Computational RNA Structure Prediction

Samuel Jackson, University Of Aberystwyth

***Abstract*—The abstract goes here.**

***Index Terms*—RNA, literature review, dynamic programming, context free grammars, bioinformatics, computational biology.**

I. Introduction

Ribonucleic acid (RNA) is a fundamental biological macromolecule that play an important part in many biological processes. RNA folding, like DNA and protein folding can be classified into three different categories: primary, secondary, and tertiary structure. The primary structure refers to the sequence of building blocks of RNA. These are the nucleotides guanine, uracil, adenine, and cytosine, commonly abbreviated to the first letter of each molecule (G, U, A, C). While DNA molecules frequently form the eponymous double helix between two strands, RNA molecules are single stranded and form complex structures by folding upon themselves.

A. RNA secondary structure

The way in which nucleotides in an RNA strand form hydrogen bonds with one each is referred to as the RNA's secondary structure. In RNA molecules bonds are usually formed between the base pairs A-U, C-G, and U-G. The patterns of the structures that are created through base-pairing can be classified into a number of different sub structures. Reference [1] provides a comprehensive introduction. Commonly encountered structural motifs frequently used in RNA secondary structure prediction are:

- **Base pair stacks** - The most common structural element. Formed by an RNA strand folding on itself and forming hydrogen bonds between complementary bases. Bonds in base pair stacks form between two parts of the RNA each running in an anti-parallel direction to one another.
- **Hairpin loops** - A collection of unpaired nucleotides at the terminus of a base pair stack. So called because the strand loops back and binds with itself.
- **Symmetric and asymmetric loops** - a collection of unpaired nucleotides between two base pair stacks. Symmetrical if the number of nucleotides on each side is equal, asymmetrical if not.
- **Bulges** - Similar to loops but with one side having no unpaired nucleotides.
- **Junctions** - The point at which multiple base pair stacks meet is referred to as a junction.
- **Pseudoknots** - Pseudoknots are formed between the unpaired nucleotides on a hairpin loop with the unpaired nucleotides on an adjacent strand. So called because the structure shows some resemblance to a mathematical knot. The properties of pseudoknots and similar structures propose a particular challenge to prediction due to the complex, interwoven, long range base pairing.
- **Kissing hairpins** - Similar to pseudoknots, but directly between the unpaired nucleotides of two hairpin loops.

B. RNA tertiary structure

II. RNA secondary structure prediction

The challenge of accurately predicting the secondary structure of RNA has a long and varied history. There are two major schools of thought in secondary structure prediction, with folding algorithms being loosely categorised as being either thermodynamic or probabilistic using stochastic context free grammar (SCFG) models. Most of the approaches to folding share deep similarities in how structure is determined. The major differences are in the scoring schemes and the parametrisation used. This section provides an overview of history prediction starting from early thermodynamic models and working forwards chronologically to more recent probabilistic models.

A. Thermodynamic Models

One of the earliest influential approaches to secondary structure prediction is the Nussinov algorithm [2]. The Nussinov algorithm is used to find the maximum base pairing of a sequence of nucleotides. The algorithm recursively calculates the maximum pairing for subsections of a RNA sequence. The recursive definition can be sped up using a dynamic programming table to yield an algorithm with $O(n^3)$ time and $O(n^2)$ space complexity. Little improvement on algorithmic complexity has been achieved since.

While the Nussinov algorithm is guaranteed to produce the structure with maximum base pairs it has some major flaws. Firstly the algorithm assumes base pairs are non-crossing and cannot handle pseudo-knotted structures. Secondly, it usually does not produce biologically plausible structures. For example, the stacking orientation of base pairs and loop length are not weighted in any way. Thirdly, the algorithm only predicts a single structure. It is known that the space of possible secondary structures will often have many plausible instances close to the optimum structure [4]. The Nussinov algorithm provides no way of differentiating between possible sub optimum structures.

