Association Analysis using Wavelets Hidden Markov Trees

Or Zuk

Modern high throughput sequencing technologies (e.g. chip-seq) enable the measurement of molecular phenotypes (e.g. binding of proteins to DNA) at single-basepair high genomic resolution. Data for each sample is typically represented as a vector of read counts at adjacent genomic positions. and may be treated as a noisy measurement of an underlying (one-dimensional) function of genomic location. Motivated by this interpretation, we develop methods for estimation of such functions for a single sample, as well as estimating and testing for differences between groups of samples, which enables association analysis of the functional data with scalar variables representing each sample (for example, genotype, age etc.).

We use a wavelet-based method to get a multi-resolution analysis of the underlying data and allowing, in principle, detection of signals up to a single base resolution, in contrast to fixed window-based approaches. We unify previous ideas proposed in different contexts such as simultaneous analysis of multiple samples, using a multi-scale Poisson model, and Hidden Markov modelling of the wavelets coefficients into a single framework. We get improved performance compared to fixed window-based approach and to a simpler, previously proposed wavelet-based approach, for both detection of differences, as well as for estimation and localization of effect size.

Joint work with H. Shim and M. Stephens