New Jersey Institute of Technology
Digital Commons @ NJIT

Theses | Theses and Dissertations

Spring 2004

Review of algorithms for RNA secondary structure prediction with pseudoknots

Ingrid Helene Nielsen
New Jersey Institute of Technology

Follow this and additional works at: https://digitalcommons.njit.edu/theses

 Part of the Computer Sciences Commons

Recommended Citation
Nielsen, Ingrid Helene, "Review of algorithms for RNA secondary structure prediction with pseudoknots" (2004). *Theses*. 557.
https://digitalcommons.njit.edu/theses/557

This Thesis is brought to you for free and open access by the Theses and Dissertations at Digital Commons @ NJIT. It has been accepted for inclusion in Theses by an authorized administrator of Digital Commons @ NJIT. For more information, please contact digitalcommons@njit.edu.

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

REVIEW OF ALGORITHMS FOR RNA SECONDARY STRUCTURE PREDICTION WITH PSEUDOKNOTS

by
Ingrid Helene Nielsen

Pseudoknots are structures that are formed from the base pairing of an RNA secondary loop structure with a complementary base which lies somewhere outside of the loop. The result is a structure, which plays a vital role in cell structure rigidity, regulation of protein synthesis, and in the structural organization of RNA complexes. Deciphering RNA folding patterns would begin to unravel some of the mysteries surrounding the cell and its functions and open a new world to scientists. Many algorithms have been written in this quest to predict RNA's secondary structure but not many have been very successful.

In this thesis, some of these algorithms are discussed and considered for their strengths and weaknesses. First those algorithms, which exclude pseudoknots and other more complex structures, are presented. The later algorithms include those, which attempt to include some of the more complex structures into their calculations.

In the end, all the algorithms are taken into consideration and their strengths and weaknesses compared so as to find some path for future direction. By using the strengths found in these variety of algorithms and avoiding some of the pitfalls encountered by others hopefully new algorithms will be developed in the future that are more successful in deciphering RNA secondary structure.

REVIEW OF ALGORITHMS FOR RNA SECONDARY STRUCTURE PREDICTION WITH PSEUDOKNOTS

by
Ingrid Helene Nielsen

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Department of Computer Science

May 2004

APPROVAL PAGE

REVIEW OF ALGORITHMS FOR RNA SECONDARY STRUCTURE PREDICTION WITH PSEUDOKNOTS

Ingrid Helene Nielsen

Dr. Michael Recce, Thesis Advisor — Date
Associate Professor of Information Systems, NJIT

Dr. Barry Cohen, Committee Member — Date
Assistant Professor of Computer Science, NJIT

Dr. Qun Ma, Committee Member — Date
Assistant Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Ingrid Helene Nielsen

Degree: Master of Science in Computational Biology

Date: May 2004

Undergraduate and Graduate Education:

- Master of Science in Computational Biology
  New Jersey Institute of Technology, Newark, NJ, 2004

- Bachelor of Arts in Molecular Biology
  State University of New York, College at New Paltz, NY, 2002

Major: Computational Biology

To my family for all their support

ACKNOWLEDGMENT

I would like to thank Dr. Michael Recce for all his help to me while researching and writing this thesis. Also to Dr. Barry Cohen and Dr. Qun Ma for being a part of my committee at the last minute and giving me some good suggestions on how to make it even better.

I would like to thank all my fellow students in the computational biology program for always working together, supporting each other and for helping to get me through the last two years.

Finally I would like to thank my parents for supporting my decisions without every understanding where I was going with them, my sister and cousin for always being there for me, and my friends and roommates for helping me deal with all the stress of the last few years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION TO RNA FOLDING

RNA folding is an essential aspect for many of the structures involved in the regulatory, catalytic and structural roles within the cell. For this reason, scientists have a big interest in deciphering the patterns by which the RNA molecule folds. For a structure of length n, it has been estimated that there are about 1.8^n possible structures. This is a huge amount of structures to consider when predicting the secondary structure of a particular sequence. For this reason, secondary structure prediction remains a very computationally intense problem (Shapiro, 1999).

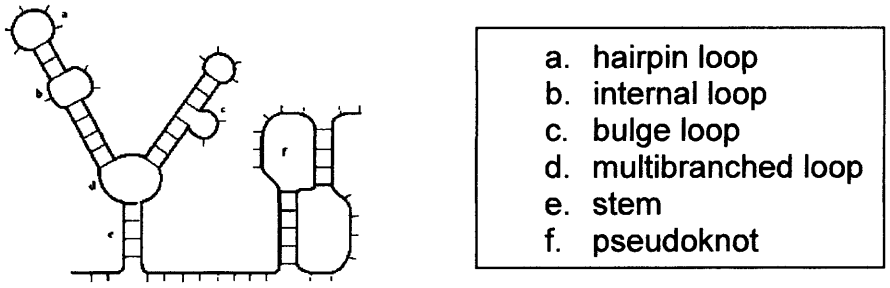

Figure 1.1 Basic diagram of some complex interactions found in RNA secondary structures

The primary structure of RNA is the series of nucleic acids, which make up its sequence. The basic secondary structure of RNA on the other hand, is dominated by its base-to-base interactions known as helices. For the most part these interactions are between complementary bases and are therefore only simple Watson-Crick pairs which are easy to predict and seemingly easy to map.

However, not all of the interactions within the RNA molecule follow the same simple pairings. Many more complex interactions, which are characterized as loops are possible within the structure. These interactions are much more difficult to predict.

Some of these more complex interactions include stacking pairs, hairpin loops, bulges, interior and multiple loops, which can be seen in figure 1.1 above. These secondary structure elements although more complex than the simple Watson-Crick base pairs can still be predicted with a certain amount of accuracy using a wide variety of algorithms already determined. On the other hand, structures such as pseudoknots are not as easily predicted by these algorithms. In fact most of the common RNA secondary structure prediction algorithms out there do not even take into account the possibility of a pseudoknot occurring within the structure. These algorithms exclude pseudoknots from the prediction in order to maintain a certain level of accuracy within a manageable amount of time.

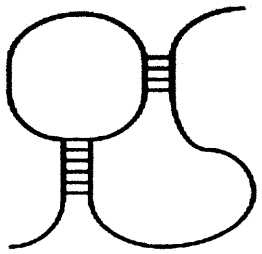

Figure 1.2 Basic pseudoknot

Pseudoknots are actually a form of RNA tertiary interactions however they influence all aspects of the RNA structure. Overall it is the combination of secondary and tertiary interactions as well as the structures interaction with

outside forces such as water, ions and proteins that results in the final 3D structure of the RNA and RNA-protein complexes. It is this final 3D structure that determines the molecules final function and therefore, it is extremely important to determine all the influences when determining the final structure (Shapiro, 1999). Pseudoknots are formed when one of the loops in the RNA's secondary structure pairs with a complementary sequence somewhere outside the loop. Another reason why pseudoknots are so different than other RNA structures is that they are not nested. As you can see in figure 1.2, the pseudoknot shown has base pairs from one loop that interact with base pairs from a separate part of the molecule, possibly another loop but not necessarily. Most RNA secondary structures follow a nested convention always maintaining pairwise correlations, meaning, "that for any two base pairs **i**, **j** and **k**, **l** (where **i<j**, **k<l** and **i<k**), either **i<k<l<i** or **i<j<k<l**" (Rivas and Eddy, 1999). Early algorithms, such as the Zucker dynamic programming algorithm or Mfold algorithm rely on this nested convention in order to predict RNA structure. They do this, by calculating the minimal energy structure recursively on progressively longer subsequences. The fact that pseudoknots violate this nesting convention is yet another reason why many algorithms chose to ignore their existence in order to get a faster and simpler algorithm.

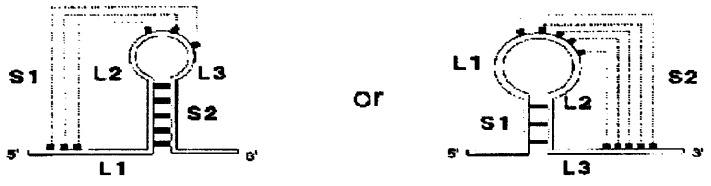

Figure 1.3 Common H-type pseudoknots

At this time almost all known pseudoknots are made up of at least two stems and at least three base pairs. For this reason, if you take into account all four types of single stranded loop regions that are possible in RNA structures the result is that there are theoretically 14 possible types of pseudoknots. Not all of these pseudoknots are sterically possible however. The most common pseudoknot found is the H-type pseudoknot, which is the result of hairpin loops binding with single stranded regions and can been seen above in figure 1.3 (Shapiro, 1999). The other common types of pseudoknots include the I-type, which, is formed between interior loops and single stranded regions and the B-type, which is formed between bulge loops and single stranded regions and can be seen below in figure 1.4.