A much more biologically feasible criteria of determining whether two bases will pair is to minimise the free energy exhibited by a structure. This is the method proposed by Zuker and Stieglar [3]. The underlying algorithm shares a very similar formulation as the Nussinov method but with a few key differences. Firstly their algorithm associates energy with the regions between bonds, as opposed to the bonds themselves (which is effectively what Nussinov uses). Secondly two energy functions are defined for subsequences of the string

TABLE I
SUMMARY OF APPROACHES TO SECONDARY STRUCTURE PREDICTION

Paper	Year	Criteria	Weighting	Contributions
Nussinov [2]	1980	Maximum Base Pairing	Binary	Dynamic programming algorithm for base pairs
Zuker and Stieglar [3]	1981	Minimum Free Energy	Thermodynamic	Minimum free energy algorithm
McCaskill [4]	1990	Partition Function Probability & MFE	Thermodynamic	Partition function, base pair probability matrix, melting behaviour description.

of nucleotides. These are the energy of the subsequence with and without base pairing between two given indices. Energy for the structure is recursively computed in a bottom-up fashion by taking the minimum energy at each point. The final computation should yield the secondary structure with minimum free energy (MFE).

The MFE formulation can be used to produce much more biologically plausible structures in contrast to base pair maximisation. The thermodynamic weights are used to push the algorithm away from impossible or implausible structures (such as very short hairpin loops) and towards the correct structure by giving them highly positive weights. The method also has a certain biological backing because of its basis in thermodynamics which is more realistic to how cell processes work than base pair maximisation.

However, this method and thermodynamic based approaches in general, are limited by the accuracy of experimental studies of RNA. Many approaches rely on custom scoring rules or simply ignore aspects of reality in the model. For example, sequence dependance in RNA loop structures are often ignored due to the lack of experimental tools for assessing their free energy contribution [5]. Many more recent prediction algorithms utilise the thermodynamic parameters used by Turner's group [6] as opposed to the weights used in the original paper.

Furthermore, this still shares some of the limitations of [2]. The MFE algorithm cannot handle pseudo-knotted structures and can only produces a single structure rather than a distribution of likely structures. Despite these limitations thermodynamic models based on this approach are still used in abundance for secondary structure prediction and produce some of the best available results [7] [8].

Moving forward in time, another key contribution to the area was the equilibrium partition function formulation by McCaskill [4]. The aim of this paper was to not only produce the MFE structure for a given sequence but to produce a visual picture of the full ensemble of alternative equilibrium structures and provides a practical method for computing probability of bases pairing.

McCaskill describes the ensemble of RNA structures using the partition function

$$Q = \sum_{s} e^{-(E(A)/kT)} \tag{1}$$

where A is a specific structure, E is the energy of a structure, T is the absolute temperature in Kelvin, and k is the Boltzmann constant. The probability of a specific structure A given sequence S is then given by

$$P(A|S) = \frac{1}{Q} e^{-(E(A)/kT)} \tag{2}$$

Finally the probability of two bases (i, j) pairing is given by

$$P((i,j)|S) = Q_{ij}/Q \tag{3}$$

where

$$Q_{ij} = \sum_{(i,j)\in A} e^{-(E(A)/kT)} \tag{4}$$

More complicated interactions where bases pair at hairpin loops, internal loops, and junctions are handled in further derivations excluded for brevity. McCaskill also outlines how to reduce the computational time and space complexity of the final algorithm to be $O(n^3)$ and $O(n^2)$ respectively.

Further contributions by the paper include the "box matrix" plot visualise the probabilities for each base pair predicted by the algorithm alongside the predicted optimal and experimental pairings. This takes the form of a matrix where each element is the probability that bases i and j will pair shown on a logarithmic scale in the upper left corner of the matrix. The lower right side of the matrix is then used to show the optimal pairings and optionally where the experimentally confirmed structure differs.

The partition function formulation also encodes information about the phase transitions for the ensemble with respect to change in temperature. This provides another window into the structural properties of a RNA sequence.

