Computational prediction of shape from sequence for non-coding ribonucleic acids

Jim Beck

May 1, 2020

Abstract

Limited by an experimentally intractable problem space, biologists develop computational models to predict the physical structures found within non-coding ribonucleic acids ("ncRNA") [5]. This class of RNA serves an array of important biological functions that are impacted by their associated tertiary structures [1]. Individual ncRNAs can be characterized in terms of their molecular sequence, shape, and function [11]. This characterization fuels the hope that an ncRNA's three-dimensional shape may be predicted by correctly interpreting the molecule's sequence of nucleotides. But this challenge has proven insurmountable for all but the simplest of predictions [12]. Existing models, largely based on methods employed for protein modeling, have fared poorly when predicting larger structures or those lacking homologous templates [7]. Choices made in the interest of computational efficiency have degraded model accuracy. And determinations of model performance are speculative at best given the lack of a common definition for accuracy. Despite these obstacles, recent advancements in solving the challenges inherent to longer sequences as well as the potential for more integrative approaches provides hope for improved predictive accuracy.

1 Introduction

Computational biologists use predictive models to expand their understanding of non-coding ribonucleic acid ("ncRNA") structures [5]. Historically, biologists seeking to understand an ncRNA's three-dimensional ("3D") structure have been limited to expensive and time-consuming crystallographic experiments [3]. But a growing backlog of sequenced ncRNAs has increased the appeal of in silico models for structural prediction [3]. By incorporating computational models, biologists hope to improve the speed and depth of their understanding for this important class of RNA molecules. But before we can realize these benefits, computational biologists must first improve the accuracy of existing predictive models.

At the heart of each ncRNA model is a one-dimensional ("1D") nucleotide sequence that can be viewed as a set of folding instructions [13]. An ncRNA's 3D shape is an implementation of these instructions capable of performing a particular function [11]. Predicting an ncRNA's shape can, therefore, be thought of as the mapping of a 1D sequence space to a 3D structure space. This mapping takes an ncRNA's primary structure (in the form of a nucleotide sequence of length L with 4^L possible combinations), identifies secondary structures (composed of subsequences) as well as interactions between secondary structures, and produces a folded, 3D tertiary structure (see Figure 1). Mutations occurring within an ncRNA sequence may cause structural alterations and corresponding changes in function (see Figure 2). These relationships are important and direct - sequence determines shape and shape defines function.

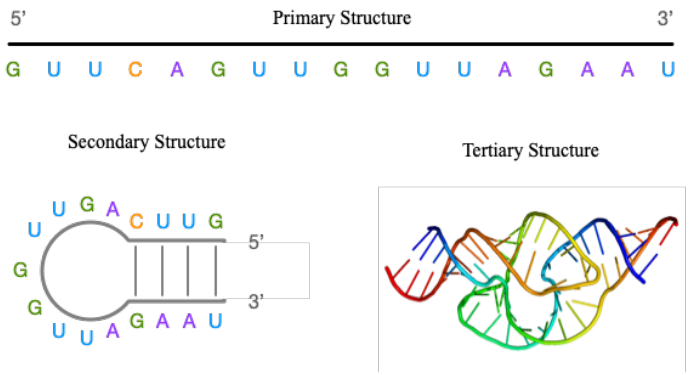

Figure 1: Primary and Secondary structures of D-Loop. Tertiary Structure of Hammerhead Ribozyme by William G. Scott https://commons.wikimedia.org /w/index.php?curid=2257831

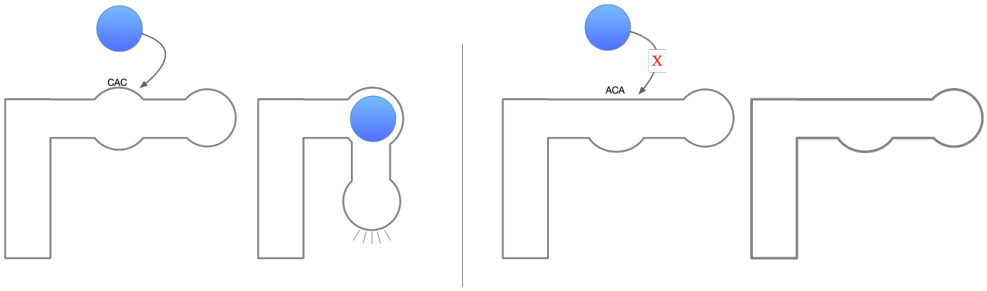

Figure 2: In this example, the ligand (blue) is able in one instance to bind to an RNA aptamer causing a conformational change that initiates catalysis. A mutation causes the aptamer to take on a new shape, making it unable to accept the ligand. Consequently, no conformational change occurs.

