A context dependent pair hidden Markov model for statistical alignment of biological sequences

Catherine Matias

joint work with Ana Arribas-Gil

We propose a novel approach to statistical alignment of nucleotide sequences by introducing a context dependent structure on the substitution process in the underlying evolutionary model. We propose to estimate alignments and context dependent mutation rates relying on the observation of two homologous sequences. The procedure is based on a generalized pair-hidden Markov structure, where conditional on the alignment path, the nucleotide sequences follow a Markov distribution. We use a stochastic approximation expectation maximization (saem) algorithm to give accurate estimators of parameters and alignments. Results on simulated as well as real data are provided. In particular, we investigate the performance of the method for aligning a set of sequences of vertebrates pseudogenes, known to have a high mutation rate from CG dinucleotide (which are denoted CpG).