Structure, Search and Sampling: Computational Learning of Biological Networks

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 embodies precise conditional independence statements regarding variables under study. In biological problems, it is very often the case that questions of interest concern the conditional independence graph itself. However, the problem of making inferences regarding graphical model structure (or "structure learning") is well known to be a daunting one, and for real-world problems inevitably involves computationally-intensive inference.

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 approaches rooted in Bayesian statistics (see e.g. Mukherjee & Speed, 2007). Here, the aim is to characterize the posterior distribution over graphs, using that distribution to make probabilistic statements about questions of interest (e.g. whether or not a certain pathway plays an especially important role in a particular kind of cancer). However, Bayesian inference in this setting inevitably involves resorting to methods for approximate inference (such as Markov chain Monte Carlo) which can be complicated and computationally intensive.

Given the trade-offs involved, and the practical importance of network inference in, for example, systems biology and cancer research (e.g. Mukherjee et al., 2007), it is important to investigate the substantive pros and cons of search- and sampling-based methods in this context. **The aim of this project will be to investigate, primarily by means of computer simulation, the ability of these contrasting approaches to uncover the structure of biological networks.** The project will exploit work described in detail in Mukherjee & Speed (2007) and will benefit from the availability of associated MATLAB code.

Skills developed: The project will develop your understanding of machine learning and computational statistics 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, ideally in MATLAB and at least some understanding of statistical inference, ideally including Bayesian methods and graphical models. Absolutely no prior knowledge of molecular biology is required: indeed, depending on your interests, the project could be carried out with a methodological emphasis and little or no reference to biology.

PhD opportunities: The mini-project could lead very naturally to a PhD project in network inference, either from the point of view of computational statistics or with a focus on biological networks in cancer.

(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.

(Mukherjee & Speed 2007) S. Mukherjee and T. P. Speed. Markov chain Monte Carlo for Structural Inference with Prior Information. Technical report #729, Department of Statistics, U. C. Berkeley, 2007.

(Mukherjee et al. 2007) S. Mukherjee et al. ERK-pathway connectivity is dependent on tumor sub-type. Proceedings of the American Association for Cancer Research (AACR) Annual Meeting 2007, 2007.