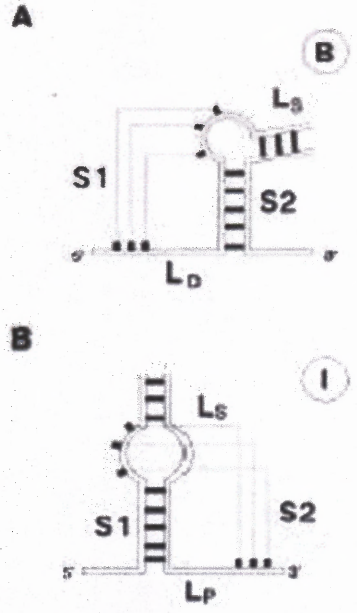

Figure 1.4 **(A)** Common B-Type pseudoknot
(B) Common I-Type pseuodknot

Pseudoknots provide the RNA's secondary structure with a certain amount of rigidity that doesn't come from some of its other more fluid structures. It is able to do this because of the branching affect that occurs due to the overlapping of base pairs. These overlapping base pairs are a result of the regular double stranded helices of the structure connecting with some of the more flexible structures of the RNA molecule. Usually pseudoknots are found in ribosomal structures where they are essential for 3D enzymatic shape due to their contribution to this structural rigidity. They can also be found in the catalytic core of group I introns, RNase P RNAs, and in mRNA-ribosome interactions during the initiation of translation and during frameshift regulation (Xayaphoummine et al., 2003). Recently, pseudoknots have been found to be a common structural motif in viral RNAs where they are thought to mimic certain tRNA structure.

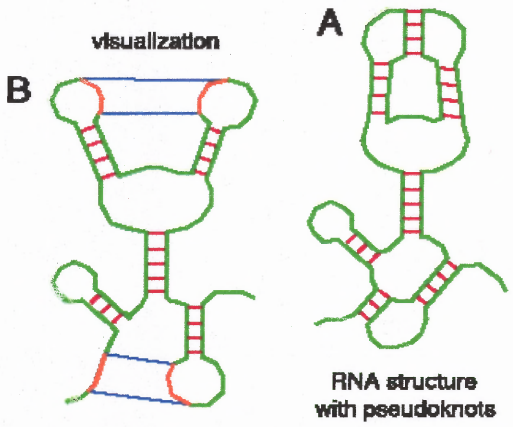

Figure 1.5 Binding of a more complex pseudoknot

When pseudoknots are found in coding regions they tend to have a very large impact. For example, they are known to stimulate ribosomal frameshifting

and translational read-through during elongation. Frameshifting is a process that allows the cell, usually viral, to code for more then one protein using the same sequence. Pseudoknots stimulate such processes by combining with so-called "slippery" heptanucleotide sequences resulting in a slipping of the ribosome during translation, which leads to a frameshift. However, the impact of these structures is not limited to only coding regions; even non-coding regions are highly influenced by the presence of pseudoknots. In such regions, pseudoknots have been shown to initiate translation in one of two ways. The first is when they fold at the 5" end of non-coding regions. Here they become part of the internal ribosomal entry sites or IRES during translation control. These sites mediate end-independent ribosomal attachment to a specific internal position in the mRNA. The second is when the pseudoknot folds at the 3' non-coding region. Here their influence is as a translational enhancer as well as providing a signal for replication (Han, Lee and Kim, 2002). The final area where pseudoknots seem to influence the cell is when they are found in molecules, which have some sort of catalytic activity. In these types of molecules, pseudoknots are found in the core of the tertiary fold and involve interactions between nucleotides that are extremely far apart in the RNA's sequence.

Pseudoknots that are involved in translational control at the 5' end seem to adopt one of two roles. In the first case, translation control is the result of a specific recognition of the pseudoknot by a particular protein. In the second, the presence of the folded pseudoknot is the only necessity for control with no restrictions on the nucleotide sequence. Pseudoknots that are found in core

positions otherwise known as core pseudoknots are necessary for the formation of reaction centers in ribosomes. Most enzymatic RNA's that contain core pseudoknots are involved in specific activities which are usually cleavage or self-cleavage reactions (Stadler and Haslinger, 1997).

The complexity of these structures leads to two main reasons why it becomes an issue to include pseudoknots into RNA structure prediction algorithms. First of all there is the issue of structural modeling. Modeling nested structures is simpler because a database of known structures and energies is available to pull from when running the algorithm. This is not the case for pseudoknots, there is no database of known energies and for this reason an algorithm including pseudoknots most use some other form of description for the structure. The second issue in pseudoknot modeling is that of computational efficiency. Simple RNA structure prediction algorithms excluding pseudoknots are able to compute a structure in polynomial time (Xayaphoummine et al., 2003). The best of these algorithms is the Mfold algorithm which has a time complexity of $O(|n|^3)$ with space $O(|n|^2)$. It makes sense to say that since Mfold algorithms have the best time and space complexities and are unable to predict pseudoknots, prediction with pseudoknots would take at least as much time and space to compute and in reality take much more. Lyngso has recently shown, (Lyngso, 2000) that prediction with pseudoknots is in fact an NP-complete problem and that there is really little hope that an algorithm that takes into account all pseudoknots will have a polynomial time complexity. Therefore at this time the only way to get a polynomial time complexity for an RNA secondary

structure prediction algorithm with pseudoknots is to either limit the types of legal pseudoknots allowed in the model or ignore the interactions that occur between the neighboring base pairs. Finally if we compromise our level of accuracy we can use heuristics for structure prediction on pseudoknots resulting in structures of low energy but not necessarily of lowest energy.

Here we will be looking at a series of different algorithms, which predict the secondary structure of RNA sequences. In the next chapter, basic algorithms are presented that predict structures while excluding the presence of pseudoknots. These algorithms all have polynomial time and space complexities and return structures that are optimal based on the legal structures allowed for that particular algorithm. In the third chapter, the algorithms are more complex and all include pseudoknots into their legal structures even though each has its own definition of what types of pseudoknots are allowed. These algorithms all have slightly larger time and space complexities but all of them show that the idea of including pseudoknots into structure prediction algorithms is in fact feasible even if it does take slightly more time and space to compute the necessary extra calculations. All of these algorithms are considered in the final chapter where a discussion put forward on what might be the best route to take in the future for new RNA secondary structure prediction algorithms.

CHAPTER 2

BASIC ALGORITHMS WITHOUT PSEUDOKNOTS

In this chapter, four simple algorithms are described which predict RNA secondary structure while excluding the presence of pseudoknots in their calculations. All of these algorithms, state that limiting the amount of legal structures possible in the structure allows the algorithm to keep a time and space complexity which is within a reasonable range for today's computers to handle. Whether or not these algorithms would be able to handle pseudoknots were they included is not addressed although studying these more basic algorithms is a good background for evaluating those algorithms, which do include pseudoknots.

2.1 Comparative Modeling Approach

Currently the best way of determining the secondary structure of RNA is to use a technique known as comparative modeling. This technique requires the availability of several related structured with a homology of at least 30% and works by identifying pairs of positions where mutations occur that result in retaining base pair capability. The model finds structures with similar base pairing and infers the unknown structure from the known related structure. However since several related structures must be known for the model to work it is not always possible to determine an accurate model. Another problem with this model is that it is difficult to fully automate due to the fact that intervention is

often necessary to identify the common mutations. So although, Protein servers such as Swiss-Prot have been able to deal with the problems of comparative modeling in a reasonable way this is still a ways off for RNA modeling since it would require many components and access to a know database of RNA structures (Lyngso, 2000).

2.2 Genetic Algorithm Approach

Genetic algorithms (GA's) are stochastic optimization techniques that work on populations of possible solutions changing them through series of steps to become closer to a thermodynamically fit model. These series of evolutionary steps involved in the algorithm involve changing some of the solutions known here as a GA mutation as well as the recombining of certain features of the parental solution known as the GA crossover and finally the GA selection which models natures survival of the fittest mechanism. The next generation is then selected based on predefined fitness criteria. The steps of the algorithm are then repeated until no further improvement can be found. Although the results produced by the genetic algorithm approach are not necessarily the optimal solution, the solutions have been shown to perform well under certain tests.

The genetic algorithm approach deals with only a single RNA sequence and uses free energy as its fitness criterion. Free energy is the energy associated with a chemical reaction that is used to do work and is defined as the sum of the systems enthalpy plus the product of the temperature and the entropy. The population used to perform the genetic operations is made up of an

array of elements where each of the elements represents a base-paired region known as a stem (Shapiro, 2003).

In (Chen et al., 2000) a genetic algorithm approach for RNA secondary structure prediction is shown which takes into consideration not only the structural energy but also the structural similarity among different sequences. The method they propose is able to predict a common RNA structure without finding the alignment of the sequences, which genetic algorithms usually require for optimal results. The algorithm begins by generating a list of all possible stems based on a particular sequence. It achieves this by applying a free energy GA to a random population of structures until a certain level of stability is reached. From this list, stems, which are compatible with those already in the structure, are added one at a time in a stepwise manner as long as the resulting structure is more stable than the parent structure. This process is repeated until no stems can be found that will increase the stability of the parent structure.

