6.047/6.878 Lecture 08: RNA Folding

Guest Lecture by
Stefan Washietl (wash@mit.edu)
Scribed by Sam Sinaei (samsinai@mit.edu) (2010)
Scribed by Archit Bhise (archit@mit.edu) (2012)

October 19, 2012

Contents

List of Figures

1 Motivation and Purpose

RNA **(Ribonucleic acid)** is a very important molecule. The aim for this chapter is to understand the methods that can explain, or even predict the secondary structure of RNA.

To accomplish this, we first look at RNA from a biological perspective and explain the known biological roles of RNA. Then, we study the different methods that exist to predict RNA structure. There are two main approaches to the RNA folding problem: 1) predicting the RNA structure based on thermodynamic stability of the molecule, and looking for a thermodynamic optimum 2) probabilistic models which try to find the states of the RNA molecule in a probabilistic optimum.

Finally, we can use evolutionary data in order to increase the confidence of our predictions by these methods.

2 Chemistry of RNA

RNA is consisted of a 5-carbon sugar, ribose, which is attached to an organic base (either adenine, uracil, cytosine or guanine). The biochemical difference between DNA and RNA is in two places : 1) the 5-carbon sugar has no hydroxyl group in the 5 position 2) the uracil presence in the RNA which is the non-methylated form of thymine instead of just thymine. The presence of ribose in RNA makes its structure more flexible than DNA, therefore the RNA molecule is able to fold and make bonds within itself which makes the single stranded RNA more stable than single stranded DNA.

3 Origin and Functions of RNA

The initial belief about the RNA was that it acts as an intermediate between the DNA code and the protein, which is true. However, in early 80s, the discovery of catalytic RNAs (ribozymes) expanded the perspective on what this molecule can actually do in living things. Sidney Altman and Thomas Cech discovered the first ribozyme, RNase P which is able to cleave off the head of tRNA. Self-splicing introns (group I introns) were also one of the first ribozymes that were discovered. They do not need any protein as catalysts to splice. Single or double stranded RNA also serves as the information storage and replication agent in some viruses.

The **RNA World Hypothesis** proposed by Gilbert which tries to explain the origin of life, relies on the fact that RNA can have both information storage, and catalytic activity at the same time, which are the fundamental characteristics that a living system needs to have. In short, it says in the beginning RNA molecules were the first replicators, and because they were catalytic at the same time, it was possible for them to replicate without dependency on other molecules. Although to this day, there are no natural self-replicating RNA found in vivo, self-replicating RNA molecules have been created in lab via artificial selection. For example, a chimeric construct of a natural ligase ribozyme with an in vitro selected template binding domain has been shown to be able to replicate at least one turn of an RNA helix.

Through evolution, RNA has passed its information storage role to DNA, because it is more stable and less prone to mutation and acted as an intermediate between DNA and proteins, which took over some of RNAs catalytic role in the cell. Thus, scientists sometimes refer to RNA as molecular fossils. Nevertheless, RNA still plays an important catalytic role in the living organisms. For instance, the catalytic portion of the ribosome i.e. the main functional part of the ribosomal complex consists of RNA. RNA also has regulatory roles in the cell, and basically serves as an agent for the cell to sense and react to the environment.

3.1 Riboswitches

Regulatory RNAs have different families, and one of the most important ones are **riboswitches**. Riboswitches are involved in different levels of gene regulation. In some bacteria, important regulations are done by simple RNA families. One example is the thermosensor in Listeria, a riboswitch that blocks the ribosomes at low temperature (since the hydrogen bonds are more stable). The RNA then forms a semi-double stranded conformation which does not bind to the ribosome and turns the ribosome off. At higher temperatures (37 C), the double strand opens up and allows ribosome to attach to a certain region in the riboswitch, making translation possible once again. Another famous Riboswitch is the adenine Riboswitch (and in general purine riboswitches) , which regulate protein synthesis. For example the ydhl mRNA which has a terminator stem at the end and blocks it from translation, but when the Adenine concentration increases in the cell, it binds to the mRNA and changes its conformation such that the terminator stem disappears.

3.2 microRNAs

There are other sorts of RNAs such as **microRNAs**, a more modern variant of RNA (relatively). Their discovery unveiled a novel non-protein layer of gene regulation (e.g. the EVF-2 and HOTAIR miRNAs). EVF-2 is interesting because its transcribed from an ultra conserved enhancer, and separates from the transcription string by forming a hairpin, and thereafter returns to the very same enhancer (along with a protein Dlx-2) and regulates its activity. HOTAIR RNA induces changes in chromatin state, and regulates the methylation of Histones, which in turn silences the HOX-D cluster.

