1-1-2014

The Rna Newton Polytope And Learnability Of Energy Parameters

Elmirasadat Forouzmand
Wayne State University,

THE RNA NEWTON POLYTOPE AND LEARNABILITY OF ENERGY PARAMETERS

by

ELMIRASADAT FOROUZMAND

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2014

COMPUTER SCIENCE

Approved By:

_____
Advisor   Date

www.manaraa.com

Table of Contents

List of Figures

Introduction

The RNA world hypothesis suggests that RNA was the main player in the cell in the origin of life, and later it evolved to DNA and different proteins [1, 2]. The fact that RNA conveys the genetic information like DNA and also works as a catalyzer in chemical reactions, similar to proteins and enzymes, supports this hypothesis.

Since the discovery of key regulatory roles of RNA in the cell, RNA related research has earned even more attention [3, 4, 5, 6, 7, 8]. One fundamental aspect of RNA is the folding process, which leads to the secondary structure of RNA. The proved biological significance of RNA secondary structure has cleared the necessity of tool development for RNA structure determination or prediction.

Due to the complexity of experimental methods for RNA structure determination, similar to other experimental fields and measurement processes in the wet lab, computational RNA structure prediction methods have emerged and evolved during the past four decades.

Although the development of different novel methods, based on the thermodynamic features of RNA [9, 10, 11, 12, 13] and machine learning techniques [14, 15, 16], made a noticeable progress in RNA structure prediction, still the accuracy of existing tools is not satisfying.

Chapter one of this thesis reviews some of the known RNA structure prediction algorithms and methods to date. Numerous tools and techniques have been published to address this problem, but here we chose the most novel and superior algorithms, which could change the common perception in their time.

Chapter one contains two main sections. The first section or background includes the motivation of the work with addressing biological role, secondary structure significance, and the essentiality of computational RNA structure prediction.

In methods review, comparative and dynamic programming based algorithms are explained; however, our focus here is the latter. Some of the most popular algorithms are mentioned, and their improvements over their previous ones are justified.

Chapter two starts with addressing the gaps, possible improvements and available areas for work in RNA secondary structure prediction. However, the focus of this part is on the intrinsic limitation of energy models as one of those gaps. Accordingly, a method is introduced that helps to discover the intrinsic limitation of an energy model. This section focuses on the concept of learnability of the parameters of an energy model, which helps to check the capability of the model. The necessary condition for learnability and the dynamic programming algorithm to verify this condition is provided in the rest of this context. Results and conclusions are the last sections of this thesis. The majority of the content of chapter two has been published in [63].

CHAPTER 1

Background

RNA: biological role

Messenger RNAs (mRNAs) might be the most popular family of RNAs, but RNA role in the cell is not summarized to only an intermediate state of the information transition process between DNA and protein [3, 4, 5, 6]. Discovery of non-coding RNAs (ncRNA), which unlike mRNAs are not translated to protein, started a few years ago and by coincidence at first. Everyday a larger number of ncRNAs are identified in different species.

ncRNAs can be categorized based on different features. Their lengths vary between ~22nt for microRNAs (miRNAs), 100nt to 200nt for small RNAs (sRNAs) and to more than 10,000nt for long ncRNAs in evolved eukaryotes [3, 8]. Obviously these different families of ncRNAs function differently.

miRNAs play a significant role in translation process. With binding to mRNAs, they can prevent translation while keeping the mRNA stable in the environment. In this case, ncRNAs regulate gene expression in the cell. In plants, microRNAs usually bind to a perfect complementary strand of mRNA. In animals, miRNA and the target mRNA pairing follows a pattern but it is not as perfect as what happens in plants. More than one third of human genes are expected to be regulated by miRNAs [7].

On the other hand, some ncRNAs known as small interfering RNAs (siRNAs) are responsible for mRNA degradation. These RNAs also control gene expression through a process called RNA interference. siRNAs are small fragments of double stranded RNAs (dsRNAs) that lead a chaperone protein to the target mRNA to silence its expression.

Some non-coding RNAs inhibit the transcription process by binding to the transcription factor. As an example in human body, 7SK RNA binds to P-TEFb and suppresses the transcription. ncRNAs affect the RNA modification process. Some of them control the pre-mRNA splicing and others bind to RNA to modify the methylation. Moreover, it has been shown that the effect of ncRNAs in protein stability and transportation is significant [3]. It is known that the RNA sequence is not the only important piece of information in these scenarios. RNA structure also affects the chemical reactions and pairing processes.

Secondary and tertiary structure

RNA bases have the tendency to pair with each other; this base pairing changes the strand of RNA to a structured molecule. RNA secondary structure is simply the list of base pairs.

Tertiary structure of RNA is the three dimensional shape of RNA molecule and its atoms locations in the space. Different experimental techniques have been developed for tertiary structure determination. Figure 1 is an example of the secondary and tertiary structures.
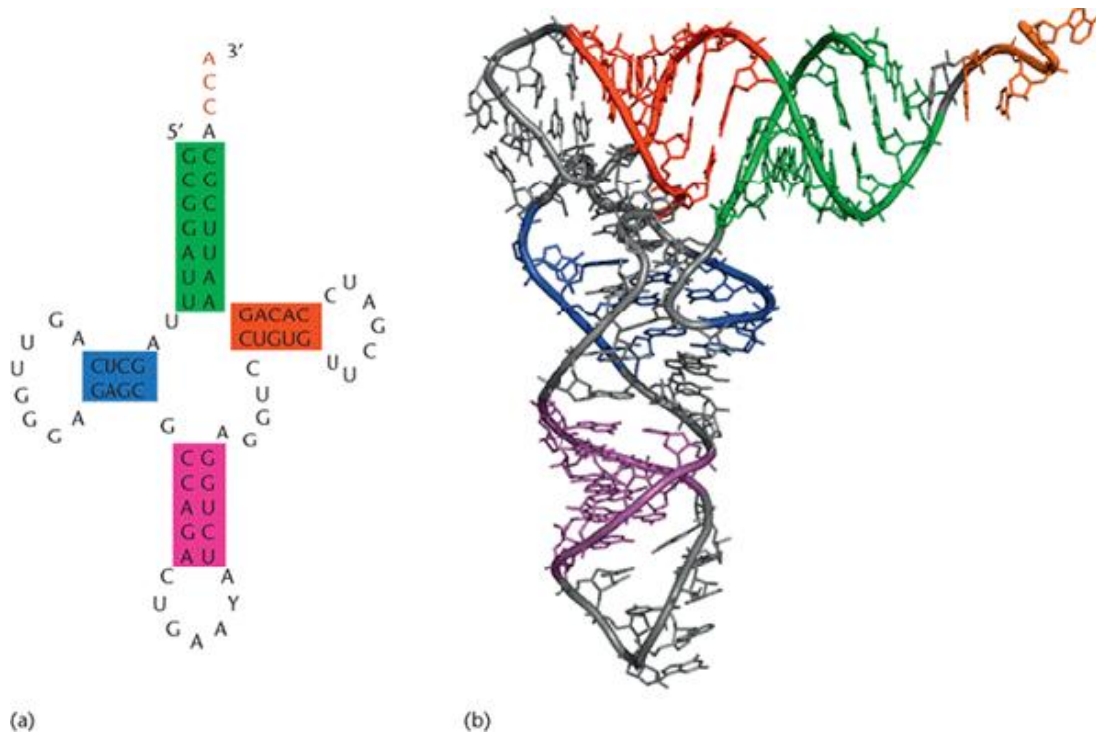
Figure 1 - The secondary and tertiary structures of yeast tRNAPhe. Colors show the corresponding parts of the structures [58].

Clearly, RNA structure affects its functionality. In more accurate words, after pairing some parts are less likely to interact, and some parts have more inclination to play a role in chemical reactions.

RNA structure importance

Identifying RNA structure helps to understand RNA functionality mechanism, which is of importance due to RNA's significant role in biological processes. This information can also be used for synthetic RNA design to fulfill specific roles in a designed environment [20]. The domain of synthetic biology advances every day, and building novel cells and organisms is on the way. However, without complete knowledge of each constituent particle, reaching a perfect design is not possible.

RNA secondary structure can help to interpret the tertiary structure, and provides us with that part of the information, which is required to discover the influence and importance of the structure for RNA.

Also, the energy correspondent to the tertiary structure is less than the energy involved in the secondary structure creation, which means that the secondary structure is more stable and effective [21].

Why computational prediction

Similar to most other experimental methods, determining RNA secondary structure in the wet lab is time consuming and costly. Some of the high accuracy methods are X-ray crystallography, Nuclear Magnetic Resonance (NMR), and Cryo-electron microscopy. There also exist some techniques with lower resolution such as chemical or enzymatic probing, thermal denaturation, mass spectrometry, and RNA engineering [22].

As a consequence, computational methods and algorithms have been developed during last four decades to predict RNA secondary structure. In some cases, the result of experimental process can be given to a computational method as a part of input. Chemical modification techniques [23, 24], which use a special chemical with the ability to interact only with specific types of paired or unpaired nucleotides, are in this group of experiments. These techniques work based on the fact that paired nucleotides are less likely to interact. SHAPE or Selective 2'-Hydroxyl acylation Analyzed by Primer Extension also uses a chemical such as N- methylisotoic anhydride (NMIA), which reacts with the backbone of RNA, and this reaction is more likely in the flexible part of RNA or the single stranded part [25].

Methods Review

RNA secondary structure prediction methods can be categorized in two general groups:

     i.   Comparative methods,

     ii.   Dynamic programming based methods.

However, these two classes of RNA structure prediction techniques have a fair intersection and are not completely separate. One can be augmented by the other one, or help to improve the result of the other one in a pipeline. Here, our concentration is on the second group of algorithms.

Comparative methods

It has been observed that RNA structure is more stable than the sequence through the evolution. Like other strands of genetic information over time, RNA bases change in different ways. Although mutations change the nucleotides, it has been seen that this change happens in a way that the pairing potential of bases stays untouched in most of the cases, for instance C changes to A and G changes to U, so CG converts to AU. These types of sites in RNAs, which are different in strand but the same in pairing pattern are called co-varying sites [18].

Hence, if a set of homologous RNA sequences, which came from different species or even one organism is available, then valuable information for RNA secondary structure of that line of RNA can be extracted from their alignment.

The comparative method is still the most trusted one. For ribosomal RNA, the accuracy of the comparative method is about 97% of predicted pairs [27]. However, one important drawback of this method is that a big set of homologous sequences is necessary to predict the structure of a new member of the group. Additionally, comparative method is a mostly manual one, due to the required human supervision for the alignment step. It is important to notice that this alignment is

not just the sequence alignment, but the structures need to be consistent with the sequence alignment too. Figure 2 shows an example. In this figure, each piece of aligned sequence is correspondent to the piece with the same color in 2D structure.

This picture illustrates how sequence alignment alone can mask some important information [18].
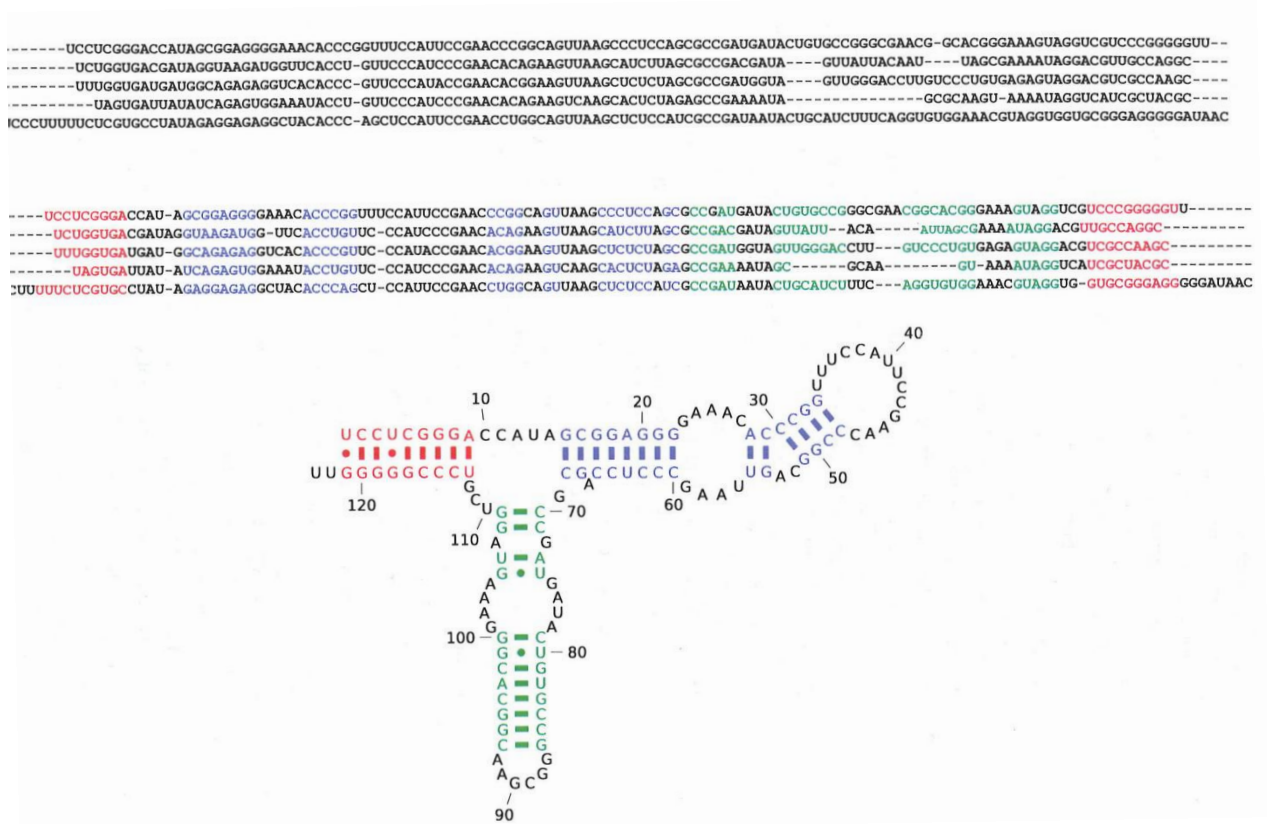
Figure 2 - A sequence alignment [59, 60] by MUSCLE [61], the structural alignment of 5S rRNA sequences [59], and the secondary structure of the first sequence [18].

