

BEYOND HMMS: A HIERARCHICAL MODEL FOR INHOMOGENEOUS TILED ARRAY OBSERVATIONS;

Lian H¹, Noble, WS^{2,3}, Thompson, W¹, Stamatoyannopoulos, J², and Lawrence CE¹. (1) Center for Computational Molecular Biology and Division of Applied Mathematics, Brown University, (2) Departments of Genome Sciences and (3) Computer Science and Engineering, U. of Washington

Hidden Markov models offer an appealing technology for the analysis of dye intensities from tiled arrays. Applications of this technology using multi-state HMMs with each state described by a Gaussian model have already shown themselves to be very useful. However, inhomogeneity in the dye intensities of probes can lead to poor fits of these models. When sufficient array data are available probe-specific models of intensities provide a useful means to address probe-to-probe variation. Here we described an alternate procedure to address these inhomogeneities using hierarchical change point models that can be applied even when no replicates are available. This model seeks to capitalize on the premise that sonicated fragments that span multiple probes induce similar intensities across substrings of tiled array probes. This change point model assumes that within-state sequences are not homogeneous, but instead are characterized by substrings of unknown length each of which is homogeneous. All substring models within each state are drawn from a continuous mixture of Gaussians. Hierarchical models of the means and variances within each state capture the similarities of the observations of a given state, and provide for differences among the states. The analysis of residuals from an application to data from high throughput studies of DNase I hypersensitivity shows that the poor fit of Gaussian models is addressed well by these hierarchical models. Preliminary findings on three states of DNase I hypersensitivity are given. The model also yields a principled means to span the missing data of repeat masked gaps that goes beyond the implicit geometric length distributions of HMMs. Marginalization over substring end points using recursions and integrals over substring means and variances greatly reduce dimensionality and enable empirical Bayesian inferences.