3.3 Other types of RNA

We can also look at types of **noncoding RNAs**.

piRNAs are the largest class of small non-coding RNA molecules in animals. They are primarily involved in the silencing of transposons, but likely have a lot of functions. They are also involved in epigenetic modications, and post-transcriptional gene silencing.

lncRNAs are long transcripts produced that operate functionally as RNAs and are not translated into proteins. Many studies implicate lncRNAs in epigenetic modications, maybe acting as a targeting mechanism or as a molecular scaffold for Polycomb proteins. lncRNAs are likely to possess numerous functions, many are nuclear, many are cytoplasmic.

4 RNA Structure

We have learned about different functions of RNA, and it should be clear by now how fundamental the role of RNA in living systems is. Because it is impossible to understand how RNA actually does all these activities in the cell, without knowing what its structure is, in this part we will look into the structure of RNA.

RNA structure can be studied in three different levels 1:

1. *Primary* structure: the sequence in which the bases (U, A, C, G) are aligned.
2. *Seconary* structure: the 2-D analysis of the [hydrogen] bonds between different parts of RNA. In other words, where RNA becomes double-stranded, where RNA forms a hairpin or a loop or other similar forms.

3. *Tertiary* structure: the complete 3-D structure of RNA, i.e. how the string bends, where it twists and such.

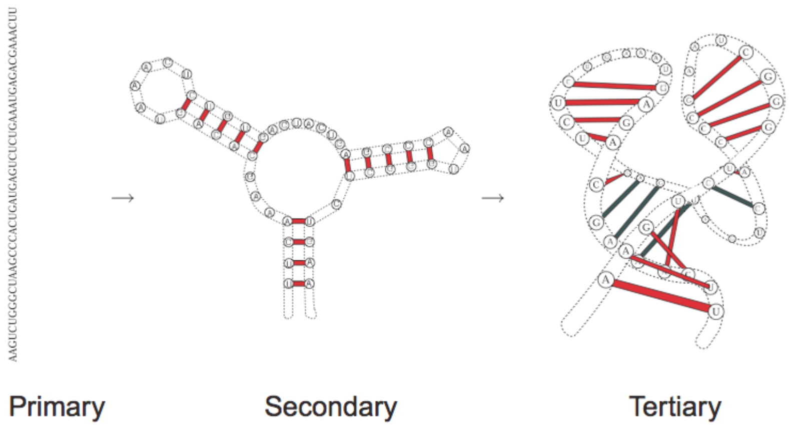


Figure 1: Graphical representation of the hierarchy of RNA strucure complexity

As mentioned before, the presence of ribose in RNA enables it to fold and create double-helixes with itself. The primary structure is fairly easy to obtain through sequencing the RNA. We are mainly interested in understanding the secondary structure for RNA: where the loops and hydrogen bonds form and create the functional attributes of RNA. Ideally, we would like to study the tertiary structure because this is the final state of the RNA, and what gives it its true functionality. However, the tertiary structure is very hard to compute and beyond the scope of this lecture.

Even though studying the secondary structure can be tricky, there are some simple ideas that work quite well in predicting it. Unlike proteins, in RNA, most of the stabilizing]free energy for the molecule comes from its secondary structure (rather than tertiary in case of proteins). RNAs initially fold into their secondary structure and then form their tertiary structure, and therefore there are very interesting facts that we can learn about a certain RNA molecule by just knowing its secondary structure.

Finally, another great property of the secondary structure is that it is usually well conserved in evolution, which helps us improve the secondary structure predictions and also to find ncRNA (non-coding RNA)s. There are widely used representations for the secondary structure of RNA:

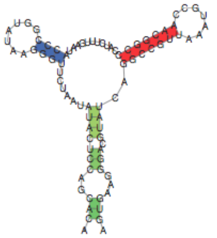


Figure 2: The typical representation of RNA secondary structure in textbooks. It clearly shows the secondary substructure in RNA.

Formally: A secondary structure is a vertex labeled graph on n vertices with an adjacency matrix $A = (a_{ij})$ fulfilling:

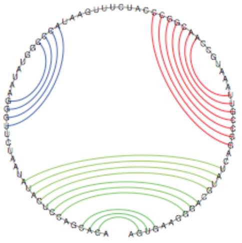
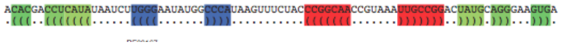


Figure 3: Graph drawing where the back-bone is a circle and the base pairings are the arcs within the circle. Note that the graph is outer-planar, meaning the arcs do not cross.

- $a_{i,i+1} = 1 for 1 \leq i \leq n1$ (continuous backbone)
- For each $i, 1 \leq i \leq N$ there is at most one $a_{ij} = 1$ where $j \gneqq i + / - 1$(a base only forms a pair with one other at the time)
- If $a_{ij} = a_{kl} = 1 and i < k < j then i < l < j$ (ignore pseudo knots)



```
ACACGACCUCAUAUAAUCUUGGGAAUAUGGCCCAUAAGUUUCUACCCGGCAACCGUAAAUUGCCGGACUAUGCAGGGAAGUGA
.(((..(((((((......((((.......))))...........(((((((.......)))))))..)))).)))...))).
```

Figure 4: A machine readable dot-bracket notation, in which for each paired nucleotide you open a bracket(and close it when you reach its match) and for each unpaired element you have a dot.