Once a structure is determined the three genetic operations; crossover, mutation, and selection, are repeated using free energy as the only criterion until the most stable structure is determined. The probability that a structure will be selected for a crossover event is proportional to its free energy with the offspring being a random selection of stems from each of the two parents. Stems that close unstable regions are subject to the mutation operation. These stems are removed from the structure and new stem is added in a completely random manner. In Chen's algorithm if there are **n** structures in the population then **n** structures are mutated and **n** pairs of structures are subject to crossover. This

results in a population that increases three-fold each time an iteration of the genetic algorithm is complete. For this reason, the selection operation is a necessity. This is because the selection operation chooses the next generation, by maintaining the size of the population as **n** and keeping the time and space complexity of the algorithm consistent. Selection takes into account both the structural stability as well as the structural distance between each of the solutions to prevent the solutions chosen from converging prematurely to a local favorable solution. After the operations are complete a conservation score is determined for each of the structures in the population and the genetic algorithm is repeated until the structures converge to one final optimal solution.

Table 2.1 Genetic algorithm accuracy for RNA secondary structure prediction

RNA	Nucleotides	Base pair	Correctly predicted base pair (%)			
			Rank 1	Rank 10	Best structure	Any structure
tRNA	1556	432	87.7 ± 12.4	81.2 ± 12.5	98.8 ± 2.7	99.8
5S rRNA	3004	910	95.3 ± 7.0	87.9 ± 7.3	98.6 ± 4.3	98.7

Source: Chen, J., Le, S. & Maizel, J. (2000) Prediction of common secondary Structures of RNAs: a genetic algorithm approach. Nucleic Acids Research, 28, 991-999.

The optimal solutions in Chen's version of the genetic algorithm for RNA secondary structure prediction have been shown to correctly predict on average about 87.7% of the known base pairs of a tRNA structure with at least one of the top ten structures predicting on average about 98.8% of the known base pairs. All together the top ten structures predict an average of about 99.8% of the known base pairs within a structure, a very good result.

Chen's genetic algorithm takes $O(n^2)$ comparison time where **n** represents the maximum number of stems from all the structures that were considered. The computational time required is $O(n^2m^2N^2)$ where **N** is the number of sequences and **m** is the maximum number of sequences found. Chen characterizes his algorithms as being an alternative for poorly determined structural domains or for molecules with few well-determined domains (Chen et al., 2000).

Table 2.2 Summary of algorithms complexities.

Algorithm	**Time Complexity**	**Comparison Time**
Genetic Algorithm without pseudoknots	$O(n^2m^2N^2)$	$O(n^2)$

In the case of prediction without pseudoknots, Chen's genetic algorithm is a very good example of an algorithm that is very accurate but is also highly computational. Most genetic algorithms are run simultaneously using several processors in order to accomplish this massive task of creating these generations. A single computer can easily handle one generation but in order to achieve high levels of accuracy it is often necessary to compute hundreds of generations over several runs of the genetic algorithm. For this reason, although highly accurate, genetic algorithms may not be the most efficient way of predicting structures especially very large structures.

2.3 Energy Minimization Models

Energy minimization models otherwise known as Mfold type algorithms or Zucker algorithms based on their originator, determine the secondary structure of RNA by calculating free energy. From these calculations, the optimal structure is

considered to be that which has the lowest free energy. The time and space necessary for computations in such an algorithm depend entirely on the type of legal structures allowed, and their free energies. The basic Mfold algorithm which excludes pseudoknots runs in $O(n^3)$ time with a space complexity of $O(n^2)$ where **n** is the length of the sequence being determined.

These types of algorithms are recursive in nature and consist of two parts, the fill and the traceback. The fill component of the Mfold algorithm takes most of the total computational time and space. During this part, the algorithm calculates and stores the minimum folding energy for each of the fragments contained in the structure. Smaller fragments such as pentanucleotides are used as a basis for this energy determination. The resulting calculations are saved in a matrix whose size is dependent on the length of the sequence being determined. The traceback component of Mfold type algorithms basically assembles the final structure by searching back through the folding energy matrix and adding one base at a time to the evolving structure. The time needed for a single traceback is very negligible as compared to the entire algorithm however several thousand tracebacks may be necessary in order to completely assemble the optimal structure. The reason for this large amount of traceback is because the amount of tracebacks just like the size of the energy matrix is dependent on the size of the sequence being determined as well as on the energy range being explored. Therefore, for large sequences and for wide thresholds in energy large amounts of tracebacks must take place in order for the optimal structure to be determined.

Mfold type algorithms are traditionally designed to output a single structure as the solution but can be modified by the user to output more if necessary. Since these algorithms work recursively to find the optimal folding for each sub fragment of the entire structure, the optimal folding of the entire structure is easily determined without any additional cost to time or space. This is the main reason pseudoknots are ignored in these types of algorithms singe they are made up of at least two stems and as a result influence the folding of all the sub-fragments involved. Excluding pseudoknots allows the algorithm to run must faster and much more efficiently computationally (Shapiro, 1999).

Table 2.3 Summary of algorithms complexities.

Algorithm	**Time Complexity**	**Space**
Mfold	$O(n^3)$	$O(n^2)$
Genetic Algorithm without pseudoknots	$O(n^2m^2N^2)$	

Mfold type algorithms represent the most computationally efficient class of algorithms for prediction without pseudoknots. The best of these algorithms runs the fastest with the least amount of pull on memory resources of any secondary structure prediction algorithm developed as of yet. However, it is necessary to have some prior knowledge of energy within structures in order to receive the best results. For this reason, including pseudoknots in such types of algorithms immediately begins to complicate things. This is mainly because very little is known about the energies involved in pseudoknot interactions. Even with this lack of knowledge though, it seems using the ideas from Mfold algorithms would

be the best starting place for developing new algorithms for structure prediction that do include pseudoknots.

2.4 Stochastic Context-free Grammars and Evolutionary History

In (Knudsen and Hein, 1999) a model was that incorporated evolutionary history into the prediction of RNA secondary structure. This model uses an alignment of RNA sequences as its input and the results are output as one single common structure. There are two main parts to this model; the first is the stochastic context-free grammars (SCFGs) that are able to give a probability distribution of structures. The second part is the evolutionary model begins with the creation of a set of phylogenetic trees. These trees are built using maximum likelihood estimations (ML) from the model and serve as a way to relate the sequences. Using these trees can reveal information about structures by tracking the mutation patterns in RNA sequences. After the trees have been built, maximum a posteriori estimation (MAP) is used to predict the structures of the sequences. This sort of estimation basically finds the most likely structure based on all the information known thus far. The accuracy of this algorithm is due mainly to the use of the prior distributions, which were determined by the SCFGs earlier. However, this accuracy is usually limited to small sequence sets. As long as the sequence is modest in length, these distributions will ensure accurate structure predictions from even just a small amount of related sequences.

One of the issues with this algorithm is that it needs a sequence alignment to perform optimally. The reason for this is because the MAP calculation is

performed on a series of sequences that are assumed to have a structural alignment. Also multiple sequences are necessary to create the phylogenetic trees, which are used by the MAP calculations. Another big issue with this model is that pseudoknots cannot be modeled. The reason for this is not because of time or space constraints but rather constraints on the use of the SCFGs. These SCFGs can model most long range interactions found in RNA structures but cannot model any type of crossing interactions. Since pseudoknots are mainly composed of crossing interactions they are unable to be modeled using SCFGs and cannot therefore be predicted using this algorithm. Another one of the limitations with this algorithm due to the use of SCFGs is that loop and stem lengths are automatically considered to be geometrically distributed. Some of these issues can be fixed by using hidden Markov models (HMMs) instead of the SCFGs however, more issues can develop along with these changes and the accuracy of the model is reduced.

The accuracy of this model was found to be comparable to that of Mfold algorithms having only slightly lower accuracy levels based on single sequence input which is the only way to input data for Mfold. The accuracy improves when a series of structures are used, especially if they are properly aligned. Knudsen and Hein found that most of the inaccuracy in the models came from regions where pseudoknots were known to occur and that taking into account the phylogenetic data helped improve their accuracy level by approximately 5%. (Knudsen and Hein, 1999)

Using what has been seen to work and even what has failed new algorithms are being introduced which try to conquer this task of including pseudoknots. Most of these newer more complex algorithms find their basis in these simpler algorithms, which have been discussed here. A few of these algorithms that include pseudoknots are discussed next in the following chapter.

CHAPTER 3

ALGORITHMS WITH PSEUDOKNOTS

In chapter 2, simple algorithms were described which predicted RNA secondary structures while excluding pseudoknots. In this chapter, five algorithms are discussed which do take into account the presence of pseudoknots in RNA structures. Although each of these algorithms has a certain time and space complexity, they all contribute to the theory that RNA secondary structure prediction with pseudoknots is in fact possible.

3.1 Energy Based Predictions

In a recent paper Rivas and Eddy use a very specific representation to diagram the RNA secondary structures commonly formed (Rivas and Eddy, 1999). This representation allows them to handle the more complex structures like pseudoknots for algorithmic purposes. A very important feature in their representation is that there can be no more then two bases that are interacting at once. The RNA's backbone is represented as a straight line with the 5' end always placed on the left. Secondary interactions are represented as wavy lines that connect the two positions that are interacting along the backbone. Figure 3.1 shows this representation with common structures found in RNA, **H** represents a hairpin, **S** a stem, **B** a bulge, **IL** are internal loops and **M** are multiloops. In this representation you can easily see that any non-nested structure is considered to be a pseudoknot.