A more modern free energy minimisation approach was created by Deigan et al. [9] which incorporates additional experimental information from SHAPE experiments into there approach. Selective 2'-hydroxyl acylation analysed by primer extension (SHAPE) experiments report differences in local nucleotide flexibility. Base pairing reduces the local flexibility associated with a nucleotide which can be related to the probability that a particular nucleotide will form a base pair. The author's propose a "pseudo-free energy change" term which can be added to a regular free energy model. The term has the form

$$\Delta G_{SHAPE}(i) = m \cdot ln[SHAPEreactivity(i) + 1] + b \tag{5}$$

where i is the nucleotide number, m is a parameter that penalises base pairing in nucleotides with high SHAPE reactivities and b is parameter which is negative and represents an increment in free energy for nucleotides which exhibit low SHAPE reactivity. The author's fit these parameters against 23S rRNA which exhibits a large number of distinct structural motifs.

The author's report a high degree of accuracy compared with conventional thermodynamic parameters. The author's

note that the errors in prediction are not only less, but are generally of a shorter range.

Another method based on the thermodynamic viewpoint is the work of Ding et al. [10], [11]. Their work can largely been seen as a logical extension of McCaskill's work in [4]. Their first paper [10] presents a method for drawing a statistically representative sample from the Boltzmann ensemble of possible secondary structures. This method allows them to gain valuable insights RNA structure from a statistical mechanics point of view. The sampled distribution allows them to calculate information relating to RNA:RNA interaction sites, density of states, and predict alternative structures.

In [11] the author's continue their work to include prediction of the "best" secondary structure using the Boltzmann ensemble samples using cluster centroids. They first generate a sample from the ensemble using the method developed in [10]. The samples are then clustered using a top-down method from [12]. They note that they use the CH index to choose the number of clusters and the base pair distance as the distance metric. They define the cluster centroid as the instance which has the shortest possible distance to all others in the in the cluster.

Interestingly the authors note that there appears to be a fixed number of clusters regardless of sequence length. Furthermore the author's concluded that the MFE approach breaks down when the MFE structure is in the wrong cluster from the true structure. The limitation of this method is that the ensemble centroid is likely to be quite far removed from many sampled structures. The centroid for the cluster containing the correct structure will be far more accurate. However, it is difficult to determine the cluster containing the "correct" centroid without prior knowledge. On the other hand, probable sub optimal structures are also of great interest and the list of centroids provides yet another method for accessing this information.

B. Probabilistic Models

The major alternative school of thought for RNA secondary structure prediction is through the use of Stochastic Context Free Grammars (SCFGs). Before diving into how SCFGs are applied to RNA secondary structure prediction it is useful to define what a CFG and therefore proceed to define a SCFG.

According to Giegerich [13] a context free grammar is a formal system of rules G that produce a language L from finite set of symbols (including the empty string ϵ) called an alphabet and denoted $\mathcal{A}$. A language is simply a combination of multiple elements from $\mathcal{A}$. A grammar G is a collection of V non terminal symbols and a set of production rules of the form $X \rightarrow \alpha$ where $X \in V$ and $\alpha \in \{V \cup \mathcal{A}^*\}$ and where $\mathcal{A}^*$ is set of all combinations of A.

An example grammar from [8] which expresses the Nussinov method [2] of RNA folding discussed previously is

$$S \rightarrow Sa|SaS\hat{a}|\epsilon \tag{6}$$

Where a and $\hat{a}$ are paired bases of some string of bases S. The vertical bar represents logical OR for brevity.

Checking whether a word $w \in \mathcal{A}^*$ exists in language $L(G)$ can be achieved by creating a parse tree for w. If such a tree exists then $w \in L(G)$ else it does not. If more than one parse tree exists for a given w the language is said to be ambiguous (unlike the grammar in equation 6 which is unambiguous).

In order for a parsing algorithm to choose between multiple potential parse trees some form of scoring function must be used. One such function might favour the smallest possible parse tree for example. If the scoring function is based on probabilities then the CFG is said to be a SCFG. More formally, each production rule r has a probability π_r associated with it. The probability of one possible parse tree is the product of π_{ri} for all uses of r_i. The probability of w is then given as the sum of the probability of a parse tree over all possible parse trees for w.