The relationship between sequence, structure, and function underlies computational predictions of ncRNAs' tertiary structures. Current models rely on this relationship to catalog secondary structures that tend to occur within particular sequence spaces. And because similar sequential, structural and functional relationships are also found in proteins, ncRNA models rely upon methods used to successfully predict protein structures from amino acid sequences to aggregate ncRNA secondary structures. While these methods have produced highly accurate predictions for simple ncRNA sequences, they

have failed when confronted by longer sequences or those lacking homolog templates [7]. Additionally, the combinatorial complexity inherent to the problem requires researchers to strike a difficult balance between computational tractability and model accuracy. New techniques are needed to overcome these predictive deficiencies.

2 Modern Methods

Modern computational predictions of ncRNA structures rely upon techniques used to predict protein structures [6]. These models commonly organize known structural elements contained within a sequence into a low energy configuration by iteratively manipulating the atomic positions of the various secondary structures [12]. Informed by the idea that structural divergence is much slower than sequence divergence, models use template libraries of homologous sequences (e.g. ModeRNA) to recognize folds within an associated sequence [7, 12]. The secondary structures derived from these templates are then applied to a force field model (e.g., SimRNA, Rosetta or VFold) that evaluates core alignment or atomic distance and torsion angle constraints [12]. Of course, the implementation details vary from model to model, with some applying coarse grained energy functions to secondary structures (the Bujnicki Group), others combining homology modeling with step-wise, all-atom energy functions (the Das Group), and still others iterating between secondary structures, energy landscapes and a tertiary structure buildup (the Chen Group) [7]. Importantly, current ncRNA models are less accurate than their protein counterparts and this may be a result of molecular dissimilarities.

The prospect of using protein folding models for ncRNA predictions is appealing given their molecular similarities with each possessing a backbone and side chains, covalently bound within their primary structures, containing secondary structures stabilized by hydrogen bonds , and having 3D structures that contain many long-range interactions amongst their constituent substructures [12]. But ncRNA structures are significantly different from proteins in that they have greater structural diversity [6], more complex packing arrangements [6], secondary structures bound at different locations [12], and different mechanisms driving molecular compaction [12]. These differences significantly increase the ncRNA model's degrees of freedom, making accurate prediction appreciably more complex than in comparable protein models [6]. And in fact, most current ncRNA models produce inconsistent and inaccurate predictions.

ncRNA models fare particularly poorly when predicting long sequences lacking structural templates. For short sequences, accurate models can be obtained by manipulating the full array of atomic representations [12]. For sequences with template homologs, accurate models can be produced by judiciously accumulating secondary structures into an aggregate 3D shape [7]. But for sequences above 70 nucleotides or without homology, the result is quite different [15, 7]. Model inaccuracy for

these molecules are correlated with their sequence size and the volume of additional experimental data available to further identify their secondary structures [7]. Because long sequences have the potential for both a greater diversity of secondary structures as well as an increased volume of intra-structural interactions, a solid understanding of secondary structures is essential to the prediction of ncRNA's 3D shape.

3 Secondary Structure

Secondary structures are the fundamental building blocks upon which all ncRNA predictive models are built. Schuster et al. described secondary structure as a coarse-grained representation of ncRNA's tertiary structure and identified key predictive properties within their sequence space [13]. First, the variety of distinct secondary structures follow a power-law distribution, with a few common and many uncommon structures [13]. Next, secondary structures are distributed evenly throughout their underlying sequence space [12]. And finally, secondary structures exist within mutational neighborhoods, with each neighborhood possessing the potential to form all common shapes (see Figure 3) [13]. This information is useful because modern algorithms rely on secondary structures as the building blocks for their predictions. Understanding how structures exist within a given sequence space allows us to compare similar structures across homologs and rely on the availability of common structures across a variety of sequence spaces. But secondary structures are, at their core, a function of base pairing. And the prevalence of non-canonical base pairings, creates a significant obstacle to the accurate identification of many secondary structures.

