Poster L-32

Gene Expression Trees in Blood Cell Development

Authors:
Ivan G. Costa Filho (*Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics*)
Stefan Roepcke (*Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics*)
Alexander Schliep (*Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics*)

Short Abstract: We present a statistical framework -- mixtures of conditional probabilistic trees - designed to cluster gene expression in the course of cell development. Additionally, we include microRNA target prediction in our framework. We recover well-known biological facts and also identify novel putative but convincing egulatory and functional assignments of microRNAs.

Long Abstract:

The regulatory processes that govern cell proliferation and differentiation are central to developmental biology. Particularly well studied in this respect is the hematopoietic system. Gene expression data of cells of various distinguishable developmental stages fosters the elucidation of the underlying molecular processes, which change gradually over time and lock cells in certain lineages. Large-scale analysis of this data requires a computational framework for tasks ranging from visualization, querying, and finding clusters of similar genes, to answering detailed questions about the functional roles of individual genes and their similarities and differences.

In development, we have temporal sequences, in which one cell type undergoes a given differentiation process and turns into a new cell type. Furthermore, there can be branching points at which cells at particular developmental stages can differentiate in two or more cell types, and follow different differentiation paths from this point on. These latter processes are represented by a tree. Second, we know that the complexity of an organism is better reflected by the combinatorial nature of regulatory molecules then by the size of its genome. This is supported by experimental results that indicate that a single gene often participates in several differentiation processes, performing distinct functions in distinct biological conditions, hence participating in a variety of molecular mechanisms. Consequently, gene expression patterns cluster into overlapping groups.

To address this biological reality, we choose to represent mRNA expression data during cell development and differentiation with conditional probabilistic trees, which reflect dependencies during differentiation, and to combine several of these models in a mixture. Together, we obtain a robust and flexible statistical model for analyzing and clustering genome-wide mRNA expression data sets, in which the inherent dependencies between stages can be seen and overlapping clusters are allowed. For this, we extend conditional trees, originally proposed for discrete variate, for continuous variates. Additionally, we

combine sequence information and genes with similar expression profiles to perform microRNA target prediction.

Our work concentrates on two detailed studies covering several stages of the B- and T-cell development and a tree containing three lineages of lymphoid cells. Furthermore, we also use simulated data sets in order to evaluate the method. The results with simulated data shows the superiority of our method, in relation to k-means, for finding clustering of genes profiles with tree dependencies. For the biological data, we recover well-known biological facts and also identify putative but convincing regulatory elements, genes and functional assignments. In special, our results suggests that some microRNAs, which have been previously related to hematopoiesis, have a regulatory role in reducing the transcript levels of genes that are important for cell proliferation.