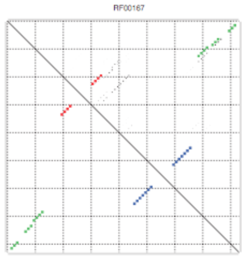


Figure 5: A matrix representation, in which you have a dot for each pair.

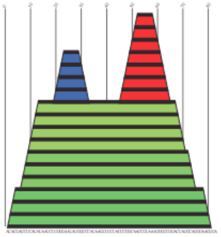
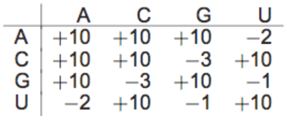


Figure 6: Mountain plot, in which for pairs you go one step up in the plot and if not you go one step to the right.

5 RNA Folding Problem and Approaches

Finally, we get to the point where we want to study the RNA structure. The goal here is to predict the secondary structure of the RNA, given its primary structure (or its sequence). The good news is we can find the optimal structure using dynamic programming. Now in order to set up our dynamic programming framework we would need a scoring scheme, which we would create using the contribution of each base pairing to the physical stability of the molecule. In other words, we want to create a structure with minimum free energy, in in our simple model we would assign each base pair an energy value. 7



	A	C	G	U
A	+10	+10	+10	−2
C	+10	+10	−3	+10
G	+10	−3	+10	−1
U	−2	+10	−1	+10

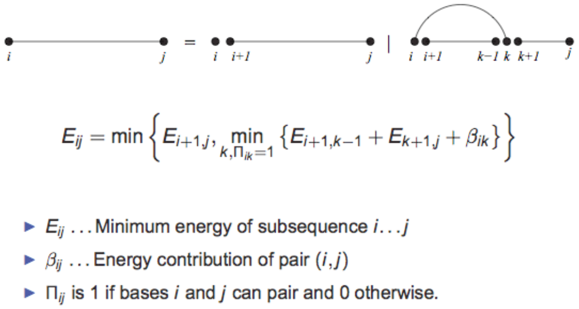
Figure 7: Example of a scoring scheme for base pair matches. Note that G-U can form a wobble pair in RNA.

The optimum structure is going to be the one with a minimum free energy and by convention negative energy is stabilizing, and positive energy is non-stabilizing. Using this framework, we can use dynamic programming (DP) to calculate the optimal structure because 1) this scoring scheme is additive 2) we disallowed pseudo knots, which means we can divide the RNA into two smaller ones which are independent, and solve the problem for these smaller RNAs.

We want to find a DP matrix E_{ij}, in which we calculate the minimum free energy for subsequence i to j. The first approach to this is Nussinov's algorithm.

5.1 Nussinov's algorithm

The recursion formula for this problem was first described by Nussinov in 1978. 8



$$E_{ij} = \min\left\{E_{i+1,j}, \min_{k,\Pi_{ik}=1}\{E_{i+1,k-1} + E_{k+1,j} + \beta_{ik}\}\right\}$$

- E_{ij} . . . Minimum energy of subsequence $i \ldots j$
- β_{ij} . . . Energy contribution of pair (i, j)
- Π_{ij} is 1 if bases i and j can pair and 0 otherwise.

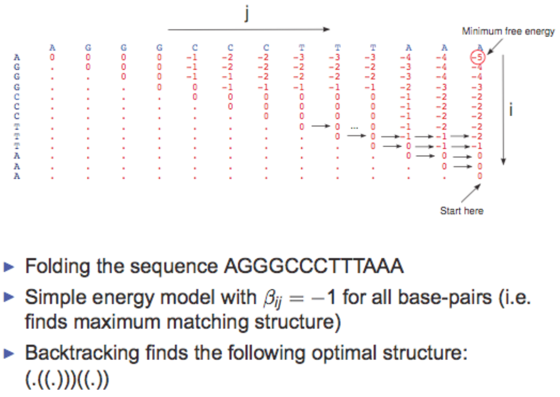
Figure 8: The recursion formula for Nussinov algorithm, along with a graphical depiction of how it works.