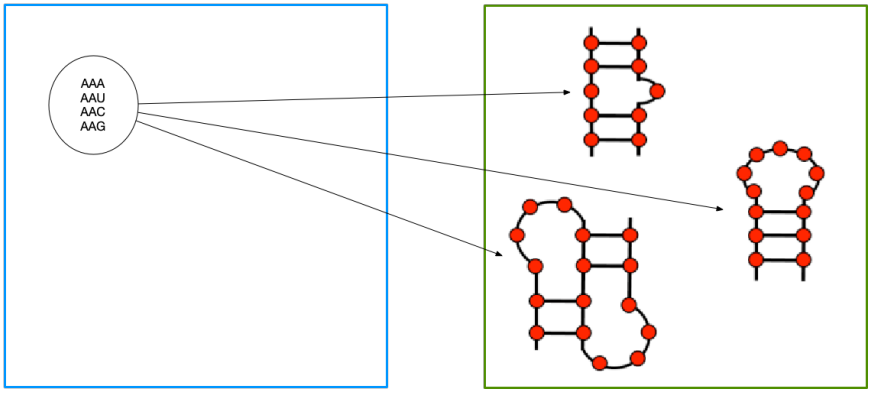

Figure 3: A sequence neighborhood (blue box)is generally able to create all common structures (examples within circles in the secondary structure space(green box)) and many uncommon structures.

Non-canonical base pairs are significant contributors to predictive accuracy [7]. These pairings participate in the irregular loop segments commonly found amongst ncRNA's secondary structures and provide stiffness and deformability within those structures [7, 9]. Yet correctly interpreting the structural impact of non-canonical pairs can be

challenging due their high variability and unique covariation rules [7]. The simplest way to assess the structural deformations caused by non-canonical pairings is by considering their spatial arrangement [9]. The relative orientation of ncRNA bases with respect to each other includes three translational parameters (shear, stretch, and stagger) and three rotational parameters (buckle, propellor, and opening). The many spatial degrees of freedom combined with the variety of potential pairings has led to significantly decreased predictive accuracy even in those models with correct global folds [14].

Current models that correctly predict non-canonical configurations are computationally expensive. Watkins et al. provided a high-resolution mechanism for evaluating non-canonical base pairings through the use of a stepwise optimization process (see Figure 4) [14]. The optimization process evaluates the energy state for each parameter's configuration of a non-canonical pairing. This evaluation guarantees a unique solution but comes with a large computational cost - often tens of thousands of CPU hours for even small sequences [14]. To reduce this cost, they further employed a stepwise Monte Carlo (SWM) optimization that abandons the full enumeration of each non-canonical configuration in favor of stochastically chosen nucleotide additions. Each addition is iteratively reconfigured until an energy reducing step produces a decrease that falls below a random fluctuation threshold. Despite the fact that the stochastic iteration converges approximately two orders of magnitude faster than the fully enumerative process, the authors still found the method computationally costly [14]. Consequently, they resolved themselves to the idea that this high-resolution approach might only serve as an ancillary method to other low-resolution models [14]. But their process is interesting for another reason in that it represents a solution to a targeted model deficiency. One might conclude that accurate predictions of ncRNA tertiary structures require many such approaches, each solving a different portion of the predictive problem.

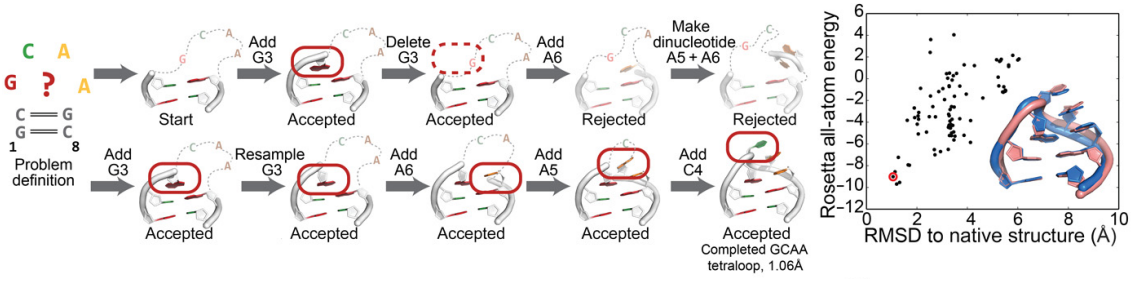

Figure 4: Stepwise iterative optimization to identify low energy configurations wherein each addition is stochastically chosen and positional adjustments made until a sufficiently low energy state is achieved. Graphic by Watkins et al.[14]