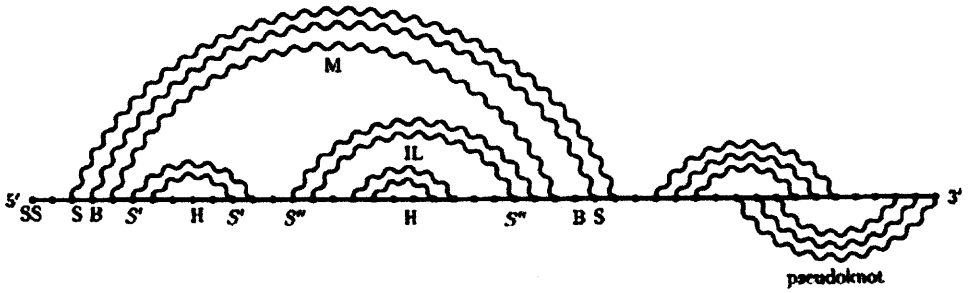

Figure 3.1 Representation of most relevant RNA secondary structures including Pseudoknots.

(Rivas and Eddy, 1999), describe their nested algorithm using two triangular matrices of size N x N. The matrix vx(i,j) is defined as the score of the optimal folding between the paired positions **i** and **j**. The other matrix, wx(i,j), is defined as the score of the optimal folding of positions **i** and **j** whether or not they are paired. For these pairs, the position **i** is defined as less than or equal to **j** since **i** is always found at the 5' end and **j** at the 3' of a fragment. After the algorithm has defined these two matrices they are filled recursively with the appropriate scores. For the matrix **vx** the scores added are highly influenced by the presence of hairpins, bulges, internal loops and multiloops found within the structure since **vx** is calculated using irreducible surfaces. The reason for this influence is that hairpin loops represent the irreducible surface of with an order of one. An order of one represents only one secondary interaction occurring such as with stems and bulges. Internal loops represent all the irreducible surfaces with an order of two otherwise known as multiloops. The optimal function for vx(i, j) is given below in figure 3.2. **EIS** stands for the scoring function of the irreducible surfaces, $\mathbf{P_I}$ represents the score for closing the multiloop and **M** is the score for starting the multiloop.

$$vx(i,j) = \text{optimal}\begin{cases} EIS^1(i,j) &] \ IS^1 \\ EIS^2(i,j:k,l) + vx(k,l) &] \ IS^2 \\ P_I + M + wx_I(i+1,k) + wx_I(k+1,j-1) &] \ \text{multiloop} \end{cases}$$

$$[\forall k,l \quad i \leqslant k \leqslant l \leqslant j]$$

Figure 3.2 Optimal recursive function for the vx matrix.

The **wx** matrix, on the other hand, is influenced by single-stranded nucleotides, external pairs and bifurcations. This matrix is used only when there are no external base pairs to represent. The recursive function used to fill this matrix can be seen below in figure 3.3. In this recursion **P** represents the score given to the external base pairs, and **Q** is the score for the single stranded nucleotides. Both **Q** and **P** in most cases are given approximate values of zero.

$$wx(i,j) = \text{optimal}\begin{cases} P + vx(i,j) &] \ \text{paired} \\ Q + wx(i+1,j) \\ Q + wx(i,j-1) &] \ \text{single-stranded} \\ wx(i,k) + wx(k+1,j) \quad [\forall k, \quad i \leqslant k \leqslant j]. &] \ \text{bifurcation} \end{cases}$$

Figure 3.3 Optimal recursive function for the wx matrix.

Since pseudoknots are non-nested structures it is impossible to represent them using only the two matrices described earlier so Rivas and Eddy added gap matrices to represent these more complex structures. The use of such gap matrices can be seen in figure 3.4 below which shows the graphical representation used by Rivas and Eddy for a simple pseudoknot structure. In

this diagram, two gap matrices are represented that have complementary holes. Putting the two matrices together forms the simple pseudoknot desired.

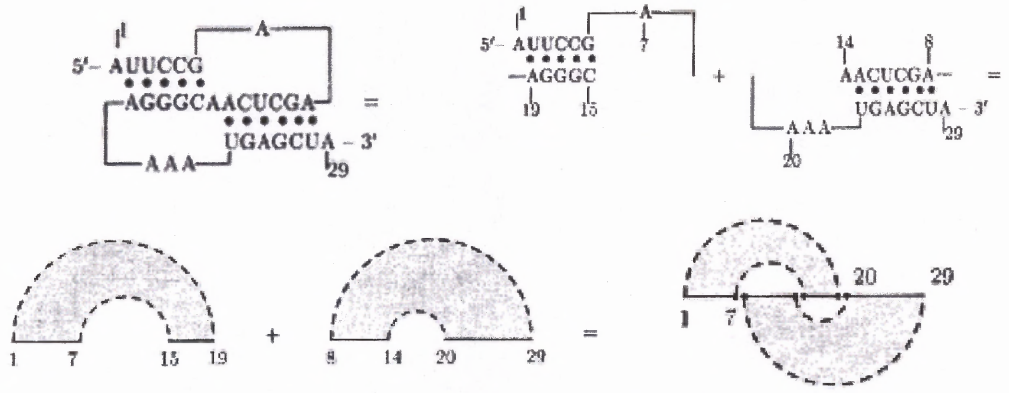

Figure 3.4 Construction of a simple pseudoknot using Rivas and Eddy's gap matrices.

In order to create the gap matrices the algorithm uses generalized forms of the **wx** and **vx** matrices known as whx(i,j:k,l) and vhx(i,j:k,l). The matrix whx(i,j:k,l) is defined as the optimal folding connecting the segments [i,k] and [l,j] and is used when the relationship between the four points is undetermined. On the other hand, vhx(i,j:k,l) is defined as the optimal folding connecting segments [i,k] and [l,j] when **i** and **j** are paired and **k** and **l** are paired. Since there are two other possible situations that can arise from the interactions of these four points two matrices are added into the algorithm at this point to cover all these possibilities. These two new matrices are yhx(i,j:k,l) which is defined when **k** and **l** are paired but the relationship between **i** and **j** is undetermined and zhx(i,j:k,l) which is defined when **i** and **j** are paired but the relationship between **k** and **l** is undetermined. Although these four new matrices are the basis for the

pseudoknot algorithm the original two matrices are still considered as special cases of **whx** and **vhx** when no hole is present in the gap matrix.

Using combinations of these matrices, Rivas and Eddy's algorithm is able to solve most RNA configurations that include pseudoknots as long as they can be broken down into a series of gap matrices that follow a set of defined recursion rules. Structures become unmanageable when they require more then two gap matrices for representation since this requires a higher order of calculations which at this point the algorithm is unable to handle (Rivas and Eddy, 1999). This particular algorithm can however handle overlapping pseudoknots and some non-planar pseudoknots as well as pseudoknot chains of any length. Whether or not an algorithm can handle chains of pseudoknots is a good judge of the algorithms ability. This is because a chain of pseudoknots can be thought of as a good representation of classes of much more complex structures. The fact that this particular algorithm can handle lengths of any size is a good testament to its ability to determine complex structures even though the algorithm cannot handle all complex knotted structures (Lyngso, 2000).

Rivas and Eddy's pseudoknot algorithm has a worst case time complexity of $O(N^6)$ with a storage of $O(N^4)$. These complexities are the main drawback for this particular algorithm since the large amount of time and space needed leads to constraints on the size of RNA that can be analyzed as well as on the complexity of the structures to be determined. Dynamic programming algorithms to determine RNA secondary structure all have some sort of inaccuracy within them due to some of the approximations that are necessary to

carry out all of the necessary calculations. This dynamic programming algorithm is no different; in fact it actually has a slightly higher level of inaccuracy due to the larger amount of legal structures the algorithm allows. This is mainly due to the fact that there is very little thermodynamic information available for pseudoknotted structures. Implementing more complete set of parameters as well as deciphering more thermodynamic information would take care of the inaccuracy of this algorithm leaving its complexity as its only drawback for structure prediction. (Rivas and Eddy, 1999)

Table 3.1 Summary of algorithms complexities.

Algorithm	**Time Complexity**	**Space**
Energy minimization with pseudoknots	$O(N^6)$	$O(N^4)$
Mfold	$O(n^3)$	$O(n^2)$
Genetic Algorithm without pseudoknots	$O(n^2m^2N^2)$	

All in all Rivas and Eddy's energy minimization model is a very good algorithm for structure prediction with pseudoknots. Much of the basis of this algorithm can be traced back to the Mfold algorithms and the predictions made there based also on energy minimization. Rivas and Eddy took the basic Mfold algorithm one step further and expanded it to include a wide variety of much more complex structures and interactions. In doing so they increased the time and space due to much more intense computations and lowered the average accuracy yet they still maintained an algorithm which accomplished the goals intended. This algorithm is a good starting point for any algorithm that includes pseudoknots because it gets the job done. If improvements could be made on its

time complexity and accuracy levels this algorithm would be the best choice for structure prediction with pseudoknots.

