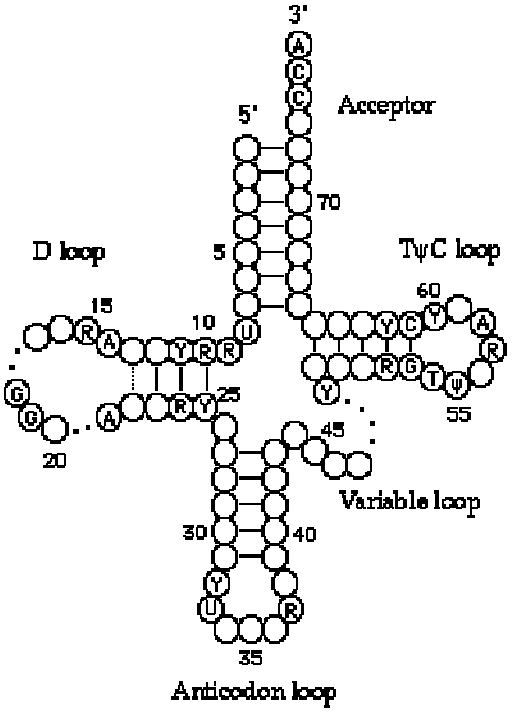
Hermann V. Klann

Bioinformatic Tutorial:

RNA STRUCTURE PREDICTION

Completion to the lecture of Oktober 20th, 2003



TU/LMU München

0. Arrangement

RNA Structure Prediction

1. Introduction

In this paper you will get an Overview on RNA structure and its representation, typical structure elements, free energy calculation, structure prediction methods and at the end You will see a practical excursion how to predict RNA structure with an online tool generated by Mr Zuker called “mfold”.

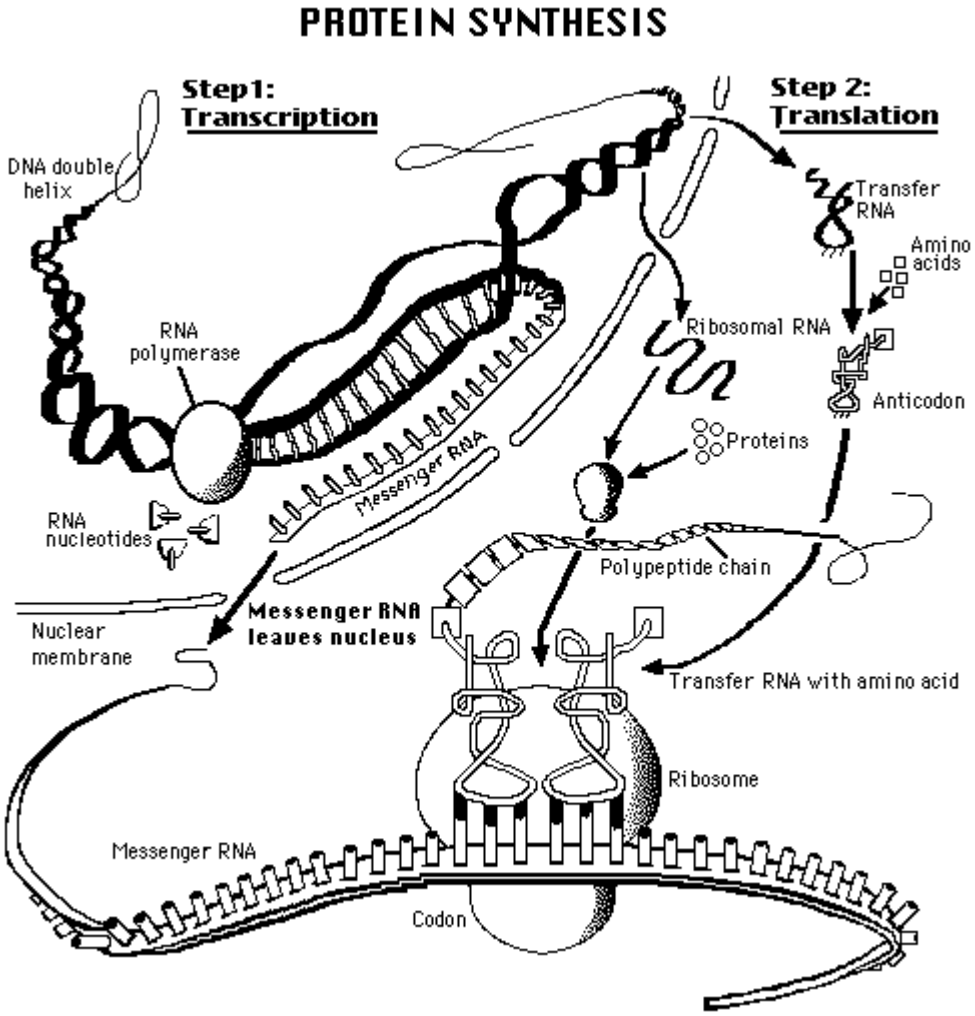


Figure 1: RNA´s and the protein synthesis

What is RNA???

The biological polymers RNA (ribonucleic acid) is a long string of monomer units known as nucleotides. Each nucleotide is composed of a base, a sugar and, a phosphate group. The sugar and phosphate groups form a contiguous unit known as the sugar-phosphate backbone, they perform a structural role.
In contrast, the order of the different bases protruding from the sugar-phosphate backbone codes for genetic and sometimes structural information of the molecule.

The process of transforming strings from a DNA alphabet to strings from a twenty-letter alphabet is known as translation. Translation of DNA into protein is heavily influenced by

RNA. Usually a DNA template is first transcribed into messenger RNA (mRNA) then the mRNA is translated into protein, much of this step is mediated by several stable species of functional RNA´s or fnRNA´s (these are also known as non-coding RNA).

