Pop, Y. Zhao, HFold: RNA pseudoknotted secondary structure prediction using hierarchical folding, in: Algorithms in Bioinformatics, Springer, Berlin, Heidelberg, 2007, pp. 323–334.
[20] G.F. Joyce, Directed evolution of nucleic acid enzymes, Annu Rev Biochem 73 (2004) 791–836.
[21] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, E. Segal, The role of site accessibility in microRNA target recognition, Nat Genet 39 (10) (2007) 1278–1284.
[22] R.B. Lyngsø, Complexity of pseudoknot prediction in simple models, in: J. Díaz, J. Karhumäki, A. Lepistö, D. Sannella (Eds.), ICALP, in: Lecture Notes in Computer Science, vol. 3142, Springer, 2004.
[23] R.B. Lyngsø, C.N. Pedersen, RNA pseudoknot prediction in energy-based models, J. Comput. Biol. 7 (3–4) (2000) 409–427.
[24] D.H. Mathews, Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization, RNA 10 (8) (2004) 1178–1190.
[25] D.H. Mathews, M.E. Burkard, S.M. Freier, J.R. Wyatt, D.H. Turner, Predicting oligonucleotide affinity to nucleic acid targets, RNA 5 (11) (1999) 1458–1469.
[26] D.H. Mathews, J. Sabina, M. Zuker, D.H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J. Mol. Biol. 288 (5) (1999) 911–940.
[27] J.S. McCaskill, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers 29 (6–7) (1990) 1105–1119.
[28] W. Melchers, J. Hoenderop, H. Bruins Slot, C. Pleij, E. Pilipenko, V. Agol, J. Galama, Kissing of the two predominant hairpin loops in the coxsackie B virus 3' untranslated region is the essential structural feature of the origin of replication required for negative-strand RNA synthesis, J. Virol. 71 (1) (1997) 686–696.
[29] U. Mückstein, H. Tafer, J. Hackermuller, S.H. Bernhart, P.F. Stadler, I.L. Hofacker, Thermodynamics of RNA–RNA binding, Bioinformatics 22 (10) (2006) 1177–1182.
[30] R. Nussinov, A.B. Jacobson, Fast algorithm for predicting the secondary structure of single-stranded RNA, PNAS 77 (11) (1980) 6309–6313.
[31] B. Rastegari, A. Condon, Parsing nucleic acid pseudoknotted secondary structure: Algorithm and applications, J. Comput. Biol. 14 (1) (2007) 16–32.
[32] J. Reeder, R. Giegerich, Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics., BMC Bioinformatics 5 (104).
[33] M. Rehmsmeier, P. Steffen, M. Hochsmann, R. Giegerich, Fast and effective prediction of microRNA/target duplexes, RNA 10 (10) (2004) 1507–1517.
[34] E. Rivas, S. Eddy, A dynamic programming algorithm for RNA structure prediction including pseudoknots, July 1998.
[35] Y. Shao, C.Y. Chan, A. Maliyekkel, C.E. Lawrence, I.B. Roninson, Y. Ding, Effect of target secondary structure on RNAi efficiency, RNA 13 (2007) 1631–1640.
[36] D.W. Staple, S.E. Butcher, Pseudoknots: RNA structures with diverse functions, PLoS Biol. 3 (6) (2005) e213.
[37] I. Tinoco, C. Bustamante, How RNA folds, J. Mol. Biol. 293 (2) (1999) 271–281.
[38] Y. Uemura, A. Hasegawa, S. Kobayashi, T. Yokomori, Tree adjoining grammars for RNA structure prediction, Theor. Comput. Sci. 210 (2) (1999) 277–303.
[39] F.H. van Batenburg, A.P. Gultyaev, C.W. Pleij, Pseudobase: Structural information on RNA pseudoknots, Nucleic Acids Res. 29 (1) (2001) 194–195.
[40] Z. Weinberg, J.E. Barrick, Z. Yao, A. Roth, J.N. Kim, J. Gore, J.X. Wang, E.R. Lee, K.F. Block, N. Sudarsan, S. Neph, M. Tompa, W.L. Ruzzo, R.R. Breaker, Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline, Nucleic Acids Res. 35 (14) (2007) 4809–4819.
[41] W.C. Winkler, R.R. Breaker, Regulation of bacterial gene expression by riboswitches, Annu. Rev. Microbiol. 59 (2005) 487–517.
[42] C. Witwer, I.L. Hofacker, P. Stadler, Prediction of consensus RNA secondary structures including pseudoknots, IEEE/ACM Trans. Comput. Biol. Bioinformatics 1 (2) (2004) 66–77.