However, researchers have tried to improve the automation degree of the comparative method [27, 28, 29, 31, 32, 56]. These semi-automated approaches can be classified in three categories, which are shown in Figure 3 [26].
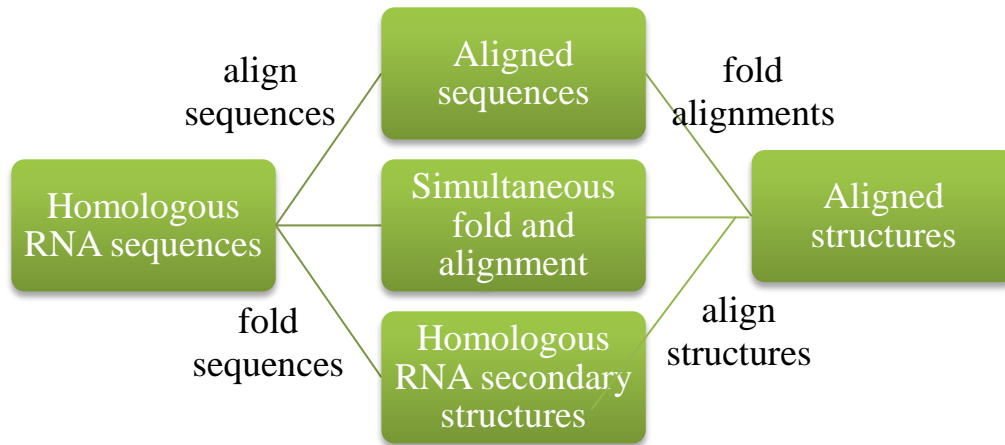
Figure 3 – Automated approaches for comparative prediction of RNA structure [26]

In the first category, the result of multiple sequence alignment is used to find a consensus structure for all the relevant sequences. Generally, the output of sequence alignment provides us with some information about the conserved base pairs, and this information combined with the thermodynamics, for instance in RNAalifold [28] or probabilistic models based on stochastic context-free grammars like in Pfold [29], gives a comprehensive result.

In this approach an initial alignment is required; this is the main weakness of these methods because of the strict dependency of the result quality on the multiple sequence alignment accuracy.

In the second category, alignment and finding the consensus structure for a set of homologous sequences happen at the same time. This family of algorithms is very time consuming (usually time complexity of $O(n^6)$) and needs huge amount of memory ($O(n^4)$). A well-known algorithm for simultaneous alignment and folding is the Sankoff algorithm [30], which has been used in FOLDALIGN [31] and Dynalign [32].

When little conserved is in the sequences, their structures are to be predicted first, and then those structures are aligned, but a method to predict those structures is required. Most methods in the second category are not effective for a novel ncRNA; just a few of them such as evefold [33] and RNAz [34] can be used for genome wide search and prediction of the structure of a new RNA [54]. However even for these methods, the fact that their efficacy is dependent on the initial alignment remains unchanged.

Dynamic programming based methods

For those situations that no set of homologous or relevant sequences are available, development of *de novo* methods is inevitable. For the last few decades, different algorithms have been introduced to address this issue; some of these algorithms are discussed here.

Nussinov method or base pair maximization

The Nussinov's algorithm [11] uses the simple idea of base pairing maximization. Now this way of RNA structure prediction may seem very inefficient and meaningless; however, in 1978 it was a major step forward for computational techniques in this field. That method considers canonical base pairs CG and AU and the wobble base pair GU, and the goal is to find the structure with the maximum possible number of base pairs. To have a biologically meaningful structure as the outcome of this algorithm, some constraints are considered. Based on chemical and biological observation, for the vast majority of cases, each base may be involved with only one other nucleotide in pairing.

Consider $S = \{i.j \mid 1 <= i <= j <= N\}$ is the secondary structure of sequence R with the length of N when $i.j$ are the structure base pairs ($i.j$ means the nucleotide in position i and the one in position j are paired.)

For a problem to be solved using a dynamic programming algorithm, the problem should be divisible to smaller but similar sub-problems, and this fact imposes a limitation: if a nucleotide is inside a loop (an unpaired part of the strand which ends with base pairs forming a double helix), it cannot pair with a base outside that loop, i.e. $if\ i < g < j < h$ or $g < i < h < j$, then $i.j$ and $g.h$ cannot happen at the same time. This situation will be explained more in pseudoknot section. The score of base pairing is shown with $\delta(i, j)$. If there is a possible base pair between i and j, $\delta(i, j) = 1$, and if there is not any base pair there, $\delta(i, j) = 0$.

A straightforward dynamic programming algorithm can be used to find the structure with maximum number of possible base pairs. An $N \times N$ table keeps the folding information of sub-strands. The following recursion provides the result:

$$\gamma(i, j) = \max \begin{cases} \gamma(i+1, j) \\ \gamma(i, j-1) \\ \gamma(i+1, j-1) + \delta(i, j) \\ \max_{i<k<j}[\gamma(i,k) + \gamma(k+1, j)] \end{cases}$$

(1)

Where $\gamma(i, i-1) = 0$ and $\gamma(i, i) = 0$, i.e. there can be no base pair between a nucleotide and its neighbor nor itself.

The first case corresponds to the situation that we know i is not involved in any pair. The second one shows the same thing for j. The third case happens when $i.j$ is a pair. The last case, which is known as bifurcation, considers breaking the structure into two sub-structures, when there is no base pair between g and h if $i < g < k$ and $k + 1 < h < j$. These four cases are also shown in Figure 4.
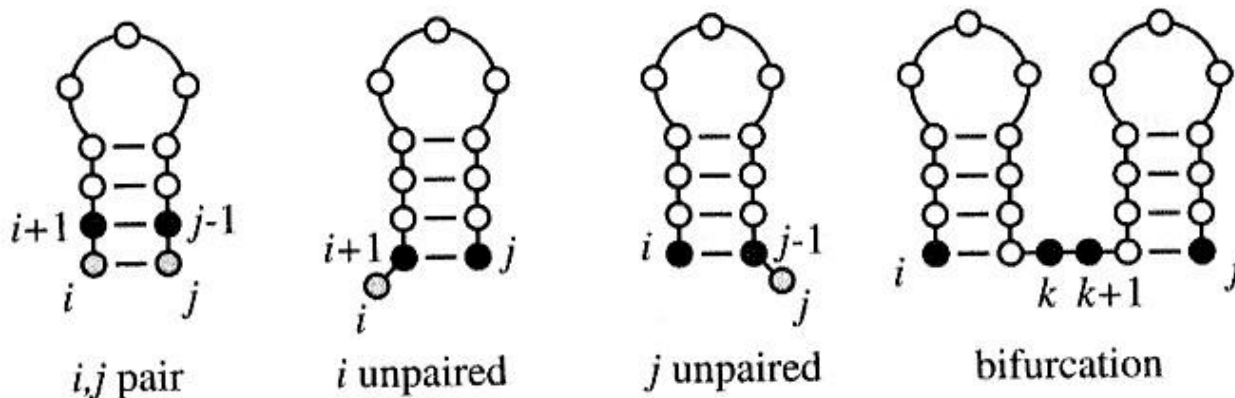
Figure 4 – Four possible cases in the Nussinov's algorithm [17].

Although this algorithm really yields the structure with the maximum possible base pairs, and it satisfies the mentioned limitations, the output structure is not usually biologically relevant in practice. Therefore, the necessity of improving the computational methods based on biological and chemical insight became undeniable.

Minimum Free Energy (MFE)

It was mentioned before that RNA folding which leads to RNA secondary structure is a chain of chemical reactions like base pairing. Similar to any other chemical reaction, the structure of RNA in equilibrium is the one with minimum free energy, in most cases [35].

Gibbs free energy

The Gibbs free energy is that portion of the energy of a system that can do non-mechanical work. The change of the Gibbs free energy, when an unpaired RNA strand converts to its secondary structure, represents the spontaneity of the relevant reactions. The Gibbs free energy is

$$G(T,P) = H - TS, \qquad (2)$$

where T is the temperature and P is the pressure, H shows enthalpy and S represents entropy [35].

In general, when $\Delta(G) <= 0$, the reaction is spontaneous; when $\Delta(G) = 0$, the system is at equilibrium; and when $\Delta(G) >= 0$, the reaction is not spontaneous, where $\Delta(G)$ is the difference between energy after a reaction and before that reaction.

$\Delta(G) <= 0$ means the products of the process are more stable than the reactants, or they are in a lower level of energy. Usually, lower free energy is equivalent to more stability; however, the energy level of RNA can be in a locally minimum point, and still RNA may be very stable.

Energy model

To estimate RNA secondary structure free energy, researchers decompose the structure to a set of sub-structures or building features. The free energy of each sub-structure has been measured in the wet lab using very short strands of RNA which fold into the studied structure. Figure 5 shows an instance of these features. The choice of these building features together with their energies is the energy model. The most popular energy model (Figure 5) is the Turner or Nearest Neighbor energy model [36]. In the Nearest Neighbor model, the free energy is determined based on the base pairs and their close neighbors.
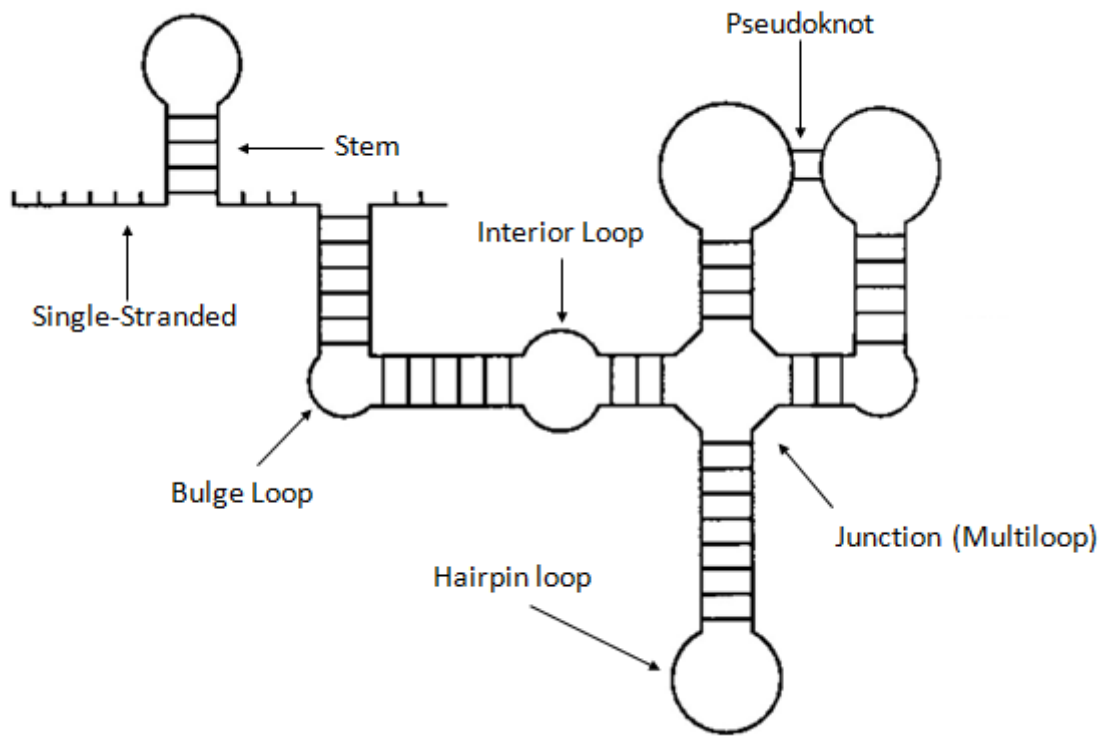
Figure 5 – RNA building blocks in the Turner energy model

The free energy of RNA structure is the sum of the free energies of its sub-structures. That means the structure can be decomposed to its building blocks, and the energies of those blocks are independent. Figure 6 shows an example of how this computation works.
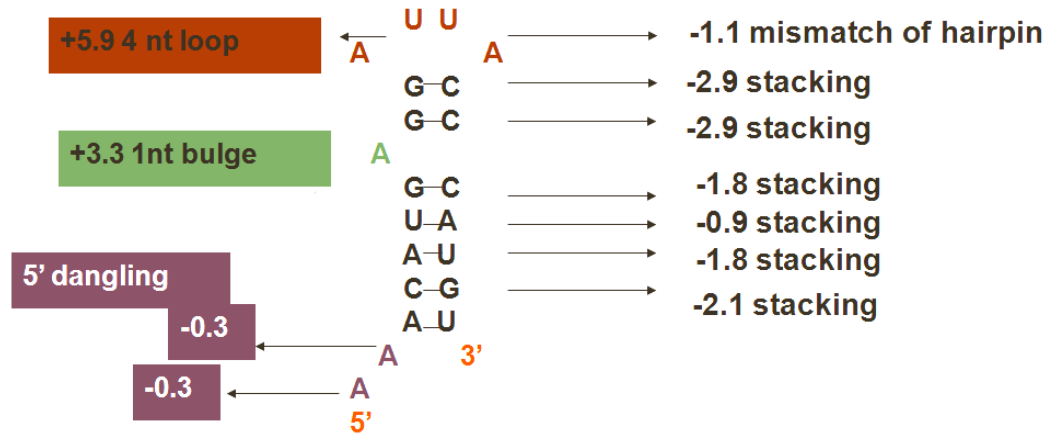
Figure 6 –Different sub-structures have different amount of free energies.

The Zuker's algorithm

One of the first algorithms which considered thermodynamic characteristics of different features of RNA secondary structure was the Zuker's algorithm [37, 38]. This algorithm was a dynamic programming solution with running time complexity of $O(n^4)$ first, and then it was improved to a version with the complexity of $O(n^3)$.

Zuker considered hairpin loops, stacked pairs (stems), internal loops, external bases or single stranded parts, and multi-loops, using nearest neighbor model. The Zuker's algorithm uses two tables $w(i,j)$ and $v(i,j)$ and pre-specified free energy for different sub-structures. $w(i,j)$ keeps the minimum free energy of all feasible structures for sub-sequence (i,j), and $v(i,j)$ is the minimum free energy of all the possible structures for sub-sequence (i,j) where $i.j$ is a base pair. Free energy relations for different features are specified below.