Examples of these fnRNA´s are the ribosomal RNA´s (rRNA´s) that play a crucial role in ribosome structure and architecture and the transfer RNA´s (tRNA´s) that is the adapter molecule recognized by the enzyme responsible for synthesizing protein and the anti-codon on mRNA. Other fnRNA´s that are not directly involved in protein synthesis are the small nuclear RNA´s (snRNA´s), which are components of the splice some. Some exciting, newly discovered fnRNA´s are RNAse P RNA, telomerase RNA, meiRNA, and mammalian Xist RNA which is heavily involved in X dosage compensation.

Why we do RNA structure prediction:

Now, that we can find the minimum free energy structure of a sequence in computationally tractable time, we should ask ”What does the optimum tell us”?
That is, there may be more than one structure with the optimum free energy, or there may be many structures within 5% to 10% of the minimum free energy, and these may be topologically very different. A minimum energy folding algorithm will return only one secondary structure, though there are many candidates for the natural structure. To address this, some software packages (such as Zuker's mfold) will display a number of suboptimal folds.
Inferring what structure is truly representative of the natural structure requires additional information. Phylogenetic information is often used to constrain the search by identifying highly conserved motifs. Some programs allow the user to specify constraints on the secondary structure, by specifying paired, single-stranded, or non-pair able regions, or by actively participating in the folding process.

2. Overview – RNA

2.1 Structure and Function of RNA

What is RNA structure:

RNA is often single stranded and forms structures driven by hydrogen bonding and stacking of the bases. Base-pairs within RNA shapes are usually of the canonical Watson-Crick type, that are formed by three hydrogen bonds between C and G and two hydrogen bonds between A and U. Wobble pairing can also occur, this is a non-canonical pairing between G and U that is often found in RNA secondary structure. Probes of RNA structure using X-ray diffraction, NMR and thermodynamic studies have revealed other non-canonical pairings such as G with A, and U with C pairs.
RNA structure is not limited to primary and secondary forms, tertiary structure is also present. Tertiary interactions are usually defined as non-nested base interactions such as pseudoknots and base triples.

RNA has the same primary structure as DNA . It consists of a sugar-phosphate backbone, with nucleotides attaches to the 1' carbon of the sugar. The differences between DNA and RNA are that:

RNA has a hydroxyl group on the 2' carbon of the sugar (thus, the difference between deoxyribonucleic acid and ribonucleic acid.
Instead of using the nucleotide thymine, RNA uses another nucleotide called uracile:

Because of the extra hydroxyl group on the sugar, RNA is too bulky to form a a stable double helix. RNA exists as a single-stranded molecule. However, regions of double helix can form where there is some base pair complementation (U and A , G and C), resulting in hairpin loops. The RNA molecule with its hairpin loops is said to have a secondary structure.

In addition, because the RNA molecule is not restricted to a rigid double helix, it can form many different tertiary structures. Each RNA molecule, depending on the sequence of its bases, can fold into a stable three-dimensional structure.

2.2 dsRNA & ssRNA
(double & single stranded RNA)

RNA has more biological functions than DNA. Like DNA, which is the genetic material in cells, double or single stranded RNA is the genetic material of RNA-Virus. In a Retrovirus, RNA is the matrix for DNA synthesis. (Steger G.)
Normally RNA is single stranded. (**Figure 2**)

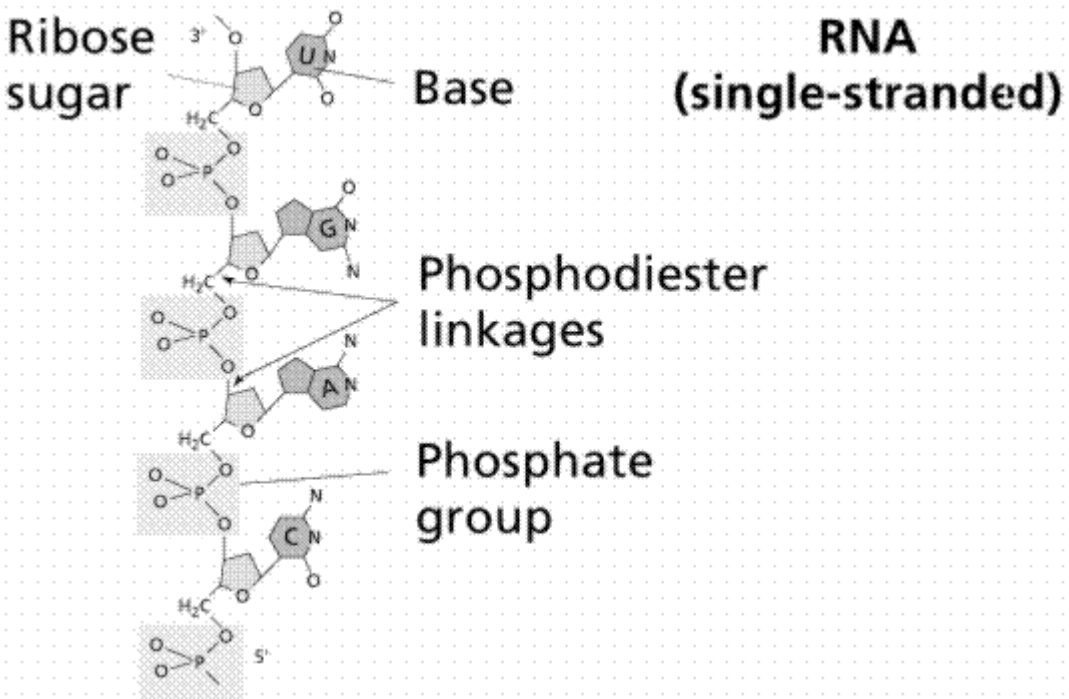


Figure 2: Single stranded RNA

2.3 mRNA:
(messenger RNA)

Messenger RNA is the type of RNA familiar to most people.
The information stored in mRNA is used to make proteins. (**Figure 3**)
When mRNA is first created in eucaryotes it is called precursor mRNA because it needs to be modified before it can pass on the information it has for the formation of protein. The first two modifications are capping and the addition of a poly A tail. These are discussed in Transcription. The third type of modification involves the removal of introns and the splicing together of exons.

