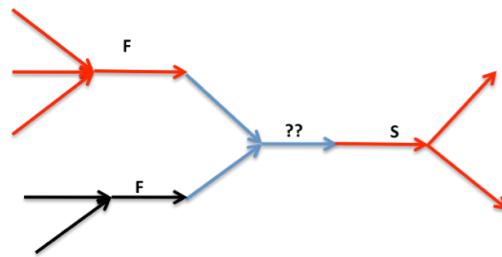


Markov Random Field model for fast/slow loss-gain of reactions in a metabolism

15.6.09

Motivation and Background. Continuous Time Markov Chain has been used to model the evolution of DNA and protein sequences in the late 60s (Jukes-Cantor, Neyman, Dayhoff). They are normally formulated in terms of a rate matrix, Q , and time between the two homologous sequences to be analyzed, t . Q has dimensions $k \times k$, where k is the size of the state space. q_{ij} is the rate with which the process jumps from i to j . These models have been extended seriously to incorporate a long series of observed biological factors, such as rate heterogeneity, context dependency, insertion-deletion events, hidden structures (annotation by genes for instance) and more general rate matrices (Yang, 2006). If k is of moderate size, the transition probabilities, $p_{ij}(t)$ [in matrix form $P(t) = \{p_{ij}(t)\}$], can be calculated by matrix exponentiation, $P(t) = \exp(tQ)$.

Models have now moved ahead and attempts to model phenomena such as structure, networks and shape. These objects are more structured than sequences and there will be less data. This poses many interesting challenges. One step forward in sequence analysis was to introduce variation in rates with which sequences evolve. Yang (1994) used independent rates for different positions while Felsenstein and Churchill (1996) used a Hidden Markov Model (HMM) to describe rate variation. The latter approach has the additional advantage that it creates a neighbour structure that is observed in real sequences. It would be very natural to use the analogous model to HMM on network evolution. We will focus on metabolic pathways. The natural framework for this would be to use Markov Random Fields for this. It is easy to describe conditional probability on graphs in terms of conditional distributions. For simplicity we will assume that there are only two rate categories, fast and slow (F and S), and that the metabolic pathways is given in terms of some hypergraph $-(\{E_i\}, \{V_j\})$. The V 's are labelled vertices [metabolites] and the E 's are sets of vertices [reactions]. We will assume that reactions are reversible and thus edges will have two sets of nodes associated, but it is not defined what is input and which is output. Let the probability that a give edge is F or S depend on the states of its neighbours, like $P(H(E_i)=F) = f(|F|/|\text{neighbors}|)$. $H(E_i)$ is the hidden state of E_i and the probability is here chosen to be a function of the fraction of neighbors having fast hidden states. This defines a probability distribution of F and S on all the reactions of this network.

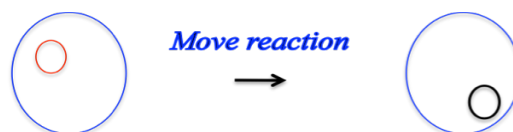


This metabolism has 4 reactions and we want to assign the probability of F/S hidden state for the reaction with "??". It has 3 neighbors [shares a metabolite/node with these reactions]. two of the three neighbors have F and probability of F for the reaction in question is $f(2/3)$.

To use this model for data analysis a few problems must be solved:

- An evolutionary model of the metabolism will have to be chosen conditional on hidden states. The conditional models are 2 state models, which need 2 parameters [rate from absence, rates from presence]. Thus 4 parameters in total. The conditional rate matrix is Q_H and associated probability matrix $P_H(t)$. The probability of the data is $\sum_H P(D|H)P(H)$, where $P(D|H)$ is the probability of the metabolisms given the hidden configuration and $P(H)$ the probability of the hidden configuration.

