# TRANSAT: a method for detecting evolutionarily conserved helices in alignments of RNA sequences and its application in identifying transient or alternative RNA structures

by

Nicholas J. P. Wiebe

B.Sc., The University of Alberta, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2010

# Abstract

The secondary structure of RNA molecules is often critical to their proper functioning, and so prediction of those structures has been a focus of bioinformatics research for many years. RNA folds as it is transcribed, and it has lately become apparent that the sequences of structures that an RNA adopts (its *folding path*) is vitally important for the RNA to fold into its proper structure.

Analysis of the evolution of a group of related RNAs is useful for identifying conserved and therefore functionally important secondary structures. In theory, functional but transient secondary structures which play a role in the folding pathway should also be detectable in this way. Moreover, folding may be affected by a variety of factors in the cellular environment, which presents a challenge to the existing methods of RNA pathway prediction via simulation. Evolutionarily conserved helices are implicitly the ones formed *in vivo*, and so is a useful means of accounting for this problem.

Here we present Transat, a method of identifying evolutionarily conserved elements of RNA secondary structure, including transient structures, from an alignment of related RNAs. We evaluate Transat's performance on a wide variety of alignments, present some examples of its predictions, and show how its predictions may be useful for predicting folding pathways. We also present a method of generating simulated alignments, and use these alignments to examine Transat's performance in ways that are challenging with alignments of real sequences.

# Preface

A paper based on this work has been published:
N. J. P. Wiebe and I. M. Meyer, "TRANSAT — a method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures," *PLoS Comp Bio*, vol. 6, p. e1000823, 2010.

For this paper, N. J. P. Wiebe implemented TRANSAT, conducted the computational experiments, and made the figures. Both authors contributed ideas to the conception and fine tuning of TRANSAT and analyzed its results. I. M. Meyer wrote the paper.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgements

I would like to thank my supervisor Irmtraud Meyer, the members of my supervisory committee, Anne Condon and Steven Hallam, and the others I consulted with on various aspects of this project, particularly Christopher Thachuk and Leon French. I would also like to thank Rodrigo Goya, Anamaria Crisan, and Oana Sandu, with whom I worked with on the class project that grew into this work, and Yaojie Chen, who worked with me on a related project. Above all, thank you to my parents, whose unwavering support has made this possible.

# Chapter 1

# Introduction

## 1.1 RNA secondary structure

Ribonucleic acid (RNA) is a biopolymer that plays a vital role in many of the most basic cellular processes. RNA molecules are composed of nucleotides, like DNA. Unlike DNA, the sugar in its phosphate-sugar backbone is ribose rather than 2-deoxyribose; the extra oxygen in RNA makes it more chemically active (and less stable) than DNA. Four types of nucleotide are found in RNA sequence: adenosine (A), cytosine (C), guanine (G), and uracil (U). The sequence of these nucleotides gives rise to the RNA's function. RNAs serve a wide variety of purposes in the cell, some examples of which are:

- Messenger RNA (mRNA) encodes protein sequence information, and is used as the template for protein synthesis.
- Transfer RNAs (tRNA) are used in the translation process to 'read' mRNA codons and associate them with the appropriate amino acid.
- Micro RNA (miRNA) act as regulators of gene expression by binding to target mRNA molecules and triggering their degradation [1].

A host of other non-coding RNA (ncRNA) have been identified [2, 3], and new evidence indicates that, in mammals, the majority the genome is transcribed, hinting that there may yet be more ncRNAs to be discovered [4]. The functional diversity of RNAs, and in particular the ability of certain RNAs, called ribozymes, to act as enzymes, has led some to hypothesize that RNA formed the basis of the first self-replicating ancestors to life as we know it (this is known as the *RNA World* hypothesis [5, 6]).

For many of the roles that it performs, an RNA's structure is critical to its proper function. Single-stranded RNAs fold into complex three-dimensional structures in solution. In the late 1950s and early 1960s, RNA structures were found to be composed of helices similar to the helices of double stranded DNA, formed from intramolecular pairing of regions of the RNA sequence [7, 8]. To bring these regions together, the RNA backbone loops back, and so regions that pair must be reverse complements (fig. 1.1). As with DNA, paired bases are hydrogen-bonded, and so only certain combinations of bases — those that can form the appropriate hydrogen bonds — can be paired. For RNA, the canonical base-pairs are AU, GC, and GU (certain non-canonical base-pairings also occur, but much less frequently [9]). Depending on

the sequence, different sets of helices are formed; and the three-dimensional structure of the RNA is composed from these helices, positioned in three-dimensional space.

These early studies conceptually divided RNA structure into two parts, secondary and tertiary structure. Secondary structure is the set of intramolecular hydrogen-bonded base-pairs, and the loops (unpaired regions) that connect them together. Adjacent base-pairs form double-stranded helices which are stabilized by relatively strong stacking interactions. Tertiary structure specifies the three-dimensional co-ordinates of the secondary structure 'domains'. Because secondary structure captures the most salient features of the overall structure, it is on this level of abstraction that RNA structure is usually studied (the use of 'structure' in the proceeding text refers to secondary structure unless otherwise specified).

Figure 1.1 shows several visual representations of secondary structure. For our purposes, an RNA's secondary structure is defined as the set of base-pair positions. Adjacent base-pairs form helices (position-pairs adjacent to pair $(i, j) := (i + 1, j - 1)$ and $(i - 1, j + 1)$), while the stretches of unpaired sequence are referred to as loops.

If an RNA secondary structure is pseudoknot-free, it contains only nested base-pairs (i.e. there are no two pairs $i : j$ and $i' : j'$ in the structure whose sequence positions are $i < i' < j < j'$). At a single moment in time, a single position may be paired with only one other position. However, as we shall see, an RNA's secondary structure does not remain fixed, and so over an RNA's lifetime, a single position may pair with more than one other position.

## 1.2 RNA structure prediction

Predicting RNA secondary structure from sequence is a venerable problem in molecular biology. At first, computational approaches to RNA secondary structure prediction involved human analysis of computed dot-matrices [11, 12]. Later, fully computational methods were developed [13, 14], and since then, many programs for prediction of RNA secondary structure have been produced.

Mfold [15], one of the most commonly used programs for RNA secondary structure prediction, identifies the thermodynamically most stable pseudoknot-free secondary structure. The search for this so-called minimum free energy (MFE) structure characterizes many methods of RNA structure prediction. There exists an efficient dynamic programming algorithm to calculate the pseudoknot-free MFE structure, given an energy model [13]. These methods are predicated on the assumption that the RNA molecule exists for the most part in thermodynamic equilibrium, so its most likely structure would be the MFE structure [11]. However, RNA sequences *in vivo* do not necessarily exist in thermodynamic equilibrium. For some RNA molecules, it may take a long time, longer than the lifetime of the RNA, to fold into its minimum free energy structure [16]. Additionally, RNAs may interact with proteins, other RNAs, or other cell contents, altering its energy landscape [17, 18].

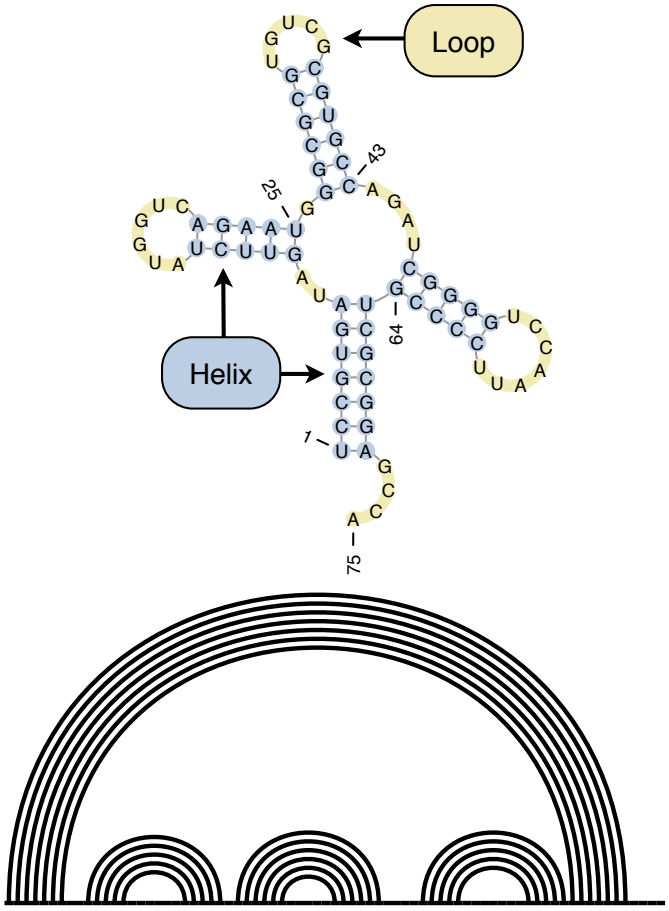


Figure 1.1: Two pictorial representations of the secondary structure of a tRNA. The upper image is generated by PseudoViewer [10]. The lower image shows the structure as an arc diagram. In this image, the horizontal line represents the sequence and each arc represents a base-pair.

A common approach to improve RNA structure prediction performance is to predict a common secondary structure for a set of related sequences rather than for a single sequence [19, 20]. If RNA structure is functional, then it will be conserved in evolution, and this conservation of structure can be leveraged to improve prediction. The telltale sign of structure conservation is the pattern of covariation of base-paired homologous positions of each sequence (fig. 1.2). The nucleotides in these alignment columns may not be identical, but if there is a mutation at one position, its pairing partner is likely to also mutated in such a way as to allow these positions to pair. For alignments with many related sequences, the analysis of covarying positions has also been used to identify certain tertiary structure motifs [21].

Because the conserved structure is likely to be the same as the functional structure *in vivo*, the secondary structure information contained in the evolutionary signal is not subject to the same concerns over cellular context that are problematic for purely MFE-based approaches. Interactions with other molecules within the cell may affect the RNA's structure, but evolution acts on the structure as it occurs in nature with these interactions taken into account.

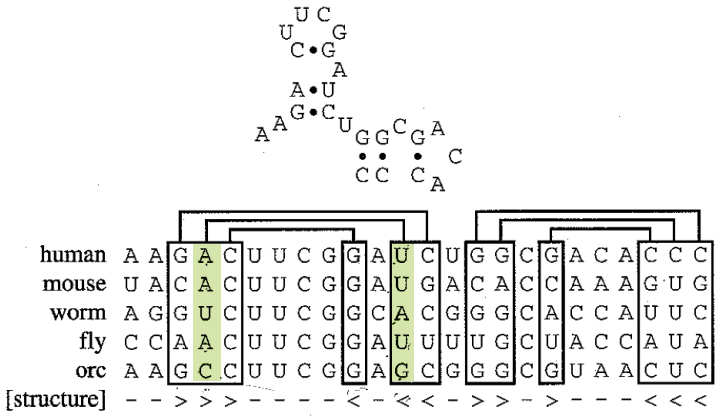


Figure 1.2: Hypothetical alignment showing covariation of base-paired columns. In the highlighted columns, the worm and orc sequences are different from the other sequences, but in both cases, the change still allows those positions to pair. Mutations which disrupt pairing are disfavored by evolution, whereas mutations that preserve pairing potential are likely to neutral. Because of this evolutionary pressure, columns of the alignment which are base-paired in the functional structure do not evolve independently; instead, they covary to preserve pairing potential. Adapted from Durbin *et al.* [22].

Many programs which predict the secondary structure of a group of related RNAs are available, of which I will describe only three:

1. RNAalifold [23] is an example of a program that combines a MFE-based approach with information on sequence conservation from an alignment. It takes as input a fixed alignment of RNA sequences, and predicts a pseudoknot-free consensus secondary structure for that alignment. It does so by adjusting the energy parameters of specific

base-pairs based on their pattern of sequence conservation. With these new alignment-specific energy parameters, it calculates the MFE structure using the standard dynamic programming techniques.

2. The program PFOLD [24, 25] also predicts a pseudoknot-free structure for a fixed alignment of related sequences and a phylogenetic tree. The advantage of PFOLD is that it uses explicit probabilistic models of sequence evolution and secondary structure, endowing its predictions with a probabilistic interpretation. It models secondary structure with a stochastic context-free grammar (SCFG), and uses mutation rate matrices to model sequence evolution for paired and unpaired columns (these models are described in detail in section 2.4.1). PFOLD uses the CYK algorithm to identify the most probable structure, given the alignment and tree. Like RNAALIFOLD, PFOLD's performance is highly dependent on the quality of the alignment.

3. The program SIMULFOLD [26] can predict a phylogenetic tree, alignment, and secondary structure for a set of related sequences. It takes a Bayesian approach to structure prediction, drawing structure/alignment/tree $(S, A, T)$ triples from the posterior probability of these triples given the input sequences. It uses a Markov chain Monte Carlo (MCMC) method to draw this sample, wherein it starts with an initial $S, A, T$ triple and moves through the space of possible structures, alignments and trees by proposing small local changes to the structure, alignment, or tree and then either accepting or rejecting those changes. At well-spaced intervals, it adds its current $S, A, T$ 'position' to the pool of samples. This pool of samples can then be summarized into a single predicted consensus structure, alignment, and tree, and also used to ascertain the uncertainty in the various parts of this prediction. SIMULFOLD can predict pseudoknotted structures.

   The rationale for predicting structure, tree, and alignment simultaneously is that these tasks are interconnected; partial information on any of the three help to constrain the search for the other two [27]. By predicting all three simultaneously, SIMULFOLD uses the information from all three tasks to produce mutually consistent, high-quality structure, tree, and alignment predictions. Several other programs also combine the tasks of alignment and structure prediction, including RNA SAMPLER [28] and DYNALIGN [29].

## 1.3 RNAs with multiple functional conformations

At physiological temperatures, an RNA will not permanently fold into a single secondary structure; base-pairs and helices are added or removed stochastically over the course its lifetime. These structural fluctuations are usually ignored because they are assumed to be mostly minor and inconsequential. The point at issue is whether there are multiple *functional* structures that an RNA folds into over the course of its lifetime. A precise definition of

what counts as a functional structure is not entirely easy to pin down, but essentially it is one that, if disrupted, would adversely affect the ability of the RNA to do its job, or, at a slightly higher-level, adversely affect the survival of the organism it occurs in. In the following section, we describe some examples of RNAs that, at different times, adopt different functional structures.

### 1.3.1 Riboswitches

Many examples of RNA switches, which alternate between two structures with different functions, have been characterized. Riboswitches, the first of which was identified in 2002 [30], are perhaps the most prominent example of this class. Riboswitches are found in the untranslated regions of some prokaryotic and, less commonly, eukaryotic mRNAs [31, 32]. They are composed of two overlapping domains: the aptamer domain and the expression platform. The aptamer domain forms a structure which can bind a metabolite ligand, while the expression platform affects the mRNA's expression in some way. Expression platforms may form terminator stems to halt transcription [33], block the ribosome binding site to prevent translation [34], or affect the splicing efficiency of an intron [35]. Together, these domains fold into two possible structures (the ON and OFF structures), each of which have different effects on the expression of the mRNA, depending on the particular expression platform. Crucially, ligand-binding favors the ON structure, allowing the riboswitch to 'sense' the environment and adjust expression accordingly. Figure 1.3 outlines this mechanism diagrammatically.

Similarly, thermosensors are hairpins which affect the mRNA's expression, and are destabilized at high temperatures [36]. Protein-mediated RNA switches have also been identified, which act in a similar manner to riboswitches except that in this case, a protein binds directly to a primary-sequence binding site on the RNA to favor one functional secondary structure over the other. Recently, such a protein-mediated RNA switch was identified in a human mRNA [37]. Synthetic ribozymes have also been constructed which can fold into two alternative structures, each catalyzing a different reaction [38].

### 1.3.2 Folding pathways

Denatured RNA, when allowed to refold *in vitro*, often misfolds into inactive structures which are rare or absent in *in vivo* populations of the same RNA [39]. Observations such as this have spurred interest in folding pathways, the sequence of structures that are formed as an RNA folds from its original unfolded state. *In vivo*, folding takes place co-transcriptionally [40]. Partially transcribed RNAs have limited folding 'options', but as transcription progresses, the set of *possible* structures that the RNA can fold into increases rapidly. However, some of those options may be blocked by a requirement to rearrange parts of the region that has already folded. The folding pathway is therefore thought to be critical to the formation of many RNA's functional structures *in vivo* [41–43].

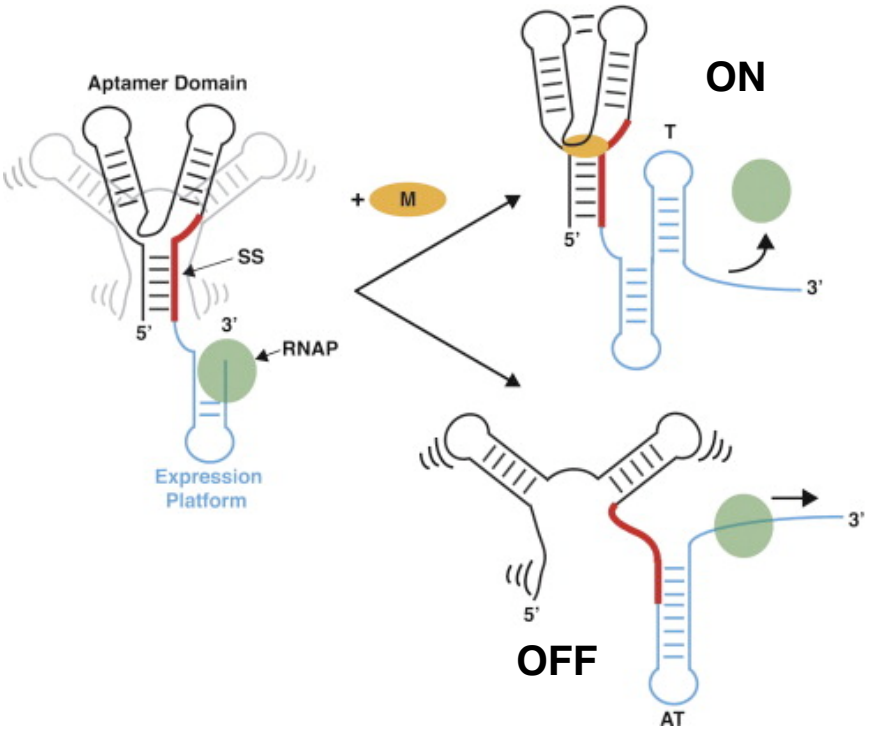


Figure 1.3: Diagram of the mechanism for riboswitch regulation of transcription. If the metabolite (M) binds to the aptamer domain, leading to the formation of a terminator stem (T). In the absence of the metabolite, a helix of the aptamer domain is displaced by the antiterminator stem (AT), which prevents the formation of the terminator stem and allows the RNA polymerase (RNAP) to continue transcription. The switching sequence region (SS), highlighted in red, is base-paired in differently in the alternative structures. Adapted from Garst and Batey [31].

Co-transcriptional folding pathways often involve transient helices, which are formed and subsequently disrupted as transcription progresses. These transient structures may have functions of their own; for instance, the metastable structure formed as the potato spindle tuber viroid (PSTVd) is transcribed is necessary for replication [44].

Transient structures may also be functional in that they may guide the folding process by blocking certain alternative folding pathways. Transient structures which increase the efficiency of co-transcriptional folding (i.e. decrease the probability of misfolding) have been identified experimentally in *E. coli* RNase P RNA [45]. Additionally, Xayaphoummine et al. [46] have shown that co-transcriptional folding paths can be encoded in transient as well as native helices by constructing synthetic RNAs with transient structural features. In a statistical analysis of possible transient helices in a wide variety of RNAs with known structure, Meyer and Miklós [47] found asymmetries in the relative proportion of transient helices which lie 5' versus 3' from the known helix it competes with, which may be due to enrichment of helices that play a role in the co-transcriptional folding pathway.

## 1.4 Identifying multiple functional structures: computational approaches

### 1.4.1 Alternative structures at thermodynamic equilibrium

Some variations of classic thermodynamics-based RNA structure prediction have been applied to the problem of identifying RNAs with alternative structures. In a population of identical RNA molecules at thermodynamic equilibrium, not all of the molecules will assume the minimum free energy structure. Instead, the population will exhibit a range of structures over the energy landscape of possible structures. The proportion of the population with structure $x$ (in other words, the probability $P(x)$) with a free energy $G_x$ follows the Boltzmann distribution, given by the formula

$$P(x) = \frac{\exp\left(-G_x/RT\right)}{Q} = \frac{\exp\left(-G_x/RT\right)}{\sum_{y \in S} \exp\left(-G_y/RT\right)} \tag{1.1}$$

where $T$ is temperature in Kelvin, $R$ is the universal gas constant, and $Q$ (the partition function) is the sum of the Boltzmann-weighted free energies of every structure in the set of possible structures $S$. If one ignores pseudoknotted structures, the partition function can be calculated efficiently with the McCaskill algorithm [48]. If two sufficiently different structures both occupy a significant fraction of the probability space, then this is taken as evidence of multiple functional structures. The McCaskill algorithm can also be used to calculate the probability that any two sequence positions are paired, which is useful for identifying specific positions that might be alternatively paired.

The program RNAsubopt [49] produces a list of all structures below a certain energy

cutoff, which can be analyzed to discover which structures occupy significant portions of the probability space. Enumerating enough structures to capture most of the structure probability, however, is not possible for longer sequences (> 100 nt). Moreover, since the total number possible structures increases exponentially with sequence length [50], the probability of any particular structure is not particularly useful for identifying potential alternative structures. Some method must be used to group similar structures together, so that one can identify groups with high total probability. Voß et al. [51] formalized this grouping process by defining abstraction functions to map structures to 'RNA shapes', and then calculating exact total probabilities for these shapes. The runtime for this method grows exponentially with sequence length, making it unsuitable for longer sequences (> 100 nt). However, for longer sequences, it is possible to sample structures from the Boltzmann distribution in polynomial time [52], and to then apply the RNA shapes abstraction to the set of samples to estimate the shape probability. This approach has been successful at recovering alternative structures for certain known RNA switches [51], and has also been reformulated as a comparative method (RNAlishapes), and has had some success at identifying evolutionarily conserved alternative structures [53].

Given enough time, a population of RNAs will eventually reach thermodynamic equilibrium. However, the energy barriers to structural rearrangement may be so high that such equilibration takes much longer than average lifespan of an RNA *in vivo* [16]. Instead, an RNA may become trapped in a local minimum of the energy landscape where the barriers to escape are high, slowing the rate of escape. Such kinetically trapped structures may represent the functional structure, or a misfolded structure; effective folding pathways might steer folding toward functional kinetic traps and/or away from misfold kinetic traps. Thermodynamics-based approaches are limited to identifying alternative structures occupying a significant portion of the Boltzmann distribution of structures. However, for kinetically trapped structures, this is not necessarily the case, since the rate of escape from such a trap is dependent on the height of the energy barrier, not the depth of the trap. These approaches also ignore the effect of interactions with other molecules (proteins, other RNAs, etc.) on the energy landscape.

### 1.4.2 Modeling RNA folding

Interest in RNA folding pathways has spurred the development of computational methods for RNA structure prediction which take the kinetics of RNA folding (i.e. the rates at which structures are folded) into account. To do so, most computational approaches directly model the physical process by which an unfolded, fully synthesized RNA folds into its functional conformation(s) as a continuous-time Markov process which allows only local rearrangements

of secondary structure [54]. This process is described by a master equation:

$$\frac{d}{dt}P_x(t) = \sum_{x \neq y} \left(P_y(t)k_{xy} - P_x(t)k_{yx}\right) \tag{1.2}$$

where $P_x(t)$ is the probability of being in state $x$ at time $t$, and $k_{xy}$ is the transition rate from state $y$ to state $x$. Given the full transition rate matrix $\mathbf{K}$ containing transition rates between all pairs of possible structures (most of which will be 0), the vector of probabilities for all states at a given time can be calculated from

$$P(t) = e^{t\mathbf{K}}P(0) \tag{1.3}$$

However, as the state space (the space of all possible secondary structures) is very large for RNAs of biological interest, it is generally not feasible to calculate the full transition matrix. However, folding trajectories can be sampled using stochastic Monte Carlo simulation of the Markov process. Several programs, including KINFOLD [55] and KINEFOLD (note the extra 'e') [56–58] employ this method, though they differ significantly in their implementation.

KINFOLD defines legal transitions as the formation, disruption, or shifting of a single base-pair, so the folding trajectories it generates are very fine-grained, specifying when each base-pair is added or removed. Transition rates from the current state to any neighboring state can be estimated by the Arrhenius equation

$$k_{xy} = \Gamma e^{-(G^{\dagger}_{xy} - G_x)/RT} \tag{1.4}$$

where $G_x$ is the free energy of structure $x$, $G^{\dagger}_{xy}$ is energy barrier between $x$ and $y$, corresponding to the free energy of the highest energy structure formed during such transition (the transition state), $\Gamma$ is a constant generally chosen to fit the timescale to experimentally validated data. KINFOLD assumes that $G^{\dagger}_{xy} = G_y$ because the moves it allows are so small that there is arguably no true transition state.

In KINEFOLD, transitions add or remove a single helices, a simplification which reduces the number of legal transitions from any state but also requires a more complex estimation of the transition state energy; it assumes the energy barrier is the energy required to nucleate three base pairs from the helix, plus the energy required to displace any helices blocking the added helix. Additionally, KINEFOLD allows pseudoknotted structures, which is more biologically appropriate but requires a more complicated energy model than the standard Turner model [59] used by KINFOLD (KINFOLD does not allow pseudoknots). Interestingly, KINEFOLD also takes into account certain topological constraints induced by pseudoknots which may kinetically trap other helices [58]. In addition to simulating the refolding of an RNA from an unfolded, fully synthesized state, both programs can be easily adapted to simulate co-transcription folding by dividing the sequence into transcribed and untranscribed

regions whose boundary shifts 5' to 3' at a certain rate, and restricting legal moves to those that form no base-pairs in the untranscribed region.