Hairpin loop: $eH(i,j)$

Multi-loop: $VM(i,j) = \min_{i<i_1<j_1<\cdots<i_k<j_k<j}\{eM(i,i_1,j_1,\ldots,i_k,j_k,j) + \sum_{l=1}^{k} V(i_l,j_l)\}$

Bulge or Internal loop: $VBI(i,j) = \min_{i<i'<j'<j \,\&\, i'-i+j-j'>2}\{eL(i,j,i',j') + V(i',j')\}$

Here, eM is the direct cost of multi loop, and eL is the cost of an internal loop.

This method is the basis of Mfold [39] and RNAfold (in Vienna package) [40] tools for RNA secondary structure prediction.

The Zuker's algorithm is a pioneer method, and like any other first, there are some drawbacks which kept the way open for other novel ideas to improve this field.

First, the Nearest Neighbor energy model is not a perfect model because the energy of each building feature of RNA is not dependent only on its closest neighbor in reality. Generally, sequence is not the only factor in RNA folding. The cell environment, other particles or chemical processes can affect the structure too.

Second, RNA is not always in its equilibrium state, and for some RNAs, such as riboswitches and tRNAs, more than one secondary structure have been observed [41, 42].

Third, due to the nesting characteristic of our RNA models that is essential for a dynamic programming algorithm, some features cannot be considered simply. One of the most significant and challenging ones is pseudoknot. The Zuker's algorithm could not consider this feature.

Suboptimal structures

The first approach to address the fact that the optimal structure may not be unique was sub-optimal structure prediction [37, 39]. Zuker *et al.* [37] suggested using specific biological observation, as prior knowledge for the algorithm, to improve the prediction. Then the structures in the range of five or ten percent of the minimum free energy are chosen and evaluated biologically to find the sub-optimal structures.

Partition function

The second and more efficient way is the utilization of partition function, besides the minimum free energy. The accuracy of the Zuker's and similar algorithms are limited. It means some of the predicted pairs exist and the rest of them have been predicted incorrectly. Partition function calculation, which provides the likelihood of correctness of a base pair, enhances the accuracy of prediction.

The equilibrium constant of a chemical reaction of $A \rightarrow B$ is calculated as below.

$$K = \frac{[B]}{[A]} \qquad (3)$$

Here, $[A]$ and $[B]$ are the concentration of A and B in the environment at equilibrium state.

For structure s_i of strand x of of RNA, $k_i = \frac{[s_i]}{[Unpaird\ RNA]} = e^{-G_i/RT}$; when $s_i \in \varepsilon(x)$ is a possible structure for the strand, $\varepsilon(x)$ is the set of all possible structures for x, and G_i is the energy level difference between s_i and the unpaired state of that RNA . T shows the temperature, and R is the gas constant.

Sum of these constants for all possible structures of one strand of RNA is the partition function.

$$Q = \sum_{s_i \in \varepsilon(x)} k_i = \sum_{s_i \in \varepsilon(x)} \exp\left(-\frac{G_i}{RT}\right) \qquad (4)$$

The probability of a specific feature, like a base pair, to happen is the sum of equilibrium constants of structures containing that feature, divided by the partition function. Those most probable base pairs, identified this way, are the ones more likely to be part of experimentally observed structure.

In 1990, McCaskill proposed a dynamic programming algorithm for partition function calculation [13]. This algorithm, which works with the time complexity of $O(n^3)$, is explained in Figure 7.
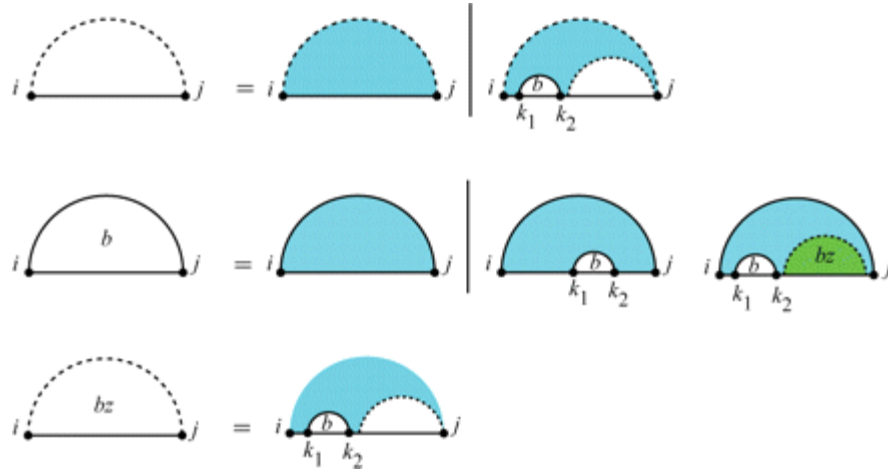
Today, the recursion diagrams used in this figure are the standard way for RNA structure prediction dynamic programming algorithm [9, 10].

Partition function integrated with free energy minimization, which helps to specify the more probable features, has been implemented in RNAstructure [51] and Vienna package [40].

Pseudoknots

Pseudoknots are one type of those features that do not follow the nesting characteristic of RNA. Very little thermodynamic information about pseudoknots exists, and this information cannot be easily measured experimentally. Different sets of parameters for pseudoknots are available based on polymer model and lattice model [43, 44].

As mentioned in the Zuker's algorithm, pseudoknot happens if both $i.j$ and $g.h$ pairs occur in the structure while $i < g < j < h$. Figure 8 shows examples of pseudoknots in different RNA.
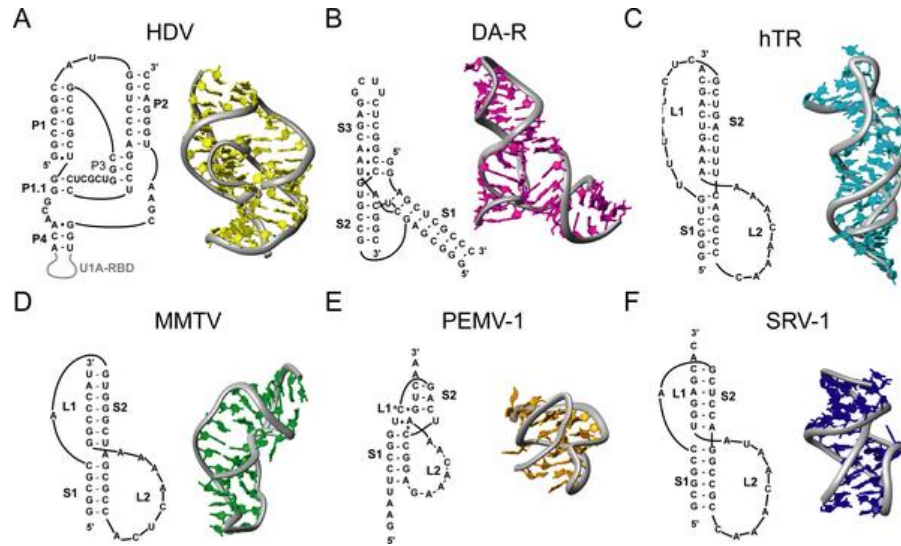
Figure 8- Pseudoknots Examples [55].

Due to this complexity as a consequence of the difference between pseudoknots and simpler features like hairpin loops, pseudoknots were not considered in several researches for RNA structure prediction at first. In fact, it has been shown that RNA structure prediction with pseudoknots using the Nearest Neighbor model is NP-hard [57]. But their existence in ribosomal RNA, ribozymes and viral RNA made it necessary to develop tools for predicting structures including pseudoknots [10].

In 1999, Rivas and Eddy addressed this gap in their paper and presented a dynamic programming for it [9]. Before that, some methods based on maximum weighted matching (MWM) [45] were introduced [46, 47]. In general, MWM builds a graph with nucleotides of RNA as the vertices. In this graph, edges are the pairing relations between two bases, and each edge has a weight. For the best outcome, the weight of an edge between two vertices can be computed using mutual information between the correspondent positions in a multiple sequence

alignment of homologous sequences. The goal is to find the set of non-conflicting base pairs, which have the highest sum of associated weights [46].

That algorithm is time and space efficient with the time complexity of $O(n^3)$, and it produces acceptable outcome; however, MWM needs a pre-alignment of sequences for the best result [9, 46]. Without this pre-alignment, MWM on a single sequence is essentially base pair maximization, which is not biologically accurate.

Hence, a technique to find the secondary structure of RNA when only one strand of RNA is available was needed.

Figure 9- General recursion for vx in right and wx in left [9].

Rivas and her collaborator used the Turner energy model as the basis of their model, but some new parameters correspondent to pseudoknots were used to boost the model. Similar to the Zuker's algorithm, they considered two matrices $wx(i,j)$ and $vx(i,j)$. $wx(i,j)$ keeps the recursion score for the strand i to j, in those situations where the relation between i and j is not determined. $vx(i,j)$ is the score for sub-sequence between i and j, when they are paired. Figure 9 illustrates the dynamic programming recursions of their algorithm without pseudoknots.

If IS or irreducible surface is a loop that cannot be decomposed into smaller ones anymore, $EIS^n(i_1, j_1; i_2, j_2; \dots; i_n, j_n)$ shows the score for an IS with the order of n, where i_k and j_k are paired. The order shows the number of secondary interaction inside a surface. Hairpins, bulges, stems and internal loops are ISs with the order of two. Multiloops which have larger order than two have an approximate score.

$$vx(i,j) = optimal \begin{cases} EIS^1(i,j) \\ EIS^2(i,j:k,l) + vx(k,l) \qquad\qquad i \leq k \leq l \leq \\ P_I + M + wx_I(i+1,k) + wx_I(k+1,j-1): multi\ loop \end{cases}$$

$$j \ (5)$$

In this relation, P_I represents the closing base pair score in a multi loop, M is the general score of multi loop. wx_I is the score corresponding to the loops inside a multi loop. wx and wx_I have the same recursion, but one of them happens inside a base pair.

$$wx(i,j) = optimal \begin{cases} P + vx(i,j): paired \\ Q + wx(i+1,j)\ or\ Q + wx(i,j-1): single\ stranded \\ wx(i,k) + wx(k+1,j) \quad i \leq k \leq j: bifurcation \end{cases} (6)$$

P is the penalty for an external base-pair, and Q stand for the single stranded nucleotide.

To add pseudoknots, they defined two new and more general matrices, gap matrices or matrices with a hole, $whx(i,j,k,l)$ and $hx(i,j,k,l)$. Figure 10 shows how a pseudoknot can be described by two hole matrices.
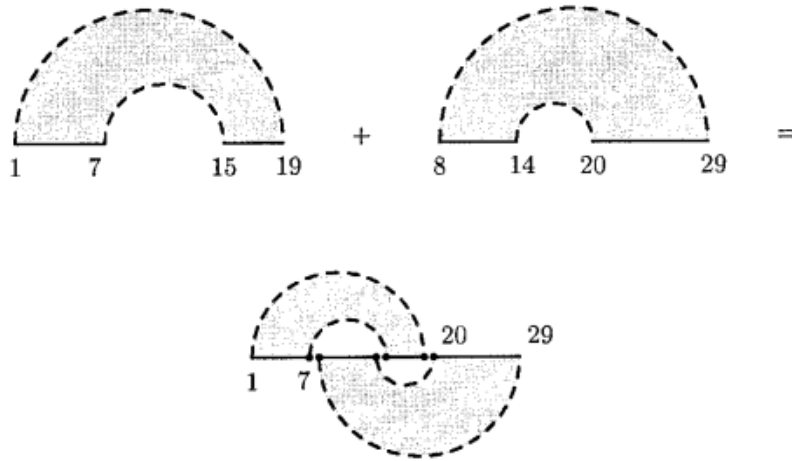
Figure 10 - Pseudoknots with two gap matrices [9].

In hx, there is a base pair between i and j and also k and l. For hx, the relation of i and j, and l and k is not known. They also introduced yhx for the situation in which there is a pair between k and l, but the relation between i and j is undetermined. zhx shows the reverse case.

Clearly, wx and vx are a specific version of the gap matrices. The point here is the augmentation of these matrices into the dynamic programming. For this purpose, another situation which shows the pseudoknots can be added to v and x. Diagrams in Figure 11 show the recursion for v and w including the pseudoknots. This algorithm has a worst-case complexity of $O(n^6)$.
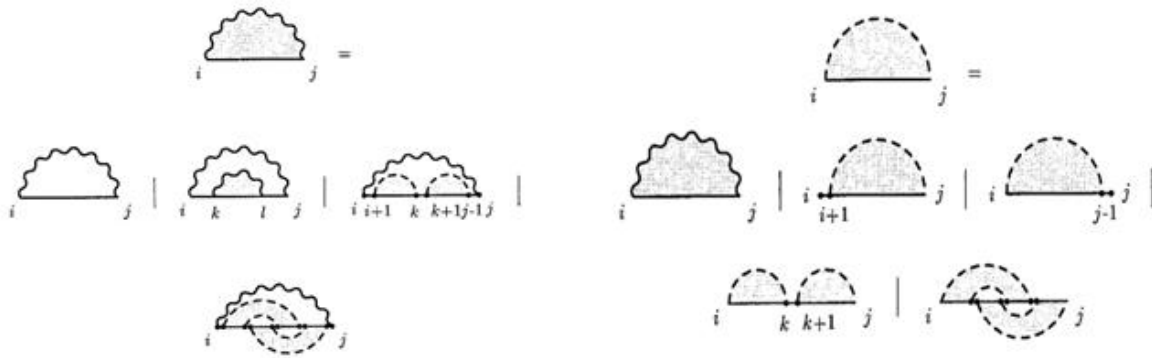
Figure 11 - Recursion for vx in right and wx in left with pseudoknot [9].

In this method, just a limited type of pseudoknots are considered, and in those cases that a knot needs more than two gap matrices to be described, or in other words when for the presentation of a pseudoknot on the paper, base pair lines cross each other, the problem is not solvable by this algorithm.

In 2003, Dirks and Pierce suggested a new partition function based algorithm for pseudoknotted RNA structure prediction by dynamic programming [10]. They considered the concept of gap matrices with more details and possible situations. Their basic algorithm had the time order of $O(n^8)$ but using a function called fastil-loop for interior loops they improved the complexity to $O(n^5)$.

One other difference between their dynamic programming and the one by Rivas *et al.* [9] is that in the recursion, they consider the right most base pair inside a surface. This small change helps to avoid redundancy and generating the same combination of features several times.

