A model of signal shift and drift for tiling array expression data

Pierre Nicolas[1], Aurélie Leduc[1], Stéphane Robin[2], and Philippe Bessières[1]

March 20, 2009

[1]INRA, Mathématique Informatique et Génome UR1077, F-78350 Jouy-en-Josas, France
[2]AgroParisTech/INRA, Mathématiques et Informatique Appliquées UMR 518, 16 rue Claude Bernard, F-75005 Paris, France

The tiling design for oligonucleotide microarrays consists of overlapping probes that provide uniform covering of the genomic sequence. Their hybridization with RNA samples, allow to assess the transcriptional activity of the whole genome of organisms such as bacteria and yeasts with high resolution.

The problem of the analysis of these data is naturally stated in terms of finding segments where the hybridization signal is relatively constant, delimited by breakpoints that are expected to correspond to biological features such as promoters, terminators or splicing sites. A variety of tools including local non-parametric smoothing and simple iterative hypothesis testing have been proposed to answer this question. Today the most popular and best statistically grounded model is the piecewise constant model with Gaussian noise. The simplicity of this approach is appealing but its use presents a number of specific difficulties, the two most obvious being the choice of the number of segments and the high time complexity of the algorithm.

In principle, embedding the segmentation model in a probabilistic setting that includes not only the noise but also the evolution of the signal can alleviate the need for the choice of a fixed number of breakpoints. In this context the problem states as the estimation of a parameter and the reconstruction of the underlying signal trajectory can integrate the uncertainty on the exact number of breakpoints. This idea stimulated the development of Hidden Markov models (HMMs). However, transcript level is a continuous quantity and none of the proposed models is satisfactory when the underlying signal is continuous. A HMM achieving this aim at a computationally affordable cost will be presented here.

The proposed model is also markedly richer than the piecewise constant model. First, it automatically accounts for differential affinity between probes via the introduction of covariates. This allows to achieve segmentation and within-array normalization in one step. Second, our model also relaxes the assumption of strictly constant underlying signal between abrupt “shifts” by also allowing progressive “drift”.