Other computational approaches consider energy landscapes in order to reduce the size of the state-space to the point where it is computationally feasible to find the solution to the master equation. The energy landscape can also be abstracted as a barrier tree, where the local minima are represented as leaves in the tree, connected to one or more gradient basins by saddle-points, the lowest energy structure that connects the gradient basins around these local minima [55, 60]. Constructing a barrier tree representation of the complete energy landscape would require the consideration of all possible structures; however, barrier trees constructed from a list of the lowest-energy structures (generated with RNAsubopt [49]) usually capture the most relevant features of the energy landscape for sufficiently short sequences ($< 100$ nt). Wolfinger et al. [61] define the folding process state-space as the basins around local minima of the energy landscape, and calculate the transition rates between adjacent basins using a variation of the flooding algorithm used to construct barrier trees. This state space is sufficiently small to solve the master equation. Barrier trees are also useful to help interpret folding trajectories sampled with Monte Carlo simulations [55].

A similar approach is taken by Tang et al. [62, 63], where the folding landscape is approximated by a probabilistic roadmap which defines the allowed transitions between states. In [63], the state-space is restricted to a set of secondary structures probabilistically sampled from the Boltzmann distribution of energies. Transitions are only allowed to the nearest $k$ neighbors, with energy barriers estimated heuristically. Ideally, these states capture the main features of the folding landscape but are few enough to solve the master equation (though it is also possible to do Monte Carlo simulation here).

Zhang and Chen [64–66] partition the structure space into clusters based on the presence or absence of certain rate-limiting base-stacks, which have particularly high energy barriers to their formation or disruption. The distribution of structures within clusters is assumed to be at thermodynamic equilibrium, so the transition rates between clusters can be calculated by summing the rates of transition between the structures at the boundaries of clusters, adjusted for the probability of the boundary structure in its cluster (the transition rate calculation in [61] requires a similar assumption).

These energy-landscape-based methods are not easily applicable to the analysis of co-transcriptional folding, since, as transcription progresses, nucleotides are added to the sequence undergoing folding, which changes the energy landscape. However, by calculating energy landscapes for all partially transcribed subsequences and then mapping the local minima from each landscape onto its successor, one can adapt landscape-based methods to the co-transcriptional folding case [67]. The recently released BacMap program [68] takes this approach, and can be used to model a variety of situations involving dynamic energy landscapes, including co-transcriptional folding or changes in temperature or ion concentration.

Long sequences are problematic for all the above methods since the space of possible secondary structures, and therefore the worst-case complexity of the energy landscape, grows exponentially with sequence length. The KINWALKER program [69] was designed to allow simple kinetics analysis for long sequences (400–1000+ nt). For this, it dispenses with simulation and instead deterministically generates a single candidate co-transcriptional folding pathway pieced together from combinations of pre-computed subsequence minimum free-energy structures. The method is kinetic in that it allows the incorporation of an MFE substructure only if the energy barrier between the current structure and the resulting merged structure that the transition would occur in a reasonable time (i.e. before the next transcription step). Calculating the energy barrier between two arbitrary structures is an NP-complete problem [70], so KINWALKER employs a heuristic for estimating such barriers. KINWALKER may be said to find the MFE structure at each transcription step, subject to the constraint that the transitions between structures be kinetically feasible.

Evaluating the performance of individual methods of predicting folding pathways is problematic for several reasons. Detailed experimental investigations of folding kinetics, usually done via temperature- or pH-jump kinetic trapping procedures [71] or single-molecule 'optical-tweezer' manipulation [72], are only available for a small number of sequences. These sequences for are generally quite short ($< 100$) nt). Also, there are no standard metrics for comparing experimental results with output from computational methods (which itself varies greatly from method to method). Consequently, most computational RNA kinetics methods are evaluated by qualitative comparison with a few chosen experimentally investigated sequences, though some more thorough performance evaluations for single methods have been undertaken (e.g. [65, 72, 73]).

From a theoretical perspective, the simulation of RNA folding is problematic because there are many external factors which may influence the folding pathway *in vivo*. Transcription speed [44, 45, 74], the binding of proteins [75, 76], metabolites [77], and other RNAs [46], and ion concentrations [78] have all been shown to influence folding pathways. Certain RNA structures formed during transcription can also influence transcription speed, which in turn influences folding [45], forming a kind of feedback loop. Incorporating this level of complexity into a simulation would be next to impossible, so any simulation will necessarily be imperfect, though it remains to be seen how much of this complexity can safely be ignored.

TRANSAT, the program we developed, is part of an attempt to break away from the simulation paradigm. Rather than attempting to simulate the folding process, our program looks for helices in an alignment of related sequences that are evolutionarily conserved. Unlike most programs for secondary structure prediction, we allow these helices to overlap with each other, since the helices which play a role in the folding pathway may not be present in the final structure of the RNA. This method by itself cannot reconstruct the full time-series of structures adopted by an RNA as it folds, but it can identify conserved alternative structures

which play a role in an RNA's folding pathway.

# Chapter 2

# TRANSAT: methods

## 2.1 Motivation

The pattern of evolution in a set of homologous RNA sequences can provide information on the structural features of those sequences [19]. If a structural feature is functionally important, the ability of a sequence to form that feature will be conserved by evolution. Over time, sequences accumulate mutations, and certain mutations at positions which are base-paired will disrupt the ability of those positions to pair (e.g. a G-to-A transition at a position that pairs with C), while others may not (e.g. a A-to-G transition at a position that pairs with U). Mutations which disrupt pairing are disfavored by evolution, while mutations that preserve pairing potential are likely to neutral. Consequently, we observe co-variation in the functionally important paired positions of an alignment (fig. 1.2).

Comparative secondary structure prediction programs, which predict a consensus structure for a set of related RNAs, provide more accurate predictions of secondary structure than methods which consider only a single sequence [20]. Since evolution acts on *in vivo* structures, comparative methods should be able to identify structures that may be obscured by methods which try to model folding directly. Comparative methods are also the only way to identify unstructured regions; one can predict the minimum free energy structure for almost any given RNA sequence, but only comparative methods are able to detect the absence of (conserved) structure. Theoretically, all functional helices, regardless of their stability or transience, should be evolutionarily conserved (though not necessarily equally conserved), and therefore comparative methods should be applicable to folding pathway prediction.

Because of the complexities of the *in vivo* folding process, we decided to focus on identifying specific structural features which are conserved rather than trying to predict entire folding pathways. In doing so, we lose the ability to predict entire secondary structures as a function of time. Instead, we predict a list of helices that are evolutionarily conserved. We expect the RNA to form each of these helices at some point in time (or under certain conditions), though not necessarily simultaneously. If, in this list, two helices are mutually incompatible, we take this as evidence that the sequences in the alignment contain functional transient structures. We developed the program TRANSAT to make this sort of prediction.

## 2.2 Strategy

Our basic strategy with TRANSAT is to examine a given input alignment at the level of individual helices rather than full structures. A very large number of possible helices could exist in any sequence of the alignment, however, so our first objective is to separate the evolutionarily conserved helices from the others.

TRANSAT takes as input a multiple sequence alignment of related RNA sequences, and a phylogenetic tree laying out the evolutionary relationships between those sequences (fig. 2.1). The starting point for the analysis is the identification of every possible helix that could form in any sequence of the alignment. After we identify these helices for each individual sequence, we ask which helices are conserved, allowing us to both evaluate how well the method captures elements of the known structure and look for structures which conflict with the known structure but are similarly conserved. We then evaluate each helix by mapping it back to the alignment and assigning it a score based on how well it conforms to our model of sequence evolution in RNA helices. This score is converted into a p-value by comparing it to a null distribution of helix scores generated from shuffled (and therefore unstructured) versions of the same alignment. The method outputs a list of helices together with their p-values. Figure 2.1 outlines TRANSAT's analysis pipeline.

## 2.3 Helix identification

We define a helix as a pair of contiguous subsequences of equal length that can form canonical base-pairs with each other. More specifically, the first (5'-most) position of the 5' region must be able to pair with the last (3'-most) position of the 3'-region, the second position of the 5' region must be able to pair with the second last position of the 3'-region, and so on. The following section spells out this definition more precisely:

### 2.3.1 Helix definition

We are given an alignment of $n$ aligned sequences $\mathcal{A} = (s_1, \ldots, s_n)$, and a tree $T$ relating those sequences together. Each aligned sequence is composed of nucleotides or gaps $s_i = (x_1, \ldots, x_m)$, and all sequences have the same length $m$. By removing the gaps from a sequence $s_i$, we get the unaligned sequence $\hat{s}_i = (\hat{x}_1, \ldots, \hat{x}_{m_i})$, where $m_i$ is the length of the unaligned sequence. For the given alignment $\mathcal{A}$, each position $j$ in the ungapped sequence $\hat{s}_i$ is mapped to its corresponding position in the aligned sequence by the function $M_i(j)$. This mapping maintains the ordering of the sequence, i.e. if $n < m$ and neither $x_n$ nor $x_m$ are gaps, $M_i(n) < M_i(m)$.

Let $\zeta_i(j,k) = 1$ if the pair of nucleotides $(\hat{x}_j, \hat{x}_k)$ in sequence $\hat{s}_i$ is in the set of canonical base-pairs $\mathcal{B} = \{AU, UA, GC, CG, GU, UG\}$, and otherwise 0. We define a helix in an

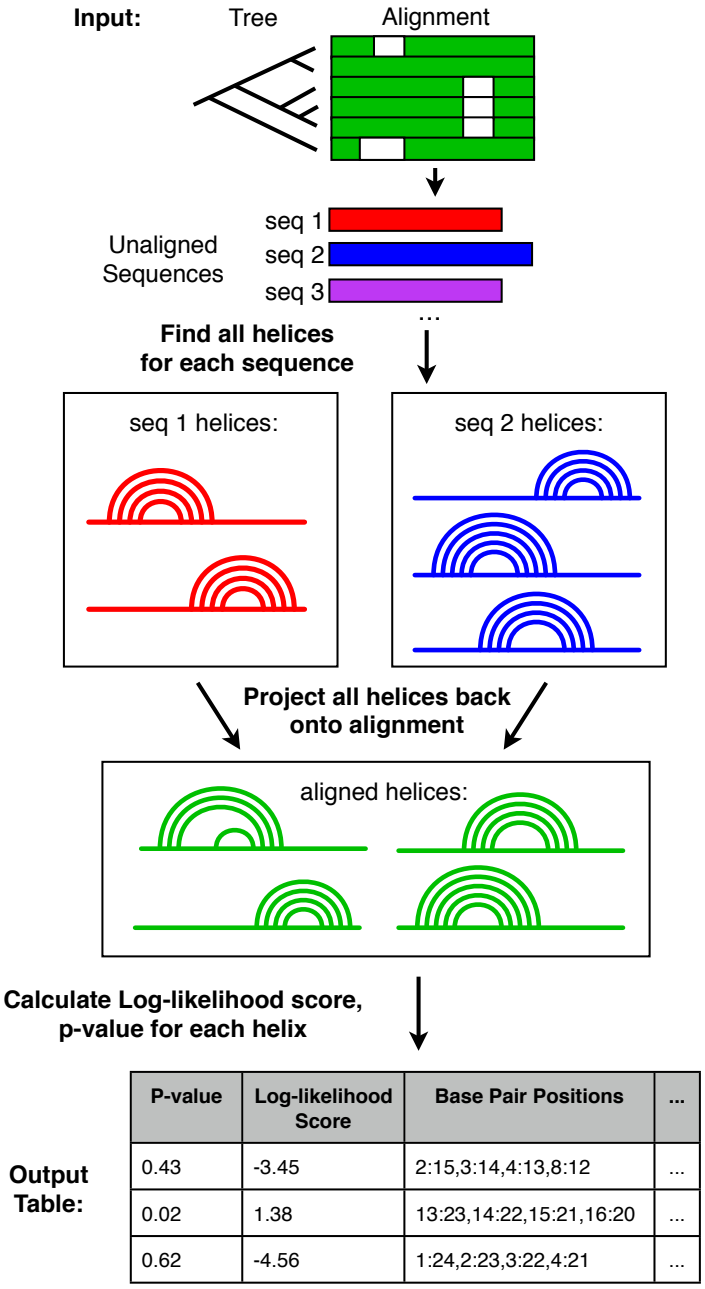


Figure 2.1: Overview of TRANSAT analysis: TRANSAT takes as input a multiple sequence alignment of RNA sequences, and a phylogenetic tree relating those sequences together. All possible helices are identified from each sequence, and projected back onto the alignment. Certain helices from different sequences may be identical when projected onto the alignment; duplicate helices are filtered out at this stage. Each helix is then assigned a log-likelihood ratio and a p-value, and outputted in the form of a table, together with other relevant information about the helix.

ungapped sequence $\hat{s}_i$ (an ungapped helix) to be a set of paired positions

$$\hat{h} := \{(j,k),(j+1,k-1),\ldots,(j+l-1,k-l+1)\} \tag{2.1}$$

where $l$ is the length of the helix in base-pairs. For $\hat{h}$ to be a valid helix, all paired positions must make valid base-pairs, i.e.

$$\sum_{(j,k)\in h} \zeta_i(j,k) = l \tag{2.2}$$

An aligned helix $h$ is the set of positions from an ungapped sequence $\hat{h}$ mapped to their corresponding positions in the alignment:

$$h := \left\{(M_i(j), M_i(k)) \text{ for all } (j,k) \in \hat{h}\right\} \tag{2.3}$$

In the following text, the terms 'helix' and 'aligned helix' are used interchangeably. Note that by our helix definition, bulges are not allowed in unaligned helices (i.e. the positions in each region of an unaligned helix must be contiguous), but the mapping procedure may introduce bulges or inner loops into the aligned helix (fig. 2.2).

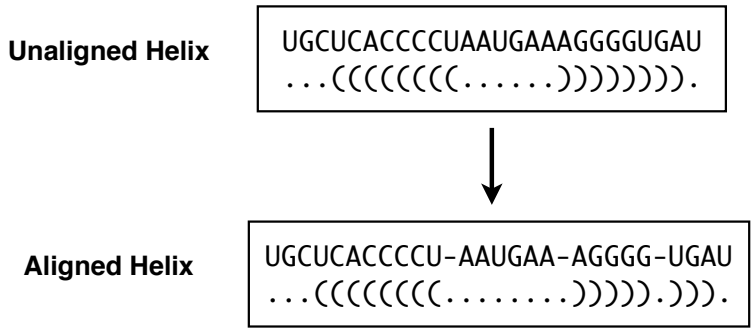


Figure 2.2: Helices identified in ungapped sequences are then projected back onto the alignment. Unaligned helices are always composed of base-pairs which are adjacent, but bulges may be introduced when the helix is projected on to the alignment.

### 2.3.2 Finding helices

Different sequences in an alignment are typically capable of forming different helices. To capture the full range of helices that may be found in the alignment, we identify possible ungapped helices in each sequence, and map them all back to the alignment. However, we

do ignore ungapped helices whose length is less than a certain threshold (default: helices of length $< 4$ are ignored — see fig. 2.3 for justification of this choice) and ungapped helices with loops of fewer than 3 nucleotides.

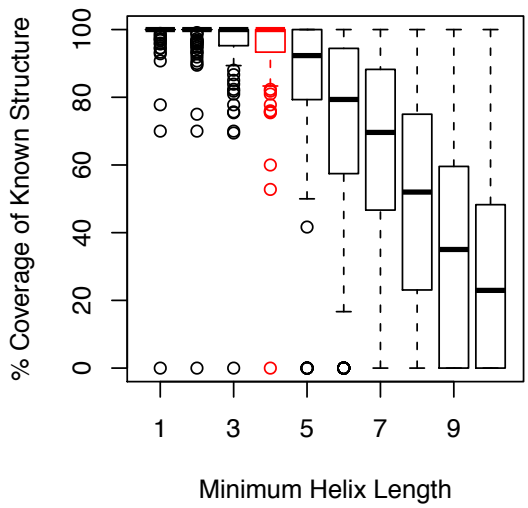


Figure 2.3: Box-and-whiskers plot showing the coverage of known structure base-pairs in alignments of the RFAM dataset (see section 3.1.3 for details) by helices identified by TRANSAT at different minimum helix length settings. A base-pair of the known structure is considered covered if TRANSAT finds a helix containing which contains that base-pair and is also composed of $> 70\%$ known structure base-pairs. Our default minimum helix length setting of 4 is highlighted in red. Shorter minimum helix lengths provide slightly better coverage, but cause TRANSAT to consider many more helices, increasing TRANSAT's running time. Each box depicts the upper and lower quartiles (the central line is the median). The whiskers extend to the maximum and minimum data points within 1.5 times the interquartile range (IQR) of the upper or lower quartile. The circles represent outlier data-points, i.e. those further than $1.5 \times$ IQR from the upper or lower quartiles. Note: for minimum helix length values of 1 to 4, there is one alignment which has zero coverage. In this alignment (RFAM family RF01288), the known structure is annotated as a single helix with 3 base-pairs, but the positions inner and outer to those base-pairs form canonical base-pairs in all the sequences of the alignment. TRANSAT identifies only maximum length helices, so the helix found by TRANSAT at this position does not meet the 70% threshold.

For each sequence in the alignment, we remove the gaps from the sequence, simultaneously storing the mapping of ungapped to aligned sequence positions. We then use a simple algorithm to identify all ungapped helices in the sequence.

The helix-finding algorithm can be phrased as a dynamic programming algorithm similar to the Nussinov algorithm [13], but greatly simplified. Let $F(i, j)$ be the length of the longest

helix whose outermost base-pair is $(\hat{x}_i, \hat{x}_j)$. Values for $F(i,j)$ can be calculated recursively:

$$F(i,j) := \begin{cases} 1 + F(i+1, j-1) & \text{if } \zeta_x(i,j) = 1 \text{ and } (j-i) \geq \text{min loop length} \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Values for $F$ are stored in an $m_x \times m_x$ upper triangular matrix. $F(i,i)$ is initialized as 0 for $i = \{1, \ldots, m_x\}$, as is $F(i, i-1)$ for $i = \{2, \ldots, m_x\}$. Then one can then simply run through the matrix left to right, filling in the values for $F$ as you go.

Helices can be identified from the $F$ matrix by finding all $\{i,j\}$ pairs where $F(i,j)$ is greater or equal to the minimum helix length. These pairs correspond to helices consisting of $F(i,j)$ base-pairs, with an outer base pair at $(i,j)$. To reduce the number of helices considered, we skip helices which are subsets of larger helices that occur in the sequence. Therefore, we search the matrix for positions $(i,j)$ where $F(i,j)$ is greater than the minimum helix length and $F(i-1, j+1)$ is either 0 or undefined (at the edges of the matrix, when $i = 1$ or $j = m_x$).

In fact, there is no need to store the entire matrix of $F$ values. Each $F(i,j)$ value depends on the value of $F$ at only one pair of positions, and there is (at most) one other pair of positions that depends on it. Therefore, as long as we calculate values of $F(i,j)$ in such as way as to follow the chains of dependent positions, we need only store one $F$ value. If we picture this process as the filling in of the upper triangular matrix, this path corresponds to traveling along each of the diagonals, heading up and to the right. Algorithm 1 outlines this process in pseudocode.

Once a valid helix has been identified in a sequence, we map it to the alignment. For each sequence, we maintain a map of the sequence positions onto the alignment positions. To project the helix onto the alignment, we must convert the sequence positions of each base-pair in the helix to alignment positions. In this conversion, bulges may be introduced into the aligned helix, though unaligned helices are always bulge-free (fig. 2.2). When an aligned helix matches exactly to an aligned helix found in another sequence, we keep only one of them in our list of aligned helices. Using a hash set, the time taken to finding these matches is proportional to the lengths of the helices, since the hash function must see all the base-pairs of the helix. In the worse case could be $m/2$, where $m$ is the length of the alignment, but in practice the helices are mostly fairly short regardless of the length of the alignment.

Identification of the helices in all sequences of the alignment requires $O(m^2 n)$ time, where $m$ is the length of the alignment and $n$ is the number of sequences in the alignment, since one must run the $O(m^2)$ time helix-finding procedure for each of the $n$ sequences (ignoring the helix matching step), and requires $O(m^2)$ memory, since one has to store a list of the helices found and the number of possible helices is proportional to $O(m^2)$.

**Algorithm 1** findHelices

**input:** ungapped sequence $\hat{s}_x$ of length $m_x$, minimum helix length $L_{\text{helix}}$, minimum loop length $L_{\text{loop}}$

**for** $k = 1$ to $2 \cdot m_x - 1$ **do**
  bpCount $\leftarrow 0$
  {Diagonals start at $(i, j)$ and $(i, j+1)$}
  $i \leftarrow \lfloor (k+1)/2 \rfloor$
  $j \leftarrow \lfloor (k+2)/2 \rfloor$
  **while** $i > 0$ and $j \leq m_x$ **do**
    **if** $j - i > L_{\text{loop}}$ and $\zeta_x(i, j) = 1$ **then**
      bpCount++
    **else**
      **if** bpCount $\geq L_{\text{helix}}$ **then**
        record helix: outer-pair $= (i+1, j-1)$, length = bpCount
      **end if**
      bpCount $\leftarrow 0$
    **end if**
    $i--$, $j++$
  **end while**
  **if** bpCount $\geq L_{\text{helix}}$ **then**
    record helix: outer-pair $= (i+1, j-1)$, length = bpCount
  **end if**
**end for**

## 2.4 Log-likelihood score

The majority of helices found using the above procedure are spurious helices of no functional significance. Functional helices are expected to be evolutionarily conserved, so our goal is to identify those helices for which there is evidence of evolutionary conservation which is in line with the evolutionary relationships specified by the input tree T. Paired positions in helices which are functionally important evolve to maintain the nucleotides pairing potential, while unpaired positions are not under such constraints. To separate conserved helices from spurious ones, we look to see how the evolutionary pattern of the helix in the alignment compares to models of evolution in paired and unpaired positions.

To evaluate each helix, we compare two hypotheses about the evolutionary history of the alignment columns of the helix: either columns evolved to maintain pairing, or they did not. These two hypotheses we capture with probabilistic models of evolution in paired or unpaired columns.

### 2.4.1 Evolutionary models

The basis of this evaluation are our two probabilistic models of sequence evolution for paired and unpaired positions, $\theta_{\text{paired}}$ and $\theta_{\text{unpaired}}$. Our models take the form a rate matrix $\mathbf{R}$, whose entries specify the rates of conversion $r_{XY}$ from nucleotide $Y$ to $X$ (i.e. the expected number of such changes per unit time), written as:

$$\mathbf{R} = \begin{bmatrix} r_{\text{AA}} & r_{\text{AC}} & r_{\text{AG}} & r_{\text{AU}} \\ r_{\text{CA}} & r_{\text{CC}} & r_{\text{CG}} & r_{\text{CU}} \\ r_{\text{GA}} & r_{\text{GC}} & r_{\text{GG}} & r_{\text{GU}} \\ r_{\text{UA}} & r_{\text{UC}} & r_{\text{UG}} & r_{\text{UU}} \end{bmatrix} \tag{2.5}$$

The $r_{XX}$ values along the diagonal are:

$$r_{XX} = -\sum_{Y \neq X} r_{XY} \tag{2.6}$$

In these models, the rate of evolution is assumed to be constant over time. Models of this type are sometimes parameterized in such a way as to reduce the degrees of freedom of the model. For instance, the Jukes and Cantor model assumes that all transitions to different nucleotides occur at the same rate; the Kimura model differentiates between tranversions (purine to pyrimidine or visa versa) and purine-purine or pyrimidine-pyrimidine transitions [22].

From these rate matrices, we can calculate the probability $P(X|Y,t)$ that nucleotide $Y$ will evolve into nucleotide $X$ over time $t$. The substitution matrix $S(t)$ contains these probabilities for each $XY$ combination. If the rate matrix $\mathbf{R}$ is time-independent, $S(t)$ is related to $\mathbf{R}$ via

the equation

$$\frac{d}{dt}S(t) = S(t) \cdot \mathbf{R} \tag{2.7}$$

Solving for $S(t)$, we get the formula

$$S(t) = e^{\mathbf{R}t} \tag{2.8}$$

Calculating the substitution matrix can be made easier by decomposing $\mathbf{R}$ into its eigenvectors $\mathbf{V}$ and eigenvalues $\mathbf{\Lambda}$. $S(t)$ may then be computed from

$$S(t) = \mathbf{V}e^{\mathbf{\Lambda}t}\mathbf{V}^{-1} \tag{2.9}$$

For unpaired positions, evolution at one position is assumed to occur independently from all other positions. Therefore, finding the probability that one sequence evolved from another sequence is a simple matter of multiplying the transition probabilities at all sequence positions together.

At base-paired positions, the assumption that positions evolve independently is incorrect; paired positions show covariation to preserve pairing potential (fig. 1.2). The model for paired position evolution take this into account by considering a pair as a single entity and providing mutation rates between pairs of nucleotides. Since there are 16 possible nucleotide pairs, these rate matrices are $16 \times 16$ rather than $4 \times 4$. Transition probabilities are calculated as before, except that the resulting substitution matrix is also $16 \times 16$.

The models and paired and unpaired sequence evolution we use are the same as those used in Pfold [25] and Simulfold [26]. These models were trained on a set tRNA and rRNA sequences with well-known structures. Mutation rates were estimated from a set of high-similarity pairwise alignments of those sequences. The training procedure is described in detail in [24].

Clearly, these models greatly simplify the evolutionary process as it occurs in nature. They fail to capture the true complexity of biological sequence evolution in several respects:

1. In our models, no unpaired position is more strongly conserved than any other. Likewise, no paired positions are more strongly conserved than any other pair (though base-paired positions are more strongly conserved than unpaired positions). In nature, certain sequence positions may be more strongly conserved than others, depending on their importance to the sequence's function. For example, the HuR protein binds to hairpins in certain mRNAs. The binding sites for this protein have a very strongly conserved U at a specific position in the loop of the hairpin [79].

