Bayesian Partitioning of Sequence Data and Identification of Regulatory Regions

Yussanne Ma, Michael P.H. Stumpf, Maria De Iorio

Elucidating the variation in the rate at which different parts of the genome evolve is biologically important. In particular, ranscription factor binding sites have been observed to evolve significantly slower than surrounding regions of non-coding DNA. Detecting such rate variation, however, remains a great challenge in statistical bioinformatics. The most promising approaches appear to be found in phylogenetic footprinting, a method for promoter prediction based on the idea that regulatory regions are evolutionarily conserved even between very distant species.

We present a novel statistical method for phylogenetic footprinting which uses a Bayesian framework combined with Markov chain Monte Carlo (MCMC) methods. A Bayesian partition model is developed to estimate evolutionary rate variation along the DNA sequence through comparison of evolutionarily distant species. The Bayesian framework allows us to incorporate in the modelling strategy "biologically meaningful" information and to account for different sources of complexity in the data. In contrast to some existing phylogenetic footprinting procedures, our approach does not rely on heuristics or *ad hoc* assumptions about sequence evolution. Full posterior inference is performed through reversible jump algorithm.

With the increased availability of sequence data, whole-genome comparisons between species have become possible. The proposed scheme has the advantages of being computationally efficient and readily applicable to vast datasets without manual intervention. We demonstrate our approach on simulated data and on real data which consist of a set of genes for which experimentally verified transcription factor binding site data is available.