Notes for lecture 17: RNA Secondary Structure Prediction

Adam Waite

The UCSC genome browser (http://genome.ucsc.edu/) is awesome. Search by position or gene name. A cartoon representation of the current chromosome appears, with a red box to indicate the current location. Next is a list of genes that are predicted to be the same, although they might have different names and come from different databases. Below that, the conservation of a given region among mammals and other organisms is displayed. All of these things are clickable/interactive, so have fun! Finally, the "Mapping and Sequencing Tracks" contains more options than you could ever use.

"RNA World" Hypothesis

A circularity regarding origins arises from the observation that DNA is transcribed into RNA, which is translated into protein, which then act to help replicate DNA: If any of these pieces were missing, the process would stop.

As RNAs are known to be able to fill the information-carrying roles of DNA (as a RNA double helix) and proteins (e.g., as a "ribozyme", or RNA that can catalyze biochemical reactions), it has been proposed that RNA was the original biopolymer.

RNA Structure

Unlike DNA, RNA is usually single-stranded. The entropic benefits of base stacking and the enthalpic benefits of base pairing result in the formation of secondary structures. Base pairing rules are the same as for DNA, except that 'T' (thymidine) is replaced by 'U' (uridine) and a G-U "wobble pair" is more likely to occur. Additionally, modified nucleotides not found in DNA are used in RNA which can also bind in non-canonical ways.

ssRNA can fold such that it forms base pairs with itself. This is known as secondary structure. The folding of this base-paired structure gives rise to tertiary structure. Tertiary structure allows RNA to act as a catalyst and perform other useful biological functions.

The three-dimensional structure of an RNA with tertiary folding is best visualized as a 3D model, however this is not convenient for many uses. Often, the secondary structure is of greatest interest and in this case, 2D representations suffice. In particular, secondary structure is much more affordable, computationally.

There are two common ways to represent the 2D structure of RNA. The first is pairing the relevant bases and leaving the unpaired ones to "loop out", resulting in a representation that looks like the floor map of an airport terminal. Another, more abstract, way of visualizing the structure is to write the RNA sequence in a circle and then connect the pairing bases with lines. While it is less obvious what the shape of the molecule is, this representation makes different topological features, such as pseudo-knots, immediately obvious.

... in Ribosomes

ribosomes are composed of about 50 proteins and 4 RNA molecules. Three of these RNA molecules are used as scaffolding and help give the ribosome its shape. The fourth RNA catalyzes the formation of the peptide bond between amino acids.

...in tRNA

The tertiary shape of this RNA is very important. One end allows for the attachment of an amino acid. The other end forms a loop of three nucleotides corresponding to the "anti-codon" of a particular DNA codon triplet. Overall, tRNA is shaped to effectively interact with the ribosome.

RNaseP is another RNA with catalytic activity that acts on RNA. It is able to specifically modify tRNA during its formation.

...in Gene Regulation

RNA secondary structure in the 5' untranslated region of a messenger RNA can determine whether or not the rest of the mRNA will be transcribed or translated. An RNA that performs this function is called a "riboswitch." An example of a riboswitch occurs in the control of genes used in glycine degradation. This gene is constitutively transcribed, but typically transcription is aborted before the coding portion of the gene is reached. However, when excess glycine is present, it binds to two different places in the 5' UTR. The conformational change brought about by this binding allows the polymerase complex to completely transcribe the gene, and thus allows the ribosome to translate the mRNA and produce the protein.

At least 5 different RNAs have been found that bind to 5-adenosylmethionine (SAM) to induce DNA binding and gene activation. Somewhat surprisingly, these 5 RNAs are very different in structure.

an RNA called 6S was discovered decades ago in E. coli and close relatives, but its function remained unknown until a few years ago. Researchers discovered that 6S is able to mimic an open promoter, effectively sequestering extra transcriptional proteins. As is the case with many functional RNAs, its structure is much more important than its sequence. Thus, its sequence has diverged greatly across different species of bacteria, which prevented other family members from being discovered until more sensitive, structure-aware, computational search procedures became available. It is now known to be present in all major bacterial groups, not just E. coli.

...in Viral Replication

The "hammerhead" ribozyme facilitates the self-splicing of a certain RNA virus that replicates by the "rolling circle" mechanism.

Structure Prediction

As can be seen by the examples, RNA structure is often more important than its sequence. This makes it hard, if not impossible, for traditional search tools, which look for sequence similarity, to locate RNAs with similar function.

Three main methods used to predict RNA structure (all ignoring pseudoknots) are:

The **max pairing** method uses dynamic programming to predict structure based on single sequence base pairing. It is simple but inaccurate.

The **minimum energy** method is similar to the Max Pairing method, but computes pairing as a function of its energy contribution to the overall structure. It is more accurate than the Max Pairing method, but can only find the optimal structure based on the energy function used.

The **loop-based energy minimization** method uses emperically-derived functions for at least five different types of loops known to form in RNA secondary structures. These functions are based on the contributions of the various ways in which neighboring base pairs stack. This method is more complex than the first two, but is more accurate. However, it can only find the single best structure based on the provided functions.