The main algorithm used to parse ambiguous SCFGs is the Cocke-Younger-Kasami (CYK) algorithm [13]–[16] . The CYK algorithm used for efficiently evaluating a SCFG is essentially the same as that which is used for finding the MFE [3]. The difference is in how the probabilities used in the "stochastic" part of a SCFG are derived.

The probabilities for the production rules can be computed from the probability of individual terminals reasonably efficiently using the inside-outside algorithm [17]. The inside-outside algorithm defines how to compute the probability of a non-terminal, a production rule, and the total probability of all parse trees of a sequence. The algorithm is used so that all parse trees need not be enumerated and can be efficiently implemented using a dynamic programming table. The fitting of the probabilities used in the SCFG can achieved using expectation maximisation.

Note that the inside-outside algorithm can be seen as equivalent to the method used by McCaskill [4] to derive an equilibrium potential function based on thermodynamic parameters. The inside-outside algorithm could be seen as a generalisation to a generic potential function. In theory the thermodynamic parameters of McCaskill's model could be replaced by appropriate probabilities and achieve similar results.

A notable early attempt at RNA secondary structure prediction using SCFGs is the work of Knudsen and Hein [18], [19] in producing Pfold. In [18] they define a grammar which is so concise that it can be stated here in full:

$$\begin{aligned} S &\rightarrow LS|S \\ F &\rightarrow dFd|LS \\ L &\rightarrow s|dFd \end{aligned} \tag{7}$$

with S producing loops, F producing stems, and L choosing between a whether a position in a loop should be a continuation of the loop or the start of a stem.

The probability associated with a production rule is created using a selection of known RNA secondary structures consisting of a number of different types of RNA. In this way Pfold uses multiple sequences (in contrast to single sequence prediction). The work in [18] first calculates the probabilities for each pairing and non-pairing columns of aligned sequences using a rate matrix to capture information about mutation between sequences. From individual columns the probability of an alignment may be obtained given a known phylogenetic

tree. Finally a MAP estimate of the RNA structure can be obtained as

$$\sigma^{MAP} = \arg\max_{\sigma} P(D|\sigma, T^{ML}, M)P(\sigma|M) \quad (8)$$

where σ is the list of all possible secondary structures, M is the model (SCFG and mutational model), D the ordered set of columns, and T^{ML} the maximum likelihood estimate of the tree. The probability of each of the production rules was found using the inside-outside algorithm and expectation maximisation.

The author's further modified there work in [19] to add a number of different enhancements to their first paper. Notable additions are further robustness to alignment and sequencing errors as well as better handling of gaps and unknown nucleotides. They also refactored their implementation to only estimate the tree once before the structure is estimated to reduce execution time. Finally the method also chooses the structure with the highest expected number of correct predictions, instead of the most likely parse reported by the CYK algorithm.

Pfold has a number of strengths in contrast to thermodynamic models and single sequence prediction methods. Firstly it is not reliant on the thermodynamic parameters obtained by experimentation. This both reduces the number of parameters needed and removes the potential limitations of experimental accuracy of the parameters. Incorporating knowledge from a full set of known sequences allows a problem formation that beings to resemble something more like a traditional machine learning problem.

However, there are some obvious limitations to this technique. Most notable is the dependance on having multiple known, aligned sequences in the first place. The accuracy of prediction from any multiple technique will be limited by the accuracy of alignment. This also raises issues such as sequencing and alignment errors which must be accounted for.

Do et al. [5] produced the CONTRAfold model that takes more inspiration from the world of natural language processing. They replace the SCFG representation with a conditional log-linear model (CLLM). CLLMs have the form

$$P(\sigma|x) = \frac{exp(\mathbf{w}^T\mathbf{F}(x,\sigma))}{\Sigma_{\sigma' \in \Omega(x)} exp(\mathbf{w}^T\mathbf{F}(x,\sigma))} \quad (9)$$

where $\mathbf{w}$ is a vector of weights to be learned and $\mathbf{F}(x,\sigma)$ is a feature vector. CLLMs are a very flexible and powerful method for using rich set of possible features to create a probabilistic model. In traditional text processing applications elements of the feature vector $\mathbf{F}(x,\sigma)$ are a collection of binary functions activated based on contextual information surrounding a word. For example $\mathbf{F}_k(x,\sigma)$ might model whether the previous word was an adjective.