4 Interactions between Secondary Structures

Tertiary contacts are another challenging, but important impediment to accurate predictive models. These contacts form key connections between folded segments of secondary structures at sequence locations often separated by large distances [15]. Consequently, these contact points are fundamental to tertiary structure but occur at distances that make identification difficult. One approach to the problem is to consider correlated mutations at distant sequences as evolutionary couples that preserve function and structure [14]. The trouble with this approach is that it tends to fixate on individual pairings and is, therefore, prone to transitivity [15]. For better accuracy, one would need to isolate participating nucleotides from correlated, but otherwise uninvolved, nucleotides [15]. Such a model would consider all pairwise contacts globally and remove the transitive participants [15].

Weinreb et al. described a technique for identifying evolutionary couplings that deconvolves transitivity while searching base-pair positions having correlated substitutions within aligned, homologous sequences [15]. Recognizing that tertiary contacts are often part of a complex network that might obscure true interactions, the authors further identified evolutionary couples by evaluating each RNA family as a sequence space distribution, with each sequence having a probability that reflects its single-site biases and its correlation with other bases (Equation 1, Figure 5) [15]. Tertiary contacts become identifiable when they involve evolutionarily coupled sequences that are constrained within a minimum energy state [15]. Using this technique to predict long-range contacts on longer sequences (70-120 nucleotides), the authors achieved significantly improved accuracy [15]. This is an important outcome, because accurate prediction of tertiary contacts allow for the combination of secondary structures into higher-ordered structures. And combining highly accurate tertiary structures into complete 3D predictions may be a path towards reducing computational complexity without a corresponding reduction in accuracy.

$$P(\sigma) = \frac{1}{Z} exp\left(\sum_{1=1}^{L} h_i(\sigma_i) + \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} J_{ij}(\sigma_i, \sigma_j)\right) \tag{1}$$

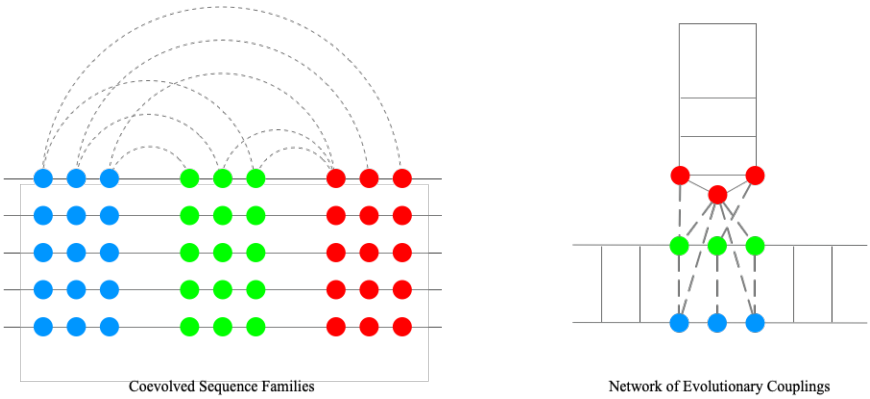

Figure 5: The probability of observing a particular sequence (σ) is a function of its single site bias hi(σ_i) and its correlation with other single site positions J_{ij}(σ_i,σ_j). The correlations between sequence examples within co-evolved RNA families are evaluated and the network of identified couplings used to predict tertiary contacts.

