Chapter 7

Discussion

Clustering is a crucial first step in the analysis of large-scale gene expression experiments. Peculiarities of gene expression data from microarray experiments, require the development of novel clustering methods. While mixture models provide a statistical framework to perform clustering, the specification of proper component density functions, which take characteristics inherent to the multi-variate data at hand, remains an open problem. In this thesis, we propose two novel component models for analyzing gene expression measured over time or developmental processes. Furthermore, we approach the problem of integrating additional sources of biological data to enhance the analysis of gene expression. This is done by using a semi-supervised method for estimating the mixture model.

In the next sections, we present the final remarks and future work of each specific contribution of this thesis.

Mixture Models and Cluster Validation

We introduce, in Chapter 3, a novel validation index for comparing overlapping partitions obtained by mixture model based clustering algorithms. This index is an extension of the well known corrected Rand (CR). In the context of mixture models, our experimental work shows that the extended corrected Rand index yields significant improvements when compared to the results obtained by the traditional corrected Rand. Finally, it is important to point out that there are still many theoretical and practical aspects of cluster validation in the context of mixture models. The definition of the extended Corrected Rand represents an initial contribution to these problems.

Analysis of Gene Expression Time Courses

We present in Chapter 4 an application of mixture models and linear HMMs for the analysis of gene expression time course data. We take advantage of several characteristics of this robust statistical model, which is of great value in the analysis of gene expression time course data. With a benchmark data set, we show that mixture of HMMs have better class recovery than model-based clustering methods with splines or autoregressive models as

components. We also evaluate different methods for model initialization. In this context, a Bayesian approach exploring the linear topology of the HMMs obtained the best practical results. Moreover, we show that the Viterbi decomposition is able to enhance the mixture of HMMs for the yeast cell cycle data set. In an anedoctal analysis with HeLa cell data, we also show that the Viterbi decomposition refines clusters in a biological meaningful way. The use of the entropy threshold for discarding ambiguous cluster assignments improve the specificity of Gene Ontology annotation, which reassures the usefulness of the soft assignments of the mixtures in detecting unambiguous clusters. Our flexible framework, combined with an effective graphical user interface implemented in the GQL application, supports interactive and exploratory knowledge discovery of gene expression time course data.

There are several aspects still to be explored in the use of linear HMMs and mixtures of linear HMMs. First, the use of other pdfs as emission functions such as the Gamma pdf, as explored in [235], could produce better results for gene expression data. One issue that have been recently explored is the fact that most data sets have few time points [73]. This can be addressed by biasing the topology learning method towards models with fewer number of states. Furthermore, an extension of our framework to perform simultaneous topology learning and mixture estimation should also improve the performance of the mixture of linear HMMs. Recently, a great deal of data sets with multiple time courses of a given species have been made available. These data sets present time course measurements over distinct gene knockouts [236], environmental conditions [82, 174] or patients [225]. In fact, such data sets pose new methodological questions about conditions at which genes are differentially expressed, or what are the temporal dynamics of these differences? For such tasks, we could extend the linear HMM to multiple linear HMM models. Then, we could apply structural learning methods to explore issues concerning detection of time-lag relationships, temporal dynamics of these differences, and finding groups displaying similar differential expression profiles.

Analysis of Gene Expression in Lymphoid Development

The regulatory processes underlying cell proliferation and differentiation are of central interest to developmental biologists and clinicians. They are frequently the focus of large-scale studies in which gene expression along paths of differentiation are investigated. To make use of these data in a principled manner, as the main contribution of this thesis, we presented in Chapter 5 a novel statistical framework, called `DTrees`, that models gene expression in the course of development. By combining `DTrees` in a mixture model (`MixDTrees`), we facilitate interactive querying and visualization of data and, more importantly, the detection of clusters of co-expressed genes, which provide a basis for the identification of key players in the regulatory mechanism and their mode of action.

In particular, with `MixDTrees` with structure set to the developmental tree as provided by biologists (`MixDTree-Dev`), we detect groups of genes not found by classical clustering

methods such as Self-organizing maps (SOM). By incorporating microRNA binding data, we show how to identify complex regulatory relationships. In comparison to an analysis based only on sequence data, we predict a manageable number of plausible microRNA targets [91]. Moreover, by the inspection of the developmental profiles of gene targets associated with microRNAs, our method offers some insights into the biological role of the predicted microRNAs.

