High-dimensional genome analysis

Complexity mini-project proposal.
Supervisors: Sach Mukherjee (Netherlands Cancer Institute, Amsterdam) & Tom Nichols
(Stats)

Hidden Markov Models (HMMs; see e.g. [1] for an overview) are widely used to model time-varying data to discover underlying changes in system state. Observations X_t are conditional on the state of an underlying Markov chain: the state of the chain is treated as a latent variable that is not itself observed (hence “hidden”). The case where X_t is scalar or low-dimensional is well studied [1]. However, in the high-dimensional setting, inference for these models remains challenging, moreso when (state-specific) network or covariance structure is of interest.

This project aims to exploit recent advances in high-dimensional inference for networks [2] and extensions to latent variable settings [3] to develop high-dimensional HMMs. The specific application that motivates the work and will provide data for the project is in genome analysis. Proteins physically bind to DNA and in this way influence biological function. It is now possible to experimentally obtain binding profiles for multiple proteins across the whole genome (i.e. at every location t along the DNA), see e.g. [4]. When the number of proteins is large, the resulting vector X_t of binding profiles is high-dimensional. Such data allow new and fundamental questions to be asked: Does protein-protein interplay depend on genomic region? Does the genome have regions that show different protein-DNA binding patterns in a multivariate sense? This project is motivated by precisely these kinds of questions. Protein-DNA binding data for 50-100 proteins across the entire *Drosophila* genome will be used to test new methods and investigate biological questions. The project could focus more on methodology or application, depending on the student’s interests. *The project will not involve any “wet” experimental work and does not require any prior knowledge of biology.*

The project will be mainly based at the Netherlands Cancer Institute (NKI) in Amsterdam. We are a world-leading biological research institute, and offer a stimulating scientific environment, with expertise in diverse areas of biology. The Mukherjee Lab in particular focuses on statistical and computational approaches in molecular and cancer biology. The project could be extended into a PhD, offering the opportunity to spend some time in Amsterdam, and the potential to exploit new ideas in inference to address some of the most fundamental questions in biology.

References

[1] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257-286, 1989

[2] J. Friedman *et al.* Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441, 2008

[3] S. Mukherjee & S. M. Hill. Network clustering: probing biological heterogeneity by sparse graphical models. *Bioinformatics* 27(7):994-1000, 2011

[4] G. J. Filion *et al.* Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in *Drosophila* Cells. *Cell* 143(2):212-224, 2010