They defined different tables including: Q, Q_b, Q_m, Q_p and Q_z. $Q(i,j)$ is the table correspondent to general situation, when the relation between i and j is not determined. $Q_b(i,j)$ keeps the score for the sub-sequence $[i,j]$, when i and j are known to be paired. For sub-

sequence $[i, j]$ inside a multi loop when there is at least one base pair or pseudoknot in this interval, the partition function is kept in $Q_m(i, j)$. $Q_p(i, j)$ on the other hand conveys the penalty of a pseudoknot filling the interval between i and j. Q_z has the same recursion as Q, with the difference that Q_z presents the partition function for the strand inside a pseudoknot. Next figure demonstrates these recursions.
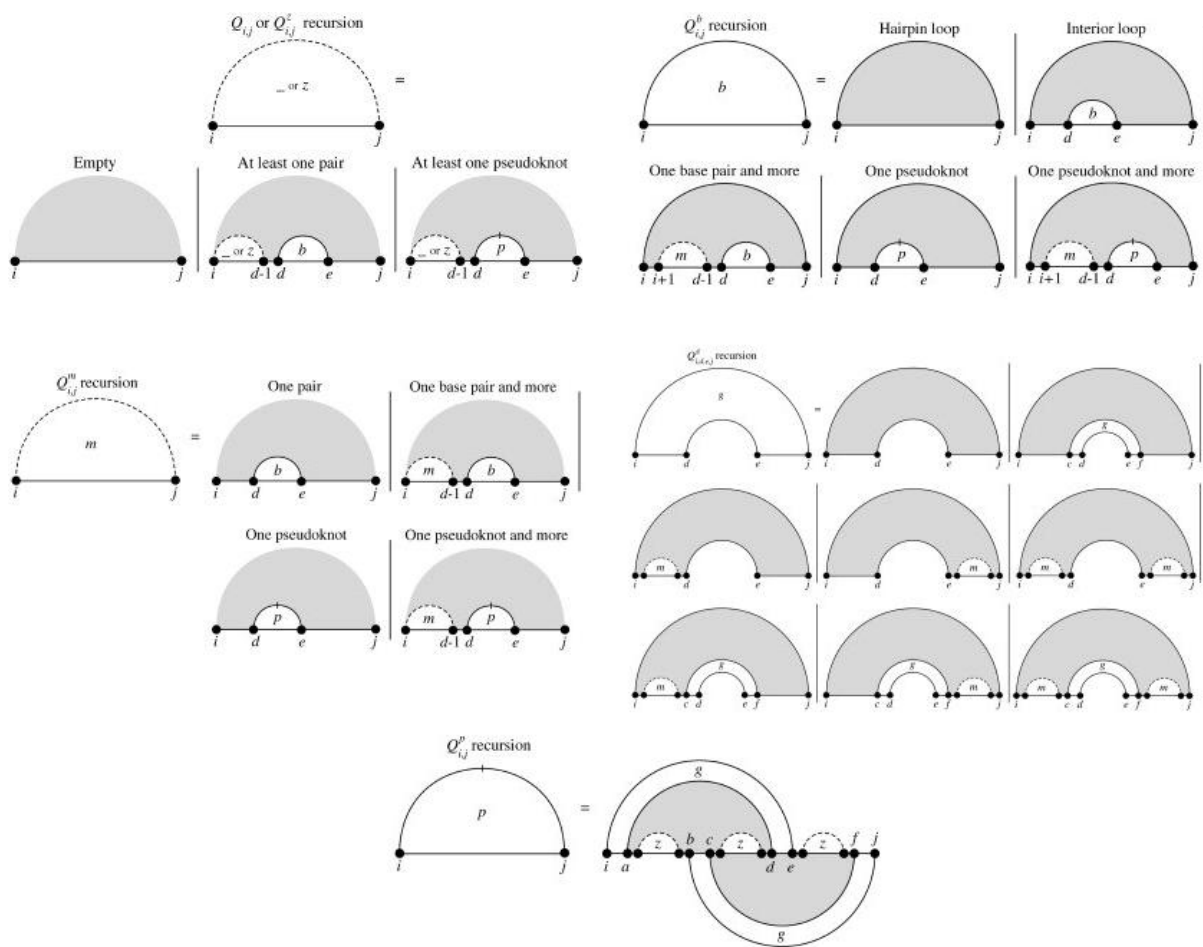
Figure 12 – Recursions of partition function tables in [10].

As mentioned before, the chemical and thermodynamic features of pseudoknots have not been determined by experiment. Also, although the Dirks *et al.* [10] algorithm is relatively comprehensive, still the types of knots that they consider are limited due to the increase in the

complexity of algorithm. Hence, the necessity of generating new scoring sets and methods was felt.

Stochastic context free grammar

Some of the RNA secondary structure algorithms function based on stochastic context free grammars [48, 49, 50]. In this family of algorithms, there are two main parts [15, 49]:

First, a set of transformation rules. One example is:

$$S \rightarrow aSu|uSa|cSg|gSc|gSu|uSg|aS|cS|gS|uS|\epsilon$$

Different rules stand for different features. For instance, $S \rightarrow aSu|uSa|cSg|gSc|$ represents the rules for canonical base pair generation.

Second, a probability value, which is associated to each rule. For example the rule $S \rightarrow aSu$ is likely with the chance of $p_{S \rightarrow aSu}$.

The set of transformation rules that produces the sequence with the highest probability provides the structure. If there is sequence $x = cgaug$ with the structure $y = ((.))$. In this representation, the matching pair of parentheses shows a base pair. For this sequence we have the parse σ :

$$S \rightarrow cSg \rightarrow cgSug \rightarrow cgaSug \rightarrow cgaug$$

Therefore, the joint probability of these rules is $P(x, \sigma) = p_{S \rightarrow cSg} \cdot p_{S \rightarrow gSu} \cdot p_{S \rightarrow aS} \cdot p_{S \rightarrow \epsilon}$.

These probability parameters can be learned and optimized for different sets of rules and input. If the data set is a set of RNA sequences x_is and their observed structures y_is, and $\theta = \{\theta_1, \theta_2, \dots, \theta_n\}$ represents the probability values for different features such as different loops,

θ_is have to be chosen such that they maximize the joint likelihood $\prod_i p(x_i, y_i; \theta)$ of the training set of sequences and structures [15].

These methods do not have a high rate of accuracy in general. Dowell and Eddy reported the accuracy of around 55% for their SCFG based methods using different grammars on different sets of RNA [49]. Mfold [39] and RNAstructure [51] algorithms, which work based on the Zuker's method, have more than 70% accuracy.

Contrafold

RNA structure prediction methods based on MFE have evolved during the past decades, but still some intrinsic characteristics of the minimum free energy technique keep the accuracy of this method limited.

Contrafold [15] uses a similar concept as SCFG, but it considers more expressive features than usual SCFG based methods. That algorithm works based on conditional log-linear models (CLLMs). The feature set can be shown by $F = \{f_1, f_2, ..., f_n\}$, and each feature has a correspondent weight w_i. If x is the sequence and σ is one of its parsings to the structure y, we have $p(x, \sigma) = \exp(w^T F(x, \sigma))$, which is the joint likelihood of sequence x and the parse σ in a log linear form.

To learn the parameters, the algorithm maximizes the conditional likelihood of the structures or $\prod_i p(y_i|x_i; w)$ and not the joint likelihood. Discriminative or conditional likelihood is superior to joint likelihood in this case since it generates the best set of weights without modeling the input distribution.

Contrafold does not consider penalty for special hairpin loops like the loop with special type of closing base pair and avoids generating different sizes of tables for internal loops to prevent over fitting. It keeps a set of more efficient features but limits the number of them.

There is another feature in Contrafold which is worth mentioning in this context. In the dynamic programming process of the algorithm, there is a coefficient γ which helps to adjust the confidence level of the method about its prediction.

Assume y' is a candidate structure and y is the true structure, the $accuracy_\gamma(y', y)$ is the number of correctly predicted single nucleotides plus γ times the number of correctly predicted base pairs. The goal is to maximize the expected value of this accuracy over different structures of a sequence. If p_{ij} represents the conditional probability to have the pair $i.j$, and q_i is the probability to have an unpaired nucleotide in the ith place of the sequence, the following recursion holds to compute $M_{1,L} = max_y(E_y[accuracy_\gamma(y', y)])$ in which L is the length of the strand.

$$M_{i,j} = max \begin{cases} q_i : if\ i = j \\ q_i + M_{i+1,j} : if\ i < j \\ q_j + M_{i,j-1} : if\ i < j \\ \gamma.2p_{ij} + M_{i+1,j-1} : if\ i + 2 \leq j \\ M_{i,k} + M_{k+1,j} : if\ i \leq k < j \end{cases} \quad (7)$$

They use the concept of maximum expected accuracy here. To find the optimal structure one can trace back this recursion. Clearly for larger value of γ, algorithm predicts more base pairs and for the smaller value, it considers more probable base pairs.

Contrafold algorithm is one of the most accurate ones in the field with the accuracy rate of $\sim 75\%$, and can be called the state-of-art algorithm.

One problem with Contrafold is that it is slow, and this is a challenge specially in the case of having large training set. Another drawback is that it does not consider any error or noise in the input, for instance the structure y may not be the minimum free energy structure for x since the feature set is not perfect. In fact, Contrafold may ignore the chemical and thermodynamics observations.

In [52], the authors mentioned these problems and suggested a constraint generation method for parameter estimation, which considers both feasibility of the predicted structures and the thermodynamic data. The structure can be found by finding the solution of a series of constraints. They reported 7% higher accuracy than the standard Turner model parameters and 5% better accuracy than Contrafold in large data sets.

Contextfold

In 2011, Zakov *et al.* published a paper on rich parameterization for RNA structure prediction [14]. They analyzed the effect of increasing the amount of information that different structure prediction models use and showed that more comprehensive and elaborated models enhance the accuracy of prediction. Their proposed model has 70,000 different features, but still the running time is manageable. They showed that their algorithm can predict the RNA structure by an accuracy of ~85%.

They defined two different categories of features: binary features and real-valued features. For binary features the occurrence value is 1 if it happens in the sequence and 0 otherwise. For real-valued features, the occurrence value can be a function of the length of the sequence of that feature. Representing these occurrence values by φ_is and considering w_is as the correspondent weights, we have:

$$G(x, y) = \sum \varphi_i w_i = \emptyset(x, y)^T . W \qquad (8)$$

Here, $G(x, y)$ is the score of sequence x and structure y, and the aim of the algorithm is to find W such that it minimizes the expected cost of having y from x. To train the system, they use a discriminative structure prediction learning algorithm based on the Collins work in [53]. These types of algorithms, which can work with a large data set, are common for natural language settings.

CHAPTER 2

The RNA Newton Polytope and Learnability of energy parameters

Various components of common tools for RNA structure prediction can be studied more, but the main aspect of these methods is their dependency on a thermodynamic based set of features or energy model. In general, the explained algorithms tried to expand the model or develop more capable parameter estimation methods; still the progress does not meet the expectation.

The first weakness of these tools is pseudoknots. Pseudoknots and other not-nested blocks of RNA structure still cannot be considered properly, with all of their details. As mentioned, existing algorithms simplify the problem and consider some special cases of them and not the general features yet. Some of the most accurate tools like Contrafold exclude pseudoknots from their models.

The second issue is the running time. Since RNA structure prediction methods, which estimate the parameters, need a large set of training data to generate an acceptable outcome, these tools are usually slow. Also running a not learning-based minimum free energy tool for a big set of RNA is time consuming. Improving the complexity of these algorithms without compromising the accuracy is necessary. One possible solution is to use approximation of partition function as a substitute of the exact value.

In [62], we explored this possibility and introduced an algorithm to compute the upper bound of partition function. The final goal of this work is to have a fast and efficient algorithm for the upper bound and lower band approximation of partition function, using sparse folding.

Still the main problem of RNA secondary structure tools is the limited accuracy. Contrafold and Contextfold as the best tools to date, consider a large set of features, and train the system to find the best set of correspondent parameters to these features, but the question here is why they cannot exceed this level of accuracy.

We believe that the conventional energy models may not have the intrinsic capability of predicting the RNA structure with higher accuracy. Hence, this potential, which shows the suitability of an energy model, should be measured or analyzed.

The rest of this thesis addresses this problem. We have defined the concept of learnability for the parameters of an energy model. We say that "the parameters of an energy model are learnable if and only if there exists at least one set of such parameters that renders every known RNA structure to date, the minimum free energy structure." In this work the notion of Newton polytope has been used to explain the necessary condition for an energy model to be learnable [63].

In most of the methods reviewed here, there is a set of alphabets or rules and a scoring function. The goal is to find the word with optimal score, and this word is correspondent to a secondary structure. For instance in Contextfold [14], the free energy is:

$$G(x,y) = \sum \varphi_i w_i = \emptyset(x,y)^T . W \qquad (9)$$

In which W is the energy model parameters, and $\emptyset(x,y)$ is the feature vector. For the sake of coordination, we use h as the notion of energy model parameters, and $c(x,s) \in Z^k$ shows the feature vector, from this point. Clearly, k denotes the number of different rules or alphabets in the model.

$$G(x, s, h) = \langle c(x, s), h \rangle \qquad (10)$$

Where $s \in \varepsilon(x)$, and $\varepsilon(x)$ is the set of all possible structures of x.

Hence, we are looking for h^* such that for every strand and its secondary structure (x, y) :

$$y = min_s G(x, s, h^*) \qquad (11)$$

Existence of such h^* means 100% accuracy is achievable, and we call this energy model a learnable one; however, such h^* may not exist. We introduce the necessary condition for existence of this h^* and a dynamic programming algorithm for its verification.

Methods

Necessary condition for Learnability

Assume y is the structure which minimizes the free energy function G. Furthermore, we have:

$$G(x, y, h^*) \le G(x, s, h^*), \forall s \in \varepsilon(s) \qquad (12)$$

If we replace $G(x, s, h) = \langle c(x, s), h \rangle$ here,

$$\langle c(x, y), h^* \rangle \le \langle c(x, s), h^* \rangle \qquad (13)$$

$$0 \le \langle c(x, s) - c(x, y), h^* \rangle \qquad (14)$$

We can write

$$0 \le \langle F(x) - c(x, y), h^* \rangle \qquad (15)$$

Where $\digamma(x)$ is the feature ensemble of sequence x or $\digamma(x) = \{c(x, s) | s \in \varepsilon(x)\}$.

