

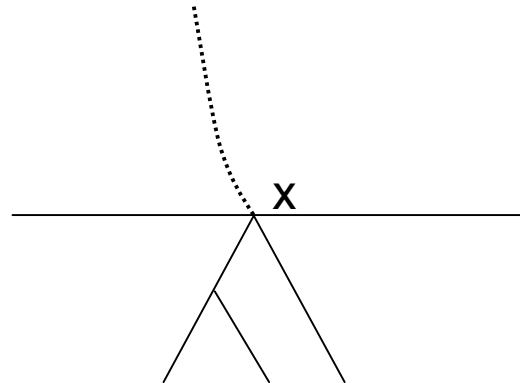
# Quantifying the Structural Value of Evolution

28.5.07

The title of this project might sound very technical, but fundamentally it is very simple. The last few years have seen the rise of comparative genomics to become a key approach to interpreting genomic information. For instance in gene finding, this was originally done for the specie of one genome only, that was scanned and regions with coding characteristics was predicted to code. With the advent of more mammalian genomes, this is done with more aligned genomes and the single genome is now complemented with the observation of how different regions evolve. Since coding regions evolve differently from non-coding regions, this will increase reliability of coding prediction. The question to be addressed in this project is to quantify the value of observing a single genome, relative to observing the genome evolve for a specified period. The comparative principle is used on several problems (also RNA or protein secondary structure prediction) and the question is likely to have different answers.

$P(x)$ :

$P(\text{Further history of } x)$ :



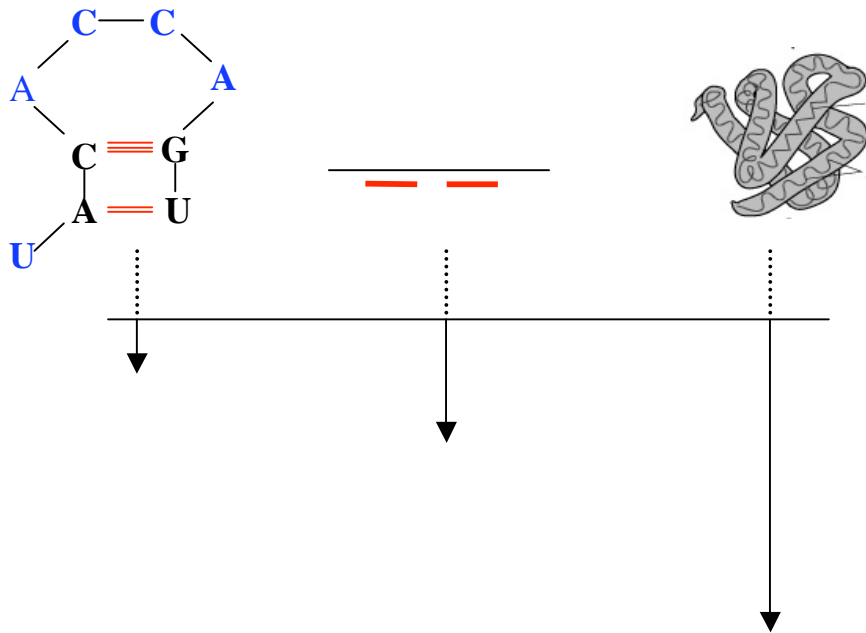
The evolution of an object (for instance sequence) can be decomposed into two parts – the evolution until the most recent common ancestor (MRCA) & the further evolution down to the present. If there is only copy of the object, then that object is the MRCA. Inference about a hidden structure can be performed if evolution is dependent on that structure and this would use a product of the 2 probability terms in the illustration, which statistically the right thing to do. However, it could be done on the two terms separately, which would lead to a decomposition of what contributed to the inference. At first, the two components are very different. The top one describing the outcome of an infinitely long history, but only one object and no differences. The lower term describes several observations and only finitely much evolution. Clearly, if branch lengths shrunk to zero, the latter would be totally non-informative.

The now standard approach to comparative genome annotation has two components: First a probabilistic description of the distribution of genes in the genome, describe prior belief. Secondly, a conditional probabilistic description of the sequence of the genome given an annotation. The genome annotation is not observable, but the genome is. Observing the genome will define a posterior distribution on annotations. To do this comparatively necessitates a description of how genomes are expected to evolved conditional on annotation. Even if only one genome is to be annotated it is natural that the prior on annotations are chosen to be the equilibrium distribution of the evolutionary process.

For RNA secondary structure prediction the slightly more advanced stochastic context free grammars (Knudsen et al., 1999).

The algorithms used in this are standard models of sequence evolution (Felsenstein, 2002) and Hidden Markov Models (Durbin et al., 1998).

The project could allow decomposition of the contributions of the static and dynamic contributions in comparative approaches. Additionally, it could also rank diverse problems according to suitable they are approaches. Biological intuition probably would say “RNA secondary structure > Gene Finding > Protein Structure Prediction”, but that has not yet been investigated in an exact manner.



Biological intuition says that the comparative principle is strong for RNA structure prediction, medium for gene finding and directly weak for protein secondary structure prediction. In this illustration it is depicted as the length of the period that would be necessary to be observed to have equivalent information as the static component.

The project would allow several decompositions: dynamic evolution versus static, possibly comparison between very different problems, but could also be used to compare inference within the same class, like “How much easier is it to predict the structure of a small RNA molecule to a large RNA molecule?”.

The sketched problem is big and it is important to define a well defined sub-problem, that can be addressed in 8 weeks. There are several points that could be refined, such as how to measure how a distribution is converging towards the truth.

Workplan. At first we will only describe this problem as if it only applied to gene finding in a genome. To define the problem it must be defined how to measure quality of the prediction. When using simulations, the probability that the prediction assigns to the “truth” could be a reasonable measure.

Simulation set:

i. Using the method described in Pedersen and Hein (2003) and parameters estimated from mammalian genomes, simulate a genome with annotation. Annotate this genome and tabulate the reliability of the annotation.

ii. Let this genome evolve in periods of 30 million years in total of 420 million years and observed. Again annotate and tabulate using the aligned simulated genomes. A genome can here be annotated in two ways, using the likelihood function conditioned on a given annotation, A: i)  $L_A(\text{root genome}) * L_A(\text{root genome} \rightarrow \text{evolved genomes})$  or only ii)  $L_A(\text{root genome} \rightarrow \text{evolved genomes})$ . The first case is the full likelihood of the data, while the second case only uses the observed evolution.

The evolutionary setup is was unrealistic as if assumed a genome could be observed over long periods, so an analysis based on a phylogeny should also be done.

iii. Make a small bifurcating tree with 8 leaves and all edges corresponding to 30 mill. years. and redo the analysis from ii.

## Literature

Durbin, Eddy, Krogh and Mitchison (1998) “Biological Sequence Analysis” Cambridge University Press

Felsenstein (2002) “Inferring Phylogenies” Sinauer

Goldman N., Thorne JL and Jones D.T. (1996) "Using Evolutionary Trees in Protein Secondary Structure Prediction and other comparative approaches. *J. Mol. Biol.* 263.196-208.

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)

Pedersen, J.S. and J.J. Hein (2003) "Gene finding with as hidden Markov model of genome structure and evolution" *Bioinformatics* 19.2.219-227.

Pollock, Goldman and Taylor, WR (1999) "Coevolving Protein Residues: Maximum Likelihood Identification and Relationship to Structure" *J.Mol.Biol.*287.187-198

Siepel A and Haussler D. (2004) "Combining phylogenetic and hidden Markov models in biosequence analysis" *J Comput Biol* 11:413-428