RNA SECONDARY STRUCTURE PREDICTION USING A COMBINED METHOD OF THERMODYNAMICS AND KINETICS

A Thesis
Presented to
The Academic Faculty

by

Minmin Pan

In Partial Fulfillment
of the Requirements for the Degree
Master in the
School of Biology

Georgia Institute of Technology
August 2011

RNA SECONDARY STRUCTURE PREDICTION USING A COMBINED METHOD OF THERMODYNAMICS AND KINETICS

Approved by:

Dr. Stephen Harvey, Advisor School of Biology *Georgia Institute of Technology*	Dr. Christine Heitsch School of Mathematics *Georgia Institute of Technology*
Dr. Roger Wartell School of Biology *Georgia Institute of Technology*	Dr. Nicholas Hud Chemistry and Biochemistry *Georgia Institute of Technology*
Dr. Joshua Weitz School of Biology *Georgia Institute of Technology*	

Date Approved: June 29, 2011

ACKNOWLEDGEMENTS

I wish to thank Dr. Stephen Harvey, my advisor, through the past few years, for supervising and supporting my research all the time. I would like to thank my committee members for advice and dedication to my thesis. I wish to thank the members in Harvery’s lab for help and discussion with my thesis fulfillment. I would also like to thank my family for supporting me all the time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MFE	Minimum Free Energy
NNM	Nearest Neighbor Model
L-R	long range

SUMMARY

Nowadays, RNA is extensively acknowledged an important role in the functions of information transfer, structural components, gene regulation and etc. The secondary structure of RNA becomes a key to understand structure-function relationship. Computational prediction of RNA secondary structure does not only provide possible structures, but also elucidates the mechanism of RNA folding. Conventional prediction programs are either derived from evolutionary perspective, or aimed to achieve minimum free energy. *In vivo*, RNA folds during transcription, which indicates that native RNA structure is a result from both thermodynamics and kinetics.

In this thesis, I first reviewed the current leading kinetic folding programs and demonstrate that these programs are not able to predict secondary structure accurately. Upon that, I proposed a new sequential folding program called GTkinetics. Given an RNA sequence, GTkinetics predicts a secondary structure and a series of RNA folding trajectories. It treats the RNA as a growing chain, and adds stable local structures sequentially. It is featured with a Z-score to evaluate stability of local structures, which is able to locate native local structures with high confidence. Since all stable local structures are captured in GTkinetics, it results in some false positives, which prevents the native structure to form as the chain grows. This suggests a refolding model to melt the false positive hairpins, probable intermediate structures, and to fold the RNA into a new structure with reliable long-range helices. By analyzing suboptimal ensemble along the folding pathway, I suggested a refolding mechanism, with which refolding can be evaluated whether or not to take place.

Another way to favor local structures over long-distance structures, we introduced a distance penalty function into the free energy calculation. I used a sigmoidal function to compute the energy penalty according to the distance in the primary sequence between two nucleotides of a base pair. For both the training dataset and the test dataset, the distance function improves the prediction to some extent.

In order to characterize the differences between local and long-range helices, I carried out analysis of standardized local nucleotide composition and base pair composition according to the two groups. The results show that adenine accumulates on the 5' side of local structure, but not on that of long-range helices. GU base pairs occur significantly more frequent in the local helices than that in the long-range helices. These indicate that the mechanisms to form local and long range helices are different, which is encoded in the sequence itself.

Based on all the results, I will draw conclusions and suggest future directions to enhance the current sequential folding program.

CHAPTER 1

INTRODUCTION

Increasingly, RNA molecules have taken center stage as key informational, structural, catalytic, and gene-regulatory molecules. Contrary to the traditional view of the passive genetic information carrier, RNAs have now been shown to be involved in a wide range of biological processes. Ribosomal RNAs catalyze and regulate protein synthesis(1,2). The crystal structures of 70S ribosome support the view that the fundamental steps of translation are based on RNA–RNA interactions. In the nucleus, small nuclear RNAs bound to proteins catalyze and regulate pre-mRNA splicing(3). Bacterial riboswitches influence transcription or translation by changing the conformations upon directly sensing metabolites or other environmental cues(4,5).

All functional RNA molecules depend on their structures to perform their respective functions, but only a few RNA structures are known. Of the structures deposited in the Protein Data Bank, proteins constitute 95%, whereas RNAs account for less than 2%(6). Nevertheless, analysis of the human genome data reveals that only about 1.5 % of the genome encodes proteins; about 60-70% of the genome produces non-coding RNA(7). The lack of RNA structures indicates that it is hard to obtain structural information through experiments(6). Nowadays, with the development of RNA/DNA sequencing, reverse transcription is a routine method to obtain RNA primary sequence. However, it is still difficult to predict RNA secondary structure.

RNA secondary structure is the base pair information within the RNA, including Watson-Crick base pairs, wobble base pairs and non-canonical base pairs. RNA secondary structures established themselves as the most significant level of description because they have a physical meaning as folding intermediates of RNA 3D structures(8,9), and are accessible to mathematical analysis, since their formation follows simple combinatorial rules. This calls for computational methods to solve the RNA secondary structure prediction problem.

1.1 Conventional RNA secondary structure predictions

Comparative analysis is considered to be the gold standard in RNA secondary structure determination. The fundamental premise of this approach is that functionally equivalent molecules will exhibit the same secondary structure in spite of variations in sequence, i.e., the two bases in a base pair should “co-vary” to preserve the base pair structure. From an evolutionary point of view, in order to keep the function of the RNA molecule, the RNA sequences have evolved to preserve certain base pairs and single stranded regions. For example, if a guanine in a G-C base pair mutates to A, in order to keep the base pair, the corresponding position of C will mutate to U.

In the 1975 paper on 5S rRNA(10), Fox and Woese stated the “covariation model” systematically. The secondary structures of 16S rRNA(11,12) and 23S rRNA(13) predicted by comparative analysis in the 1980s are the classic models in this field. These two models were substantially improved as more and more homologous sequences became available(14). Gutell *et al*. evaluated the accuracy of the comparative structure models(14) right after the high resolution crystal structures were determined(15,16). Because of the fundamental phylogenetic perspective and using thousands of homologous sequences, comparative analysis greatly succeeded in the ribosomal RNA secondary structure prediction. All of the base pairs in the 1999 16S and 23S secondary structure models, including tertiary base pairs and triplets were found in the crystal structures (14).

However, because the approach requires thousands of homologous sequences to ensure high accuracy, comparative analysis is not able to provide high quality models with few sequences. The first models of 16S rRNA (11,12) and 23S rRNA (13) which were based on about two sequences contained only 59.4% and 77.7% correct base pairs respectively, far away from the 1999 model. Moreover, comparative analysis searches for co-variation, which requires sufficient variance among the sequences. No base pair information can be deduced in the highly conserved regions, since without variation, there can be no covariation in the sequence. Comparative analysis is not useful in the case that only one sequence of its homolog species is available, or the structures of the homologous sequences are not similar.

Thermodynamic methods were first introduced(17) to maximize the number of base pairs. In this paper, Holley *et al*. suggested 3 possible secondary structures for tRNAPhe, including the cloverleaf structure(17). When the sequences of tRNASer(18) and tRNATyr(20) became available, it was clear that the cloverleaf model was consistent with all three sequences. These studies focused on finding longest possible helices subject to the constraints that the anti-codon stems must be in a single-stranded region. Now they have become the most common method for predicting secondary structure when only a single RNA sequence is known for a given function. The assumption of this method is that the RNA molecule is at equilibrium in solution so that the minimum free energy (MFE) structure is the most probable structure. The prediction of the MFE structure using nearest neighbor energy parameters began with Tinoco and colleagues

(18,19). Zuker *et al.* developed efficient algorithms for RNA secondary structure prediction (20-23) using dynamic programming methods(24). The Mfold web server and the UNAfold program(25) by Zuker's group have become some of the most popular MFE structure prediction programs. There are several other closely related programs, such as RNAfold in the Vienna Package (26,27), and RNAstructure by Matthews(28). GTfold developed by Mathuriya *et al.* (29)is a new parallel multicore and scalable program, which is one to two orders of magnitude faster than the standard programs (O(n4)) and achieves comparable accuracy of prediction. GTfold opens up a new path for the computation of MFE for large RNAs. It is available at http://gtfold.sourceforge.net/download.html.

The advantages of this approach are that it is physics-based, it gives one simple output, and it is computationally inexpensive. However, the prediction accuracy for large RNAs is not satisfying (30,31). The average prediction accuracy of Mfold 3.1 for a 16S or 23S rRNA sequence is only about 40% (30,31). One of the reasons of the limitation of the MFE method is that the thermodynamic rules are incomplete. Turner's group has been working on determining the energy parameters for over twenty years (32-34); and has recently included the stabilizing effects of different RNA motifs beyond base-paring and stacking(35-37). Nevertheless the better thermodynamic rules and parameters are not able to improve the prediction accuracy fundamentally. Mfold 3.1 offers little improvement over Mfold 2.3 for rRNAs, although it has much better parameters (31). This suggests a more fundamental reason for the limited accuracy, which is that RNA molecules adopt secondary structures that are at least partially determined by folding kinetics. This may explain why the nearest-neighbor energy parameters do work well for shorter RNA sequences such as tRNA or 5S rRNA, or for larger rRNAs when the contact distance between the base pairs is less than 100 nucleotides(31). RNA folding kinetics will be elaborated in section 1.2.

McCaskill argued that the overall structural features of the equilibrium ensemble should be more precise than the single MFE structure (38). He applied dynamic programming to calculate the full equilibrium partition function (PF) for secondary structure and featured base pair probability as a measure of confidence for the final prediction(38). Ding *et al.* utilized a statistical algorithm to sample rigorously and exactly from the Boltzmann ensemble of secondary structures(39). This provides a means to estimate the probability of any structural motif, with or without constraints. This approach provides sub-structure probability in order to deduce a higher quality secondary structure model for a given sequence. Although the base pairs predicted are very reliable, PF cannot give a complete structure. A moderate portion of structure information is missing, since some regions do not have a high base-pair or single-stranded probability. Furthermore PF is also a thermodynamic method based on the same energy model and using similar energy parameters as the MFE method, still assuming that RNA molecules are at equilibrium. PF methods neglect kinetic effects in the RNA folding process, as do the MFE methods.

The equilibrium view of RNA folding can be misleading: the time needed to reach equilibrium can be very long, perhaps even exceeding the lifetime of RNA molecule. The minimum free energy structures do not always correspond to the native structure. Mahen *et al.* (40)showed that the renatured whole length hairpin ribozyme displayed a different secondary structure from the one transcribed *in vivo*. For the self-induced riboswitches, the metastable structure is actually the functional structure(5,41). After the riboswitches gradually unfold and refold to the lowest free energy structure, they lose their functions. The formation of metastable structure during the sequential folding of potato spindle tuber viroid (-)-stranded RNA is essential for template activity during (+)-strand synthesis(41). Theoretical research also suggests that the native structures of large RNAs deviate from the MFE structures(42). By using the current thermodynamic RNA prediction program, Morgan *et al.* (42) found that the energy of native RNA structures are generally higher than the energy of the predicted MFE structures. Clearly, a purely thermodynamic model poorly predicts secondary structures of large RNAs.

1.2 RNA folding during transcription

RNAs begin to fold as they are transcribed in the cell. Folding during transcription can be modulated by properties of the RNA polymerase and the in vivo environment. Three particular properties are relevant to RNA folding during transcription(43): (1) elongation speed (41,44-46), (2) site specific pausing of RNA polymerase (47), and (3) the co-transcriptional interaction of the nascent RNA with proteins or small molecules(44). Transcription of the large ribosomal RNA of Escherichia coli by T7 RNA polymerase, which is significantly faster than the E. coli RNA polymerase, generates a folding defect in vivo(46). A non-cognate polymerase may cause RNA misfolding due to a different transcription rate or pausing sites, which suggests that kinetic effects are involved in RNA folding. A particular elongation speed and site-specific pausing provide RNA molecules with a crucial time window of intramolecular interaction for folding/refolding to the native structures.

Intramolecular interactions form within the upstream part of the RNA before the downstream part is synthesized(48,49). This has several implications for the assembly of RNA structure. First, sequential folding during transcription is expected to favor local structure over long-range interactions that require synthesis of a longer RNA chain. Second, the context of the nascent RNA is very important to its folding. Several circular permutation results show that, given a different 5'end and 3' end, the RNA molecules can misfold (50,51). In other words, co-transcriptional folding is encoded within RNA genes(52). Third, the intermediate structures formed during transcription are not static. Comparison of RNA fragments of increasing length from chain termination reactions(49), or direct observation of RNA conformers on acrylamide gels(41), found that interactions formed early during transcription can be displaced later by more thermodynamically stable interactions. The stability of the folding intermediates determines the time required to rearrange to the final structure, and hence the overall folding time. These intermediate structures can efficiently regulate or guide the folding of nascent RNA molecules into native structures.

1.3 kinetic folding programs

The characteristic of RNA sequential folding in transcription suggests a need for kinetic folding algorithms to directly model the physical folding process. These approaches are based on a description of folding in terms of a stochastic process. In general, any such model is defined by three key ingredients (53): (1) the state space, comprising the set of structures or conformations a given RNA sequence may assume, (2) a move-set, defining the elementary transitions that can occur between such conformations, (3) transition rates for each of these allowed transitions. Due to the limited kinetic experimental data of RNA folding and refolding, the calculation of transition rate becomes the hardest part of the kinetic prediction approach.

RNAkinetics

Mironov *et al.* (54) presented the formation of RNA secondary structure as a Markov process. A Markov process is one for which the likelihood of a given future state, at any given moment, depends only on the present state and the transition rates to all possible future states. The states here are the secondary structures; a transition is formation of a new helix, complete decay of an existing helix, or addition of new nucleotides. Transition rates are calculated as the kinetic rate in a chemical reaction. The number of possible transition states is equal to the number of stacking pairs in the helix, assuming that every base pair has the same probability to nucleate a helix. The transition rates of formation and dissociation of a helix are described as,

$$k^{eff}_{form} = \kappa_c \cdot N_h \cdot \exp\left(\frac{-\Delta G_{loop}}{kT}\right), \quad k^{eff}_{dis} = \kappa_c \cdot N_h \cdot \exp\left(\frac{\Delta G_{helix}}{kT}\right)$$

N_h is the number of stacking pairs in the helix. κc is an elementary event constant, determined in temperature-jump experiments(55). ΔG_{loop} is the free energy to form a loop. ΔG_{helix} is the free energy to melt a helix. Based on the computable transition rates, RNAkinetics uses kinetic Monte Carlo to simulate RNA folding. At each step in the simulation, all possible transitions are generated. The next transition is chosen according to a set of probabilities that are determined by the transition rates. Simulation time is increased by the first pass time at that step. Later on, the algorithm was refined by adding a classification of mutual positions of two candidate helices(56,57).

The problem with this program is mainly the transition model. The assumption that every base pair in the helix has the same probability to form a helix or to unfold a helix is not reasonable. Each base pair closes a different size of loop, so that the entropy to form or unfold that base pair is different. The transition rates are unrealistically simplified.

Kinefold

Isambert *et al.* (58)divided the energy barrier into two parts. One is the energy needed to destroy an existing helical region to nucleate a new helix; the other is the entropy cost to bring two parts of the new helical region together. The first part is obtained from the MFE model. They used polymer theory to calculate the entropy terms. By modeling the single-stranded sections as Gaussian chains, they analytically obtained the entropy costs for forming different structures, including pseudoknots. They also employed kinetic Monte Carlo to simulate RNA folding.