The convex hull of $\digamma(x)$ is what we call the Newton polytope of x.

$$N(x) = conv\ \{\digamma(x)\}. \qquad (16)$$

The above relations imply that $c(x, y) \in N(x)$. In other words, $c(x, y)$ places on the boundary of the convex hull of $\digamma(x)$.

Proof. Let's assume $c(x, y)$ does not lie on the boundary of $N(x)$, i.e. $c(x, y)$ is inside the Polytope. It means $\exists \delta > 0$ such that there is a sphere centered at $c(x, y)$ with the radius of δ, which completely places inside $N(x)$. If this sphere shown by $B_\delta(c(x, y))$, then

$$B_\delta(c(x, y)) \subset N(x) \qquad (17)$$

Clearly, $p = c(x, y) - \left(\frac{\delta}{2}\right)(h^*/||h^*||) \in B_\delta(c(x, y)) \subset N(x)$, and

$$\langle p - c(x, y), h^* \rangle = -\left(\frac{\delta}{2}\right)||h^*|| < 0 \qquad (18)$$

As a result, p is a linear combination of the feature vectors in $\digamma(x) = \{v_1, \dots, v_N\}$.

$$\exists\ \alpha_1, \dots, \alpha_N : \alpha_1 v_1 + \dots + \alpha_N v_N = p, \qquad (19)$$

$$\alpha_1 + \dots + \alpha_N = 1. \qquad (20)$$

Hence, at least for one $v_{i, 1 \leq i \leq N}$

$$\langle v_i - c(x, y), h^* \rangle < 0 \qquad (21)$$

But this is in contradiction with (15).

Hence, a necessary condition for existence of h^* is that the feature vector $c(x, y)$ lies on the boundary of $\mathbb{N}(x)$ the Newton polytope of x, where structure y minimizes the free energy of strand x, $G(x, s, h^*)$.

Newton Polytope

In wet lab, different thermodynamic features of RNA are measured, and one of those is melting curve. Melting curve analysis helps to improve the estimation of energy parameters, and partition function plays the role of relating the measurement and energy parameters [14].

Let $c(x, s) = (c_1(x, s), c_2(x, s), \dots, c_k(x, s))$ and $\mathbf{h} = (h_1, h_2, \dots, h_k)$, if we replace $G(x, s, \mathbf{h}) = \langle c(x, s), \mathbf{h} \rangle$ in the partition function

$$Q(x, h) = \sum_{s \in \varepsilon(x)} e^{-\langle c(x,s), \mathbf{h} \rangle / RT} \qquad (22)$$

We define Z_is as

$$Z_i = e^{-h_i/RT}, 1 \leq i \leq k \qquad (23)$$

Then, the partition function is in a polynomial form

$$Q(x, Z) = \sum_{s \in \varepsilon(x)} Z^{c(x,s)} \qquad (24)$$

The Newton polytope of a polynomial is the convex hull of its monomials power vectors. Therefore, the relation between the melting curve measurement and energy parameters turns into a set of polynomial equations, and computing their Newton polytopes is a way to solve them.

$$Newton\{Q(x, Z)\} = conv(\{c(x, s) | s \in \varepsilon(x)\}) = \mathbb{N}(x) \qquad (25)$$

Hence, the Newton polytope name is used here.

The next relations stand for two polynomials f and g,

$$Newton(fg) = Newton(f) \oplus Newton(g) \qquad (26)$$

$$Newton(f + g) = conv\{Newton(f) \cup Newton(g)\} \qquad (27)$$

Minkowski sum of two polytopes, denoted by $\oplus$ [64], is defined by

$$P \oplus Q = \{p + q | p \in P \text{ and } q \in Q\}. \qquad (28)$$

Dynamic programming algorithm

A dynamic programming algorithm needs to be defined to compute the Newton polytope. With the polytope available, we can check if the feature vector lies on the boundary.

For strand x of length L, we denote the ith nucleotide by n_i and the subsequence between ith and jth nucleotides by $n_i \dots n_j$. The Newton polytope of this subsequence is denoted by $\N(i,j) = \N(n_i \dots n_j)$.

The same dynamic programming used for calculating partition function in [10, 13, 14] can be transformed to a divide and conquer strategy for Newton polytope computation. Fig.7 illustrates the details of the recursions in partition function calculation; however, for the case of Newton polytope, the below transformations are required.

Partition Function	Newton Polytope
Multiplication	Minkowski sum
Summation	Convex hull of union
$e^{-\langle c(x,s),h\rangle/RT}$	$c(x,s)$

Here, we consider A-U, C-G, and G-U base pair counting energy model. These are the same features as those ones that Nussinov considered in [11]. The three dimensional feature vector is

$$c(x,s) = (c_1(x,s), c_2(x,s), c_3(x,s))$$

Where $c_1(x,s)$ is the number of A-U base pairs, $c_2(x,s)$ is the number of C-G, and $c_3(x,s)$ is the number of G-U base pairs in secondary structure s. Clearly, any energy model with more features can be treated similarly using above transformations.

The following dynamic programming produces the result we need.

$$\mathbb{N}(i,j) = conv\left[\cup \begin{cases} \mathbb{N}(i,l) \oplus \mathbb{N}(l+1,j), i \le l \le j-1 \\ \{[1,0,0]\} \oplus \mathbb{N}(i+1,j-1), if\ n_i n_j = AU|UA \\ \{[0,1,0]\} \oplus \mathbb{N}(i+1,j-1), if\ n_i n_j = CG|GC \\ \{[0,0,1]\} \oplus \mathbb{N}(i+1,j-1), if\ n_i n_j = GU|UG \end{cases}\right] \quad (29)$$

The base situation is $\mathbb{N}(i,i) = \{(0,0,0)\}$. First, the Newton polytope is calculated for the subsequences with the length of one, after that for the subsequences with the length of two, and it continues to the whole sequence of x.

This strategy provides us with the Newton polytope of x, $\aleph(x)$. Also, $c(x, y)$, which is the feature vector of experimentally determined structure, is available for different set of RNAs. Therefore, the problem is reduced to check if $c(x, y) \in \aleph(x)$, i.e. $c(x, y)$ places on the boundary of $\aleph(x)$.

Implementation

The proposed dynamic programming algorithm for computation of the Newton polytope has been implemented in MATLAB. Also, other related codes, which help in analysis of the result, are written in MATLAB. MATLAB has its own convex hull function, which works with one of the fastest algorithm for convex hull computation, Quick hull [66]. The Minkowski sum of two polytopes was simply implemented as the pair wise summation of vertices of those two polytopes.

It is important to note that there are two common ways to represent a polytope, and each approach has its own advantages. A polytope can be represented by its vertices, i.e. as a set of points. Also a polytope can be defined by a set of inequalities or its half planes. The former or the vertex representation, which is used here, is more convenient for the Minkowski sum calculation, but half plane representation is more efficient for convex hull of union. The most complex part in this method is the convex hull computation, which makes the worst case complexity of our algorithm exponential.

To check if $c(x, y)$ lies on the boundary of $\aleph(x)$, $r(x)$ the distance between $c(x, y)$ and the planes (or edges) that build the boundary of $\aleph(x)$ is calculated. In some cases for this calculation, function 'p-poly-dist' has been used [67]. Clearly, $r(x) = 0$ means that the feature

vector places on the boundary, and the necessary condition for learnability is satisfied. In the case that $c(x, y)$ is inside the Newton polytope, this calculated distance is positive.

As input, 2300 unpseudoknotted RNA sequences and their experimentally determined structures from RNA STRAND v2.0 database have been used (65). The lengths of those sequences vary from 4 nt to ~1000 nt. The wide range of RNA lengths in this data set makes it proper for our application. The implemented program ran on 2.5 GHz 12 Core AMD Opteron CPU.

Results

After computing the Newton polytope for each strand and extracting their feature vectors from experimentally observed structures, $r(x)$ the distance between them is calculated. Besides $\aleph(x)$ for the three dimensional energy model, Newton polygon for a two dimensional model, correspondent to A-U and C-G pairs, is also calculated.

Figures 13, 14, and 15 demonstrate the Newton polytopes from the 3D model and the Newton polygons from the 2D model for three different RNA strands. The first RNA is a ribosomal RNA with 116 nt. Using the 2D energy model, the distance between boundaries of polytope and the feature vector is 10; however, in three dimensional model $c(x, y)$ gets closer to the Newton polygon and $r(x) = 2$.

The point is NOT on the surface
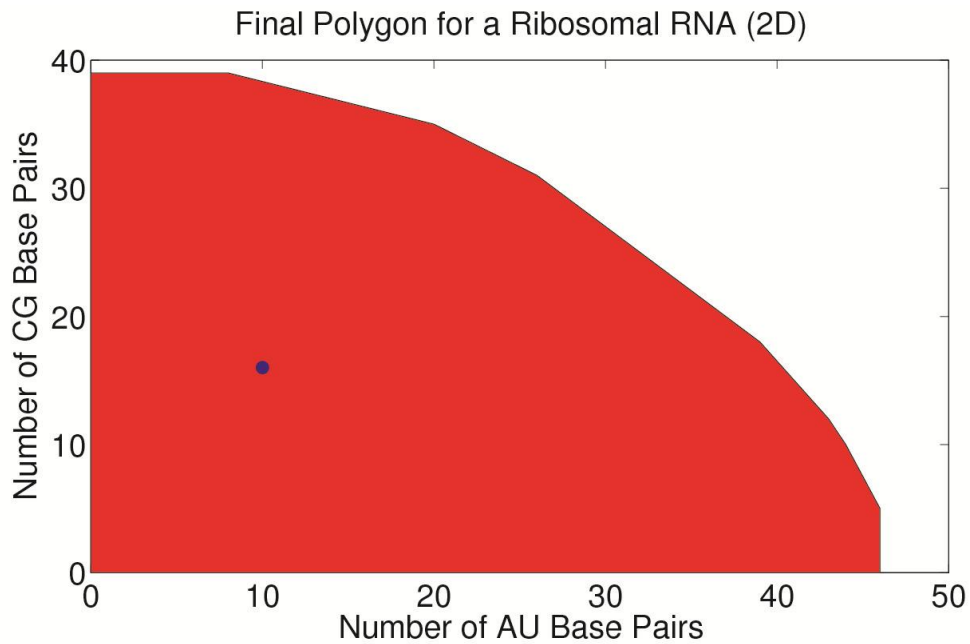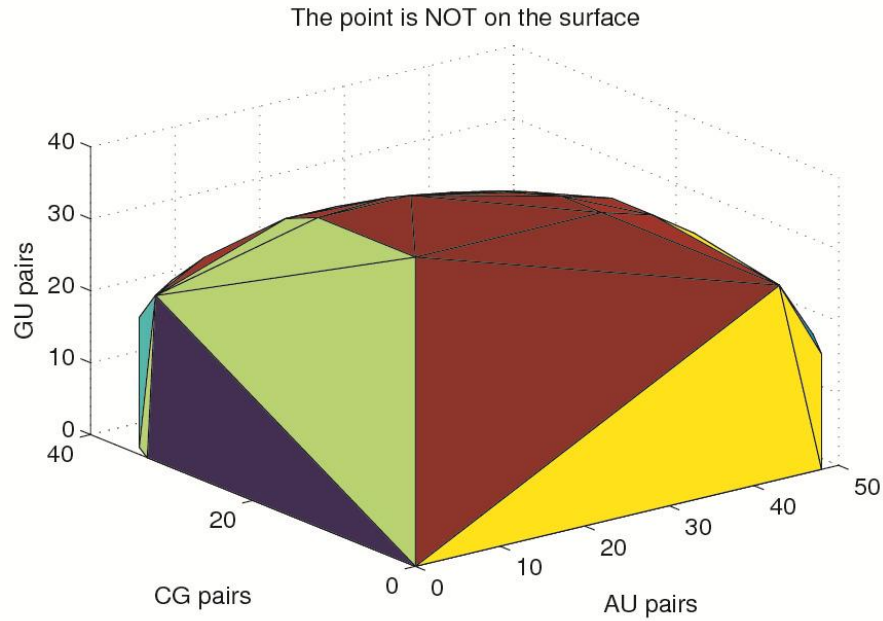
Final Polygon for a Ribosomal RNA (2D)

Figure 13 - (Top) The 3D Newton Polytope of a Ribosomal RNA, $r(x) = 2$. (Bottom) The 2D Newton Polygon of the same RNA, $r(x) = 10$.

In Figure 14, RNA is a shorter one with 32 nt in length. In that case, the feature vector lies on the boundary of the polytope in three dimensional energy model, but in 2D model $r(x) = 2$.

There is no G-U pair in the structure of this RNA, and as a result in 3D model, the feature vector places on the face $c_3(x, y) = 0$.

