

# RNA Secondary Structures

Adam Novak, Jotun Hein & Rune Lyngsø

July 23, 2010

## Background

The central dogma of molecular biology states that the flow of information is from DNA, through RNA, to proteins. What this means is that DNA is the storage medium of genomic information, while proteins are the molecules with functional and structural rôles. In this view, RNA is mostly considered a ‘supporting actor’, being the medium used for transporting information from storage to where it is activated.

RNA is short for RiboNucleic Acid, and DNA is short for 2'-DeoxyriboNucleic Acid (capitalisation just to indicate acronym). It should thus be of no surprise that the two types of molecules are similar. Both molecules consists of long chains of nucleotides, where a nucleotide is a combination of a phosphate group, a ring-formed pentose (i.e. carbohydrate with five carbon atoms), and a nitrogenous base (either a purine which consists of two aromatic ring structures and can be either adenine or guanine, or a pyrimidine that consists of a single aromatic ring structure and can be either cytosine, thymine (DNA), or uracil (RNA)). Good descriptions of RNA, DNA, nucleotides, and the central dogma are available at <http://en.wikipedia.org/>.

Where DNA molecules usually consist of two complementary strands forming the famous Watson-Crick double helix Watson and Crick [1953], RNA molecules usually just consist of a single strand. However, this strand can fold back against itself, so that local helices of consecutive base pairs are formed. Apart from the standard Watson-Crick pairings of C(ytosine) with G(uanine) and A(denine) with (U)racil, also the (G)uanine-(U)racil wobble base pairing is commonly observed in RNA structures.

The function of an RNA molecule is predominantly determined by the three-dimensional structure it forms, and information about this can therefore be very useful for studying a particular RNA. Knowing which bases form base pairs reveals a lot of information about the overall structure, so the set of base pairs is

commonly designated the secondary structure of an RNA (with the sequence of bases being the primary structure and the full three-dimensional structure being the tertiary structure). Moreover, base pair formation is the most important contributor to stabilising the structure of an RNA molecule. This led to early attempts at predicting RNA secondary structures that simply maximised the number of base pairs Nussinov and Jacobson [1980]. A short introduction to this method and extensions taking stacking, i.e. consecutivity, of base pairs and full thermodynamic model based prediction is available in Lyngsø [2008]. In this project you will be asked to use similarly simple algorithms to either count number of structures for known RNA sequences and compare it to some theoretical expression, or to modify the algorithm to take a co-transcriptional effect into account.

## Project Outline

The project will require the ability to write small programs that implement the algorithms for structure prediction and counting. A report on the progress should be expanded during the course of the project. The outline of the project is as follows

1. Introduction to programming in Python. This introduction will cover aspects such as variables, conditionals and loops and the associated block structures, complex data structures such as list, strings, dictionaries and objects, the use of functions to build structured programs, and recursive functions and dynamic programming.
2. Development and implementation of algorithms similar to the Nussinov algorithm for finding the maximum number of base pairs possible for a given sequence and for finding the maximum number of base pair stackings for a sequence. Furthermore, traceback of these numbers to generate optimal structures and test of performance on one or more real RNA sequences. A data set of sequences with known structures will be supplied.
3. The purpose of the first two parts of the project is to enable you to further investigate either
  - (a) How well the number of structures actually possible for a sequence corresponds to various analytical expressions developed by theoretical considerations Zuker and Sankoff [1984], Hofacker et al. [1998], Nebel [2004]. This will require that the algorithms and implementations developed in part 2 are modified to not only consider optimal structures

and find one such, but to count all possible structures, even if non-optimal. For the stacking context, the corresponding counting problem would be only counting the number of possible structures with all base pairs stacking, i.e. where there are no base pairs not stacking with another base pair.

To begin with, the relevant expression to compare this number with would be [Zuker and Sankoff, 1984, Eq. (5)]. For a sequence we can compute the value  $p$  in [Zuker and Sankoff, 1984, Eq. (6)] as the number of positions that can form a valid base pair divided by the total number of pairs of positions. Dividing the count with the theoretical expression should on average give 1, so the interesting thing to investigate is whether there is a deviation from this, and possibly whether to what extent this depends on the length of the sequence and  $p$ .

The main two reasons that we could see a deviation is that for real sequences we don't just have a probability  $p$  that a pair of positions form a base pair. Firstly because of the structure of valid base pairs where we just have four possible types and base pair validity is not a random thing but determined by base type. For example, we can never have that valid base pairs can be formed between any pair of positions  $i$ ,  $j$ , and  $k$  as there is no subset of three base types where all pairs are valid base pairs. Secondly, real sequences may have a bias towards either fewer or more possible structures. One way to untangle this effect would be to count number of structures both for the real sequence, and for a shuffled version of the real sequence where the base composition is preserved but the order is random.

Once the initial investigation is completed, time allowing you could look through the other two papers Hofacker et al. [1998], Nebel [2004] to find other entities to count and compare against. Beware though, that in these papers there is assumed to be a minimum separation between bases for them to be allowed to pair, so remember to include this constraint when you check whether two positions can form a valid base pair.

- (b) Incorporating an effect from so-called co-transcriptional folding into the prediction algorithm. The idea of co-transcriptional folding is that the RNA molecule is not just generated in a single operation, but is build up nucleotide by nucleotide by copying the DNA. Naturally structure will start to form as soon as we have the initial part generated, and Meyer and Miklós [2004] found evidence that real RNA sequences seem to be selected such that this structure formation works as a guide

towards the formation of the final structure. Essentially what they found was that if  $i, j$  is a base pair in the final structure, then there will be more long helices containing the base pair  $c, i$  for a  $c < i$  and fewer long helices containing the base pair  $c, j$  for  $c < i$ .

To incorporate this into the folding algorithm, you will first need to devise and implement a method that for each possible base pair determines whether it can be part of a long helix (in Meyer and Miklós [2004] long is defined to be nine base pairs). Once this is done, you should then modify the score gained for adding a base pair between  $i$  and  $j$  when computing an optimal structure such that you add  $\frac{b}{i \log(c-i)}$  for all  $c < i$  where  $c, i$  can be a base pair in a long helix and subtract  $\frac{a}{i \log(c-i)}$  for all  $c < i$  where  $c, j$  can be a base pair in a long helix. The values used for  $a$  and  $b$  should reflect the strength of this co-transcriptional effect. As we don't know what this is, you could try different pairs of values to determine which gives the best predictions on a set of sequences with known structures.

One thing to be aware of is that we are just using a simple base pair stacking criteria to predict structures. This will usually not perform too well, e.g. as compared to thermodynamic prediction. So it may be that the overall prediction is not good enough for inclusion of a co-transcriptional effect to show any noticeable effect. However, we won't know whether this is the case until we have tried it.

## References

- Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics*, 88:207–237, 1998.
- Rune Lyngsø. Lecture notes. [www.stats.ox.ac.uk/\\_\\_data/assets/pdf\\_file/0018/4626/rna.pdf](http://www.stats.ox.ac.uk/__data/assets/pdf_file/0018/4626/rna.pdf), 2008. RNA secondary structures.
- Irmtraud M. Meyer and István Miklós. Co-transcriptional folding is encoded within RNA genes. *BMC Molecular Biology*, 5:10, 2004.
- Markus E. Nebel. Investigation of the Bernoulli model for RNA secondary structures. *Bulletin of Mathematical Biology*, 66(5):925–964, 2004.
- Ruth Nussinov and Ann B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11):6309–6313, November 1980.

James D. Watson and Francis H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 248(451):765, April 1953.

Michael Zuker and David Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46:591–621, 1984.