Kinefold is able to predict pseudoknots in the RNA secondary structure, without additional parameters. This physical modeling of pseudoknots is also expected to be more widely applicable than previously proposed estimations, as it explicitly takes into account important physical-structural constraints of the RNA molecule(58). However, there is no sequence-dependent term in the entropy part.

One of the advantages of Kinefold is that pseudoknots can be predicted. The stochastic approach is a good way to simulate RNA folding in the sense that RNAs may fold by different pathways (59). However, it does not generate the same structures every time. In order to get a better prediction, users are encouraged to do several independent foldings. At present, there is no well-established way to analyze the results. Moreover, this program is always computationally expensive, since it samples a lot of possible structures at each step. Its capacity is only 400 nucleotides for the web server, or 600 nucleotides for stand-alone program.

Kinwalker

Geis *et al.* (60) employed a heuristic approach to add structural blocks sequentially onto the RNA. One MFE structure is generated for each length of the RNA chain. Every time RNA chain elongates by one nucleotide (N to N+1). The structure blocks in the MFE structure of N+1 length but not in that of N nucleotides will be evaluated whether or not to add on. This restrains the size of the conformational space. The transition of two structures are at base-pair resolution using transition paths based on the Morgan-Higgs method(61), which tries to find a direct folding path from one secondary structure to a target secondary structure where the maximum barrier height along the path is minimal. In order to find such a path, the heuristic method iteratively adds base pairs from the set of base pairs in the target structures that are not included in the current structure. The authors empirically derived a time-energy function from experiments (62) to evaluate the height of the energy barrier that can be overcome in a certain time window, $t(\Delta G) = 10(8/11\Delta G-7)$, for $\Delta G > 0$. However, the energy barriers are the highest equilibrium free energy in all transition steps, which is not the real kinetic barrier height.

Compared to stochastic simulations, the heuristic approach displays some advantages: it always gives the same secondary structure and it is computationally inexpensive. However this prediction program tends to achieve the MFE structure to a great extent, because the target structures of a transition are always MFE structures or some rearrangement of the MFE substructures. The results from this program (62) shows that this program does not identify kinetic traps very well, since it accepts almost all transition along the folding pathway.

FlexStem

Chen *et al.* developed a heuristic kinetic RNA folding algorithm with pseudoknots(63,64). Flexstem simulates the RNA folding process by successive addition of maximal stems, which are the longest helical stems to reach, in order to reduce the search space. After all the candidate stems are constructed, they are arranged in an order, which is defined as the rank of ability to decrease the free energy of the current structure. At each step, the current structure is "perturbed" by adding candidate stems in this order. However, there is no energy barrier calculation in FlexStem. As long as the free energy of the potential structure is lower than that of the current one, the potential structure will be accepted. This reduced space is constructed by the maximal stem strategy and stem-adding rules induced from elaborate statistical experiments on real RNA secondary structures. Chen *et al.* developed a novel free energy model capable of computing free energy of pseudoknots. They validated FlexStem by testing it on tRNAs, 5S rRNA and a large number of pseudoknotted structures. They also compared it with other well-known algorithms such as RNAfold, PKNOTS(65), PKnotsRG (66), and ILM (67). FlexStem significantly increases the prediction accuracy through its local search strategy.

I summarize the transition function in each program in Table 1.1 for comparison.

Other kinetic folding programs

There are several other kinetic folding programs (53). They are similar in principle, but differ in the move-set (a helix, or a base pair) and the transition rate model.

Most of them develop equations to compute transition rates. However, given the sparsity of kinetic data compared to the abundance of free energy parameters, the addition of refolding steps has not yet produced an overall improvement in prediction accuracy. An interesting alternative is the analysis of energy landscapes(68). The resulting barrier trees provide a convenient summary of possible folding scenarios without the need to sample trajectories from different initial states. In addition, barrier trees form the basis for a coarse graining such that the folding dynamics can be solved exactly in the reduced conformation space. However, the barrier tree alone cannot provide a folding pathway and it starts from the whole length sequence, so it does not mimic folding during transcription.

Fortunately, the renewed interest in RNA as a versatile biomolecule has also inspired diverse experimental approaches to measure folding kinetics in detail, ranging from classical temperature jump experiments (62,69)to time-resolved NMR spectroscopy (70,71)and single molecule methods(72). With the development of experimental techniques, more and more kinetic data will advance the RNA folding prediction, just as more accurate thermodynamic parameters improve the quality of the MFE computations. At present, since the kinetic data is limited, an empirical but realistic RNA folding algorithm is needed to suggest some underlying mechanisms of RNA folding and refolding.

Table 1.1 Summary of three kinetic RNA folding programs.

program	Transition Rate	Energy barrier	Target structure
RNAkinetics	$k^{eff}_{form} = \kappa_c \cdot N_h \cdot \exp\left(\frac{-\Delta G_{loop}}{kT}\right)$	$\Delta G_{\text{loop}} = T\Delta S$	Chosen at random with a probability proportional to its rate
Kinefold	$k_{+} = k^{\circ} * \exp(-\Delta G_{+}/kT)$	$\Delta G_{+} = \Delta G_{\text{free-up-existing-helix}}$ $+T\Delta S_{\text{bring-two-strand-new-helix}}$	Chosen at random with a probability proportional to its rate
Kinwalker	$t(\Delta G) = 10^{(\frac{8}{11}\Delta G)-7}$	Saddle point, $t(\Delta G_{\text{max}})$, a next transcription event can surpass	An MFE structure or rearrangement of some MFE sub-structures

1.4 Accuracy of the leading kinetic folding programs

The performance of the four leading kinetic programs on tRNA, 5S rRNA, RNase P and the 5' domain of 16S rRNA are shown in Table 1.2. The accuracy is measured in two ways, sensitivity and specificity. Sensitivity is the percentage of true positives in the real structure, while specificity is the percentage of true positives in the predicted model.

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \times 100\%$$

$$Specificity = \frac{TruePositives}{TurePositives + FalsePositives} \times 100\%$$

Each program was tested with its default settings. In order to only compare pseudoknot-free structures I turned off the 'pseudoknot' function in Kinefold and Flexstem.

UNAfold and RNAfold are MFE RNA secondary structure prediction programs, serving here as controls for comparison. For tRNA and 5S rRNA, the kinetic folding programs work better than MFE programs overall, while for the larger RNAs, their results are worse than those MFE programs. This is unexpected, since the kinetic folding programs were developed to resolve the mistreatment of kinetic traps in the folding of large RNAs by MFE programs. The results indicate that the transition functions in the kinetic programs, which were established to cope with refolding, are not very realistic. The theories behind these transition functions are not able to correctly explain the behavior of RNA chains. One of the problems in the theories is the inaccurate estimation of the heights of energy barriers. These programs all apply some oligonucleotide experimental data in the empirical equations; however these data may be not suitable for a complex system, since there are many more interactions in larger RNAs than in smaller ones. While these kinetic folding programs pioneered a kinetics-driven track in the RNA folding problem, they are far from elucidating real RNA folding. A more realistic kinetic folding program is still required. The remainder of this thesis describes my efforts at developing an empirical folding algorithm that combines thermodynamic and kinetic effects.

1.5 Other interactions

Besides the nearest-neighbor interactions, which correspond to the proximity along the sequence, RNAs also have intramolecular tertiary interactions. Pseudo-knots, base triplets, and GNRA tetraloop interactions are all in this category. These interactions play a very important role in stabilizing RNA structure. For example, in tRNAs, the pseudoknots between the D-loop and the T-loop make tRNAs adopt a unique pattern of folding (73). van Batenburg *et al*. have established a database containing over 250 pseudoknots obtained in the past 25 years through crystallography, NMR, mutational experiments and sequence comparisons(74,75).

Nevertheless, several observations suggest that 3D architecture results from the compaction of separate preexisting and stable secondary structural elements(8). Such structural elements as hairpins are known to form as autonomous entities that interact with each other later to form the 3D structure. RNA unfolding experiments show that RNAs unfold in a series of discrete steps. The multiplicity of intermediate states represent the breaking down of the folded structure into localized regions of the structure, i.e., tertiary interactions destroyed first, followed by melting of the secondary structure. This suggests that the secondary structure interactions are stronger than tertiary interactions. Most of the time, secondary structures do not change after tertiary interactions; only the weakest of the secondary structure elements may change, leaving the rest largely unaffected(9).

There exist several prediction programs that include pseudo-knots(58,63,65,76). These prediction programs generally require comprehensive computation, O(N^6), which sample all possible interactions. Given this computational complexity, the inclusion of tertiary interactions is not feasible for large RNAs (>1000 nt), so these interactions are excluded from my proposal.

1.6 Overview of this research

In this research, I first introduce a new sequential folding program called GTkinetics. Chapter 2 describes the algorithm, presents results of folding test sequences and discusses the implications of these results. In addition, I suggest a refolding mechanism by analyzing suboptimal ensemble along the folding pathway. In Chapter 3, I evaluate a distance function in MFE calculations, as a possible method, as a possible method to improve accuracy of prediction. In Chapter 4, I examine differences in composition between local and long-range helices, testing a hypothesis about folding mechanisms. In Chapter 5, I conclusions and suggest future directions for RNA secondary structure prediction.

Table 1.2 Sensitivity and Specificity of thermodynamic and kinetic RNA folding programs tested on different RNAs. Sensitivity is the percentage of true positives in the real structure. Specificity is the percentage of true positives in the predicted model. ND, not determined.

			Thermodynamic programs		Kinetic programs			
Molecule	length	accuracy	UNAfold	RNAfold	RNA kinetics	Kinefold	Kinwalker	Flexstem
Yeast tRNA^{phe}	76	sensitivity	30%	20%	95%	68%	30%	95%
		specificity	30%	20%	100%	63%	30%	100%
E.coli 5S rRNA	120	sensitivity	25%	25%	35%	42%	43%	25%
		specificity	27%	26%	82%	50%	50%	26%
E. coli RNase P	377	sensitivity	60%	38%	ND	48%	52%	31%
		specificity	66%	40%	ND	52%	53%	34%
E. coli 5'domain of 16S rRNA	560	sensitivity	80%	69%	ND	37%	68%	65%
		specificity	82%	70%	ND	36%	67%	68%

CHAPTER 2

GTkinetics

2.1 Overview

GTkinetics is an RNA sequential folding program. For a given RNA, it predicts a secondary structure and a series of RNA folding trajectories. It treats the RNA as a growing chain, and adds stable local structures sequentially. At this stage, the time scale of successive events is not determined, but the program proposes a folding pathway and predicts kinetic traps on the folding pathway.

GTkinetics mimics the RNA transcription and folding *in vivo*. When RNAs fold during transcription, the nascent nucleotides will fold into some structures before the whole sequence is transcribed. There are four structural possibilities for nascent nucleotides (Figure 2.1):

(1) becoming an independent hairpin,

(2) extending a pre-existing helix,

(3) pairing with a single stranded region to form a long-range helix,

(4) unfolding a part of an existing structure and folding these with the nascent nucleotides to a new structure.

The first three possibilities are simply adding more base pairs onto the structure without interrupting any existing base pairs, while the last one represents a refolding event. At present, GTkinetics is designed to assign the nascent nucleotides either to adopt possibility (1) or to adopt (2) or to leave them temporarily unassigned, i.e. it can also adopt possibility (3) and (4). In other words, at each step, GTkinetics adds some more nucleotides onto the chain and makes a decision whether these new nucleotides should be used to form a hairpin, or to form an extension, or to leave them free to form any other structures later. This is a starting point, since it does not deal with the transition rates for refolding, which are difficult to establish realistically. In this version, the algorithm captures the stable structures as kinetic traps when the RNA chain gradually elongates.

At each step, there is generally more than one possible hairpin or extension, which can form along the nascent sequence. These possible hairpins and extensions overlaps, so compete with each other; I call a collection of such overlaps a competition cluster. In next section, a competition cluster will be elaborated mathematically. In Section 2.2, I introduced Z-score as a criterion to estimate relative stabilities of competing hairpins. GTkinetics utilizes Z-scores to select optimal candidates and then evalutes them by folding the growing chain; the algorithm is described in Section 2.3. The results of applying GTkinetics to a series of test cases are given in Section 2.4. In the results, I identified possible intermediate structures along a pathway to the native state that includes refolding events (Section 2.5). The chapter closes with a possible evaluation method for refolding events, which is suggested by suboptimal structures along the folding pathway (Section 2.6).

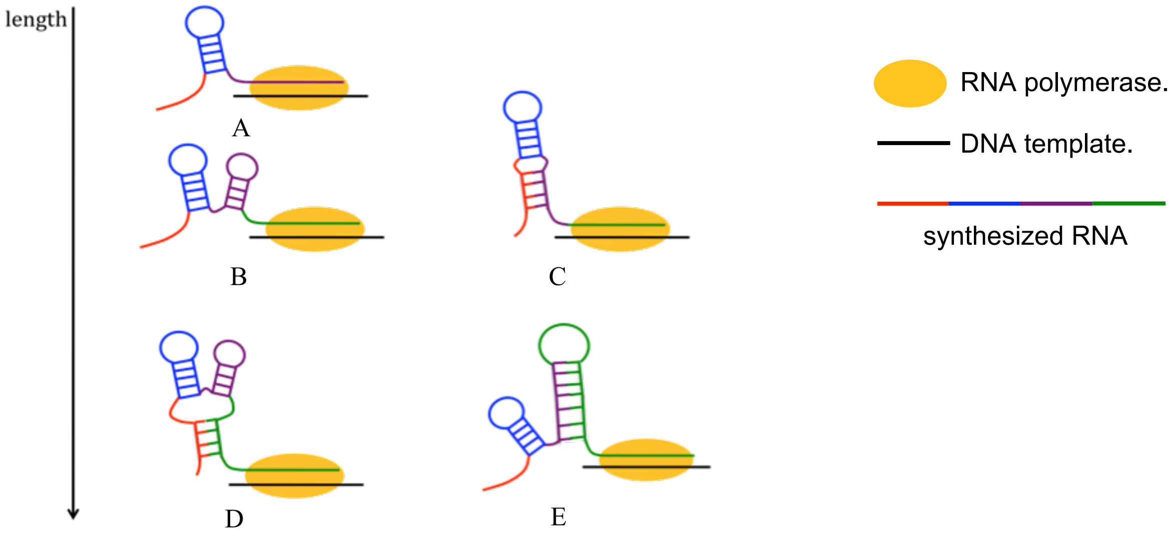

Figure 2.1 Structural possibilities for nascent nucleotides.
(A) a synthesizing RNA
(B) nascent nucleotides (purple) form an independent hairpin
(C) nascent nucleotides (purple) form an extension of a previous hairpin
(D) nascent nucleotides (green) form a long range helix
(E) nascent nucleotides (green) lead to a refolding event

2.2 Z-score

While RNA is transcribed, several potential local hairpins may compete with each other along the sequence. The key to correctly predicting local hairpins is a good scoring function to evaluate the relative stabilities of competing local hairpins. One simple criterion is free energy; however, the free energies of longer RNA segments are generally lower than those of shorter RNAs (Figure 2.2), because longer segments have more bases to pair and to stack.

As a consequence, using free energy for the scoring function biases predictions toward more thermodynamically stable structures, long hairpins at the expense of shorter hairpins, which are favored kinetically. In order to credit kinetic effects with a higher weight, I normalize free energy according to the lengths of segments. Z-score is introduced here to normalize ΔG with ensemble of sequences of a given length.

$$Z = \frac{\Delta G - \mu(w)}{\sigma(w)}.$$

ΔG is the folding free energy of the MFE structure of a given RNA segment; w is the window size, i.e., the sequence length of that segment. Z-score indicates how many standard deviations an observation is above or below the mean μ(w). It is a normalized dimensionless quantity derived by subtracting the population mean μ(w) from the raw score and then dividing the difference by the population standard deviation σ(w). The lower the Z-score, the more stable a helix is, in terms of folding free energy per base pair.