3.2 Exactly Clustered Stochastic Simulations

Xayaphoummine proposes that pseudoknots can be predicted using long-time-scale RNA-folding stochastic simulations by modeling the opening and closing of RNA helices (Xayaphoummine et al., 2003). Since stochastic simulations have the tendency to become inefficient, Xayaphoummine developed a generic algorithm that accelerates the simulations by exactly clustering the main short cycles that lie among the folding paths that have already been explored. Comparing these findings with non-clustered simulations helps predict the actual prevalence of pseudoknots within the RNA structures.

It was shown in chapter 2 how many algorithms exclude pseudoknots from calculations of RNA structure prediction due to their complexity. The presence of these more complex structures would have resulted in a much higher time and space complexity for the algorithm, along with the need for a more complex algorithm to deal with them. Xayaphoummine takes pseudoknots into account by using polymer theory to model the 3D constraints associated with these structures. Within this algorithm the entropy costs of such structures as pseudoknots and loops are evaluated all on the same basis. This is achieved by modeling these structures as a series of stiff rods that are connected to polymer strings. The rods are meant to model the helices found within these structures with the strings representing the unpaired regions. The only set back to using

such a method is that some hardcore interactions are ignored between bases and structures that could actually stereochemically prohibit certain other structures from forming. In practice however, this has been found to be rather limited and depends mainly on the G + C content of the sequence. Overall it has been seen to affect only about <1-10% of the predicted structures (Xayaphoummine et al., 2003).

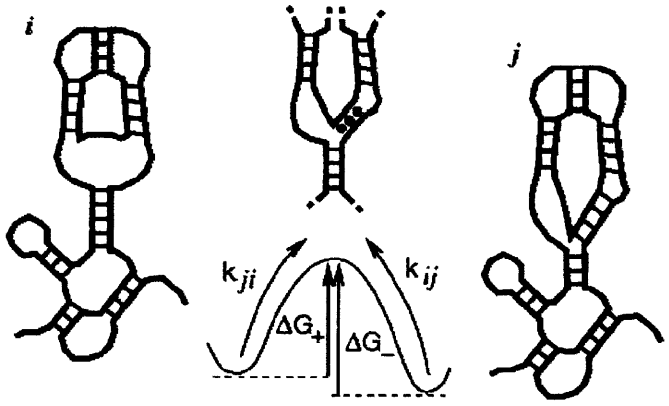

Figure 3.5 Stochastic transitions needed to open and close a helix

This stochastic modeling approach works by following a particular pathway involving the opening and closing of one helix at a time. The result is the possibility of either the elongation or the reduction of that particular helix in order for the structure to reach equilibrium. The equilibrium is considered to be the minimum size for that particular helix. For any given RNA sequence the proportion of helices in the structure is normally equivalent to approximately the square of the length of the sequence. Figure 3.5 shows the stochastic transitions necessary to open and close a helix surmounting the thermodynamic barrier to form a new helix.

The ECS algorithm involved in these stochastic simulations works to reduce the computational stain, which is usually involved with trying to simulate RNA-folding dynamics. In this algorithm, this is accomplished by clustering. Recently explored configurations are combined together into one cluster and can then be collectively revisited during further stochastic simulations. The clustered configurations loose their stochasticity in terms of individuality instead they take on a level collectively, which is proportionate to the number of configurations present within the cluster as a whole.

It has been shown that this approach to secondary structure prediction is efficient for predicting pseudoknots even though there is no actual optimal solution found. Instead, low-energy structures are determined and visited through simulations as the helices open and close. Xayaphoummine was able to use the algorithm to come to some conclusion about the percentage of pseudoknots present in RNA structures in general. Based on the results, pseudoknots are distributed from a few percent to over 30% in areas with a high G + C content. These results were consistent with experimentally determined RNA structures. Also it was shown that the average proportion of pseudoknots increases with the amount of G + C content in the sequence and independent of the length of the sequence. Finally, the algorithm showed that excluding pseudoknots from the legal structures creates a larger amount of inaccuracy then the small amount of inaccuracy that is possible from adding the additional calculations necessary to include them (Xayaphoummine et al., 2003).

Based on the results that Xayaphoummine found, it seems likely that using the ECS algorithm with other simpler structural models discussed earlier would be the best way to benefit from this algorithms strengths. Combining a more computationally efficient and accurate algorithm for structure prediction with this algorithm using it to predict where and how many pseudoknots may occur could result in a relatively effective model to predict RNA secondary structure with pseudoknots

3.3 Simple Dynamic Programming Algorithm with Pseudoknots

Several simple dynamic programming algorithms have been determined that does in fact take into account the presence of pseudoknots within the RNA secondary structure. Akutsu presented one such algorithm in (Akutsu, 2000). The basic problem Akutsu's algorithm looks at is the maximizing of the number of base pairs. This combined with some of the ideas used by a tree adjoining grammar approach suggested by Uemura in (Uemura et al., 1999) were combined in order to improve the basic secondary structure prediction algorithm using energy minimization in terms of accuracy of optimization. Tree adjoining grammars are used to generate trees by replacing all of the internal nodes of the current tree with a new tree, which is based on the rules of grammar. This approach is similar to the way context-free grammars generate strings by replacing all non-terminals with strings from the current string (Lyngso, 2000).

The basis of Akutsu's algorithm is a basic structure prediction algorithm that includes basic pseudoknots by maximizing the number of base pairs whenever a pseudoknot is encountered. This algorithm by itself has a time

complexity of $O(n^5)$ but this can be reduced to $O(n^{5-\delta})$ with a small modification. Scoring in this algorithm is based on the number of base pairs where ε, δ are fixed constants with the criteria that $0<\varepsilon$ and $\delta<1$. This modified algorithm is able to compute for most RNA sequences, a structure that has a score of at least $1-\varepsilon$ of the optimal. Although this complexity is still rather high, Akutsu suggests that combining these ideas with heuristics could lead to a more manageable algorithm, which contains very useful techniques for predicting pseudoknotted structures (Akutsu, 2000).

In Akutsu's algorithm the score function is defined as the total number of base pairs that are found within the secondary structure. This score represents a very generalized estimation of the free energy of the structure. The algorithm has a set of defined conditions that are meant to represent simple pseudoknots. A set of base pairs represented as $\mathbf{M}_{io,ko}$ are considered a simple pseudoknot only if these conditions are met and a set of base pairs called **M** is an RNA secondary structure with simple pseudoknots if $\mathbf{M}_{ih,kh}$ is a simple pseudoknot for a consecutive subsequence. (Akutsu, 2000)

The general recurrence scheme that Akutsu uses can be represented as an S-like shape since the involved pair is folded along the backbone to form an S pattern. The pseudoknot is then formed by pairing a base in the middle of the S form with one on the outer portion of the form. Crossing is not allowed however more complex recursive pseudoknots can be formed by allowing two bases on the same stem to pair. Figure 3.6 shows the basic recurrence scheme found in Akutsu's algorithm. The gray areas of the structures are defined structures

where the black areas are not. It is in these black areas where the addition of new base pairs takes place.

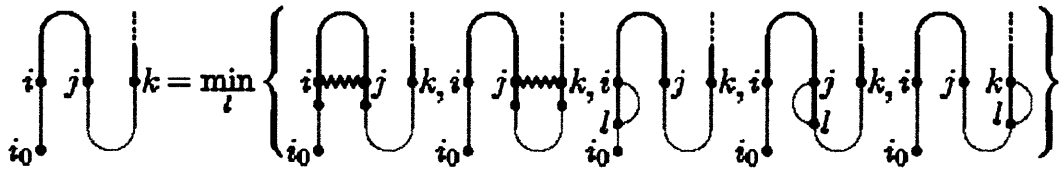

Figure 3.6 Recurrence scheme for Akutsu's algorithm

Using this type of S-shaped representation of the structure and recurrence scheme allows Akutsu's algorithm to find the structure, which possesses the least amount of energy. The basic element in Akutsu's recurrence is to find the minimum energy between a series of base pairs; these include **i** and $\mathbf{i_o}$ in one of the outer stems, as well as **j** and **k** in the other outer as well as the middle stem. All of these pairs can be seen in their S shape pattern in figure 3.6 above. In order to evaluate the structure either a base pair is added between the middle stem and one of the outer stems or an additional minimum energy structure is added to one of the stems. Recursive elements are built in this way with one element depending on another if and only if they share the same leftmost base represented as $\mathbf{i_o}$ in figure 3.6.

For the most part, Akutsu's recurrence is very similar to Rivas and Eddy's discussed earlier. However, Akutsu's runs in a time of $O(|n|^5)$ using $O(|n|^3)$ space both slightly faster and smaller than Rivas and Eddy's. One limitation is that Akutsu's algorithm can only handle pseudoknots in chains of exactly two. Also since Akutsu's recurrence is so similar to Rivas and Eddy's it makes sense that it cannot deal with any pseudoknot that are not included in Rivas and Eddy's

model. The fact that this algorithm is slightly faster than other algorithms that include pseudoknots with less space needed makes it easy to understand why the pseudoknots allowed are much more general then in the other algorithms. Decreasing time and space leads to an increase in the generality of the pseudoknots included. (Lyngso, 2000)

Table 3.2 Summary of algorithm complexities.