Segments of DNA that contain information for the formation of proteins are called exons. Exons typically have other segments DNA separating them from each other. These segments are called introns. The precursor mRNA contains both the exons and introns. The introns need to be cut out and the exons need to be connected back together. Some human genes for proteins are split up into as many as 79 different exons.
A spliceosome is a complex of proteins and small RNA molecules, and is where the removal of introns and the splicing together of exons takes place.

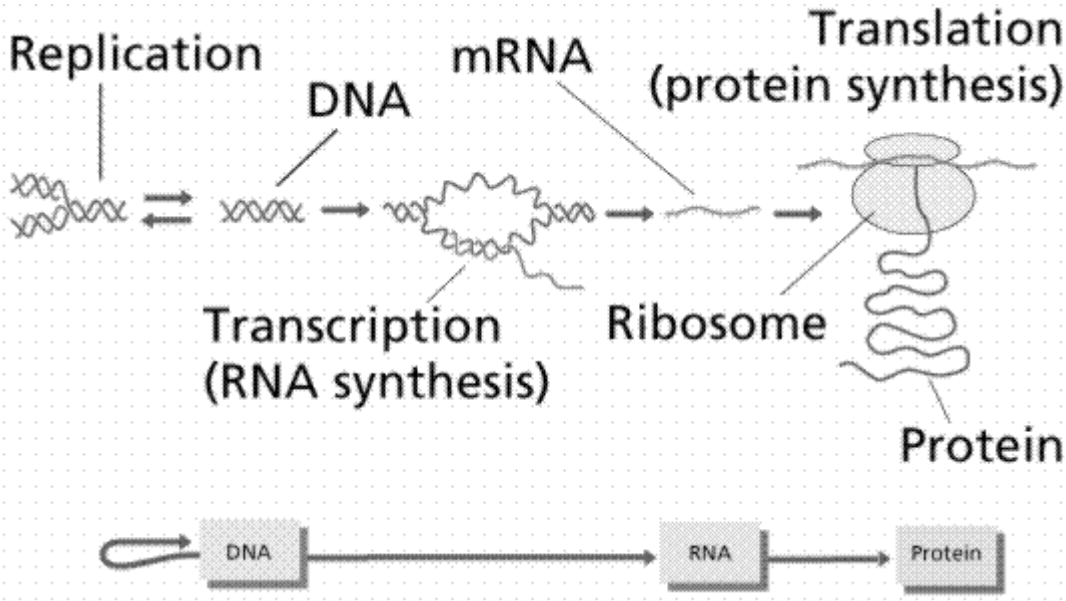


Figure 3: mRNA is used to make proteins

Messenger RNA makes up only about 5% of all RNA in a typical cell and is made up of small amounts of thousands of different mRNA molecules.
In bacteria mRNA is modified very little if at all. Since bacteria do not have a nucleus, translation starts before transcription even ends so there is no time for RNA splicing, or a need as procaryotic genes are not split into separate exons.

2.4 tRNA:
(transfer RNA)