2. Rates of evolution are assumed to be constant in our models, whereas in nature, rates of evolution may differ greatly between lineages and change over time [80]. This does not pose a large problem for this study, since we are not interested in absolute times,

so this effect may be compensated for by assigning longer branch lengths to regions of the phylogenetic tree that are evolving more quickly. The method that we use for estimating phylogenetic trees should accomplish this.

3. The model only takes into account dependencies within a sequence, but RNA-RNA interactions between two different molecules can induce dependencies between sequences. For example, in the *hok* system, the target loop (the loop of stem 5 in fig. 3.1) is conserved to maintain complementarity with the 5'-end of the *sok* target RNA [81].

4. The models only take into account substitutions, not insertions or deletions. Instead, gaps are treated as missing information. Methods exist for including insertions and deletions into such models [82, 83], but we judged that the added complexity of these kinds of models was not suitable for our purposes.

5. It is very unlikely that two simultaneous mutations would occur at specific locations in nature, so substitutions under the paired model which change both positions do not represent a single action. Instead, they stand for two (or more) mutations which occur in a finite time interval.

6. The model parameters are derived from experimental observations, and are necessarily subject to error.

### 2.4.2 Felsenstein algorithm

Using these models, we can calculate $P(\mathcal{S}|\theta, T)$, the likelihood that a set of alignment columns $\mathcal{S}$ was produced by the model $\theta$ with a given phylogenetic tree $T$. This probability is calculated with the Felsenstein algorithm [84].

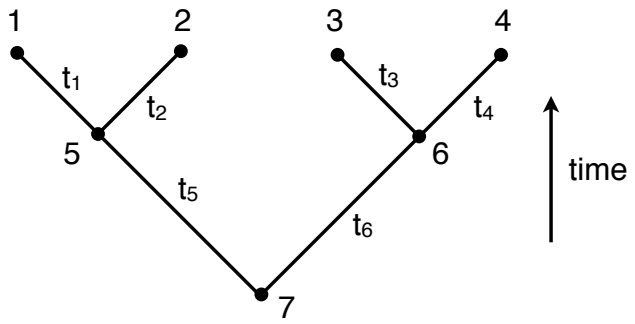


Figure 2.4: This diagram shows a hypothetical phylogenetic tree for a four-sequence alignment. Nodes 1-4 (the leaf nodes) represent the sequences in the alignment from organisms as we observe them today, while nodes 5-7 (the inner nodes) represent the ancestral sequences that gave rise to the alignment sequences. The root node is node 7.

The sequences in an alignment represent the product of evolution along a phylogenetic tree; each sequence corresponds to a leaf node in the tree. The inner nodes of the tree represent ancestral sequences. For the example shown in figure 2.4, if $\mathcal{S}$ contains a single column $C$ using the unpaired model (or a single pair of columns using the paired model), the likelihood calculation would be straightforward if we knew the ancestral sequences:

$$\begin{aligned} P(C|\theta,T) = P(s_7|\theta)P(s_5|s_7,t_5,\theta)P(s_1|s_5,t_1,\theta)P(s_2|s_5,t_2,\theta) \\ P(s_6|s_7,t_6,\theta)P(s_3|s_6,t_3,\theta)P(s_4|s_6,t_4,\theta) \end{aligned} \tag{2.10}$$

We obtain the $P(s_i|s_j,t,\theta)$ from the substitution matrix $S(t)$ of the model $\theta$.

However, since the ancestral sequences are unknown, we must sum over all possible assignments of those sequences:

$$\begin{aligned} P(C|\theta,T) = \sum_{s_5\in\mathcal{A}}\sum_{s_6\in\mathcal{A}}\sum_{s_7\in\mathcal{A}} P(s_7|\theta)P(s_5|s_7,t_5,\theta)P(s_1|s_5,t_1,\theta)P(s_2|s_5,t_2,\theta) \\ P(s_6|s_7,t_6,\theta)P(s_3|s_6,t_3,\theta)P(s_4|s_6,t_4,\theta) \end{aligned} \tag{2.11}$$

$\mathcal{A}$ is the alphabet of possible sequence assignments (i.e. A, U, G, or C in the unpaired model, or all pairs of two nucleotides for the paired model).

For even moderately sized trees, it is cumbersome to enumerate all the combinations of inner node assignments, but because the relationships between the sequences are structured as a tree, the summations can be moved rightwards to simplify the calculation:

$$\begin{aligned} P(C|\theta,T) = \sum_{s_7\in\mathcal{A}} P(s_7|\theta) \left[\sum_{s_5\in\mathcal{A}} P(s_5|s_7,t_5,\theta)P(s_1|s_5,t_1,\theta)P(s_2|s_5,t_2,\theta)\right] \\ \left[\sum_{s_6\in\mathcal{A}} P(s_6|s_7,t_6,\theta)P(s_3|s_6,t_3,\theta)P(s_4|s_6,t_4,\theta)\right] \end{aligned} \tag{2.12}$$

In general, to calculate the probability of seeing a particular nucleotide $k$ at node $x$ in the tree, one needs only the probabilities of each nucleotide at the children of node $x$ (we refer to these nodes as nodes $i$ and $j$ and the distance separating them from node $x$ as $t_i$ and $t_j$, respectively):

$$P(s_x = k|\theta,T) = \left[\sum_{s_i\in\mathcal{A}} P(s_i|s_x = k,t_i,\theta)P(s_i|\theta,T)\right]\left[\sum_{s_j\in\mathcal{A}} P(s_j|s_x = k,t_j,\theta)P(s_j|\theta,T)\right] \tag{2.13}$$

The $P(s_i|s_x = k,t_i,\theta)$ terms are known from the substitution matrix $S(t)$. At the leaf nodes, the probability of the nucleotide that appears in the node's sequence is 1, and the probability of the other nucleotides are 0. By traversing the tree from the leaves to the root, we can use

this formula to recursively calculate the likelihoods of each nucleotide for each inner node in the tree. The overall probability of the alignment column is taken by summing the product of the probability of each nucleotide at the root node and the nucleotide's prior probability:

$$P(C|\theta,T) = \sum_{s_{\text{root}} \in \mathcal{A}} P(s_{\text{root}}|\theta) P(s_{\text{root}}|\theta,T) \tag{2.14}$$

The Felsenstein algorithm performs a post-order traversal of the phylogenetic tree, calculating at each node the probabilities of each nucleotide (or pair of nucleotides, in the case of the paired model). The transition probability matrices for each branch need only be calculated once for each model. $P(s_{\text{root}}|\theta)$ is the prior probability of each nucleotide (or nucleotide pair). It is equal to the equilibrium frequency of the nucleotides, i.e. the entries along the diagonal of the substitution matrix $S(t)$ as $t$ approaches infinity.

In the unpaired model, columns are assumed to evolve independently of each other, so the likelihood of an alignment region is the product of the likelihoods the columns in that region:

$$P(\mathcal{S}|\theta,T) = \prod_{C \in \mathcal{S}} P(C|\theta,T) \tag{2.15}$$

In the paired model, paired columns are assumed to evolve independently of all other paired columns, so the likelihood of an alignment region is the product of the likelihoods of all pairs of columns.

Ordinarily, gaps in an alignment column are treated as missing information. When using the paired model, however, we treat a gap paired with a non-gap nucleotide as if the gap were a nucleotide which does not form a canonical base-pair with the non-gap nucleotide. For example, the gap in a gap-A pair would be treated as if it were a G or C, but not a U. This ensures that gap/non-gap pairs are properly penalized in the paired model likelihood calculation.

### 2.4.3 Calculating log-likelihood ratios for individual helices

We assign each helix $h$ identified in the alignment a log-likelihood ratio $\Lambda_h$, capturing the relative likelihood that the alignment columns of the (aligned) helix resulted from evolution on base-paired or unpaired nucleotides:

$$\Lambda_h = \log_2 \left( \frac{P(\mathcal{S}_h|\theta_{\text{paired}},T)}{P(\mathcal{S}_h|\theta_{\text{unpaired}},T)} \right) \frac{1}{l} \tag{2.16}$$

where $l$ is the length (number of base-pairs) of the helix and $\mathcal{S}_h$ is the set of columns in the region of the alignment covered by helix $h$ (fig. 2.5). We normalize the log-likelihood ratio by helix length in order to be able to compare the log-likelihood ratios of different helices. Normalizing for helix length is equivalent to taking the mean log-likelihood ratio of

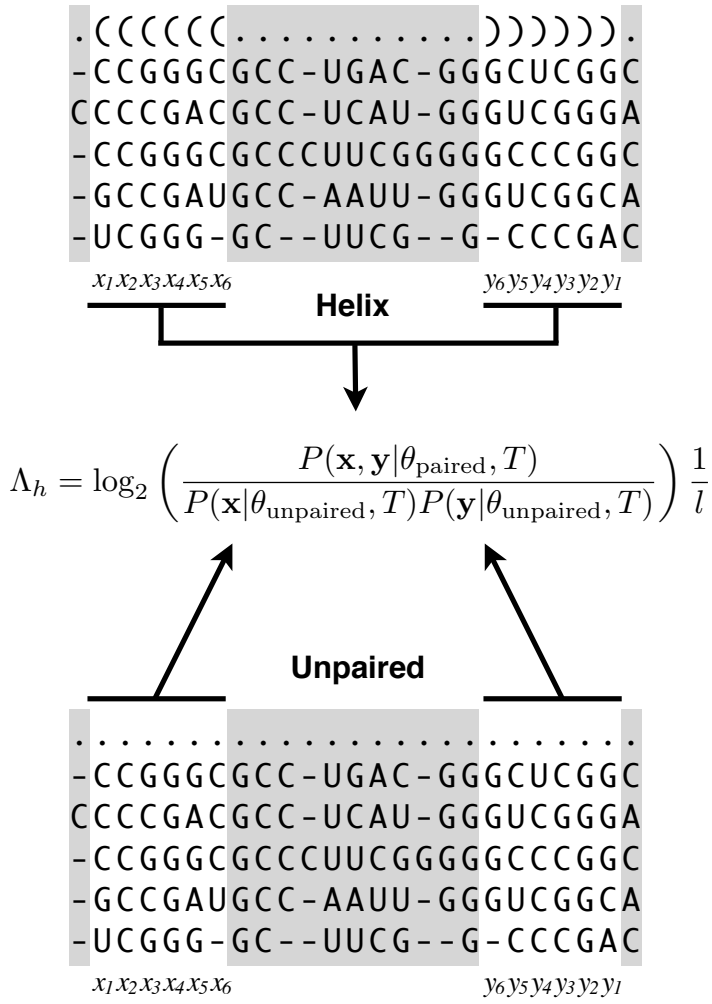


Figure 2.5: The log-likelihood score is calculated from the likelihood of observing the columns in the alignment given two alternate hypotheses about sequence evolution. The helix covers 12 columns, labelled $x_1, \ldots, x_6$, and $y_1, \ldots, y_6$ (this region is referred to as $\mathcal{S}_h$ in the text). In the unpaired model, columns evolve independently. In the paired model, pairs of alignment columns are assumed to evolve together to maintain base-pairing potential. The loop region is ignored in the calculation of a helix's log-likelihood ratio.

all base-pairs in the helix.

Since many of the possible helices share positions or base-pairs, we store the log-likelihood values of each alignment column or pair of alignment columns as we compute it, thereby saving us from having to compute it more than once.

## 2.5 Assigning p-values

The log-likelihood ratio score has a relatively straightforward probabilistic interpretation: a helix with a score greater than 0 is more likely to be base-paired than unpaired. However, the distribution of these scores varies widely for different alignments. This distribution is affected by, among other things, alignment size, alignment length, total tree length, and the amount of sequence variation in an alignment.

In order to be able to compare log-likelihood scores meaningfully across alignments, we convert log-likelihood scores into p-values by estimating the probability of finding high-scoring helices in shuffled, non-structured alignments. Our approach is outlined in figure 2.6. This approach is similar to the approach used in [85] to identify structured RNA genes in genomic sequence alignments. By shuffling the columns of the original alignment, we produce a set of alignments which are not expected to contain functional RNA structure. These randomized alignments do preserve the original alignment's nucleotide frequency. In [85], such randomized alignments were used as a baseline from which to assess the structure-forming potential of the original alignment (usually windows of a fixed length extracted from the genomic alignment); if the original alignment's minimum free energy (as estimated with RNAalifold [23]) is far outside the distribution of minimum free energies observed in the randomized set, the alignment is likely to contain functional RNA structure.

Here, we identify all possible helices from the randomized alignments in the same manner as for the original alignment (i.e. using the same minimum stem and loop length criteria). These helices we assume to be non-functional, and have occurred by chance. We calculate the log-likelihood ratio for each of these helices, and use this pool of scores as a sample of the null-distribution of non-functional helix log-likelihood scores. We then assign each helix found in the original alignment a p-value which reflects the probability that one would find a helix with a higher log-likelihood score by chance.

Often, secondary structure is explicitly taken into account in the construction of an alignment, either manually (as is the case for the Rfam seed alignments — see section 3.1.3) or as part of the structure prediction process [26, 28]. For this reason, the alignment quality may vary across the alignment, and certain columns may be optimized for pairing. Such columns may bias the null distribution, since shuffling does not alter the composition of columns, only their order. To ensure that the alignment used for randomization is based entirely on primary sequence characteristics, we realign the sequences using T-coffee (version 7.97) [86]. We

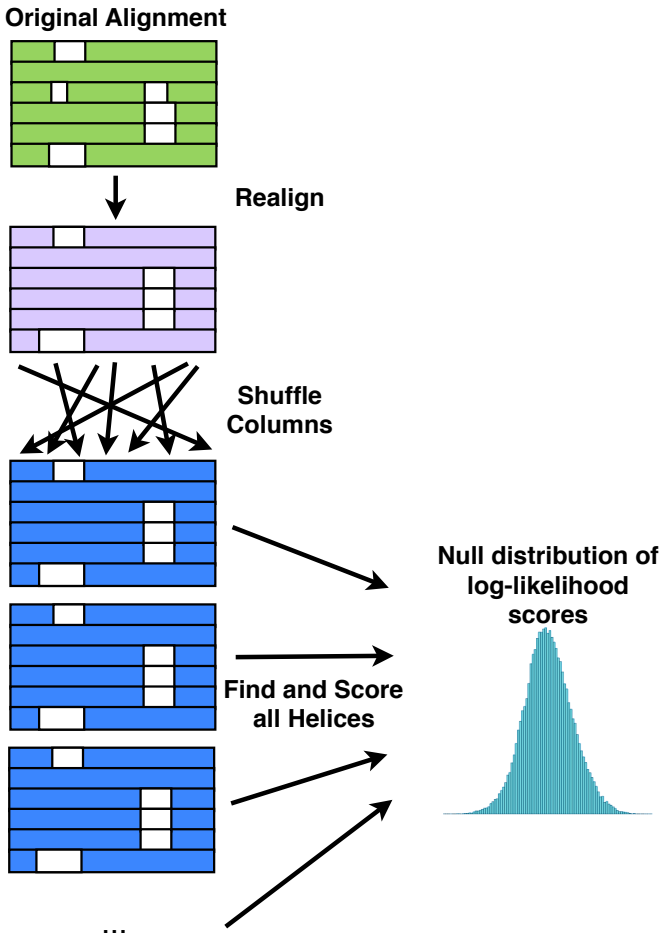


Figure 2.6: To estimate the probability of observing a helix of a given log-likelihood score by chance, we generate a set of shuffled alignments (the alignments coloured dark blue). These alignments should have no evolutionarily conserved helices. We use the scores of helices found in these alignments as our null distribution. A given input alignment (green alignment) is first realigned using T-COFFEE (light blue alignment) to ensure that the alignment is determined by primary sequence conservation only. Its columns are then shuffled (dark blue alignments), and all helices from these shuffled alignments are scored. Then the p-value of each helix in the original alignment is calculated as the fraction of helices from the shuffled alignments with log-likelihood scores above its log-likelihood score.

chose T-COFFEE for the realignment procedure because it performs well on a wide variety of sequences [87]. The randomization procedure is performed on the realigned sequences.

Randomization is done using the shuffling algorithm described in [85]. This algorithm preserves the original pattern of primary sequence conservation in the alignment by binning the alignment columns by mean pairwise identity (MPI), and shuffling only with these bins. Columns are swapped only with others that have a similar level of sequence conservation, so original pattern of sequence conservation is maintained. Maintaining the original pattern of sequence conservation is important because the evolutionary models we use are not neutral with respect to primary sequence conservation; in our models, paired columns mutate slightly slower than unpaired columns, reflecting the added evolutionary constraints imposed by the base-pairing requirement. This algorithm is implemented as a Perl script distributed with the RNAz package (version 1.0) [88]. We use this script, set to round column MPIs to one digit, to generate our shuffled alignments.

This randomization procedure does not, however, preserve the original alignment's dinucleotide frequency, i.e. the frequency of the various adjacent nucleotide pairs in the sequence. Dinucleotide frequency is relevant because the strength of the stacking interactions between adjacent base-pairs depends on the identities of the participating nucleotides. For instance, a C-G base-pair stacked on an A-U base-pair is more stable than a G-C base-pair stacked on (i.e. adjacent to) an A-U base-pair. Altering the dinucleotide frequency of a sequence during shuffling might therefore change the strength of the sequence's potential stacking interactions. For studies comparing the minimum free energies of original and randomized sequences, it is considered important for the randomization procedure to preserve dinucleotide frequency [89, 90]. There does exist an algorithm for shuffling alignments that preserves approximate dinucleotide frequency [91], but we chose to use the simpler shuffling approach that ignores dinucleotide frequency on the grounds that our method for calculating the log-likelihood ratio ignores stacking interactions.

For every randomized version of a given input alignment, we apply the same helix finding approach as to the original alignment in order to identify the helices in the alignment and to calculate their log-likelihood scores. Rather than storing the entire set of helices from all randomized alignments, we iteratively update the p-value of each helix from the original alignment with the contribution of the helices from each randomized alignment and then discard that alignment before generating a new one. We weigh randomized alignment equally. The formula for calculating the p-value of helix $h$ with a log-likelihood ratio of $\Lambda_h$ is

$$P(\Lambda_X > \Lambda_h) = \left( \sum_{A \in \mathcal{R}} \frac{|\forall\, x \in H(A) \text{ where } \Lambda_x > \Lambda_h|}{|H(A)|} \right) \Bigg/ |\mathcal{R}| \tag{2.17}$$

where $\mathcal{R}$ is the set of randomized alignments and $H(A)$ is the set of helices in randomized alignment $A$.

To calculate the contribution of each randomized alignment $A$ to the p-value, we first sort the helices of the randomized alignment by log-likelihood score in ascending order. We then find where each helix from the original alignment would be inserted in this sorted array. An alignment of length $m$ may have $O(m^2)$ helices, so the sorting step takes $O(m^2 \log(m))$ time. The fraction of helices which are before this position is fraction of helices in randomized alignment, and we update the running total of these fractions with this value. Currently, we find positions in the sorted array using a binary search, so this step also takes $O(m^2 \log(m))$. However, by sorting the set of original helices first, we could reduce this time to $O(m^2)$. Finally, once we have generated the desired number of randomized alignments, we divide these totals by the number of randomized alignments to get the p-value for each original helix. For all our experiments, we generated 500 randomized alignments for each input alignment, from which we derive the p-values.

Figure 2.7 shows an example null distribution of log-likelihood ratio scores for one particular alignment. In this example, log-likelihood ratio scores seem normally distributed. This suggests that one could model the null distribution as a normal distribution, and use the randomized alignments only to estimate the mean and standard deviation. Estimating those quantities would require fewer randomized alignments, and eliminate the need to sort log-likelihood ratios, which might substantially reduce TRANSAT's running time. However, this feature remains to be implemented, and we have yet to check whether log-likelihood ratios are similarly distributed for a larger set of alignments.

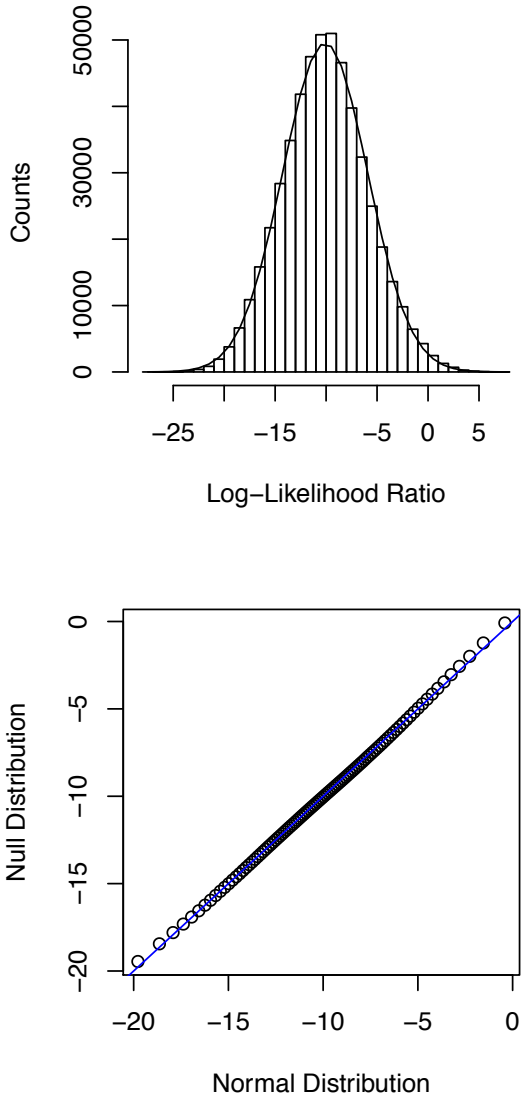


Figure 2.7: An example null distribution of helix log-likelihood ratios, generated from 500 randomization of the *hok* alignment. The upper figure shows the histogram of log-likelihood ratios. The black line corresponds to a normal distribution with the same mean and standard deviation. The lower figure is a quantile-quantile plot of the null distribution and a normal distribution. The line $y = x$ is shown in blue. The two distributions match closely.

# Chapter 3

# Performance evaluation

## 3.1 Datasets

New methods of RNA structure prediction are generally evaluated by predicting structures for a set of sequences whose structure is already known (the test set), and comparing the predictions to the known structure. Ideally, the test set should be large and contain a wide variety of RNA structures. If the prediction method has free parameters which were learned from known structures of a training set, there should be no overlap between the test and training sets.

Unfortunately, there are relatively few alignments available with confirmed functional alternative helices. We have found two such alignments, the *hok* alignment and the *trp*-attenuator alignment. To supplement these alignments in our test set, we use a set of alignments from the Rfam database [92–94]. These alignments are annotated with a conserved secondary structure (which may comprise a pseudoknot), but lack any information on alternative helices.

The mutation rates of our evolutionary models were estimated from a set of aligned tRNAs and large subunit ribosomal RNAs [24]. As far as we know, none of those sequences are part of our datasets.

### 3.1.1 *hok* alignment

In the R1 plasmid from *E. coli*, the *hok/sok* system is responsible for maintaining the plasmid's presence through successive generations (i.e. plasmid stabilization) [95]. It is composed of three genes: *hok* ('host-killing') encodes a protein toxin, *mok* ('modulation of killing') is required for *hok* translation, and *sok* ('suppression of killing') blocks the translation of *mok*, thereby repressing *hok* [81]. These genes are transcribed on two constitutively expressed transcripts; the *hok* and *mok* reading frames overlap on a single transcript (referred to as the *hok* transcript), while *sok* is an RNA gene, part of a larger class of RNA antitoxins [96]. The system stabilizes the plasmid by killing daughter cells that lack the plasmid after fission from the plasmid-containing parent, termed post-segregational killing. This works because the constitutively-expressed *hok* transcript is relatively stable, and so will be present in the daughter cell. *sok* is also constitutively expressed, but degrades quickly, so the daughter cell will soon run out of *sok* RNAs to suppress *hok* translation. The structure of the *hok*

transcript is key to this mechanism.

As *hok* is transcribed, the transcript forms a metastable structure, which blocks the ribosomal binding sites for *hok* and *mok*, preventing premature ribosome loading. Once the whole transcript has been produced, the transcript adopts a stable inactive structure (fig. 3.1, hairpins 1-3). However, 3'-end processing of the transcript allows the transcript to rearrange into the active structure (fig. 3.1, hairpins 4-5), which is translationally active unless *sok* is bound to it. The metastable structure likely also helps to guide folding into the stable inactive structure (which has a 'long-distance' helix that pairs a region at the 5' end of the transcript to a region near the 3' end), preventing premature formation of the active structure. A review of the *hok* mechanism can be found in [81].

Several related toxin/antitoxin systems have been identified, and the alignment of their transcripts reveals covariation patterns consistent with each of these structures [97]. For our dataset, we analyze the alignment from [81], which contains several more members of the *hok* family (9 sequences total). As this alignment provides only an outline of the helices, we filled in the exact consensus structure manually. The alignment does not cover the entire length of the *hok* transcript, but does cover the regions with the most structural rearrangement.

