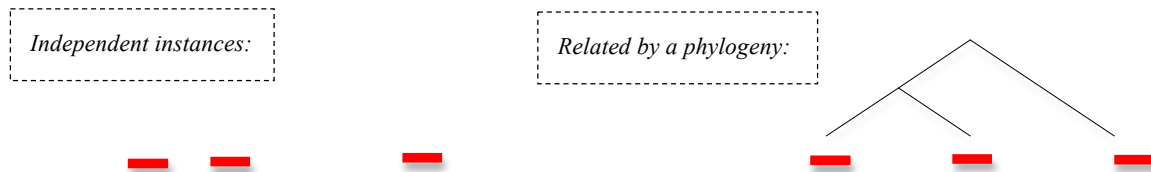


Evolutionary Models corresponding to Complex Patterns

5.2.10

Background and Motivation. Signal description is done quite differently in single genomes and in multiple genomes. In single genomes very complex models are often designed to describe the common features of a set of patterns, where these patterns are evolutionary unrelated and assumed to be independent instances from the same underlying distribution. Major classes of techniques are i. Position Weight Matrix (PWM), ii. Hidden Markov Models (HMMs), iii. Bayesian Networks (BNs), iv. Neural Networks (NNs) and v. Vector Support Machines (VSMs).

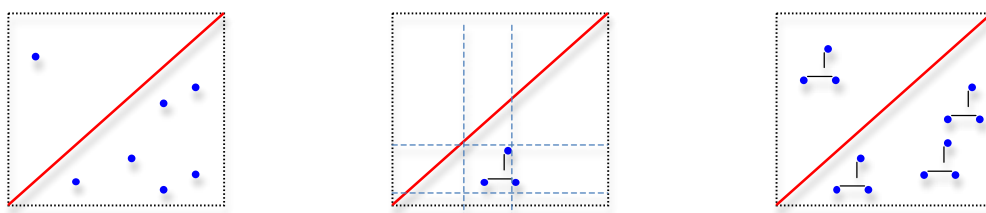
In multiple genomes comparative methods and evolutionary models are used, but the equilibrium of these models, what is what would be relevant in a single sequence, is typically much simpler than the non-evolutionary models will assume.



Left: a set of independent instances of a pattern is given and we are interested in give a probabilistic description of the pattern. If the pattern for instance is 10 bp long, this would mean $4^{10}-1$ (1.048.575) real numbers. Since is it unrealistic to get enough patterns to evaluate so many probabilities independently, models of patterns are of much lower dimensions. A very simple approach is the so called Position Weight Matrix, where each position gets its own probability matrix. In the above example, a reduction to 30 real numbers has been achieved.

Right: The set of patterns are now related by a phylogeny and now the pattern are dependent and an evolution model must be described. Clearly it would be natural to combine the two classes of methods.

The main idea of this project is to try to unify these approaches – ie can we devise evolutionary models that give equilibrium distributions corresponding to complex models. Goldman and Whelan (2002) and Knudsen and Miyamoto (2004) have considered this for simpler models not coupled to machine learning techniques. Halpern and Bruno (1998) developed a method that given a neutral biochemical transition rate matrix and an observed equilibrium frequency distribution post-selection, could evaluate the selection need to create the modified rate matrix. The problems of defining a Q (quadratic in state space size) from a Π (linear in state space size) is sometimes called the shadow problem.



Independent signal observation will assume signals to be distributed uniformly in the positive area and avoiding the negative area. Homologous signals will be related by a phylogeny and thus would typically be found as small clusters and only explore a small part of the P area. However, if an independence assumption is used between position, patterns space will be explored position by position (demonstrated by x and y axis here). Independent signal sampling would intuitively seem to be more efficient in defining P/N areas than homologous signals. However, it is harder to verify that the function – and thus equilibrium distribution - is the same for independent signals. Additionally, independent signal sampling only gives information about the equilibrium distribution, while homologous potentially allows characterisation of the full molecular evolution process. Clearly, the two kinds will yield very different information and would be of interest to combine.

