

2. Statistical alignment with a sequence evolution model allowing heterogeneous evolution behaviors along the sequence

Ana Arribas Gil

*Equipe Probabilités, Statistique et Modélisation
Batiment 425, Université de Paris-Sud, 91405 Orsay Cedex, France
ana.arribasgil@math.upsud.fr*

We present a stochastic sequence evolution model based on TKF models (Thorne, Kisino and Felsenstein 1991, 1992) and the fragment insertion-deletion model (Metzler 2003) where heterogeneous behaviors along a DNA sequence are possible. We propose two algorithms to estimate evolution parameters and to align sequences from our model.

The topic of this paper is the estimation of alignments and mutation rates from a stochastic sequence-evolution model that allows two possible evolution behaviors along a DNA sequence in order to consider its heterogeneity and to determine conserved regions. The sequence is seen as being divided into slow and fast evolution regions. The boundaries between these sections are not a priori known and must also be estimated from the data. This evolution model is based on a fragment insertion and deletion process that works on fast regions. Slow regions are conserved along the time and only substitutions are allowed. The same substitution process applies with different rates on each kind of region.

The evolution model induces a pair hidden Markov structure at the level of alignments and conserved regions and thus makes possible efficient statistical alignment algorithms. We propose two complementary approaches for estimation: a MCMC method and a maximum likelihood technique. Accurate results on simulated and real data are shown where the algorithms provide consistent estimations for the mutation rates and plausible alignments and sequence segmentations.