It calculates the best substructure for the subsequences and then builds them up to larger sequences till it finds the structure of the whole sequence. There are basically only two cases to get from one step to the next in the recursion: 1) The newly added base is unpaired 2) It is paired with some base k. For the latter case the base pair (i, k) divides the problem into two subproblems which can be then recursively solved the same way. Below 9 is example for the DP Matrix after running this algorithm:

When you calculate the minimum free energy, you are interested in the sequence which corresponds to this particular energy, a helper matrix is filled to backtracking over the sequence,

Here is the code for the backtracking:

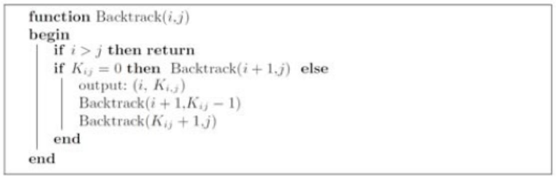
This model is very simplistic and there are some limitations to it. Most importantly, stacking interaction



	A	G	G	G	C	C	C	T	T	T	A	A	A
A	0	0	0	0	−1	−2	−2	−3	−3	−3	−4	−4	−5
G	.	0	0	0	−1	−2	−2	−2	−2	−2	−3	−4	−4
G	.	.	0	0	−1	−1	−2	−2	−2	−2	−3	−4	−4
G	.	.	.	0	0	−1	−1	−1	−1	−1	−2	−3	−3
C	.	.	.	.	0	0	0	0	0	0	−1	−2	−2
C	.	.	.	.	.	0	0	0	0	0	−1	−2	−2
C	.	.	.	.	.	.	0	0	0	0	−1	−2	−2
T	.	.	.	.	.	.	.	0	0	0	−1	−2	−2
T	.	.	.	.	.	.	.	.	0	0	−1	−1	−2
T	.	.	.	.	.	.	.	.	.	0	0	−1	−1
A	.	.	.	.	.	.	.	.	.	.	0	0	0
A	.	.	.	.	.	.	.	.	.	.	.	0	0
A	.	.	.	.	.	.	.	.	.	.	.	.	0

- Folding the sequence AGGGCCCTTTAAA
- Simple energy model with $\beta_{ij} = -1$ for all base-pairs (i.e. finds maximum matching structure)
- Backtracking finds the following optimal structure: (.((.)))((.))

Figure 9: The algorithm starts by assigning the diagonal for the matrix as 0 (since you cannot pair with yourself) and then works through the recursion up and right, the minimum free energy is the top-rightmost element in the matrix. The minimum length for a loop is 1 here (usually 3). As i counts backward form n to 1, j counts from 1 to n. In this example we simply assign -1 for a pair, and 0 for a non-pair.



```
function Backtrack($i$,$j$)
begin
    if $i > j$ then return
    if $K_{ij} = 0$ then Backtrack($i+1$,$j$) else
        output: ($i$, $K_{i,j}$)
        Backtrack($i+1$,$K_{ij}-1$)
        Backtrack($K_{ij}+1$,$j$)
    end
end
```

Figure 10: The helper array K_{ij} is filled during the recursion that holds the optimal secondary structure when k is paired with i for a sub-sequence $i..j$. If i is unpaired in the optimal structure, K_{ij} is filled during the recursion that holds the optimal secondary structure when k is paired with i for a sub-sequence $i...j$. If i is unpaired in the optimal structure, $K_{ij} = 0$.

between neighboring pairs is a very important factor (even more important than the hydrogen bonds)in RNA folding which is not considered by the Nussinovs model. 11

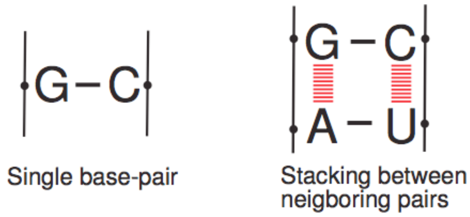


Figure 11: Stacking between neighboring base pairs is RNA. The flat aromatic structure of the base causes quantum interactions between stacked bases and changes its physical stability.

Therefore people have thought of methods to integrate such biophysical factors into our prediction. One improvement for instance is that instead of assigning energies to single base pairs, we assign them to faces of the graph (structural elements in 12). In order to find out the total energy of the structure, we have to find the free energy of each substructure, and simply add them up. The stacking energies can be calculated by melting oligonucleotides experimentally.

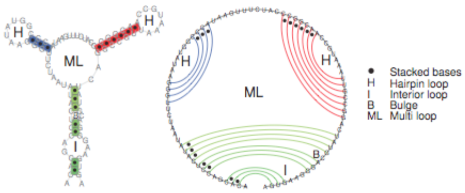


Figure 12: Various internal substructures in a folded RNA. A hairpin is consisted of a terminal loop connected to a paired region, an internal loop is an unpaired region within the paired region. A Bulge is a special case of an interior loop with a single mis-pair. a Multi loop is a loop which consists of multiple of these components (in this example two hairpins and a paired region, all connected to a loop).

5.2 Zuker Algorithm

Theefore, we use a variant which includes stacking energies to calculate the RNA structure. This is called the Zuker algorithm. Like Nussinovs, it assumes that the optimal structure is the one with the lowest equilibrium free energy. Nevertheless, it includes the total energy contributions from the various substructures which is partially determined by the stacking energy. Some modern RNA folding algorithms use this algorithm for RNA structure predictions.

