

LIFE SCIENCE INTERFACE/ DOCTORAL TRAINING CENTRE

First Year Project Proposal - 2004

Supervisor: Jotun Hein

Summer Project

Title of Project: Measurement of Selection on RNA molecules.

Autumn Project

(delete if appropriate)

Description of project:

Motivation and Background: When analyzing protein coding genes it is very useful to measure the strength of selection. Genes that are not selected are often non-functional (pseudo genes). For selected genes, the strength and distribution along the sequence often gives functional clues to what the coded protein does and which positions are important. The most used such measure is the Ka/Ks ratio – that measures the relative amounts of amino acid changing versus “not amino acid changing” nucleotide substitutions. There is not an analogue of this for RNA molecules and the aim of this project is to consider this.

There are at least 2 approaches to this: Firstly, a pure counting approach, where all possible substitutions that could happen to the present molecule is considered and classified into those that changes and those that doesn't change the structure. When a set of homologous molecules are analyzed then the changes are estimated and do they avoid structure changing events this is taken as indication of selection. Secondly, the set of molecules can be analyzed under 2 scenarios: that is molecule has a RNA structure and that it hasn't. The hypothesis of pseudo gene could be tested by a likelihood ratio test and parameters in the molecular evolution process could be estimated. The two approaches to measuring selection are related. The second is superior as it allows estimation and hypothesis testing and is the one that we will suggest to be pursued.

Workplan: A large set of proposed homologous RNA genes must be assumed given from for instance the mouse and human genome.

i. The counting approaches could be done by aligning RNA structures from Human and Mouse for instance and classify whether a substitution would change the structure or not. Structure changing substitutions could for instance remove a base pairing at the end of beginning or a series of base pairing (called a stem) and thus shortening the stem. Alternative outside the boundaries of a stem substitutions could allow base pair formations and thus elongate the stem. Are such substitution low in frequency this could be the signature of selection.

b. In for instance the Knudsen and Hein model only 3 rates would be of interest, doublet rates, singlet rates and junk-rates. There are other alternative models that could be used. To characterize the differences between these rates, all homologous pairs of RNA genes are analyzed and a series of hierarchical tests performed to group RNA genes within classes that have the same molecular evolution. This would allow a good estimation of doublet, singlet and junk rates, that would be interesting in itself and used to quantify the amount of selection. Given a proposed RNA gene, a very natural test for whether it is functional or not would be to see if the three rates above very significantly different.

Comment: It is an unpleasant assumption of most, possibly all present methods that the structures does not change, but choosing closely related species like mouse and man should alleviate this.

Literature

Eddy, S. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet.* 2001 Dec;2(12):919-29.

Eddy, S. “Computational genomics of noncoding RNA genes” *Cell.* 2002 Apr 19;109(2):137-40.

Knudsen, B. and J.J.Hein (1999) "Using stochastic context free grammars and molecular evolution to predict RNA secondary structure (*Bioinformatics* vol 15.5 15.6.446-454)

Knudsen, B. and J.J.Hein (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* 2003 Jul 1;31(13):3423-8.

Moulton et al. (2000) Metrics on RNA Secondary Structures *J.Compu.Biol.* 7.1/2.277-

Perriquet et al.(2003) Finding the common homologous structure shared by two homologous RNAs. *Bioinformatics* 19.1.108-116.

Location: OCGF