In the application to RNA structure prediction the elements of the feature vector correspond to a scoring related to contextual information from the RNA sequence. For example the score for a hairpin between i and j accounts for terminal mismatch interactions, hairpin length, and the loop base. The feature vectors are derived from known thermodynamic weights such as those from [6].

CONTRAfold also diverts from the use of MFE/CYK approach to recovering the best structure. Instead the authors propose a method of Maximum Expected Accuracy (MEA). MEA incorporates a parameter γ which controls a sensitivity vs. specificity tradeoff. This is defined as

$$\hat{y}_{mea} = \arg\max_{\hat{y}} \mathbb{E}[accuracy_{\gamma}(y, \hat{y})] \quad (10)$$

where $\hat{y}$ is a candidate structure and y is the true structure. $accuracy_{\gamma}$ is defined as the number of correctly unpaired positions plus the product of γ and the number of correctly paired positions.

C. Handling Pseudoknots

The majority of methods mentioned in the preceding two sections make any realistic attempt at handling pseudoknotted structures within RNA sequences. This is mostly due to the algorithmic time increase required to handle such structures. For example, an early attempt by Rivas and Eddy [20] was able to predict a restricted subset of pseudoknots but with the associated time and space complexity of $O(n^6)$ and $O(n^4)$ respectively. Obviously such an approach is intractable for anything but the most short sequences. Despite being hard to predict, these structures are of great biological interest as they often play a key role in biological processes [**?**] such as ... In this section two methods which tackle the pseudoknot prediction from different paradigms are presented.

Reeder et al. [21] produced a method that was largely based on the MFE method by Matthew's et al. [6] but with added support for what they term *canonical simple recursive pseudoknots*. The note that the majority of known examples of pseudoknots are fairly simple in structure. This allows them to make some simplifying assumptions: only two stems are allowed, bugles and internal loops are disallowed within the pseduoknot and stems at either end of the knot must be maximal. If the stems overlap then one stem is prioritised over another.

These simplifications allow them to utilise a $O(n^4)$ loop to compute the maximal length of both stems within the subsequence bounded by locations i and j for all interior pointer k and l such that $i < k < l < j$. The total energy of the pseudoknot can then be computed from the energy of the two loops and two stems for a given k and l to obtain the total energy for the pseudoknot. The values for the pseudoknot are then treated like any other term in the MFE algorithm.

Another more recent attempt at pseudoknot prediction is CyloFold by Bindewald et al. [22]. CyloFold uses a coarse grained 3D simulation of pseudoknotted structures. Their method starts by recovering all the stem structures containing > 3 base pairs from the nucleotide sequence by conventional MFE (the authors use the ViennaRNA package [23]).

Once a list of stems has been obtained a number of simulation runs are performed. Stems are added to the simulation one at a time according to Boltzmann weighted probability. Each stem in the simulation is represented by capsule with length proportional to the length of the sequence. The position of the capsules three dimensions are initialised randomly.

Single stranded regions between cylinders are represented as distance constraints between the hemispherical ends of each capsule. Distances are constrained by a minimum and maximum bounds. The existing and newly added capsules then optimised to satisfy distance constraints and minimise collisions. If a newly added cylinder collides with existing stems it is reinitialised several times until a threshold where it is detailed to be a failure. Likewise if after optimisation there the capsule still collides it is removed.

The authors showed that their method offered some improvement on several criteria over existing methods such as pknotsRG. The time complexity of the method is difficult to estimate, but the authors suggest that it is roughly proportional to $O(n^4)$ making it comparable to [21]. The noted benefits of this method are twofold: 1) it avoid some of the simplifying constraints associated with approaches pknotsRG (but possibly does not solve more complex pseudoknots), and 2) provides an automatic check for distance constraints between structures (steric feasiblity).