5 Higher-Order Structures

ncRNAs have higher order structures that may improve predictive models [6]. The idea that predictive models are aggregations of structural building blocks can be extended to the use of higher order structural motifs - structures that incorporate multiple secondary structures [6]. This notion of modular componentry is reinforced by the finding that structural motifs often appear together, in a seemingly collaborative relationship [6]. Laing et al. considered structural motifs as graph representations of RNA sequences (see Figure 6) capable of reducing computational complexity [6]. By applying concepts from graph theory, ncRNAs secondary and tertiary structures can be surveyed for relationships that may prove useful in predictive models [6]. For example, by categorizing structures within their graph representations , Lang et al. determined that the universe of RNA structures is dominated by pseudoknots and that statistical clusters differentiate RNA-like shapes from artificial configurations. The authors further proposed that graph isomorphisms could be a useful mechanism for comparing structural motifs [6]. This is an interesting approach to identifying topological similarities in that it should be computationally efficient to implement at the sequence level.

By changing the granularity of their observations, researchers can use secondary structures and tertiary motifs to reveal additional information about ncRNA while also attending to the combinatorial complexity of the problem space [6]. The natural extension of such a graphical representation is to connect the organization of variously grained structures with homolog templates to extend their utility. We know from the literature that the availability of templates dramatically improves predictive accuracy [7]. So maybe the identification of tertiary patterns within homolog templates could expand their application to sequences without homologs and the resulting increase in template accessibility could reduce the computational complexity inherent to the prediction.

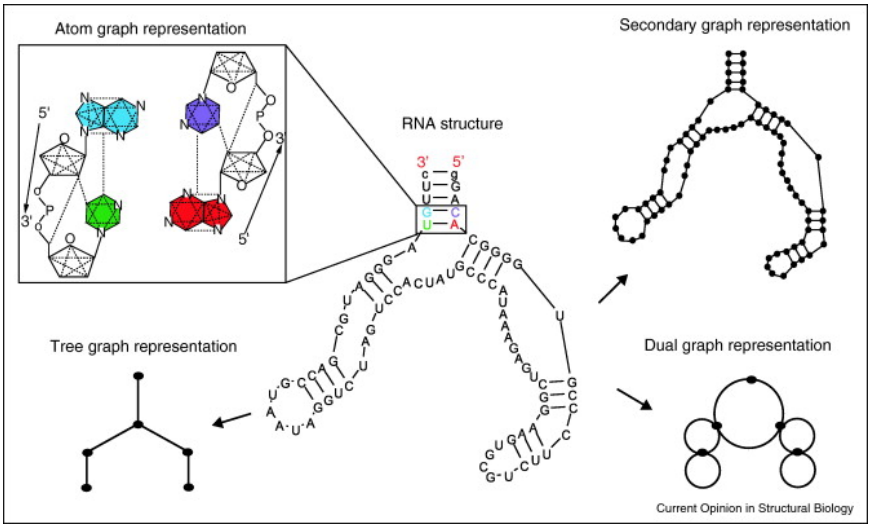

Figure 6: Graph representations of RNA structure at various levels of abstraction by Laing et al.[6]

6 Computational Complexity

The unsatisfying trade off between model accuracy and computational complexity remains unresolved. Predicting 3D shape from a 1D sequence is a combinatorially complex problem. As a threshold matter, each additional sequence member increases the total sequence space by a factor of 4 and the free energy determination increases by $3N_{atoms} - 5$ degrees of freedom for each evaluated nucleotide [12]. The sequence space component of the problem is addressed through the use of homolog templates, thus limiting the volume of free energy states to evaluate. But models for sequences lacking known homologs can't avail themselves of this technique. The free energy component of the problem is computationally taxing because the force field algorithm must select from a large number of similarly low energy conformations.

To reduce the size of the analysis, predictive models have two choices: they can simplify their coordinate systems by transforming bond lengths and angles to a restricted set of idealized values or they can form a single, atomic interaction center rather than using all available atoms to reduce the problem's degrees of freedom [12]. These solutions seem to work adequately for the less demanding prediction of proteins. But coarse-grained models and other dimensionality reduction schemes lead to reduced model accuracy for ncRNA, particularly those without homolog templates [12]. Accordingly, most models produce an array of potential conformations, each reflecting a different choice within the dimensionality reduction scheme [7]. These predictions are like a strategic guess that deals with a dense problem space by first making an approximation using higher level structures before producing many possible detailed solutions. This strategy produces models of highly variable accuracy. Of course, this assumes we know what accuracy means - and we might not.