Combining the above classes of models with evolution is not so difficult for the first 2 classes of models. PWM is a description of the equilibrium frequency of each position. HMM is a natural extension of PWM that would necessitate an evolutionary process having neighbor dependencies, which has been investigated exhaustively (Jensen and Pedersen, 2000; Siepel and Haussler, 2003; Hobolth and Jensen, 2005). Choi et al. (2008) has investigated models that resulted in pre-given HMM and variable length Markov Chain distributions. This last paper is much in the spirit of what is suggested here. We basically want to investigate patterns defined by other machine

learning methods as well. One class, BN, as for instance used by Ben-Gal et al. (2005) and Reddy et al. (2007) can most likely be handled by the techniques described by Choi et al. as BN is simply a generalisation of HMMs to allow a more general class of dependencies.

Both NN and SVM are given a set of positive (P) and negative (N) instances, and uses a geometric procedure to carve up the space into a P-area and N-area. They can have a variety of architectures and thus create P/N areas of very different geometries. SVM and NN create sharp boundaries between P and N areas and do not inherently define a distribution. It could be natural to assume a uniform distribution on the P area and zero probability/density on the N area. Such a distribution might be hard to reproduce by a substitution process with superimposed selection. To envision this, imagine point close to the P/N boundaries. They will have less neighbors than the one in the middle of P. It would be natural to create P and N areas by adding a fitness/selection factor to events when you are in these two areas. NN and SVMs might also create non-connected areas that cannot be bridged by an evolutionary process with based on 1-nucleotide substitution events. For instance, if two areas are more than 2 events apart. Any Π will then depend on the initial probabilities of starting in different areas in contrast to irreducible Markov Processes.

Project – there are several challenges. For larger state spaces, there will be much freedom in defining a Q matrix and finding the relevant constraints will be interesting. Defining P that corresponds to given SVM and NN seems to be a new problem and it is difficult to predict, which problems will be encountered.

Plan:

- Read the papers from the reference list.
- Implement basic substitution background model (Jukes-Cantor, Kimura 2P, F81, HKY85).
- Implement the most basic PSM, HMM, BN, SVM and NN algorithm to infer patterns.
- Implement evolutionary process corresponding to background model, but with superimposed selection on P and N areas. Investigate the equilibrium distribution as function of selection strength and distance into P and N areas. Start with patterns defined on very simple state spaces and by simple NN/SVM architectures.

References:

- Ben-Gal et al (2005) Identification of transcription factor binding sites with variable-order Bayesian networks *Bioinformatics* 21(11):2657-2666
- Bishop (2005) *Pattern Recognition and Machine Learning* Springer
- Choi, S.C., Redelings, B.D., and Thorne, J.L. (2008) Basing population genetic inferences and models of molecular evolution upon desired stationary distributions of DNA or protein sequences. *Phil. Trans. R. Soc. B*
- Durbin et al. (1998) *Biological Sequence Analysis* CUP
- Goldman and Whelan (2002) A Novel Use of Equilibrium Frequencies in Models of Sequence Evolution *Mol. Biol. Evol.* 19(11):1821–1831.
- Halpern and Bruno (1998) Evolutionary Distances for Protein-Coding Sequences: Modeling Site-Specific Residue Frequencies *Mol. Biol. Evol.* 15(7):910–917.
- Hobolth et al. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Stat appl Genet Mol Biol*, 4, 1
- Jensen et al. (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution *Adv. Appl. Prob.* 32.2 499-517.
- Knudsen and Miyamoto (2004) Using equilibrium frequencies in models of sequence evolution *BMC Evolutionary Biology* 2005, 5:21
- Lawrence et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262.208-14.
- Reddy et al. (2007) Binding Site Graphs: A New Graph Theoretical Framework for Prediction of TFBS. *PLoS compu biol* 3.5.844-
- Siepel and Haussler (2003) Combining phylogenetic and hidden Markov models in biosequence analysis *Recomb* 277 - 286
- Valen et al. (2009) Discovery of Regulatory Elements is Improved by a Discriminatory Approach *PLoS Comput Biol* 5(11): e1000562
- Wasserman & Sandelin (2004) Applied bioinformatics for the identification of regulatory signals *Nature Reviews Genetics* 5, 276-287
- Wikipedia: Sequence motif, Position-specific scoring matrix, vector support machine, neural networks,
<http://jaspar.genereg.net/>