III. RNA tertiary structure prediction

Acknowledgment

The authors would like to thank...

References

[1] J. Nowakowski and I. Tinoco, "Rna structure and stability," in *Seminars in virology*, vol. 8, no. 3. Elsevier, 1997, pp. 153–165.

[2] R. Nussinov and A. B. Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded rna," *Proceedings of the National Academy of Sciences*, vol. 77, no. 11, pp. 6309–6313, 1980.

[3] M. Zuker and P. Stiegler, "Optimal computer folding of large rna sequences using thermodynamics and auxiliary information," *Nucleic acids research*, vol. 9, no. 1, pp. 133–148, 1981.

[4] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for rna secondary structure," *Biopolymers*, vol. 29, no. 6-7, pp. 1105–1119, 1990.

[5] C. B. Do, D. A. Woods, and S. Batzoglou, "Contrafold: Rna secondary structure prediction without physics-based models," *Bioinformatics*, vol. 22, no. 14, pp. e90–e98, 2006.

[6] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure," *Journal of molecular biology*, vol. 288, no. 5, pp. 911–940, 1999.

[7] C. Laing and T. Schlick, "Computational approaches to 3d modeling of rna," *Journal of Physics: Condensed Matter*, vol. 22, no. 28, p. 283101, 2010.

[8] E. Rivas, "The four ingredients of single-sequence rna secondary structure prediction. a unifying perspective," *RNA biology*, vol. 10, no. 7, pp. 1185–1196, 2013.

[9] K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks, "Accurate shape-directed rna structure determination," *Proceedings of the National Academy of Sciences*, vol. 106, no. 1, pp. 97–102, 2009.

[10] Y. Ding and C. E. Lawrence, "A statistical sampling algorithm for rna secondary structure prediction," *Nucleic acids research*, vol. 31, no. 24, pp. 7280–7301, 2003.

[11] Y. Ding, C. Y. Chan, and C. E. Lawrence, "Rna secondary structure prediction by centroids in a boltzmann weighted ensemble," *Rna*, vol. 11, no. 8, pp. 1157–1166, 2005.

[12] P. J. Rousseeuw and L. Kaufman, *Finding Groups in Data*. Wiley Online Library, 1990.

[13] R. Giegerich, "Introduction to stochastic context free grammars," *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, pp. 85–106, 2014.

[14] J. Cocke, "Programming languages and their compilers: Preliminary notes," 1969.

[15] D. H. Younger, "Recognition and parsing of context-free languages in time n 3," *Information and control*, vol. 10, no. 2, pp. 189–208, 1967.

[16] T. Kasami, "An efficient recognition and syntaxanalysis algorithm for context-free languages." DTIC Document, Tech. Rep., 1965.

[17] K. Lari and S. J. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm," *Computer speech & language*, vol. 4, no. 1, pp. 35–56, 1990.

[18] B. Knudsen and J. Hein, "Rna secondary structure prediction using stochastic context-free grammars and evolutionary history." *Bioinformatics*, vol. 15, no. 6, pp. 446–454, 1999.

[19] ——, "Pfold: Rna secondary structure prediction using stochastic context-free grammars," *Nucleic acids research*, vol. 31, no. 13, pp. 3423–3428, 2003.

[20] E. Rivas and S. R. Eddy, "A dynamic programming algorithm for rna structure prediction including pseudoknots," *Journal of molecular biology*, vol. 285, no. 5, pp. 2053–2068, 1999.

[21] J. Reeder, P. Steffen, and R. Giegerich, "pknotsrg: Rna pseudoknot folding including near-optimal structures and sliding windows," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W320–W324, 2007.

[22] E. Bindewald, T. Kluth, and B. A. Shapiro, "Cylofold: secondary structure prediction including pseudoknots," *Nucleic acids research*, vol. 38, no. suppl 2, pp. W368–W372, 2010.

[23] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "Viennarna package 2.0," *Algorithms for Molecular Biology*, vol. 6, no. 1, p. 1, 2011.