In order to obtain the population mean μ(w) and standard deviation σ(w), which are functions of window length, 20,000 random sequences with equal probability of for nucleotides were generated for each window size (8≤w≤110), followed by MFE calculations with UNAfold. The average value and the standard deviation are plotted against window size in Figures 2.3 and 2.4.

The average MFE generally decreases linearly with window size. The standard deviation, on the contrary, increases linearly with window size. I regressed $\mu(w)$ and $\sigma(w)$ to linear functions of window size,

$\mu(w) = -0.2558w + 3.9684$

$\sigma(w) = 0.0512w + 1.23$.

Combining these, the Z-score for a segment of length w whose energy is ΔG is given by,

$Z = (\Delta G - (-0.2558w + 3.9684))/(0.0512w + 1.23)$.

Figure 2.5A shows the correct secondary structure of tRNAphe, while Figure 2.6 shows all possible hairpins in the molecule, with their free energies. Three red lines represent the three native hairpins in the tRNA. The longest possible hairpin in blue, which has the lowest free energy, is not part of the native structure.

Figure 2.7 shows the Z-scores of all possible helices in the tRNA. The three native helices stand out with this normalization method. If one forces these three stable helices to form when running UNAfold, one obtains an essentially correct cloverleaf structure (Figure 2.5B). This structure only misses two base pairs, one of which is a non-canonical GA pair. Sensitivity and specificity are 95% and 100% respectively, which are significantly higher than those of UNAfold and RNAfold without any constraints (Table 1.2).

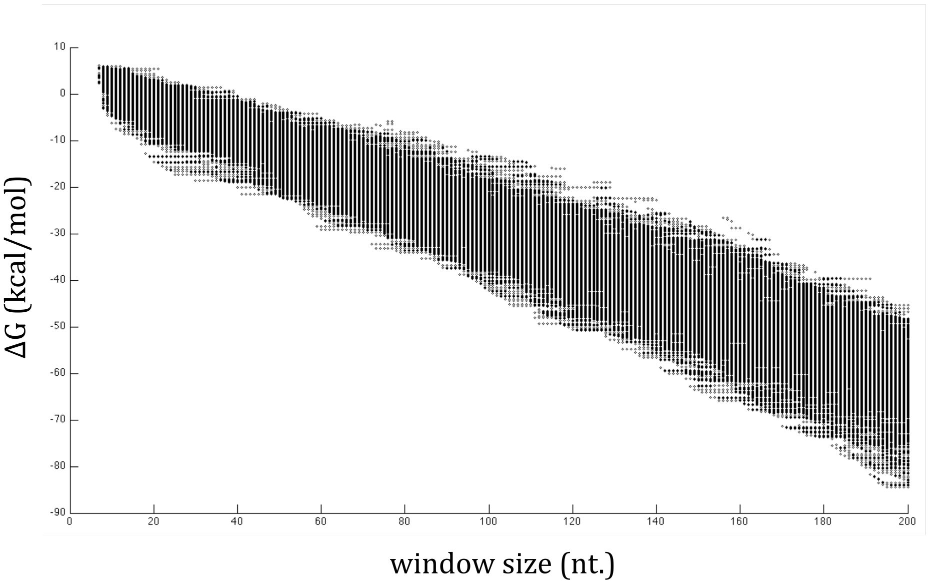

Figure 2.2 MFEs of all sub-sequences with length of 8 to 200 in 16S rRNA as a function of window size. Each dot represents the free energy of the MFE structure for one sub-sequence.

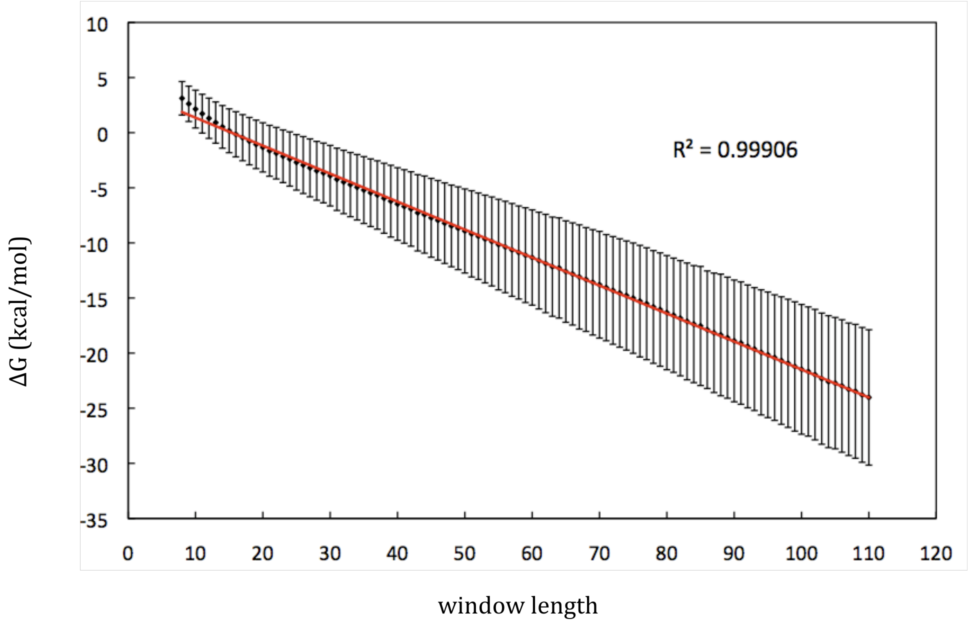

Figure 2.3 The population mean ΔG of each window size plotted against window size. Error bars show the standard deviation. The red line is the linear regression of average ΔG. Correlation coefficient is also shown.

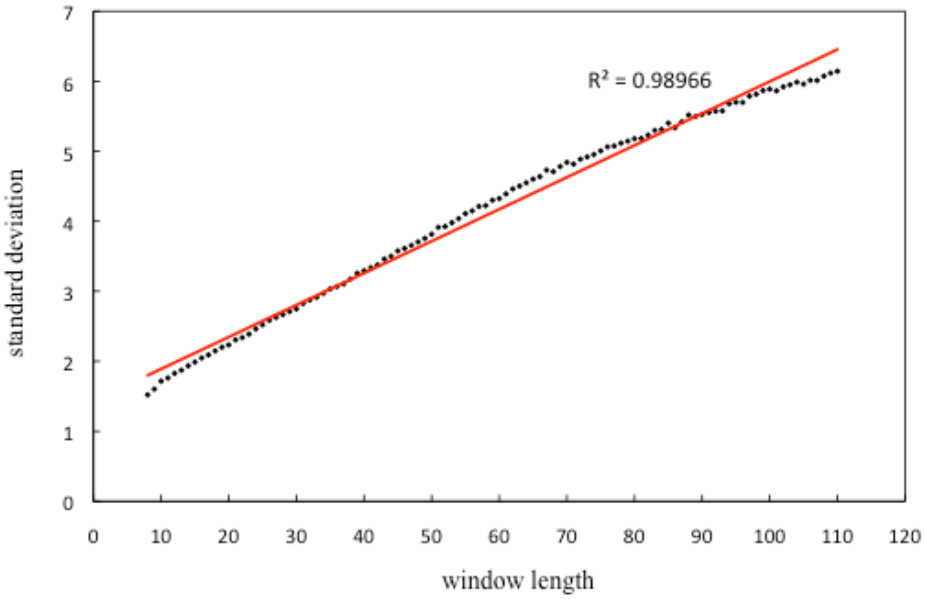

Figure 2.4 The population standard deviations of ΔG plotted against window size. The red line is the linear regression of the standard deviations. Correlation coefficient is also shown.

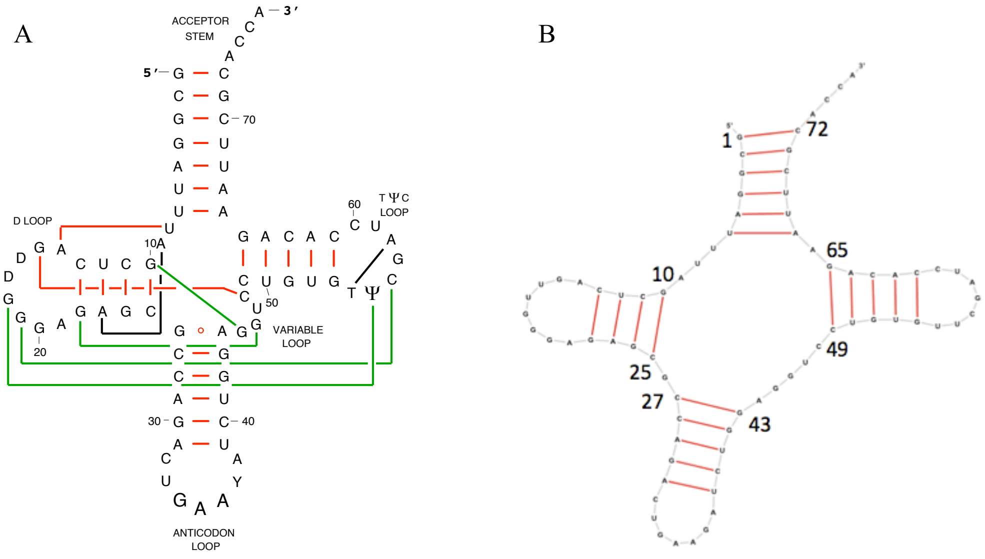

Figure 2.5 Secondary structure of tRNA^{phe} of yeast.
(A) Real structure, cited from Comparative RNA web site, http://www.rna.icmb.utexas.edu**.**
(B) The structure predicted by constraining the three hairpins with lowest Z-score in Figure 2.7.

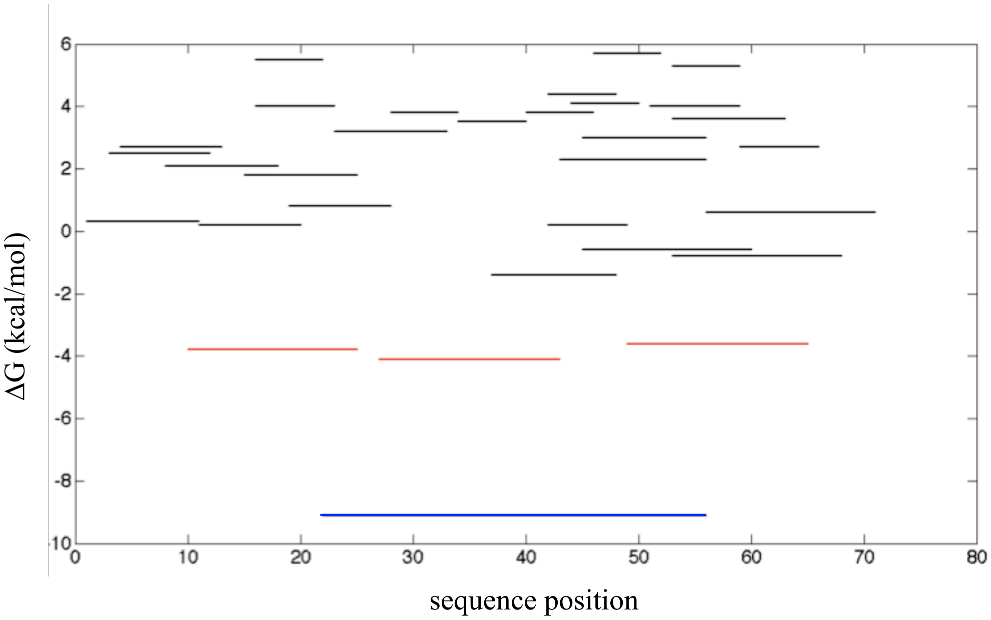

Figure 2.6 Free energy (ΔG) of all possible hairpins along tRNA^{phe}. X-axis is the sequence position that the hairpins occupy. The red lines are three real hairpins (D stem, anticodon stem and T stem respectively); the blue line is the lowest free energy hairpin.

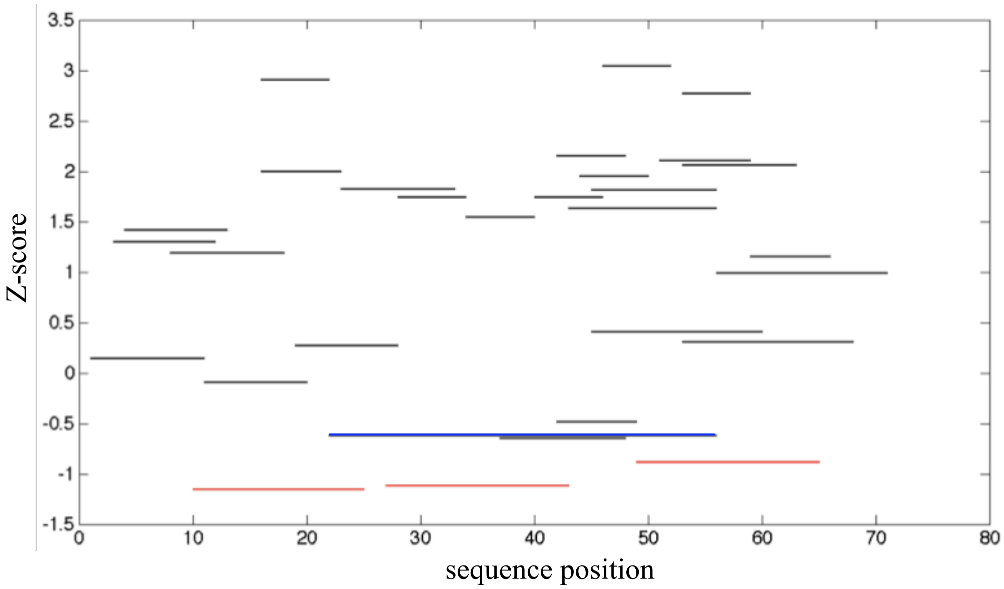

Figure 2.7 The Z-scores of all possible hairpins along tRNA^{phe} sequence. X-axis is the sequence position that the hairpins occupy. Color code is same as in Figure 1.

2.3 Algorithm

The program works in the following steps.

Step 0, a pre-process step, all possible hairpins, ListH and extensions ListEX are generated for selection.

Step 1, initialize folding, transcribed length N is 0; folded structure S set to empty; constraining List C is empty; Z-score of the preceding hairpin, Z_{ex} set to 99999; the 5' boundary of competition cluster, b5 is 0; the 3' boundary of competition cluster, b3 is the 3' end of the first ending hairpin in ListH.

Step 2, generate a competition cluster. Find hairpins whose 5' ends fall in the range of b5 and b3. Put the Z-scores of these hairpins and Z_{ex} into the competition cluster

Step 3, select optimal structure. The helix (m-n) with the lowest Z-score in the competition cluster is selected as a hypothetical helix.

Step 4, elongate the chain and fold the RNA. Then the RNA chain elongates to the 3' end of the hypothetical helix, N= n+1. Fold (1- N) by UNAfold with all helices in List C constrained and obtain new structure S.

Step 5, make decision. If the hypothetical structure survives from the competition with all other possible folding, it becomes a new stable helix. It is then added into List C. If the hypothetical structure does not survive this competition, that piece of RNA will be left free to adopt any conformation.