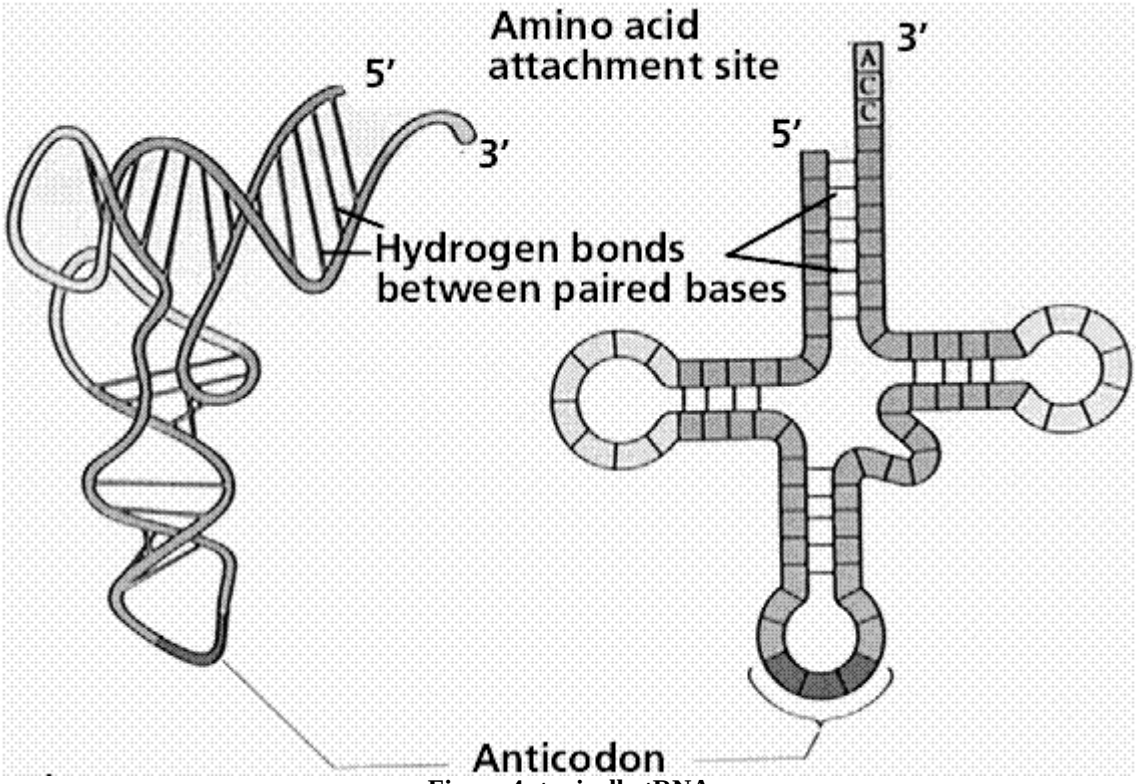


Figure 4: typically tRNA

There are at least 32 different kinds of tRNA in an eucaryotic cell. They are relatively small molecules, each one is made up of only 73-93 ribonucleotides. Although tRNA is a single strand of RNA, it bends around in certain places resulting in some ribonucleotides pairing up with others in the same chain, forming three loops. Each tRNA molecule has one amino acid attached to its 3' end. (**Figure 4**) Since there are only 20 amino acids and around 32 different kinds of tRNAs, some amino acids are carried by more than one type of tRNA. On one of the three loops is what is called an anticodon. Anticodons are made up of three bases and are involved in translation. The particular amino acid attached to a tRNA molecule is determined by its anticodon sequence.

2.5 rRNA:
(ribosomal RNA)

Ribosomes are made of protein and ribosomal RNA (rRNA) and are where translation of RNA to protein takes place.
In E. coli ribosomes contain three kinds of rRNA - 23S, 16S and 5S.
In eucaryotes, there are four kinds of rRNA - 18S, 28S, 5.8S, and 5S. One 18S molecule is used to make the small subunit of the ribosome, with the help of several proteins. The 28S, 5.8S, and 5S rRNA molecules are involved with the construction of the large subunit of the ribosome. The 28S, 18S, and 5.8S molecules are made from the processing of a single precursor RNA.

3. RNA structures

3.1 Primary Structure

The primary structure of RNA is nearly identical to the primary structure of DNA. The basics are the base, the ribose and the phosphate. The main difference is the ribose instead of the desoxyribose, that is the basic to the different conformations of the helices of RNA and DNA and the chemical instability of RNA in comparison with DNA.

The four different bases are two purines, called adenine and guanine and the two pyrimidines cytosine and uracile. Instead of thymine in DNA is uracile in RNA
The difference is only the 5-methylegroup which is mainly the reason, why the thermodynamic stability of double stranded RNA is higher than those of DNA.
Bases, nucleosides (base & ribose) and the nucleotide (nucleoside & phosphate) are mostly recommend noticed as AGCU.

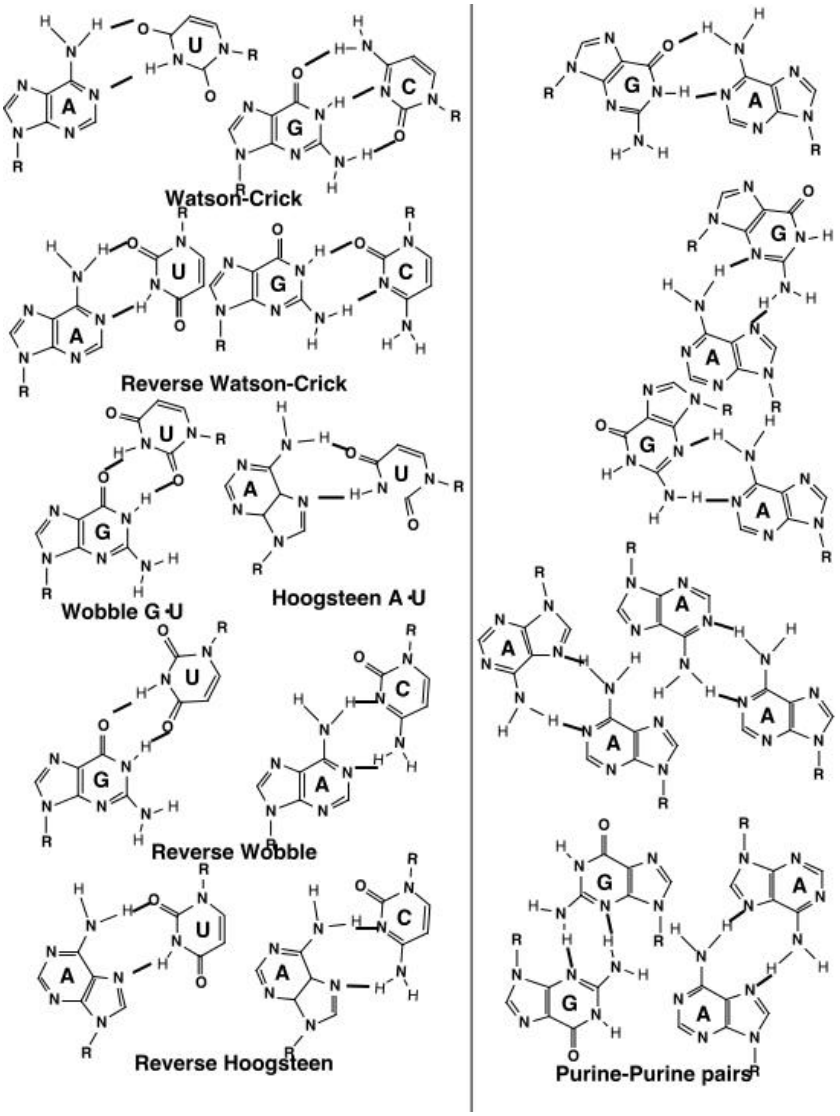


Figure 5: RNA – Base Pairs

3.2 Secondary Structure

The secondary structure of RNA can be described as a set of base pairings in the three dimensional structure of the molecule. RNA can fold like DNA in complex structure. The bases to all these structure of higher order is the skill of the bases to build hydrogen bonds and stackings. Hydrogen bonds only connect non-neighbored bases.