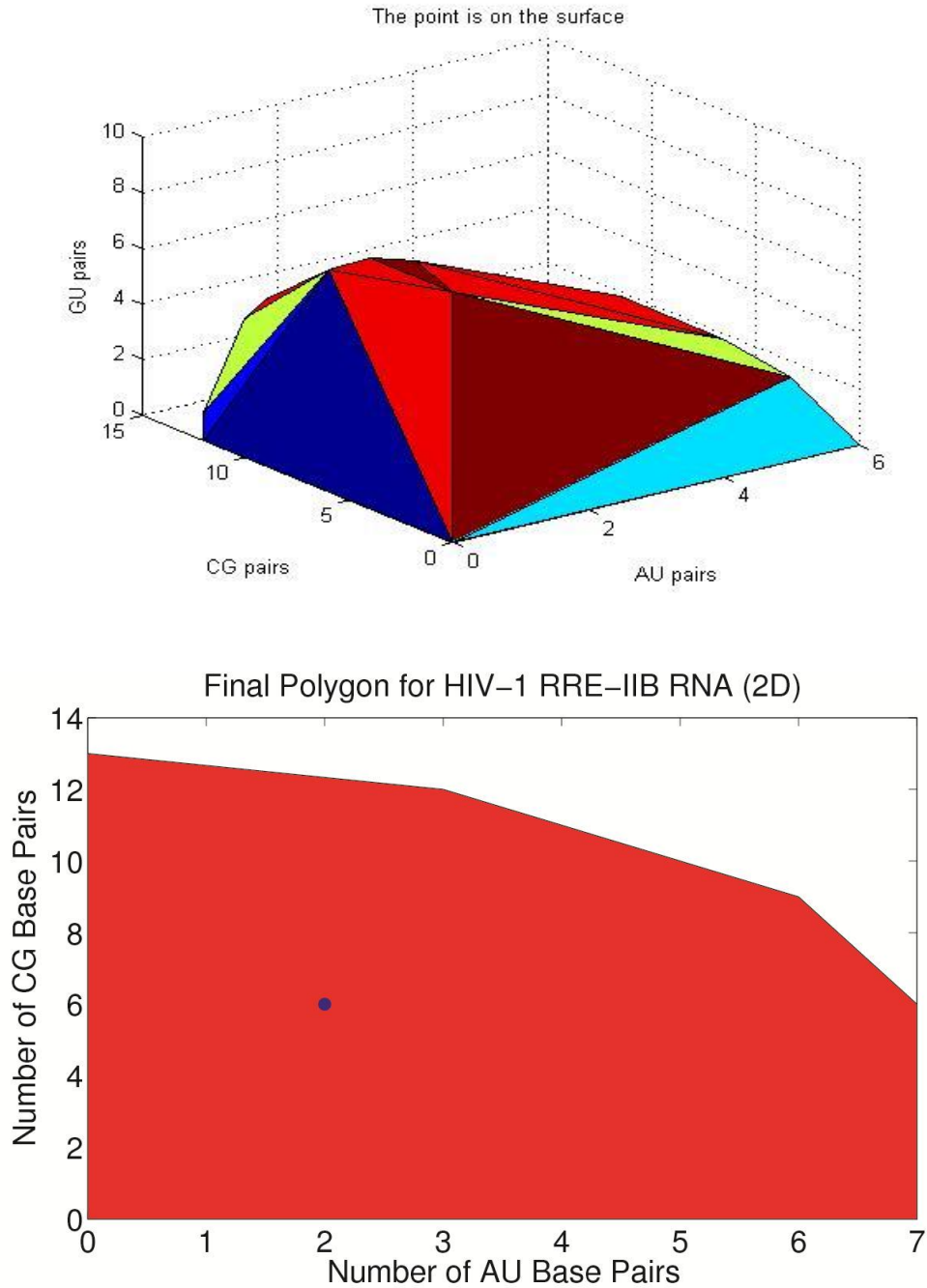Figure 14 - (Top) The 3D Newton Polytope of HIV RRE-IIB RNA, $r(x) = 0$. (Bottom) The 2D Newton Polygon of the same RNA, $r(x) = 2$.

The third RNA in Figure 15 is a 121 nt long *E. coli* 5s Ribosomal RNA. In this example, the distance is not different in the two energy models.
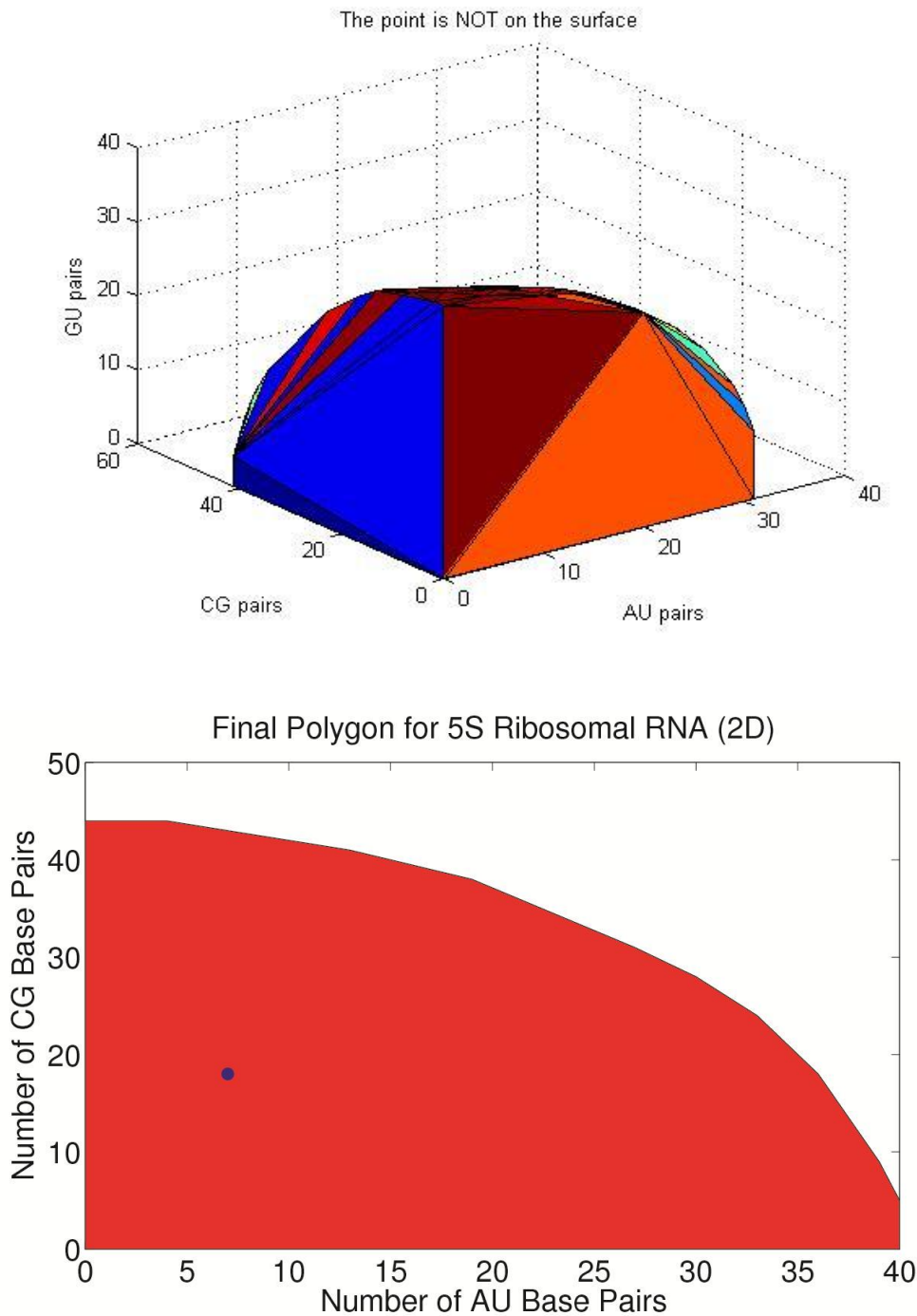

Figure 15 - (Top) The 3D Newton Polytope of 5S Ribosomal RNA, $r(x) = 7$. (Bottom) The 2D Newton Polygon of the same RNA, $r(x) = 7$.

Clearly in 3D model, we expect a 3D volume as the result; however, there are some exceptional cases that produce 2D polygons or just a line as the Newton polytope in 3D energy model. The reason is that one or two types of base pairs cannot happen in the secondary structure, for instance when the strand does not include one or two types of the bases.

The histograms of the calculated distance $r(x)$ are demonstrated in Figure 16 and Figure 18. Other two histograms in Figure 17 and Figure 19 are correspondent to the normalized distance. In 3D model, distance is normalized with the third root of the polytope volume, and in 2D model the normalization factor is the square root of polygon area.
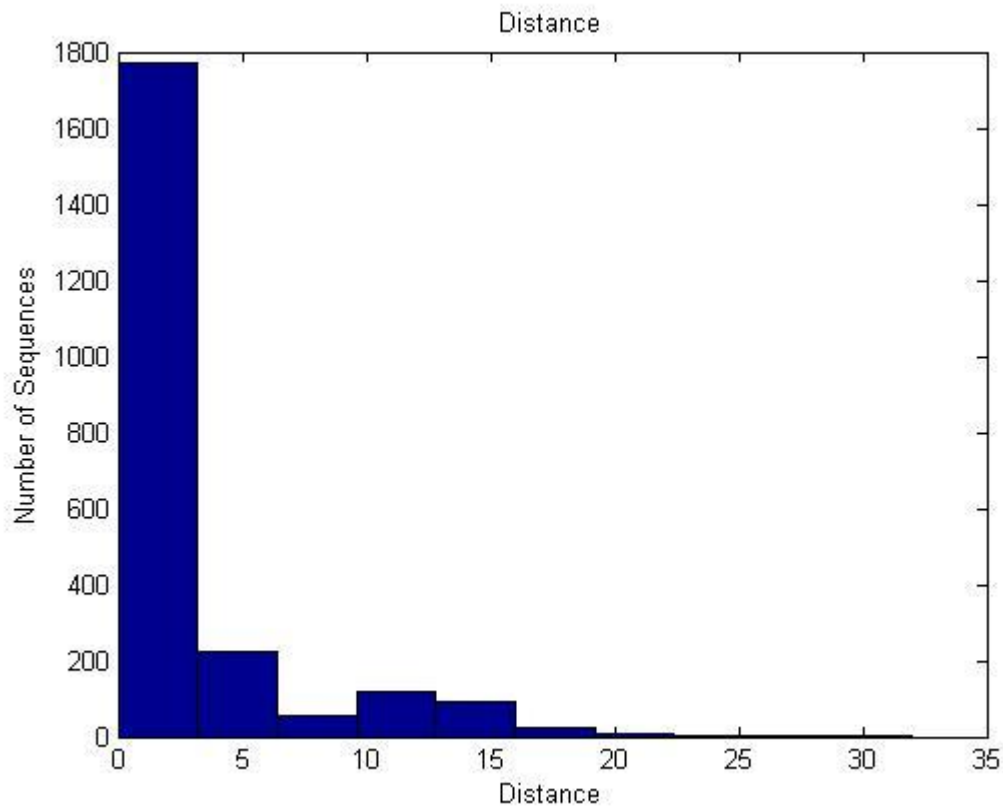

Figure 16-Histogram of $r(x)$ in the 3D energy Model.

Normalized Distance
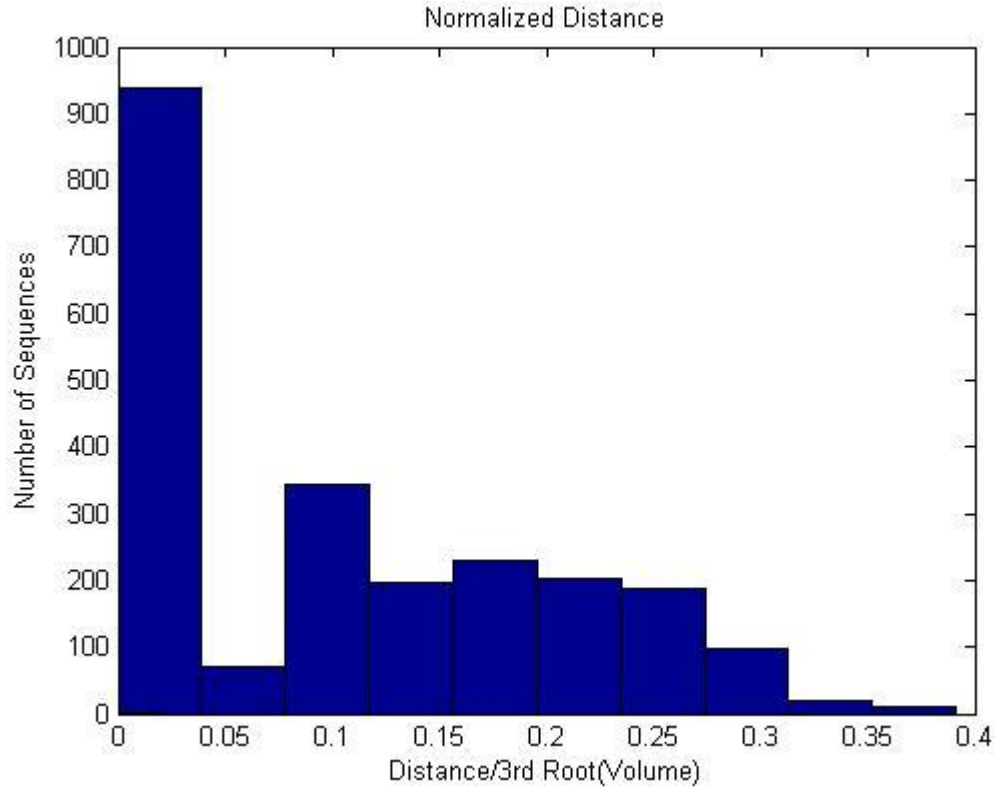
Figure 17- Histogram of normalized $r(x)$ in 3D model.