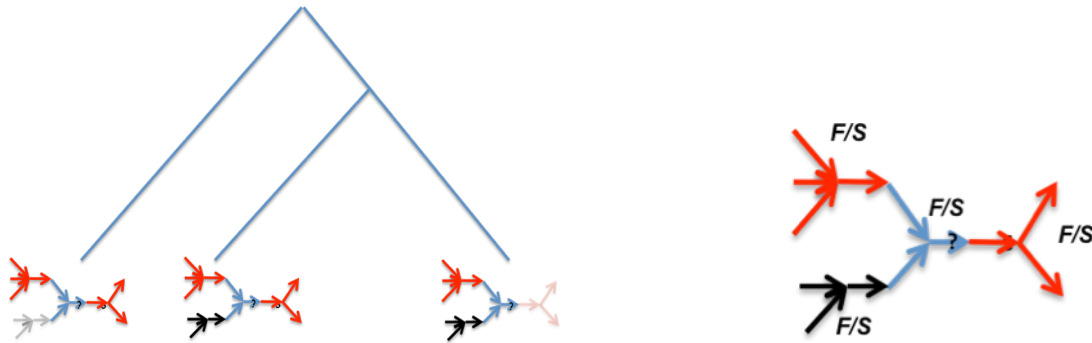
We then need the analogues of Forward, Viterbi, Forward-Backward and Baum-Welch from HMMs. For graphs there are no simple algorithms for this and we suggest a Gibbs sampler,



The Gibbs sampler will have a present "reaction" whose hidden state can be sampled correctly conditional on the hidden states of the all other reactions. The present "reaction" is then moved around. This will create a trajectory in the set of hidden states.

Following the trajectory of the Gibbs sampler would allow:

1. Summing over all hidden configurations to get the probability of data.
2. Mapping the most likely hidden configurations.
3. Mapping the marginal probability for each hidden state.
4. Re-estimating the parameters in the model [Emission Probabilities and MRF parameters].



3 metabolisms related by a phylogeny have been observed (left). If the reactions added/deleted independently, then the likelihood of a single reaction can be calculated easily given the hidden state. The likelihood of all three metabolisms can be calculated by independence given all hidden states. The challenge is to sum over all configurations of hidden state according to their probability (right).

A large number of metabolic annotations is available in public databases such as Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa et al., 2006) and BioCyc (Karp et al., 2005). Here we will use metabolic annotations from KEGG for a subset of bacteria belonging to class gamma-proteobacteria. Metabolic annotations for over 200 gamma-proteobacterial genomes are available in KEGG. These bacteria are further classified as Gamma-enterobacteria (about 80 genomes) and those belonging to miscellaneous Gamma orders including Pseudomonadales, Xanthomonadales and many others. To focus on a dataset which is manageable we will limit our analysis to individual amino acid metabolisms which range from 6 (D-Alanine Metabolism) to 105 (Tyrosine Metabolism) reactions. These datasets can be obtained from KEGG. For example, the universal Glycine, serine and threonine metabolism containing 68 reactions is available at http://www.genome.jp/dbget-bin/get_pathway?org_name=map&mapno=00260 and the corresponding metabolism in E. coli K-12 with 30 out of 68 reactions present can be seen at http://www.genome.jp/dbget-bin/get_pathway?org_name=eco&mapno=00260.

Work schedule:

- Read the key papers
- Implement simulator of metabolisms
- Implement algorithms 1 - 4.
- Test them on toy example where calculations can be done exhaustively. The number of hidden configurations is limiting ($2^k - k$ is the number of reactions). So the toy example should have at most 12 reactions. The equilibrium probabilities can then be calculated solving the set of linear equations. It would be advantageous to investigate a phylogeny with very few homologous metabolisms (like 3), where calculations can be followed manually, and many (>50) where the evidence for F/S structure of the metabolism is more obvious.
- Use model to analyze the real data set.
- Possible extensions would include making goodness of fit of the model, extending to more hidden states, testing for correlated mutations,..

Comments. The proposed model should be much faster computationally and simpler to implement, than the context dependent model of Mithani et al. (2009). Both this and the Mithani model has propensity to favour connected components, which is observed in real metabolisms.

References.

- Clifford, P "Markov Random Fields in Statistics"
- J. Felsenstein and G. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol., 13:93--104, 1996.
- Kanehisa, M. et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res., 34, D354-D357.
- Karp, P. et al. (2002) The Pathway Tools Software, Bioinformatics, 18, S225-232.
- Mithani et al. (2009) A stochastic model for the evolution of metabolic networks with neighbor dependence Bioinformatics
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306-314