The energy earning because of hydrogen bonds-pairings is relatively minimal as there are similar good hydrogen bonds with the around water possible. Stackings connect neighbored or non-neighbored bases or base-pairs. The interaction is a dipole which is energetic good.

Double stranded RNA is build from two anti parallel single stranded which is most recommend for double stranded RNA-virus. Mostly RNA is single-stranded, so the simplest secondary structure is a hairpin, which is a helix and a hairpin-loop. RNA is only in rare cases complete self compliment. Non-complement parts a called loops. In comparison to helical parts the building of loops is energetic bad. All loops destabilize RNA structure.

Different visualization variants are used:

- Point and mark notation
- Squiggles plot
- Mountain plot

3.3 Tertiary structure

On the basis of RNA secondary structure a tertiary RNA structure can build, which is stabilized by hydrogen bonds and stackings similar to the secondary structure. Tertiary structure is very important to the biological activity of many RNAs. Tertiary structure elements often are interaction between two loops e.g. 2 x hairpin loop in t-RNA, hairpin loop with an internal loop or hairpin loop and free end (pseudo knots).
If a base in a hydrogen bond interaction is involved with an other base we call it a triplet strand.

3.4 Additional

RNA secondary structure include much more and better interaction as you need to build tertiary structure. The building of a secondary structure is much more faster than the building of a tertiary structure.

This arguments let you think that in the structure-building-process first secondary-structure is created and then tertiary-structure is build without influencing the basic secondary-structure.
So, you have to detect the secondary-structure first with experimental and theoretic methods before modeling a tertiary structure with a much higher complexity.
Further more the tertiary-structure can be build on a energetic sub optimal secondary-structure.
The sub optimal structure may be less energetic build then the optimal secondary structure.

4. Structure Elements

4.1 Base pairs or Stacks

Principally everyone of the four bases can build up to three hydrogen bonds with an other base.

The standard base pairs are canonical or Watson Creek , in short WC-Base pairs. Like adenine and uracile (adenine and thymine in DNA) and guanine and cystidine. Normally base pairs are ordered in sequence-structure or stems in secondary structure. WC pairs are isoster, that means that you can build regular helices of different length. Watson–Crick pairs constitute the most basic unit encountered in RNA structures. As such, it is of paramount importance to gain a precise view of their hydration. Besides, data gathered for Watson–Crick
pairs are the basis for a sound comparison in the evaluation of modifications of the hydration shell associated with insertions of non-Watson–Crick pairs and modified nucleotides into RNA structures.

In RNA an other base pair is recommend. The so called Wobble Base Pair guanine and cystidine.

Wobble pairs are isoster, that means that you can build regular helices of different length. Other base pairs are so called non canonical base pairs which are not isoster and not self isoster.
To add a base pair to a helix is the kinetically most fast process for RNA structure building. The non-Watson–Crick pairs play major structural and functional roles in RNA architectures and are associated with specific hydration patterns.

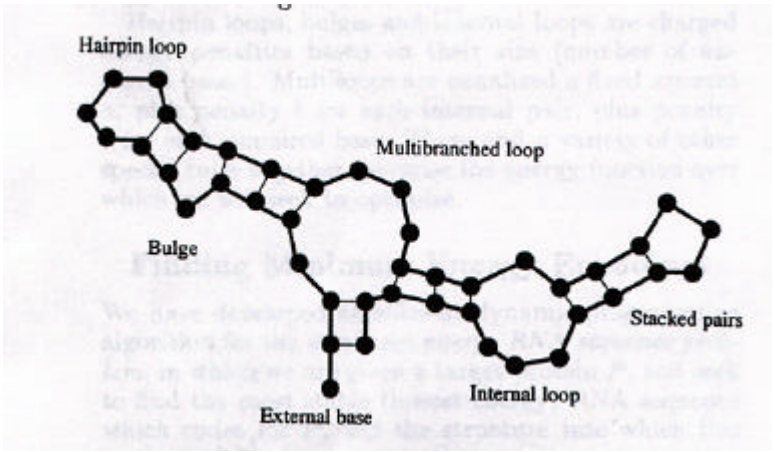


Figure 6: RNA - Structures

4.2 Hairpins

A hairpin is a helix with a loop, which closes the helix. Hairpin loops can be build fast. The thermodynamic stability of hairpin loops belongs to the length of the loop, to the loop sequence and the type of the loop building sequence. Generally, loops with five unpaired bases are less destabilizing. Loops with three base pairs or less are to small to close a helix.

A special case are the so called tetra loop-hairpins. They are only 4 base pairs long, but the first and the last base of the loop create a base pair that stacks at the last regular base pair of the helix. Tetra loops are the most less destabilizing loops as possible. In rRNA, extra stabile tetra loop hairpins are most recommended.

Grater hairpins can have a special structure, too. For example in tRNA the anticodon loop has seven bases.

4.3 Bulge Loops

Bulge loops have unpaired bases in one strand of a double stranded region, while the other strand is completely base paired.

A bulge loop may be only one base long but there is principally no limit for its length. The confirmation of a bulge loop depends on the type of the unpaired nucleotides and of the neighbor base pairs. Bulge loops may bend a stem loop and influence at this way the tertiary structure

4.4 Internal loops

The difference of internal loops to bulge loops is, that they have unpaired bases in both strands of a double stranded region.

They are called symmetric if they have in both strands the same number of bases. A symmetric internal loop is called mismatch.

The thermodynamic stability of an internal loop belongs to the number and the type of the unpaired bases and the type of the neighbor base pairs.

If there is the speech of unpaired bases in an internal loop that may be incorrect in most cases. That means that the loop may be like every other loop by stacking of unpaired bases or hydrogen bonds between bases relatively strong stabilized.