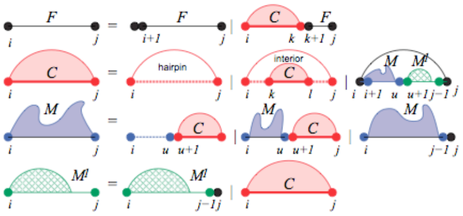
In the Zuker algorithm, we have four different cases to deal with. Figure 13 shows a graphical outline of the decomposition steps. The procedure requires four matrices. F_{ij} contains the free energy of the overall optimal structure of the subsequence x_{ij}. The newly added base can be unpaired or it can form a pair. For the latter case, we introduce the helper matrix C_{ij}, that contains the free energy of the optimal substructure of x_{ij} under the constraint that i and j are paired. This structure closed by a base-pair can either be a hairpin, an interior loop or a multi-loop.

The hairpin case is trivial because no further decomposition is necessary. The interior loop case is also simple because it reduces again to the same decomposition step. The multi-loop step is more complicated. The energy of a multi loop depends on the number of components, i.e. substructures that emanate from the loop. To implicitly keep track of this number, there is a need for two additional helper matrices. M_{ij} holds the free energy of the optimal structure of x_{ij} under the constraint that x_{ij} is part of a multi loop with at least one component. M^1_{ij} holds the free energy of the optimal structure of x_{ij} under the constraint that x_{ij} is part of a multi-loop and has exactly one component closed by pair (i, k) with $i < k < j$. The idea is to decompose a multi loop in two arbitrary parts of which the first is a multi-loop with at least one component and the second a multi-loop with exactly one component and starting with a base-pair.

These two parts corresponding to M and M^1 can further be decomposed into substructures that we already know, i.e. unpaired intervals, substructures closed by a base-pair,or (shorter) multi-loops. (The recursions are also summarized in 13.

In reality, however, at room temperature (or cell temperature), RNA is not actually in one single state, but rather varies in a Thermodynamic ensemble of structure. Base pairs can break their bonds quite easily, and although we might find an absolute optimum in terms of free energy, it might be the case that there is another sub-optimal structure which is very different from what e predicted and has an important role in the cell. To fix the problem we can calculate the base pair probabilities to get the ensemble of structures, and then we can have a much better idea of what the RNA structure probably looks like. In order to do this, we utilize the Boltzman factor:

This gives us the probability of a given structure, in a thermodynamic system. We need to normalize the temperature using the partition function Z,which is the weighted sum of all structures, based on their



$$F_{ij} = \min\left\{F_{i+1,j},\ \min_{i<k\leq j} C_{ik} + F_{k+1,j}\right\}$$

$$C_{ij} = \min\left\{\mathcal{H}(i,j),\ \min_{i<k<l<j} C_{kl} + \mathcal{I}(i,j;k,l),\ \min_{i<u<j} M_{i+1,u} + M^1_{u+1,j-1} + a\right\}$$

$$M_{ij} = \min\left\{\min_{i<u<j} (u-i+1)c + C_{u+1,j} + b,\ \min_{i<u<j} M_{i,u} + C_{u+1,j} + b,\ M_{i,j-1} + c\right\}$$

$$M^1_{ij} = \min\left\{M^1_{i,j-1} + c,\ C_{ij} + b\right\},$$

Figure 13: F describes the unpaired case, C is described by one of the three conditions : hairpin,interior loop, or a composition of structures i.e. a multi loop. M^1 is a multi loop with only one component, where are M might have multiple of them. The | icon is notation for *"or"*.

$$\text{Prob}(\mathcal{S}) = \frac{\exp(-\Delta G(\mathcal{S})/RT)}{Z}$$

Boltzman factor:

$$Z = \sum_{\mathcal{S}} \exp(-\Delta G(\mathcal{S})/RT)$$

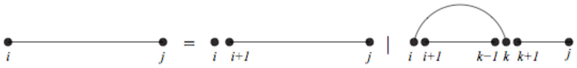
We can also represent this ensemble graphically, using a dot plot to visualize the base pair probabilities. To calculate the specific probability for a base pair (i, j) , we need to calculate the partition function, which is given by the following formula :

$$p_{ij} = \frac{\hat{Z}_{ij} Z_{i+1,j-1} \exp(-\beta_{ij}/RT)}{Z}$$

To calculate Z (the partition function over the whole structure), we use the recursion similar to the Nussinovs Algorithm (known as McCaskill Algorithm).The inner partition function is calculated using the formula:

$$Z_{ij} = Z_{i+1,j} + \sum_{\substack{i+1\leq k\leq j \\ \Pi_{ik}=1}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\beta_{ik}/RT)$$

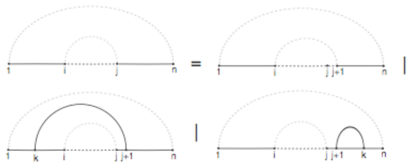
With each of the additions corresponding to a different split in our sequence as the next figure illustrates. Note that the addition are multiplied to the energy functions since it is expressed as a exponential.