Algorithm	**Time Complexity**	**Space**
Dynamic Algorithm	$O(\|n\|^5)$	$O(\|n\|^3)$
Energy minimization with pseudoknots	$O(N^6)$	$O(N^4)$
Mfold	$O(n^3)$	$O(n^2)$
Genetic Algorithm without pseudoknots	$O(n^2m^2N^2)$	

3.4 Motif Algorithms

Motif algorithms such as RNAmotif (Macke et al., 2000) make use of both computations and public databases. RNAmotif is able to describe the structural elements found within RNA structures and then search through any available nucleotide database in order to find other organisms that contain the same structures. RNAmotif uses a rather flexible language to define its structures and can therefore specify any type of base-to-base interaction. A user-controlled section allows the user to add criteria and patterns that they choose to suite their needs. The algorithm can be classified as robust yet flexible since it is able to deal with any structural features found in RNA secondary structures from basic

base pair interactions to loops and hairpins all the way to pseudoknots and tetraloops.

Within the RNAmotif algorithm the structures are defined at the lowest level to be a set of descriptors, which distinguishes paired and unpaired positions in the structures. These descriptors can be given parameters such as length of a sequence, base pairings, mismatches and mispairings, which the user inputs along with the sequence. If necessary the user may define to have more then just base pairs, they can include both base triples and base quadruples as legal interactions since RNAmotif allows for all 16 types of legal base pairings

The actual algorithm for RNAmotif uses two stages; the first is known as the compilation stage. The compilation stage converts the users descriptor into a tree, which is based on the helical nesting of the motif. Every duplex that the user defines can be represent as a binary tree with the root of the tree representing the helix, the left sub-tree representing the motif on the interior of the helix and the right sub-tree representing the motif that immediately follows the defined helix. At this stage all of the consistency factors are checked to make sure they are correct. If the descriptor is not valid or no solution exists the algorithm will not continue. If the descriptor makes it through all the tests, the lengths of the motif elements are found and the limits to where each element begins and ends are computed. The second stage then moves on to perform a depth first search on the compiled tree. Table 3.3 shows the types of trees that can be created from user inputted descriptors and the order of symbols resulting from the depth first search on that tree.

The last row in Table 3.3 shows the representation of a pseudoknot using this descriptor motif. Pseudoknots are easily dealt with in RNAmotif by collecting all the duplexes involved in the formation of that particular knot into a single tree. The root of the tree is the helix at the 5' end of the structure, the left sub-tree represents all the helices contained within the pseudoknot and the right sub-tree represents all helices immediately following the pseudoknot.

Table 3.3 Conversion of structure motifs into an RNAmotif search list.

Symbol	Motif	Tree	Search List
ss	Single Stranded.	$ss\rightarrow$	ss
h5 ss h3	Standard helix forming a hairpin.	$h5\rightarrow$ $\downarrow$ $ss\rightarrow$	h5, ss
$h5_1$ ss_1 $h5_2$ ss_2 $h3_2$ ss_3 $h3_1$	Two standard helices arranged as an internal loop containing a hairpin.	$h5_1\rightarrow$ $\downarrow$ $ss_1\rightarrow$ $h5_2\rightarrow$ $ss_3\rightarrow$ $\downarrow$ $ss_2\rightarrow$	$h5_1$, ss_1, $h5_2$, ss_2, ss_3
$h5_1$ ss1 $h3_1$ ss_2 $h5_2$ ss_3 $h3_2$	Two consecutive standard helices each forming a hairpin.	$h5_1\rightarrow$ $ss_2\rightarrow$ $h5_2\rightarrow$ $\downarrow$ $\downarrow$ ss_1 $ss_3\rightarrow$	$h5_1$, ss_1, ss_2, $h5_2$, ss_3
$h5_1$ ss1 $h5_2$ ss_2 $h3_1$ ss_3 $h3_2$	Two standard helices arranged as a pseudoknot.	$h5_1\rightarrow$ $\downarrow$ $h5_2\rightarrow$ $ss_1\rightarrow$ $ss_2\rightarrow$ $ss_3\rightarrow$	$h5_1$, $h5_2$, ss_1, ss_2, ss_3

Source: Macke, T. et al. RNAmotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Research, 29, 4724-4735

After the formation of these binary trees based on the necessary descriptors the tree traversal begins by testing the first position of the target sequence against the left-most sub-motif of the descriptor for similarity. The algorithm is then called recursively to find all the solutions possible for the sub-motif in the left-most interior region. Once all the interior regions have been exhausted the region following the left-most motif is searched. The point is that

the algorithm is searching to find a solution that matches the original descriptor. For each complete solution obtained it is scored, ranked and stored. The scoring allows the user to add certain constraints on the motif that were omitted before due to their complexity in programming. It also helps reduce the number of candidate structures by automatically eliminating structures that fall below a certain threshold level. When all candidates at a certain level have been exhausted, the search is repeated moving one position over in the target sequence and starting from the top of the tree until the entire sequence has been searched. The structures with the highest score and lowest energy is returned as the optimal solution (Macke et al., 2001).

Motif algorithms work wonderfully when a well-maintained database is available to pull from and horribly when there is not. Like comparative modeling introduced back in chapter 2 the limitations of this model are dependant on the amount of experimental research available. If a lot is known about a certain type of RNA and its structures then using a motif algorithm is the easiest and best route to take but if experimental research is scarce this type of model should be avoided so as not to give results based on little or no knowledge.

3.5 Genetic Algorithms including Pseudoknots

In 1996 Shapiro introduced a version of the genetic algorithm that was able to deal with the presence of pseudoknots within the RNA structure specifically with H-type pseudoknots (Shapiro 1999). The algorithm itself is non-deterministic and works by iterating in parallel a three-step procedure that is supposed to mimic

evolution. Like the genetic algorithm presented earlier in chapter 2, these steps are selection, mutation, and crossover all of which work in the same basic way as the earlier algorithm.

The algorithm starts by generating a stem pool, which consists of all of the stems both fully, and partially zipped that are possible given the original sequence. This pool is then initialized by the processor as a series of small structures. For each generation, the algorithm selects two of these initialized structures and its eight neighbors using a ranking rule, which is biased towards lower free energies. These two structures labeled P1 and P2 are then mutated by addition of randomly chosen stems from the stem pool to form C1 and C2 the children of the original structures. In this step, stems that have the ability to form pseudoknot type interactions are included which was not the case for the previous genetic algorithm presented. A crossover operation is then performed between the two parents and the two children by taking stems from the parental structures and distributing them between the children. From here, the child with the lowest energy is selected as the best. For each of the iterations of the genetic algorithm there are 16,384 new structures created. The algorithm is finished when a stable population if formed.

In order to add in the feature of including stems that can form pseudoknots, a few issues have to be dealt with. First off and most importantly, some sort of definition has to be given as to what types of stems can and cannot form a pseudoknot. In this case this is defined as any two stems, which are not conflicting or overlapping can form a pseudoknot if and only if one side of one of

the stems is outside the bounds of the 5' and 3' ends of the other stem while the other side of the stem is within bounds on both ends. In order to implement this definition the algorithm keeps two lists of stems on each processor in such an order so that their 5' ends are always presented first. The first list contains the stems that form the secondary structure of the RNA sequence and the other contains all the stems that can complete the formation of pseudoknots. For each generation in the algorithm the child structures are based on the stems from the first list forming an intermediate secondary structure. Once these structures are formed, the compatible stems from the second list are added whenever possible in order to form H-type pseudoknots. H-type pseudoknots are the most common for of pseudoknots and are the result of a hairpin loop binding with a single stranded region. Not all structures will have possible pseudoknots so for this reason in each generation, structures with and without pseudoknots will compete to be the optimal structure.

Another issue with including pseudoknots in the genetic algorithm is how to deal with their free energy. Since the best structure from every generation is selected based on their free energy it becomes very necessary to have a set way of dealing with pseudoknots in terms of energy. Since further experiments need to be performed on pseudoknots before a set energy rule can be determined Shapiro deals with this issue by using a simple energy rule. This simple rule assumes that a stable H-type pseudoknot can have connecting loops that are no longer then 16 nucleotides. With this assumption a value of 4.2 kcal/mol is assigned to each of the loops a positive value since the loop destabilizes the

structure. The free energy of the pseudoknot itself is then defined as the sum of the energies of the two stems that are involved along with their stacking energies and the energies of the two connecting loops.

Although the two main issues of including pseudoknots into a genetic algorithm are dealt with here there are still many limitations to using it, mainly the fact that only H-type pseudoknots can be determined at this time. This is most likely due to the fact that there is not a lot of experimental data on pseudoknots out there. As more research is completed and more pseudoknot structures analyzed more rules can be defined and added to such existing algorithms to form more complete and accurate structure prediction tools. This particular algorithm for example, is used as part of an RNA structure analysis workbench called Structurelab, which uses a series of algorithms to determine the secondary and tertiary structures of RNA (Shapiro, 1999).