Step 6, reset the boundaries of competition cluster. b5 set to N+1, the end of the transcribed sequence. If a new constraint has been added in Step 5, then search for the shortest extension of that constraining structure. If the shortest extension (m'-n') is found, b3 set to n’ and Z_{ex} becomes the Z-score of that extension. In all other cases, b3 is the first ending hairpin in List H and Z_{ex} is set to 99999. Go to Step 2.

In this algorithm, local stable helices are favored over other helices; meanwhile, competition between local helices, long-range helices, and single stranded regions are incorporated. These two characteristics are both suggested by the nature of sequential folding. The pseudo-code is shown in Figure 2.8.

```
**Input:** RNA sequence of length N
**Output:** folding trajectories and a secondary structure
Compute $C_{ij}$ for (i, j) with $i<j\leq n$, $7\leq(j-i+1)\leq 100$;
Create List H of all hairpins closed by (i, j), List EX of the all hairpin extensions closed by (i, j)
S $\leftarrow \emptyset$
Constraints $\leftarrow \emptyset$
B5 $\leftarrow$ 0  /* 5'end of a competition cluster */
B3$\leftarrow$ 3'end of first end hairpin starting from B5  /*3'end of a competition cluster */
$Z_{ex} \leftarrow$ 99999 /* z-score of extension initially set to 99999 */
**While** B3 $\leq$ N **do**
    Cluster $\leftarrow$ hairpins with B5 < i $\leq$ B3 /* competition cluster */
    $Z_h \leftarrow \min\{Z(\text{cluster})\}$ /* pick the lowest Z-score */
    **If** $Z_h < Z_{ex}$ **then** /* hairpin is better than extenstion */
        CandidateHelix($i_h$, $j_h$) $\leftarrow$ hairpin ($i_h$, $j_h$) with $Z_h$
        $N_t \leftarrow (j_h+1)$ /* elongate the RNA chain */
        S' $\leftarrow$ fold (1, $N_t$) /* fold the elongated RNA chain with UNAFold */
        **If** S' contains hairpin ($i_h$, $j_h$) **then**
            S $\leftarrow$ S'
            Next5 $\leftarrow$
            Bp $\leftarrow$ basepairs in hairpin ($i_h$, $j_h$)
            Constraints = Constraints + (F $i_h$ $j_h$ bp)
            B5 $\leftarrow$ $j_h$
            If hairpin ($i_h$, $j_h$) has extension then
                Ex0 ($i_{ex0}$, $j_{ex0}$) $\leftarrow$ shortest extension ($i_{ex0}$, $j_{ex0}$)

        $N_t \leftarrow (j_h+1)$  /* elongate the RNA chain */
     S' $\leftarrow$ fold (1, $N_t$) /* fold the elongated RNA chain with UNAFold */

```

Figure 2.8. Pseudo code of GTkinetics

2.4 Test cases

tRNAphe

Figure 2.9 shows the folding pathway of tRNAphe predicted by GTkinetics. The hairpins predicted by Z-scores were (10-25), (27-43), and (49-65), which are the D-stem, anti-codon stem and T-stem respectively (Figures 2.6, 2.7). The D-stem and the anti-codon stem were constrained one after the other, since they did occur in the foldings of corresponding lengths. The results agreed with an NMR study of successively longer tRNAphe fragments with a common 5' end (48). In this study, the T-stem and anticodon stem sequentially formed as the tRNA elongated to the corresponding lengths. However, the T-stem did not arise in the folding of (1-66); instead, a long-range helix formed. Since piece 49-65 is free to adopt any conformation, later on this long-range helix unfolded when a longer sequence was available. Figure 2.9D is a predicted transient structure, which can be tested by experiment, such as time-resolved NMR(70), which is able to monitor the folding and refolding of the transient structure at atomic resolution. Finally the cloverleaf structure is achieved, which is the same as the previous prediction (Figure 2.5B).

The 5' domain of 16s rRNA

The secondary structure of the 5' domain of 16S rRNA (1-560) predicted by GTkinetics is shown in Figure 2.10. The sensitivity and specificity are 52% and 48% respectively. Twelve of the predicted helices are true positives. In contrast to the case of tRNA, GTkinetics does not work as well as UNAfold, which achieves a higher specificity and sensitivity. Zuker *et al.* showed that MFE programs are able to provide a high quality prediction if the intact domain is given(77). In order to assess this possibility, I tested nucleotides 1-600 of the 16S rRNA with UNAfold and GTkinetics, since this segment does not comprise an actual domain. The additional 40 nucleotides dramatically change the result of UNAfold (Figure 2.11). The specificity of UNAfold drops from 80% for [1-560] to 43% for [1-600], while that of GTkinetics decreases only slightly, 52% to 49%. This suggests that the high accuracy of UNAfold resulted from specifying the actual domain. For GTkinetics, the local structures (hairpins and extensions) are added sequentially; so the structure of the longer sequence is not very different from the shorter sequences.

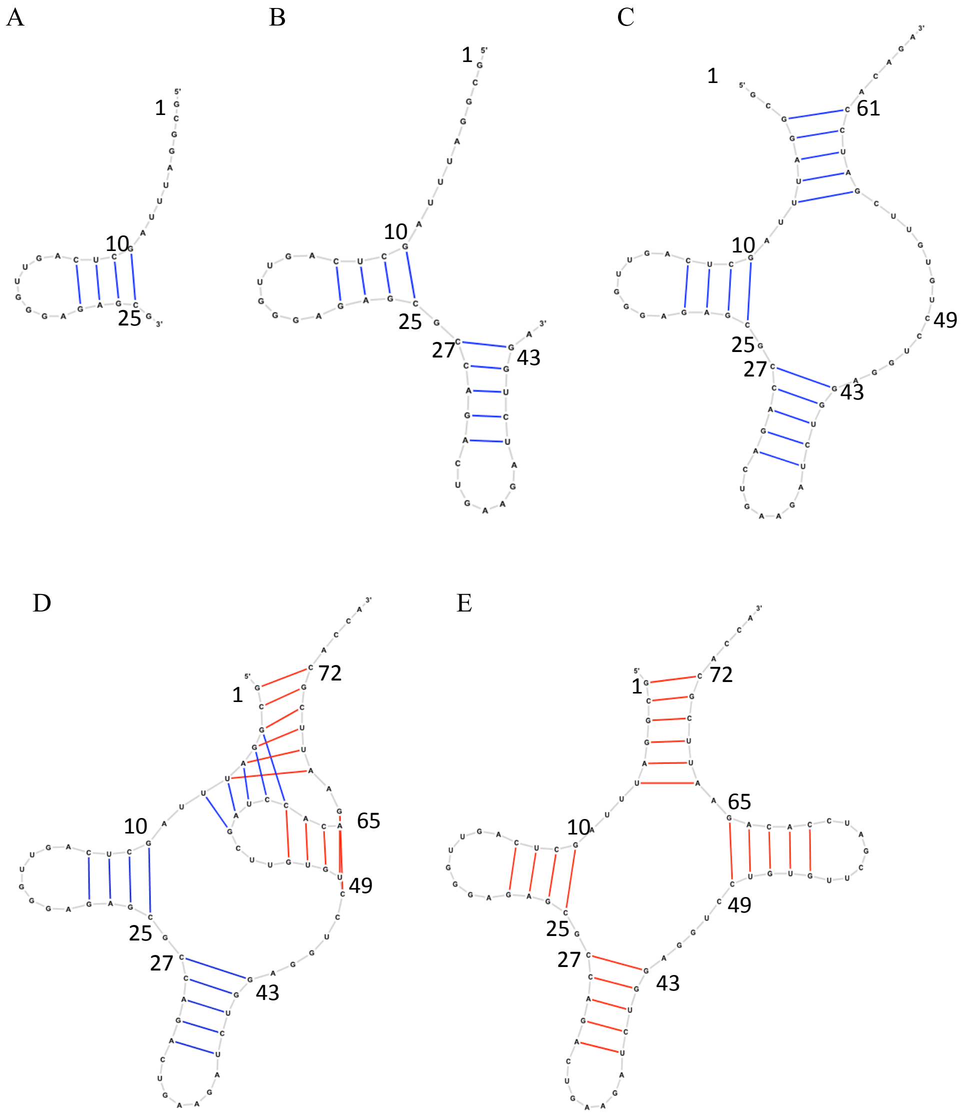

Figure 2.9 Folding pathway of tRNA^{phe} predicted by GTkinetics.
A. Formation of D-stem
B. Formation of anticodon stem, result without refolding
C. The predicted transient structure, result without refolding
D. The predicted refolding event. Blue base pairs are predicted in the previous step and are broken during formation of the red base pairs as the RNA elongates to its full length
E. Final cloverleaf structure

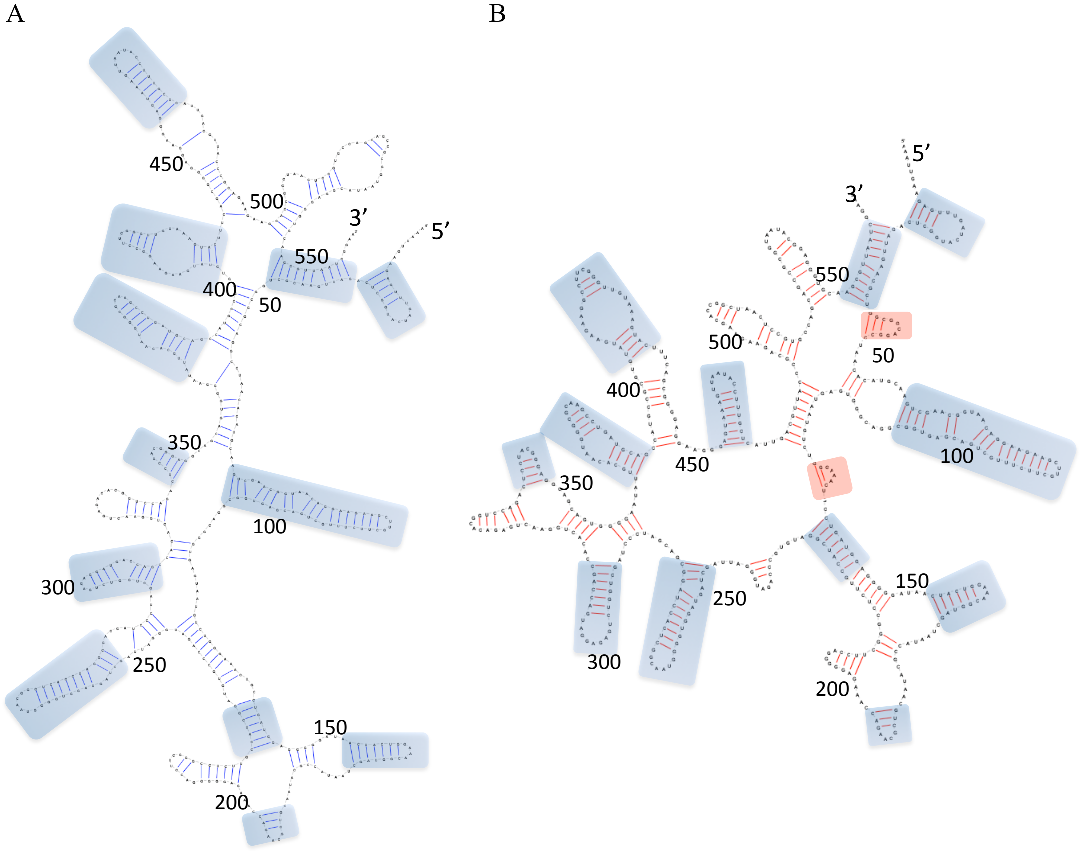

Figure 2.10 Secondary structure of the 5' domain of 16S rRNA. (A) The real structure, from the Comparative RNA web site, http://www.rna.icmb.utexas.edu. (B) This structure predicted by GTkinetics. Blue highlighted regions are correctly predicted by GTkinetics. Red highlighted regions represent the possible intermediate structures.

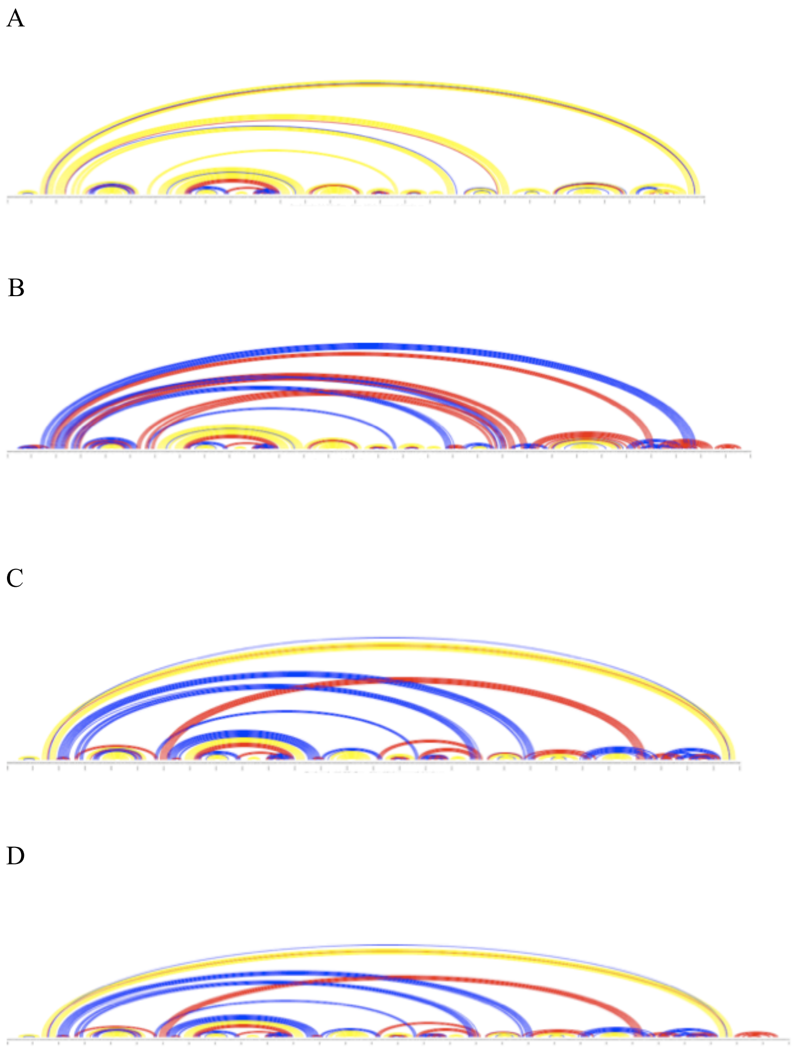

Figure 2.11 The predictions for different lengths of 16S rRNA with UNAfold and GTkinetics without refolding. Each arch represents a base pair connecting the two positions.
(A) The structure of 1-560, real vs UNAfold prediction
(B) The structure of 1-600, real vs UNAfold prediction
(C) The structure of 1-560, real vs GTkinetics prediction
(D) The structure of 1-600, real vs GTkinetics prediction.
Yellow, true positive; red, false positive; blue, false negative.

2.5 Intermediate structures

By examining the constrained structures given by GTkinetics, I found that one small false positive hairpin (38-48) occupied the nucleotides that are actually part of a long-range helix, which led to great error proliferation (Figure 2.10B). This caused three false negative helices, and two false positive helices. This suggests that the problem might be fixed by removing the constraint of (38-48). Figure 2.12B displays a significantly improved prediction after the constraint was removed. Another false positive example is hairpin (126-133). Once both constraints of (38-48) and (126-133) were taken off, the predicted secondary structure becomes much closer to the real structure (Figure 2.11). This kind of error cannot be avoided in the first version of GTkinetics, since all the stable hairpins and extensions are constrained all the time after confirmed. This indicates the necessity of a good refolding model, by which some local structures are able to refold properly.