These histograms are all based on the computed distance for 2300 strands of RNA. In Figure 16, which illustrates the distance histogram for 3D model, we can see that for 934 or 41% of strands, $r = 0$. For 439 (20%) of RNAs, the distance between the feature vector and the Newton polytope is less than or equal to one and not zero. Only for less than 1% of strands, this distance goes larger than 18. In 2D distance histogram, for 99 strands of RNA, r is larger than 15. For 361 strands the feature vector places on the boundary of polygon.
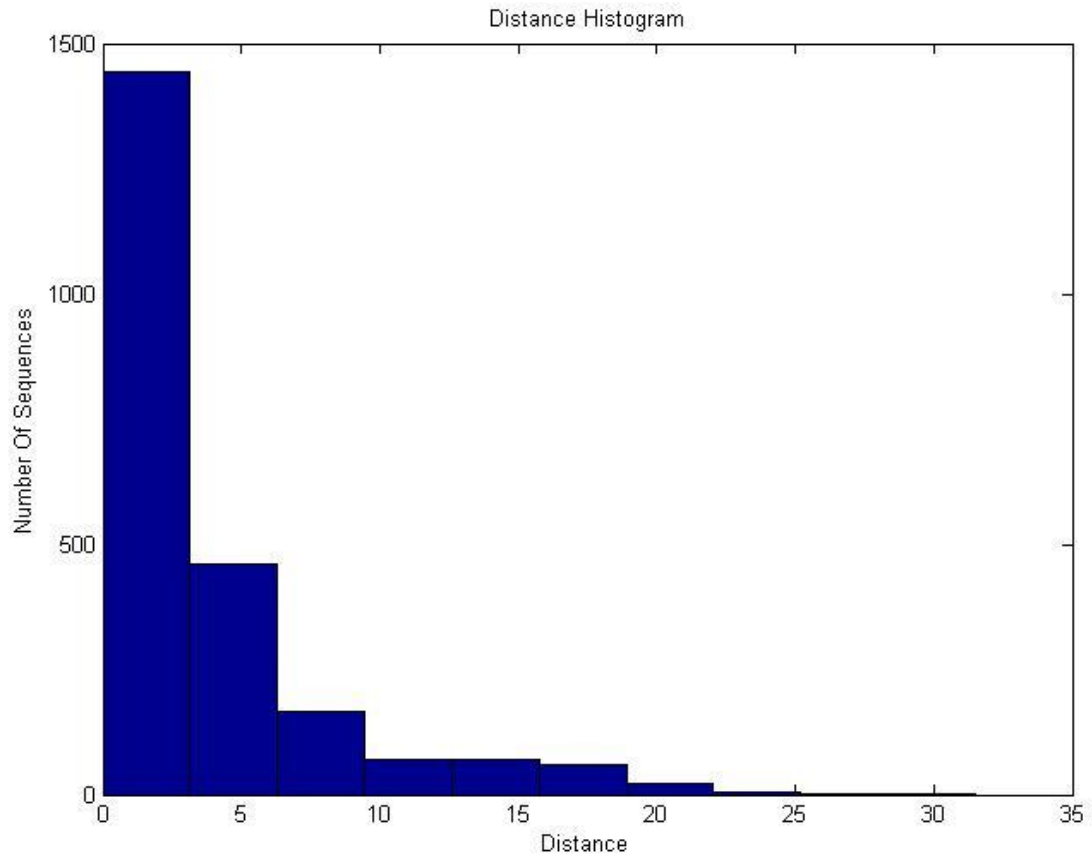
Figure 18- Histogram of $r(x)$ in the 2D energy Model.
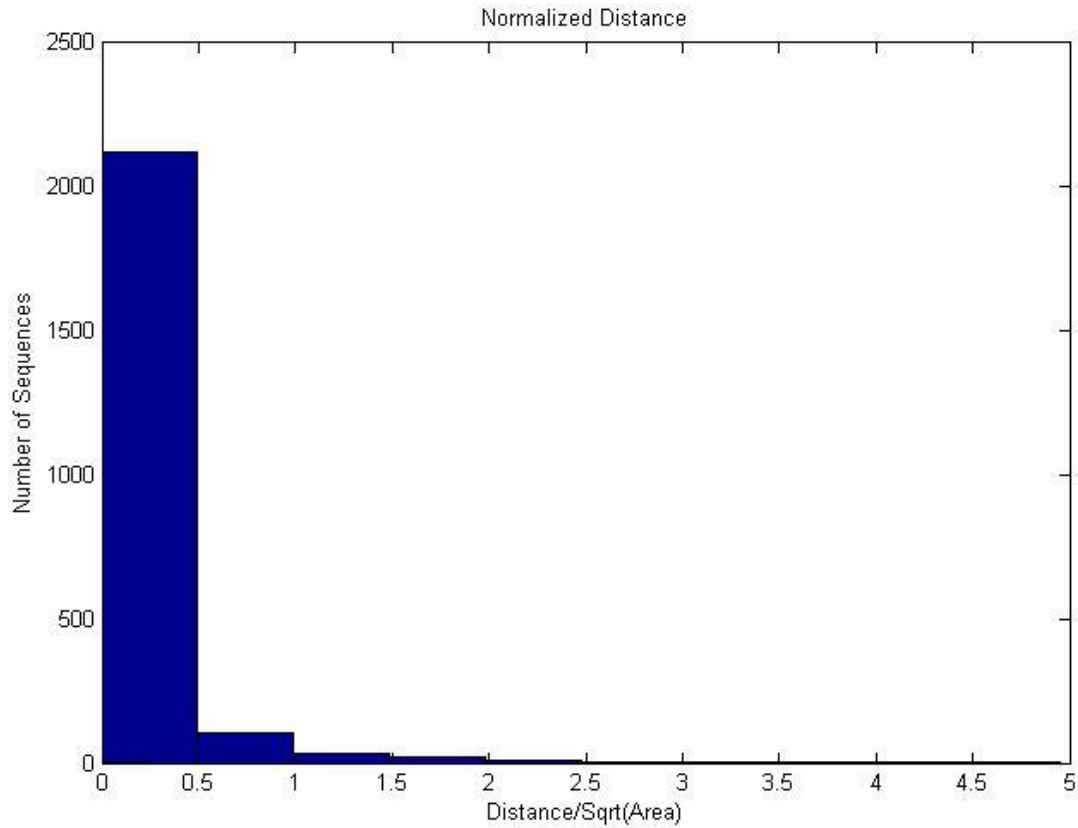
Figure 19 - Histogram of normalized $r(x)$ in the 2D model.

The number of faces for the Newton polytope in the three dimensional model is computed. Figure 20 demonstrates the histogram of number of faces. This number can range from 5 to more than 75. More than 58% of RNAs in this dataset produce polytopes with less than 20 faces.
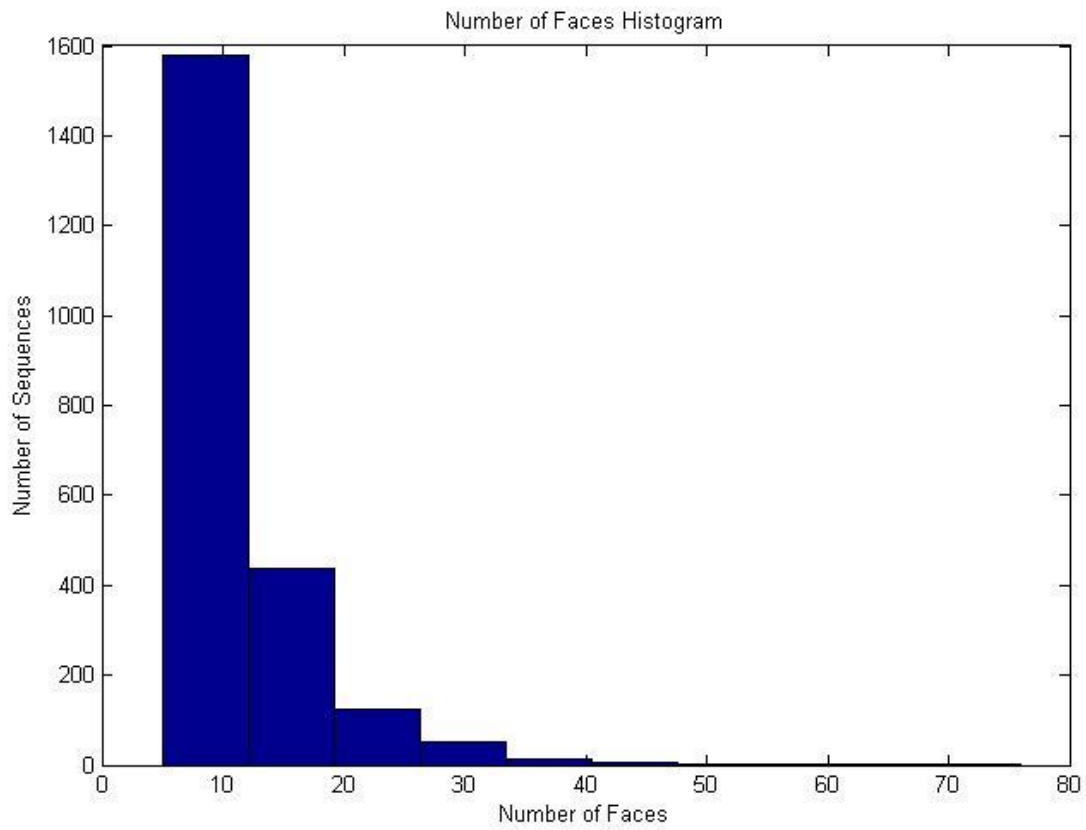
Figure 20 - Histogram of number of faces of the 3D Polytope.

The last two figures demonstrate the relation between the length of strands and number of vertices in Newton polytopes.
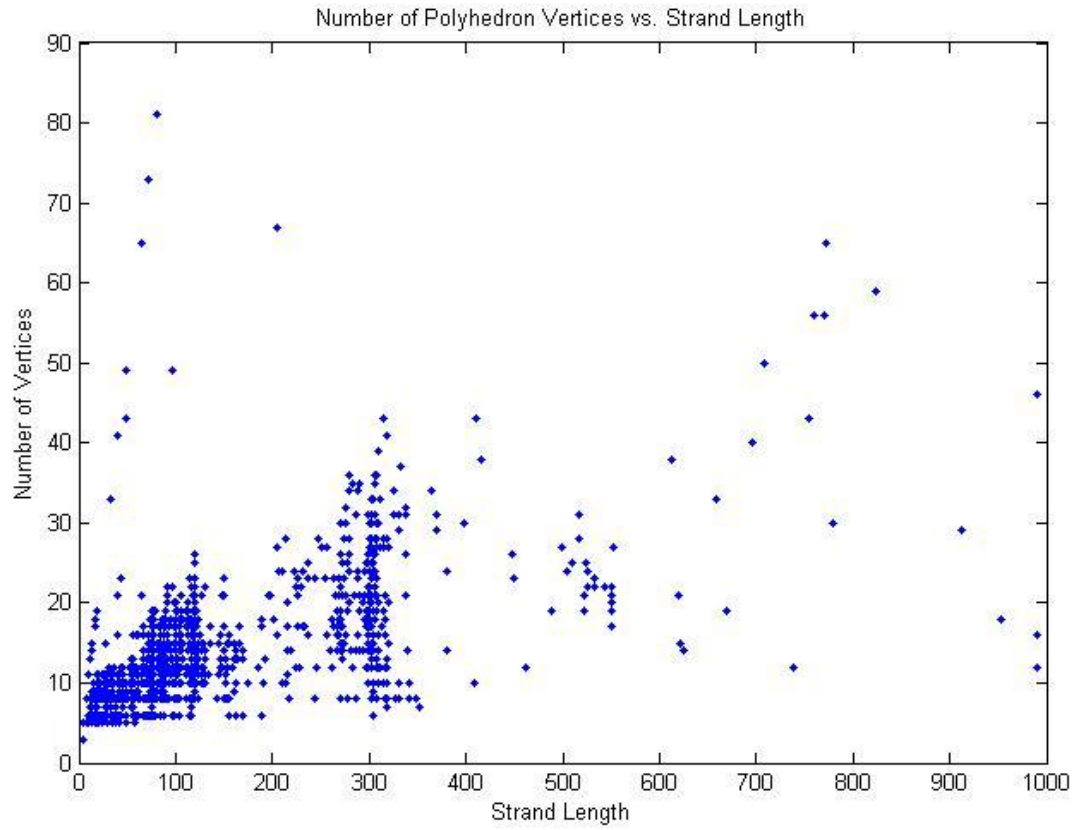
Figure 21 - Number of vertices vs. strand length in 3D model.
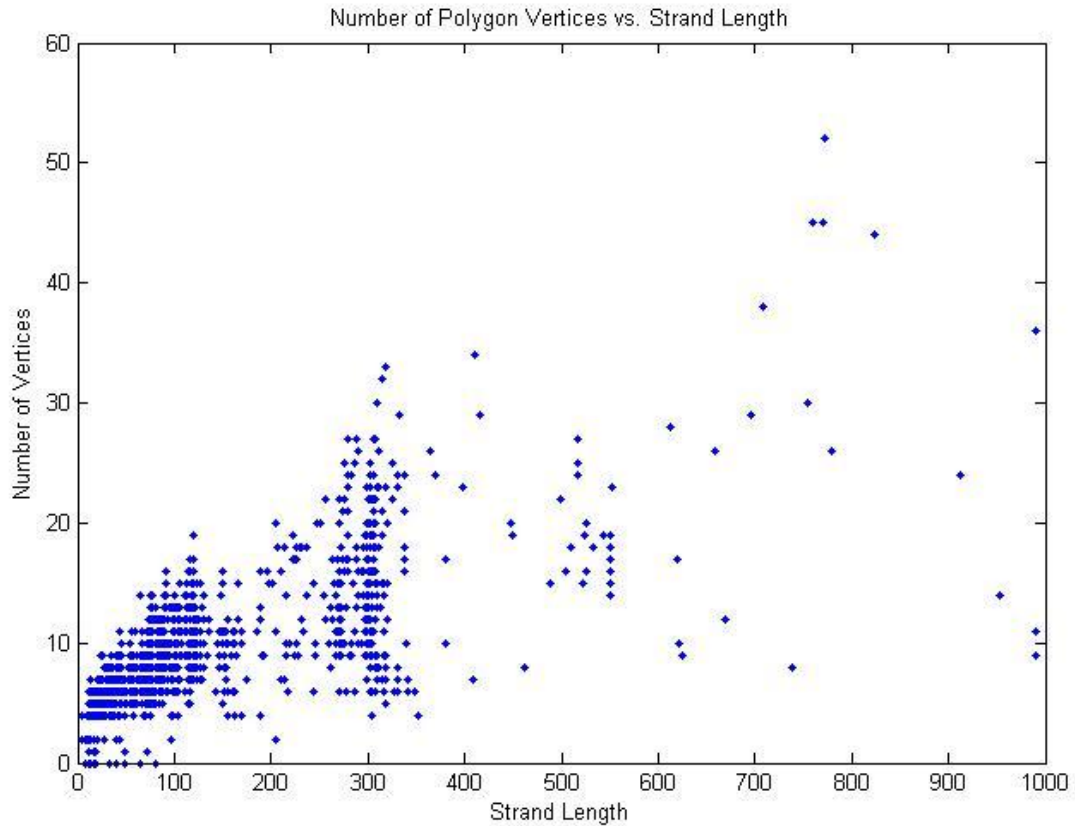
Figure 22 - Number of vertices vs. strand length in 2D model.

Conclusion and future work

This thesis started with a review on different RNA secondary structure prediction techniques, addressing their advantages and drawbacks. Based on these methods and their characteristics, few gaps and possible ways to improve the field were addressed.

In the next part, the focus was on the inherent limitation of energy model, which makes achieving high accuracy with the existing methods impossible. The notion of learnability was introduced to measure the potential of energy models. The necessary condition for a learnable model was defined, and the required dynamic programming to verify this condition, which works based on the computation of the Newton polytope of the partition function, was purposed. To

examine the suggested method, we applied this theory on a 3D energy model, including A-U, C-G and G-U counts. For 40% of the input strands, the condition was satisfied. For almost 20% of the RNAs in the dataset, the condition was not satisfied, but the violation is small. Hence, we suggest that expanding the energy model may help to satisfy the condition for these RNAs. For the rest of strands, the necessary condition was violated significantly. These cases are the subjects of future investigations.