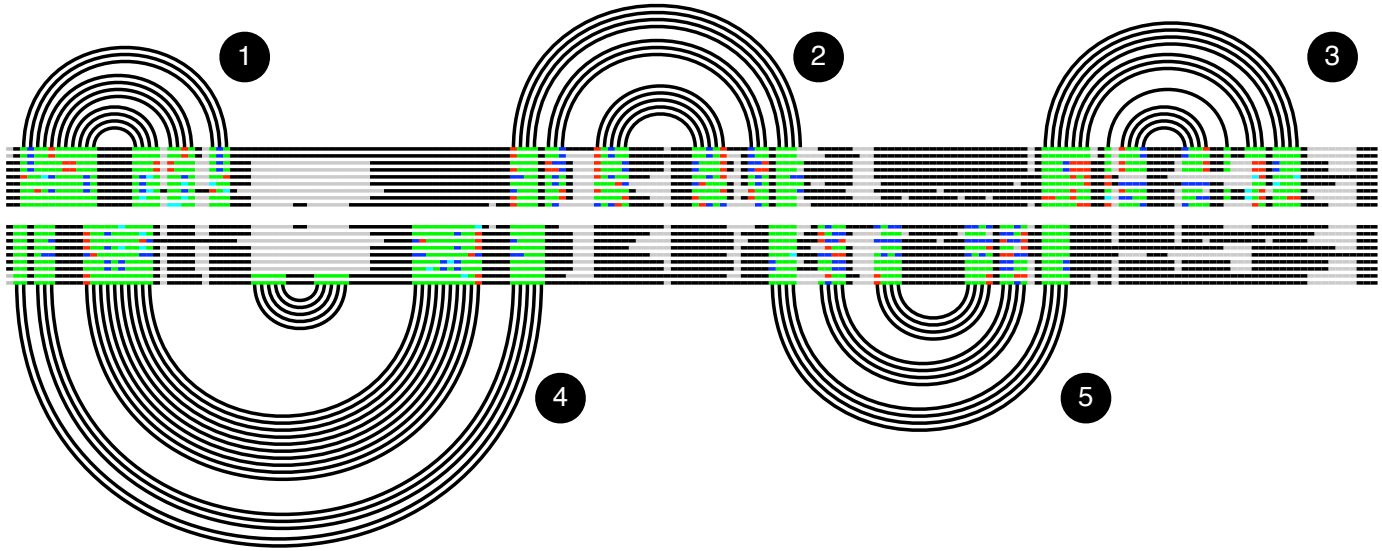


Figure 3.1: The *hok* alignment contains 5 hairpins, numbered 1 through 5 in this diagram. The arcs represent the base-pairs of the known structure for this alignment. The horizontal lines represent the sequences of the alignment, coloured to highlight covariation of paired positions in the known structure; the upper block shows the covariation pattern for hairpins 1–3, while the lower block shows the covariation pattern of hairpins 4 and 5. For each paired position, the most common canonical base-pair is coloured green. Single nucleotide substitutions that preserve pairing potential (e.g. A:U to G:U) are coloured cyan. Double nucleotide substitutions that preserve pairing potential (e.g. A:U to G:C) are coloured dark blue. Non-canonical base-pairs are coloured red.

### 3.1.2 *trp*-attenuator alignment

The *trp*-attenuator regulates transcription of the *trp*-operon in *E. coli* [98]. This switch is situated in the leader peptide region of the *trp*-operon transcript, and may form three possible helices (fig. 3.2). Two helices, involving the 1:2 and 3:4 stems (the numbers denote regions in the transcript that are paired to form the helix), are mutually compatible. The other helix, involving stem 2:3, overlaps both other helices. The 3:4 stem is a terminator stem, and ends transcription if it is allowed to form immediately. During transcription, the formation of helix 1:2 causes the RNA-polymerase to pause. If a ribosome then binds to the transcript and starts translation, it will disrupt helix 1:2 once it reaches that part of the transcript, freeing the RNA-polymerase. If tryptophan is limited, the ribosome will pause at the tryptophan codon in region 1, allowing helix 2:3 to form, disfavoring the formation of the 3:4 terminator stem. This will allow the trp-operon to be fully transcribed. If tryptophan is not limiting, then the ribosome will disrupt the 2:3 stem, allowing the 3:4 stem to form and end transcription. Several protein-mediated RNA switches which regulate trp-operon activity have been identified in *Bacillus subtilis* [99]. Comparative analysis of *trp*-operons from several species of Actinobacteria show similar features to the *E. coli trp*-attenuator [100].

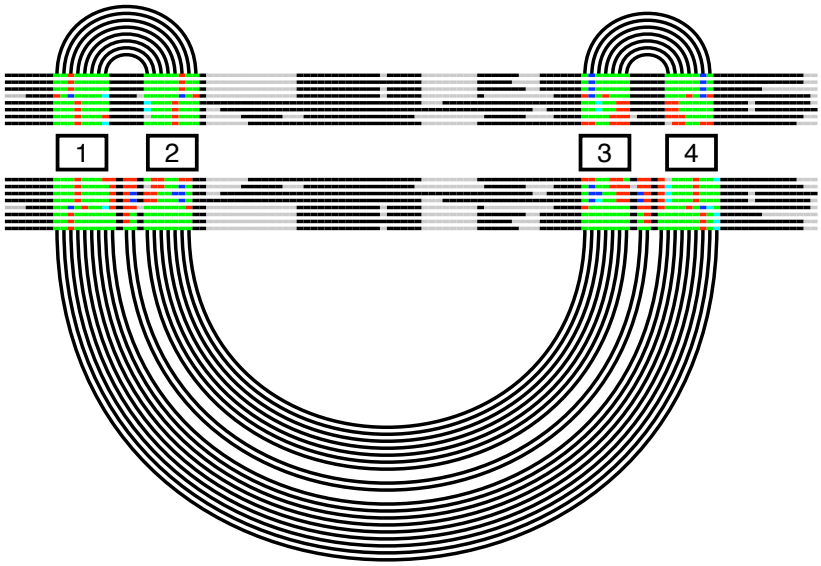


Figure 3.2: An arc diagram showing the known structure of the trp-attenuator alignment, with labeled regions 1–4. The arcs represent the known structures of the alignment, while the horizontal lines represent the sequences of the alignment, coloured to highlight covariation of paired positions in the known structure. For details on the colour scheme, see fig. 3.1.

The RNAlishapes program [53] successfully identified all three helices from an alignment of the Actinobacterial *trp*-attenuator candidates. Here, we use the alignment produced in that study (8 sequences, with a total length of 117), and compare our method to the predictions

by RNAlishapes.

### 3.1.3 Rfam dataset

The Rfam database [92–94] contains multiple sequence alignments for a wide variety of RNA gene families. For each family, Rfam stores a manually curated seed alignment and a conserved secondary structure for that alignment. Each seed file is used to generate a covariance model [101, 102] for the corresponding RNA family. Using this covariance model, Rfam compiles a full alignment consisting of the seed sequences plus nucleotide sequences from EMBL [103] that score above a certain threshold with respect to the covariance model.

| Type | Dataset | All of Rfam |
|---|---|---|
| gene | 72.39 | 84.62 |
| snRNA | 44.78 | 36.44 |
| snoRNA | 38.81 | 35.64 |
| Cis-reg | 26.87 | 15.23 |
| C/D-box | 25.37 | 24.78 |
| H/ACA-box | 13.43 | 9.69 |
| miRNA | 8.96 | 33.09 |
| riboswitch | 8.96 | 1.09 |
| splicing | 5.97 | 0.73 |
| ribozyme | 5.22 | 0.80 |
| sRNA | 5.22 | 5.39 |
| IRES | 2.99 | 1.97 |
| leader | 2.99 | 0.73 |
| rRNA | 2.24 | 0.29 |
| intron | 0.75 | 0.15 |
| thermoregulator | 0.75 | 0.22 |

Table 3.1: Table showing the percentage of each type of RNA for our Rfam dataset and the entire set of seed alignments in the Rfam database. The type terms used are the ones used by Rfam to categorize families. A family may have several types (e.g. Rfam labels the family RF00002 with the terms 'gene' and 'rRNA'), so the percentages add up to more than 100%.

We have chosen to analyze a subset of seed alignments from the Rfam database v. 9.1 [94] which meet a several criteria. Alignments must score well on two quality control statistics used by Rfam, mean fraction of canonical base pairs ($> 0.8$) and covariation ($> 0.2$) [104]. The mean fraction of canonical base-pairs (FC) measures the proportion of nucleotides pairs which are canonical base pairs (AU, GC, and GU) in paired columns of the consensus structure. The covariation is a measure of how consistently mutations in paired columns maintain pairing potential (see appendix A for precise definitions of these two measures). The seed alignments in the Rfam database have FC values ranging from near 0 (almost no base-pairs in any sequence are canonical base-pairs) to 1 (all base-pairs from every sequence are canonical

base-pairs). Covariation values for the RFAM database seed alignments range from -2 to 2. Alignments with FC and covariation values of greater than 0.4 and 0, respectively, are considered 'good' [104]; here, we use slightly more stringent criteria. Additionally, we require alignments to contain at least 5 sequences, because with fewer sequences, there is less evolutionary information to work with. We also require alignments in our dataset to have a length greater than 100, since very short sequences are less likely to show complex folding behavior. Table 3.1 shows the types of RNA sequences represented in our RFAM dataset.

### 3.1.4 Phylogenetic trees

In addition to an alignment, TRANSAT requires as input a phylogenetic tree representing the evolutionary history of sequences of alignment. To estimate these phylogenetic trees, we used the maximum likelihood method phylogenetic tree included in the PFOLD package[24, 25]. This approach is a simplified version of the commonly used maximum likelihood (ML) approach to phylogenetic tree inference [105, 106]. This method uses the neighbor-joining algorithm to infer the topology of the tree, and then optimizes the tree's branch lengths to maximize the likelihood of of the tree, i.e. the probability that the alignment would be produced with the tree, given a model of evolution [25]. Unlike other methods of ML phylogenetic tree prediction, this method neither optimizes the tree topology nor the evolutionary model.

## 3.2 Performance measures

TRANSAT produces a list of helices, together with their p-values which reflect our confidence that the measured amount of evolutionary conservation is not due to chance. We therefore treat the problem as one of binary classification: helices are classified as either conserved or not based on their p-value. We can then evaluate the TRANSAT's performance by comparing its predictions to the helices of the known structure(s).

To perform this classification, we choose a p-value threshold; TRANSAT classifies helices with a p-value below a given threshold as conserved, and all other helices as unconserved. It may also be useful to know and quantify the strength of conservation, since certain helices or base-pairs may be more important to the function of the RNA molecule, and therefore might be more strongly conserved. Analogous information on the relative strength of primary sequence conservation in alignments of protein sequences is useful for identifying regions which are particularly important for to a protein's function [107]. Information of this type may be obtained by looking at the log-likelihood ratio. For now, however, we ignore this aspect to simplify the problem to one of binary classification.

Classification at the helix level involves an additional complication in that certain helices may match a helix of the known structure only partially (i.e. if it contains some base-pairs that are found in the known structure, and others which are not). Because the log-likelihood

score is essentially an average of the log-likelihood ratios of all base-pairs of the helix, a helix with mostly conserved (high-scoring) base-pairs but a few unconserved (low-scoring) base-pairs is likely to be indistinguishable from helices composed entirely of conserved helices. We therefore define a true helix as one in which at least 70% of the helix's base-pairs are found in the known structure.

| | | actual | |
|---|---|---|---|
| | | helix has $> 70\%$ known base-pairs | helix has $\leq 70\%$ known base-pairs |
| predicted | p-value $< c$ | True positive (TP) | False positive (FP) |
| | p-value $\geq c$ | False negative (FN) | True negative (TN) |

Table 3.2: Definitions for evaluating the helix-level performance of TRANSAT at a p-value threshold of $c$.

The performance evaluation can also be performed at the level of individual base-pairs by assigning the base-pair the minimum p-value of the helices that contain that base-pair (or 1, if no helices contain that base-pair). This formulation sidesteps the problem of evaluating helices that match the helices of the known structure only partially.

| | | actual | |
|---|---|---|---|
| | | base-pair in known structure | base-pair not in known structure |
| predicted | min p-value $< c$ | TP | FP |
| | min p-value $\geq c$ | FN | TN |

Table 3.3: Definitions for evaluating the base-pair-level performance of TRANSAT at a p-value threshold of $c$. The p-value for a base-pair is the minimum p-value of all helices with that base-pair.

From the above definitions in tables 3.2 and 3.3, we can define several useful measures of performance:

$$\text{Sensitivity} := \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.1}$$

$$\text{Specificity} := \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.2}$$

$$\text{False Positive Rate (FPR)} := \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{3.3}$$

$$\text{Positive Predictive Value (PPV)} := \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.4}$$

Sensitivity may be thought of as how much of the known structure we recover, while specificity measures how well we exclude helices which are not part of the known structure. Positive predictive value (PPV) measures how 'useful' a positive prediction is (i.e. how likely it is that a predicted helix/base-pair is part of the known structure).

Additionally, we use the F-measure, the harmonic mean of sensitivity and PPV as a useful summary measure of performance.

$$\text{F-measure} := \frac{2 \cdot \text{Sens} \cdot \text{PPV}}{\text{Sens} + \text{PPV}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}} \tag{3.5}$$

## 3.3 Performance on alignments with known alternative structures

For the *hok* and *trp*-attenuator alignments, whose structures have been studied in detail, we can be fairly confident that our known structures capture nearly all functional conserved helices of the alignment. As their structures both contain overlapping helices, testing on these alignments is our primary means of assessing TRANSAT's most interesting feature, the ability to identify alternative helices.

### 3.3.1 *hok* alignment

Figure 3.3 shows the performance of TRANSAT for the *hok* alignment. Overall, TRANSAT's performance on this dataset is encouraging. The ROC curve rises sharply, indicating that helices of the known structure can distinguished from spurious ones fairly well on the basis of the p-value produced by TRANSAT. However, because the classes are so highly imbalanced (for *hok*, TRANSAT identifies 1040 possible helices with varying p-values, of which only 40 correspond to known helices), it is more difficult to achieve a high PPV, since even a low false positive rate produces a large number of false positive helices relative to the number of true positive helices. Even so, TRANSAT achieves a reasonably high PPV without losing much sensitivity, indicating just how well TRANSAT's p-values separate true helices from false ones for this alignment. The p-value threshold with the highest F-measure (0.78) is $5.5 \cdot 10^{-3}$.

To take a closer look TRANSAT's predictions, we developed a method of visualizing the predicted helices using arc diagrams, two examples of which are shown in figure 3.4. In these diagrams, the horizontal line represents the entire alignment, with the 5' end on the left and the 3' end on the right. The arcs above the line represent the base-pairs of the known structure. Base-pairs from the known structure that are present in TRANSAT's set of predicted helices are coloured (i.e. non-black), while unpredicted base-pairs of the known structure are left black. The arcs below the line are base-pairs present in at least one of TRANSAT's predicted helices, but not part of the known structure. The arc colouring gives an indication of the p-value of the helix in which the base-pair occurs (if a base-pair occurs in more than one helix, we assign it the lowest p-value of those helices). For these and subsequent arc diagrams, the colour scheme is as follows: helices with p-values $< 10^{-5}$ are green, $< 10^{-4}$ blue, $< 10^{-3}$ orange, and $<$ (p-value threshold) red.

Closer inspection of the helices predicted by TRANSAT for the *hok* alignment shows that

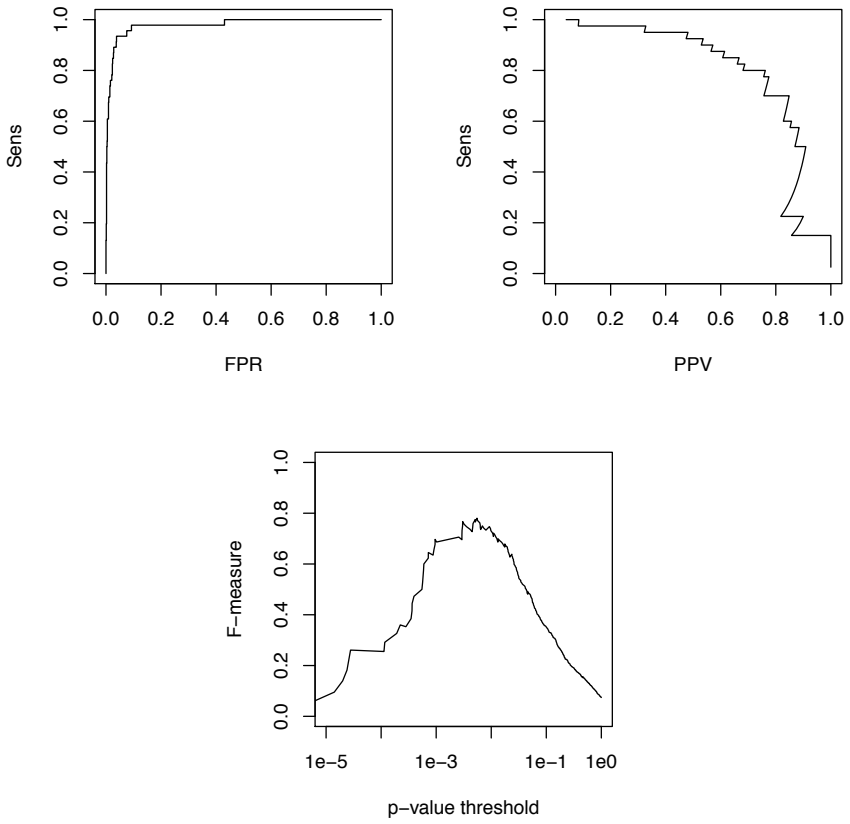


Figure 3.3: Helix-level performance of TRANSAT for identifying known functional helices in the *hok* alignment. The figure on the upper-left shows the tradeoff between sensitivity (Sens) and false-positive rate (FPR) as we raise the p-value threshold (ROC curve). The upper-right figure shows the tradeoff between sensitivity and positive predictive value (PPV). The bottom figure shows F-measure as a function the p-value threshold.

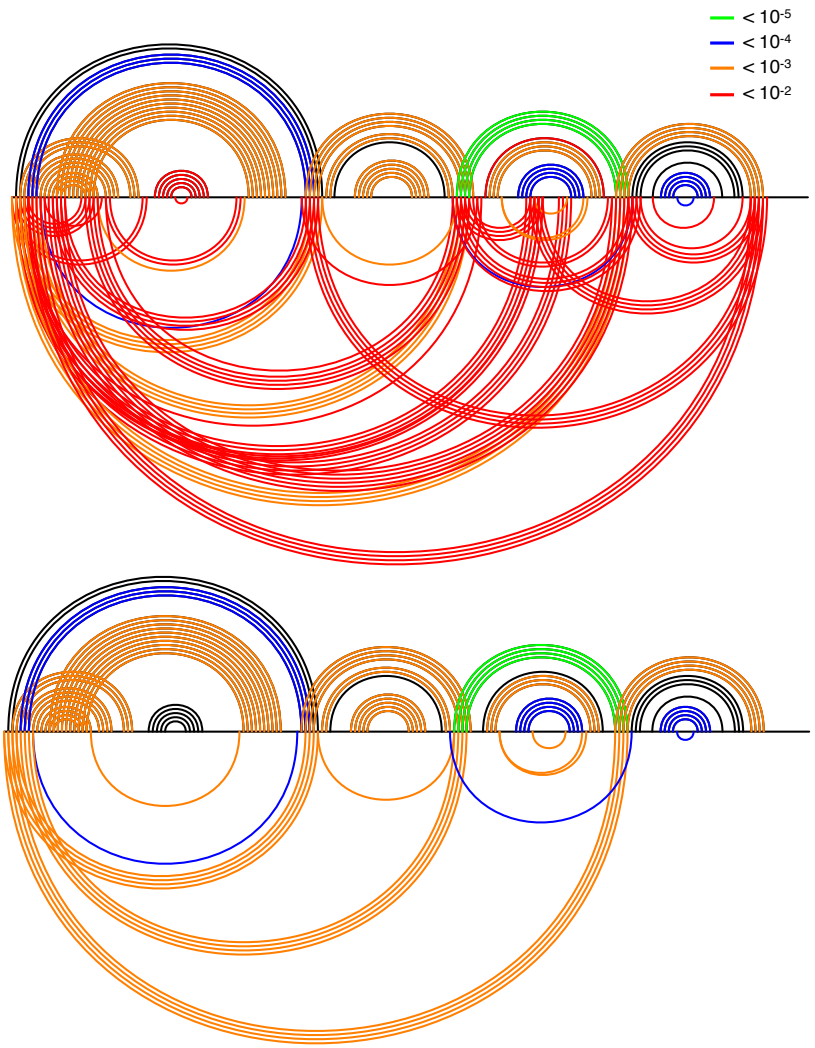


Figure 3.4: Arc diagrams of TRANSAT predicted helices for the *hok* alignment. In these two diagrams, the base-pairs of the known structure is plotted above the horizontal line. The coloured arcs above the line are base-pairs in the known structure that were also predicted by TRANSAT. The arcs below the line represent base-pairs predicted by TRANSAT but not part of the known structure. In the upper diagram, we plot the base-pairs from all helices with p-values below $10^{-2}$. In the lower diagram, we use a p-value threshold of $10^{-3}$. Arcs are coloured according to minimum p-value of the helices in which they occur: $< 10^{-5}$ green, $< 10^{-4}$ blue, $< 10^{-3}$ orange, and $< 10^{-2}$ red.

TRANSAT is largely successful at identifying the alignment's known helices, including the overlapping helices. With the p-value threshold set to $10^{-2}$, we capture almost every base-pair of the known structure, but also pick up a large number of other helices (fig. 3.4, upper diagram). At the (default) p-value threshold of $10^{-3}$, the prediction loses a few base-pairs of the known structure, but still manages to capture most of them (fig. 3.4, upper diagram). A few helices which are not in the known structure remain in the predicted structure, but far fewer than are found using the less stringent p-value threshold of $10^{-2}$.

The base-pairs predicted by TRANSAT but not present in the known structure of the *hok* alignment fall into two categories:

1. Several base-pairs immediately adjacent to and compatible with known helices are included in the TRANSAT predictions as part of helices composed of mainly base-pairs from the known structure, which we call *extending base-pairs*. Helix D of Figure 3.5 shows one such example. The inner base-pair of that helix is not part of the known structure, and is largely unsupported by sequence covariation. However, because this position can form a valid base-pair in a few of the sequences, it is included in a helix predicted for that location in the alignment. The log-likelihood ratio assigned to this helix is the average log-likelihood of all its base-pairs, so while the log-likelihood ratio of that particular base-pair is low, the helix is 'rescued' by the high ratios of the other base-pairs. We observe extending base-pairs in other alignments as well (e.g. fig. 3.11, lower arc diagram).

   With respect to the problem posed by extending base-pairs, it may be worthwhile to incorporate some method of identifying and excising them, preferably before the p-value calculation step. This filter might work by checking the edges of a helix for low-likelihood base-pairs, or by looking for helices which differ by only one base-pair and eliminating the lower scoring one. However, we have not yet implemented such a filter, nor have we fully fleshed out a method for doing so.

   Conceivably, such base-pairs may be indicative of structural evolution; the base-pair might have been acquired after the divergence of the sequences in the alignment, in which case it would only appear in some of the sequences. In this case, they may be of interest.

2. Three helices in the predicted set contain no base-pairs of the known structure (fig. 3.5, helices A, B, and C). Looking at the pattern of sequence conservation for these novel helices, we see several instances of possible covariation. Covariation events seem slightly less frequent in these helices than the helices of the known structure (compare with fig. 3.1), but only marginally so. Helix A and the outer helix from stem 4 are very closely positioned, suggesting that perhaps the structure is annotated incorrectly and helix A is the correct one. Helices B and C may be spurious, though it is interesting

to note that the formation of either of those helices would at least partially block the formation of hairpins 4 and 5, the translationally active structure in the absence of *sok* (see section 3.1.1). In general, although TRANSAT provides evidence that a novel helix is functional, it is difficult to infer the specific function of that helix.

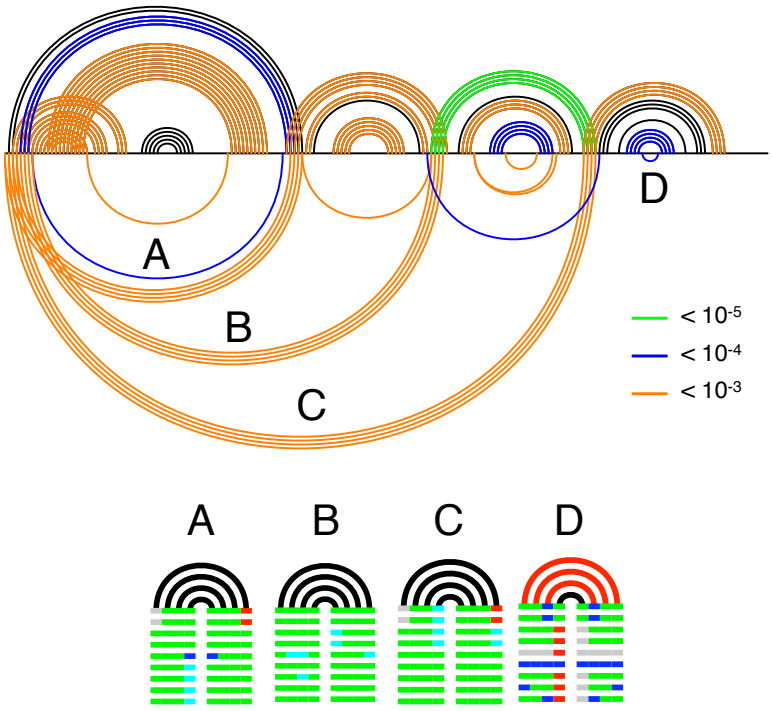


Figure 3.5: Arc diagram of the *hok* alignment showing TRANSAT predictions with a p-value threshold of $10^{-3}$ (as in fig. 3.4, lower image), together with diagrams showing the pattern of covariation in four selected helices. In the covariation diagrams, the coloured lines represent the sequences of alignment positions covered by that helix. Arcs colored in red are base-pairs found in the known structure, while arcs colored in black are novel base-pairs. Positions are coloured to highlight covariation of paired positions in the known structure. For each paired position, the most common canonical base-pair is coloured green. Single nucleotide substitutions that preserve pairing potential (e.g. A:U to G:U) are coloured cyan. Double nucleotide substitutions that preserve pairing potential (e.g A:U to G:C) are coloured dark blue.

### 3.3.2 *trp*-attenuator alignment

TRANSAT's performance for the *trp*-attenuator alignment is less convincing than for the *hok* alignment. As with the *hok* alignment, the ROC curve is fairly sharp (fig. 3.6). A high PPV is hard to achieve, however, and requires sacrificing most of the sensitivity. The F-measure

plot peaks at a relatively high p-value threshold (0.045), and declines sharply after that, suggesting that the helices of the known structure are assigned relatively high p-values.