It is possible that these structures may be transient structures along the folding pathway. They could guide the RNA folding by lowering the transition energy or eliminating misfolding pathways(47). For example, the 5' side of a long-range helix might need to be sequestered before the 3' side is synthesized. The 5' side could fold to some intermediate structure at first and refold later. The stability of intermediate structures should be moderate, high enough to avoid undesired premature refolding but low enough to permit melting and refolding to the final structure. Both hairpins (38-48) and (126-133) are on the 5' side of the long-range helices; both are small hairpins with a good Z-score. I hypothesize that helices (38-48) and (126-133) occur early in the folding process and unfold later. This prediction could be also tested by time-resolved NMR analysis or by temperature-gradient gel electrophoresis(41) using partially transcribed 16S rRNA. The possible existence of such intermediate structures again emphasizes the need to develop an algorithm that does allow refolding and to test this algorithm.

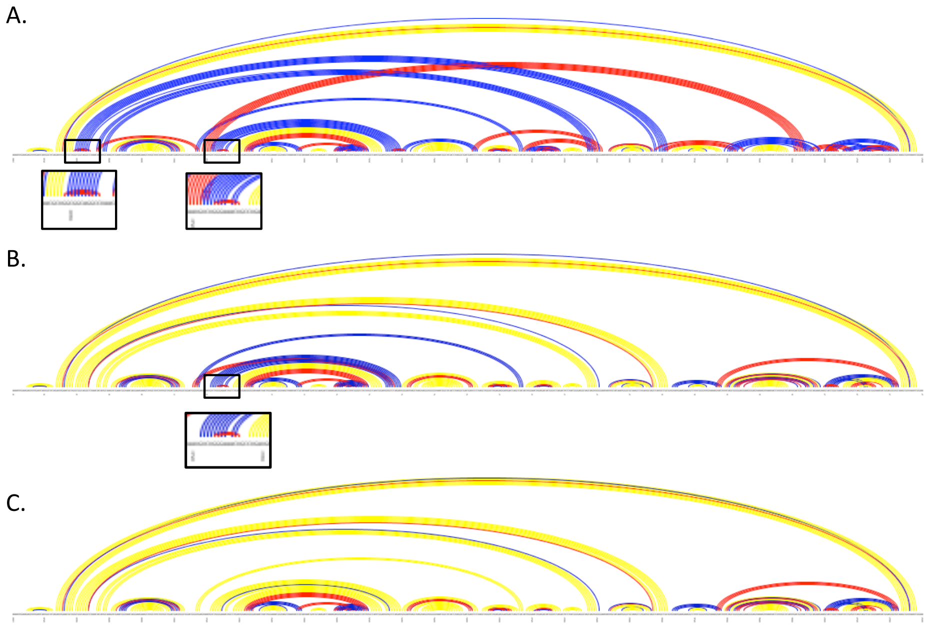

Figure 2.12 The performance of GTkinetics is greatly improved by allowing refolding. The improvement by removing the constraints of two hairpins.

(A) GTkinetics prediction without refolding.

(B) The folding by GTkinetics removing the constraint of hairpin (38-48) and allowing refolding. Nucleotides (38-48) become part of a long-range helix.

(C) The folding by GTkinetics removing the constraints of two hairpins (38-48), (126-133). Again, refolding leads to the formation of long-range helices.

The insets show the enlargement of the two wrong hairpins in the black boxes.

Yellow, true positive. Red, false positive. Blue, false negative.

2.6 Refolding suggested by suboptimal structures

One possible basis for refolding is that the ensemble of all possible secondary structures may not be dominated by the MFE structure as the RNA chain grows. If the free energy of suboptimal structures is close to that of a correct local MFE structure, the ensemble might be unstable and "leak" to the MFE structure of a longer fragment. It is worthwhile to examine the suboptimal structures along the folding pathway for local structures first, in order to see how refolding leads to the native structure.

The hypothesis of this analysis is that the native local structures, which are the MFE structures of certain pieces of sequence, dominate the ensemble suboptimal structures during transcription. I took the sequence segments containing the optimal hairpins according to the Z-scores and generated suboptimal ensembles for the elongating segments. I measured the stability of the MFE structures by free energy, Boltzmann weight and the number of structural clusters. A structural cluster is defined as a group of compatible structures without any conflicting base pairs. The procedure of clustering is described below. All the secondary structures in an ensemble are sorted from low to high energy in a list. At every step, the clustering starts from the top of the list. The first one is taken as the central structure of a new cluster. And then scan the list; if a structure is compatible with the central structure, it will be grouped into the cluster and removed from the list. The clustering runs iteratively until all the structures in the ensemble are assigned to a cluster. After clustering, the suboptimal ensemble is divided into several clusters. The central structures are the representatives of the ensemble structures. Boltzmann weight is also calculated according to the corresponding clusters, instead of every single structure. According to this hypothesis, we expected to observe true positive hairpins with low free energy, high Boltzmann weight and few competing clusters.

Methods

1. Locate the local optimal structures (closing base pair i-j) with minZ in the 16s rRNA of E.coli.

2. Fold the sequence starting with a segment 10 nucleotides long and add one more nucleotide at a time, until hit the nucleotide j+5.

3. Calculate the MFE and suboptimal structures within 2 kcal/mol using RNAsubopt, a program predicting suboptimal structures in Vienna package.

4. Group all the suboptimal structures into clusters and calculate the Boltzmann weight of cluster centered the MFE structure.

Results:

The local optimal structures selected by Z-score can be categorized into four classes from this analysis.

Class I (Figure 2.13): The local optimal structure forms exclusively. There are no alternative structures competing with the optimal local structures along the pathway. The Boltzmann weight of the MFE cluster is close to 1 (>0.95).

Class II (Figure 2.14): Alternative structures before the local optimal structure are unstable, i.e. the free energy is higher than -2 kcal/mol, and/or the Boltzmann weight is lower than 0.6.

Class III (Figure 2.15): The alternative structure before the local optimal structure is stable ($\Delta G < -2$ kcal/mol and Boltzmann weight > 0.6)

Class IV (Figure 2.16): The local optimal structure does not occur, or not stable ($\Delta G > -2$ kcal/mol or Boltzmann weight < 0.6)

The data below show the distribution of true and false positives in four classes:

	True+	False+
Class I	14	7
Class II	9	10
Class III	3	1
Class IV	2	9

The hairpins of Class I and II are able to form with little or no kinetic cost, while those of Class III and Class IV have a much higher barrier to overcome or are thermodynamically unstable. The results show that the true positive hairpins are overall more stable and more dominant among the suboptimal ensembles than the false positive ones. For the true positives, there are 23 out of 28 (82%) hairpins are in Class I and II; while for the false positives, there are only 17 out of 27 (63%) hairpins in these two classes. The chi-square test shows that the two distributions are significantly different from each other ($p< 0.001$).

This analysis indicates that the suboptimal structures may play an important role in helix formation as in Class II. The alternative hairpins could provide a nucleation platform by bringing the two ends of the real hairpins close to each other. Since they are not thermodynamically stable, the real ones finally replace them. Contrasted with Class II, the alternative hairpins in Class III are not able to melt easily, which suggests that the refolding is unlikely to happen. This can be used to evaluate the possibility of refolding. Given an initial structure and a target structure, a folding pathway can be plotted as demonstrated. Refolding is more likely to happen if there are a lot of suboptimal structures with close free energy along the folding pathway.

Unfortunately, the results here are not sufficient to guide us directly to modify GTfold to incorporate refolding events. In particular, it's not clear how to determine when to allow a local helix to refold, or to unfold and allow formation of a long-range helix. In Chapter 3, I will examine the possibility to incorporate entropic effects into MFE calculation, which favors short-range helices over long-range helices. In Chapter 4, I will examine sequence characteristics of short-range and long-range helices to identify factors that may suggest the design of future algorithms.

hairpin	1087	1098			
end	no.	ΔG	Bolt-w	clusters	
1093	0	0.00	1.0000	1	....
1094	0	0.00	1.0000	1	....
1095	0	0.00	1.0000	1	....
1096	0	0.00	1.0000	1	....
1097	0	0.00	0.4883	4	....
1098	1	-2.70	1.0000	1	..........(((......)))
1099	1	-4.40	1.0000	1	..........(((......))).
1100	1	-4.40	1.0000	1	..........(((......)))..
1101	2	-4.70	0.5009	3	.....(((.(((......))))))
1102	2	-5.50	0.4954	3	.....(((.(((......)))))).
1103	2	-5.50	0.4954	3	.....(((.(((......))))))..

Figure 2.13 Example of Class I hairpins. The local optimal structure forms exclusively. There are no alternative structures competing with the optimal local structures along the pathway. The Boltzmann weight of the MFE cluster is close to 1. Yellow highlighted, the local optimal hairpin with minZ.

hairpin	9	25			
end	no.	ΔG	Bolt-w	cluster no.	
7	0	0.00	1.0000	1	....
8	0	0.00	1.0000	1	....
9	0	0.00	1.0000	1	....
10	0	0.00	1.0000	1	....
11	0	0.00	1.0000	1	....
12	0	0.00	1.0000	1	....
13	0	0.00	1.0000	1	....
14	0	0.00	0.9350	2	....
15	0	0.00	0.7575	2	....
16	0	0.00	0.7575	2	....
17	0	0.00	0.6428	3	....
18	0	0.00	0.5829	4	....
19	1	0.00	0.3521	5	........((......)).
20	1	0.00	0.3434	5	........((......))..
21	1	0.00	0.3434	5	........((......))...
22	1	0.00	0.3434	5	........((......))....
23	1	0.00	0.3385	5	........((......)).....
24	1	0.00	0.2366	5	........((......))......
25	2	-2.60	1.0000	1	........(((((.......)))))
26	2	-4.30	1.0000	1	........(((((.......))))).
27	2	-4.30	1.0000	1	........(((((.......)))))..
28	2	-4.30	1.0000	1	........(((((.......)))))...
29	2	-4.30	1.0001	1	........(((((.......)))))....
30	2	-4.30	0.9506	2	........(((((.......))))).... .

Figure 2.14 Example of Class II hairpins. The local optimal structure does not occur, or not stable (ΔG > -2 kcal/mol or Boltzmann weight < 0.6). Yellow highlighted, the alternative structure before the local optimal.

hairpin	1506	1529			
1510	0	0.00	1.0000	1	....
1511	0	0.00	0.5889	2	....
1512	0	0.00	0.5889	2	....
1513	0	0.00	0.5889	2	....
1514	0	0.00	0.5889	2	....
1515	0	0.00	0.5746	3	....
1516	0	0.00	0.3243	4	....
1517	1	-1.10	0.6534	5	........((....)).
1518	1	-1.10	0.4690	5	........((....))..
1519	1	-1.10	0.4690	5	........((....))...
1520	2	-1.40	0.6981	5	....((..((....))..))
1521	3	-4.90	1.0000	1	...(((..((....))..)))
1522	4	-6.60	1.0000	1	..((((..((....))..))))
1523	4	-7.30	1.0000	1	..((((..((....))..)))).
1524	5	-7.50	0.6486	3	..........(((((....)))))
1525	6	-9.90	0.8578	2	.........((((((....))))))
1526	7	-13.40	1.0000	1	........(((((((....)))))))
1527	8	-14.90	1.0000	1	.......((((((((....))))))))
1528	9	-15.70	1.0000	1	......(((((((((....)))))))))
1529	10	-16.70	1.0000	1	.....((((((((((....))))))))))
1530	10	-17.50	1.0000	1	.....((((((((((....)))))))))).
1531	10	-17.50	1.0000	1	.....((((((((((....))))))))))..
1532	10	-17.50	1.0000	1	.....((((((((((....)))))))))))...
1533	10	-17.50	1.0000	1	.....((((((((((....)))))))))))....
1534	10	-17.50	0.6539	2	.....((((((((((....)))))))))).....

Figure 2.15 Example of Class III hairpins. The alternative structure before the structure is stable (ΔG < -2 kcal/mol and Boltzmann weight > 0.6). Yellow highlighted, the alternative structure before the local optimal.

```
15      391     400     4
395     0       0.00    0.8319   2      ..........
396     0       0.00    0.8319   2      ...........
397     0       0.00    0.8319   2      ............
398     0       0.00    0.7988   2      .............
399     0       0.00    0.7343   2      ..............
400     1       -1.00   0.7552   2      .....(((....)))
401     1       -1.80   0.9580   2      .....(((....))).
402     1       -1.80   0.9251   2      .....(((....)))..
403     1       -1.80   0.8048   3      .....(((....)))...
404     2       -2.40   0.6517   2      .....((.((...)).)).
405     3       -3.40   0.9064   2      ...((((.((...)).))))
```

Figure 2.16 Example of Class IV hairpins. The local optimal structure does not occur, or not stable (ΔG > -2 kcal/mol or Boltzmann weight < 0.6). Yellow highlighted, the local optimal structure.

CHAPTER 3

FOLDING WITH DISTANCE PENALTY FUNCTION

In order to favor local structures over long-distance structures, we introduced a distance penalty function into the free energy calculation. This penalty function serves as an ‘entropy’ term in the free energy. Classical statistical mechanics suggests that there is an entropic penalty for forming base pairs between nucleotides that are far apart in the primary sequence. The penalty cannot rise without limit, however, because this will absolutely forbid base pairing between nucleotides beyond some distance, since the entropic penalty will exceed the favorable enthalpy. For this reason, I have chosen to model the distance penalty as a sigmoidal function. It is added to every base pair, together with the stacking energy and loop energy. The result from the dynamic programming is a sum of the original nearest neighbor model (NNM) and this distance penalty. If the parameters of the penalty function are well estimated, we expect that it should provide a significant improvement in the prediction of RNA secondary structure.

3.1 Methods

The distance function is a sigmoid function with three parameters.

$$\Delta G_{DISTANCE}(d) = \frac{V_{\max}}{1 + e^{-\beta(d-d_0)}}.$$

In this expression, d is distance (number of nucleotides) between nucleotide i and j in base pair ij; V_{max} is the amplitude of the maximum penalty; β controls the steepness of the slope; d_0 is the critical distance, at which $\Delta G_{DISTANCE}$ is half of V_{max}. It may be easier to think of the width (w) of the region centered by d_0 where the penalty rises from about 0.05Vmax to about 0.95Vmax, for this particular function, w = 6/β. The curves of the functions using two different β are shown in Figure 3.1. As β increases, the curve becomes steeper at d_0. This function has been incorporated in the free energy calculation for a base pair within the source code of GTfold(29), in which the three parameters can be specified as an option.

Optimum values of V_{max}, β and d_0 were determined for the 16S and 23S rRNA of *E.coli*, using a grid search method. I sampled Vmax in the range of [0, 2.5], β in the range of [0, 0.1] and d_0 in the range of [100, 1000], 4560 combinations in total. Sensitivity of the predicted structure using each combination is then calculated. An important question is whether the optimum parameters for the two rRNAs cases are similar or significantly different.

