6.047 Computational Biology Project Proposal

Novel RNA Folding Method Using Classification And Clustering Techniques

Nicole Power and Lily Seropian

November 4, 2013

Introduction

RNA is a very biologically important molecule. It is the intermediate between DNA and proteins, thus vital to gene expression. It also plays catalytic roles in the cell, most notably in ribosomes. The many different types of RNA—mRNA, tRNA, rRNA, piRNA, microRNA, and others—have diverse functions as a result of their diverse forms. Unlike DNA, RNA is capable of forming single-stranded doubled helices, a result of bonding with itself. RNA function would be better understood if we knew its form, but the tertiary structure is complex to predict. Secondary structure prediction, however, is within the grasp of our current tools, and can greatly expand our understanding of RNA. From properties of the secondary structure, functional attributes of the specific RNA molecule can be gleaned. As with proteins, in RNA function follows form, so looking at RNA's form helps researchers understand how specific RNA molecules bind and interact with other molecules. RNA secondary structure is also useful to study because it is evolutionarily well preserved, allowing us to better find and understand non-coding RNAs.

Existing algorithms that have sought to predict RNA secondary structure have been based either on thermodynamics or probabilities. Nussinov's Algorithm uses Dynamic Programming to recursively calculate the substructure of subsequences of the RNA that minimize free energy. The Zuker Algorithm is also structured to minimize free energy, but it also factors in the contribution of stacking energy. The existing algorithms which approach the RNA folding problem probabilistically are McCaskill, CYK, and Inside-Outside Algorithms. The McCaskill Algorithm uses a partition function and pair probabilities, while the CYK and Inside-Outside Algorithms use SCFGs. These algorithms all run in $O(n^3)$ time.

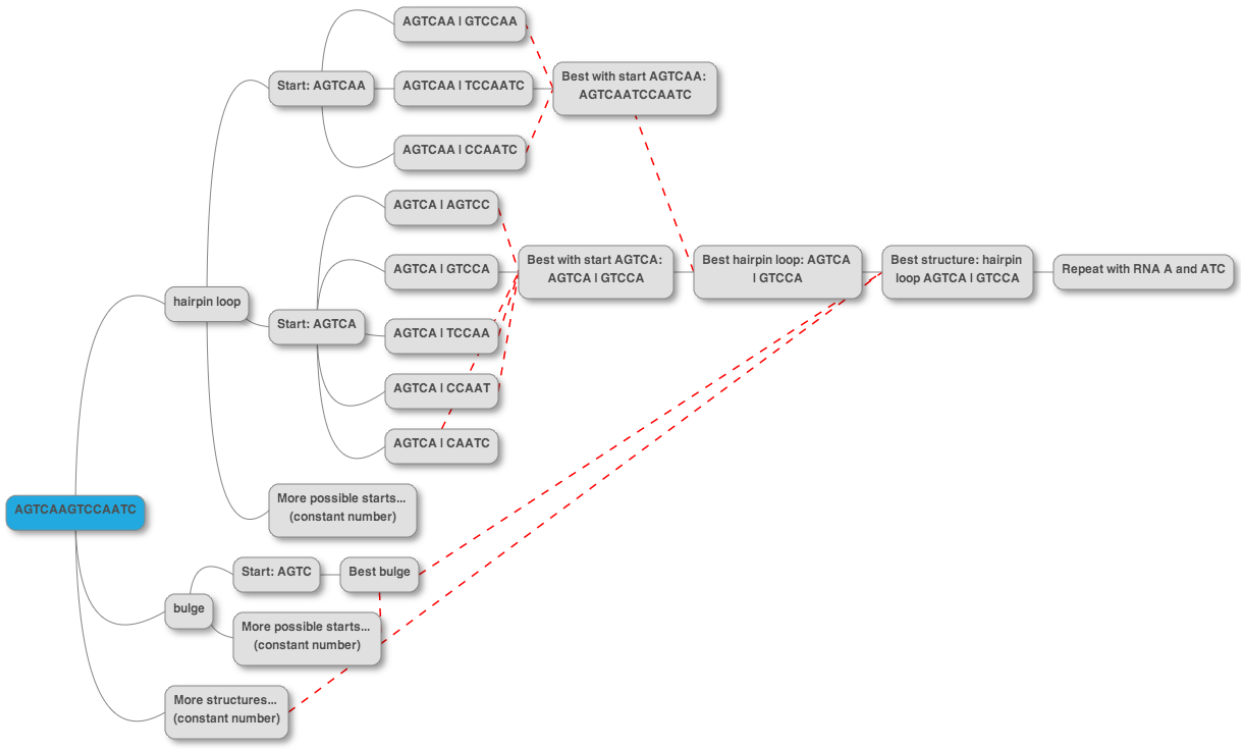
We would like to approach the RNA Folding problem in a new way, using classification and clustering. We will first develop a partitioning algorithm to break RNA into viable subsequences, which we will then seek to classify into secondary structures (loops, bulges, etc). We believe that this approach, by looking at sections of RNA rather than the whole, will improve the runtime of RNA folding prediction, with an ideal runtime of $O(n^2)$.