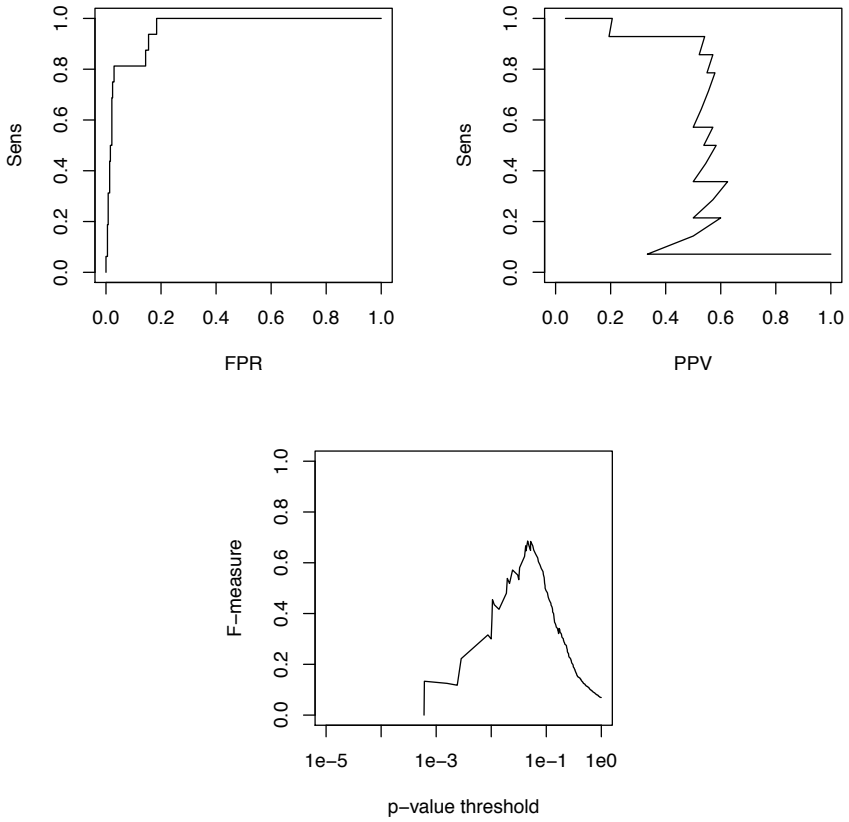


Figure 3.6: Helix-level performance of TRANSAT for the *trp*-attenuator alignment. The figure on the upper-left shows the tradeoff between sensitivity (Sens) and false-positive rate (FPR) as we vary the p-value threshold (ROC curve). The upper-right figure shows the tradeoff between sensitivity and positive predictive value (PPV). The bottom figure shows the F-measure as a function the p-value threshold.

The arc diagram (fig. 3.7) of the TRANSAT predictions at a p-value threshold of 0.05 (a higher threshold than the default one of $10^{-3}$) shows that while TRANSAT captures most of the base-pairs of the known structure, it also finds a large number of novel helices. Many of these helices are positioned similarly to helices of the known structure, and may be artifacts of errors in the alignment. At least two helices with p-values below the threshold are not all similar to helices in the known structure. These helices, both pairings of the 5' end with the middle of the alignment, are probably false positives. Looking at the covariation pattern of the known helices of this structure (fig. 3.2), we see that the known structure contains many non-canonical base-pairs and relatively little covariation, indicating that the alignment may

contain errors, so it is perhaps not surprising that the performance of TRANSAT is worse for this alignment.

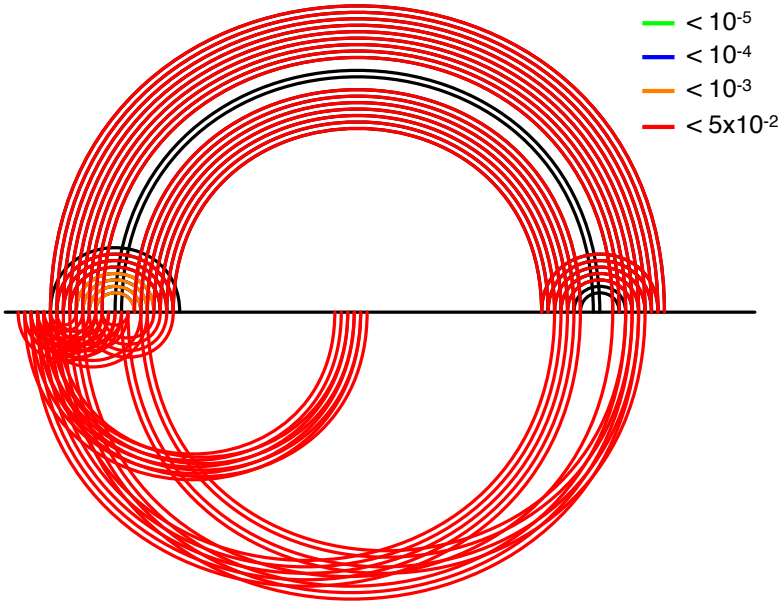


Figure 3.7: Arc diagrams of TRANSAT predicted helices for the *trp*-attenuator alignment at a p-value threshold of 0.05. See figure 3.4 for more information on arc diagrams.

## 3.4 Performance on RFAM dataset

For our RFAM dataset, we are less confident that *all* functional structures are included in the known structure provided by RFAM. In particular, there may be certain conserved transient structures which are not annotated, since RFAM only aims to annotate a single functional structure for each family; functional helices which conflict with that structure are left out. In the context of performance evaluation, this amounts to an underestimation of the number of true helices/base-pairs.

Although a small (but growing) number of RNA sequences have been investigated in detail with regard to their folding pathway, there are only a few RNA sequence *alignments* available which have been annotated with folding pathway information. This obviously poses a significant problem for evaluating the performance of TRANSAT on the RFAM dataset, as we cannot be confident that the structural annotations are complete. We assume, however, that the structure annotation provided by RFAM is correct.

However, because TRANSAT predicts all functional helices, including transient ones, we can evaluate how well TRANSAT predicts the helices of the known structure provided by RFAM. The purpose of this analysis is to test whether or not TRANSAT can reliably recover

many of these known structures. Competing helices identified by TRANSAT with comparable p-values to the helices of the known structure are also of interest, and may represent functional transient helices left out of the RFAM annotation.

We find that TRANSAT does indeed assign low p-values to helices with base-pairs from the RFAM known structure. The ROC curve of TRANSAT's performance on the RFAM dataset shows that TRANSAT's p-values separate base-pairs and helices of the known structure from the vast majority of other base-pairs and helices quite successfully (fig. 3.8, top left). Note that although we hope that some of the helices which are not found in the known structure are functional and evolutionarily conserved, any novel helices predicted by TRANSAT will, for the purposes of performance evaluation, be considered to be false positives. We assume that these helices will not greatly affect the validity of our performance measures, however, since the vast majority of the helices identified are expected to be spurious.

### 3.4.1 P-value threshold selection

Because the base-pairs of the known structure constitute only a very small fraction of the possible base-pairs for an alignment, the base-pair classes (*True* vs. *False* base-pairs) are highly imbalanced. The sharp ROC curve therefore does not translate into high PPV values, since even at low false positive rates, the small fraction of the false base-pair class may contain many more base-pairs than the entire true base-pair class. This is not necessarily cause for concern; we expect the RFAM known structure to not contain all the conserved functional base-pairs. For the purposes of selecting a default p-value threshold, however, we choose a value that maximizes the F-measure. For the RFAM dataset, the F-measure maximum occurs at $0.47 \cdot 10^{-3}$ for the base-pair-level performance and at $1.9 \cdot 10^{-3}$ for the helix-level performance (fig. 3.8).

For users of TRANSAT, we suggest they use the threshold $10^{-3}$ for analyzing alignments whose structure is not known. If a functional structure is known, and one is looking for competing helices, we suggest choosing a p-value threshold which maximizes the F-measure for that alignment.

### 3.4.2 Variability of performance

The previous section considered the performance of TRANSAT averaged over our entire RFAM dataset. TRANSAT's performance, however, varies widely from family to family. At the default p-value threshold of $10^{-3}$, F-measures for individual RFAM families range from 0 to 1. This performance appears to be uncorrelated with the quality of the alignment as measured by the RFAM group (fig. 3.9).

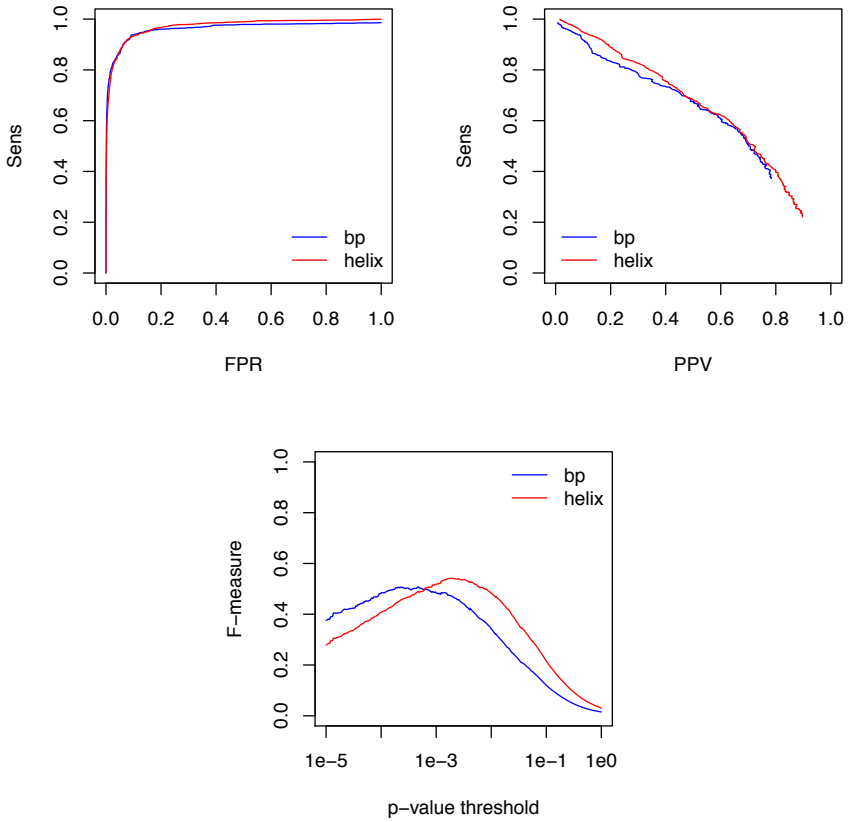


Figure 3.8: Performance of TRANSAT for identifying known functional helices in the RFAM dataset for individual base-pairs (blue line) and entire helices (red line). The top-left figure shows the ROC curve, showing the tradeoff between sensitivity and false positive rate (FPR). The top-right figure shows a plot of sensitivity versus positive predictive value (PPV). The bottom figure shows a plot of the F-measure as a function of the p-value threshold. Sensitivity, FPR, and F-measure are averaged over all alignments in the RFAM dataset.

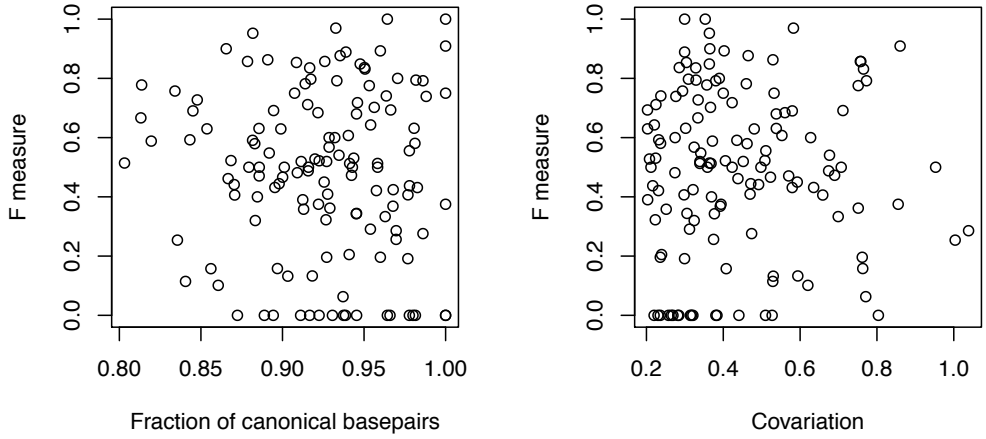


Figure 3.9: Performance of TRANSAT for individual RFAM families plotted against two measures of alignment-quality, mean fraction of canonical base-pairs (left) and covariation (right) (see appendix A for details on these measures). Base-pair level performance for each family is summarized as F-measure at the default p-value threshold of $10^{-3}$. Neither measure of alignment quality appears to be correlated with performance.

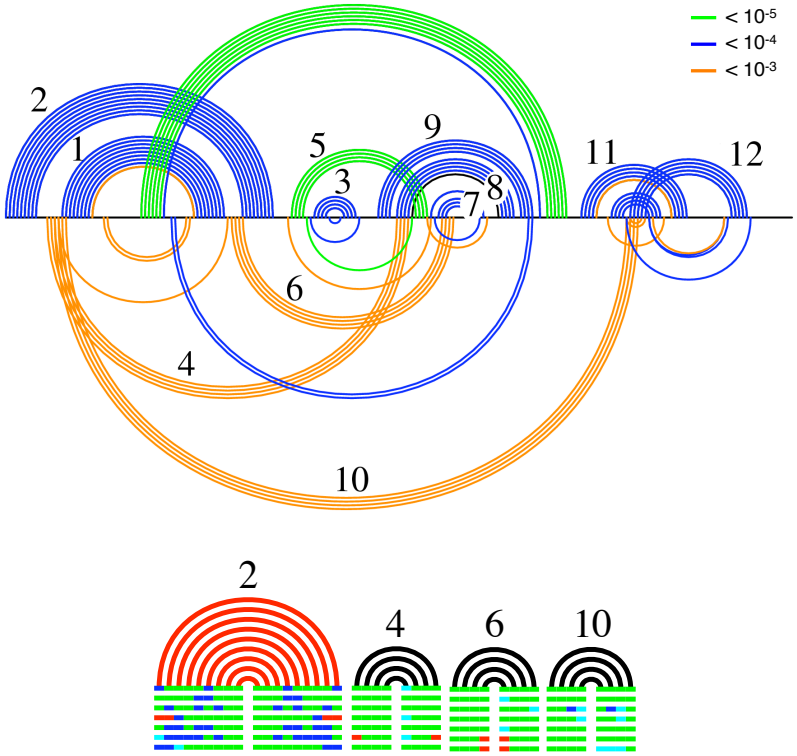


Figure 3.10: Arc diagram for Rfam family RF00485, the Cripavirus internal ribosome entry side (IRES), showing Transat predictions using the default p-value threshold of $10^{-3}$. The covariation patterns of helix 2 from the known structure and novel helices 4, 6, and 10 are shown as covariance diagrams (see fig. 3.5 for details) below the arc diagram. Helices 4, 6, and especially 10 show some evidence of covariation, but not to the same extent as helix 2 or others of the known helices (data not shown). See figure 3.4 for more information on arc diagrams.

### 3.4.3 Examples

In this section, we discuss a few RFAM families and their TRANSAT predictions in detail. Figure 3.10 shows TRANSAT's predictions for the alignment of Cripavirus internal ribosome entry sites (IRES), RFAM family RF00485. TRANSAT correctly identifies all the base-pairs of the known structure, as well as three helices which are not present in that structure. Each of these novel helices conflicts with at least one helix of the known structure. The novel helices display a certain amount of covariation (though the pattern is less pronounced than for many of the helices in the known structure). One can imagine that these novel helices could correspond to transient helices, i.e. that they could form and then be displaced as the RNA is transcribed, since they occur (mostly) 5' of the helices they compete with. Such 'just-so' stories are of value only as hypotheses, however; to confirm these hypotheses, dedicated experiments would be required.

Figure 3.11 shows TRANSAT's predictions for two alignments of telomerase RNA sequences, one for vertebrates (RF00024) and one for ciliates (RF00025). In the alignment of vertebrate sequences, TRANSAT finds the known structure fairly well, but also predicts a large number of additional helices. In particular, it predicts a large set of helices which link the two hairpin-like structures, some of which are assigned quite low p-values (see the novel helices coloured blue in fig. 3.11, top). We do not expect all new helices to be functional, but a few of them seem reasonably well supported by the alignment.

In the alignment of ciliate sequences, all the base-pairs of the known structure are assigned low ($< 10^{-5}$) p-values. All predicted helices with a p-value below $10^{-3}$ contain at least one base-pair of the known structure, although many contained additional base-pairs. These are likely to be extending base-pairs, described in section 3.3.1. As opposed to the alignment of vertebrate sequences, there are no 'linking' helices predicted for the alignment of ciliate sequences.

In both vertebrate telomerase RNA alignment (fig. 3.11) and the IRES alignment (fig. 3.10), TRANSAT successfully identifies helices which render the known structure pseudoknotted. TRANSAT predicts individual helices rather than entire RNA structures and is thus no biased in favour or against pseudoknotted configurations of helices. Figure 3.12 shows two families for which TRANSAT predicts novel helices which would introduce pseudoknots into the known structures. In these cases, the predicted helices may have been overlooked in the annotation of the known structure, since many secondary structure prediction programs ignore pseudoknots, and the regions forming these helices are far apart along the alignment and thus more difficult to find by human annotation.

Interactions between an RNA and proteins or other molecules may prevent the formation of secondary structure in certain regions of the RNA. Because TRANSAT relies exclusively on the evolutionary signal of RNA structure, it is also capable of identifying regions which have no conserved structure without requiring detailed knowledge of the *in vivo* environment.

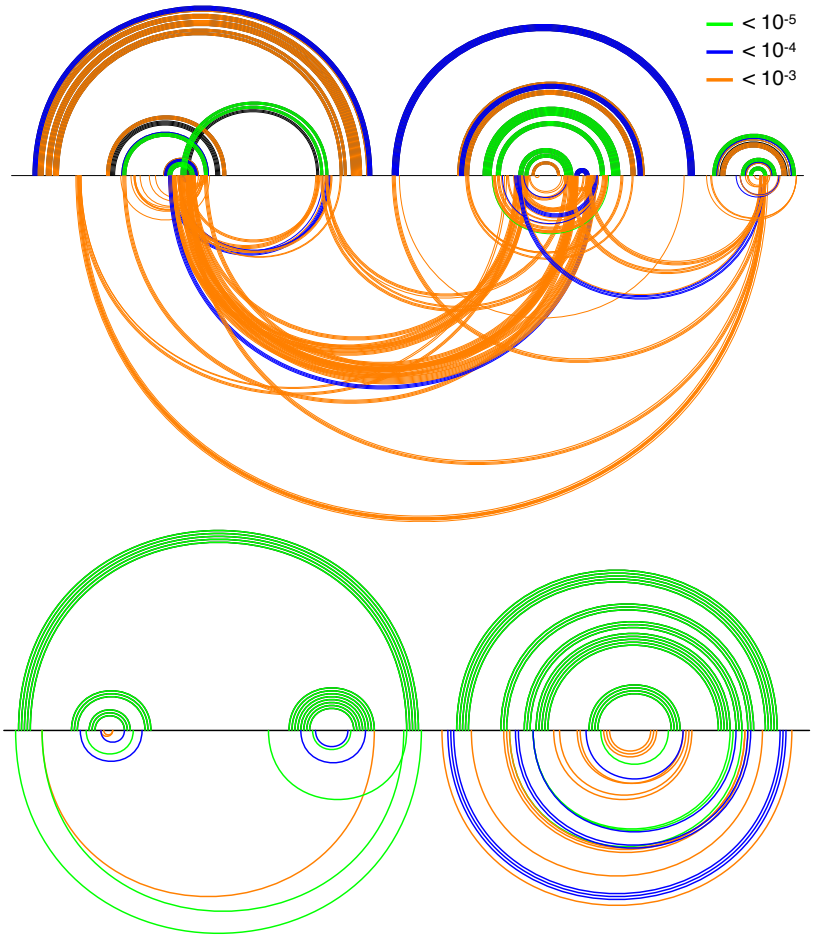


Figure 3.11: Arc diagrams of TRANSAT predictions (using the default p-value threshold of $10^{-3}$) for two alignments of telomerase RNA, vertebrates (top, RF00024) and ciliates (bottom, RF00025). See figure 3.4 for more information on arc diagrams.

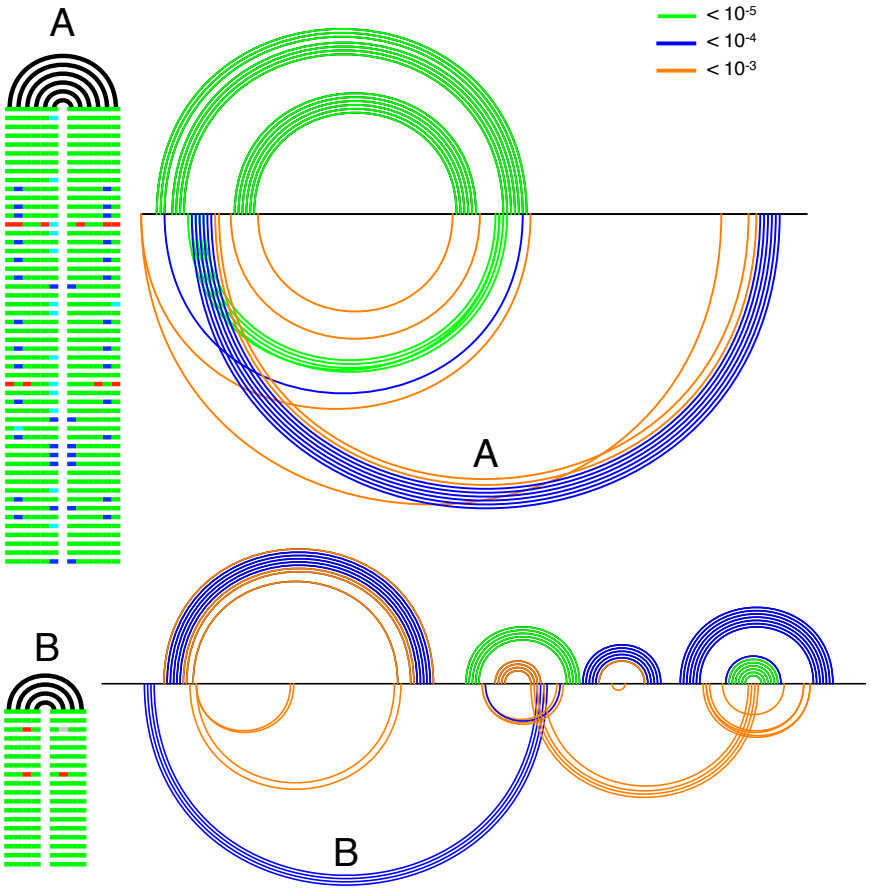


Figure 3.12: Arc diagrams of TRANSAT predictions (using the default p-value threshold of $10^{-3}$) for the S-adenosyl-L-homocysteine riboswitch family (top, RF01057), a riboswitch found on certain bacterial mRNAs, and the glmS glucosamine-6-phosphate activated ribozyme (bottom, RF00234), a bacterial ribozyme. See figure 3.4 for information on arc diagrams. The helices labelled A and B are novel helices which would introduce a pseudoknot into the known structures of their respective alignments. Diagrams of the covariation pattern of each of these helices is included on the left. There is little variation at the positions of helix B, but helix A is strongly supported by the pattern of covariation. See fig. 3.5 for information on covariation diagrams.

Figure 3.13 shows arc diagrams of TRANSAT predictions for the bacterial tmRNA family (RF00023), which contains an open reading frame (ORF) which is must be unstructured to function. At the more stringent p-value threshold of $10^{-4}$ (fig. 3.13, bottom), no helices are predicted in the ORF region, but nearly all the helices of the known structure are captured. This p-value threshold is more appropriate for this alignment than the default p-value threshold of $10^{-3}$ (fig. 3.13, top).

Finally, figure 3.14 illustrates with greater precision the problem mentioned previously, wherein novel helices predicted by TRANSAT are more strongly conserved at the primary sequence level than the helices of the known structure (see figures 3.5 and 3.10). Because the alignment of flavin mononucleotide (FMN) riboswitch sequences (RFAM family RF00050) has a large number of sequences (147) and a high degree of sequence variation, this effect is particularly pronounced. In helix A, the primary sequence is not strongly conserved, but the base-pairs show a large amount of covariation. In contrast, very little covariation is observed in helix B, but the primary sequence is well conserved. What little variation there is in helix B is more often than not unsupportive of its predicted base-pairs. TRANSAT assigns these two helices similar p-values, however.

This effect likely caused by the slight preference of the paired model of evolution for sequence conservation relative to the unpaired model. This makes sense for the original purpose of the models; because of the imperative to conserve pairing potential, mutations will often only be observed in subsequent generations if both pairing partners are changed, and such events are rare. However, this becomes problematic in the case that two regions of strong primary sequence conservation happen to be compatible for base-pairing. On the basis of their strong primary sequence conservation, the 'helix' they form is inherently favored by the paired model.

One conceivable solution to this problem is to scale the input tree (i.e. multiply all the branch lengths in the tree by some overall scaling factor) differently for different helices or alignment columns. For columns where the primary sequence conservation is high, one would apply a scaling factor in the range $(0, 1)$ to the tree to favor primary sequence conservation. Conversely, if the primary sequence conservation is low, one would apply scaling factor that is $> 1$.

However, helices with high sequence conservation might be over-represented in the set of predicted novel helices for the following reason. In the absence of structure information, alignment is done on the basis of primary sequence conservation. Unknown helices are therefore much more likely to be properly aligned if they have high primary sequence conservation. Consequently, if the goal is to predict novel helices, an overzealous solution to this problem might prove to be more of a hindrance than a help.