3.2 Results

A 3D sensitivity landscape for 16S rRNA using different combinations is plotted in Figure 3.2. The color of each point indicates the sensitivity: the redder the color, the higher the sensitivity. It is interesting to observe that ‘sweet spot’, which is the best estimation of the parameters, is not a small area, but a large ‘sweet zone’. One boundary of the ‘sweet zone’ is ‘soft’, where the sensitivity is close to the highest value, while the other boundry is very ‘hard’, which is around 0.3 lower than the highest value. The zone occupies the space with Vmax > 0.5, β > 0.02 and d0 > 600. Within the zone, all the combinations of the parameters lead the structure with the maximum sensitivity (0.72) and maximum specificity (0.71). The circular arc representation of this secondary structure is shown in the Figure 3.3. The overall structure is very similar to the real structure, except for three major false long-range helices and two major local helices.

This structure is significantly better than that prediceted by NNM alone (sensitivity = 0.49, specificity = 0.49).

Compared to the 16S rRNA, the distance penalty function affects 23S rRNA to a smaller extent (Figure 3.4). With the distance function, sensitivity of 16S rRNA spans from 0.2 to 0.7, while that of 23S rRNA only spans from 0.2 to 0.65 and most falling between 0.5 and 0.65. The 'sweet zone' is also much smaller than that of 16S rRNA, which has 'soft boundaries' on both sides. The structure with highest sensitivity (sensitivity = 0.64, specificity = 0.61) is shown in Figure 3.5. However, this does not coincide with the highest specificity structure (sensitivity = 0.56, specificity = 0.65) (Figure 3.6). Both structures are better than that from NNM alone (sensitivity = 0.50, specificity = 0.47). The best specificity structure has most of the local structures correctly, but it does miss a lot of long-range ones. On the contrary, the best sensitivity structure has several true large domains, but falsely predicted other long-range helices. It is expected, since d0 of the highest specificity is 100, which prevents from forming long-range helices, resulting in fewer base pairs in the final structure.

When the 'sweet zone' of 23S rRNA is overlapped onto 16S rRNA landscape, the sensitivity of 16S rRNA in this zone is also around 0.65. This zone may be able to use for other RNAs. I tested this on the 16S and 23S rRNA of *Thermus thermophilus*. I selected one combination $V_{max} = 1$, $\beta = 0.08$, $d_0 = 750$. The sensitivity and specificity of the 16S rRNA with and without the distance function are 0.60, 0.57 and 0.54, 0.52 respectively. The sensitivity and specificity of the 23S rRNA with and without the distance function are 0.68, 0.62 and 0.59, 0.53 respectively. The predictions for both RNAs are improved by adding the distance function, but only modestly. In the next chapter, I will examine the possibility that the composition of short-range helices is different from that of long-range helices.

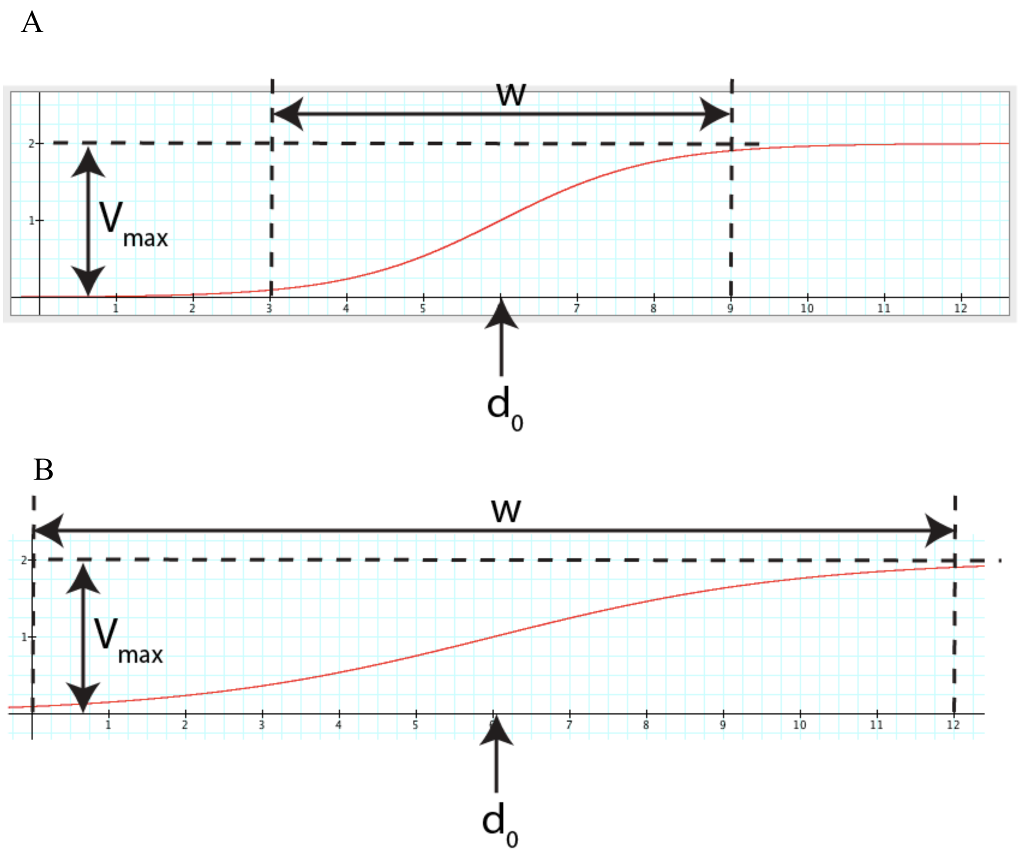

Figure 3.1 Curves of sigmoidal function using different β.
In both curves, $V_{max} = 2$, $d_0 = 6$. (A) $\beta = 1$, $w = 6/\beta = 6$. (B) $\beta = 0.5$, $w = 6/\beta = 12$.

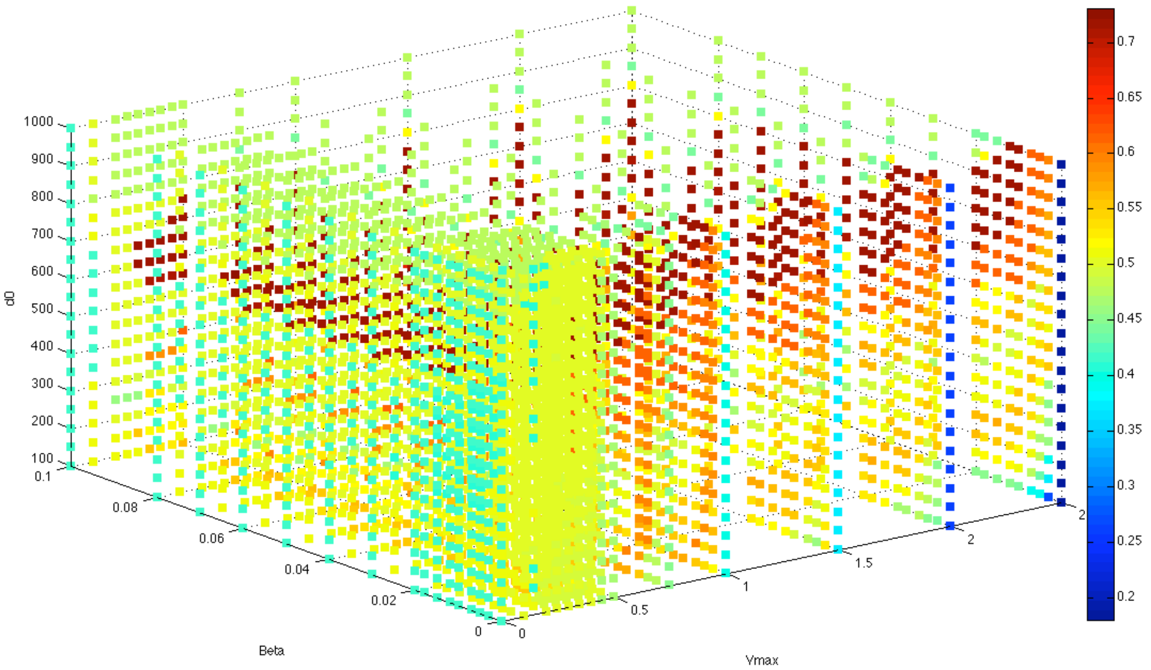

Figure 3.2 Sensitivity landscape of 16S rRNA using different combinations of Vmax, β and d0. Color code is on the right.

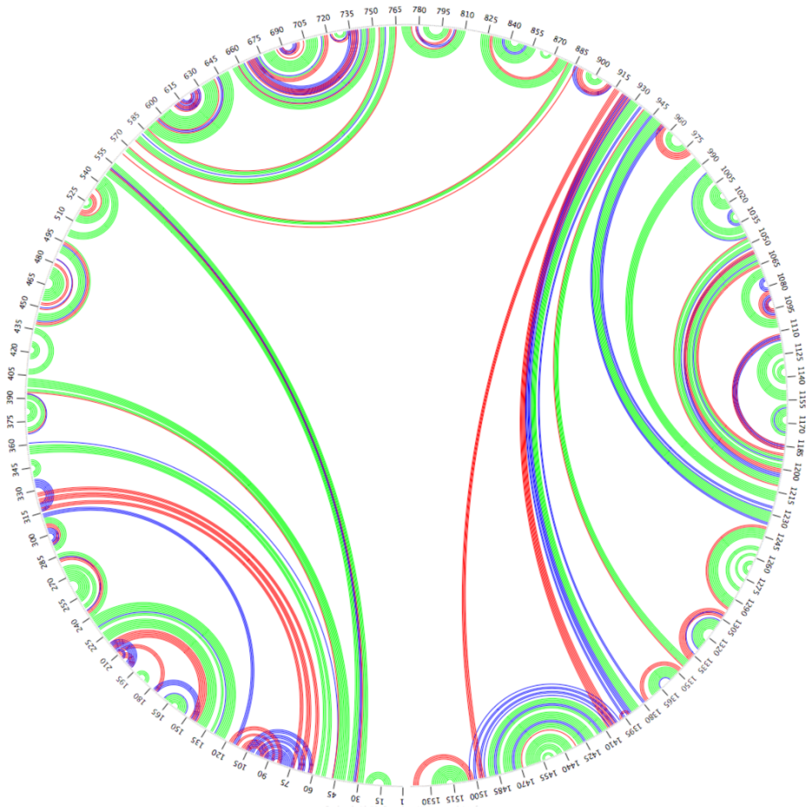

Figure 3.3 Secondary structure of 16S rRNA with highest sensitivity (0.72) and specificity (0.71). Green, true positive; Red, false positive; Blue, false negative.

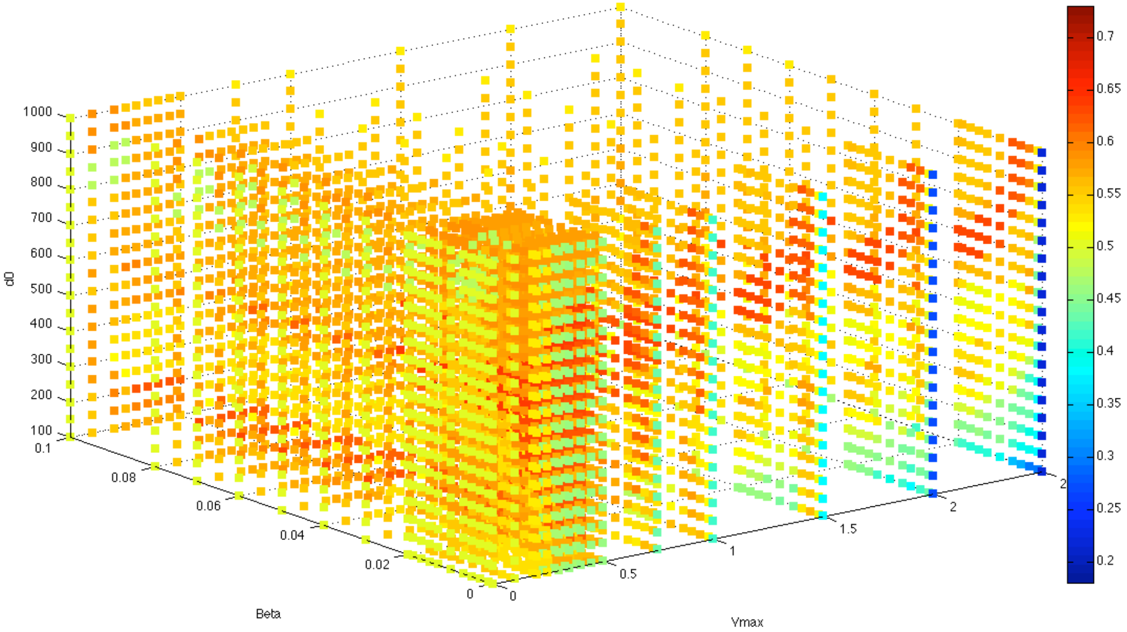

Figure 3.4 Sensitivity landscape of 23S rRNA using different combinations of Vmax, β and d0. Color code is on the right.

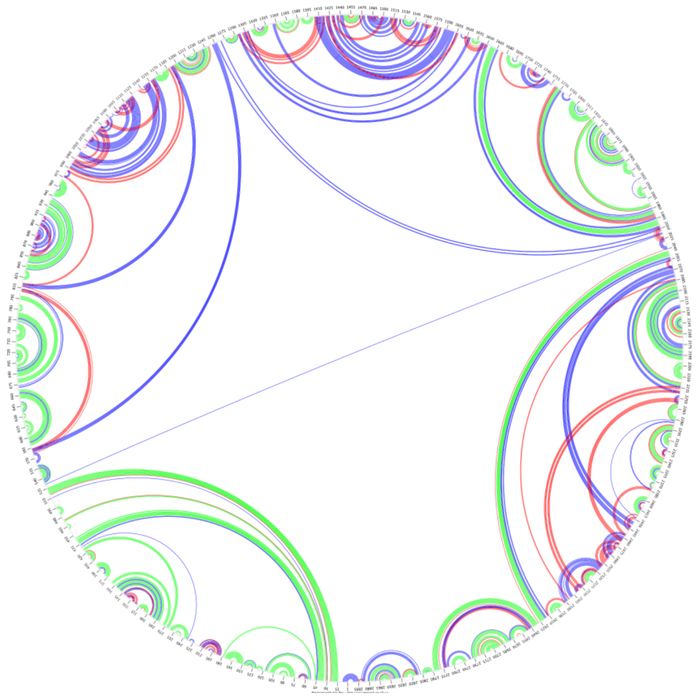

Figure 3.5 Secondary structure of 23S rRNA with highest specificity (0.65). Green, true positive; Red, false positive; Blue, false negative.

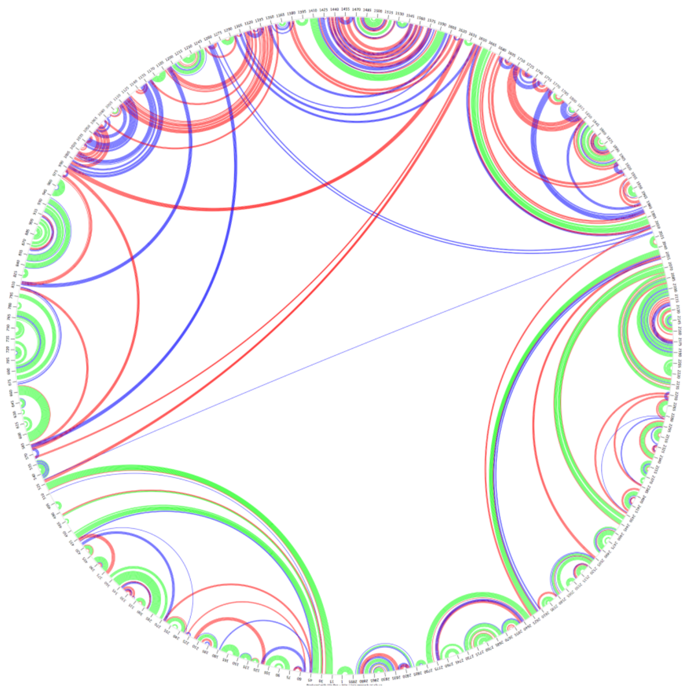

Figure 3.6 Secondary structure of 23S rRNA with highest sensitivity (0.64). Green, true positive; Red, false positive; Blue, false negative.