Approach

Overview:

- Start with an RNA sequence
- For each structural element (loop, bulge, etc.):
    - Find a constant number of possible lengths (in base pairs) for that element based on a probability distribution learned from annotated sequences
    - For each length:
        - Fix the first half of the subsequence of that length at the beginning of the RNA
        - For each possible second half in the remainder of the RNA:
            - Use the classifier/clusterer to score the subsequence composed of the two halves. A higher score indicates that the two halves are more likely to be part of the given structural element.
        - Based on the scores, determine the best ending half for the start half
    - Based on the scores, determine the best length
- Based on the scores, determine the best structural element
- Remove the RNA used in the best solution from the sequence, and repeat on the remaining part of the sequence until all of the original sequence has been classified.

Example:



Specific Aims

Aim 1: Develop a partitioning algorithm that enables a significant improvement in runtime over existing RNA folding algorithms.

Traditional RNA folding algorithms consider the entire segment of RNA. While these are able to achieve ever increasing success, they also require the time to classify the entire, potentially long, segment, and typically take $O(n^3)$ time to compute. We will develop a partitioning algorithm that will split the RNA up into segments representing structural elements, such as stems, hairpin loops, and bulge loops. This will be done by learning a probability distribution from an annotated dataset over the length of each different structural element. The first half of a subsequence will be taken from the beginning of the RNA, and for the second half all subsequences of that size in the remainder of the RNA will be considered. The small segments resulting from the partitioning will be classified individually, which takes significantly less time than folding the entire RNA, since the size of the problem becomes constant instead of linear in the length of the RNA. The total runtime is therefore $O(n^2 \cdot T_{classification})$, but since the segments to be classified are of constant length, it is $O(n^2)$. The accuracy may suffer slightly because information about transitions between structural elements will be ignored, but because a structural element is a biologically logical unit of RNA, the accuracy should not suffer significantly.

Aim 2: Implement classification and clustering algorithms to use in conjunction with the partitioning algorithm to compute RNA folding.

The partitioning algorithm will break the RNA up into segments that will then be classified. We will experiment with both a classification and a clustering approach. With classification, we will train a Naive Bayes classifier on our labeled dataset, and then use the classifier to identify the structure of a subsequence of RNA. With clustering, we will build a generative model by creating clusters corresponding to each structural shape from our dataset. We then compute the probability that our subsequence was generated by each cluster, assigning the subsequence to the cluster with maximum generative probability. In both cases, we simply look at all of the assignments for each subsequence, make a decision about the current segment, and move on to the next segment, to get our folding for the entire sequence.

Aim 3: Determine parameters and approach that yield the best results, and compare accuracy and runtime with existing algorithms.

There is room for experimentation in both how the partitioning is done and how the generative models are constructed. We will experiment to find good results, determine whether the classification or clustering approach is more successful, and do a comparative analysis between the two, as well as with existing algorithms. We will use cross-validation to judge the performance of our algorithms, and percent accuracy to compare with existing algorithms.

Timeline

Due Date	Task
11/4/13	Revise proposal, have datasets
11/18/13	Midcourse report: partitioning algorithm and Naive Bayes should be implemented, analysis should be started
11/22/13	Finish analysis of Naive Bayes, implement clustering
11/28/13	Analyse clustering
12/7/13	Final report
12/11/13	Presentation

Resources

Unfortunately, we have not yet found a mentor.

For this project, we will be using data from http://www.rnasoft.ca/strand/.

The relevant lectures and chapters we will consult are as follows:

- Lecture 8: RNA Folding, Four Russians, and Stochastic Grammars
- Recitation 5: Stochastic Context-Free Grammars
- Chapter 10: RNA folding
- Chapter 15: Gene Regulation II: Classification

Collaboration

Nicole will implement the Naive Bayes classifier and do most of the analysis. Lily will implement the partitioning algorithm and the generative cluster model. Both will contribute equally to all reports.

References

Washietl, S. *et al.* (2012) RNA folding with soft constraints: reconciliation of probing data and

thermodynamic secondary structure prediction. *Nucleic Acids Res.,* **40**, 4261-4272.

Nussinov R, Jacobson AB, Fast algorithm for predicting the secondary structure of single-stranded RNA.Proc Natl Acad Sci U S A. 1980 Nov; 77:(11)6309-13

Zuker M, Stiegler P Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981 Jan; 9:(1)133-48

McCaskill JS The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990; 29:(6-7)1105-19

Dowell RD, Eddy SR, Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics. 2004 Jun; 5:71

Do CB, Woods DA, Batzoglou S, CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006 Jul; 22:(14)e90-8