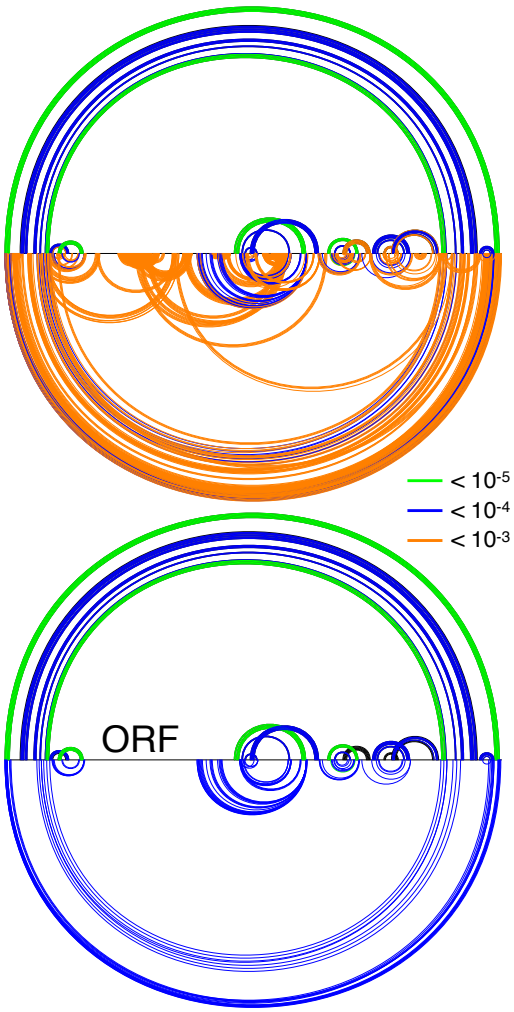


Figure 3.13: Arc diagrams of TRANSAT predictions (using the default p-value threshold of $10^{-3}$ (top) and $10^{-4}$ (bottom)) for the bacterial tmRNA family (RF00023). tmRNA contain a reading frame ending in a stop codon (labelled ORF on the bottom figure) which is important for its proper function. It serves as an ersatz reading frame to free mRNA from a stalled ribosome[108], and as such, it must be unstructured to function properly. At the less stringent threshold of $10^{-3}$ (top), some helices are predicted from in this region. The helices in this region, however, have p-values substantially higher than the p-values of the helices of the known structure. In this case, the p-value threshold of $10^{-4}$ may be more appropriate. At this p-value threshold, several novel helices are also identified, but none lie in ORF region. See figure 3.4 for information on arc diagrams.

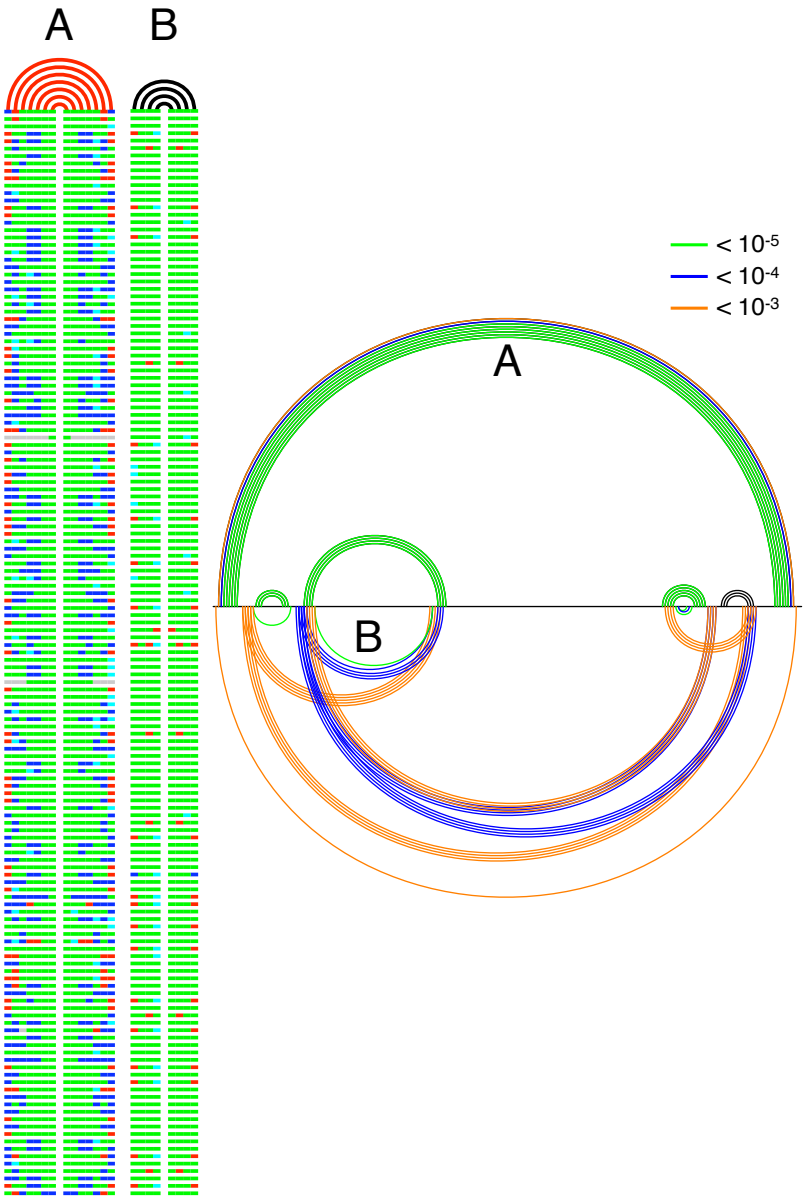


Figure 3.14: Arc diagram of TRANSAT predictions (using the default p-value threshold of $10^{-3}$) for the FMN riboswitch RFAM family (RF00050). These riboswitches are found in the 5'-untranslated regions of certain prokaryotic mRNAs encoding proteins related to flavin mononucleotide (FMN) biosynthesis. The covariation diagrams for Helix A from the known structure and novel helix B are shown on the left (see fig. 3.5 for details on covariation diagrams).

# Chapter 4

# Conserved competing helices

In the folding process, an RNA molecule may form temporary (transient) helices which are eventually displaced by the helices of the final structure. These transient helices may still be functional in that they guide the folding process in some way [45] or in that they are required to interact with other molecules [109].

For the Rfam datasets, we have at least partial knowledge of the functional structure. However, if transient structures that serve to guide the folding process exist, they are not annotated in Rfam. We therefore look for conserved competing helices. These are helices in which at least one of the positions has a different pairing partner in the known structure. Such conserved competing helices may be functional transient helices.

## 4.1 Competition definitions

Meyer and Miklós [47] defined four classes of competition between a given known structure and novel helices: 5'-cis, 3'-cis, 5'-trans, and 3'-trans (table 4.1). These classes are defined based on the relative positions of the known and competing base-pairs. We define an additional two competition classes (fig. 4.1); these classes were ignored in [47] because the loop regions of the helices in the known structures analyzed in that study were typically too short to accommodate alternative helices.

Identification of such events is straightforward in a single sequence once all possible helices are identified. Helix identification is performed as described in section 2.3 (using the same minimum helix length and loop length requirements of 4 and 3, respectively). First, all helices which contain at least one base-pair of the known structure are discarded (since here we are interesting only in competing helices). Then the base-pairs of each remaining helix are inspected to see if one or both members of a pair are base-paired in the known structure. If such is the case, then the helix can be considered a competing helix. The sequence positions of the helix base-pair and known base-pair together define a competition triplet consisting of three positions: the position that takes part in both the known and competing base-pair ($a$), its pairing partner in the known structure ($b$), and its pairing partner in the competing helix ($c$). The order in which these positions occur in the sequence determine a competition triplet's type (fig. 4.1).

For each class, Meyer and Miklós [47] define statistics to assess the 5'-to-3' symmetry for these competition classes in a single sequence. These statistics, Cis, Trans, and Mid, are

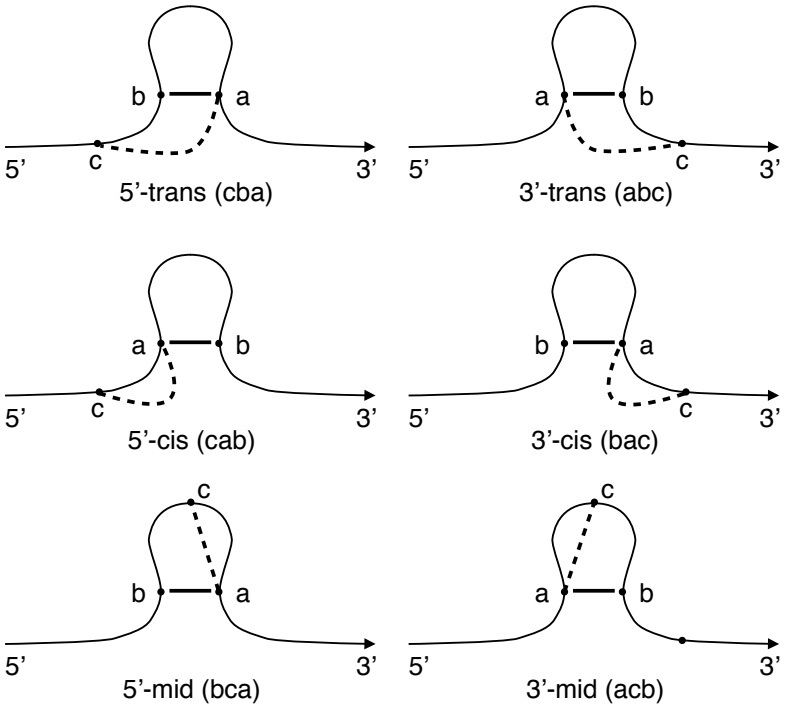


Figure 4.1: Pictorial definitions of 6 types of competing base-pairs. In the known structure, position $a$ pairs with position $b$. If position $c$ also pairs with position $a$ in a novel helix (and $c \neq b$), then this helix is incompatible with the helix containing the $a$:$b$ pair in the known structure. We call such helices *competing helices*. Note that a single base-pair can compete in both a 5' and 3' direction if both positions of the base-pair are paired with different positions in the known structure. A single helix may contain base-pairs that compete in many different configurations.

calculated from the positions of the three participating positions in a competition triplet, summed over the set of all such triplets found in a sequence, $C_t$, where $t$ is the competition class. Each competition triplet contributes to its respective class statistic with a weight proportional to $1/(d \log(l))$, where $d$ is the distance between competing position and the known base-pair and $l$ is the length of the subsequence available for competition of that class. For the Cis and Trans classes, $l$ is the length of the subsequence 5' of the 5'-most position or 3' of the 3'-most position of the known base-pair; for mid classes, $l$ is the distance between positions of the known base-pair. Competition triplets are weighted by the factor $1/d$ to penalize competing positions which are very distant from the base-pairs they are in competition with. The factor $1/\log(l)$ is a normalizing factor, adjusting the weights so that the relative competition of the classes are not biased by differing availability between 5' and 3' classes (the expected sum of $1/d$ terms over a subsequence of length $l$ is proportional to $\log(l)$, since $\int_1^l 1/x \; dx = \log(l)$). Such differing availability arises because the base-pairs of

the known structure are not necessarily evenly distributed throughout the sequence. For instance, imagine a 10-nucleotide sequence with single base-pair between positions 3 and 5; only positions 1 and 2 may compete with the base-pair in a 5'-cis manner, but positions 4-10 may compete with the base-pair in a 3'-cis manner. Table 4.1 contains the definitions of the Cis, Trans, and Mid statistics. The Cis and Trans statistics for single sequences are defined as they are in [47]. The Mid statistic definition is novel, but analogous to the other two. We define Cis($A$), Trans($A$), and Mid($A$) statistics for an alignment $A$ as the average value of the Cis, Trans, and Mid statistics for the sequences in that alignment.

| Statistic | Definition |
|---|---|
| $5'$-cis$(a,b,c)$ | $\frac{1}{(a-c)\log(a+1)}$ |
| $5'$-cis$(a,b,c)$ | $\frac{1}{(c-a)\log(L-a)}$ |
| $5'$-trans$(a,b,c)$ | $\frac{1}{(b-c)\log(b+1)}$ |
| $3'$-trans$(a,b,c)$ | $\frac{1}{(c-b)\log(L-b)}$ |
| $5'$-mid$(a,b,c)$ | $\frac{1}{(a-c)\log(a-b)}$ |
| $3'$-mid$(a,b,c)$ | $\frac{1}{(c-a)\log(b-a)}$ |
| $\text{Cis}_x$ | $\sum_{\{a,b,c\}\in C^x_{5'\text{-cis}}} 5'\text{-cis}(a,b,c) - \sum_{\{a,b,c\}\in C^x_{3'\text{-cis}}} 3'\text{-cis}(a,b,c)$ |
| $\text{Trans}_x$ | $\sum_{\{a,b,c\}\in C^x_{3'\text{-trans}}} 3'\text{-trans}(a,b,c) - \sum_{\{a,b,c\}\in C^x_{5'\text{-trans}}} 5'\text{-trans}(a,b,c)$ |
| $\text{Mid}_x$ | $\sum_{\{a,b,c\}\in C^x_{3'\text{-mid}}} 3'\text{-mid}(a,b,c) - \sum_{\{a,b,c\}\in C^x_{5'\text{-mid}}} 5'\text{-mid}(a,b,c)$ |
| Cis($A$) | $\frac{\sum_{x\in A}\text{Cis}_x}{\lvert A\rvert}$ |
| Trans($A$) | $\frac{\sum_{x\in A}\text{Trans}_x}{\lvert A\rvert}$ |
| Mid($A$) | $\frac{\sum_{x\in A}\text{Mid}_x}{\lvert A\rvert}$ |

Table 4.1: Definitions of Cis, Trans, and Mid statistics. A single competition triplet consists of three positions, $a$, $b$, and $c$, where $a$ pairs with $b$ in the known structure and $a$ pairs with $c$ in at least one competing helix. $L$ is the length of the sequence, and positions are indexed from 0. The first section outlines how a single competition triplet is weighted. The second section defines Cis, Trans, and Mid statistics for a single sequence. $C^x_t$ is the set of all competition events of type $t$ in sequence $s_x$. The last section shows the formulas for calculating Cis($A$), Trans($A$), and Mid($A$) statistics for an alignment $A$.

The Cis($A$), Trans($A$), and Mid($A$) statistics are used to assess the symmetry between

competing helices that are 5' or 3' to a helix of the known structure. If transient helices played no part RNA folding, one would expect the competing helices to be randomly distributed and therefore equally likely to contribute to a 5' or 3' class. If that were the case, then one would expect the statistics be normally distributed with a mean of zero. If they deviate significantly from zero, this is an indication that the competing helices found in the alignments are distributed non-randomly, implicating them in the folding process. Meyer and Miklós [47] found an imbalance toward 5'-cis and 3'-trans competition in a dataset of rRNA sequences. A 5'-cis helix may be formed before the helix it competes with is fully transcribed, so the overabundance of 5'-cis helices suggests that this kind of transient helix plays a role in RNA folding. 3'-Trans competition is theorized to be favored over 5'-trans competition because, in the 5'-trans, the position with two possible pairing partners is transcribed after both its pairing partners, so it will have two possible pairing partners when it is transcribed. This situation may lead to the frequent formation of the wrong helix. These asymmetries were only detectable in sequences that were full transcripts; in their dataset of sequences which were only a subsequence of their original transcript, no asymmetry was detected.

To calculate Cis, Trans, and Mid statistics for an alignment, we need: a) a known reference structure and b) a set of competing helices. The set of competing helices we can gather from TRANSAT by setting a p-value threshold. From the set of helices with p-values below that threshold, we extract those that contain no known structure base-pairs and use these helices as the set of competing helices needed to calculate each of these statistics. The procedure for calculating these statistics is as follows:

1. For each base-pair in each competing helix, we determine if it shares a position with a base-pair from the reference structure, and if so, what competition triplet(s) it participates in. At this stage, the positions of these triplets are with respect to the alignment.

2. We project each triplet onto each sequence of the alignment. If any of the projected positions are gaps, or if either of the two base-pairs of the triplet form non-canonical base-pairs in this sequence, we ignore its contribution statistics for this sequence.

3. We calculate the contribution of each triplet to the statistics for each sequence which passed the test in the previous step. We store a running total of each statistic for each sequence while we process all the competing helices.

4. Once all the competing helices have been processed, we calculate the alignment statistics by averaging each statistic over all the sequences in the alignment, giving every sequence equal weight.

Competing helices that play a role in co-transcriptional folding should not be randomly distributed with respect to the reference structure (assuming that the reference structure is

final 'target' structure). Therefore, if TRANSAT is successfully identifying conserved competing helices that are functional in co-transcriptional folding pathways, then one would expect the statistics to deviate significantly from 0 at low p-value thresholds.

## 4.2 Competing conserved helices in RFAM alignments

For our RFAM dataset, we have a single functional structure for each family, but no prior knowledge of additional functional transient helices. By looking at the distributions of Cis, Trans, and Mid statistics for the alignments in this dataset, we hoped to find evidence which would implicate the competing helices identified by TRANSAT in the co-transcriptional folding process.

For each alignment in the dataset, we calculated Cis, Trans, and Mid statistics for a range of p-value thresholds. At each p-value threshold, we took the mean of each statistic over all alignments, and calculated the 95% confidence interval for that mean.

Figure 4.2 shows how the mean of each statistic changes as we lower the p-value threshold, with the 95% confidence intervals shaded in green. Within this confidence interval, all three statistics do not diverge significantly from zero, even at low p-value thresholds.

This lack of a sequence signal may be interpreted in several ways. If transient helices which function in co-transcriptional folding are rare, if they are sequence specific, or if these transient helices are poorly aligned and therefore difficult for TRANSAT to detect, then it is possible that our sample size (consisting of 134 alignments) is not large enough to give us the statistical power to reliably detect the sequence signal. At present, we know of no studies which have attempted to estimate the prevalence of functional transient helices, which is unsurprising given that the prevalence of functional structures that are determined by folding kinetics (rather than equilibrium thermodynamics) is still an open question [54].

It is also possible that many of the sequences in these alignments are not full-length transcripts, i.e. they correspond only to subsequences of the transcripts that were originally transcribed from the genome. Some of the co-transcriptional 'context' of non-full-length transcripts is missing, and so such sequences may be misleading for this analysis. Meyer and Miklós [47] found that it was important to remove sequences that were not full-length transcripts from their dataset in order to detect the asymmetry in competing helix distribution they observed. Unfortunately, we could not find this information for many of the alignments in the RFAM dataset, and so we did not attempt to create a clean set of alignments known to be composed of full length transcripts.

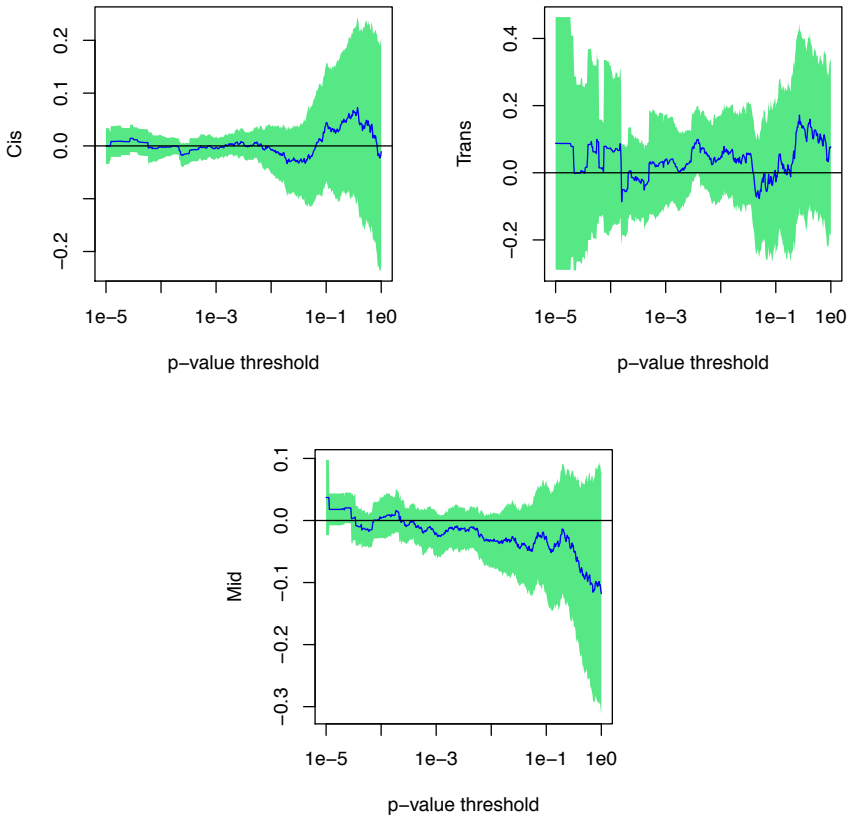


Figure 4.2: Cis, Trans, and Mid plots of the RFAM dataset as a function of the TRANSAT p-value threshold. The blue line corresponds to the mean values of the statistics over all alignments in the RFAM dataset. At each p-value, we also calculate the 95% confidence interval (shaded in green) for the corresponding mean value. For all three statistics, the zero line is encompassed by the 95% confidence interval at all p-value thresholds, suggesting that the competing helices assigned a low p-value threshold by TRANSAT have no asymmetry 5' or 3' to the helices of the known structure that one would expect if they were involved in co-transcriptional folding.

# Chapter 5

# Generated datasets

## 5.1 Motivation

The models of paired and unpaired sequence evolution introduced in section 2.4.1 can also be used as generative models. This implies that we can employ them to generate synthetic alignments by simulating the process of sequence evolution for a specified evolutionary tree and conserved RNA secondary structure. We use these generated alignments because it provides us with a source of alignments which are perfect in the sense that they contain no alignment errors and that they are perfectly in line with the given phylogenetic tree. More importantly for us, we know that the structural annotation of these alignments is both correct and complete.

The weaknesses of the models discussed in section 2.4.1 also apply to the synthetic sequences, so the generated alignments fail to capture the true complexity of biological sequences. In particular, our models only aim to model sequence conservation induced by RNA secondary structure, and not any other factors which may introduce other evolutionary pressures (e.g. conservation of protein binding sites [79]). In the generated alignments, all positions are under identical evolutionary pressure; no unpaired positions are more strongly conserved than any other unpaired position, and no paired positions are more strongly conserved than any other paired position. Our models also do not take into account insertions or deletions, so the generated alignments contain no gaps.

These generated datasets are nevertheless useful in that they allow us to evaluate Transat independently from concerns about alignment quality, tree quality, and the correctness and completeness of the annotation. The performance of Transat on these generated alignments can be considered Transat's best case performance. Synthetic alignments also allow us to systematically test how certain sequence alignment characteristics like alignment length affect Transat's performance.

## 5.2 Algorithm

The Felsenstein algorithm calculates the probability that alignment or alignment region would be produced by the model with a given phylogenetic tree [84]. It does so by considering the unknown sequences at the inner nodes of the tree, which represent the common ancestor sequence of two alignment sequences (or groups of sequences). Since we do not know the

common ancestor sequences, the Felsenstein algorithm sums the probabilities over all possible assignments the inner node sequences.

To generate new alignments, we reverse this process and turn the deterministic Felsenstein calculation into a probabilistic sampling procedure. For any single unpaired alignment column, we pick an initial nucleotide for the root node and then traverse the given tree, probabilistically assigning nucleotides to the inner nodes, until we reach the leaf nodes. The initial nucleotide at the root node is drawn from the prior probability distribution of nucleotides in unpaired columns (i.e. the equilibrium frequency of entries in the substitution matrix). As we move from a node to one of its two children, we draw the child node's nucleotide from the corresponding entries of the substitution matrix $S(t)$ for unpaired columns, using the node and its child as $t$. This operation simulates the mutation process as time progresses. The nucleotides at the leaf nodes correspond to the entries in one alignment column.

The process is analogous for paired alignment columns, only now we are evolving pairs of nucleotides rather than individual, unpaired nucleotides. We draw the initial pair of nucleotides from the prior probability distribution of nucleotide pairs, and choose new nucleotide pairs for inner nodes based on the evolutionary model for base-pairs.

**Algorithm 2** Reverse Felsenstein algorithm

**input:** root tree node $T$, structure $S$
$\Sigma_{\text{unpaired}} = \{\text{A}, \text{U}, \text{G}, \text{C}\}$
$\Sigma_{\text{paired}} = \{\text{AA}, \text{AU}, \text{AG}, \text{AC}, \text{UA}, \text{UU}, \ldots\}$
**for all** unpaired positions $i$ in $S$ **do**
  draw $n_{\text{start}}$ from prior probability $P(X = x) = \forall x \in \Sigma_{\text{unpaired}} : P(x)$
  reverseFels$(T, \{i\}, n_{\text{start}}, \Sigma_{\text{unpaired}})$
**end for**
**for all** paired positions $\{i, j\}$ in $S$ **do**
  draw $n_{\text{start}}$ from prior probability $P(X = x) = \forall x \in \Sigma_{\text{paired}} : P(x)$
  reverseFels$(T, \{i, j\}, n_{\text{start}}, \Sigma_{\text{paired}})$
**end for**

**function:** reverseFels
**input:** tree node $T$, position(s) $K$, starting nucleotide(s) $n$, alphabet $\Sigma$
$t \leftarrow$ branchLength$(T)$
Transition probabilities $P(n \rightarrow X = x) = \forall x \in \Sigma : e^{\mathbf{Q}t}[n, x]$
draw $m_{\text{new}}$ from $P(n \rightarrow X)$
**if** $T$ is leaf node **then**
  $s \leftarrow$ sequence which corresponds to T
  $s[K] \leftarrow m_{\text{new}}$
**else** {$T$ is inner node}
  **for all** $C \in$ children of $T$ **do**
    reverseFels$(C, K, m_{\text{new}}, \Sigma)$
  **end for**
**end if**

Algorithm 2 outlines the algorithm in pseudocode. It is essentially the same algorithm as used in Rose [110], a program which generates simulated alignments of DNA, RNA, or protein sequences. Rose does not allow the user to specify a conserved RNA structure, however.

## 5.3 Experiments

Using generated alignment datasets, we looked at two factors which may affect Transat's performance, alignment length and total tree length.