Workbenches such as this are a good way of finding a compromise between existing algorithms. If one algorithm is known to do well in one area and another in a completely different then use both together to determine the structure and your accuracy will increase. Running multiple algorithms can obviously be computationally draining for a single processor but if they can be automated into a server such as the one that Structurelab runs on they will serve as the best way of determining structure both with and without pseudoknots.

CHAPTER 4

PSEUDOKNOT DATA SET

Very few pseudoknots have been experimentally determined in the lab. This is one of the main reasons why it is so hard to predict these structures using automated algorithms. Without the experimental data about energy and folding patterns for these structures it becomes a trial and error game to determine what other information is available that will be useful in determining structure prediction patterns. This chapter will introduce some of the pseudoknots that have been experimentally determined using such techniques as X-ray crystallography and nuclear magnetic resonance (NMR). This small sample set can be used in the future for testing both the present algorithms presented here and future algorithms to determine how efficient they are for predicting RNA structures that contain pseudoknots.

4.1 Turnip Yellow Mosaic Virus

The turnip yellow mosaic virus was determined by NMR technique in 1998 and contains 1056 residues. Within the structure a simple pseudoknot was found on the acceptor arm of the structure containing 44 of the bases and no water molecules. The structure is very reminiscent of tRNA structures. The pseudoknot's nucleic acid sequence was determined to be: G-G-G-A-G-C-U-C-A-A-C-U-C-U-C-C-C-C-C -C-C-U-U-U-U-C-C-G-A-G-G-G-U-C-A-U-C-G -G-A-A-C-C-A. Figure 4.1 shows a 3D representation of the pseudoknot determined

from the NMR deciphering of the turnip yellow mosaic virus (http://www.rcsb.org/pdb/).

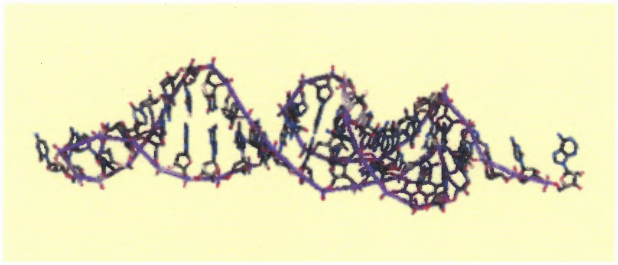

Figure 4.1 3D representation of the turnip yellow mosaic virus.

4.2 Simian Retrovirus Type-1

This particular simian retrovirus was determined using the NMR technique in 2000. The structure contains 540 residues total with 36 being involved in the pseudoknotted structure. This small fragment in chain A contains the sequence: G-C-G-G-C-C-A-G-C-U-C-C-A-G-G-C-C-G-C-C-A-A-A-C-A-A-U-A-U-G-G-A-G-C-A-C. This particular pseudoknot is involved in ribosomal frameshifting and can be seen in the 3D representation below in figure 4.2 (http://www.rcsb.org/pdb/).

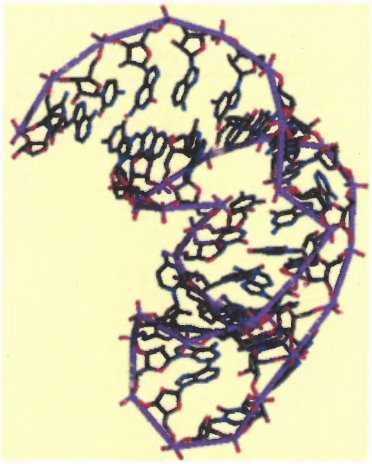

Figure 4.2 3D representation of simian retrovirus type-1

4.3 HIV Type-1

HIV Type-1 is considered here to be a complex of nucleotidylitransferase and RNA, which was deciphered experimentally by X-ray diffraction in 1998. This extremely large complex contains multiple chains and a huge amount of varied motifs however approximately 120 residues are involved in a complex with a 33 base pseudoknot. This extremely complex structure gives a variety to a data set in that this is not a simple structure to decipher yet the pseudoknot contained within is relatively small. This type of structure helps to test the limits of an algorithm in terms of size and complexity as well as in terms of inclusion of pseudoknots. Figure 4.3 below shows a 3D representation of the various chains involved in this HIV complex (http://www.rcsb.org/pdb/).

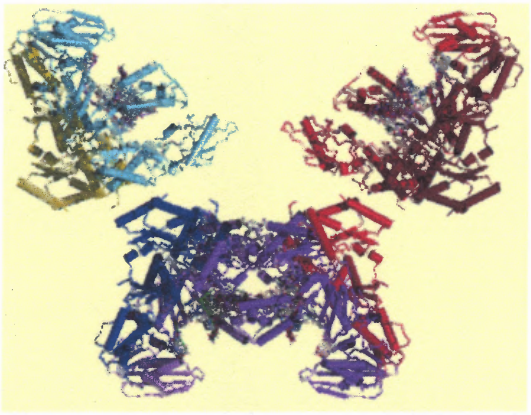

Figure 4.3 3D representation of HIV type-1

4.4 Mouse Mammary Tumor Virus

The mouse mammary tumor virus was deciphered in 1996 using the NMR technique. This frameshifting pseudoknot contains 32 bases which are: G-G-C-G-C-A-G-U-G-G-G-C-U-A-G-C-G-C-C -A-C-U-C-A-A-A-A-G-C-C-C-G. This

is a relatively small structure in terms of bases which can seen represented along the purple line in the 3D representation in figure 4.4. However, just because this structure has relatively few bases does not mean that the structure is not complex as can also be seen below. This makes this structure a good addition to a data set as you try to test the limits of an algorithm and its accuracy (http://www.rcsb.org/pdb/).

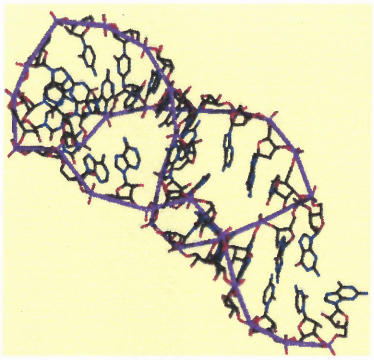

Figure 4.4 3D representation of mouse mammary tumor virus

4.5 Pea Enation Mosaic Virus Type-1

The pea enation mosaic virus type-1 was determined in 2002 experimentally using NMR. The entire structure contains a total of 405 residues however, chain A contains the pseudoknot and consists of only 28 bases, which are: U-C-C-G-G-U-CH-G-A-C-U-C-C-G-G-A-G-A-A-A-C-A-A-A-G-U-C-A. The pseudoknot contained in this chain is p1-p2 frameshifting pseudoknot, which is naturally occurring. 15 models in all have been determined for this particular viral structure all by use of NMR. From these models a 3D

representation has been created which can be seen below in figure 4.5 (http://www.rcsb.org/pdb/).

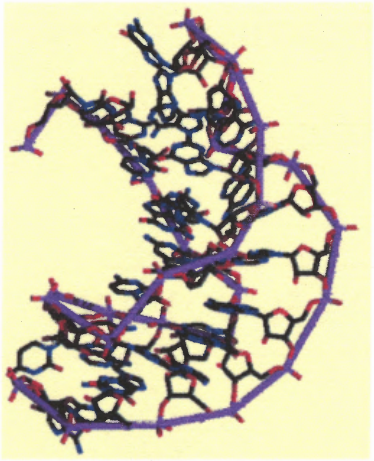

Figure 4.5 3D representation of pea enation mosaic virus type-1

4.6 Beet Western Yellow Virus

This RNA sequence was taken from the beet western yellow virus and contains 28 bases which are: GTP-G-C-G-C-G-G-C-A-C-C-G-U-C-C-G-C-G-G -A-A-C-A-A-A-C-G-G. This structure was deciphered in 2002 using X-ray diffraction techniques. This molecule is not very large but in contains some features that the previous examples have not and that is that it has metal ions within the molecule. These metal ions are not a part of the sequence obviously but they do influence the structures shape. Including diversified molecules in a data set is essential in order to completely test something like a structure prediction algorithm. Without the diverse data set there is no way of knowing what the limitations are of such algorithms. Without knowing these limitation we have no way of knowing how accurate our solutions actually are. Figure 4.6 below shows

a representation of the beet western yellow virus and you can see where the large metal ions are inserted inside the molecule influencing the final shape of the structure (http://www.rcsb.org/pdb/).

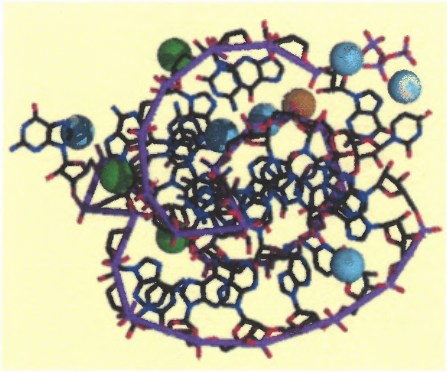

Figure 4.6 3D representation of the beet western yellow virus

CHAPTER 5
DISCUSSION OF FUTURE PATHS FOR ALGORITHMS

In the past few chapters' algorithms have been presented which predict RNA secondary structure with and without the inclusion of pseudoknots. The result was shown that algorithms that do not include pseudoknots will have a smaller time and space complexity but are not necessarily more accurate then those that do include pseudoknots. Comparative modeling seems to be the best way overall to predict secondary structures, as it is the most accurate. However, there are many problems and obstacles that go along with using a model such as this. First off, many known structures must be retained before such a model can be successful and for this reason, at this time this model can not be successful with pseudoknots since so little is known about them and so few structures have been determined experimentally. In the future, this algorithm will be a very useful tool for RNA prediction and so for now other types of algorithms must be focused on.