For example the loop E in 5S rRNA, which has 9 bases, which are all stacked and paired. This is an ideal interface to proteins.

4.5 Multibranched loops

Multibranched loops are also called junctions or bifurcations. Multibranched loops connect more than two helices. Between the helices may be unpaired bases, 4-helix-junctions are relative often, but there are other helix-numbers, too.

Unpaired nucleotides in multibranched loops decide in which way the helices will stack and they influence the tertiary RNA confirmation.

Thermodynamic parameters to junctions are not very popular, as there are so many possibilities of structure and sequence.

4.6 Pseudoknots

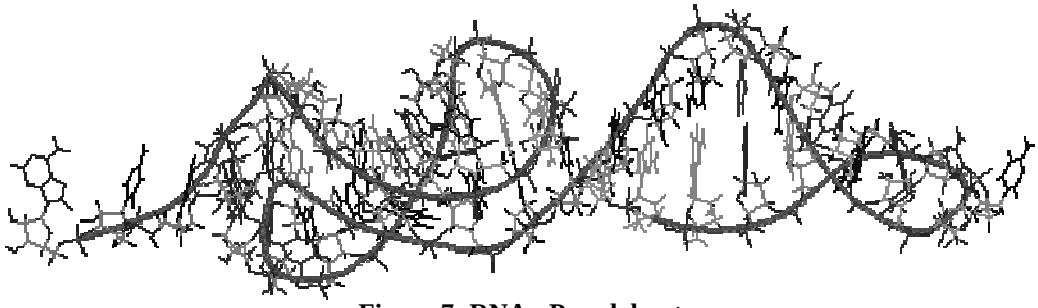


Figure 7: RNA - Pseudoknot

Pseudoknots are real existent and to the 3D structure of RNAs relevant constructs. But there is no concept to handle them by dynamic programming. So, they are enclosed of algorithms.n Without knots a secondary structure is a planar graph.

There are alternative methods to 2D-structure prediction which do not base on dynamic programming. Like STAR or other genetic algorithms which are able to handle pseudo-knots.

5. Free Energy Minimization

5.1 Assumptions

To do a structure prediction of RNA one possibility would be to calculate with free energy. Therefore we can use different algorithms and methods like recursive algorithms or dynamic programming. One of the favorite persons on predicting RNA structure is Michael Zuker. He decided to solve the problem of free energy minimization by dynamic programming. But first we have to look at some assumptions:

–Secondary structure has the lowest possible energy
–Free energies of stems depend only on the nearest neighbor sequences of base pairs only
–Stem and loop free energies additive

5.1 Values

Free energies of stems and loops used by the program come from the experimentally measured values on oligonucleotides.

5.2 Energy Minimization Algorithms

Before starting structure prediction by energy minimization we should think about what our algorithm has to do, what is our input and what should be our output.

Input:

The input has to be a primary RNA sequence, which can be downloaded from one of the DNA/RNA libraries such as the GenBank RNA library.

Output:

The output has to be a predicted secondary structure of our input sequence. The free energy should be minimized while maximizing the number of consecutive base pairs.

What type of algorithm to use?

An algorithm that uses a primary RNA sequence to produce lowest energy structure. The algorithm should enumerate all possible structures, then we choose one with lowest energy.

Problem:

If n base pairs, 2^n ways of piecing it together !
Many RNA molecules are tens of thousands of pairs long. Even a Supercomputer would not be able to evaluate the molecule in a reasonable amount of time.
For 400 base sequence (2^{200}), number of possible combinations is larger than the number of atoms in the universe.

Solution

We divide our large problems in smaller problems in other words we use a dynamic programming algorithm like mfold (M. Zuker).
Mfold predicts optimal and suboptimal secondary structure for RNA molecule using energy minimization method of Zuker. Solution of a long sequence is decomposed into solutions to smaller problems.

5.3 Dynamic Programming

So how does the this algorithm actually work?
Independent base pair algorithm is very simplified and it comes from the fact that helical (base pairing) sequences have negative free energies.

A recursive algorithm finds the lowest free energy structure from the subfragments of a sequence (Zuker and Stiegler, 1981). It starts a systematic search in all subfragments for the lowest free energy structure containing at least one base pair. The first subfragments considered are those capable of forming a hairpin loop closed by a single base pair. So in a first pass it will find the lowest free energy structures for all pentanucleotides in the sequence. The energy of the lowest free energy structures for each fragment are calculated with and without the constraint that the terminal nucleotides are paired, and stored in matrices V and W. The following pass will find the lowest free energy structures for all hexanucleotides in the sequence. In every step one nucleotide is added to the subfragments, and the best structure is searched for this new subfragment. To do this it makes use of the already calculated values for the smaller subfragments. This procedure is repeated until the lowest free energy for the entire sequence is found. The final structure is found by tracing back the steps taken. The major advantage of this method is speed. However it does not allow structures like pseudoknots, and it is difficult to incorporate more elaborate rules than the nearest-neighbor model.
Thermodynamic modeling always involves approximations, and it is quite possible that lowest free energy structures are not the only important ones. Therefore structure prediction

programs have been adopted to also generate *suboptimal structures* within a certain scope (Williamsen Tinoco, 1986).