CHAPTER 4

SEQUENCE AND STRUCTURAL ANALYSIS OF RIBOSOMAL RNA

GTkinetics is able to capture stable local structures, but it also predicts a lot of false positive local structures. These false positives extensively affect the accuracy of the prediction. We hypothesize that some of the false positives are intermediate structures, which might exist temporally to sequester the 5' side for its real partner. In addition, other false positive local structures may occur off the folding pathway. These are resulted from the false signal of Z-score. In other words, we need some other criteria to evaluate the local structures, in addition to the Z-score. If we are able to eliminate the false positives, particularly those that lie off the folding path, the accuracy of GTkinetics will be improved. In order to obtain some characteristics to discriminate the real local structures from the false positive ones, I have carried out several analyses on the sequences and secondary structures of some ribosomal RNAs.

4.1 Standardized local nucleotide composition

The structure of the HIV RNA genome put forward by Watts *et al.* (78) has a remarkably high degree of single strandedness: only 41% of the 9142 nucleotides are involved in either Watson-Crick or wobble base pairing. This is in sharp contrast to the ~60% base pairing found in ribosomal RNAs and predicted MFE structures of random sequence RNAs. Our laboratory has recently completed an analysis of the HIV genome, finding that the unusual secondary structure has two sources. First, the HIV genome is very A-rich (36%) and C-poor (18%). Second, the composition of single-stranded regions is even more A-rich (48%) and C-poor (12%) than the genome as a whole. Since the 5' sides of each helix should have a propensity to remain single-stranded until the complementary 3' side of the helix is synthesized, I hypothesize that there will be more adenines on the 5' side of helices than on the 3' side, or that there will be a gradual build-up of adenines on the 5' sides of helices.

Because the absolute number of adenines increases monotoically, we used a normalized value to characterize the richness and poorness of A. We assume that all the adenines are scattered uniformly along the sequence, so that we have an expected number of adenine for every length, <A(length)>= sum(A)/total(length) * length. By subtracting the expected value from the real number of adenines at every length, we are able to visualize the A-rich and A-poor regions. Figure 4.1 shows the standardized local adenine composition of the 5’ domain of 16S rRNA. The two peaks circled are two typical examples where the hypothesis holds true. The regions with increasing adenines are the 5’ sides. The two peaks are located in the hairpin loops. The decreasing adenine regions are the 3’ sides of the hairpin helices. However, this is not always true along the entire sequence. Many of the long-range helices do not have this pattern. In order to test this hypothesis from a statistical perspective, I counted the adenines of both sides of all helices in the 16S rRNA to verify whether or not adenine at the 5’ sides are significantly more than those at the 3’ sides. The results show that the number of adenines at the 5’ sides (n = 2152) is even smaller than that of the 3’ sides (n = 2253).

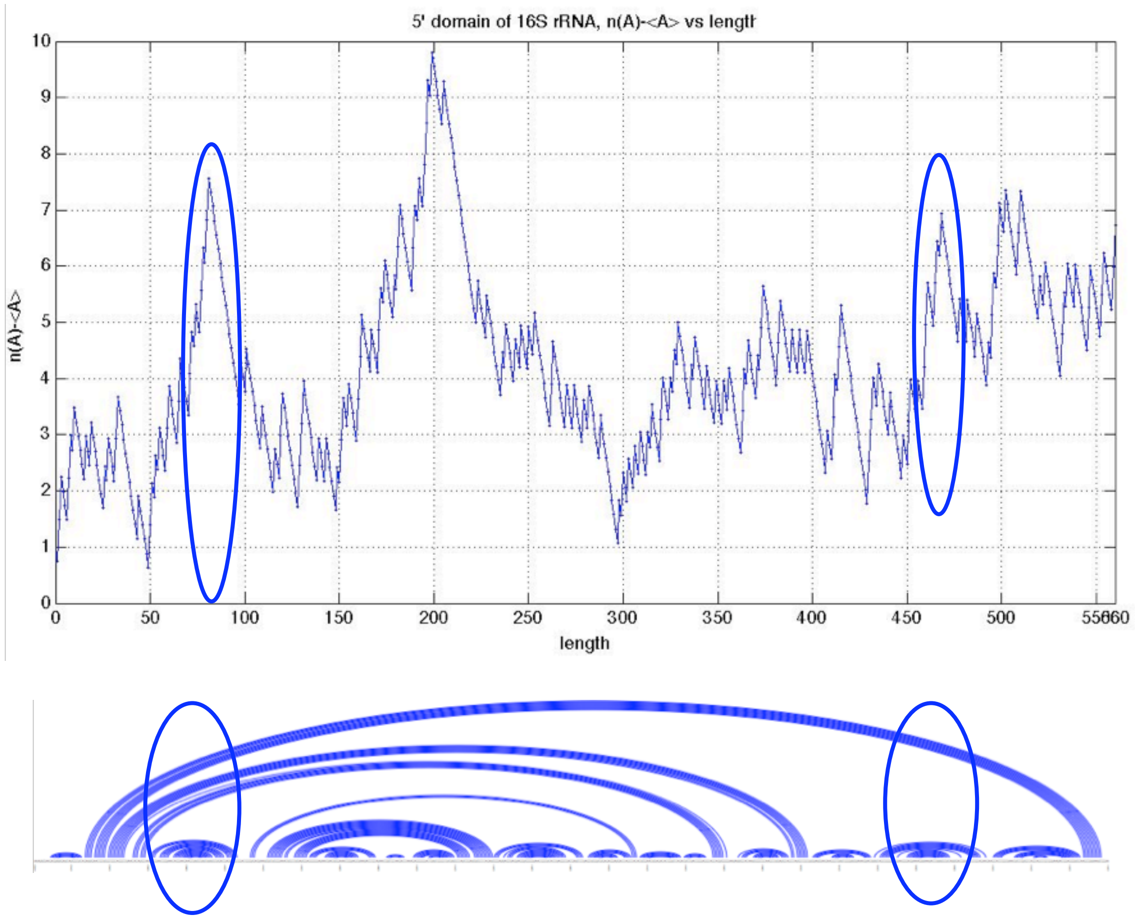

Figure 4.1 Standardized local adenine composition of the 5' domain in the 16S rRNA.Top, The standardized local adenine composition of the 5' domain of the 16S rRNA. Bottom, the secondary structure of the 5' domain of the 16S rRNA. The two peaks circled are two hairpins; whose 5' sides are adenine rich and 3' sides are adenine poor.

4.2 Base pair composition analysis

From the standardized local nucleotide composition analysis, we found some signals for local helices but not for long-range helices. This suggested that folding of long-range helices is different from that of local ones. Intermediate structures have been already suggested as a strategy to protect the 5’ side before its 3’ complement is synthesized. Another direct perspective to examine the differences between long-range and local helices is to examine the base pair composition. The null hypothesis is that the distribution of four kinds of base pairs (A-U, G-C, G-U and non-canonical (nc) pairs) in the local helices is the same as that in long-range (l-r) helices.

I tested 24 ribosomal RNAs from 3 domains of life (Table 3.1). From the chi-square test (Table 3.2), it is obvious that the base pair compositions between local and long-range helices are significantly different, p=3.23E-11. The biggest contributor is from long-range GU base pairs. The number of GU base pairs in long-range helices is smaller than the expected number, while GU base pairs in local helices are more frequent than expected. This is also true for non-canonical base pairs. This indicates that the high-energy GU and non-canonical base pairs are not favored by long-range helices, which may suggest long-range helices need lower average base pair energy to overcome the higher entropy penalty. The next biggest contributor is the long-range AU base pair. AU base pair appears more than expected in the long-range helices, but less than expected in the local helices. The distribution of GC base pair is almost the same in local and long-range helices. It is also interesting that AU and GU base pairs are mutually impairing. When the number of AU increases, the number of GU decreases, and vise versa. This suggests that the promiscuity of uracil facilitates the formation of local and long-range helices via different means.

Table 4.1 The rRNA sequences in the test of base pair composition

Species	Type	CRW ID	Length	Domain
Escherichia coli	16S	CRW_00110	1542	Bacteria
Escherichia coli	23S	CRW_00492	2904	Bacteria
Deinococcus radiodurans	16S	CRW_00105	1504	Bacteria
Deinococcus radiodurans	23S	CRW_00490	2880	Bacteria
Staphylococcus aureus	16S	CRW_00196	1555	Bacteria
Staphylococcus aureus	23S	CRW_00515	2923	Bacteria
Thermus thermophilus	16S	CRW_00252	1518	Bacteria
Thermus thermophilus	23S	CRW_00519	2915	Bacteria
Haemophilus influenzae	16S	CRW_00129	1539	Bacteria
Haemophilus influenzae	23S	CRW_00496	2897	Bacteria
Arabidopsis thaliana	16S	CRW_00303	1808	Eucaryota
Arabidopsis thaliana	23S	CRW_00524	3539	Eucaryota
Drosophila melanogaster	16S	CRW_00330	1995	Eucaryota
Drosophila melanogaster	23S	CRW_00536	1335	Eucaryota
Xenopus laevis	16S	CRW_00415	1826	Eucaryota
Xenopus laevis	23S	CRW_00545	1640	Eucaryota
Carelia paradoxa	16S	CRW_00327	1807	Eucaryota
Carelia paradoxa	23S	CRW_00547	2926	Eucaryota
Saccharomyces cerevisiae	16S	CRW_00389	1800	Eucaryota
Saccharomyces cerevisiae	23S	CRW_00529	3554	Eucaryota
Thermococcus celer	16S	CRW_00470	3029	Archaea
Thermococcus celer	23S	CRW_00042	1487	Archaea
Haloarcula marismortui	16S	CRW_00022	1473	Archaea
Haloarcula marismortui	23S	CRW_00467	2925	Archaea

Table 4.2 Base pair composition analysis of local and long-range helices. A. Base pair composition of local and long-range (l-r) helices. B. The actual number and expected number of all kinds base pairs. Green, actual number is larger than expected; Red, actual number is smaller than expected. C, the chi-square test of the two distributions. Degree of freedom is 3, chi-square = 51.8458. P-value is 3.23E-11.

A

	A-U	G-U	G-C	NC	total
local	2470	1320	5414	602	9806
l-r	1364	495	2684	222	4765
total	3834	1815	8098	824	14571

B

	Local		Long-range	
	Expected	Actual	Expected	Actual
A-U	2580.21	2470	1253.79	1364
G-U	1221.46	1320	593.54	495
G-C	5449.80	5414	2648.20	2684
NC	554.54	602	269.46	222

C

	local	l-r
A-U	4.7073	9.6872
G-U	7.9497	16.3598
G-C	0.2351	0.4839
NC	4.0626	8.3604

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

As the functional importance of RNA has become more apparent in recent years, research on RNA secondary structure determination has been developing very rapidly. There are two big questions in RNA secondary structure prediction. One is that given a certain sequence, what is the final structure? The other is, given the sequence and the final structure, what is the folding path?

My work here does not only aim at more accurate secondary structure prediction, but also try to understand RNA folding *in vivo*. GTkinetics is designed to serve these two targets. In GTkinetics, the Z-score is the key to the selection of local structures. With the Z-score criterion, there are 10 out of 13 hairpins correctly predicted in the 5' domain of the 16s rRNA. One of three false hairpins (199-209) is also suggested a refolded region in SHAPE analysis. Z-score is a normalized free energy, which compares the stability of local structures of different lengths. It facilitates the selection for local competitions, which suggest that the relative stability plays a more important role than the absolute free energy in the local structures. It is worthwhile to point out that Z-score is still a thermodynamic term, which means that the local structures fold thermodynamically.

In GTkinetics, Z-score is not the only criterion of local structure selection. The length of the sequence elongation favors the lowest Z-score structure, but the structure is not forced to form. Instead, it is under the competitive pressure from other local structures and possible long-range interactions with other parts along the sequence. In this way, the program partially mimics the folding *in vivo*, which is dynamic but also is encoded in the sequence. If a local minimum Z-score structure is able to form under such competition, it is confirmed as a stable one. This step is also determined by thermodynamics, since the structure computed is from energy minimization.

The kinetic effect involved in GTkinetics is to constrain those stable local structures, which assumes an infinitely high-energy barrier to refold them. This is not very realistic, but it is a good starting point to evaluate the selection method for local structures and to learn some characteristics of the RNA folding. Some small hairpins occur on the 5' side of the long-range helices, are these may be intermediate structures (Section 3.3). Meyer *et al.* (54) did a series of statistical tests and pointed out that the formation of transient structures, which may serve as guidelines for the co-transcriptional folding pathway, is encouraged. Wong *et al.* (49) found a non-native structure during transcription, which facilitates the RNA folding. Alternative structures or non-native structures are no longer considered as obstacles in RNA folding, but as intermediate or transient structures that guide and regulate the RNA folding. Both of the proposed intermediate structures in the 16S rRNA stand on the 5' sides of the long-range helices, by which the 5' sides are protected from premature interactions with other parts of the sequence. Their free energies are -2.7 kcal/mol and -1.6 kcal/mol respectively, which are not so high as to prevent refolding. They have the characteristics of intermediate structures: relatively stable (low Z-score) and relatively refoldable (high free energy). GTkinetics enables us to capture these possible intermediate structures; however, the question remains as to when and how to refold them. A refolding model is required within every kinetic folding algorithm, which we need to establish in GTkinetics.

By analyzing the suboptimal structures along the folding pathways of hairpins, I have concluded that real hairpins are usually more stable and dominant in the suboptimal ensemble than the false positive hairpins. It is interesting to observe that half of the real hairpins fall into Class II, which feature alternative transient structures before the formation of real structures. This suggests a mechanism to evaluate refolding, but it will require a new searching method for target structures, particularly long-range helices. Also the method for evaluating possible refolding events needs to be quantified.

Instead of incorporating a refolding model in the program, the other way to solve the long-range base pairs falsely predicted by MFE program is to introduce a distance-dependent penalty function into the program. The sensitivity landscapes of 16S and 23S rRNA seem to be very different with regard to the optimal parameters in the variation and ‘sweet spot’ location. This indicates that there is no ‘optimal’ combination for all RNAs, but it is possible to obtain a better prediction if an appropriate set of parameters is selected. From the test on *Thermus thermophilus* rRNA, we saw that a compromise set of parameters from the training sets did improve the prediction to a moderate extent.

From the analysis of local nucleotide composition and base pair composition, it is apparent that RNAs utilize different strategies to form local and long-range helices. The local nucleotide composition of adenine accumulates on the 5’ sides of some local structures, but not on the 5’ sides of long-range helices. In the base pair composition analysis, the A-U pairs are more frequent than expected in the local helices, but less frequent than expected in the long-range helices. It has been pointed out that GC content is higher in stems than in the loop regions(80). This also agrees with my analysis. The number of G-C pairs is more than half of the number of total base pairs. Furthermore, the G-C pair has the same frequency in local and long-range helices. This suggests that GC characterizes helical regions and adenine characterizes loop regions.

