Chapter 5

Summary and outlook

Genome-scale gene silencing screens pose novel challenges to computational biology. At present, RNA interference appears to be the most efficient technology for producing large-scale gene intervention data. This dissertation developed methodology to tackle two problems peculiar to gene silencing data:

1. Gene perturbation effects cannot be controlled deterministically and have to be modeled stochastically. The uncertainty of intervention effects in a noisy environment is modeled by choosing informative prior distributions for the relationship between regulators and their targets. We formalize this approach in the general framework of conditional Gaussian networks in chapter 3.
2. Direct observations of intervention effects on other pathway components are often not available. Large-scale datasets may only contain observations of secondary downstream effects. Learning from secondary effects is implemented via a two-leveled model of an unobserved pathway with observable downstream reporters. In chapter 4 we develop a Bayesian scoring function to evaluate models with respect to data.

Each of these two problems becomes aparent in different modeling situations. Accounting for stochasticity of interventions is of special importance when reconstructing transcriptional regulatory networks from microarray data. In this setting we assume that expression states of gene coding for transcription factors are good approximations of the activation state of the trascription factor protein. Under this assumption, the correlation structure of genes in different conditions allows conclusions about transcriptional regulators and their targets. Silencing a gene leads to primary effects at other genes in the model and increases the accuracy of network reconstruction.

The second challenge is learning from indirect information and secondary effects. This becomes important when inferring signal transduction pathways from phenotypical changes after interventions. In the cell, a signal is propagated on protein level and mRNA concentrations mostly stay constant for all pathway components. Thus, interventions do not lead to primary effects observable at other pathway components. Instead, reflections of signaling activity perceived in expression levels of downstream genes after pathway perturbations can be used to reconstruct non-transcriptional

features of signaling pathways. Single reporter genes below the pathway of interest can be used as transcriptional phenotypes. Subset patterns on observed phenotype changes allow inference of regulatory hierarchies. In simulation studies we confirmed small sample size requirements and high reconstruction accuracy for the Bayesian score devised to evaluate candidate models. The usefulness of our approach on real data was shown by analyzing a study of *Drosophila* innate immune response.

Non-transcriptional phenotypes In chapter 4 we used reporter genes downstream the pathway of interest to reconstructed a regulatory hierarchy. Expression changes of reporter genes can be interpreted as transcriptional phenotypes. In fact, any other kind of binary phenotype could also be used in our analysis. The only requirement is that the number of phenotypes is large enough and contains a meaningful subset structure. We plan to extend our approach to data from large-scale screens in *C. elegans* [102, 52]. Phenotypes measured there include "no developing embryos seen 48 hours after dsRNA injection", "Reduced fecundity of injected worm", "osmotically pressure sensitivity", or "multiple cavities". Until now, genes in the *C. elegans* genome have only been clustered according to phenotype similarities [53]. Elucidating regulatory hierarchies remains an open question.

Scaling up model size In its present form, the algorithm proposed in chapter 4 can be applied to filter (several thousands of) pathway hypotheses and to find those well supported by experimental data. The hypotheses build on existing biological expertise. This constrained search space can be interpreted as the result of a rigid structure prior focussing on biological relevant hypotheses. To apply our method to large-scale intervention data with thousands of silenced genes and little biological prior knowledge, model search will have to be improved. There seem to be two promising avenues for further research. One could combine optimal subnetworks to big networks, as it is done in quartett-puzzling algorithms in phylogeny [132]. Another strategy is to define a neighborhood relation on the set of silencing schemes and use techniques of combinatorial optimization to explore the score landscape. The contribution of this thesis is a scoring function to link data with models. Efficient search heuristics are the topic of future research.

The need for a holistic view The internal organization of the cell comprises many layers. The *genome* refers to the collection of information stored in the DNA, while the *transcriptome* includes all gene transcripts. On the next level the *proteome* covers the set of all proteins. The *metabolome* contains small molecules—sugars, salts, amino acids, and nucleotides—that participate in metabolic reactions required for the maintenance and normal function of a cell. Results of internal reactions are features of the cell like growth or viability, which can be taken as *phenotypes* to study gene function. To understand the complexity of living cells future research will need to build models including all these layers. Statistical inference on parts of the system will not provide the mechanistic insights functional genomics is seeking for. Recent research concentrates on combining information from genome, transcriptome and proteome, *e.g.* by building models jointly on expression and protein-DNA binding data. This is a necessary step into the right direction. However, these models will still

be fragmentary if they not include (and predict) phenotypical changes of interventions into the normal course of action in the cell. We will only understand what we can break.

It's the biology, stupid! This thesis explored how to recover features of cellular pathways from gene expression data. All in all, this thesis shows: pathway reconstruction is not an issue of more advanced models and more sophisticated inference techniques. Pathway reconstruction is a matter of careful experimental planning and design. Well designed experiments focus on a pathway of interest and probe information flow by interventions. Only a small sample size and simple statistics are then needed to extract the relevant information from data.