### 5.3.1 Alignment length

For a sequence of length $m$, the number of possible base-pairs for that sequence is proportional to $m^2$. One might therefore expect that Transat's performance would worsen as alignment length increases, since longer alignments would be expected to contain a higher ratio spurious to conserved helices. To test the effect of alignment length on Transat's performance, we generated a set of alignments with a wide range of alignment lengths and analyzed them with Transat.

We used known RNA secondary structures from the RNA STRAND database [111] as the input structures for alignment generation, selected at random from the unique structures in the database. A generated alignment's length corresponds to the length of the known structure (specified in dot-bracket notation) used to generate the alignment. In order to ensure that a wide range of alignment lengths was represented, we selected 50 structures at random (with replacement) from each of 9 length bins ($100-199, 200-299, \ldots, 900-999$), and generated one alignment from each structure, for a total of 450 alignments. This random selection is with replacement, since the RNA STRAND database contains less than 50 unique structures within the length ranges of certain bins.

The phylogenetic tree used to generate all of these alignments was a balanced binary tree with 10 leaf nodes (each alignment therefore contained 10 sequences). Each branch has the same length, and the total tree length is 4 (the median length of the Pfold trees used with the Rfam dataset is 3.3). We use these generated alignments, together with the corresponding trees, as input for Transat to analyze them in the same way as with the biological datasets.

Figure 5.1 outlines Transat's performance on the generated alignments, grouped by alignment length. In general, Transat performed very well on these alignments, identifying the helices of the known structure while excluding most other helices. The performance does not degrade substantially for longer alignments, which is surprising because in such alignments, the ratio of spurious helices to real ones will increase. The covariation signal seems to be robust enough to discriminate between spurious and conserved helices, even when this ratio is very high. The F-measure plot does show a shift to the left for longer alignments, suggesting that it may be advisable to use a lower p-value threshold for longer

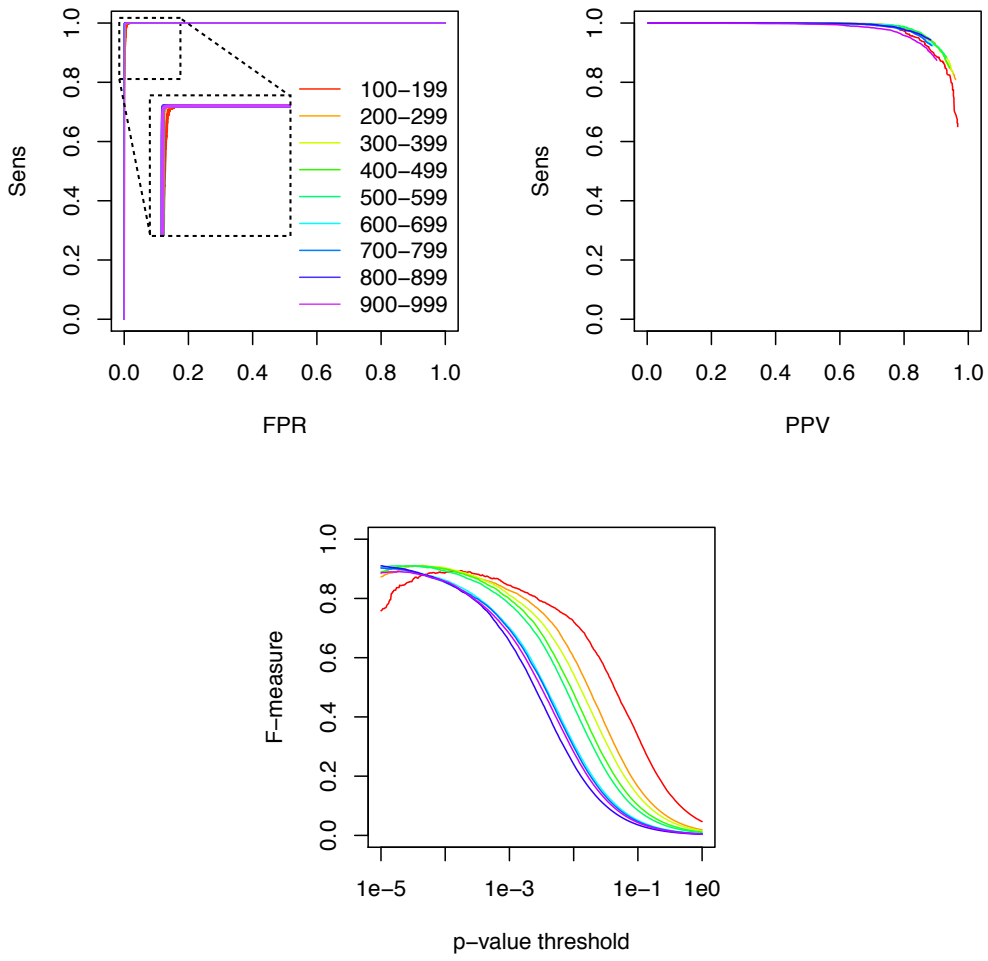


Figure 5.1: Helix-level performance of TRANSAT on generated alignments, coloured according to alignment length. Top left, the ROC curve, showing the tradeoff between sensitivity and false positive rate (FPR). Top right, the PPV/sensitivity curves. Bottom, a plot of F-measure as a function of the p-value threshold. Sensitivity, FPR, and F-measure are averaged over all alignments in a length bin at every p-value threshold.

alignments.

### 5.3.2 Tree length

The total tree length is the sum of all branch lengths, and gives an indication of how much sequence variation an alignment contains. Tree distance represents evolutionary time: the shorter the distance, the less time a sequence has to acquire mutations. An alignment corresponding to a tree with a low total length will therefore show less sequence variation than an alignment corresponding to a tree with a high total tree length. Since TRANSAT relies on covariation signals in the alignment, we expect that shorter tree lengths will adversely affect TRANSAT's performance.

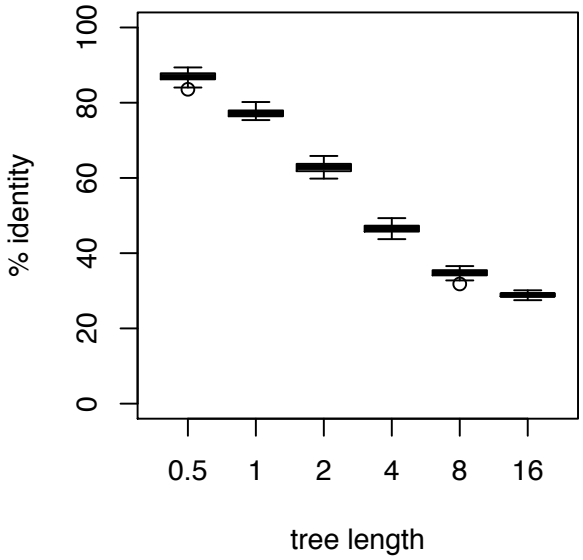


Figure 5.2: Box-and-whiskers plot showing the distribution of mean pairwise sequence identity of alignments generated from trees with different total lengths. All alignments comprise 10 sequences. Each box depicts the upper and lower quartiles (the central line is the median). The whiskers extend to the maximum and minimum data points within 1.5 times the interquartile range (IQR) of the upper or lower quartile. The circles represent outlier data-points, i.e. those further than $1.5 \times$ IQR from the upper or lower quartiles.

To test the effect of total tree length on TRANSAT's performance, we generated sets of alignments from the same structure but with varying total tree lengths. Again, we constructed balanced binary trees with 10 leaf nodes. As before, all branches in the same tree have equal lengths. We created six trees, with total tree lengths on a log scale between $2^{-1}$ and $2^4$, and selected 10 structures from each of the 9 alignment length bins (described the previous

section) to use as the input for generating alignments. For each structure, an alignment was generated with every tree. In total, 540 alignments were generated for this experiment. Figure 5.2 shows the effect of total tree length on the generated alignments' average pairwise sequence identity, a commonly used measure of sequence similarity in an alignment.

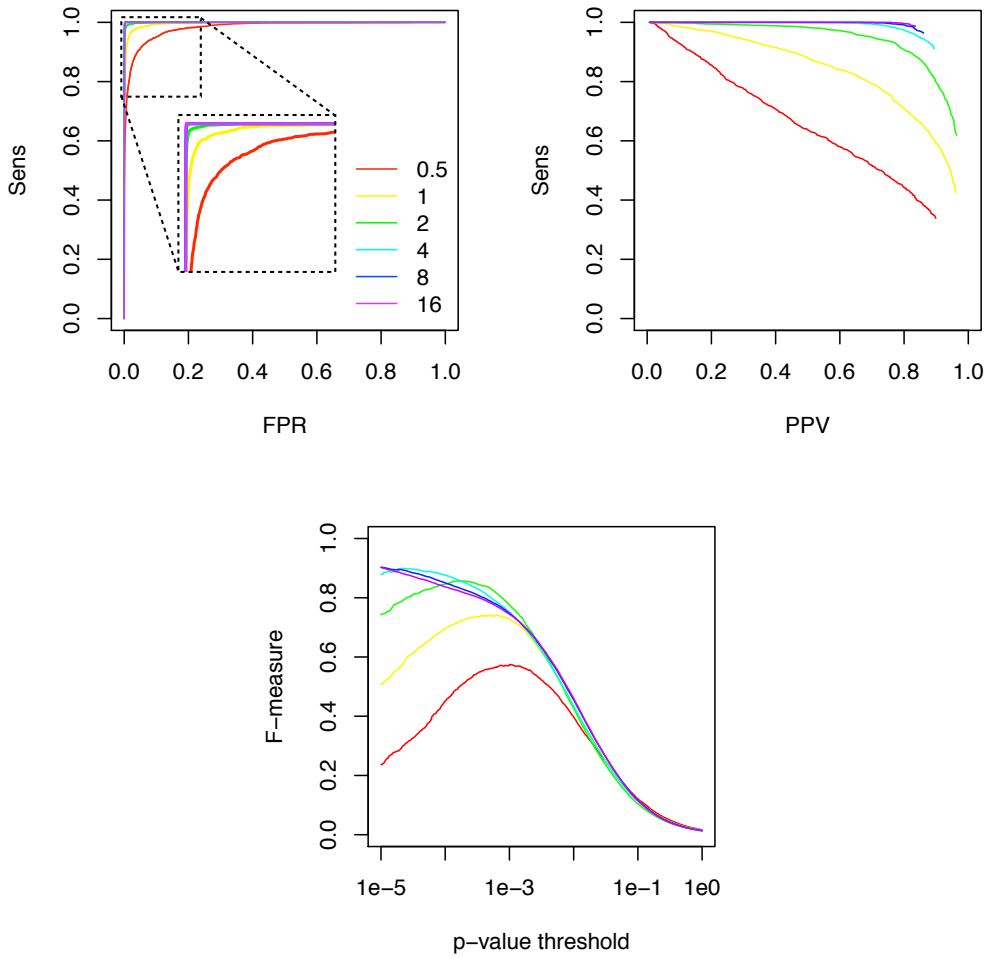


Figure 5.3: Helix-level performance of TRANSAT on generated alignments, coloured according to the total length of the tree used to generate the alignment. Top left, the ROC curve, showing the tradeoff between sensitivity and false positive rate (FPR). Top right, the PPV/ sensitivity curves. Bottom, a plot of F-measure as a function of the p-value threshold. Sensitivity, FPR, and F-measure are averaged over all alignments at every p-value threshold.

We used these alignments (and corresponding trees) and inputs to TRANSAT, and analyzed its performance on alignments generated from trees of different total lengths (fig. 5.3). For shorter trees, TRANSAT's performance declines substantially, as expected since the co-

variation between pairs of alignment columns on which TRANSAT's predictions are based disappears as we reduce the total tree length.

However, for a set of real sequences, a high degree of sequence variation also poses a problem. More distantly related sequences are more difficult to align correctly based on primary sequence conservation only. The performance of RNA secondary structure prediction methods which take a fixed input alignment (e.g. RNAALIFOLD [23]) declines relative to the performance of methods which allow some flexibility in the alignment for alignments with sequence identity below 70% [28]. Alignments like those produced from the trees of length 16, where the percent sequence identity approaches 25% (fig. 5.2), would be very difficult to align correctly without taking RNA structure into account.

In general, TRANSAT performs better on these generated alignment datasets than on the biological datasets. This is unsurprising, since multiple sequence alignment is not an easy task, and is a stumbling block for many programs that predict RNA structure [20, 26]. By starting with a perfect alignment and tree, we give ourselves a considerable head-start, bypassing the problems of alignment errors, tree errors, and incomplete structural annotation present in the datasets of biological sequences. Nevertheless, the fact that TRANSAT performs so well on these generated alignments demonstrates the theoretical viability of our approach.

## 5.4 Alignments with competing helices

With the models for paired and unpaired positions, we can generate alignments with a single conserved secondary structure that is non-overlapping (i.e. each position in the alignment is paired with at most one other position). However, since the strength of TRANSAT is that it can find competing helices, we would also like to assess this feature.

In our models, unpaired alignment positions evolve independently of all other positions, so we need only a $4 \times 4$ matrix to capture all the possible mutations at one position. Paired positions evolve independently of all positions except that of their pairing partner, and so we need a $16 \times 16$ rate matrix to capture the transition rates between all 16 possible pairs of nucleotides. In alignments with competing helices, certain positions will be dependent on more than one other position (fig. 5.4). Modeling arbitrary dependencies quickly becomes infeasible, since the number of permutations of a nucleotide $n$-tuple is $4^n$, so the size of the corresponding rate matrix would be $4^n \times 4^n$. However, the rate matrix necessary to model triples is sufficiently small ($64 \times 64$) to be practical. With such a model, we could model cases where single position pairs with two other positions, i.e. where two helices are overlapping.

There are not enough high-quality alignments with known competing helices available to train such a model from data, so instead we infer the model from the model of paired positions. For a nucleotide triple $\{a, b, c\}$ where $a$ pairs with $b$ and $b$ pairs with $c$, substitution

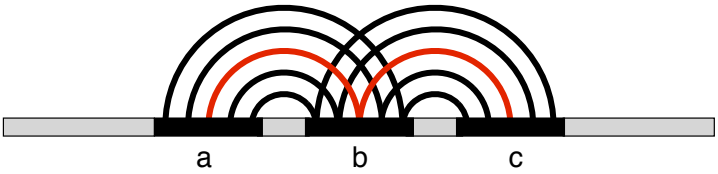


Figure 5.4: Arc diagram of two competing helices. Because position b is paired with position a in one helix and position b in the other, the three positions do not evolve independently of each other.

probabilities can be calculated in the following manner:

$$P\left(abc \to a'b'c'|t\right) = P\left(ab \to a'b'|t\right) \cdot P\left(c \to c'|ab \to a'b', t\right) \tag{5.1}$$

Nucleotide $c$ is not paired with $a$, and so we assume it is independent from $a$ given $b$, allowing us to state the probability entirely in terms of substitution probabilities for individual pairs:

$$\begin{aligned} P\left(abc \to a'b'c'|t\right) &= P\left(ab \to a'b'|t\right) \cdot P\left(c \to c'|ab \to a'b', t\right) \\ &\approx P\left(ab \to a'b'|t\right) \cdot P\left(c \to c'|b \to b', t\right) \\ &\approx P\left(ab \to a'b'|t\right) \cdot \frac{P\left(bc \to b'c'|t\right)}{P\left(b \to b'|t\right)} \\ &\approx P\left(ab \to a'b'|t\right) \cdot \frac{P\left(bc \to b'c'|t\right)}{\sum_i P\left(bc \to b'c_i'|t\right)} \end{aligned} \tag{5.2}$$

This assumption of conditional independence only be an approximation. One can imagine that if the helix which contains the $\{b, c\}$ base-pair displaces the helix containing the $\{a, b\}$ base-pair, the relative stability of the two helices would impact their function. In the absence of adequate training data, however, we feel that it is a reasonable assumption.

The rate matrix $\mathbf{R}$ is derived from the matrix of substitution probabilities $S(t)$ for branch length $t = 1$, and can then be used to calculate substitution probabilities for arbitrary branch lengths:

$$\mathbf{R} = \ln\left(S(t=1)\right) \tag{5.3}$$

Prior probabilities for triples at the root node are inferred with the transition probabilities from an arbitrary triple $a, b, c$ to any triple $a', b', c'$ at a sufficiently large time $t$:

$$P(a', b', c') = \lim_{t\to\infty} P(abc \to a'b'c'|t) = \lim_{t\to\infty} e^{\mathbf{R}t}\left[abc \to a'b'c'\right] \tag{5.4}$$

Note that the order of nucleotides $a$, $b$, and $c$ matters. In the paired position model, $P(ab \to a'b'|t) \neq P(ba \to b'a'|t)$, so different models are required for different orderings of

$a$, $b$, and $c$. There are three possible orderings, corresponding the three possible position of nucleotide which is participates in two pairs ($b$): $b$ can be located between $a$ and $c$ (order: $(a, b, c)$), $b$ can be the 5'-most position (order: $(b, a, c)$), or $b$ can be the 3'-most position (order: $(a, c, b)$). These models are probably not wildly different from one another, however.

With this additional evolutionary model, we can generate alignments with competing helices using essentially the algorithm as was outlined in Algorithm 2, modified to subdivide the positions of the structure into unpaired, paired, and triple categories. For convenience, we have implemented only the model for order: $(a, b, c)$, and we use it here only to generate alignments with structures whose triples conform to that ordering.

## 5.5 Experiments on generated alignments with competing helices

In order to test TRANSAT's ability to identify competing helices, we generated a set of simulated alignments with conserved competing helices. For simplicity, we generate alignments with a simple structure containing two competing helices (fig. 5.5). From this structure, we generated ten alignments of ten sequences each with trees of six different total lengths ($2^{-1}$ to $2^4$, as in section 5.3.2), a total of 60 alignments. TRANSAT was then given these alignments along with their corresponding trees as input to predict conserved helices.

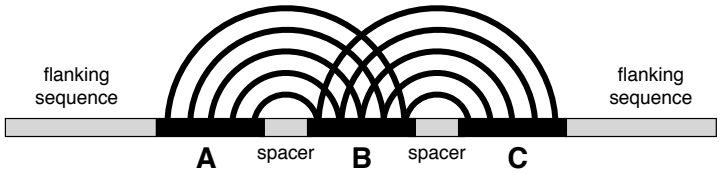


Figure 5.5: Arc diagram of the structure used to generate alignments. The structure has two helices of 5 base-pairs each. Because the helices overlap, each position in region B has two pairing partners. The spacer regions are 5 nucleotides long, and each the 5' and 3' flanking sequences is 50 nucleotides long. In total, the alignment is 125 nucleotides long.

Figure 5.6 shows the performance of TRANSAT on this set of generated alignments. As before, TRANSAT is relatively successful in recovering the helices of the original structure. No performance drop-off was observed at shorter total tree lengths, which is different from what we observed in our experiments with generated alignments with no competing helices (section 5.3.2), where TRANSAT's performance degraded on alignments generated from trees with low total lengths.

When testing TRANSAT on alignments of biological sequences, we estimated phylogenetic trees using the maximum likelihood tree prediction program included with PFOLD (see section 3.1.4). To test whether this procedure affects TRANSAT's performance, we generated PFOLD

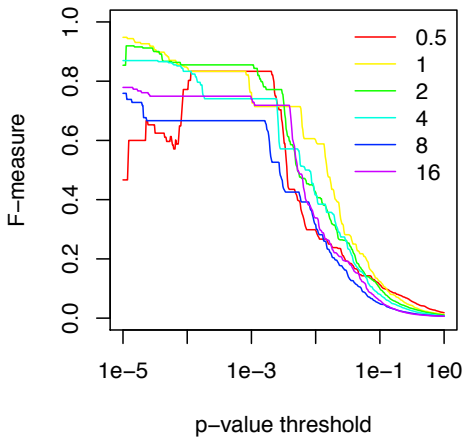


Figure 5.6: F-measure plot of TRANSAT's base-pair-level performance on generated alignments with competing helices. The lines show the performance averaged over ten alignments generated the same tree, labelled according to its total length. TRANSAT's performance is strong for all tree lengths.

trees for each of the generated alignments used here and compared TRANSAT's performance using the true tree from which the alignment was generated and the PFOLD-estimated trees. TRANSAT's performance given the estimated trees was nearly identical to its performance with the true tree (fig. 5.7), suggesting that the PFOLD trees we use elsewhere are accurate enough for the purposes of TRANSAT.

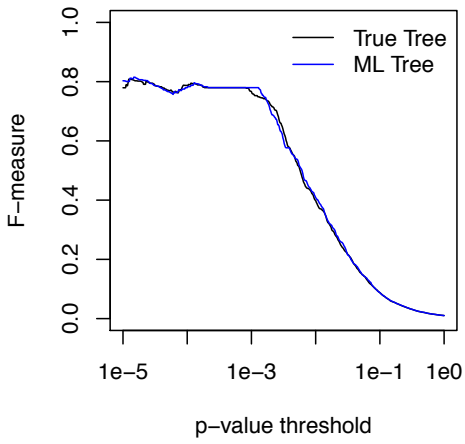


Figure 5.7: F-measure plot of TRANSAT's mean base-pair-level performance for generated alignments with competing helices, given either the true tree or the maximum likelihood (ML) tree inferred with PFOLD [25]. TRANSAT performs equally well with either tree.

# Chapter 6

# Conclusion and future work

We designed our program TRANSAT to identify evolutionarily conserved helices. It relies exclusively on evolutionary information to rank the helices that it identifies, allowing us to interpret high-scoring, statistically significant helices as being functionally important. Since TRANSAT relies so heavily on evolutionary information, it requires large, high-quality alignments to be accurate. Because TRANSAT places no restrictions on the number of pairing partners a single position might have, it is capable of detecting functional alternative structures, such as the transient helices involved in the folding pathway of *hok*. This approach is advantageous in that it implicitly takes into account interactions between the RNA and other molecules in the cell, in contrast to folding pathway prediction programs which rely on simulation of RNA folding. By taking this approach, we lose the ability to predict folding pathways for an RNA molecule as a function of time. We can, however, predict functional structural features of folding pathway without requiring specific knowledge about the cellular environment or the possible interactions with other molecules *in vivo*, knowledge which is at present largely unattainable.

Testing the quality of TRANSAT's predictions proved somewhat challenging, because of the lack of alignments of RNA sequences with known alternative helices. We tested TRANSAT on two alignments with known alternative helices, the *hok* alignment and the *trp*-attenuator alignment (section 3.3). On the higher quality *hok* alignment, TRANSAT performed well, identifying most of the known helices and relatively few extra ones. TRANSAT performed less well on the *trp*-attenuator alignment, especially when compared with RNAALISHAPES, which could identify the alternative structures of this alignment [53]. TRANSAT's weakness on this alignment is indicative of the importance of alignment quality to TRANSAT's performance; the *trp*-attenuator alignment contained less covariation than the *hok* alignment and more non-canonical base-pairs (compare fig. 3.2 to fig. 3.1).

Because of the lack of RNA sequence alignments with known alternative structures, we also tested our method on a large set of alignments from the RFAM database with a single non-overlapping known structure (section 3.4). TRANSAT was capable of identifying the known structure in these alignments quite well on average, though there was a good deal of variation in its performance. We would not recommend using TRANSAT as an exclusive means of predicting an RNA's secondary structure if one suspects that the RNA has only one functional secondary structure, but it is nonetheless useful for analyzing sequences with

known structure to look for evidence of conserved alternative structures. We did not find evidence of asymmetry in the 5'-to-3' distribution of low p-value novel competing helices which overlapped with the known structure of these alignments (section 4.2), but careful analysis of TRANSAT's predictions on certain families in the RFAM dataset revealed several interesting novel helices (section 3.4.3). TRANSAT's predictions may provide a useful starting point for experimental studies, since helices can be ranked by p-value, allowing one to test the most promising structures first by means of mechanical unfolding [18, 72] or other methods [45].

To test TRANSAT more thoroughly, we also evaluated its performance on several sets of simulated alignments. TRANSAT's performance proved to be sensitive to total tree length (a proxy for sequence variation), as expected since TRANSAT relies on sequence variation to make its predictions, but relatively independent of alignment length. We also developed a method for generating alignments with conserved competing helices, and showed that TRANSAT is capable of identifying such helices as well as non-competing helices.

In the absence of additional well-studied RNA sequence alignments with known alternative structures, folding simulation programs provide another method by which to test TRANSAT's performance, by comparing it with the predictions of such programs. Some work on this topic has been started by Yaojie Chen as part of a rotation project with Irmtraud Meyer.

## 6.1 Improving TRANSAT

In the previous chapters, we have suggested several ways in which TRANSAT might be improved:

- Modeling the null distribution of log-likelihood scores with a normal distribution would require fewer randomized alignments, improving TRANSAT's running time (fig. 2.7).
- The evolutionary models used by TRANSAT pose a problem in that primary sequence conservation is favoured by the paired model. For alignments with many sequences, this causes TRANSAT to score helices with high primary sequence conservation highly, even if they are unsupported by covariation (fig. 3.14). This problem might be addressed by either altering the models to reduce this bias, or by scaling the tree used in the calculation of the unpaired model likelihood. The tree-scaling approach is similar to the strategy used by the program RNA-DECODER to account for increased primary sequence conservation induced by conservation of codon positions in protein-coding RNA sequences [112, 113].
- The problem of helices with extending base-pairs (base-pairs on the inner or outer margin of a helix that are less well conserved than the rest of the helix — see section 3.3.1) in TRANSAT's predicted helices is would probably be straightforward to address by filtering out base-pairs that fit that description. The prevalence of these base-pairs,

however, raises interesting questions about the nature of our predictions. TRANSAT, along with all alignment-based structure prediction programs discussed here, predicts common structures for the alignment as a whole. The underlying assumption is that every sequence in the alignment shares a common functional structure. It is conceivable, however, that over the course of evolution, the common ancestor of only some of the sequences in the alignment evolved a slightly different, but still functional, structure. In such a scenario, TRANSAT would likely miss this structure, since it is not conserved over the whole alignment. Partially conserved structures, such as the extending base-pairs we observe, could conceivably represent this sort of structural divergence within an alignment, and it might therefore be worthwhile to develop a method for identifying such structures. In the absence of such a method, there is an inherent tension in the level of alignment sequence diversity which is optimal for structure prediction; one would like a diverse set of sequences so as to be able to observe a reasonable amount of covariation, but a more diverse set of sequences is more likely to contain this sort of structural diversity.