We show that the `DTree` inferred from the complete Lymphoid data set approximates the dependencies intrinsic to Lymphoid development well. Furthermore, by combining the methods for mixture estimation and for the inference of the `DTree` structure, we find `DTrees` structures specific to groups of co-regulated genes. These groups display different differentiation pathways reflected by the distinct estimated dependence structures. Furthermore, groups have a lineage specific expression pattern. Enrichment analyses of gene annotation using KEGG and GO indicate development-specific function of the groups found.

For simulated data, `MixDTrees` compares favorably to other methods widely used for finding groups of co-expressed genes, even for data arising from variable dependence structures. In particular, our method is not susceptible to over-fitting, which is otherwise a frequent problem in the estimation of mixture models from sparse data.

Interesting extensions to our analysis are possible, even when one only considers gene expression data and the basic method. None of the currently publicly available data sets offer both a tree with a large number of branches and a detailed view of all development stages. An interesting compendia of gene expression data from lymphoid cells [105], concentrates on mature and immature cells in final development stages. The creation of an expression compendium such as the one in [105], where many intermediary stages of differentiation of the developmental tree are present, will be of great value as computational methods can exploit characteristics intrinsic to cell development.

It is also important to point out that developmental biologists are still redrawing developmental trees with the discovery of new intermediary stages and "alternative" paths of development [25, 140, 177]; a particular developmental stage might also be formed by a mixture of distinct cell types not yet well-characterized. An example of an alternative path is the fact that DN1 T cells can be originated not only from the lymphoid progenitor as depicted in Figure 5.1, but also from the earlier multi-potent progenitor cells [25]. It is an interesting prospect to extend the structure estimation approach to infer confidence values for branches and stages of a developmental tree from gene expression; as well as to estimate graphs of arbitrary structures. The estimation of graphs with arbitrary structures has already been explored. For example, see [40, 213] for approaches based on graphical models and [182] for an approach based on estimation of covariance matrices. However, in contrast to the method used for the estimation of `DTree` structures, those methods do not provide an efficient and exact solutions for inferring dependence graphs.

Clustering with Constraints for Integration of Heterogeneous Biological Data

If high-quality secondary data is available, semi-supervised learning is an effective framework for the analysis of heterogeneous data as previous experiments using class labels demonstrate [185, 187]. In our experiments based on yeast cell cycle data set (Chapter 6), we use biological information routinely used to support cluster validity as secondary data, i.e., GO annotation and location analysis. Surprisingly, this data can deteriorate cluster quality drastically, if parameters are not chosen properly. Furthermore, we can show that the addition of noise can drastically reduce the performance of clustering with constraints. Although there are other parameter choices to explore, further theoretical questions to address, and more data sets to perform experiments, a main point of our analysis remains valid and clear: secondary data can have little power for clustering, unless it is of good quality, free of errors and have no ambiguities.

These issues discussed in the previous paragraph indicate the need for methodological improvements in clustering with constraints methods. One possible solution for the selection of parameters is the use of cross-validation procedures, as suggested in [45]. Moreover, the inclusion of a step to evaluate the constraint "quality" during clustering execution can be an interesting strategy for preventing problems related to noise in the constraints.

For the Drosophila development case study, we show how to automatically fuse temporal and spatial gene expression patterns by clustering with few high quality constraints derived from in-situ data. Our results demonstrate that the clusters found are biologically meaningful and that we can improve the detection of syn-expressed genes. In particular, the cluster results, obtained after applying the constraints, are better at recovering the functional annotation of ImaGO terms than the clustering solution without constraints. Inferred groups are worthwhile targets for further investigation, either with classical biological analysis or as input for methods to infer gene networks.

There are several open questions regarding the detection of syn-expressed genes. One direction is to improve the image processing pipeline by, for example, using higher quality images, such as 3D models from [96, 117] or images with sub-cellular localization [127]. In relation to the constraints it would be desirable to model the temporal nature of the constraints derived from the in-situ images. A quite challenging problem is to combine an automatic image annotation of expressed cellular compartments with a tree describing the Drosophila development. This would allow us to obtain a detailed developmental profile of genes for this complex multicellular organism, i.e., at which tissues a particular gene is expressed. Hence, we could use Mixture of Dependence Trees, as proposed in Chapter 5, to analyze gene expression of Drosophila development.