Automated Computational RNA Structure Prediction

Samuel Jackson, University Of Aberystwyth

***Abstract*—The prediction of secondary and tertiary RNA structure from single or multiple sequences is a challenging open problem in bioinformatics. The conformational space of an RNA strand is both extremely large and complex. This article provides a review of automated computational approaches to RNA structure prediction (i.e. those which do not require human intervention). This paper begins with an introduction for non-biologists and outlines the core techniques used in predicting secondary and tertiary structure. Lastly, a summary of the general trends in the field and directions for future work are suggested.**

***Index Terms*—Bioinformatics, Computer Simulation, Dynamic Programming, Molecular Biophysics**

I. Introduction

Ribonucleic acid (RNA) is a diverse set of fundamental biological macromolecules that play an important part in many biological processes. An understanding of the structure of different types of RNA molecules can further our understanding of the molecular machinery at work inside our body as well as have direct practical applications to modern medicine by controlling protein synthesis, transcription, and virus replication [1].

The primary structure of nucleotides in a strand of RNA is fairly easy to obtain experimentally through rapid sequencing methods. Experimental determination of RNA secondary and tertiary RNA structure to an atomistic level of resolution is typically performed using either by X-ray crystallography or NMR spectroscopy. However secondary and tertiary structure determination has a bottleneck due to the diverseness of RNA molecules and the cost of experimental determination methods [2]. This leaves a high number of RNA with determined primary structure but little or no information about their the secondary or tertiary structure.

RNA structure prediction is the field of bioinformatics which aims to accurately predict the secondary and tertiary structure of RNA from the known primary structure. It provides a complementary method of analysis to experimental methods. However, this is a difficult problem to model due to the non-trivial structures that can be formed as an RNA molecule folds and binds with itself. Another limitation is the high number of potentially feasible but sub-optimum structures that inhabit the conformational space close to the true structure. Structure prediction also suffers from a “catch-22” scenario where the quality of predictions could potentially be made better if only more known structures could be determined. Most approaches, both for secondary and tertiary structure focus on predicting the structure which exhibits the minimum of free energy of the chemical structure, but this is still an imperfect simplification.

This paper provides a review the methods used to predict the secondary and tertiary structure of RNA molecules. Section II provides an overview of the chemical structure of RNA from the basic building blocks up to the full 3D structure. Section III discusses various methods applied to predicting the secondary structure of RNA. Section IV provides the complementary discussion prediction of the tertiary structure of RNA. Finally, A summary and discussion of future challenges is provided in section V.

II. RNA Structure

RNA molecules come in a wide variety of different shapes, sizes, and functions. One of the most important types of RNA is messenger RNA (mRNA) which is responsible for carrying information about the amino acid sequence required to build a protein from DNA strands in the cell nucleus to the ribosome for protein synthesis. However, many types of RNA are non-coding, i.e. they do not play a direct part carrying the coding information for protein translation. Instead they for part of the complex system of cellular machinery. An example of non-coding RNA is ribosomal RNA (rRNA) [3] which act as the catalysts in the ribosome for protein synthesis. Transfer RNA (tRNA) [4], [5] are yet another part of the translation process which delivers an amino acid to the ribosome according to a codon of mRNA. RNAs can also act in a regulatory role by down-regulating gene expression in the translation process. RNA molecules can also act as the genome for viruses, such as Ebola, HIV, and SARS which can utilise proteins to replicate themselves.

RNA is typically found as single stranded molecules in nature where parts of the of the strand is folded upon itself. This is in contrast to DNA which are typically composed of multiple strands which bond with one another. RNA folding, the process by which an RNA strand bonds with itself to form complicated 3D structures, can be classified into three categories: primary, secondary, and tertiary structure. The remainder of this section defines the different categories of structure and what they mean in terms of RNA folding.

A. Primary Structure

The primary structure refers to the sequence of building blocks of RNA. Each strand of RNA is made up of a sequence of organic chemicals called nucleotides. Nucleotides can be split into a three major components: a five carbon sugar ring (ribose), a phosphate group, and a nitrogenous base. The nitrogenous base can be one of four biological compounds: the nucleobases adenine, cytosine, guanine, and uracil, commonly abbreviated to the first letter of each molecule (A, C, G, and U). Nucleotides form the repeating units of a strand of RNA. A collection of nucleotides form a nucleic acid through a phosphodiester bond connecting the $5'$ carbon on the ribose ring of one nucleotide to the $3'$ carbon of the next. Typically as

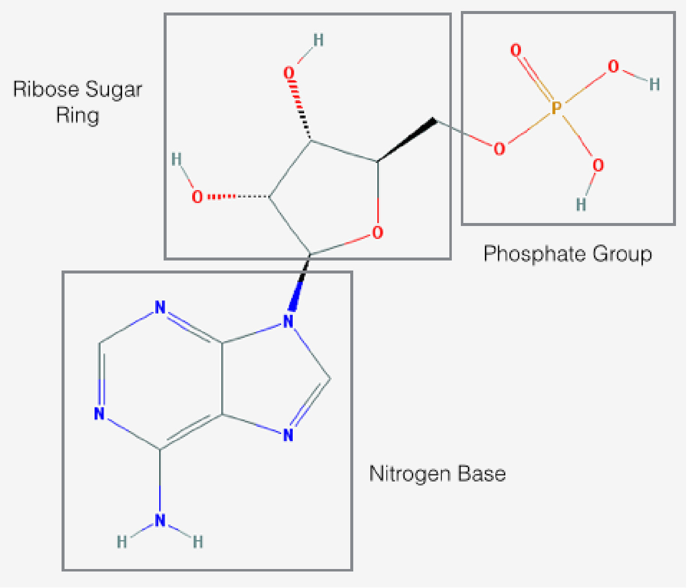


Fig. 1: 2D chemical structure of a nucleotide (adenosine 5′-monophosphate) with the adenine base (lower left), ribose sugar ring (middle) and phosphate group (upper right). 2D chemical structure source: PubChem [6].

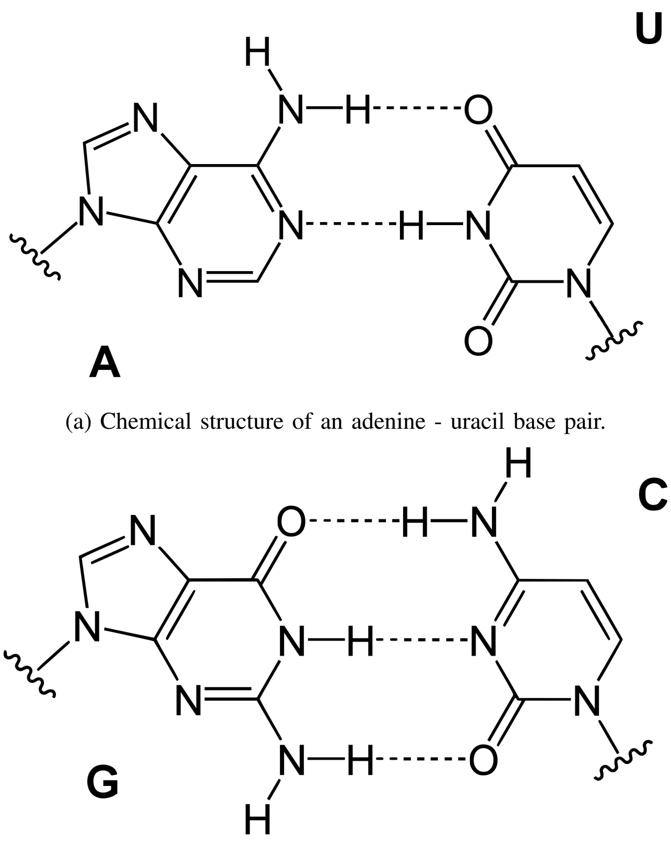
strand of RNA is described using the abbreviated bases from the 5′ carbon end of the sequence towards the 3′ carbon end.

B. Secondary Structure

Bases can be further split into two categories based on their chemical structure. Purine bases (A and G) contain a two nitrogen rings while pyrimidines (C and U) contain just a single ring. The differences in the structure of the nitrogen base allow hydrogen bonds to form between complementary pairs of nucleotides. The classic Watson-Crick base pairing scheme causes A to bond with U and C to bond with G. Other pairing schemes are possible such as Hoogsteen base pairs where the purine base is rotated and wobble base pairs where G-U base pairing is possible. The type of base pairing is important to structure prediction because different types of base pairs will have different thermodynamic stabilities associated with them which is a major factor in predicting how a molecule will fold.

The pattern of the structures that are created through base-pairing can be classified into a number of different structures. Reference [7] provides a comprehensive introduction. Commonly encountered patterns frequently found in RNA secondary structure 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 in a unpaired section of bases.
- **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** - A variation of pseudoknots, but directly between the unpaired nucleotides of two hairpin loops.



(a) Chemical structure of an adenine - uracil base pair.

(b) Chemical structure of a guanine - cytosine base pair.

Fig. 2: Watson-Crick base pairing in RNA. A-U and G-C form base pairs using two or three hydrogen bonds (shown as dashed lines) depending on the pair type.

Traditionally secondary RNA structure prediction has focussed on attempting to generate a combination of these structures based on the thermodynamic stability of the structure and the type of base pairs within the structure.

C. Tertiary Structure

While the secondary structure of an RNA molecule can reveal much about it's role, it is the tertiary structure that is the

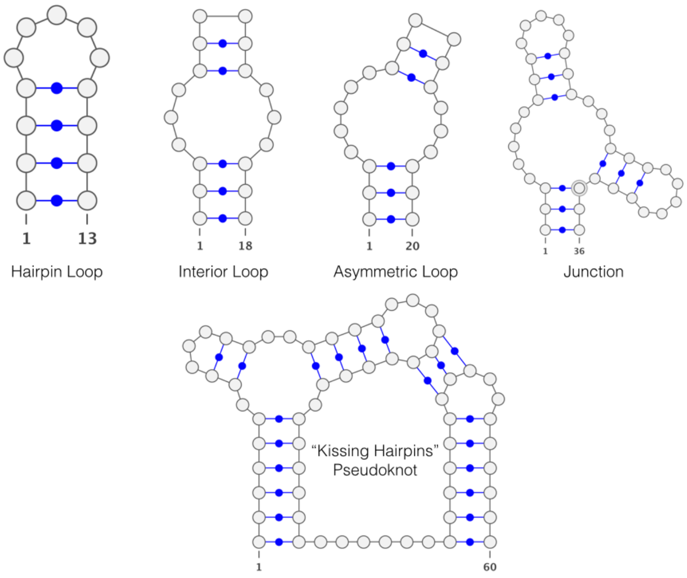


Fig. 3: Various types of motif present in the secondary structure of RNA molecules. The kissing hairpin pseudoknotted structure exhibits stems, hairpin loops, a junction (top left) and a pseudoknot (top right). All diagrams made with the aid of VARNA [8].

ultimate goal of structure prediction. 3D RNA folding is still very much an open problem in the bioinformatics community, but as will be shown in the following pages much progress has already been made. The additional spatial dimension brings a wide variety of potential conformations beyond those of two dimensions.

The hydrogen bonds connecting complementary base pairs are not rigidly fixed. Individual base pairs in an RNA strand can exhibit are variety of movements (see figure 4). The geometry of the base pairs can be described by a coordinate system with one axis parallel to the direction of the helical stem, one perpendicular along the direction of the bond, and one perpendicular to the bond itself:

- **Stretch, Shear, and Stagger** - Each base can be translated in all axes relative to one another.
- **Buckle** - The dihedral rotation along the hydrogen bond causing a kink to occur in the connection between the two bases.
- **Propeller Twist** - The dihedral rotation about the hydrogen bond connecting a pair of bases relative to the normal of each base. The bases appear to twist like a propeller or bow-tie.
- **Opening** - A scissoring effect rotating the bases about the helical axis.

The bonded base pairs can also move together relative to the helical axis. These movements can be one of:

- **X/Y Displacement** - Translation of the bonded bases in either the x or y directions relative to the z axis.
- **Inclination** - Given by the angle between the long axis and the helical axis.
- **Tip** - The complement to inclination. This is the angle between the short axis and the helical axis.

The geometry of two adjacent pairs of bases will also affect one another. This can be seen in the third part of diagram 4.

- **Shift, Slide, and Rise** - Displacement between two adjacent pairs of bases along one of the three coordinate axes.
- **Tilt, Roll, and Twist** - Rotation about the same axes as shift, slide, and rise respectively.

Beyond the flexibility of the bonded between base pairs the rest of the nucleotide can exhibit movement in all three dimensions. The ribose sugar ring in a single nucleotide exhibits sugar puckering which introduces non-planarity to the ring structure. This can be parameterised by 5 torsion angles for the bonds in the ring typically denoted by τ_i where i is the i^{th} torsion angle starting from the O4$'$ to C1$'$ bond. The glycosidic bond between the ribose sugar and the base can rotate and is typically parametrised by another torsion angle labelled χ. Finally, the backbone of an RNA molecule from the 5$'$ end connected to the phosphate group through the ribose sugar to the 3$'$ end of the molecule is parameterised by a set of 6 torsion angles over each bond (each shown in figure 5).

Despite the high number of parameters describing just a single nucleotide the space of potential realistic conformations is mercifully much smaller than the space of theoretical values for every parameter. For example, the values that the 6 backbone torsion angles are limited by the need to avoid steric clashes.

Folded RNA molecules also exhibit a variety of structural elements in three dimensions. These of course include all of structure types present in two dimensions and described in the previous section. Higher order tertiary structural components include:

- **Triplexes and Quadruplexes** - Single stranded portions of RNA can bond with a helical stack through hydrogen bonding with non-Watson-Crick pairs to form a triplex. Triplexes are a common feature in RNA, particularly at the junction between two helical stacks. Quadruplexes, where four nucleotides bond with one another, are also possible.
- **Coaxial stacking** - The formation of a more stable structure by the creation of a pseudoknot. This is where two single stranded regions bond with one another next to an existing stem.
- **Tetraloops** - Three dimensional hairpin loop structures capped with four loop nucleotides. These offer a chemically stable structure while offering a site for potential RNA-RNA or RNA-Protein interactions.

For a more detailed discussion on the conformations of nucleotides in three dimensions the reader is directed to references [1], [7], [9].

III. 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

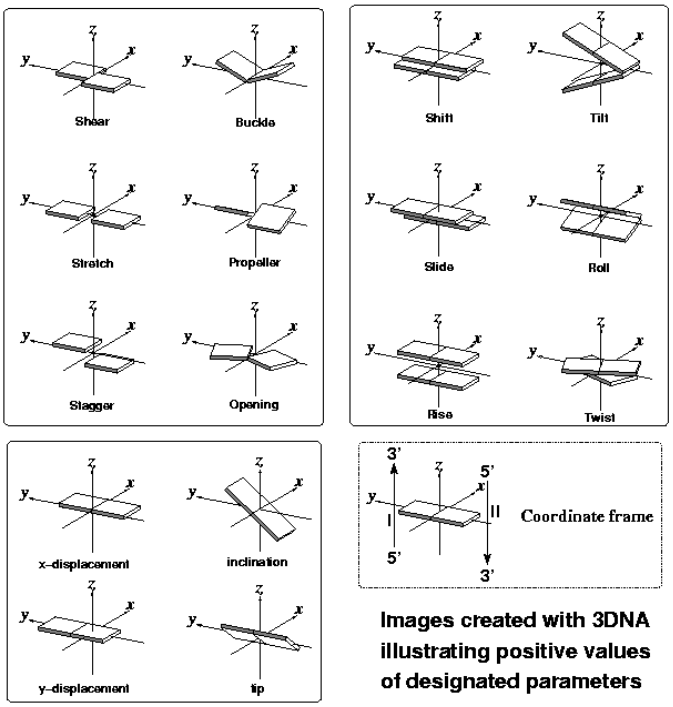


Fig. 4: Base pair for nucleotide interactions in three dimensions. Top left and bottom left: individual base pair movements relative to one another (top) and relative to the helix (bottom). Top right: possible interactions between two adjacent base pairs. Image source: Olsen et al. Copyright 2001 by Elsevier Inc. [10].

folding share deep similarities in how structure is determined. The major differences are in the scoring schemes and the parametrisation used. From a computer science perspective the SCFG formulation is perhaps the most interpretable. Before diving into secondary structure prediction methods it is useful to define context free grammars (CFGs) and therefore SCFGs.

According to Giegerich [11] 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 [12] which expresses the Nussinov method [13] of RNA folding discussed previously is

$$S \rightarrow Sa|SaS\hat{a}|\epsilon \qquad (1)$$

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 1 which is unambiguous).

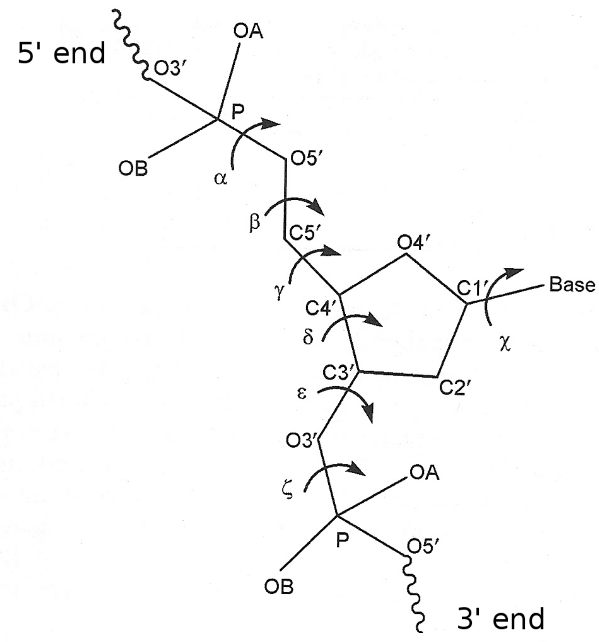


Fig. 5: Nucleic acid backbone torsion angles for a single nucleotide denoted by α, β, γ, δ, ϵ, and ζ. The torsion angle on glycosidic bond between the ribose sugar and base is also show as the parameter χ. Image source: Stephen Neidle Copyright 2008 by Elsevier Inc. [9].

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 [11], [14]–[16] . The CYK algorithm used for efficiently evaluating a SCFG is essentially the same as that which is used for finding the MFE [17]. 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 [18]. 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 [19] 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.

The rest of this section provides an overview of various secondary structure prediction methods starting from early thermodynamic models and working forwards roughly chronologically to more recent probabilistic models.

A. Thermodynamic Models

One of the earliest influential approaches to secondary structure prediction is the Nussinov algorithm [13]. 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 an 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 [19]. 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 [17]. The underlying algorithm shares a very similar formulation to 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 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 [29]. Many more recent prediction algorithms utilise the thermodynamic parameters used by Turner's group [23] as opposed to the weights used in the original paper.

Furthermore, this still shares some of the limitations of [13]. 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 [36] [12].

Moving forward in time, another key contribution to the area was the equilibrium partition function formulation by McCaskill [19]. The aim of this paper was 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

$$Q = \sum_{s} e^{-(E(A)/kT)} \tag{2}$$

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

$$P(A|S) = \frac{1}{Q} e^{-(E(A)/kT)} \tag{3}$$

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

$$P((i,j)|S) = Q_{ij}/Q \tag{4}$$

where

$$Q_{ij} = \sum_{(i,j)\in A} e^{-(E(A)/kT)} \tag{5}$$

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

TABLE I: Summary of Approaches to Secondary Structure Prediction

Paper	Year	Name	Criteria	Parameterisation	Description
Nussinov & Jacobson [13]	1980	Nussinov Algorithm	Maximum Base Pairing	Binary	Dynamic programming algorithm for computing base pairing
Zuker & Stieglar [17]	1981	Mfold	MFE	Thermodynamic	Minimum free energy (MFE) algorithm
McCaskill [19]	1990	-	MFE	Thermodynamic	Partition function, base pair probability matrix, melting behaviour description.
Hofacker et al. [20]	1994	RNAfold	MFE	Thermodynamic	Parallelised algorithm, sequence comparison tools, inverse folding tools.
Shapiro & Wu [21], [22]	1997	MPGAfold	MFE	Thermodynamic	Parallelised Genetic algorithm. Handles pseudoknots.
Matthews et al. [23]	1999	Mfold	MFE	Thermodynamic	More advanced thermodynamic parameters.
Knudsen & Hein [24], [25]	1999	Pfold	MAP estimate	Probabilistic	Multiple sequence method based on using SCFGs and a MAP estimate for the optimum structure.
Ding et al. [26], [27]	2005	Sfold	Cluster centroids	Thermodynamic	Clustering of statistical sampling from Boltzmann ensemble.
Cao & Chen [28]	2005	Vfold	MFE	Thermodynamic	Coarse-grained model. Estimation of the sequence dependance of loop free energy.
Do et al. [29]	2006	CONTRAfold	MEA	Probabilistic	A highly configurable CLLM which proposed using MEA for the optimum structure.
Bindewald & Shapiro [30]	2006	KNetFold	MFE	Probabilistic	Comparative method using a mutual information measure.
Reeder et al. [31]	2006	pknotsRG	MFE	Thermodynamic	Pseudoknot prediction using MFE terms.
Deigan et al. [32]	2009	RNAstructure	MFE	Thermodynamic	SHAPE chemistry contraints.
Hamada et al. [33], [34]	2009	Centroidfold	γ-centroid	Probabilistic	Centroid based estimator. Can utilise probabilities from either CONTRAfold or McCaskill partition function
Bindewald et al. [35]	2010	Cylofold	MFE	Thermodynamic	Coarse-grained model. Prediction of pseudoknots by simulated stem folding.

change in temperature. This provides another window into the structural properties of a RNA sequence.

Both McCaskill's partition function method and Zucker and Stieglar's MFE method remain to this day as the bedrock of many successful approaches to RNA structure prediction. Two notable extensions of these works which should be mentioned for completeness are the papers by Matthews et al. [23] (often referred to as the Turner group) in producing Mfold and Hofacker et al. [20] in producing the RNAfold and ViennaRNA packages.

Matthews et al. improved on the set of thermodynamic parameters by extrapolating free energies obtained from analysis of representative molecules for loop structures and comparative sequence analysis for stability of tetraloops and estimate junction initiation parameters. Hofacker et al. contribute a collective package (ViennaRNA) that incorporates not only RNAfold, a parallelised MFE algorithm based on Zucker's, but also tools for inverse folding and comparison of secondary structures. Mfold in particular is often used as a baseline for secondary structure prediction. These methods are also often used in the calculation of secondary structure for the prediction of tertiary RNA structure [36]–[38].

A more modern free energy minimisation approach was created by Deigan et al. [32] 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

$$\Delta G_{SHAPE}(i) = m \cdot ln[SHAPEreactivity(i) + 1] + b \quad (6)$$

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. They also 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. [26], [27]. Their work can largely been seen as a logical extension of McCaskill's work in [19]. Their first paper [26] 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 [27] 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 [26]. The samples are then clustered using a top-down method from [39]. 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.

Cao and Chen [28] created the Vfold model. They represent each nucleotide using a coarse-grained model where the seven torsion angles in a nucleotide are replaced by a simplified 3-vector representation. They generate all possible base stacks in a sequence and use partition function similar to McCaskill's to obtain the probability of a structure. A key difference is that their model is able to estimate, to some extent, the sequence dependance of loop free energy which is unobtainable from experimental data. They achieve this estimation by explicit enumeration of all possible conformations by the log ratio of the frequency of loop to coil conformations multiplied by the Boltzmann constant.

Using a partition function developed to support their coarse grained representation of nucleotides and an algorithm based on Mfold [23] they find the structure with the lowest free energy. From the distribution of possible structures they can compute the free energy landscape for a structure containing n native base pairs and m non-native base pairs. The minima of the landscape then represents the stablest states.

B. Probabilistic Models

A notable early attempt at RNA secondary structure prediction using SCFGs is the work of Knudsen and Hein [24], [25] in producing Pfold. In [24] 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 [24] 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

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

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 refined their work in [25] 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 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. [29] 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

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

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 [23].

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

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

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.

Bindewald and Shapiro [30] created KNetFold which uses an entropy based measure that captures the mutual information between two aligned columns and a hierarchical network of k-nearest neighbour classifiers to infer secondary structure.

Their mutual information measure is the difference between information in aligned columns R_i with the information of a column pair R_{ij}. The formula for the individual information in a column i is

$$R_i = H_g(i) + \Sigma_{k=1}^{4} P_k(i) log_2 P_k(i) \tag{11}$$

And the formula for two columns (i, j) is

$$R_{ij} = H_g(i, j) + \Sigma_{k=1}^{16} P_k(i, j) log_2 P_k(i, j) \tag{12}$$

where $P_k(i)$ is approximated by the ratio of observed characters k and the number of characters in the column. H_g is the expected uncertainty of the alignment in column i. Given enough sequences this term will approach the number of bits needed to represent the alphabet of characters multiplied by the number of columns considered (i.e. 2 bits for one column, 4 bits for a pair of columns) but can be approximated to correct for sampling noise for a low number of sequences.

Feature vectors formed from a combination of the mutual information of aligned two columns, and the fraction of pairing nucleotides using the four nearest neighbour columns, both diagonally and anti-diagonally. Nine Gaussian weighted k-nearest neighbour classifiers are built using the AdaBoost algorithm to handle the dimensionality of the feature space. Subsequent layers in the network reduce the number of classifiers by a factor of 3 and inputs are randomly chosen from the previous level.

Finally, one last classifier is created which takes the input from the single classifier from the previous level along with a thermodynamic consensus matrix. This matrix is created by calculating MFE for each sequence, the elements for which are then aligned, averaged, and weighted proportionally to the number of nonzero prediction probabilities for the given element.

The authors report that for a very low number of aligned sequences (5) the predictions made are almost entirely based on the consensus matrix, but for a larger number of sequences their method enhanced those predicted from the consensus matrix. Notably there method was also able to successfully predict two pseudoknot interactions in a test sequence.

Hamada et al. [33], [34] produced CENTROIDfold, another comparative analysis method that can either use the output of CONTRAfold or the McCaskill probability matrix. The main contribution of the paper is a novel γ-centroid estimator which attempts to maximise the expected number of base pairs in opposition the the maximum likelihood estimate provided by MFE approaches.

The γ-centroid measure is defined for a single sequence:

$$G_{\gamma}(\sigma, y) = \gamma TP + TN \tag{13}$$

Where γ is a trade off between the sensitivity and selectivity of the algorithm. From this their method maximises the quantity

$$\hat{y} = \arg\max_{y} \sum_{\sigma \in \Omega(x)} G_{\gamma}(\sigma, y) p(\sigma|x) \tag{14}$$

They also provided several further derivations that both prove the validity of their approach and shows a generalisation of equation 14 for multiply aligned sequences.

They propose that their method also has benefits over the MEA estimator. The MEA estimator maximises the expected accuracy with respect to each base, while the γ-centroid measures the expected accuracy with respect to each base pair. The authors note that this measure has then benefit that the best base pairs are supported by evidence provided by the many sub-optimal structures found in the distribution of potential structures rather than relying on very weak probabilities for many near optimal candidates in given by the conventional MFE/probabilistic models.

They demonstrate through their experiments that the estimator performs better than prediction by conventional MFE/ML estimates and also show improvement over the MEA method of CONTRAfold [29].

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 [40] 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 [41], [42]. In this section two methods which tackle the pseudoknot prediction from different paradigms are presented.

Shapiro and Wu [21] implemented support for predicting basic types of pseudoknots using a massively parallel genetic algorithm previously created by Shapiro [22]. Their algorithm is carried out on a super computer consisting on 16,384 cores each representing a single candidate solution to the RNA folding problem.

In the original GA paper [22] the algorithm used stem list which contained all maximally sized stems. Each stem is represented by its start and stop position along with its size and thermodynamic energy parameter. A region list is maintained by each core which contains a sorted list of all stems currently in the structure. The fitness function used was the negative of the free energy associated with each structure. Each processor is initialised by randomly choosing from the stem list. Selection is achieved from sampling from the logical local neighbourhood of adjacent processor cores with toroidal wrap around and taking the top two as parents. Uniform crossover of stems between parents is used and stems are only accepted if there is no conflict. Mutation is achieved by randomly selecting stems from stem table and adding them to the region table.

In the later paper [21] they make several adjustments to this approach. Firstly they implemented a new mutation annealing operator where the probability of mutation decreases proportionally to the size of the stem. Secondly they introduced a second stem list called the "pseudoknot stem list". Like regular stems, pseudoknots have an energy term associated with them. At each iteration, after the initial structure is formed, possible pseudoknotted structures are added by traversing the structure and computing the free energy terms.

Reeder et al. [31] produced a method that was largely based on the MFE method by Matthew's et al. [23] 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. [35]. 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 [43]).

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 determined to be a failure. Likewise if after optimisation 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 [31]. The noted benefits of this method are twofold: 1) it avoids some of the simplifying constraints associated with approaches like pknotsRG (but possibly does not solve more complex pseudoknots), and 2) provides an automatic check for steric feasibility.

IV. RNA TERTIARY STRUCTURE PREDICTION

While determining the secondary structure of RNA molecules provides valuable insights into their properties, the true goal of RNA folding is the determination of the tertiary structure of the molecule. Tertiary structure not only incorporates the secondary structure but also adds important long range interactions between bases and shows how it contorts in three dimensions. This provides key insights not only to the molecule's final structure, but also its biological function. Like secondary structure, tertiary structure prediction is difficult due to the extraordinary size of the conformational space from which to choose potential structures.

Broadly speaking the computational approaches to the prediction of the tertiary structure of RNA can be broken into two categories: those based on fragment assembly and those based on folding simulation, although there is some overlap between the two. This section covers several different methods for predicting the tertiary structure of a molecule from both paradigms.

Fragment assembly approaches the problem by attempting to combine subsections of several known portions of RNA sequences into the correct final structure. This typically employs a database of previously determined RNA structure motifs from which to choose from. The resulting structures can be further refined with a folding simulation. The probing of potential structures in fragment assembly can either be carried out as a constrain satisfaction problem (such as with MC-fold/MC-Sym [44]) or by Monte Carlo sampling of candidate structures by the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm generates new candidate solutions by randomly modifying the existing structure by moving or rotating one or more atoms. The energy of the new structure generated by the modification is calculated and the structure is either accepted or rejected with probability according to the Metropolis criterion:

$$f(\Delta E) = \begin{cases} 1 & \text{if } \Delta E \leq 0 \\ e^{-\frac{\Delta E}{kT}} & \text{if } \Delta E > 0 \end{cases} \tag{15}$$

where ΔE is the change in energy, T is the temperature and k is the Boltzmann constant. The algorithm is often combined with simulated annealing of the temperature parameter as sampling progresses to encourage convergence towards the global optimum.

Methods using folding simulations attempt to simulate how chains of nucleotides interact in order to fold into the correct structure using physics based potential functions. In contrast with fragment assembly based approaches the conformational space is typically sampled using molecular dynamics. Molecular dynamics approaches represent the molecular structure of an RNA as a three dimensional model parameterised by similarly to the definitions given in section II (although usually with simplifications). The positions of atoms in the system are updated over time according to potential functions that define the forces acting on both bonded and non-bonded portions of the molecule. For short range non-bonded interactions the Lennard-Jones potential is common which is highly repulsive over a small distance and has an attractive well as the distance increases:

$$V_{lj} = 4\epsilon\left[\left(\frac{\sigma}{r}^{12}\right) - \left(\frac{\sigma}{r}^{6}\right)\right] \tag{16}$$

where r is the distance between two given particles and σ and ϵ are parameters governing the distance at which the potential is zero and the depth of the well respectively. Bonded interactions between atoms in a nucleotide are typically represented as a linear combination of different potential energy terms representing the parameters for the bonded distance, angle, and torsion angle with more complicated terms and interactions can be defined if necessary:

$$E_{bonded} = E_{distance} + E_{angle} + E_{torsion} \tag{17}$$

Sets of parameters governing the individual contributions of each energy term are application dependant and are typically empirically based.

A. *Fragment Assembly*

Ras and Baker [45] created FARNA (Fragment Assembly of RNA). FARNA is a fragment assembly method which is *de novo* in its approach and does not utilise experimental data or precomputed secondary structures as input. FARNA represents an RNA as a collection of trinucleotides with seven corresponding torsion angles and a pucker amplitude. For simplicity the authors only differentiate between pyrimidine and purine bases instead of using separate definitions for each base type. Sample fragments are taken from a single large example of RNA structure represented by *Haloarcula marismortui* [50]. Each fragment has a corresponding energy potential specifically designed for RNA and imposed over the centroid of heavy atoms of the base. The potential function favours compactness in the resulting structure but heavily penalises steric clashes. Their potential also includes terms which enforce coplanar base pairing.

To predict the structure of the RNA sequence fragments are drawn using the Metropolis-Hastings algorithm. Each draw chooses a random position in the molecule and replaces parameters of the segment with the parameters of a random segment. After the initial burn-in, the fragments are accepted according to the Metropolis criterion. Terms weighting coplanarity in the energy function are slowly stepped up over the course of the simulation.

The authors note in their conclusions that the major limitations of their method are the MC algorithm used to sample the conformational space and the potential function. The work in [45] is limited to sequences containing less than 40 nucleotides. They propose that incorporating secondary structure information may lead to performance gains for longer sequences. More importantly they state that the limiting factor of the model for short sequences is the accuracy of their potential function. The authors later extended this work to produce FARFAR (fragment assembly of RNA with full-atom refinement) [51]. This work aimed to correct inaccurate ranking by the low resolution potential by performing a full-atom molecular dynamics simulation. In this work they observed that the accuracy of the refinement dropped relative to the length of the sequence used. Sequences which did not converge to the native structure were chalked up to poor conformational sampling.

Parisien and Major [44] created the MC-Fold/MC-Sym pipeline for secondary and tertiary structure prediction. In their method the MC-Fold program first generates a collection of sub-optimal secondary structures by combining multiple predefined structural motifs, referred to in the paper as nucleotide cyclic motifs (NCMs). All possible NCMs are enumerated but many potential structures are discarded as infeasible. MC-Fold uses traditional free energy minimisation and a scoring function determine the most likely secondary structures.

The predicted secondary structure is then used as input for the tertiary structure prediction program MC-Sym. MC-Sym uses a predetermined 3D library of motifs using the same premise as secondary structure determination. As the enumeration of all possible 3D fragments is not computationally feasible a Las Vegas algorithm is used to sample potential motifs. The difference between a Las Vegas algorithm and a traditional Monte Carlo algorithm is that a Las Vegas algorithm always produces a valid candidate structure. Each 3D fragment is represented as a full atom model. The fragments are added such that they optimise the score generated during secondary structure determination.

Cao and Chen [48] modified their secondary structure prediction tool Vfold [28] (reviewed in section III-A) to predict tertiary structure using a combination of fragment assembly and molecular dynamical simulation. Vfold first predicts the secondary structure using a coarse grained model of RNA nucleotides. Using the secondary structure they search through a database of tertiary structural motifs for matches closely related to fragments of the 2D structure classified into hairpins, bulges, junctions etc. Based on the coarse grained 3D model built from motif fragments a full-atom model is created and then refined using energy minimisation in the AMBER [52] molecular dynamics program.

B. *Folding Simulations*

An notable attempt at using molecular dynamics for RNA structure prediction is the nucleic acid simulation tool (NAST) [47]. NAST uses a coarse-grained representation of RNA nucleotides by approximating individual atoms with a single pseudo-atom in the location of the central C3$'$ atom. The NAST energy function makes the assumption that the geometry of the non-bonded regions in the secondary structure will follow a distribution closely following known RNA structures.

TABLE II: Summary of Automated Tertiary Structure Prediction Methods

Paper	Year	Name	Type	Representation	Description
Ras and Baker [45]	2007	FARNA/ FARFAR	Folding Simulation & Fragment Assembly	Full atom	Coarse grained fragment assembly refined in full atom MD simulation.
Ding et al. [46]	2008	iFoldRNA	Folding Simulation	Coarse-grained. Three pseudo-atoms. ”Bead on a string” model.	MD simulation with discrete potentials.
Parisien and Major [44]	2008	MC-fold/MC-Sym	Fragment Assembly	Full atom	Uses constraint satisfaction to choose fragments
Jonikas et al. [47]	2009	NAST	Folding Simulation	Coarse-grained. Single pseudo-atom	Pure MD. Energy functions derived from distributions of known structures.
Cao and Chen [48]	2011	Vfold	Folding Simulation & Fragment Assembly	Coarse-grained. Full atom refinement.	Combination of both Fragment assembly and MD refinement with AMBER.
Rother et al. [49]	2012	SimRNA	Folding Simulation & Fragment Assembly	Coarse-grained. Three pseudo-atoms. Full atom refinement.	Uses Monte Carlo for simulation. A fully atomistic reconstruction is then generated from a database and refined via conventional MD.

Based on this assumption geometries between 2, 3, and 4 sequential nucleotides in known structures are used to create probability distributions for the angle, distance, and dihedral parameters. The energy function (E) is then derived based on the Boltzmann relationship:

$$E(x) = -RTlnP(x) \tag{18}$$

Non-bonded interactions are modelled using the classic Lennard-Jones potential to prevent steric overlap for nucleotides separated by a distance greater than three. The geometry of the model is further constrained using the known secondary structure. Helices are constrained using one distance, one angle, and two dihedral parameters to help fix the model to the ideal helical shape. Long range tertiary interactions are modelled using and additional term in the energy potential the strength of which is determined by the data source (known crystal structure: strong, experimental data: weak).

Ding et al. [46] took a discrete molecular dynamics (DMD) approach to tertiary RNA structure modelling and created the iFoldRNA web server. In their method they approximate an RNA molecule using a “bead on a string” model where the backbone of the nucleotide is represented as simple collection of sugar, phosphate, or base molecules. This gives the final nucleotide model three “beads” attached to a “thread” of covalent bonds. Angular and dihedral constraints are also included in the model.

