34. Measuring Selection in RNA molecules.

Naila Mimouni, Rune lyngsoe and Jotun Hein.

University of Oxford, UK

Measuring selection on RNA will help predict RNA function, identify new RNAs and pseudogenes. RNA does not have DNA periodicity and there is no equivalent of the genetic code. Therefore, we aim to extract selection information from conservation of secondary structure of a large number of RNA sequence alignments from different species, for different RNA families.

RiboNucleic Acid (RNA) is a polymer with a ribose sugar backbone. Each sugar has one of the four bases adenine (A), cytosine (C), guanine (G) and uracil (U) linked to it as a side group. Messenger RNA (mRNA) is one of the early discovered RNAs; it codes for protein. There is a wealth of other types of RNA families, called noncoding RNA (ncRNA) which play catalytic, regulatory, or structural roles in the cell. These do not get translated into protein and rather function directly as RNA. They are diverse and have a range of sizes from 21 nt (microRNAs) to >10,000 nt (Xist). An interesting example is microRNAs (miRNAs). These are a family of ncRNAs, 21-25 nt long, which are known to negatively regulate the expression of protein-coding genes. Recently, miRNAs have been found experimentally to be linked to cancer [1] through either downregulation of tumour suppressor genes or upregulation of oncogenes. Understanding the selectional constraints acting on different RNA families will certainly help discovery of new RNAs, identification of pseudogenes, as well as RNA function prediction.

Unlike DNA, RNA is typically produced as a single-stranded molecule which then folds intramolecularly to form secondary structure. Secondary structure is composed of stems (or helices) and loops. Stems occur wherever there are two parts of the sequence that are complementary;C-G,A-U (and G-U). Loops are those single stranded regions between stems.

Secondary structure of an RNA is quite informative. It is crucial for the RNA to conduct its function, and would therefore allow positions of functional sites to be identified in the sequence. Some studies have investigated structure conservation within RNA families. The process of compensatory mutations has been observed [2], whereby a mutation in one base is often followed by the mutation far down the sequence of its complementary base. However, we are still far from understanding selection in RNA molecules.

Measuring selection on protein molecules exploits the periodicity of the 3 bases forming the amino acid and the genetic code to calculate Ka/Ks. Ka is the number of non-synonymous substitutions per non-synonymous site, and Ks is the number of synonymous substitutions per synonymous site. The ratio of the two has been used to identify functional genes (protein coding genes) and pseudogenes. We aim to do the same for RNA.

There is no analog of Ka/Ks for RNA because RNA does not have DNA periodicity and there is no equivalent of the genetic code. We aim to extract selection information from conservation of secondary structure of alignments of homologous RNA sequences from different species, for different RNA families. It worth noting at this point, that RNA evolution is guided by the process of mutation on the DNA (before it is translated to RNA) and selection acting on the RNA to preserve structure stability and function.

We are using the Rfam dataset consisting of 503 different RNA families. To our knowledge, this is the largest dataset used for investigating RNA selection. The initial focus will be on miRNAs. Currently, the counting approach is being undertaken. For a number of 47 miRNA families, each containing aligned sequences with conserved structure, the number of substitutions is counted and classified according to location. For example, differences can occur as a singlet, in loop structures. This results in a 4*4 matrix of the four bases, where each entry records the number of times a base changes into another base. Alternatively, it could be a doublet, participates in stem structures and results in a 16*16 matrix. Each entry records the count of base pairs changing into another base pair. Finally, it could be that the observed differences do not participate in any structure at all (only 55%-60% of bases participate in RNA structure.). This is referred to as junk. MiRNAs, because they are short, do not have junk differences. At a later stage, it would be desirable to measure selection for longer and more complicated RNA molecules. Unfortunately, the structure of most of these is not known. Using algorithms like Pfold [4] to predict RNA structures is an alternative.

Devising a molecular model is central to the next stage. The rates would consist of singlet, doublet and junk rates. To obtain these rates, the set of homologous RNAs needs to be partitioned into classes of genes with similar homology, as explained in [2]. This would be novel in its own right.

Finally, these rates, used in a likelihood test, would give an indication of selection. Given a proposed RNA gene, determining whether it is a pseudogene would be based on the likelihood that it fits the pseudogene model, as compared to the functional gene model.

References:
----

[1] Caldas, C. & Brenton, J. D. 2005. Sizing up miRNAs as cancer genes, Nature Medicine 11, 712-714.
[2] Higgs, P. G. 2000. RNA secondary structure: Physical and computational aspects. Quarterly Reviews of Biophysics, 33,199-253.
[3] Griffiths-Jones, S. et. al. 2003. Rfam: an RNA family database, Nucleic Acids Research, 33 (1), 439-441.
[4] Knudsen, B. & Hein, J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research, 31(13), 3423-8.