Genetic algorithm approaches both with and without pseudoknots were discussed previously. Addition of pseudoknots to the algorithm added to its complexity by adding an additional level of computations to each generation in the genetic populations. The basis of both algorithms are essentially exactly the same they both uses lists of stems to hold the possible structures that can be formed the only difference was that in Shapiro's version of the genetic algorithm (Shapiro, 1999) an extra list was created to hold the stems which were able to form pseudoknots. After the initial formation of the population a pseudoknot stem

was added wherever it was possible to form a new population, which contain some structures with and some without pseudoknots. The addition of these two extra steps in the algorithm resulted in a population that was more accurate to real RNA structures with just a minimum amount of extra calculations. These two algorithms are a very good example of making the necessary changes in order to accommodate more possible legal structure. But even in this case, Shapiro's genetic algorithm does not cover all of the possible structures that can be formed, even though it does shows where to go in order to improve existing algorithms to include more and more structures as they are deciphered experimentally.

Energy based models are the most common types of algorithms found for RNA secondary structure prediction. Tons of algorithms can be found that tweak different calculations in different areas to form slightly new algorithms with perhaps slightly higher accuracy or maybe even slightly lower time and space complexities. Algorithms not including pseudoknots are mostly known as Mfold type algorithms based on the original algorithm presented by Zucker back in the 1970's. These algorithms have evolved since then but the current Mfold algorithms are still the fastest of all secondary structure prediction algorithms and are usually the most accurate of those not including pseudoknots. The time complexity of $O(n^3)$ and space complexity of $O(n^4)$ is considered to be the best possible complexities for prediction algorithms. The reason for this is that these algorithms use simple recursive functions to calculate the energy of structures found within the final structure and from there use basic traceback methods to

assemble the final structure outputted by the program. The algorithm needs very little space since it doesn't save any unnecessary information and it keeps its calculations as basic as possible in order to keep a short running time. How is it then that these simple algorithms can be so accurate keeping everything so basic? The reason is first because it calculates the aspect of the structure, which influences the folding of the structure the most, its energy. Secondly because it avoids more complicated calculations by ignoring interactions whose energies have not been determined, in other words pseudoknots. If the energy to form a particular structure is extremely high it wont be formed in nature and it wont be chosen in this program. It keeps things simple and as a result it gets the job done in most cases having most of its inaccuracies in areas where pseudoknots are commonly found. Rivas and Eddy, (Rivas and Eddy 1999) built on this theory in order to create an algorithm, which takes care of some of these ignored interactions. Like Shapiro's genetic algorithm, Rivas and Eddy had to first come up with a new way of representing the structures with pseudoknots. After that, they could use energy-based calculations like those in the Mfold algorithms to determine the most optimal structures. They went about this by using a series of matrices to represent both structures with and without pseudoknots. The result was an algorithm with a time complexity of $O(n^6)$ and space of $O(n^4)$ much slower than the normal Mfold algorithms but using the same amount of space. However, although Rivas and Eddy's algorithm is much slower it takes into account many structures ignored by Mfold algorithms. This algorithm can handle regular pseudoknots, overlapping pseudoknots, some non-planar pseudoknots

and even chains of pseudoknots of any length, certainly much more than Mfold algorithms and even much more than Shapiro's genetic algorithm.

Akutsu uses recursive elements similar to Rivas and Eddy's along with ideas for using tree-adjoining grammars to create a more general pseudoknot algorithm. The result is that the algorithm runs in slightly less time and uses less space having a time complexity of $O(|n|^5)$ and space of $O(|n|^3)$ (Lyngso, 2000). The problem as a result of this smaller time and space is that the amount of pseudoknots included is much less than that of Rivas and Eddy. For example, only pseudoknots in chains of no more than two can be handled whereas Rivas and Eddy's model can handle chains of any length.

Knudsen and Hein propose (Knudsen and Hein, 1999) that using a combination of stochastic free grammars and evolutionary history is the best method for predicting RNA secondary structure however, evidence does not necessarily support this. The results of Knudsen and Hein's research showed an accuracy level that was slightly lower than that of Mfold algorithms and this reduction in accuracy is not the result of added calculations for pseudoknots. In fact, this algorithm cannot be used for pseudoknots at all and it cannot be modified to incorporate them either. Knudsen and Hein found that most of the inaccuracy of the algorithm was in areas where pseudoknots were known to be found. The reason this algorithm cannot handle pseudoknots is because the stochastic context free grammars are not equipped to handle crossing interactions, the interaction which defines a pseudoknot. However, even though it is impossible to use this particular algorithm to determine pseudoknots its still

remains a highly useful algorithm. The reason for this is because of the incorporation of evolutionary history into the algorithm. This algorithm not only predicts structures but it keeps track of structures in a tree form that is very useful for further research on the distribution of structures found in RNA.

Xayaphoummine uses exactly clustered stochastic simulations to predict pseudoknots by modeling the opening and closing of RNA helices (Xayaphoummine et al., 2003). This particular algorithm is very insightful in finding regions where pseudoknots can be found however it is not as great when it comes to determining a final optimal solution. Instead low energy structures are returned but no one conclusive solution is found. This algorithm is good at illustrating however, that exclusion of pseudoknots creates a larger issue then actually incorporating them might result in since the amount of inaccuracy in the model can fall quite drastically when such structures are omitted. The main reason for this is because it has been shown that pseudoknots can make up over 30% of the structures in areas where the G + C content is very high. If these structures were to be ignored in such areas, this would lead to at least 30% of the structure being inaccurate a huge disadvantage for such algorithms (Xayaphoummine et al., 2003).

Motif algorithms make use of computations and public databases to come up with an optimal structure (Macke et al., 2001). The calculations in this algorithm describe the structural elements found within the structure, which are then taken and compared against nucleotide databases in order to find organisms that contain similar structures. The result of this program is a series of

structures that match the original descriptors. The use of a ranking function that is user controlled results in the elimination of certain structures that fall below a particular threshold. At this point the program continues recursively until one final optimal structure is returned. Using this type of algorithm has its benefits but it also creates a lot of issues. On the plus side this algorithm combines two types of models and can be therefore very accurate and very useful if the correct information is around. On the other hand, it requires the presence of databases that have the correct information and since not much is known about RNA secondary structure experimentally this can be a problem and can lead to some level of inaccuracy.

In the future the best path for creating new algorithms for RNA secondary structure prediction is to combine what has been found in the past. Using aspects of both algorithms with and without pseudoknots or even different types of pseudoknot predicting algorithms can result in much higher levels of accuracy. In terms of time and space, it seems logical that no algorithm will be found that includes pseudoknots that will be any faster with less space being used then the optimal Mfold algorithm. However, it has been shown through various types of algorithms that pseudoknots can be included in a model while retaining a time complexity that is polynomial. Whether or not this would remain the case if all possible structure were included in an algorithm remains to be proven. However, for the time being a reasonable algorithm can be determined using information from these past algorithms that would result in a fairly accurate and fast algorithm for secondary structure prediction with pseudoknots.

REFERENCES

Akutsu, T. (2000). Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. Discrete Applied Mathematics, 104, 45-62.

Chen, J., Le, S., Maizel, J. (2000). Prediction of common secondary structures of RNAs: a genetic algorithm approach. Nucleic Acids Research, 28, 991-999.

Han, K., Lee, Y., Kim, W. (2002). PseudoViewer: automatic visualization of RNA pseudoknots. Bioinformatics, 18, S321-S328.

Knudsen, B., Hein, J. (1999). RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics, 15, 446-454.

Lyngso, R.B., Pedersen, C.N.S. (2000). Pseudoknots in RNA secondary structures. Proceedings of the 4th Annual International Conference on Computational Molecular Biology. Tokyo, April 2000.

Lyngso, R.B., Pedersen, C.N.S. (2000). RNA pseudoknot prediction in energy based models. Journal of Computational Biology, 7(3/4), 409-428.

Macke, T.J., Ecker, D.J., et. al. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Research, 29, 4724-4735.

Rivas, E., Eddy, S. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of Molecular Biology, 285, 2053-2068.

Rivas, E., Eddy, S. (2000). The language of RNA: A formal grammar that includes pseudoknots. Bioinformatics, 16, 334-340.

Shapiro, B.A., Bengali, D., Kasprzak, W., Wu, J.C. (2003). Exploring RNA intermediate conformations with the massively parallel genetic algorithm. In J.T.L. Wang, C.H. Wu, P.P.Wang (Ed.), Computational Biology and Genome Informatics (pp.1-34). Toh Tuck, Singapore: World Scientific Publishing Company.