Similarly the outer partition function is calculated with a the same idea using the formula:

corresponding to different splits in the area outside the base pairs (i, j).

$$\begin{aligned}\widehat{Z}_{ij} = \widehat{Z}_{i,j+1} \quad &+ \sum_{\substack{1\leq k<i \\ \Pi_{k,j+1}=1}} \widehat{Z}_{k,j+1} \exp(-\beta_{k,j+1}/RT) Z_{k+1,i-1} \\ &+ \sum_{\substack{j+2\leq k\leq n \\ \Pi_{k,j+1}=1}} \widehat{Z}_{i,k} \exp(-\beta_{k,j+1}/RT) Z_{j+2,k-1}\end{aligned}$$



6 Evolution of RNA

It is useful to understand the evolution of RNA structure, because it unveils valuable data, and can also give us hints to refine our structure predictions. When we look into functionally important RNAs over time, we realize their nucleotides have changed at some parts, but their structure is well-conserved.

In RNA there are a lot of **compensatory mutations and consistent mutations**. In a consistent mutation, the structure doesnt change e.g. an AU pair mutates to form a G pair. In a compensatory mutation there are actually two mutations, one disrupts the structure, but the second mutation restores it, for example an AU pair changes to a CU which does not pair well, but in turn the U mutates to a G to restore a CG pair. In an ideal world, if we have this knowledge, this is the be the key to predict the RNA structure, because evolution never lies. We can calculate the **mutual information content** for two different RNAs and compare it. In other words, you compare the probabilities of two base pair structures agreeing randomly vs. if they have evolved to be conserve the structure.

The mutual information content is calculated via this formula:

$$MI_{ij} = \sum_{X,Y} f_{ij}(XY) \log \frac{f_{ij}(XY)}{f_i(X) f_j(Y)}$$

If we normalize these probabilities, and store the MI in bits, we can plot it in a 3D model and track the evolutionary signatures. In fact, this was the method for determining the structure of ribosomal RNAs long before they were found by crystallography.

The real problem is that we dont have so much information, so what we usually do is combine the folding prediction methods with phylogenetic information in order to get a reliable prediction. The most common way to do this is to combine to Zuker algorithm with some covariance scores. For example, we add stabilizing energy if we have a compensatory mutation, and destabilizing energy if we have a single nucleotide mutation.

7 Probabilistic Approach to the RNA Folding Problem

RNA-coding sequence inside the genome Finding RNA-coding sequences inside the genome is a very hard problem. However there are ways to do it. One way is to combine the thermodynamic stability information, with a normalized RNAfold score and then we can do a Support Vector Machine (SVM) classification, and compare the thermodynamic stability of the sequence to some random sequences of the same GC content and the same length and see how many standard deviations is the given structure more stable that the expected value.
We can combine it with the evolutionary measure and see if the RNA is more conserved or not. This gives us (with relative accuracy) an idea if the genomic sequence is actually coding an RNA.

We have studied only half of the story. Although the thermodynamic approach is a good way (and the classic way) of folding the RNAs, some part of the community like to study it from a different aspect.

Lets assume for now that we dont know anything about the physics of RNA or the Boltzman factor. Instead, we look into the RNA as a string of letters for which we want to find the most probable structure. We have already learned about the Hidden Markov Models in the previous lectures. They are a nice way to make predictions about the hidden states of a probabilistic system. The question is can we use Hidden Markov models for the RNA folding problem? The answer is yes.

We can represent RNA structure as a set of hidden states of dots and brackets (recall the dot-bracket representation of RNA in part 3). There is an important observation to make here: the positions and the pairings inside the RNA are not independent, so we cannot simply have a state of an opening bracket without any considerations of the events that are happening downstream.

Therefore we need to extend the HMM framework to allow for nested correlations. Fortunately, the probabilistic framework to deal with such a problem already exists. It is known as stochastic context-free grammar (SCFG).

Context Free Grammar in a nutshell
You have:

- Finite set of non-terminal symbols (states) e.g. $\{A, B, C\}$ and terminal symbols e.g. $\{a, b, c\}$
- Finite set of Production rules. e.g. $\{A \to aB, B \to AC, B \to aa, \to ab\}$
- An initial (start) nonterminal

You want to find a way to get from one state to another (or to a terminal). $A \to aB \to aAC \to aaaC \to aaaab$
In a stochastic CFG, the only difference is that each relation has a certain probability.e.g.$P(B \to AC) = 0.25$ $P(B \to aa) = 0.75$

Phylogenetic evaluation is easily combined with SCFGs, since there are many probabilistic models for phylogenetic data. The Probabilistic models are not discussed in detail in this lecture but the following picture basically gives an analogy between the Stochastic models and the methods that we have see so far in the class.