## 6.2 TRANSAT in relation to other methods of structure prediction

One may view the type of structure prediction done by TRANSAT as a step in a succession of increasingly complex structure prediction problems. For the problem of predicting a single pseudoknot-free structure, the search space is limited by the requirements that a position can be paired with only one other position, and that base-pairs must be nested. The search space expands when we allow pseudoknotted structures. TRANSAT predicts structures from an even more expanded search space, where we remove the restriction that a position may only pair with one other position. For a sequence (or alignment) of length $m$, there are $m(m-1)/2$ possible base-pairs, and so the number of possible structures of this type is $2^{m(m-1)/2}$. Folding pathway prediction is even more complex, since it adds the time dimension.

Since our problem is more challenging than 'ordinary' structure prediction, one avenue to improving the quality of TRANSAT's predictions is to incorporate some amount of flexibility in the alignment, as is done in the leading programs for alignment-based structure prediction [26, 28, 29]. For this work, we have used only manually curated alignments, but such alignments are rarely available, and not compatible with high-throughput analyses. Indeed, the difficulty of acquiring high quality alignments of sequences with known alternative structures has been problematic for evaluating TRANSAT's performance.

From this perspective, it is clear that our problem may be phrased as one of binary classification, where each possible base-pair is classified as either pairing or non-paring. This phrasing suggests an alternative approach to that of TRANSAT, focussing on classification

of individual base-pairs rather than helices. There exists wide array of machine learning techniques that may be applied to this type of classification problem [114]. One might also phrase it as a metric labeling problem [115], since base-pair stacking imposes a relationship between a pair of positions and its neighbors. Set in this light, the problem is also somewhat reminiscent of the earliest days of RNA structure prediction using dot-matrices [11, 12].

This hierarchical perspective of RNA structure prediction problems also suggests that methods such as TRANSAT may be useful as a foundation to build upon in order to tackle the folding pathway prediction problem. The most obvious scheme would be use TRANSAT as part of a helix-based folding simulation, such as KINEFOLD [57], to favor the formation of well-conserved helices as they become available. This approach may be a fruitful means of overcoming the limitations of purely simulation-based methods.

# Bibliography


[1] M. Inui, G. Martello, and S. Piccolo, "MicroRNA control of signal transduction," *Nat Rev Mol Cell Biol*, vol. 11, pp. 252–63, Apr 2010.

[2] S. R. Eddy, "Computational genomics of noncoding RNA genes," *Cell*, vol. 109, pp. 137–40, Apr 2002.

[3] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Hum Mol Genet*, vol. 15 Spec No 1, pp. R17–29, Apr 2006.

[4] FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), "The transcriptional landscape of the mammalian genome," *Science*, vol. 309, pp. 1559–63, Sep 2005.

[5] N. R. Pace and T. L. Marsh, "RNA catalysis and the origin of life," *Orig Life Evol Biosph*, vol. 16, no. 2, pp. 97–116, 1985.

[6] W. Gilbert, "Origin of life: The RNA world," *Nature*, vol. 319, p. 618, 1986.

[7] P. Doty, H. Boedtker, J. R. Fresco, R. Haselkorn, and M. Litt, "Secondary structure in ribonucleic acids," *Proc Natl Acad Sci U S A*, vol. 45, pp. 482–99, Apr 1959.

[8] J. R. Fresco, B. M. Alberts, and P. Doty, "Some molecular details of the secondary structure of ribonucleic acid," *Nature*, vol. 188, pp. 98–101, Oct 1960.

[9] W. K. Olson, M. Esguerra, Y. Xin, and X.-J. Lu, "New information content in RNA base pairing deduced from quantitative analysis of high-resolution structures," *Methods*, vol. 47, pp. 177–86, Mar 2009.

[10] Y. Byun and K. Han, "PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots," *Bioinformatics*, vol. 25, pp. 1435–7, Jun 2009.

[11] I. Tinoco, Jr, O. C. Uhlenbeck, and M. D. Levine, "Estimation of secondary structure in ribonucleic acids," *Nature*, vol. 230, pp. 362–7, Apr 1971.

[12] W. Fitch, "Considerations regarding the regulation of gene transcription and messenger translation," *J Mol Evol*, vol. 1, no. 2, pp. 185–207, 1972.

[13] R. Nussinov, G. Pieczenik, J. Griggs, and D. Kleitman, "Algorithms for loop matchings," *SIAM J Appl Math*, vol. 35, no. 1, pp. 68–82, 1978.

[14] M. Waterman, "Secondary structure of single-stranded nucleic acids," *Adv math suppl studies*, vol. 1, pp. 167–212, 1978.

[15] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.," *Nucleic Acids Res*, vol. 9, pp. 133–148, Jan 1981.

[16] H. Isambert, "The jerky and knotty dynamics of RNA," *Methods*, Jun 2009.

[17] E. Buratti and F. E. Baralle, "Influence of RNA secondary structure on the pre-mRNA splicing process," *Mol Cell Biol*, vol. 24, pp. 10505–14, Dec 2004.

[18] W. J. Greenleaf, K. L. Frieda, D. A. N. Foster, M. T. Woodside, and S. M. Block, "Direct observation of hierarchical folding in single riboswitch aptamers," *Science*, vol. 319, pp. 630–3, Feb 2008.

[19] G. W. Fox and C. R. Woese, "5S RNA secondary structure," *Nature*, vol. 256, pp. 505–7, Aug 1975.

[20] P. P. Gardner and R. Giegerich, "A comprehensive comparison of comparative RNA structure prediction approaches," *BMC Bioinf*, vol. 5, p. 140, Sep 2004.

[21] R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo, "Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods," *Nucleic Acids Res*, vol. 20, pp. 5785–95, Nov 1992.

[22] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabalistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press, 1998.

[23] I. L. Hofacker, M. Fekete, and P. F. Stadler, "Secondary structure prediction for aligned RNA sequences," *J Mol Biol*, vol. 319, pp. 1059–1066, Jun 2002.

[24] B. Knudsen and J. Hein, "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history," *Bioinformatics*, vol. 15, pp. 446–54, Jun 1999.

[25] B. Knudsen and J. Hein, "Pfold: RNA secondary structure prediction using stochastic context-free grammars.," *Nucleic Acids Res*, vol. 31, pp. 3423–3428, Jul 2003.

[26] I. M. Meyer and I. Miklos, "Simulfold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework.," *PLoS Comput Biol*, vol. 3, p. e149, Aug 2007.

[27] D. Sankoff, "Simultaneous solution of the RNA folding, alignment and protosequence problems," *SIAM Journal on Applied Mathematics*, vol. 45, no. 5, pp. 810–825, 1985.

[28] X. Xu, Y. Ji, and G. D. Stormo, "RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment," *Bioinformatics*, vol. 23, pp. 1883–1891, Aug. 2007.

[29] A. O. Harmanci, G. Sharma, and D. H. Mathews, "Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign," *BMC Bioinf*, vol. 8, p. 130, 2007.

[30] W. Winkler, A. Nahvi, and R. R. Breaker, "Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression," *Nature*, vol. 419, pp. 952–6, Oct 2002.

[31] A. D. Garst and R. T. Batey, "A switch in time: Detailing the life of a riboswitch," *Biochim Biophys Acta*, Jul 2009.

[32] N. Sudarsan, J. E. Barrick, and R. R. Breaker, "Metabolite-binding RNA domains are present in the genes of eukaryotes," *RNA*, vol. 9, pp. 644–7, Jun 2003.

[33] M. Mandal and R. R. Breaker, "Gene regulation by riboswitches.," *Nat Rev Mol Cell Biol*, vol. 5, pp. 451–463, Jun 2004.

[34] W. Winkler, S. Cohen-Chalamish, and R. Breaker, "An mRNA structure that controls gene expression by binding FMN," *Proc Natl Acad Sci U S A*, vol. 99, pp. 15908–15913, Dec. 2002.

[35] M. T. Cheah, A. Wachter, N. Sudarsan, and R. R. Breaker, "Control of alternative RNA splicing and gene expression by eukaryotic riboswitches," *Nature*, vol. 447, pp. 497–500, May 2007.

[36] J. Johansson, P. Mandin, A. Renzoni, C. Chiaruttini, M. Springer, and P. Cossart, "An RNA thermosensor controls expression of virulence genes in listeria monocytogenes," *Cell*, vol. 110, pp. 551–61, Sep 2002.

[37] P. S. Ray, J. Jia, P. Yao, M. Majumder, M. Hatzoglou, and P. L. Fox, "A stress-responsive RNA switch regulates VEGFA expression," *Nature*, vol. 457, pp. 915–9, Feb 2009.

[38] E. A. Schultes and D. P. Bartel, "One sequence, two ribozymes: implications for the emergence of new ribozyme folds," *Science*, vol. 289, pp. 448–52, Jul 2000.

[39] O. C. Uhlenbeck, "Keeping RNA happy.," *RNA*, vol. 1, pp. 4–6, Mar 1995.

[40] J. Boyle, G. T. Robillard, and S. H. Kim, "Sequential folding of transfer RNA. a nuclear magnetic resonance study of successively longer tRNA fragments with a common 5' end," *J Mol Biol*, vol. 139, pp. 601–25, Jun 1980.

[41] C. K. Ma, T. Kolesnikow, J. C. Rayner, E. L. Simons, H. Yim, and R. W. Simons, "Control of translation by mRNA secondary structure: the importance of the kinetics of structure formation," *Mol Microbiol*, vol. 14, pp. 1033–47, Dec 1994.

[42] S. L. Heilman-Miller and S. A. Woodson, "Effect of transcription on folding of the Tetrahymena ribozyme," *RNA*, vol. 9, pp. 722–33, Jun 2003.

[43] L. Zhang, P. Bao, M. J. Leibowitz, and Y. Zhang, "Slow formation of a pseudoknot structure is rate limiting in the productive co-transcriptional folding of the self-splicing Candida intron," *RNA*, vol. 15, pp. 1986–92, Nov 2009.

[44] D. Repsilber, S. Wiese, M. Rachen, A. W. Schröder, D. Riesner, and G. Steger, "Formation of metastable RNA structures by sequential folding during transcription: time-resolved structural analysis of potato spindle tuber viroid (-)-stranded RNA by temperature-gradient gel electrophoresis," *RNA*, vol. 5, pp. 574–84, Apr 1999.

[45] T. N. Wong, T. R. Sosnick, and T. Pan, "Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures," *Proc Natl Acad Sci U S A*, vol. 104, pp. 17995–18000, Nov 2007.

[46] A. Xayaphoummine, V. Viasnoff, S. Harlepp, and H. Isambert, "Encoding folding paths of RNA switches.," *Nucleic Acids Res*, vol. 35, no. 2, pp. 614–622, 2007.

[47] I. M. Meyer and I. Miklós, "Co-transcriptional folding is encoded within RNA genes.," *BMC Mol Biol*, vol. 5, p. 10, Aug 2004.

[48] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, no. 6-7, pp. 1105–19, 1990.

[49] S. Wuchty, W. Fontana, I. Hofacker, and P. Schuster, "Complete suboptimal folding of RNA and the stability of secondary structures," *Biopolymers*, vol. 49, pp. 145–165, Feb. 1999.

[50] P. Stein and M. Waterman, "On some new sequences generalizing the Catalan and Motzkin numbers," *Discrete Math*, vol. 26, no. 3, pp. 261–272, 1979.

[51] B. Voß, R. Giegerich, and M. Rehmsmeier, "Complete probabilistic analysis of RNA shapes," *BMC Biol*, vol. 4, p. 5, 2006.

[52] Y. Ding and C. E. Lawrence, "A statistical sampling algorithm for RNA secondary structure prediction," *Nucleic Acids Res*, vol. 31, pp. 7280–301, Dec 2003.

[53] B. Voß, "Structural analysis of aligned RNAs," *Nucleic Acids Res*, vol. 34, no. 19, pp. 5471–81, 2006.

[54] C. Flamm and I. Hofacker, "Beyond energy minimization: approaches to the kinetic folding of RNA," *Monatsh Chem*, vol. 139, no. 4, pp. 447–457, 2008.

[55] C. Flamm, W. Fontana, I. L. Hofacker, and P. Schuster, "RNA folding at elementary step resolution," *RNA*, vol. 6, pp. 325–38, Mar 2000.

[56] H. Isambert and E. D. Siggia, "Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme," *Proc Natl Acad Sci U S A*, vol. 97, pp. 6515–20, Jun 2000.

[57] A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert, "Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations," *Proc Natl Acad Sci U S A*, vol. 100, pp. 15310–5, Dec 2003.

[58] A. Xayaphoummine, T. Bucher, and H. Isambert, "Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots.," *Nucleic Acids Res*, vol. 33, pp. W605–10, Jul 2005.

[59] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure," *Proc Natl Acad Sci U S A*, vol. 101, pp. 7287–92, May 2004.

[60] C. Flamm, I. Hofacker, P. Stadler, and M. Wolfinger, "Barrier trees of degenerate landscapes," *Z Phys Chem*, vol. 216, pp. 155–173, 2002.

[61] M. Wolfinger, W. Svrcek-Seiler, C. Flamm, I. Hofacker, and P. Stadler, "Efficient computation of RNA folding dynamics," *J Phys A: Math Gen*, vol. 37, pp. 4731–4741, Apr. 2004.

[62] X. Tang, B. Kirkpatrick, S. Thomas, G. Song, and N. Amato, "Using motion planning to study RNA folding kinetics," *J Comp Biol*, vol. 12, pp. 862–881, July 2005.

[63] X. Tang, S. L. Thomas, L. Tapia, and N. M. Amato, "Tools for simulating and analyzing RNA folding kinetics," in *RECOMB* (T. P. Speed and H. Huang, eds.), vol. 4453 of *Lecture Notes in Computer Science*, pp. 268–282, Springer, 2007.

[64] W. Zhang and S. Chen, "Analyzing the biopolymer folding rates and pathways using kinetic cluster method," *J Chem Phys*, vol. 119, pp. 8716–8729, Oct. 2003.

[65] W. Zhang and S. Chen, "Exploring the complex folding kinetics of RNA hairpins: I. General folding kinetics analysis," *Biophys J*, vol. 90, pp. 765–77, Feb 2006.

[66] S. Cao and S. Chen, "Biphasic folding kinetics of RNA pseudoknots and telomerase RNA activity," *J Mol Biol*, vol. 367, pp. 909–24, Mar 2007.

[67] C. Heine, G. Scheuermann, C. Flamm, I. L. Hofacker, and P. F. Stadler, "Visualization of barrier tree sequences," *IEEE Trans Vis Comput Graph*, vol. 12, no. 5, pp. 781–788, 2006.

[68] I. L. Hofacker, C. Flamm, C. Heine, M. T. Wolfinger, G. Scheuermann, and P. F. Stadler, "BarMap: RNA folding on dynamic energy landscapes," *RNA*, May 2010.

[69] M. Geis, C. Flamm, M. T. Wolfinger, A. Tanzer, I. L. Hofacker, M. Middendorf, C. Mandl, P. F. Stadler, and C. Thurner, "Folding kinetics of large RNAs.," *J Mol Biol*, vol. 379, pp. 160–173, May 2008.

[70] J. Manuch, C. Thachuk, L. Stacho, and A. Condon, *Lecture Notes in Computer Science*, vol. 5877, ch. NP-completeness of the direct energy barrier problem without pseudoknots, pp. 106–115. Springer Berlin, 2009.

[71] J. H. A. Nagel, A. P. Gultyaev, K. J. Oistämö, K. Gerdes, and C. W. A. Pleij, "A pH-jump approach for investigating secondary structure refolding kinetics in RNA," *Nucleic Acids Res*, vol. 30, p. e63, Jul 2002.

[72] S. Harlepp, T. Marchal, J. Robert, J.-F. Léger, A. Xayaphoummine, H. Isambert, and D. Chatenay, "Probing complex RNA structures by mechanical force," *Eur Phys J E Soft Matter*, vol. 12, pp. 605–15, Dec 2003.

[73] J. H. A. Nagel, C. Flamm, I. L. Hofacker, K. Franke, M. H. de Smit, P. Schuster, and C. W. A. Pleij, "Structural parameters affecting the kinetics of RNA hairpin formation," *Nucleic Acids Res*, vol. 34, no. 12, pp. 3568–76, 2006.

[74] B. Lewicki, T. Margus, J. Remme, and K. Nierhaus, "Coupling of rRNA transcription and ribosomal assembly in vivo: Formation of active ribosomal subunits in Escherichia coli requires transcription of rRNA genes by host RNA polymerase which cannot be replaced by Bacteriophage T7 RNA Polymerase," *J Mol Biol*, vol. 231, no. 3, pp. 581–593, 1993.

[75] D. Herschlag, "RNA chaperones and the RNA folding problem," *J Biol Chem*, vol. 270, pp. 20871–4, Sep 1995.

[76] R. Russell, "RNA misfolding and the action of chaperones," *Front Biosci*, vol. 13, pp. 1–20, 2008.

[77] W. C. Winkler, A. Nahvi, A. Roth, J. A. Collins, and R. R. Breaker, "Control of gene expression by a natural metabolite-responsive ribozyme," *Nature*, vol. 428, pp. 281–6, Mar 2004.

[78] A. M. Pyle, "Metal ions in the structure and function of RNA," *J Biol Inorg Chem*, vol. 7, pp. 679–90, Sep 2002.

[79] I. López de Silanes, M. Zhan, A. Lal, X. Yang, and M. Gorospe, "Identification of a target RNA motif for RNA-binding protein HuR," *Proc Natl Acad Sci U S A*, vol. 101, pp. 2987–92, Mar 2004.

[80] D. Graur and W. Martin, "Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision," *Trends Genet*, vol. 20, pp. 80–6, Feb 2004.

[81] K. Gerdes, A. P. Gultyaev, T. Franch, K. Pedersen, and N. D. Mikkelsen, "Antisense RNA-regulated programmed cell death," *Annu Rev Genet*, vol. 31, pp. 1–31, 1997.

[82] J. L. Thorne, H. Kishino, and J. Felsenstein, "An evolutionary model for maximum likelihood alignment of DNA sequences," *J Mol Evol*, vol. 33, pp. 114–24, Aug 1991.

[83] I. Holmes, "A probabilistic model for the evolution of rna structure," *BMC Bioinf*, vol. 5, p. 166, Oct 2004.

[84] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach.," *J Mol Evol*, vol. 17, no. 6, pp. 368–376, 1981.

[85] S. Washietl and I. L. Hofacker, "Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics," *J Mol Biol*, vol. 342, pp. 19–30, Sep 2004.

[86] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: A novel method for fast and accurate multiple sequence alignment," *J Mol Biol*, vol. 302, pp. 205–17, Sep 2000.

[87] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Curr Opin Struct Biol*, vol. 16, pp. 368–73, Jun 2006.

[88] S. Washietl, I. L. Hofacker, and P. F. Stadler, "Fast and reliable prediction of noncoding RNAs," *Proc Natl Acad Sci U S A*, vol. 102, pp. 2454–2459, Feb 2005.

[89] C. Workman and A. Krogh, "No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution," *Nucleic Acids Res*, vol. 27, pp. 4816–4822, Dec 1999.

[90] P. Clote, F. Ferre, E. Kranakis, and D. Krizanc, "Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency.," *RNA*, vol. 11, pp. 578–591, May 2005.

[91] P. Anandam, E. Torarinsson, and W. L. Ruzzo, "Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies," *Bioinformatics*, vol. 25, pp. 668–669, Mar 2009.

[92] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: an RNA family database," *Nucleic Acids Res*, vol. 31, pp. 439–41, Jan 2003.

[93] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: annotating non-coding RNAs in complete genomes," *Nucleic Acids Res*, vol. 33, pp. D121–4, Jan 2005.

[94] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman, "Rfam: updates to the RNA families database.," *Nucleic Acids Res*, vol. 37, pp. D136–40, Jan 2009.

[95] K. Gerdes, J. E. Larsen, and S. Molin, "Stable inheritance of plasmid R1 requires two different loci," *J Bacteriol*, vol. 161, pp. 292–8, Jan 1985.

[96] K. Gerdes and E. G. H. Wagner, "RNA antitoxins," *Curr Opin Microbiol*, vol. 10, pp. 117–124, Apr. 2007.

[97] A. P. Gultyaev, T. Franch, and K. Gerdes, "Programmed cell death by hok/sok of plasmid R1: coupled nucleotide covariations reveal a phylogenetically conserved folding pathway in the hok family of RNAs," *J Mol Biol*, vol. 273, pp. 26–37, Oct 1997.

[98] C. Yanofsky, "Transcription attenuation: once viewed as a novel regulatory strategy," *J Bacteriol*, vol. 182, pp. 1–8, Jan 2000.

[99] P. Gollnick, P. Babitzke, A. Antson, and C. Yanofsky, "Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*," *Annu Rev Genet*, vol. 39, pp. 47–68, 2005.

[100] A. V. Seliverstov, H. Putzer, M. S. Gelfand, and V. A. Lyubetsky, "Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria," *BMC Microbiol*, vol. 5, p. 54, 2005.

[101] S. R. Eddy and R. Durbin, "RNA sequence analysis using covariance models," *Nucleic Acids Res*, vol. 22, pp. 2079–88, Jun 1994.

[102] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy, "Infernal 1.0: inference of RNA alignments," *Bioinformatics*, vol. 25, pp. 1335–7, May 2009.

[103] G. Cochrane, R. Akhtar, P. Aldebert, N. Althorpe, A. Baldwin, K. Bates, S. Bhattacharyya, J. Bonfield, L. Bower, P. Browne, M. Castro, T. Cox, F. Demiralp, R. Eberhardt, N. Faruque, G. Hoad, M. Jang, T. Kulikova, A. Labarga, R. Leinonen, S. Leonard, Q. Lin, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, S. Plaister, S. Robinson, S. Sobhany, R. Vaughan, D. Wu, W. Zhu, R. Apweiler, T. Hubbard, and E. Birney, "Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database," *Nucleic Acids Res*, vol. 36, pp. D5–12, Jan 2008.

[104] P. P. Gardner. Personal communication, Dec 2008.

[105] J. Sullivan, Z. Abdo, P. Joyce, and D. L. Swofford, "Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation," *Mol Biol Evol*, vol. 22, pp. 1386–92, Jun 2005.

[106] F. S. Brinkman, *Bioinformatics: a practical guide to the analysis of genes and proteins*, ch. 14: Phylogenetic analysis, pp. 365 – 392. John Wiley and sons, third ed., 2005.

[107] D. Blanco and R. Guigó, *Bioinformatics: a practical guide to the analysis of genes and proteins*, ch. 6: Predictive Methods Using DNA, pp. 115 – 142. John Wiley and sons, third ed., 2005.

[108] C. S. Hayes and K. C. Keiler, "Beyond ribosome rescue: tmRNA and co-translational processes," *FEBS Lett*, vol. 584, pp. 413–9, Jan 2010.

[109] T. M. Henkin, "Riboswitch RNAs: using RNA to sense cellular metabolism," *Genes Dev*, vol. 22, pp. 3383–90, Dec 2008.

[110] J. Stoye, D. Evers, and F. Meyer, "Rose: generating sequence families," *Bioinformatics*, vol. 14, no. 2, pp. 157–63, 1998.

[111] M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon, "RNA STRAND: the RNA secondary structure and statistical analysis database," *BMC Bioinf*, vol. 9, p. 340, 2008.

[112] J. S. Pedersen, I. M. Meyer, R. Forsberg, P. Simmonds, and J. Hein, "A comparative method for finding and folding RNA secondary structures within protein-coding regions," *Nucleic Acids Res*, vol. 32, no. 16, pp. 4925–36, 2004.

[113] J. S. Pedersen, R. Forsberg, I. M. Meyer, and J. Hein, "An evolutionary model for protein-coding regions with conserved RNA structure," *Mol Biol Evol*, vol. 21, pp. 1913–22, Oct 2004.

[114] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer New York:, 2001.

[115] J. Kleinberg and E. Tardos, "Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields," *JACM*, vol. 49, no. 5, pp. 616–639, 2002.

# Appendix A

# RFAM quality control measures

The mean fraction of canonical base-pairs, $FC$ is defined as:

$$FC := \frac{\sum_{a=1}^{M} \left( (\sum_{S_{ij}} \Pi_{ij}^{a}) / |S_{ij}| \right)}{M}$$

The covariation statistic, $C$, is defined as:

$$C := \frac{\sum_{a=1,b=1,a<b}^{M} \left( \sum_{S_{ij}} (\Pi_{ij}^{ab} H(a_i a_j, b_i b_j) - \Omega_{ij}^{ab} H(a_i a_j, b_i b_j)) \right) / (|S_{ij}|)}{\binom{M}{2}}$$

- $S_{ij}$ is the set of base-pairs $i$ and $j$ in the consensus secondary structure.
- $M$ is the number of sequences in the alignment.
- $H(a_i a_j, b_i b_j)$ is the Hamming distance between the strings $a_i a_j$ and $b_i b_j$.
- $\Pi_{ij}^{ab}$ is an indicator function such that if $a_i$ and $a_j$ can form a canonical base-pair, and $b_i$ and $b_j$ can also form a canonical base-pair, $\Pi_{ij}^{ab} = 1$ (otherwise $\Pi_{ij}^{ab} = 0$).
- $\Omega_{ij}^{ab}$ is an indicator function such that if $a_i$ and $a_j$ and/or $b_i$ and $b_j$ cannot for a canonical base-pair, $\Omega_{ij}^{ab} = 1$ (otherwise $\Omega_{ij}^{ab} = 0$).

Adapted from [104]. These definitions can be applied to helices by considering $S_{ij}$ as just the base-pairs of the helix.