Because of the computation of convex hull in the suggested algorithm, it has an exponential complexity; however, we hope to decrease this complexity by dimensionality reduction techniques.

The next step is to investigate the sufficient condition for a set of parameters to be learnable and the generalization power of a learnable set.

To the best of our knowledge, the introduced approach in [63] is the first systematic way to analyze the suitability of an energy model, and it can be a beginning point for further research. Eventually, this method can help to find an optimal set of features, which includes the entire required sub-structures for the RNA structure prediction. A sufficient number of RNA strands, which cover all of these features, can be designed and built synthetically to provide us the necessary thermodynamic measurements, more efficiently.

References

1. W. Gilbert, "The RNA world," Nature, vol. 319, 1986.

2. D. Bartel and P. Unrau," Constructing an RNA world," Trends in Cellular Biology, vol. 9, no. 12, 1999.

3. G. Storz, "An expanding universe of noncoding RNAs," Science, vol.296, 2002.

4. G. Hannon, "RNA interference," Nature, vol. 418, 2002.

5. D. Bartel, "MicroRNA: Genomics, biogenesis, mechanism, and function," Cell, vol.116, 2004.

6. T. Mercer, M. Dinger, and J. Mattick, "Long non-coding RNAs: insights into functions," Nature reviews (Genetics), vol. 10, 2009.

7. P. Zamore and B. Haley, "Ribo genome: The big world of small RNA," Science, vol. 309, 2005.

8. J. Wilusz, H. Sunwoo, and D. Spector, "Long non-coding RNA: functional surprises from the RNA world," Genes and development, vol. 23, 2009.

9. E. Rivas and S. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots," Journal of Molecular Biology, vol. 285, no. 5, 1999.

10. R. Dirks and N. Pierce, "A partition function algorithm for nucleic acid secondary structure including pseudoknots," Journal of Computational Chemistry, vol. 24, 2003.

11. R. Nussinov, G. Pieczenik, J. Griggs, and D.Kleitman, "Algorithms for Loop Matchings," SIAM Journal on Applied Mathematics, vol. 35, no. 1, 1978.

12. M. Waterman and T. Smith, "RNA secondary structure: a complete mathematical analysis," Mathematical Biosciences, vol. 42, 1978.

13.   J. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," Biopolymers, vol. 29, 1990.

14.   S. Zakov, Y. Goldberg, M. Elhadad, and M. Ziv-Ukelson, "Rich parameterization improves RNA structure prediction," Journal of Computational Biology, vol. 18:11, 2011.

15.   C. Do, D. Woods, and S. Batzoglou, "Contrafold: RNA secondary structure prediction without physics-based models," Bioinformatics, vol. 22, no. 14, 2006.

16.   M. Andronecu, A. Codon, H. Hoas, D. Mathews, and K. Murphy," Computational approaches for RNA parameter estimation," RNA, vol. 16, 2010.

17.   R. Durbin , S. Eddy, A. Krogh, and G. Mitchison, "Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids," Cambridge University Press, 1998.

18.   M. Steen and D. Mathews, "RNA structure prediction: an overview of methods," Methods in Molecular Biology, vol. 905, 2012.

19.   H. Chitsaz, R. Salari, S. Sahinalp, and R. Backofen, "A partition function algorithm for interacting nucleic acid strands," Bioinformatics, vol. 25, no. 12, 2009.

20.   D. Gibson, J. Glass, C. Lartigue, V. Noskov, R. Chuang, M. Algire, G. Benders, M. Montague, L. Ma, M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. Denisova, L. Young, Z. Qi, T. Segall-Shapiro, C. Calvey, P. Parmar, C. Hutchison III, H. Smith, and J. Venter, "Creation of a bacterial cell controlled by a chemistry synthesized genome," Science, vol. 329, 2010.

21.   I. Tinoco and C. Baustamante, "How RNA folds," Journal of Molecular Biology, vol. 293, 1999.

22.   B. Felden, "RNA structure: experimental analysis," Current Opinion in Microbiology, vol. 10, 2007.

23.   K. Miura, S. Tsuda, T. Udea, F. Harada, and N. Kato, "Chemical modification of guanine residues of mouse 5s ribosomal RNA with kethoxal," Biochimica et Biophysica Acta, vol. 739, 1983

24.   P. Rocca-Serra, S. Bellaousov, A. Birmingham, C. Chen, P. Cordero, R. Das, L. Neulander, C, Duncan, M. Halvorsen, R. Knight, N. Leontis, D. Mathews, J. Ritz, J. Stombaugh, K. Weeks, C. Zirbel, and A. Laederach, "Sharing and archiving nucleic acid structure mapping data," RNA, vol. 17, 2011.

25.   E. Merino, K. Wilkinson, J. coughlan, and K. Weeks, "RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE)," Journal of the American Chemical Society, vol. 127, 2005.

26.   P. Gardner and R. Giegerich, "A comprehensive comparison of comparative RNA structure prediction approaches," BMC Bioinformatics, vol. 5, no. 140, 2004.

27.   J. Parsch, J. Braverman, and W. Stephan, "Comparative sequence analysis and patterns of covariation in RNA secondary structures," Genetics, vol. 154, 2000.

28.   I. Hofacker, M. Fekete, and P. Stadler, " Secondary structure prediction for aligned RNA sequences," Journal of Molecular Biology, vol. 319, no. 5, 2002.

29.   B. Knudsen and J. Hein, "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history," Bioinformatics, vol. 15, no. 6, 1999.

30.   D. Sankoff, "Simultaneous solution of the RNA folding, alignment and protosequence problems," SIAM Journal on Applied Mathematics, vol. 45, 1985.

31.   J. Gorodkin, L. Heyer, and G. Stormo, "Finding the most significant common sequence and structure motifs in a set of RNA sequences," Nucleic Acids Research, vol. 25, no.18, 1997.

32.   D. Mathews and D. Turner, "Dynalign: an algorithm for finding the secondary structure common to two RNA sequences," Journal of Molecular Biology, vol. 317, no. 2, 2002.

33.   J. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. Lander, J. Kent, W.   Miller, and D. Haussler, "Identification and classification of conserved RNA secondary structures in the human genome," PLoS Computational Biology, vol. 2, 2006.

34.   S. Washietl, I. Hofacker, and P. Stadler, "Fast and reliable prediction of noncoding RNAs," Proceedings of the National Academy of Sciences, vol. 102, 2005.

35.   I. Tinoco, O. Uchlenbeck, and M. Levine, "Estimation of secondary structure in ribonucleic acids," Nature, vol. 230, 1971.

36.   D. Mathews, J. Sabina, M. Zuker, and D. Turner, " Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," Journal of Molecular Biology, vol. 288, 1999

37.   M. Zuker and P. Steigler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," Nucleic Acids Research, vol. 9, no. 1, 1981.

38.   M. Zuker and D. Sankoff, "RNA secondary structure and their prediction," Bulletin of Mathematical Biology, vol. 46, no. 4, 1984.

39.   M. Zuker, "On finding all suboptimal foldings of an RNA molecule," Science, vol. 244, No. 4900, 1989.

40.   I. Hofacker, "Vienna RNA secondary structure server," Nucleic Acids Research, vol. 31, 2003.

41.   M. Mandal and R. Breaker, "Gene regulation by riboswitches," Nature reviews Molecular Cell Biology, vol. 5, no. 6, 2004.

42. J. Soukup and G. Soukup, "Riboswitches: the oldest mechanism for the regulation of gene expressions?," Trends in Genetics, vol. 20, 2004.

43. D. Aalberts and N. Hodas, "Asymmetry in RNA pseudoknots: observations and theory," Nucleic Acids Research, vol. 33, 2005.

44. S. Cao and S. Chen, "Predicting RNA pseudoknot folding thermodynamics," Nucleic Acids Research, vol. 34, 2006.

45. J. Edmonds, "Maximum matching and polyhedron with 0, 1- vertices," Journal of research of the National Bureau of Standards, vol. 69, 1965.

46. R. Cary and G. Stormo, "Graph- theoretic approach to RNA modeling using comparative data," ISMB, 1995.

47. J. Tabaska, R. Cary, H. Gabow, and G. Stormo, " An RNA folding method capable of identifying pseudoknots and base pair triples," Bioinformatics, vol. 8, 1998.

48. B. Knusden and J. Hein, "Pfold: RNA secondary structure prediction using stochastic context-free grammars," Nucleic Acids Research, vol. 2003, no. 13, 2003.

49. R. Dowell and S. Eddy, "Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction," BMC Bioinformatics, vol. 5:71, 2004.

50. M. Brown and C. Wilson, "RNA pseudoknots modeling using intersections of stochastic context free grammars with applications to database search," Pacific Symposium on Biocomputing, 1996.

51. J. Reuter and D. Mathews, "RNAstructure: software for RNA secondary structure prediction and analysis," BMC Bioinformatics, vol. 11, 2010.

52. M. Andronescu, A. Condon, H. Hoos, D. Mathews, and K. Murphy, "Efficient parameter estimation for RNA secondary structure prediction," Bioinformatics, vol. 23, 2007.

53. M. Collins, "Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithm," Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, Association for Computatioal Linguistics, 2002.

54. E. Torarinsson, J. Havgaard, and J. Gorodkin, "Multiple structural alignment and clustering of RNA sequences," Bioinformatics, vol. 23, no. 8, 2007.

55. D. Staplle and S. Butcher, "Pseudoknots: RNA structures with diverse functions," PLOS Biology, vol. 3, no. 6, 2005.

56. S. Siebert and R. Backofen, "MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons," Bioinformatics, vol. 21, no. 16, 2005.

57. T. Akutsu, "Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots," Discrete Applied Mathematics, vol. 104, 2000.

58. Y. Chen and G. Varani, "RNA Structure," Encyclopedia of Life Sciences, 2010.

59. M. Szymanski, M. Barciszewska, V. Erdmann, and J. Barcizewski, "5S ribosomal RNA databse," Nucleic Acids Research, vol. 30, 2002.

60. F. Crick, "On protein synthesis," Symposia of the Society for Experimental Biology, vol. 12, 1958.

61. R. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," Nucleic Acids Research, vol. 32, 2004.

62. H. Chitsaz, E. Forouzmand, and G. Haffari, "An efficient algorithm for upper bound on the partition function of nucleic acids," Journal of Computational Biology, 2013.

63. E. Forouzmand and H. Chitsaz, "The RNA Newton polytope and learnability of energy parameters," ISMB 2013, Jan 2013.

64. I. Emiris, "Sparse elimination and applications in kinematics," Ph.D. thesis, UC Berkeley, Berkeley, CA.

65. M. Amdronescu, V. Bereg, H. Hoos, and A. Condon, "RNA STRAND: the RNA secondary structure and statistical analysis database," BMC Bioinformatics, vol. 9, 2008.

66. C. Barber, D. Dobkin, and H. Huhdanpaa, " The quickhall algorithm for convex hulls," ACM Transactions on Mathematical Software, vol. 22, 1996.

67. M. Yosphe, "Distance from a point to a 2D polygon," http://www.mathworks.com/matlabcentral/fileexchange/12744-distance-from%-a-point-to-polygon, 2006.

Abstract

The RNA Newton Polytope and Learnability of Energy Parameters

by

Elmirasadat Forouzmand

August 2014

Advisor : Dr. Hamidreza Chitsaz

Major: Computer Science

Degree: Master of Science

Computational RNA secondary structure prediction has been a topic of much research interest for several decades now. Despite all the progress made in the field, even the state-of-the-art algorithms do not provide satisfying results, and the accuracy of output is limited for all the existent tools. Very complex energy models, different parameter estimation methods, and recent machine learning approaches had not been the answer for this problem. We believe that the first step to achieve results with high quality is to use the energy model with the potential for predicting accurate output. Hence, it is necessary to have a systematic way to analyze the suitability of an energy model. We introduced the notion of learnability to measure this suitability. A learnable energy model has at least one subset of parameters that can render every known RNA to date the minimum free energy structure, which means 100% accuracy. We also found the necessary condition for a model to be learnable and implemented the dynamic programming based algorithm to asses this condition for a set of RNAs. This algorithm computes the convex hull of all possible feature vectors for a sequence. With the partition function as a polynomial, this convex hull is also the Newton polytope of the partition function. To the best of our knowledge, this is the first systematic approach for evaluating the inherent capability of an energy model.

Autobiographical statement

My way towards an academic life began with my entrance to Farzanegan High School, which is a branch of NODET (National Organization for Developing Exceptional Talents). Studying among talented students and experienced teachers in this school helped me greatly in developing my skills in my favorite fields such as mathematics and programming. This background helped me to rank 57th among 400,000 students in the national University Entrance Exam (Konkoor) in 2006. In that year, I started my undergraduate studies in Electrical Engineering (EE) Department of Sharif University of Technology.

According to my rank in Konkoor, I was able to choose any major to study. I wanted to study in a field which is highly correlated with math and programming and also provides me opportunities for research and innovation. After examining all features, I decided to continue my studies in EE department because of the great academic atmosphere it possesses due to highly motivated students and brilliant professors. However, later I found out that if I had more information about computer science major, it could be my choice. The first 2 years of university provided me with basics and also introduced different areas of specialization in electrical engineering. In the 3rd year with deeper understanding of these areas I chose Control as my specialization. Studying in Control gave me the opportunity to know about AI, Neural Network, machine learning, and robotics. I became familiar with Bioinformatics and later Computational Biology as an application for Machine Learning and clustering. In that time, the idea of being an engineer who works in Biology with creation and designing new tools to analyze Biological data was very interesting for me.

When I decided to apply abroad for my graduate studies, I thought with this opportunity to continue my studies in a proper environment, with facilitated lab and under the supervision of the best professors, I can continue my studies in computer science, which covers my interests in a better way. Although I knew with a different academic background, it would be more difficult to find a good research position; I tried to do my best to continue my academic way in computer science with a focus on Bioinfornatics. In 2011, I started my graduate studies in the Computer Science department of Wayne State University under supervision of Prof. Hamidreza Chitsaz in Algorithmic Biology lab. Bioinformatics is the strongest group in this department. I have been doing research in different areas of bioinformatics including next generation sequencing and RNA secondary structure prediction in the last three years. A big part of what I have done during this period has been reported in my master thesis and also published as a paper in 2013.

Finishing my master studies, now my ultimate academic goal is to achieve a PhD degree and work as a researcher in academia or industry after that.