7 Model Accuracy

The lack of a common definition for accuracy is an impediment to realizing a robust model for predicting ncRNAs' 3D shape. The most common metric of structural accuracy is the root mean squared deviation ("RMSD") between the predicted conformation and the actual crystal structure [7]. This measure of topological accuracy can be thought of as the positional similarity averaged across all atoms. Lower RMSDs tend to be highly correlated with the use of template homologs and this seems to make intuitive sense recalling the idea that spatial changes lag sequential mutations [7]. But other measures, such as nucleotide interaction, torsion angle similarity, and atomic harmony, also describe model accuracy and these metrics could be useful when evaluating certain structures [7]. And, in fact, structural models do compare differently across the array of metrics, making the identification of a superior model highly subjective [7].

So, we are left with a question - must a model be accurate in all metrics or are accuracy metrics specific to a particular application. Whether a model is globally or situationally accurate is a useful consideration for predictive modeling in that accuracy may be a construct that depends on the analytic perspective of the researcher. The literature typically reports accuracy by evaluating each model against an array of metrics, suggesting that there is some future model that will be robustly accurate [7]. But maybe this is a rabbit hole and we should instead pursue accuracy using metrics that reflect the question being asked. Or maybe a "best" composite metric can be identified to reflect a universal consensus for model accuracy.

In this regard, functional accuracy is curiously missing from all definitions of model accuracy. A measure of functional similarity seems like it should be the most important outcome of any predictive model - does the thing act as predicted? But how one would conduct a functional assessment of a computational model is not obvious. Perhaps some indication that an existing metric correlates well with functional performance might reveal a suitable surrogate. Maybe prioritized locations exist within an ncRNA sequence such that one could weight the RMSD function to reflect functionally important structures. Or maybe corollaries exist within homolog templates that could be empirically tested for functional comparisons. Regardless of the options, if we seek to identify a prediction as accurate, we'll need to be able to clearly define what it means to be accurate.

8 Integrated Approaches

The functional performance data found within fitness landscapes may help improve the accuracy and speed of de novo structural predictions. Functional improvements can be viewed as a navigation of mutational pathways in search of evolutionary optimization [10]. For close to a century, biologists have described these evolutionary advancements in terms

of a fitness landscape with each landscape consisting of three components:

1. a configuration set of sequence variants, secondary structures, motifs, etc.,
2. a configuration space that organizes the set: nearness, hamming distance, accessibility, etc., and
3. a function that assigns fitness to the set members [4].

By studying the interaction of these components and their multi-dimensional mapping from 1D mutational sequences, one can highlight features of the fitness landscape that reveal how mutational events impact functional performance (see Figure 7). The relationship between fitness values and mutationally related sequence values seems like an important, and possibly unused, connection to ncRNA structure. Leveraging the delay between sequence mutation and structural adaptation, computational biologists might look for connections between ncRNA fitness values and structures. These connections could conceivably expand the utility of existing template libraries, thereby improving the accuracy and speed of existing algorithms. Additionally, fitness landscapes may provide a computationally convenient mechanism for describing the accuracy of structural models, where a model's accuracy is proportional to its functional similarity with an identically sequenced wild type molecule.

