Poster C-44

Bayesian Partitioning for Phylogenetic Footprinting

Authors:
Yussanne Ma (*Division of Molecular Biosciences, Imperial College London*)
Dr. Maria De Iorio (*Division of Epidemiology, Public Health and Primary Care, Imperial College London*)
Dr. Michael Stumpf (*Division of Molecular Biosciences, Imperial College London*)

Short Abstract: Reliable promoter prediction is one of the big challenges in bioinformatics today. Phylogenetic footprinting is a method which uses the property of evolutionary conservation of promoters to identify them through the comparison of distant species. We present a novel algorithm for phylogenetic footprinting which uses a Bayesian hierarchical framework combined with Markov chain Monte Carlo method. Our method uses Bayesian partitioning to estimate evolutionary rate variation along the DNA sequence.

Long Abstract:

One of the most complex biological processes in living organisms is their response to external stimuli and surroundings. These responses include multi-step biochemical pathways to produce often apparently instantaneous shifts in hormone levels, physical changes or slower acting permanent adaptations to environmental shifts. At a cellular level, these changes are predominantly brought about by altering the expression levels of specific genes and proteins.

The process of protein expression through transcription, translation and post-translational modification is beginning to be understood qualitatively. However, there is much less information known about the mechanisms involved in transcriptional regulation, the process which determines gener and protein expression levels. Transcriptional regulation is responsible for much of the complexity and diversity in living organisms and therefore holds the key to understanding their biological function at a molecular level.

There are several mechanisms for transcriptional regulation, but perhaps the primary and currently best understood one is through the binding of transcription factors to the regions upstream of genes in the DNA sequence. Transcription factors are proteins which bind to cis-regulatory promoter regions in the non-coding area upstream of the gene. The promoters, or transcription factor binding sites, are 10-15 base pairs long. Cis-regulatory regions in E. Coli are about 300 base pairs long and often contain several transcription factor binding sites. The binding of different combinations of transcription factors to their cognate sites results in differential expression of genes in response to external stimuli and/or in different tissues or at different stages in the organism's development.

Despite the many recent experimental and computational advances in evolutionary genetics and bioinformatics, the molecular machinery underlying transcriptional regulation remains poorly characterised. One of the main roadblocks is the identification of transcription factor binding sites in the genome. Much progress has been made in the sequencing and

identification of orthologous genes, but comparatively few reliable results have been obtained as far as regulatory regions are concerned.
Many regulatory elements have been shown to be evolutionarily conserved, even between very distant species. Mouse-human comparisons demonstrate 1.5% of highly conserved non-repetitive non-coding DNA. Comparison of the upstream regions of orthologs in distant species effectively filters out the non-functional sequences which have diverged during evolution, leaving the promotors which have been preserved due to their functional importance. Promotor prediction by comparison between distant species is called phylogenetic footprinting.
With the increased availability of sequence data, whole-genome comparisons between species have become possible, increasing the popularity of comparative analysis techniques in recent years. We present a novel algorithm for phylogenetic footprinting which uses a Bayesian hierarchical framework combined with Markov chain Monte Carlo (MCMC) methods.

In our model we seek to identify regions of relatively high conservation in the genome by estimating the evolutionary rate variation along a sequence. We use a technique called Bayesian partitioning which separates a sequence into segments of different evolutionary rate.

The ultimate goal of our approach is not only to estimate the rate variation along the DNA sequence, but also the evolutionary time and ancestral states in a phylogenetic tree. A Bayesian framework was developed to provide a unifying structure for dealing with the potential sources of uncertainty in the data in a consistent manner and to incorporate prior knowledge into the model. The unknown quantities, such as rate variation, evolutionary time and ancestral sequences, are organized into levels of hierarchies which can be treated in a straightforward Bayesian framework.

Through Bayes' theorem, the posterior distribution of the parameters, or the unknown variables in the model, can be obtained by combining the prior distribution on the unknowns and the likelihood of the data.

The joint posterior distribution of the ancestral states and the evolutionary rate cannot be calculated in analytical form. However, the posterior probabilities of the quantities of interest can be approximated by drawing the evolutionary rates and the ancestral configuration from their posterior distribution.

In our algorithm we run a Gibbs sampler on (i) the evolutionary rate variation along the sequence, (ii) the rate along each branch of the phylogenetic tree, and (iii) the ancestral states. The phylogenetic tree is updated at the transition from one state to the next in the Markov chain, moving through the levels of hierarchy. At each level, a new state is proposed by drawing a new random value for the unknown variable. The posteriors of the present state and proposed state are calculated by combining the priors on the unknowns and the likelihood of the data and all other variables.

The model was tested on data sequences created by simulating the mutation of a nucleotide sequence over several million years. The rate variation was built into the mutation process, including a conserved region with lower mutation rate. Initial testing of our model on

simulated data sequences shows good convergence of all model parameters and an accurate estimate of the rate variation across the sequence.

Our approach uses no biological heuristics and makes no assumptions beyond the standard phylogenetic premises. It can therefore be applied de novo to any sequence data. The method allows a parsimonious representation of rate variation along the genome using only the data and very general and non-informative priors. In the special case of two species comparisons, the model can be simplified using exact analytical results, which leads to a large reduction of computational time.

Bayesian partition formalism readily lends itself to formal model selection. That is, one can decide whether or no there is evidence for significant statistical rate variation across the genome. The Bayesian approach goes well beyond the p-values (or associated information criteria based on likelihood) at the centre of frequentist approaches, but allows us to express measures for the probability that the evolutionary rate in a given region is in fact reduced.

While these advantages are generally within the realm of statistical theory, this new approach will allow us to determine whether or not present sequence information from related species is in fact sufficient to detect regulatory regions. Previous approaches have employed heuristics which may have only picked up the low hanging fruit. Generally there appears to have been notable lack of success in finding and confirming new sites involved in transcriptional regulation using comparative approaches.

Therefore from a biological point of view, the outcome of a throrough statistical analysis will provide important insights into the likely success of in silico approaches for detecting regulatory sequence elements. However, even if this should not be possible in general, our approach allows us to systematically identify and map regions conserved between species and evolutionary intuition tells us that these are worth further investigation.