5.3.1 Independent Base Pairs

The quickest and easiest route to RNA structure prediction is through the use of simple energy rules. One way is to assign an energy to each base pair in a secondary structure. That is, there is a function e such that e(ri, rj) is the energy of a base pair I,j. The energy, E(S) , of the entire structure, is then given by this formula, which simply says that the energy E of the system is the sum of energies of the individual base pairs. Even though this is method is attractive for its simples ness, it does not take into account the destabilizing effects of the various loops. So we need a little more complicated algorithm to compute minimum energy of folding. (Paul P. Gardner 2001)

$$E(S) = \sum_{i,j \in B} e(r_i, r_j).$$

Figure 8: Energy formula

5.3.2 Base Pair Dependent Events

This is to say that the total energy of folding is calculated over the whole secondary structure, containing both base pairs as well as loops.
When all the sub fragments are identified the program searches all the sub fragments for the lowest free energy containing at least one base pair. First pass it will find the lowest free energy for all the pentanucleotides in the sequence (if less no 2 structure formed).
The following time the lowest free energy for all the hexanucleotides, then heptanucleotides and so on are found. To do this already calculated values for sub fragments are considered. This is repeated until the lowest energy for the entire structure is found.
There is only one possible solution predicted by algorithm, corresponding to lowest energy state for the structure.
However, RNA in the cell may be associated with proteins or other cellular elements, so that the “lowest energy” state does not hold.
Pseudoknots are known to exist in RNA structure. However, they are often excluded from the algorithms to make calculations more efficient. But, it is known that pseudoknots do exist, and are important for RNA structure and function.

5.4 Example: Zuker´s Algortihm

Zuker's algorithm is based on rules of structure determination by free-energy calculations proposed by Waterman, who noted that newer approaches of structure determination revolved around the fact that the most probable structure should be the one with the lowest energy.

Waterman's free-energy calculations:

Let h (i, j) be the minimum free-energy (single hairpin) secondary structure on $a_i\, a_{i+1} \ldots a_j$, $i < j$, where a_i and a_j form a base pair and there is a single ended loop.
If a_i and a_j cannot form a base pair,
h(i, j) = +a (Waterman, M.S 1995)

The free energy functions are of the form:

a (a, b) = free-energy of an a, b base pair
x (k) = destabilization free-energy of an end-loop of k bases.
h = stacking energy of adjacent bases
b (k) = destabilization free-energy of bulge of k bases
g (k) = destabilization free-energy of an interior loop of k bases

Zuker, like Waterman, postulated that Biophysics rather than counting and maximizing the number of base pairs dictate RNA folding. He developed the energy minimization algorithm – which assigns free energy score to all possible structures. The structure with the lowest equilibrium free energy ? G is the correct one. The following rules are followed:

1. Stacking contributions are added from stems, not individual base pair contributions.

2. Energies are assigned to loops, stems, and all the structural elements, and are added up to give the overall ? G value of the structure.

3. Pseudoknots are not taken into account.

6. Excursion

In this last section, we will locate an RNA sequence and send it to a server running Michael Zuker's *mfold* program.

Step 1:

Locate an RNA sequence. One way to do this is to use the

Entrez browser/nucleotide sequence search.

Using the search terms "hepatitis ribozyme" I found 15 sequence records. One of these was L35894, titled "hammerhead ribozyme." I selected "FASTA report" which gave me the sequence on-screen. I selected the lines containing sequence (not the title line) and used the browser's Edit/Copy command to copy the sequence.

Step 2:

Go to one of the *mfold* servers. You can find one at Rensselaer. This gives you a form to fill out.

- Give your sequence a name. I called mine "hep hammerhead"
- Put your cursor in the sequence box and paste in the sequence which you previously copied.
- Indicate whether your sequence is linear or circular, using the linear/circular button.
- Change parameters if you wish. I used default values.
- If your sequence is 500 bases or less, you can select "immediate" calculation, and the results will come to your browser in a short time. If it's bigger, you will have to run in batch and have results sent to you by e-mail. In either case, enter your e-mail address in the appropriate box.
- Click on the "Fold RNA" button to send your sequence to the server. It might take a couple of minutes. Be patient!

Step 3:

Take a look at your results. First, you will have an energy dot-plot. This is a triangular graph showing bases which can form complementary pairs. Here is a dot-plot for the hepatitis virus hammerhead ribozyme (**Figure 9**)

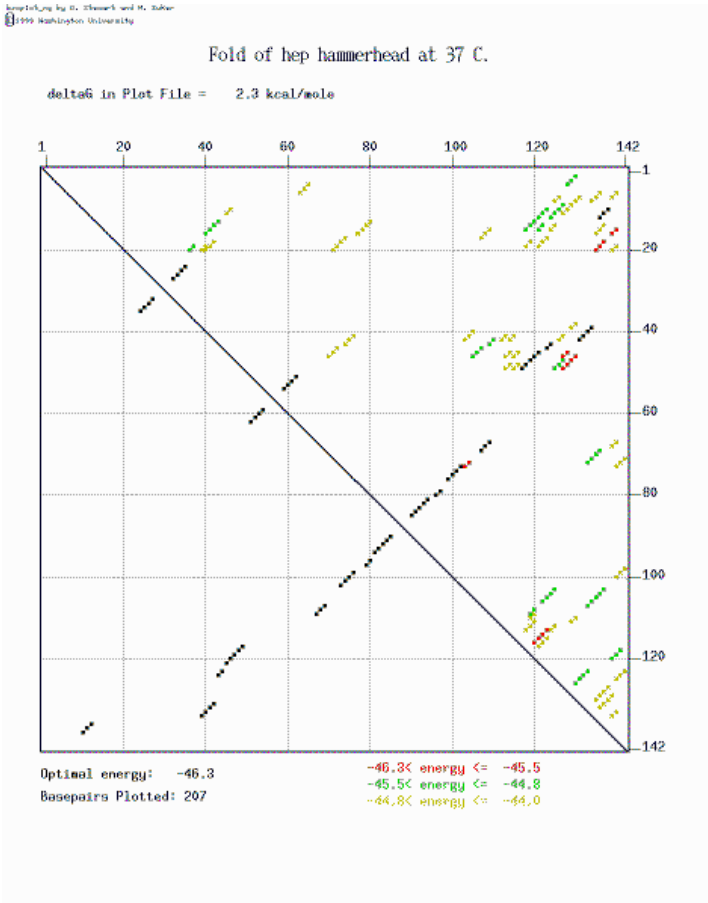


Figure 9: Dot-plot for the hepatitis virus hammerhead ribozyme

Next you will have a series of 2-dimensional representations of folded structures. These are arranged in order of increasing free energy (i.e., most stable first). Remember that these are only approximate free energies! Here is one of the structures calculated for the *E. histolytica* ribosomal RNA small subunit (**Figure 10**)

