Chapter 1

Introduction

This thesis is concerned with signaling pathways leading to regulation of gene expression. I develop methodology to address two problems specific to gene silencing experiments: First, gene perturbation effects cannot be controlled deterministically and have to be modeled stochastically. Second, direct observations of intervention effects on other pathway components are often not available. This first chapter gives a concise background on gene regulation and cell signaling and explains the experimental technique of RNA interference (RNAi). Gene silencing by RNAi has drastically reduced the time required for genome-wide screens for gene function, but no work has been done so far to adapt statistical methodology to the specific needs of RNAi data.

1.1 Signal transduction and gene regulation

The success of genome sequencing projects has led to the identification of almost all the genes responsible for the biological complexity of several organisms. The next important task is to assign a function to each of these genes. Genes do not work in an isolated way. They are connected in highly structured networks of information flow through the cell. Inference of such cellular networks is the main topic of this thesis.

Eukaryotic cells Eukaryotes are organisms with cells containing nuclei, in which the genetic material is organized. