Algorithmic Approaches to Modelling and Predicting RNA Structure

Literature Review

Carlos Gonzalez Oliver
Supervisor: Jérôme Waldispühl

August 27, 2017

Abstract

Ribonucleic acid (RNA) is a chain-like molecule akin to a string of characters which is found ubiquitously in every living cell. RNA effects crucial functions as an information carrier like DNA, and in catalyzing reactions and regulating interactions like proteins. A defining characteristic of RNA function is that it is determined by the spatial organization (2D and 3D) of its polymeric chain which in turn is encoded in its sequence. This property gave rise to three major problems in the way of understanding and manipulating RNA function: RNA structure modelling, prediction, and classification. For the past four decades that the field has been active, efficient and expressive mathematical solutions to these problems have led to major advances in the biological and medical fields. In this review, we introduce the hierarchical nature of RNA structure and provide a brief overview of the major computational contributions at each level of the hierarchy.

1 RNA Biology

RNA are a large class of bio-molecules that are crucial in proper functioning of cells of all living organisms [2]. Similar to proteins and DNA, RNA are

highly structured molecules on various levels. However, unlike the two other central biomolecules, RNA structure has been severely under-studied (due largely to technical difficulties in RNA biochemistry) despite their crucial importance in an ever growing number of biological functions. In one dimension, RNA are polymers made up of sequences of linked monomers known as nucleotides (or bases) which can be any of the four types: adenosine (A), uracil (U), cytosine (C), guanine (G). Canonical pairings between bases (A-U, C-G, G-U) in primary sequence give rise to interactions that define a 2D shape, known as the secondary structure. And finally, in 3 dimensions, higher order interactions between the nucleotides and secondary structure elements give rise to a 3D structure **Fig. 1**.

Information about an RNA's structure at all levels is critical to understanding its function. The primary sequence of a messenger RNA is read directly by special proteins and other RNAs to synthesize the appropriate protein during the process of translation. The formation of secondary structure elements can be used to regulate the process of translation by physically blocking access to the RNA's primary sequence by the translation machinery. Furthermore, the secondary structure of an RNA serves as a scaffold for the higher order 3D interactions which directly mediate RNA function as an independently catalytic molecule [1]. As can be seen in **Fig. 1**, there are

It is clear that a set of computational tools for efficiently modelling and predicting the nature of these interactions at every level is crucial to the aim of understanding key biological processes and engineering solutions when these processes fail. In this text we will review some of the major developments in computational modelling of RNA structure on the 2D and 3D level and its applications for structure prediction and classification.

2 RNA Secondary Structure

It is widely accepted that the majority of information necessary for forming RNA structure is encoded in its primary sequence [1]. We define an RNA sequence, or primary structure, as a string from a 4 letter alphabet

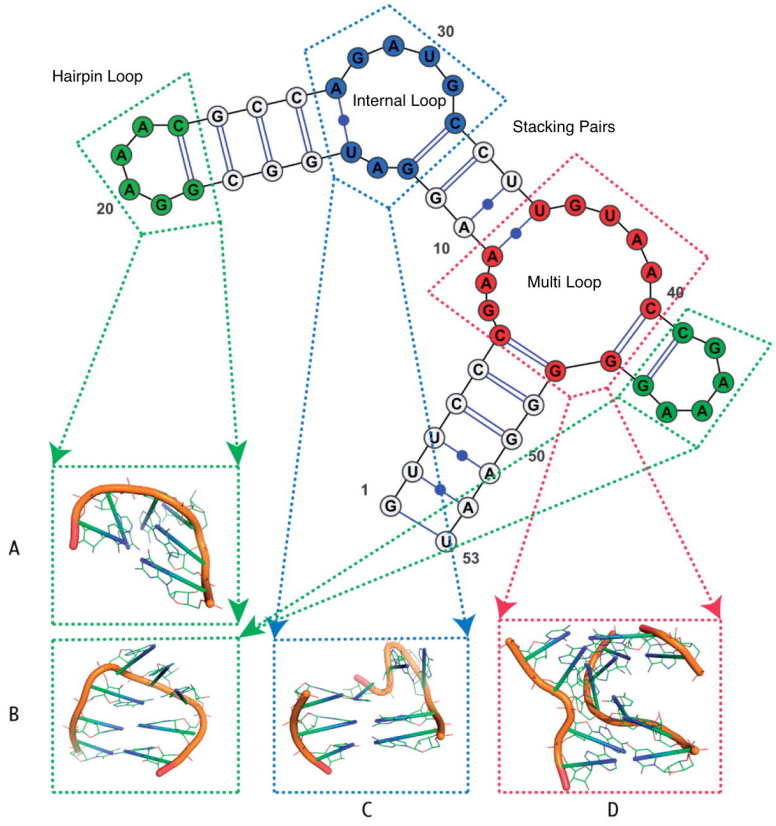

Figure 1: Example of an RNA represented in the three structural levels: nodes in the graph represent a continuous chain of nucleotides (primary structure) and compose the sequence, planar pairwise interactions between nodes define secondary structure elements (labeled by name), and boxes (labeled A-D) represent the 3D geometry of structural elements at an atomic level. Figure taken from [13].

consisting of the 4 RNA bases $\omega \in \{A, U, C, G\}^+$. One of the first tasks in understanding RNA structure was to develop algorithms to efficiently predict the secondary structure given a particular sequence. We define a secondary structure $\mathcal{S} := \{(i_1, j_1), ..., (i_k, j_k), ...(i_N, j_N)\}$ as a set of pairs of indices in ω where N is the length, or number of bases in the sequence. A valid secondary structure is a set of pairings in the form (i, j) between indices obeying the following conditions:

i.) (i, j) must form a valid Watson-Crick pairing

ii.) Each base i can pair with 1 or 0 other bases

iii.) $i < j - p$. Pairing bases must have at least p unpaired bases between them. Typically $p = 3$. (Minimal loop size)

iv.) If pairs (i, j) and (k, l) are in the structure, then (i_l, j_l), $j_k < i_l$ or $j_k > j_l$ must hold. This eliminates crossing interactions (see below).

As we are dealing with structure in two dimensions, we do not allow any crossing pairs. Although crossing interactions, or pseudoknots, are known to be present in real structures, there are currently no reliable energy measurements which can be used in efficient structure prediction methods. Most 2D folding algorithms only permit canonical `A-U`, `C-G` pairs, and 'wobble' `G-U`. Although it is known that other pairings are possible, their energetic contributions are not well established, and it is typically the case that canonical and wobble interactions are most involved in secondary structure formation. Finally, we enforce that a base can only be involved in one pairing interaction with another base.

We note that this review will not cover physics based simulation approaches due to their prohibitive computational costs and lack of convergence guarantees.

2.1 Predicting RNA 2D Structure

One of the first algorithmic approaches to the problem of secondary structure prediction, which laid the groundwork for the current state of the art was the

Nussinov algorithm, proposed in 1980 [11]. The Nussinov algorithm aims at finding the structure with the maximum number of base pairs, which uses the heuristic that base pairs have a stabilizing effect on the overall structure energy. Using this simple assumption, the Nussinov algorithm is able to efficiently identify maximally pairing structures using dynamic programming (DP). The problem is broken down in two major steps:

i.) Compute and store $OPT(i,j) \forall i,j \in [1,N]$ where $OPT(i,j)$ is the score of the lowest energy structure between indices i and j.

ii.) Retrieve the final pairing by tracing back the recursive calls in the DP table from entry $OPT(1,N)$.

The definition of secondary structure above then leads to a natural decomposition of secondary structure which can be written using the following recursion:

$$OPT(i,j) = \begin{cases} OPT(i,k-1) + OPT(k+1,j-1) + 1 & j \text{ paired with} \\ & k \in [i,j-1] \\ OPT(i,j-1) & i \text{ and } j \text{ unpaired} \end{cases}$$

Where $OPT(i,j)$ contains the optimal score of a structure on indices between i and j. Filling the table OPT in increasing order allows us to use previously computed results in adjacent cells for efficient computation. We use the minimal loop size and the crossing criteria as stopping conditions in the recursion. The computation of maximal base pairings for every subsequence can therefore be achieved in time $\mathcal{O}(N^3)$ where N is the length of the input sequence. The Nussinov algorithm was a simple, yet very large conceptual step forward in secondary structure prediction as its predecessors all required user interaction or expensive atomistic simulations. However, in its founding around base pairs it was unsuited for incorporating experimental energy that was only for larger secondary structure elements.

Soon after, Zuker [15] proposed a similar dynamic programming framework that would consider the energetic contributions of *the space between*

bonds instead of Nussinov's method of counting bonds. This shift allowed the algorithm to incorporate energetic values from experimental measurements made on various RNA 'building blocks', or structural elements **Fig. 1**. These elements include: stacks (consecutive base pairs), hairpin loops (loops closed by a single base pair), interior loops (loops closed by two base pairs), and bifurcation loops (loops closed by two or more base pairs). From experimental measurements, energy values can be associated to each building block and so the problem of identifying the MFE structure corresponds to finding the structure with the minimal sum of block energies. With this problem definition, similar recursive formulas as those used by Nussinov can be used to populate the dynamic programming table of all possible energies on the subsequences ω_i, ω_j and backtrack to reconstruct a optimal solution again in $\mathcal{O}(n^3)$. While accuracy of Zuker's folding protocol validated against experimentally solved structures still showed room for improvement, Zuker provided a key insight into the limitations of a purely theoretical approach:

> A program based solely on conformational rules and thermodynamics will not yield a biologically meaningful folding of a molecule on its own. There are too many different structures with similar energies. More and different kinds of additional information must be incorporated into the algorithm as well. – Zuker, 1981

Approaches which take advantage of evolutionary information are discussed in the next section address this remark.

2.2 Classifying RNA Structural Families

The efficient computational framework laid out from *de novo* structure prediction efforts then led to the design of structure aware probabilistic models for sequence classification. Because function is determined by an RNA's structure, sequence is generally less conserved than structure, and pairs in the structure are preserved through compensatory mutations. Statistical models are able to take advantage of this information in sequences and allow us to build what are known as RNA Structural Families. We call a set of RNA molecules sharing a common structure (but often diverse set of sequences) an RNA family. Statistical models of families are crucial tools

for classification, improved structure prediction, as well as functional prediction of novel sequences. Indeed, over the years sequences belonging to many RNA families with well defined functions and structures (tRNA, rRNA, intron RNA, etc.) have been published in databases. The task is therefore to build statistical models representing each family and allow for efficient searching for matches to the model.

Covariance Models

A major step toward solving this problem was proposed in 1994 by Eddy and Durbin [5, 4] which models RNA secondary structures using stochastic context free grammars (SCFGs). A grammar is a set of rules that are able to generate sequences of a desired form. A stochastic grammar is one that generates sequences in a probabilistic manner, and a context-free grammar is one where the generation of a character in the sequence is independent of its placement in the sequence. Given an RNA alignment which contains a set of sequences whose conserved bases share an index, and a secondary structure that is thought to be shared among all sequences in the alignment (also known as the consensus structure), the SCFG more generally known as a covariance model (CM) learns a set of generation rules and probabilities that can generate sequences belonging to the family in the alignment. The CM is a very powerful tool that solves many problems in RNA sequence analysis. CMs typically start from a pre-built alignment and consensus structure which readily allows for database searches for novel instances of a family. However, CMs can also be built from initially unaligned and unknown consensus structures therefore have the potential to simultaneously perform alignment and structure prediction.

CMs encode sequence and structure simultaneously as an ordered tree. As an example, we begin from a grammar tree that is able to represent a single RNA secondary structure and sequence. Such a tree has a branching pattern that reflects the branches of an RNA structure and nodes that emit either pairs of bound bases or singlet unpaired bases. This representation can then be abstracted to model an RNA alignment, or a collection of similar

RNA sequences. Instead of nodes occupying sequence pairs, they are associated with states that can emit a portion of an RNA alignment with some probability. We model transitions between states to allow the production of different elements of an RNA. Transitions between states are also assigned a probability. States in the model of an RNA alignment more specifically correspond to elements such as sequence matches, insertions, deletions and bifurcations (branchings). The model can be visualized as a tree **Fig. 2** or more compactly summarized as a set of grammar production rules (1).

$$\begin{aligned} P &\rightarrow aWb \quad \text{pair:} a, b \in \{A, U, C, G\} \\ L &\rightarrow aW \quad \text{left or 5' insertion} \\ R &\rightarrow Wa \quad \text{right or 3'insert} \\ B &\rightarrow SS \quad \text{bifurcation} \\ S &\rightarrow W \quad \text{start} \\ E &\rightarrow \epsilon \quad \text{end} \end{aligned} \tag{1}$$

Each state contains a production rule that is used to build from the starting string W and sequentially replace it with a string whose format follows one of the production rules. A traversal through this tree or grammar would then produce a representative instance of the alignment consensus sequence and structure. The principal operation of interest given a CM to compute the likelihood $P(\omega|\theta)$ of a sequence ω and an optimal alignment to the model given a CM θ. We can simultaneously achieve both using a dynamic programming approach. The key insight here is to compute the likelihood score on growing subsequences of ω and on growing subtrees of θ. This is similar to the Nussinov style decomposition except we add an extra dimension which is the index m of the root of the subtree of the CM being considered thus generating a three dimensional table of scores $S_{i,j,m}$ where $S_{1,N,M}$ holds the likelihood score for the entire sequence over the whole model. A backtrack over this table yields the optimal alignment to the model. Similarly, when given a set of unaligned sequences, a modified Nussinov algorithm can be used to construct the optimal tree and consensus

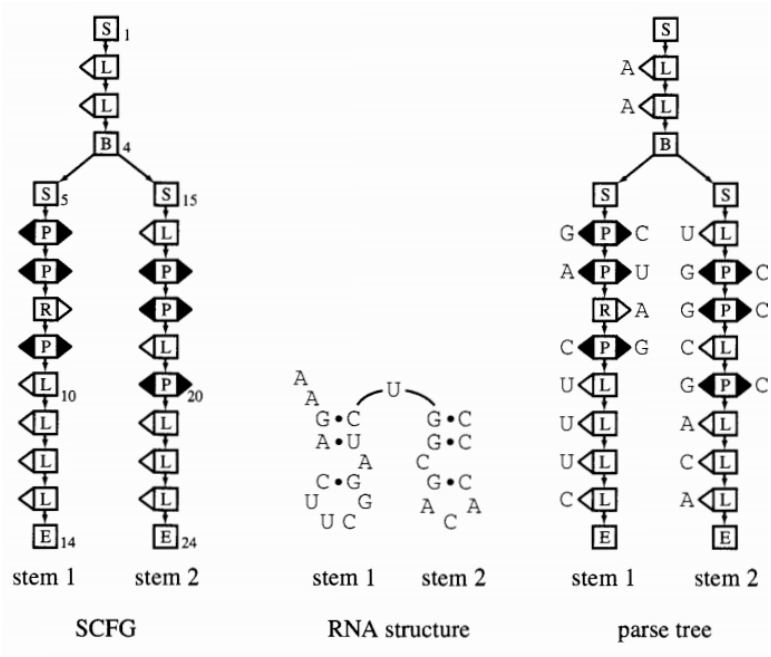

Figure 2: Example of an SCFG tree (left) representing a two hairpin RNA structure (middle) and a parse tree (right) showing the state emissions of the grammar that produce the observed sequence. Figure taken from [4].

structure by introducing the concept of mutual information over pairs of alignment columns. Pairs of columns (aligned sequence positions) i and j with high mutual information are considered as pairs in the consensus structure where mutual information is defined as follows:

$$M_{i,j} = \sum_{x_i,x_j} f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_i f_j} \tag{2}$$

The value $M_{i,j}$ can be substituted directly into the pairing score portion of Nussinov recursion to produce a consensus secondary structure. Thus the concept of correlations in sequences due to structure conservation is harnessed for structure prediction.

Building probabilistic models of RNA families allows for full genomes to be scanned in polynomial time for the presence of novel members. It is important to note that although all the algorithms are polynomial in time, the number of sequences being considered is typically high which often leads to CPU and memory expensive computations. The algorithms described were implemented in the software `infeRNAl` [10] which was used to build the popular `Rfam` database [6] which contains thousands of curated RNA families with sequence and structure information.

3 RNA Tertiary Structure

While the importance of RNA 2D structure to function motivated the development of a wide range of mathematical tools for many years, recent findings suggest that RNA also exhibit complex yet functional interaction patterns in three dimensions **Fig. 1**. Most importantly, it was found that knowledge of the 3D arrangement of atoms in an RNA yields the most information about a given RNA's function [1]. Here, we define an RNA 3D structure simply as the set of atoms and their corresponding coordinates in 3D space. Obtaining this information is often involves X-ray crystallization, a particularly challenging experiment for RNA compared to DNA and proteins.

3.1 Predicting RNA 3D Structure

The aim of 3D structure prediction most generally is to produce a 3D arrangement of all the atoms in an RNA from a given input sequence. In 2D structure, prediction algorithms relied on a fundamental structural unit such as a base pair or secondary structure element whose energy contribution was known. Although there exists such of 3D structure building block (3D motif) which will be discussed in 3.1.2 we currently do not have a set of reliable experimentally obtained energies that can be readily used in a prediction algorithm. Therefore, the class of algorithms used for 3D prediction will be more statistical in nature. For this reason, the most accurate means of producing atomistic tertiary structures currently is through physical simulations and energy minimization, known as Molecular Dynamics. However, such simulations are prohibitive in computation time, vulnerable to local minima and often require a good initial 3D structure as user input. Here we review a set of tools that take algorithmic and statistical approaches to this problem with the sacrifice of detail for gains in computational efficiency.

3.1.1 `MC-Sym`

`MC-Sym` [8] was one of the first efforts at an algorithmic approach to solve the RNA 3D structure prediction problem (3DP). The algorithm produces full 3D atomic structures from an input sequence by solving a constraint satisfaction problem. Constraints for RNA folding are extracted from a database of experimentally solved 3D structures as well as from knowledge of the sequence's secondary structure. The algorithm encodes the constraints as a set of allowed variable assignments, and produces values from the set of allowed values D to each variable. Variables correspond to 3D coordinates, as well as angles between residues in the RNA. The algorithm iterates one variable at a time through possible assignments, and if an assignment is inconsistent, a backtrack is performed and a different assignment is obtained. This is an instance of a constraint satisfaction problem using symbolic programming. It is important to note that this process does not produce a final structure, and the result of the constraint satisfaction problem is sent to an energy

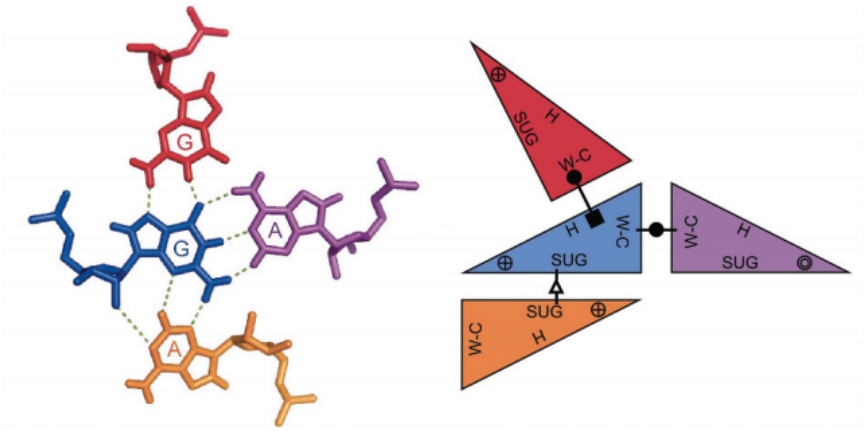

Figure 3: Leontis-Westhof nomenclature of RNA base pair interaction types. On the left an example of an RNA base (blue) interacting through each of its 3 edges with 3 other nucleotides illustrating the possible interaction geometries. The right hand side shows a schematic representation with the corresponding graphical notation.

minimization procedure to produce a refined structure. The computational efficiency of this method is limited greatly by the knowledge of structural constraints however, `MC-Sym` was still able to recreate pseudoknotted and branched structures.

3.1.2 `RMDetect`

At a higher level, Leontis and Westhof in 2001 [7] systematically defined a set of 12 possible geometrical arrangements of each of the 4 RNA bases from observations of known RNA 3D atomic structures. Leontis and Westhof found that each RNA base can be modelled as a right angled non isosceles triangle **Fig. 3** with three distinct edges: Watson, Sugar, and Hoogsten edge.

Furthermore, each of the interaction types be rotated into a *cis* or *trans* configuration. Given that there are four types of RNA base, each capable of interacting through one of its 3 edges with any other base, we get a set of 12 possible pairing geometries. Watson-Watson interactions are known as the 'canonical' base pairs which define the 2D structure, and the rest are termed

'non-canonical' and are highly prevalent in defining RNA tertiary structure. Patterns of non-canonical interactions have been shown to mediate ligand binding, catalysis and protein recognition [1]. This classification naturally gives rise to a representation of 3D structure as a graph whose nodes are nucleotides that are connected by edges which can take on any of the 12 interaction types. It has been shown that these graphs define what are known as 3D Modules which have been found to function in many unrelated RNAs [9].

Cruz and Westhof in `RMDetect` [3] then used the graphical interpretation of 3D modules to build probabilistic models of known structural patterns and accurately predict the presence of a 3D module in a given sequence, $P(\theta|\omega)$. Due to the graphical structure of the model, `RMDetect` uses Bayesian Networks to compute $P(\theta|\omega)$ where each node in the network represents a probability distribution over the 4 nucleotides conditioned on the its parent nodes (directed edges). The parameters on these distributions $P(\omega|\theta)$ are learned from a set of aligned sequences in Rfam which are known to contain a given 3D module. Once a set of modules have been trained, one can scan an input sequence and identify the Bayes net that produces the highest likelihood among the set of known modules. This approach is of course limited by the number of modules that are defined during the scanning and the availability of sequence information for each module. Regardless, `RMDetect` was able successfully predict some of the major known 3D modules and the discover of novel instances of the modules.

3.1.3 `JAR3D`

`JAR3D` in 2013 identified and aimed to resolve an important bias in the model training procedure used by `RMDetect`. That is, the assumption that all sequences in an `Rfam` multiple sequence alignment of RNA sequences for one instance of a 3D module all fold to that same module in reality. `RMDetect` relies on this assumption to parametrize the graphical models built from a crystal structure, however there is no reason to believe that this assumption will always hold. Therefore, it is important to develop models that can cap-

ture sequence variability using only known structural information. To this end, JAR3D starts from recognized structural interaction motifs (see section 3.2) to derive the novel hybrid SCFG/Markov Random Field (MRF) representations of 3D modules. MRFs, which are a generalized version of Bayesian Networks with undirected edges, are introduced to model dependencies that cannot be modelled by an SCFG. Such interactions include crossing pairs and base triples where one base has more than one partner. The models are then parametrized using a combination of isostericity knowledge and sequence variability from other known instances of the module. Isostericity in RNA 3D interactions refers to a set of discrepancy measurements from observed structures that reflect the extent to which nucleotide substitutions can preserve the same base pair geometry [14]. For each base pairing family, a 4x4 matrix with similarity measures was obtained which is used by JAR3D to model the sequence variability at each pairing position. Finally, where there exist other known instances of a 3D module, JAR3D incorporates the interaction information to better model the actual sequence variability. As a result, JAR3D was able to detect sequence instances of motifs in a more flexible manner without relying on the RMDetect alignment data. Of course, this approach is limited by the capacity of the model to efficiently represent interaction types which at the moment is limited to only internal and hairpin elements.

3.2 Classifying RNA 3D Modules

Specific 3D interaction patterns have been shown to be highly conserved and therefore contain important information regarding the function and geometry of an RNA. Therefore, building organized catalogues of modules is crucial for assisting in function prediction as well as improving structure prediction (as seen in 2.2). We consider two types of conserved interaction pattern, or 3D module as the building block of such catalogue: A *local motif* is a set of characteristic base pairing interactions that are close together in structure (i.e. located within the same secondary structure element; internal loops, hairpins, etc.), while a *long range motif* involves interactions that

connect two or more secondary structure elements. It is important to note that in both cases, the defining feature is the Leontis-Westhof interaction type. Nucleotide identity is often allowed to so long as it preserves the 3D geometry. Here we review two motif cataloguing approaches, each of which addresses one of the motif types.

3.2.1 `RNA 3D Atlas`

Petrov et. al [12] in 2013 developed an automated system for identifying and grouping *local* recurrent 3D motifs. In its current form, the 3D Atlas only catalogues motifs found within internal loops and bulges due to their richness in non-canonical interactions and their topological simplicity. The 3D Atlas automatically extracts interactions within loops labeled using the Leontis Westhof annotation from a large repository of non-redundant RNA crystal structures as the fundamental unit of the atlas. An all-against-all geometric comparison is performed which results in a matrix M containing similarity scores for every pair of structures. The next task is to identify sets of structures that are all mutually similar to each other. Such a group is deemed to be a recurrent motif. `RNA 3D Atlas` takes a graph theoretic approach to solve this problem by using M to define a graph G where nodes represent motif candidates connected by edges if they satisfy a fixed similarity score. Because the most populated motifs are of interest, the algorithm repeatedly identifies the largest motif and removes it from the graph. This problem is an iterative repetition of the maximum clique problem, where a clique is defined as a connected set of nodes in the graph and a maximal clique is the clique of the highest cardinality in the graph. The maximum clique problem has been shown to be NP-complete, however speed-ups are achieved by reducing the number of nodes to be checked using fast approximate graph coloring algorithms. The main intuition is that the number of colors used in a clique must be equal to the size of the clique. Therefore, graphs can be efficiently pre-processed by eliminating nodes *a priori* that cannot belong to the maximal clique. Once this process has been repeated sufficient times, exact maximum clique algorithms are applied. The result

is an automatically updated database which capture the major known 3D modules, as well as hundreds of novel modules. The major limitation however is that these currently only include local motifs and of the local motifs only internal and hairpin loops are modelled.

3.2.2 `CaRNAval`

Simultaneously modelling short range and long range interaction motifs is computationally challenging due to the explosion in interactions to search, and for this reason RNA 3D Motif classifies only short range motifs. `CaRNAval` [13], developed shortly after addresses the problem of identifying recurrent *long range* motifs. Similar to RNA 3D Motif, `CaRNAval` takes a graph theoretic approach to identifying groups of similar motifs. `CaRNAval` directly models each RNA structure as a graph where nodes represent nucleotides and edges represent interaction types with labels: $\{c, t\} \times \{W, S, H\}^2$ **Fig. 4**. Because the aim is to find recurrent interaction networks, nucleotide identity is not considered and the only labeled element of the graph are the edges. For each edge there is an additional binary label indicating whether an interaction is long or short range. `CaRNAval` builds the database by finding the 'maximal interaction modules' between two graphs, which it defines as a common edge-labelled subgraph of two graphs which share at least two long-range edges. The problem of finding the maximal subgraph isomorphism is NP-Hard. `CaRNAval` therefore employs a heuristic to limit the search space by iteratively building larger common subgraphs from an initial set of shared long-range interactions and limiting its search to only long range motifs. Using this technique on a set of non-redundant interaction pairs, 136 interaction motifs were identified which themselves can occur as subgraphs of other motifs forming a large network of known (such as the A-minor motif) and novel long-range interaction patterns. Currently, this approach is limited to finding interaction patterns only between pairs of secondary structure elements whereas it is possible that there could be higher order interactions at play.

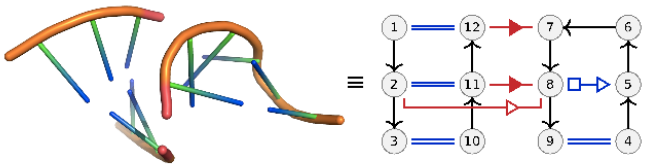

Figure 4: Graphical representation of two interacting secondary structure elements by `CaRNAval`. Interactions are labeled as edges with the corresponding Leontis-Westhof nomenclature. Nodes represent nucleotides. Blue edges are local interactions and red edges are long-range interactions. Black directed edges represent backbone interactions and point from the 5' to 3' end of the chain.

References

[1] Philippe Brion and Eric Westhof. Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997.

[2] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[3] José Almeida Cruz and Eric Westhof. Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nature methods*, 8(6):513–519, 2011.

[4] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[5] Sean R Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11):2079–2088, 1994.

[6] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R Eddy, and Alex Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research*, 33(suppl_1):D121–D124, 2005.

[7] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *Rna*, 7(4):499–512, 2001.

[8] Francois Major, Marcel Turcotte, Daniel Gautheret, Guy Lapalme, Eric Fillion, and Robert Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, pages 1255–1260, 1991.

[9] Lorena Nasalean, Jesse Stombaugh, Craig L Zirbel, and Neocles B Leontis. RNA 3D structural motifs: definition, identification, annotation, and database searching. *Non-protein coding RNAs*, pages 1–26, 2009.

[10] Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, 2009.

[11] Ruth Nussinov and Ann B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences*, 77(11):6309–6313, 1980.

[12] Anton I Petrov, Craig L Zirbel, and Neocles B Leontis. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *Rna*, 19(10):1327–1340, 2013.

[13] Vladimir Reinharz. *Algorithmic Properties of Evolved Structured RNAs*. PhD thesis, McGill University Libraries, 2016.

[14] Jesse Stombaugh, Craig L Zirbel, Eric Westhof, and Neocles B Leontis. Frequency and isostericity of RNA base pairs. *Nucleic acids research*, 37(7):2294–2312, 2009.

[15] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, 1981.