- Analogies to thermodynamic folding:
  - CYK ↔ Minimum Free energy (Nussinov/Zuker)
  - Inside/outside algorithm ↔ Partition functions (McCaskill)

- Analogies to Hidden Markov models:
  - CYK Minimum ↔ Viterbi's algorithm
  - Inside/outside algorithm ↔ Forward/backwards algorithm
- Given a parameterized SCFG (Θ, Ω) and a sequence x, the Cocke-Younger-Kasami (CYK) dynamic programming algorithm finds an optimal (maximum probability) parse tree $\hat{\pi}$:

  $\hat{\pi} = argmaxProb(\pi, x|\Theta, \Omega)$
- The *Inside algorithm*, is used to obtain the total probability of the sequence given the model summed ovver all parse trees,

  $Prob(x|\Theta, \Omega) = \Sigma Prob(x, \pi|\Theta, \Omega)$

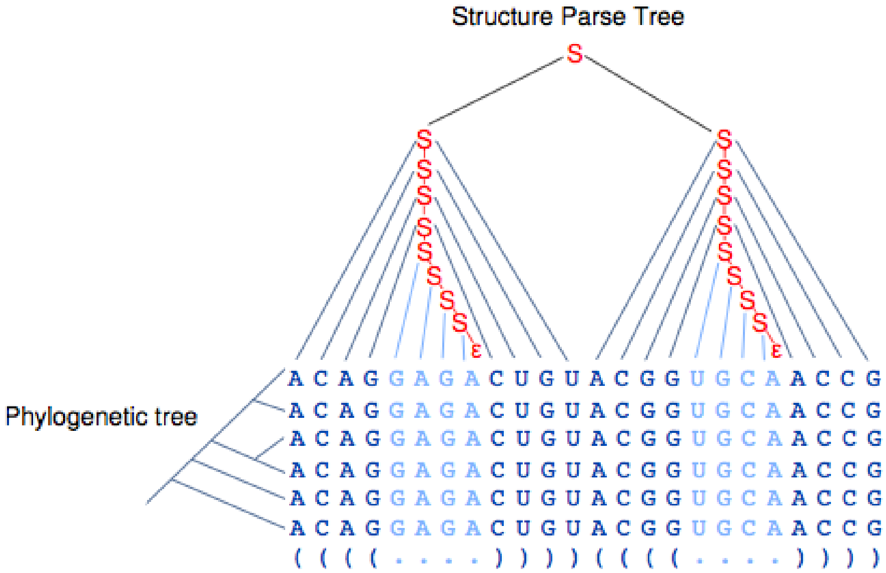


Figure 14: A) Single sequence: Terminal symbols are bases or base-pairs, Emission probabilities are base frequencies in loops and paired regions B) Phylo-SCFG: Terminal symbols are single or paired alignment columns, Emission probabilities calculated from phylogenetic model and tree using Felsenstein's algorithmWe to try to better understand RNA-RNA interactions.

7.1 Application of SCFGs

- Consensus secondary structure prediction: Pfold
  - First Phylo-SCFG
- Structural RNA gene finding: EvoFold
  - Uses Pfold grammar
  - Two competing models:
    - Non-structural model with all columns treated as evolving independently
    - Structural model with dependent and independent columns
  - Sophisticated parametrization

8 Advanced topics

There still remain a host of other problems that need to be solved by studying RNA structure. This section will profile some of them.

8.1 Other problems

Observe some of the problems depicted graphically below:

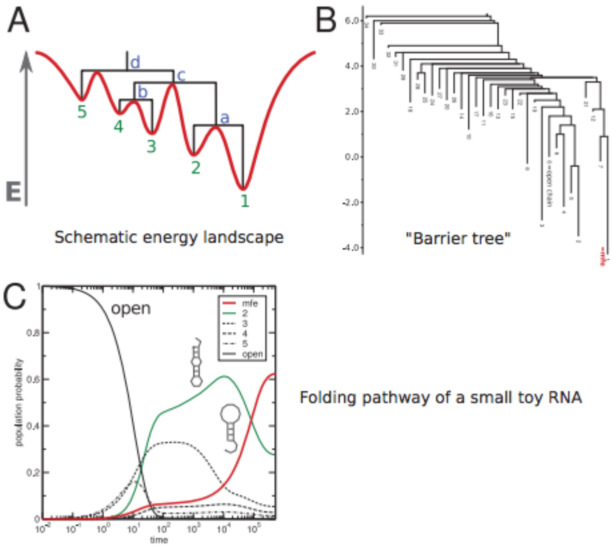


Figure 15: We can study kinetics and folding pathways in further depth.

