# *Identification of tertiary structure interactions from evolutionary analysis*

Jakob Skou Pedersen 27.8.11

BACKGROUND: The evolution of RNA and protein sequences is constrained by the conservation of structure and hence function. Because of this, structural inferences can be made from observations of the primary sequence evolution. This has long been exploited when making RNA secondary structure models for RNA sequences. RNA structure consists of base-pairing (A-U, G-C, and G-U) regions and single stranded regions. RNA base-pairs are often conserved over primary sequence, resulting in compensatory substitution, where two correlated nucleotide changes are observed (e.g., A-U to G-C). Having observed such compensatory changes in alignments of RNA sequences, thus provides strong evidence that the involved positions base-pair. This approach was formalized and used in a stochastic model for RNA secondary structure inference in [1].

However, many types of structural interactions exist in the three-dimensional structure of RNA molecules, in addition to base-pairs. These other types of interaction do not follow the base-pairing rules given above, but may potentially occur between any combination of bases. In common with the base-pair interactions, compensatory changes may be observed, where the two involved positions show a correlated pattern of evolution and often appear to change simultaneously. Observing such correlated changes thus allow general tertiary structure interactions to be inferred. This approach was taken in [2]. The value of this approach is not limited to RNA, but should also apply to protein structure prediction. For the approach to work, at least parts of the tertiary structure of the RNA or protein should be well-conserved and deep alignments encompassing many substitution events should be present.

In [2], two competing phylogenetic models are constructed, which describe the evolution of a pair of positions as either correlated or independent. The final prediction will depend on the likelihood ratio between the models. An alternative approach is to use a Monte Carlo method to sample a large number of simulated outcomes of the evolutionary process with independence between positions, which can then act as a null set. For a given pair of positions, the number of observed compensatory changes in original data set can then be compared to the distribution of the same statistic in the null set.

This approach was taken for the Evo-P method, which assigns P-values to given RNA secondary structures given multiple alignments that were not used for the structure inference [3]. In Evo-P, the given secondary structure defines which positions are hypothesized to be correlated. However, the approach can be applied to in an all-against-all manner for every pair of positions in a given input sequence

PROJECT: The goal of this project is to evaluate the Evo-P approach for identifying positions that exhibit a correlated substitution pattern due to tertiary structure interactions. The measure defined for Evo-P should be evaluated generally for tertiary structure interactions. Input data sets can be taken from or just inspired by [2].

PERSPECTIVE: One potential problem with the Evo-P approach is that it infers the branches on which substitution events happened. Compensatory changes from both sides of a position-pair fall on the same branch. This approach does not handle uncertainty in the placement of substitutions on branches optimally. Instead of counting compensatory substitutions in this way, one could define a substitution model similar to the one used in [2], and estimate the expected number of compensatory changes from that instead. If tertiary structure interactions can be inferred efficiently, the generally very hard problem of three-dimensional structure prediction is dramatically simplified.

[1] B. Knudsen and J. Hein. RNA Secondary Structure Prediction Using Stochastic Context-Free Grammars and Evolutionary History. Bioinformatics (1999) vol. 15 (6) pp. 446-454

[2] Yeang et al. Detecting the coevolution of biosequences--an example of RNA interaction prediction. Mol Biol Evol (2007) vol. 24 (9) pp. 2119-31. http://mbe.oxfordjournals.org/cgi/content/full/24/9/2119

[3] See supplemental material of: Parker et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011). In press.