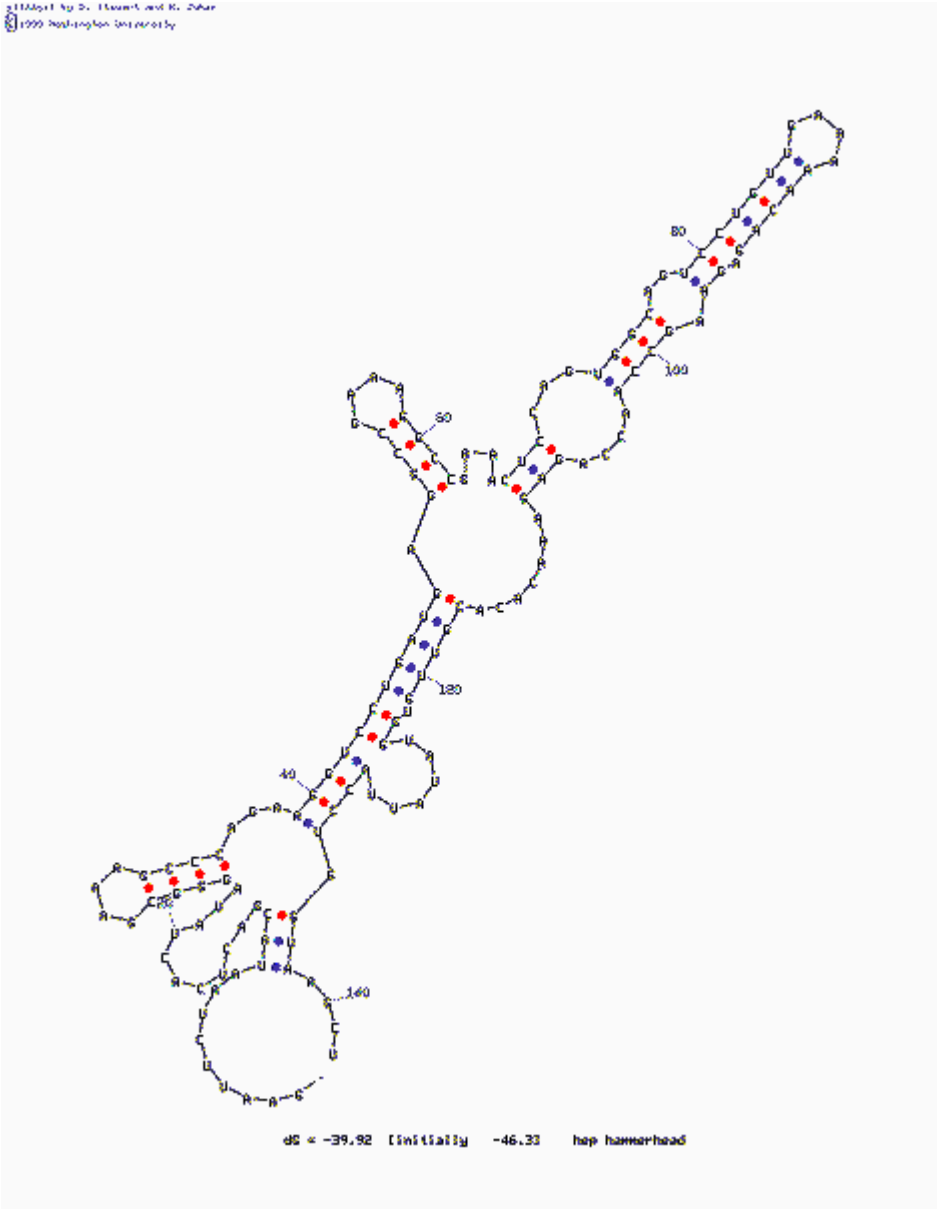


Figure 10: Calculated structures for the E. histolytica ribosomal RNA small subunit

That's it! You can try *mfold* with other RNA sequences, run it with different settings, etc. Further information is available on-line from the mfold User Manual.

7. References

Waterman, M.S 1995. Introduction to Computational Biology
Gardner, Paul P. 2001. Energy Minimization to Predict RNA Secondary Structure
Staben, Chuck 2001. RNA Structure Prediction
Patterson, Don 2000. RNA Secondary Structure Prediction
Walters, Eric 2002. Computer Applications in Biomedical Research
All Science Stuff.com 2002. RNA
RNA World 2000. Methods to study RNA structure and function
Selesniemi, Kaisa 2001, RNA Secondary Structure Prediction
The Central Dogma of Biology 2003. http://motif.stanford.edu/thesis/tRNA.html
Rivas, Elena; Eddy, Sean R. 1999. The language of RNA: a formal grammar that includes pseudoknots
Lungso, Rune B.; Zuker, Michael; Pedersen, Christian N. S. 1999.
Gerhard Steger: Bioinformatik, Birkhäuser
R. Merkl, S. Waack: Bioinformatik Interaktive, Wiley-VCH
H. Rehm, F. Hammar: Biochemie, Verlag Harry Deutsch
Andrea Hansen: Bioinformatik, Birkhäuser
G. Vogel, H. Angermann, dtv-Atlas Biologie, DTV
James Tisdall, Perl für Bioinformatik, O´Reilly
Neil A. Campbell, Biologie, Spektrum