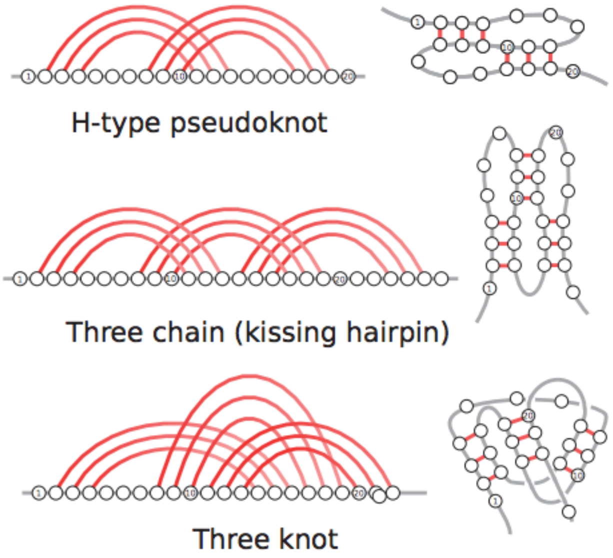


Figure 16: We can investigate pseudoknots.

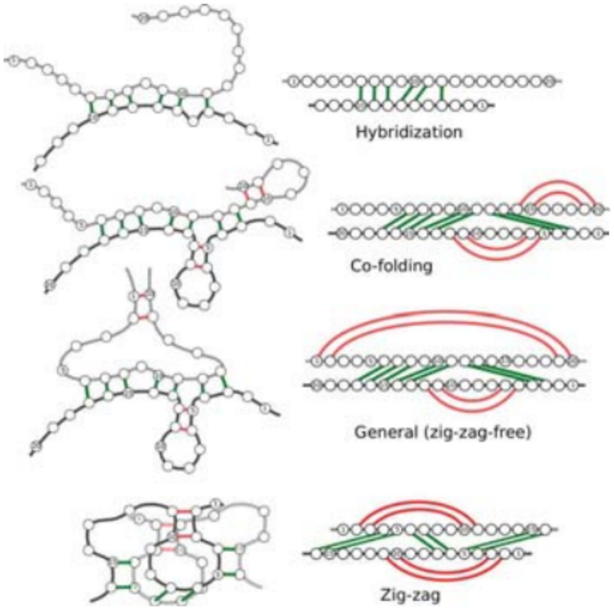


Figure 17: We can try to better understand RNA-RNA interactions.

8.2 Relevance

There are plenty of RNAs inside the cell aside from mRNAs, tRNAs and rRNAs. The question is what is the relevance of all this non-coding RNA? Some believe it is noise resulted through experiment, some think its just biological noise that doesnt have a meaning in the living organism. On the other hand some believe junk RNA might actually have an important role as signals inside the cell and all of it is actually functional, the truth probably lies somewhere in between.

8.3 Current research

There are conserved regions in the genome that do not code any proteins, and now Stefans et al. are looking into them to see if they have structures that are stable enough to form functional RNAs. It turns out that around 6% of these regions have hallmarks of good RNA structure, which is still 30000 structural elements. The group has annotated some of these elements, but there is still a long way to go. a lot of miRNA, snowRNAs have been found and of course lots of false positives. But there exciting results coming up in this topic! so the final note is, it's a very good area to work in!

9 Summary and key points

1. The functional spectrum of RNAs is practically unlimited
   (a) RNAs similar to contemporary Ribozymes and Riboswitches might have existed in an RNA world. Some of them still exist as living fossils in current cells.
   (b) Evolutionarily younger RNAs including miRNAs and many long ncRNAs form a non-protein based regulatory layer.
2. RNA structure is critical for their function and can be predicted computationally
   (a) Nussinov/Zuker: Minimum Free Energy structure
   (b) McCaskill: Partition function and pair probabilities
   (c) CYK/Inside-Outside: probabilistic solution to the problem using SCFGs
3. Phylogenetic information can improve structure prediction
4. Computational biology of RNAs is an active eld of research with many hard algorithmic problems still open

10 Further reading

- Overview
  - Washietl S, Will S. et al. Computational analysis of noncoding RNAs. Wiley Interdiscip Rev RNA. 2012, 10.1002/wrna.1134
- RNA function: review papers by John Mattick
- Single sequence RNA folding
  - 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

- Consensus RNA folding
  - Hofacker IL, Fekete M, Stadler PF, Secondary structure prediction for aligned RNA sequences. J Mol Biol. 2002 Jun; 319:(5)1059-66
  - Knudsen B, Hein J, RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics. 1999 Jun; 15:(6)446-54

- RNA gene finding
  - Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol. 2006 Apr; 2:(4)e33
  - Washietl S, Hofacker IL, Stadler PF, Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A. 2005 Feb; 102:(7)2454-9

References

[1] R Durbin. *Biological Sequence Analysis.*

[2] W. Gilbert. ”origin of life: The rna world”. *Nature.*, 319(6055):618, 1986.

[3] Rachel Sealfon, 2012. Extra information taken from Recitation 5 slides.

[4] Z. Wang, M. Gestein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.*, 10(1):57–63, 2009.

[5] Stefan Washietl, 2012. All pictures/formulas courtesy of Stefan’s slides.

[6] R. Weaver. *Molecular Biology.* 3rd edition.