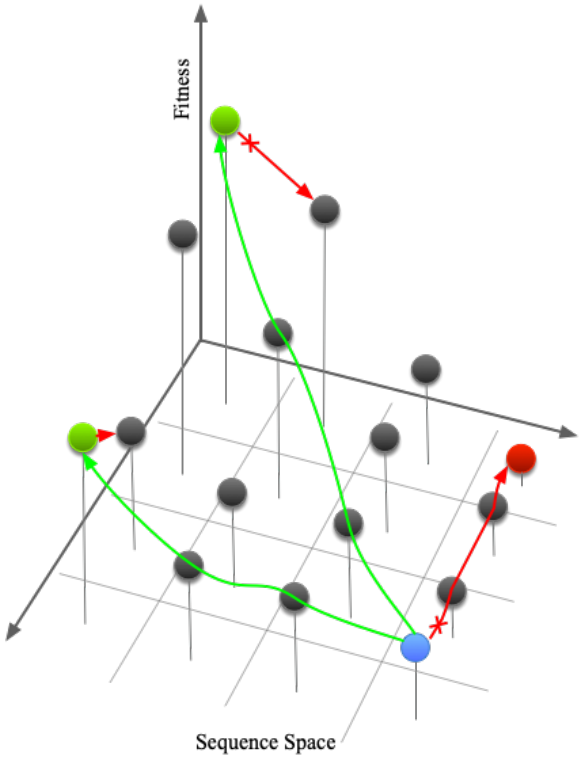

Figure 7: Successful evolutionary pathways (green arrows) are those which include sequence mutations that increase fitness. Paths that decrease in fitness (red arrows) are evolutionarily blocked. Many landscapes include multiple peaks of varying fitness (green node) as well as valleys or decreased fitness (red node).

The 1D mapping problem is curiously similar for structural and fitness predictions. Each is based on the premise that sufficient information is contained within the sequence space to accurately predict a higher dimensioned structure and fitness space. Each relies upon abstract representations to identify relationships amongst the problem elements. And each faces problems associated with scale such that neither empirical nor computational problems can be addressed without dimensionality reduction. None of these similarities should be surprising given the connection that sequence has with structure and structure has with function. This connection may provide opportunities to implement network solutions to resolve structural prediction challenges.

The individual structural and fitness problems could be connected using network theory techniques to explain the relationship amongst its constituent elements. Structural predictions often link together secondary structures into ordered root tree graphs [13], RNA motifs into dual and atomic graphs [6], and evolutionary couplings into undirected graphs [15]. Fitness predictions uniformly use a graph representation for linking sequences within a single mutation and enhance these graphs with directed paths using fitness gradients between mutated nodes [16]. Additionally, graph representations are routinely used to represent these problem structures at different levels of refinement. For example, we can graphically represent mutational differences based on their hamming distance, their fitness scores, or their membership on an obligatory path. It's also conceivable that a higher-ordered network relationship exists between sequence, structure, and function, possibly complementing existing structural prediction algorithms.

Higher-order network structures can provide important information about structural sub-graphs [2]. Benson et al. explain that sub-graphs having a particular connection pattern (see Figure 8,9) are building blocks for complex networks and that surveying a network for these patterns could explain some aspects of the network's overall structure [2]. Applied to the general sequence problem space, one could conceive of a network of sequences, fitness gradients, secondary structures, and tertiary structures exhibiting a unique high-order pattern that would aid in the de novo prediction of unknown sequence structures. Nichol et al. peripherally consider this idea in their survey of genotype to phenotype mappings, identifying the common results realized by various empirical studies and computational models operating within a framework of genotype, phenotype, environment, and fitness [8]. They further suggest that future, more diverse mappings will need to balance biological complexity with computational tractability. Necessarily, this diversity contemplates making the connections that network theory is designed to evaluate.

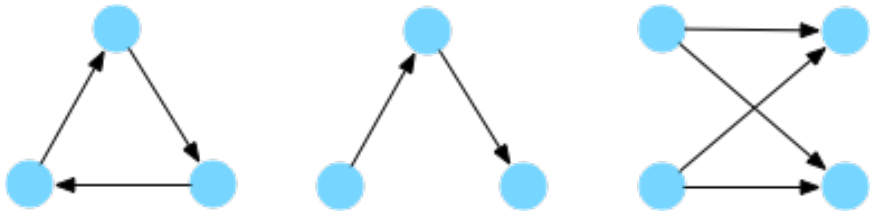

Figure 8: Examples of higher-order motifs using a directed graph.

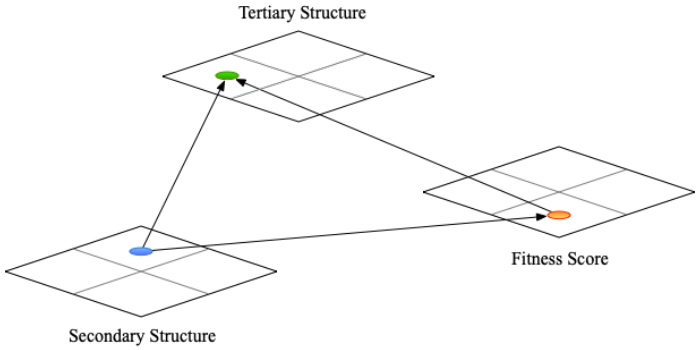

Figure 9: An applied motif is presented to demonstrate how one might search for patterns amongst various structural and fitness spaces.

