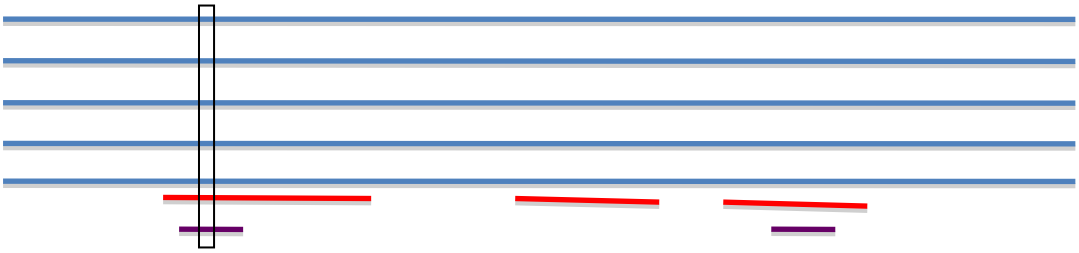
Simultaneous Genome Annotation for Multiple Levels

21.8.11

Genomes are annotated for a variety of functions/properties/….. . Three major classes are: Protein coding genes, RNA coding genes and regulatory signals, but there are others. These annotations are frequently done independently. Ie first annotate for protein genes, then for RNA coding genes etc. Ideally one should pursue one integrated approach, but doing this is no trivial matter. Annotation typically have two components: Firstly, a distribution of the possible annotations prior to observing the sequence [annotation can be interpreted as a set of hidden variables, that in contrast to sequence, can't be seen.]. Secondly, a conditional distribution describing the probability of a sequence giving the annotation. This will then define a probability of the annotation given the observed sequence.

Pedersen et al. (2004a,b) considered models of sequence evolution conditional on both RNA and protein annotation. This was done by assuming that the selective effects [reduction in rate] of protein and RNA was independent and could just be multiplied. That is probably a reasonable approximation.



5 genomes are aligned. There are 3 protein [red] and two of these have RNA structures [violet]. Given the structures [protein + RNA] it is easy to make a model of the genome. Modeling the RNA level [often stochastic context free grammars [SCFGs] or the protein level [often hidden markov models] is also easy, but doing it simultaneous is a challenge.

However, how to combine probability measures for the annotation is more of a challenge. In Pedersen et al. (2004a,b) this was circumvented by assuming that the protein level was known. It has been suggested that the concept of "Factorial HMMs" introduced in Ghahramani and Jordan (2006) could solve this. Zsuzsanna Sukosd suggested to solve the multiple annotation problem by first running one HMM that annotated the genome with combinations of {Protein [0/1], RNA [0/1], signal [0/1]} this at least 8 hidden states that would create islands on the genome of 3 kinds. Then one could use HMMs and SCFGs for this islands individually. This could be a very realistic and easily programmable approach to the problem.

There might very well be other approaches (see also Joanna Davies [2005])

Project:

Read and summarize key papers and try to make precise prescription for a 2-level annotation problem.

References

Pedersen, J.S., Meyer, I.M., Forsberg, R., P. Simmonds & J. Hein (2004) A comparative method for finding and folding RNA secondary structures in protein-coding regions. Nucleic Acids Research, 32, 4925–4936

Pedersen, J.S., Forsberg, R., Meyer, I.M. & Hein, J. (2004) An evolutionary model for protein-coding regions with conserved RNA structure. Mol. Biol. Evol., 21, 1913–1922.

Joanna Davies (2005) : Combining different grammars to make multiple annotations of a single sequence

Zoubin Ghahramani , Michael I. Jordan (1996) Factorial Hidden Markov Models

Yin, J. Jordan, M. I., and Song, Y. S.. Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data, Proceedings of ISMB 2009, Bioinformatics, 25 (2009) i231-i239