It is very interesting that uracil, which can pair with both adenine and guanine, has different composition of partners in local and long-range pairs. AU and GU base pairs are mutually impairing. When the frequency of AU increases, the frequency of GU decreases, and *vise versa*. Generally, there are fewer week base pairs (G-U and non-canonical) in long-range helices. It is reasonable to hypothesize that for local structures, because the kinetic barrier is very small, it is acceptable to adopt some high-energy base pairs. However, in the long-range helices, which either result from refolding or from directly closing a multi-branch loop, the kinetic barriers are much higher than the local ones, so more stable low-energy base pairs are favored. RNAs are the product of natural selection, in which both thermodynamics and kinetics play important roles as selection pressure. The sequences of real RNAs have evolved to form the desired structures efficiently. These compositional trends may improve the accuracy of RNA secondary structure prediction, because they allow us to compare predicted structures against known compositional preferences.

The results here point to future directions for RNA secondary structure prediction. I would propose a new GTkinetics program by adding more components to it. First, a list of possible long-range helices should be generated using a sliding window method similar to the one used for identifying candidate local structures. Once we have the lists of both local and long-range helices, we can score them for free energy, nucleotide composition, base pair composition and so on, in the hope of determining the stability of a candidate local and long-range helices. Next, as in the current GTkinetics, we can add candidate helices sequentially according to their scores. Refolding will take place if we replace a previously stable local helix with a long-rang helix. We can then generate suboptimal ensembles from the initial to target structures. The various possibilities can be evaluated from the free energy differences, Boltzmann weights and cluster numbers. The distance function can also be incorporated in GTkinetics to entropically favor local helices over long-range ones.

In order to accomplish the proposed changes to GTkinetics, I need to complete the following tasks. First a computationally inexpensive method is required for populating long-range helices. I can use the sliding window method similar to that used for generating local structures. To decrease the computing complexity, I can slide the window in steps of five nucleotides apart instead of one nucleotide. After obtaining all the possible local and long-range helices, we need a reliable scoring method. The score of each helix will evaluate thermodynamic stability (Z-score and ΔG) and the likelihood of a local structure or a long-range helix. The likelihood calculation will incorporate the analyses results in Chapter 4, such as base pair composition. How to assign the weights for stability and likelihood is not a trivial problem. I can apply a grid search method as in the distance function determination, using known local and long-range helices as a training set, to get a consistent solution. According to the stability and likelihood, I will add the highest score helix at each step. In this way, some local structures may be replaced with long-range helices as the RNA chain grows. In this scenario, we need to evaluate the possibility of refolding. As suggested by Section 2.6, I will generate suboptimal ensembles along the folding pathway from the initial structure to the target one. The free energy of the MFE structure, the Boltzmann weight of the MFE cluster and the number of total clusters will be factors to determine whether to accept refolding. The threshold of these three factors can also be determined by using a grid search method on various sets of training data set.

These steps are workable but challenging. It is possible that an optimal solution may not be found; however, a consistent suboptimal solution is also acceptable. The main goal of this research is to understand the RNA folding by learning from the mechanisms encoded in the sequences. The extent of our understanding will be tested by the capability of the program to reproduce the native structures.

REFERENCES

1. Korostelev, A. and Noller, H.F. (2007) The ribosome in focus: new structures bring new insights. *Trends Biochem Sci*, **32**, 434-441.

2. Steitz, T.A. (2008) A structural understanding of the dynamic ribosome machine. *Nat Rev Mol Cell Biol*, **9**, 242-253.

3. Bessonov, S., Anokhina, M., Will, C.L., Urlaub, H. and Luhrmann, R. (2008) Isolation of an active step I spliceosome and composition of its RNP core. *Nature*, **452**, 846-850.

4. Edwards, T.E., Klein, D.J. and Ferre-D'Amare, A.R. (2007) Riboswitches: small-molecule recognition by gene regulatory RNAs. *Curr Opin Struct Biol*, **17**, 273-279.

5. Nagel, J.H. and Pleij, C.W. (2002) Self-induced structural switches in RNA. *Biochimie*, **84**, 913-923.

6. Dayie, K.T. (2008) Key labeling technologies to tackle sizeable problems in RNA structural biology. *Int J Mol Sci*, **9**, 1214-1240.

7. Amaral, P.P., Dinger, M.E., Mercer, T.R. and Mattick, J.S. (2008) The eukaryotic genome as an RNA machine. *Science*, **319**, 1787-1789.

8. Brion, P. and Westhof, E. (1997) Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct*, **26**, 113-137.

9. Tinoco, I., Jr. and Bustamante, C. (1999) How RNA folds. *J Mol Biol*, **293**, 271-281.

10. Fox, G.W. and Woese, C.R. (1975) 5S RNA secondary structure. *Nature*, **256**, 505-507.

11. Noller, H.F. and Woese, C.R. (1981) Secondary structure of 16S ribosomal RNA. *Science*, **212**, 403-411.

12. Woese, C.R., Magrum, L.J., Gupta, R., Siegel, R.B., Stahl, D.A., Kop, J., Crawford, N., Brosius, J., Gutell, R., Hogan, J.J. *et al.* (1980) Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res*, **8**, 2275-2293.

13. Noller, H.F., Kop, J., Wheaton, V., Brosius, J., Gutell, R.R., Kopylov, A.M., Dohme, F., Herr, W., Stahl, D.A., Gupta, R. *et al.* (1981) Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res*, **9**, 6167-6189.

14. Gutell, R.R., Lee, J.C. and Cannone, J.J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol*, **12**, 301-310.

15. Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T. and Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327-339.

16. Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**, 905-920.

17. Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R. and Zamir, A. (1965) Structure of a Ribonucleic Acid. *Science*, **147**, 1462-1465.

18. Tinoco, I., Jr., Uhlenbeck, O.C. and Levine, M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362-367.

19. Tinoco, I., Jr., Borer, P.N., Dengler, B., Levin, M.D., Uhlenbeck, O.C., Crothers, D.M. and Bralla, J. (1973) Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol*, **246**, 40-41.

20. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, **9**, 133-148.

21. Zuker, M. (1989) Computer prediction of RNA structure. *Methods Enzymol*, **180**, 262-288.

22. Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48-52.

23. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406-3415.

24. Nussinov, R., Pieczenik, G., Griggs, J.R. and Kleitman, D.J. (1978) Algorithms for Loop Matchings. *Siam Journal on Applied Mathematics*, **35**, 68-82.

25. Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, **453**, 3-31.

26. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M. and Schuster, P. (1994) Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package). *Monatsh Chem*, **125**, 167-188.

27. Hofacker, I.L. (2004) RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics*, **Chapter 12**, Unit 12 12.

28. Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M. and Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A*, **101**, 7287-7292.

29. Amrita Mathuriya, D.A.B., Christine E. Heitsch, Stephen C. Harvey. (2009) GTfold: a scalable multicore code for RNA secondary structure prediction. *Proceedings of the 2009 ACM symposium on Applied Computing*

30. Konings, D.A. and Gutell, R.R. (1995) A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA*, **1**, 559-574.

31. Doshi, K.J., Cannone, J.J., Cobaugh, C.W. and Gutell, R.R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.

32. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, **288**, 911-940.

33. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T. and Turner, D.H. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A*, **83**, 9373-9377.

34. Burkard, M.E., Kierzek, R. and Turner, D.H. (1999) Thermodynamics of unpaired terminal nucleotides on short RNA helixes correlates with stacking at helix termini in larger RNAs. *J Mol Biol*, **290**, 967-982.

35. Xia, T., McDowell, J.A. and Turner, D.H. (1997) Thermodynamics of nonsymmetric tandem mismatches adjacent to G.C base pairs in RNA. *Biochemistry*, **36**, 12486-12497.

36. Chen, G. and Turner, D.H. (2006) Consecutive GA pairs stabilize medium-size RNA internal loops. *Biochemistry*, **45**, 4025-4043.

37. Bourdelat-Parks, B.N. and Wartell, R.M. (2005) Thermodynamics of RNA duplexes with tandem mismatches containing a uracil-uracil pair flanked by C.G/G.C or G.C/A.U closing base pairs. *Biochemistry*, **44**, 16710-16717.

38. McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105-1119.

39. Ding, Y. and Lawrence, C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*, **31**, 7280-7301.

40. Mahen, E.M., Harger, J.W., Calderon, E.M. and Fedor, M.J. (2005) Kinetics and thermodynamics make different contributions to RNA folding in vitro and in yeast. *Mol Cell*, **19**, 27-37.

41. Repsilber, D., Wiese, S., Rachen, M., Schroder, A.W., Riesner, D. and Steger, G. (1999) 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*, **5**, 574-584.

42. Morgan, S.R. and Higgs, P.G. (1996) Evidence for kinetic effects in the folding of large RNA molecules. *J Chem Phys*, **105**, 7152-7157.

43. Pan, T. and Sosnick, T. (2006) RNA folding during transcription. *Annu Rev Biophys Biomol Struct*, **35**, 161-175.

44. Wickiser, J.K., Winkler, W.C., Breaker, R.R. and Crothers, D.M. (2005) The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol Cell*, **18**, 49-60.

45. Chao, M.Y., Kan, M.C. and Lin-Chao, S. (1995) RNAII transcribed by IPTG-induced T7 RNA polymerase is non-functional as a replication primer for ColE1-type plasmids in Escherichia coli. *Nucleic Acids Res*, **23**, 1691-1695.

46. Lewicki, B.T., Margus, T., Remme, J. and Nierhaus, K.H. (1993) 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*, **231**, 581-593.

47. Wong, T.N., Sosnick, T.R. and Pan, T. (2007) Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proc Natl Acad Sci U S A*, **104**, 17995-18000.

48. Boyle, J., Robillard, G.T. and Kim, S.H. (1980) Sequential folding of transfer RNA. A nuclear magnetic resonance study of successively longer tRNA fragments with a common 5' end. *J Mol Biol*, **139**, 601-625.

49. Kramer, F.R. and Mills, D.R. (1981) Secondary structure formation during RNA synthesis. *Nucleic Acids Res*, **9**, 5109-5124.

50. Pan, T., Artsimovitch, I., Fang, X.W., Landick, R. and Sosnick, T.R. (1999) Folding of a large ribozyme during transcription and the effect of the elongation factor NusA. *Proc Natl Acad Sci U S A*, **96**, 9545-9550.

51. Heilman-Miller, S.L. and Woodson, S.A. (2003) Effect of transcription on folding of the Tetrahymena ribozyme. *RNA*, **9**, 722-733.

52. Meyer, I.M. and Miklos, I. (2004) Co-transcriptional folding is encoded within RNA genes. *BMC Mol Biol*, **5**, 10.

53. Flamm, C. and Hofacker, I.L. (2008) Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatsh Chem*, **139**, 447-457.

54. Mironov, A.A., Dyakonova, L.P. and Kister, A.E. (1985) A kinetic approach to the prediction of RNA secondary structures. *J Biomol Struct Dyn*, **2**, 953-962.

55. Porschke, D. (1974) A direct measurement of the unzippering rate of a nucleic acid double helix. *Biophys Chem*, **2**, 97-101.

56. Mironov, A.A. and Lebedev, V.F. (1993) A kinetic model of RNA folding. *Biosystems*, **30**, 49-56.

57. Danilova, L.V., Pervouchine, D.D., Favorov, A.V. and Mironov, A.A. (2006) RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. *J Bioinform Comput Biol*, **4**, 589-596.

58. Isambert, H. and Siggia, E.D. (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc Natl Acad Sci U S A*, **97**, 6515-6520.

59. Pan, J., Thirumalai, D. and Woodson, S.A. (1997) Folding of RNA involves parallel pathways. *J Mol Biol*, **273**, 7-13.

60. Geis, M., Flamm, C., Wolfinger, M.T., Tanzer, A., Hofacker, I.L., Middendorf, M., Mandl, C., Stadler, P.F. and Thurner, C. (2008) Folding kinetics of large RNAs. *J Mol Biol*, **379**, 160-173.

61. Morgan, S.R. and Higgs, P.G. (1998) Barrier heights between ground states in a model of RNA secondary structure. *Journal of Physics a-Mathematical and General*, **31**, 3153-3170.

62. Porschke, D. and Eigen, M. (1971) Co-operative non-enzymic base recognition. 3. Kinetics of the helix-coil transition of the oligoribouridylic--oligoriboadenylic acid system and of oligoriboadenylic acid alone at acidic pH. *J Mol Biol*, **62**, 361-381.

63. Chen, X., He, S.M., Bu, D., Chen, R. and Gao, W. (2007) A flexible Stem-based local search algorithm for predicting RNA secondary structures including pseudoknots. *IEEE*, 411-417.

64. Chen, X., He, S.M., Bu, D., Zhang, F., Wang, Z., Chen, R. and Gao, W. (2008) FlexStem: improving predictions of RNA secondary structures with pseudoknots by reducing the search space. *Bioinformatics*, **24**, 1994-2001.

65. Rivas, E. and Eddy, S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol*, **285**, 2053-2068.

66. Ren, J., Rastegari, B., Condon, A. and Hoos, H.H. (2005) HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, **11**, 1494-1504.

67. Ruan, J., Stormo, G.D. and Zhang, W. (2004) ILM: a web server for predicting RNA secondary structures with pseudoknots. *Nucleic Acids Res*, **32**, W146-149.

68. Flamm, C., Hofacker, I.L., Stadler, P.F. and Wolfinger, M.T. (2002) Barrier trees of degenerate landscapes. *Zeitschrift Fur Physikalische Chemie-International Journal of Research in Physical Chemistry & Chemical Physics*, **216**, 155-173.

69. Nagel, J.H., Flamm, C., Hofacker, I.L., Franke, K., de Smit, M.H., Schuster, P. and Pleij, C.W. (2006) Structural parameters affecting the kinetics of RNA hairpin formation. *Nucleic Acids Res*, **34**, 3568-3576.

70. Furtig, B., Buck, J., Manoharan, V., Bermel, W., Jaschke, A., Wenter, P., Pitsch, S. and Schwalbe, H. (2007) Time-resolved NMR studies of RNA folding. *Biopolymers*, **86**, 360-383.

71. Furtig, B., Wenter, P., Reymond, L., Richter, C., Pitsch, S. and Schwalbe, H. (2007) Conformational dynamics of bistable RNAs studied by time-resolved NMR spectroscopy. *J Am Chem Soc*, **129**, 16222-16229.

72. Harlepp, S., Marchal, T., Robert, J., Leger, J.F., Xayaphoummine, A., Isambert, H. and Chatenay, D. (2003) Probing complex RNA structures by mechanical force. *Eur Phys J E Soft Matter*, **12**, 605-615.

73. Kim, S.H., Suddath, F.L., Quigley, G.J., McPherson, A., Sussman, J.L., Wang, A.H., Seeman, N.C. and Rich, A. (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science*, **185**, 435-440.

74. van Batenburg, F.H., Gultyaev, A.P., Pleij, C.W., Ng, J. and Oliehoek, J. (2000) PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res*, **28**, 201-204.

75. Taufer, M., Licon, A., Araiza, R., Mireles, D., van Batenburg, F.H., Gultyaev, A.P. and Leung, M.Y. (2009) PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic Acids Res*, **37**, D127-135.

76. Abrahams, J.P., van den Berg, M., van Batenburg, E. and Pleij, C. (1990) Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res*, **18**, 3035-3044.

77. Zuker, M., Jaeger, J.A. and Turner, D.H. (1991) A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with

structures determined by phylogenetic comparison. *Nucleic Acids Res*, **19**, 2707-2714.

78. Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess, J.W., Jr., Swanstrom, R., Burch, C.L. and Weeks, K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711-716.

79. Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y. and Segal, E. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103-107.

80. Smit, S., Yarus, M. and Knight, R. (2006) Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA*, **12**, 1-14.