9 Conclusion

The sequence to structure literature has made clear that accurate ncRNA structural predictions continue to pose a significant challenge for computational models. While the scientific community has had some success in modeling simple ncRNA structures, they have struggled with long sequences for which no homolog templates are available. The combinatorial complexity of long sequences combined with the increased likelihood of distant tertiary contacts and the higher resolution requirements of non-canonical pairings has impeded the creation of a singularly robust predictive algorithm. Additionally, the lack of a generally accepted definition of accuracy makes identification of better models highly subjective. In this regard, no current definition reflects functional accuracy and this seems to be a major shortcoming of our existing array of metrics.

There does appear to be consensus on a general algorithmic approach that relies on known structures and free energy models for predicting ncRNA's 3D structures. Consequently, one could consider the accurate prediction of ncRNA's structures as a natural outcome of improved structural libraries. In this case, we should focus on building a robust template library that broadly covers the distribution of ncRNA molecules. Alternatively, we could look to refine our predictive models to include other relational data from which to better understand the mechanisms of folding. Perhaps that data might exist within other ncRNA problems, like RNA fitness landscapes. Existing approaches that rely on incremental knowledge about related sequences and supplemental empirical

results are still necessary for filling the gaps in our knowledge about folding and functional pathways. In the future, we'll possibly observe definitive structural patterns for the de novo mapping of long sequences. We should look for indications of these patterns within the connections between the sequence space and other available ncRNA solution spaces.

References

[1] Bruce Alberts et al. *Essential Cell Biology*. Garland Science, 2013.

[2] Austin R Benson, David F Gleich, and Jure Leskovec. "Higher-order organization of complex networks". In: *Science* 353.6295 (2016), pp. 163–166.

[3] Emidio Capriotti and Marc A Marti-Renom. "Computational RNA structure prediction". In: *Curr. Bioinform* 3 (2008), pp. 32–45.

[4] Luca Ferretti et al. "Evolutionary constraints in fitness landscapes". In: *Heredity* 121.5 (2018), pp. 466–481.

[5] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. "Abstract shapes of RNA". In: *Nucleic acids research* 32.16 (2004), pp. 4843–4851.

[6] Christian Laing and Tamar Schlick. "Computational approaches to 3D modeling of RNA". In: *Journal of Physics: Condensed Matter* 22.28 (2010), p. 283101.

[7] Zhichao Miao et al. "RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme". In: *Rna* 23.5 (2017), pp. 655–672.

[8] Daniel Nichol et al. "Model genotype–phenotype mappings and the algorithmic structure of evolution". In: *Journal of the Royal Society Interface* 16.160 (2019), p. 20190332.

[9] Wilma K Olson et al. "Effects of Noncanonical Base Pairing on RNA Folding: Structural Context and Spatial Arrangements of G· A Pairs". In: *Biochemistry* 58.20 (2019), pp. 2474–2487.

[10] Frank J Poelwijk et al. "Empirical fitness landscapes reveal accessible evolutionary paths". In: *Nature* 445.7126 (2007), pp. 383–386.

[11] PH Raven et al. *Biology*. New York: McGraw-Hill Education, 2017.

[12] Kristian Rother et al. "RNA and protein 3D structure modeling: similarities and differences". In: *Journal of molecular modeling* 17.9 (2011), pp. 2325–2336.

[13] Peter Schuster et al. “From sequences to shapes and back: a case study in RNA secondary structures”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255.1344 (1994), pp. 279–284.

[14] Andrew M Watkins et al. “Blind prediction of noncanonical RNA structure at atomic accuracy”. In: *Science advances* 4.5 (2018), eaar5316.

[15] Caleb Weinreb et al. “3D RNA and functional interactions from evolutionary couplings”. In: *Cell* 165.4 (2016), pp. 963–975.

[16] Nicholas C Wu et al. “Adaptation in protein fitness landscapes is facilitated by indirect paths”. In: *Elife* 5 (2016), e16965.