Fast mixing Monte Carlo for Biological Networks

Supervisors: Sach Mukherjee and Mario Nicodemi.

Recent years have witnessed rapid growth in quantitative, data-driven biology. This in turn has led to a great deal of interest in computational and statistical approaches in biology. The study of biological networks in particular has attracted a great deal of research interest. Networks of components such as genes, proteins and metabolites play a central role in biological function as well as in the molecular biology of diseases such as cancer.

Probabilistic graphical models (Jordan, 2004) have emerged as a key approach for the study of biological networks (see e.g. Friedman, 2004). Graphical models are a class of statistical model in which a graph encodes probabilistic relationships between variables under study. In biological problems, it is very often the case that questions of interest concern the graph structure itself.

A widely-used strategy is to take a search-based approach, using "greedy" algorithms to find a network that is in a certain sense the "best". These methods are appealing to practitioners because they are simple, intuitively appealing and (relatively) fast. However, approaches based on finding a single "best" graph raise serious concerns about "over-fitting" available data. Such concerns can be ameliorated by the use of sampling methods to characterize the posterior distribution $P(G \mid \boldsymbol{X})$ over graphs, given data $\boldsymbol{X}$, and then using that distribution to make probabilistic statements about questions of interest (e.g. whether or not a certain edge is present in the underlying graph).

Recently, Markov chain Monte Carlo methods, based on the well-known Metropolis-Hastings algorithm, have been used to address biological network inference (Werhli et al. 2006; Mukherjee & Speed 2008). These methods use a simple iterative scheme to draw samples from the posterior $P(G \mid \boldsymbol{X})$. The algorithm can be thought of as constructing a Markov chain, whose state space corresponds to the set of all possible graphs $\mathscr{G}$, and whose stationary distribution is the Bayesian posterior $P(G \mid \boldsymbol{X})$. However, despite asymptotic guarantees of convergence, in practice these chains may mix slowly, such that one must generate a large number of samples, at great computational cost, to draw satisfactory inferences. These issues are exacerbated in the context of practical problems in molecular biology, where sparse data and high variability lead to diffuse posteriors, with probability mass dispersed widely in the space $\mathscr{G}$.

The aim of this project will be to explore improved MCMC schemes, exploiting existing work in statistical physics and related literature in statistics, to develop fast-mixing sampling algorithms for biological network inference. A starting point will be the Swendsen-Wang (SW) method (Swendsen & Wang 1987). In statistical physics, SW was originally proposed in the context of ferromagnetic spin models and thereafter extended also to frustrated spin systems (Cataudella et al. 1994). In statistics, SW has been analyzed as a so-called auxiliary variable scheme (Higdon 1998) and more recently applied to variable selection (Nott & Green 2004), a problem similar in spirit to the network inference problem of interest here.

Skills developed: The project will develop your understanding of both statistical simulation methods and Bayesian inference and acquaint you with associated challenges in network biology. An outstanding project may lead to a scientific publication.

Abilities required: You will need to have excellent computational skills, good mathematical ability and at least some understanding of statistical inference (e.g. at the level of Co902). Absolutely no prior knowledge of molecular biology or statistical physics is required.

PhD opportunities: The mini-project could lead very naturally to a PhD project in either statistical physics or computational biology, or at the interface between the two.

(Jordan 2004) M. I. Jordan. Graphical models. Statistical Science, 19:140–155, 2004.

(Friedman 2004) N. Friedman. Inferring Cellular Networks Using Probabilistic Graphical Models. Science, 303(5659):799-805, 2004.

(Werhli 2006) A. V. Werhli et al. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics 22(20):2523:2531, 2006

(Mukherjee & Speed 2008) S. Mukherjee and T. P. Speed. Network inference using informative priors. Proceedings of the National Academy of Sciences 105(38): 14313–14318, 2008

(Swendsen & Wang 1987) R. H. Swendsen and J. S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. Physical Review Letters, 58(2):86-88, 1987.

(Cataudella et al. 1994) V. Cataudella et al. Critical clusters and efficient dynamics for frustrated spin models. Physical Review Letters, 72(10): 1541-1544, 1994.

(Higdon 1998) D. M. Higdon. Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications. Journal of the American Statistical Association, 93(442):585-595, 1998.

(Nott & Green 2004) D. J. Nott and P. J. Green. Variable Selection and the Swendsen-Wang Algorithm. Journal of Computational & Graphical Statistics, 13(1):141-157, 2004.