Chapter 7

Conclusion

Biomedical research changed tremendously during the last decades, with the emergence of biotechnologies that allow simultaneous measurements of thousands of DNA, RNA or protein sequences. Microarray and next generation sequencing generate substantial amounts of data, which may help biomedical researchers to understand the complex genetic mechanisms underlying carcinogenesis. High-dimensional data both from static and temporal experiments are generated in numerous studies to answer diverse biological questions. Here, we focus on an experiment where data are collected over time at different molecular levels, allowing the study of dynamic behavior in cells during malignant transformation. Analysis of such a complex dataset needed novel statistical methodology for integrative analysis of these platforms. Integration of data from multiple molecular levels yields a comprehensive model of carcinogenesis, allowing us to uncover meaningful relationships that will improve our understanding of cancer. The study presented in the first part (Chapter 2) of this thesis was focused on the integrative evaluation of temporal differential gene expression analysis, particularly dealing with integration of DNA copy number and gene expression data. In the second part (Chapter 3 and 4) we deal with the problem of gene regulatory network reconstruction in time-course single and multiple omics data. In the final part (Chapter 5 and 6) of the thesis the validity of the developed methodology is investigated by functional validation of the results obtained by applying the developed methodology on the data from our experiment

In our study we carefully chose the time points analyzed so that they would represent distinct phenotypic stages during HPV-induced transformation. In general, longitudinal experiments should be designed in such a way that the chosen samples are representative of all expected steps that together form the process that is being investigated. In addition, time course experiments studying the effects of certain

treatments or genetic manipulations of cells need to include time points allowing the investigator to capture early and late effects of the intervention.

Presented methodology for the analysis of data from HPV-induced transformation cell line experiment may be further extended. For `tigaR` we see several ways for further extensions both in terms of the application and methodology.

- Temporal differential expression analysis of RNA-seq count data as illustrated in Chapter 2 can be applied to other temporal high-dimensional count data, such as proteomics or metabolomics data. In addition, using the developed framework, temporal differential expression analysis on count data my be straightforwardly extended to include a time-varying covariate like DNA copy number
- As shown in Chapter 5, `tigaR` analysis can also be used to investigate miRNA-mRNA interactions over time. Currently in literature methods for miRNA target prediction are based on static experiments and are known to yield may false positive results (see [133]). Further extensions of our method may address this problem by inclusion of the temporal fluctuation in gene expression when identifying miRNA-mRNA associations. Selection of the miRNA targets can be improved by including prior information from computational target prediction data bases (see [175]).

With respect to gene regulatory network reconstruction in time-course single and multiple omics data, inclusion of prior information from steady-state gene expression measurements may improve gene-gene network reconstruction substantially. In the work of [198], they addressed this problem which showed improvements in the reconstruction of the interactions in the gene expression data. Although they use a regression-based algorithm in order to combine steady-state and time-series datasets to infer gene interaction networks, this method can be further improved. Borrowing prior information from one data set allows to model the parameters of the prior and to improve estimation of the conditional independence graph.

Another extension from the methodological perspective may address reconstruction of the time-series chain graph associated with a vector autoregressive process from RNA-seq data. Next generation sequencing rapidly replaces microarrays for genome and transcriptome profiling. Sequencing technology can identify and measure the transcripts which have not been previously annotated, offering more precise measurements, especially at the lower end of the spectrum. One of the differences between microarray and next generation sequencing techniques is in the data type generated. Sequencing data are not intensities, but counts. This type of data does not allow application of Gaussian graphical models, assuming multivariate normal

distribution, to reconstruct gene regulatory networks. In literature there are several ways to overcome this problem. Several methods are developed which first Gaussianize the data, with copula transformation (see [41, 95]), or by using non-parametric rank-based estimators (see [94]). Next generation sequencing data are also modeled employing Markov network estimation for count data, as well as local Poisson graphical models [5], which are only computationally feasible for a small number of variables. All these methods either do not suit well to the high-dimensional count of next generation sequencing data or cannot computationally deal even with medium sized networks. Thus, further improvements in statistical methodology are required.

Although currently time course experiments and corresponding molecular datasets are rare, this will change in the near future with the introduction of so-called liquid biopsies. Liquid biopsies usually refer to blood, but could also include other bodily fluids like urine or sputum. As these fluids were shown to contain circulating tumor cells and/or cell-free DNA originating from tumor cells in many cancer patients, they can be used to guide treatment decisions as they offer a unique opportunity for real-time monitoring of the disease during treatment (see [4]). One of the major challenges in the analysis of liquid biopsies however is the relative rarity of tumor-derived material in an enormous background of material derived from non-diseased tissue. To tackle this challenge one can either enrich tumor-derived material from the liquid biopsy for example by isolation of circulating tumor cells or one can adapt the statistical analysis methods. An example of the latter is provided by the use of auxiliary co-data to improve the prediction by molecular signatures [122]. Following this reasoning our `tigaR` approach combined with static datasets on pure cancer material might make this approach suitable for temporal analysis of liquid biopsies as well, thereby greatly increasing its applicability.

In conclusion, this thesis has laid the groundwork for integrative temporal differential expression analysis and temporal gene regulatory network reconstruction. The validity and applicability of our approaches are illustrated by deciphering the molecular events driving HPV-induced transformation in a cell line model. Further extensions of the framework provided here should enable applicability of the methods to RNA-seq data. `tigaR` may also prove useful in the identification of miRNA targets, an upcoming area in cancer research, and the temporal analysis of liquid biopsies that are expected to become the mainstay for monitoring and treatment decision-guidance especially in patients with metastatic cancer.