The mechanics of the model differ from traditional molecular dynamics models by using only discrete functions as potentials in the simulation. This has the advantages that computation of an atom’s velocity does not need to be recomputed at every time step. Only when the molecule jumps in relation to an interaction with a potential does the atom’s velocity get kinematically updated. The method in [46] used discrete potentials incorporating phosphate-phosphate repulsion, hydrophobic interactions, and base stacking interactions. The free energy of loop structures used in [23] are also included to push the algorithm towards more compact, less loopy structures. The loop energy change associated with a bond forming is estimated using the metropolis-hastings algorithm.

Another folding simulation based on using Monte Carlo sampling in favour of molecular dynamics to sample the conformational space is SimRNA [49]. SimRNA begins with a coarse-grained representation with three atom per nucleotide. One each for the phosphate group, C4$'$ atom, and a nitrogen atom for the base. The energy function used by SimRNA is similar to equation 18 with $P(x)$ defined as the ratio of the observed frequency of a parameter value over the expected value assuming an unbiased distribution.

Short range interactions are represented as virtual bonds with parameters for the distance along the backbone, flat angles, and torsion angles of the nucleotide. Their energy contribution is a simple linear combination of each of the individual terms. Long range interactions are captured from a representative sample of known RNA structures by measuring the spatial neighbours of a nucleotide subunit, computing the occurrence of the nitrogen atom for each combination of base types, mapping the spatial distribution to a grid and binning via a 3D histogram.

Using the defined energy function, samples are taken from the conformational space using the Metropolis-Hastings algorithm. The sampling algorithm is combined with simulated annealing to gradually reduce the temperature to help convergence. Random modification to the positions and rotations of atoms in the model were used to generate new samples. Some additional complex modifications (such as the simultaneous movement of two backbone atoms) were included to help speed up algorithm progress. Moves are applied with a probability derived from the relative mobility of atoms in a nucleotide. Final reconstruction of an RNA molecule from the reduced representation is done by comparing each coarse nucleotide with a database of known RNA fragments. Final full-atom refinement of the structure can be carried out using the same Monte Carlo procedure, but including additional Lennard-Jones (equation 16) and hydrogen bonding potential terms.

V. Discussion

RNA structure prediction is a challenging problem to undertake. The space of potential candidate solutions is both large and contains many sub-optimal minima. But despite the limitations, much progress has, and continues, to be made in the field. The increase in computational power since the first attempts at prediction were made has certainly aided progress. Perhaps more importantly is the interest in the field that has developed as our understanding of the importance of RNA as a biological entity has increased. The diversity of different proposed solutions, both for 2D and 3D structure, is immense. The progression of refinement of methods from a simple mathematical formulation of the base pairing problem to complex fragment assembly and folding simulation techniques only serves as encouragement for the progress that can still be made in the field. Most notably is the revelation that the rules governing folding of an RNA by MFE are essentially just a special case formulation of a SCFG.

There is however, much work still to be done in the field. There are many hurdles that need to be made in order to bring RNA structure prediction to the same standard are protein structure prediction methods. In the most broad terms, the core issues facing both secondary and tertiary structure prediction in the broadest terms are: the lack of experimentally confirmed data, the inherent inaccuracy present in the models used.

As mentioned in the introduction of this review, one of the reasons why the RNA structure prediction is so desirable is that the experimental determination of the structures to atomic resolution is a slow and costly processes (at least compared to computerised analysis). However, some of the best prediction systems rely on the availability of known RNA structures. Attempting to predict the final structure of RNA from just its primary structure potentially disregards a wealth of information present in other similar sequences. Perhaps there is a "critical mass" of representative experimentally confirmed structures required in order to provide a successful database for prediction techniques to be based upon. It certainly appears to be a trend across all of the papers reviewed in the article towards using known structures, or fragments of known structures to infer structure in an undetermined sequence over methods which simulate folding from first principles. Having a wide availability of diverse, known structures would bring the RNA structure prediction problem more into the sights of traditional supervised machine learning approaches which have already had great successes in other fields of bioinformatics. In such a context much focus would have to be put on achieving a high accuracy output due to the many sub-optimal structures that closely inhabit the low energy conformational space.

This issue ties in with the major next problem facing progress with RNA structure prediction which is the model used to find the optimum structure. For the longest time many of the most popular secondary structure prediction have centred around finding the structure that is thermodynamically the most stable. However, as multiple authors have noted, this technique is limited by the accuracy of the parameters used. Exactly what combination parameters are used is also a continuing topic of discussion. There is also the fact that minimum free energy structure is not necessarily the correct one. More recently secondary structure approaches have branched out in statistical techniques such as finding the maximum expected accuracy structure which has been suggested to perform better than the MFE/CYK formulation. Still, as noted by Rivas [12] even the best methods with complicated architectures only achieve around 60% for the F-measure. Several authors noted a block in prediction accuracy which has yet to be overcome. Tertiary structure prediction is also usually centred around contorting the structure into the most energetically favourable conformation predicted by thermodynamics; either through fragment assembly, folding simulations, or through sampling. This combined with the fact that the conformational landscape is usually highly rugged around the true structure perhaps that better models could be proposed.

The question then is where to go from here? The existing models obviously provide a good base, but future work could be directed towards including additional information such as the effects of long range interactions between nucleotides and the effect of RNA-protein or RNA-RNA contacts which are likely to affect the final structure. Successful attempts to include additional experimental data such as SHAPE chemistry have also shown to be successful. Indeed, any technique that supply additional information and can reduce the size of the conformational space and constrain the bounds of the solution is likely to be useful.

Another general trend across the papers reviewed is that models for tertiary structure have drifted towards a more coarse-grained approach. This appears to be partly due to necessity and partly to a better understanding of what the major contributing factors for folding are. This both speeds up slow predictions and allows for larger chains of nucleotides to be folded. However this simplification can often negatively impact the results. There is a trade off between using an accuracy model and speed. Sadly the computational intensity of the scales too fast for even coarse grained models to be quick enough at larger strands lengths. There also seems to have been a trend towards fragment assembly based approaches and a fusion of this with folding simulations. Progress in tertiary structure prediction is most likely going to be orientated towards better databases of known structural motifs combined with some form of coarse or full-atom refinement similar to FARNA. Further developments to refine the energy potential used to achieve a balance of complexity versus efficiency will no doubt lead to RNA structure prediction reaching full maturity.

Acknowledgments

The author would like to thank Edel Sherratt for her guidance on writing and for her insightful discussion on the topic of CFGs. Additional thanks should be given to Hannah Robarts for her perspective on papers using thermodynamic approaches and to Sam Nicholls for his general bioinformatics knowledge.

References

[1] T. Schlick, *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*. Springer Science & Business Media, 2010, vol. 21.

[2] S. Ya-Zhou, W. Yuan-Yan, W. Feng-Hua, and T. Zhi-Jie, "RNA structure prediction: Progress and perspective," *Chinese Physics B*, vol. 23, no. 7, p. 078701, 2014.

[3] M. M. Yusupov, G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. Cate, and H. F. Noller, "Crystal structure of the ribosome at 5.5 å resolution," *science*, vol. 292, no. 5518, pp. 883–896, 2001.

[4] J. L. Sussman, S. R. Holbrook, R. W. Warrant, G. M. Church, and S.-H. Kim, "Crystal structure of yeast phenylalanine transfer RNA: I. crystallographic refinement," *Journal of molecular biology*, vol. 123, no. 4, pp. 607–630, 1978.

[5] S. R. Holbrook, J. L. Sussman, R. W. Warrant, and S.-H. Kim, "Crystal structure of yeast phenylalanine transfer RNA: Ii. structural features and functional implications," *Journal of molecular biology*, vol. 123, no. 4, pp. 631–660, 1978.

[6] "National Center for Biotechnology Information," https://pubchem.ncbi.nlm.nih.gov/compound/6083, PubChem Compound Database; CID=6083, Accessed May 17, 2016.

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

[8] K. Darty, A. Denise, and Y. Ponty, "VARNA: Interactive drawing and editing of the rna secondary structure," *Bioinformatics*, vol. 25, no. 15, pp. 1974–1975, 2009.

[9] S. Neidle, *Principles of nucleic acid structure*. Academic Press, 2010.

[10] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.-S. Tung, E. Westhof, C. Wolberger, and H. M. Berman, "A standard reference frame for the description of nucleic acid base-pair geometry12," *Journal of Molecular Biology*, vol. 313, no. 1, pp. 229 – 237, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283601949873

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

[12] 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.

[13] 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.

[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 syntax analysis algorithm for context-free languages." DTIC Document, Tech. Rep., 1965.

[17] 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.

[18] 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.

[19] 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.

[20] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatshefte für Chemie/Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1994.

[21] B. A. Shapiro and J. C. Wu, "Predicting RNA h-type pseudoknots with the massively parallel genetic algorithm," *Computer applications in the biosciences: CABIOS*, vol. 13, no. 4, pp. 459–471, 1997.

[22] B. A. Shapiro and J. Navetta, "A massively parallel genetic algorithm for RNA secondary structure prediction," *The Journal of Supercomputing*, vol. 8, no. 3, pp. 195–207, 1994.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] S. Cao and S.-J. Chen, "Predicting RNA folding thermodynamics with a reduced chain representation model," *RNA*, vol. 11, no. 12, pp. 1884–1897, 2005.

[29] 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.

[30] E. Bindewald and B. A. Shapiro, "Rna secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers," *RNA*, vol. 12, no. 3, pp. 342–352, 2006.

[31] 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.

[32] 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.

[33] M. Hamada, H. Kiryu, K. Sato, T. Mituyama, and K. Asai, "Prediction of RNA secondary structure using generalized centroid estimators," *Bioinformatics*, vol. 25, no. 4, pp. 465–473, 2009.

[34] K. Sato, M. Hamada, K. Asai, and T. Mituyama, "CENTROIDFOLD: a web server for rna secondary structure prediction," *Nucleic acids research*, vol. 37, no. suppl 2, pp. W277–W280, 2009.

[35] 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.

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

[37] J. A. Cruz, M.-F. Blanchet, M. Boniecki, J. M. Bujnicki, S.-J. Chen, S. Cao, R. Das, F. Ding, N. V. Dokholyan, S. C. Flores *et al.*, "RNA-puzzles: a casp-like evaluation of RNA three-dimensional structure prediction," *RNA*, vol. 18, no. 4, pp. 610–625, 2012.

[38] Z. Miao, R. W. Adamiak, M.-F. Blanchet, M. Boniecki, J. M. Bujnicki, S.-J. Chen, C. Cheng, G. Chojnowski, F.-C. Chou, P. Cordero *et al.*, "RNA-puzzles round ii: assessment of rna structure prediction programs applied to three large rna structures," *RNA*, 2015.

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

[40] 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.

[41] J.-L. Chen and C. W. Greider, "Functional analysis of the pseudoknot structure in human telomerase RNA," *Proceedings of the National Academy of Sciences of the United States of America*, pp. 8080–8085, 2005.

[42] C. W. Pleij, K. Rietveld, and L. Bosch, "A new principle of RNA folding based on pseudoknotting," *Nucleic Acids Research*, vol. 13, no. 5, pp. 1717–1731, 1985.

[43] 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.

[44] M. Parisien and F. Major, "The MC-Fold and MC-Sym pipeline infers rna structure from sequence data," *Nature*, vol. 452, no. 7183, pp. 51–55, 2008.

[45] R. Das and D. Baker, "Automated de novo prediction of native-like RNA tertiary structures," *Proceedings of the National Academy of Sciences*, vol. 104, no. 37, pp. 14 664–14 669, 2007.

[46] F. Ding, S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude, and N. V. Dokholyan, "Ab initio rna folding by discrete molecular dynamics: from structure prediction to folding mechanisms," *RNA*, vol. 14, no. 6, pp. 1164–1173, 2008.

[47] M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. B. Altman, "Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters," *RNA*, vol. 15, no. 2, pp. 189–199, 2009.

[48] S. Cao and S.-J. Chen, "Physics-based de novo prediction of rna 3d structures," *The Journal of Physical Chemistry B*, vol. 115, no. 14, pp. 4216–4226, 2011.

[49] K. Rother, M. Rother, M. Boniecki, T. Puton, K. Tomala, P. Łukasz, and J. M. Bujnicki, "Template-based and template-free modeling of RNA 3d structure: Inspirations from protein structure modeling," in *RNA 3D structure analysis and prediction*. Springer, 2012, pp. 67–90.

[50] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz, "The complete atomic structure of the large ribosomal subunit at 2.4 å resolution," *Science*, vol. 289, no. 5481, pp. 905–920, 2000.

[51] R. Das, J. Karanicolas, and D. Baker, "Atomic accuracy in predicting and designing noncanonical RNA structure," *Nature methods*, vol. 7, no. 4, pp. 291–294, 2010.

[52] D. A. Case, T. Darden, T. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. Walker, W. Zhang, K. Merz *et al.*, “AMBER 11,” University of California, Tech. Rep., 2010.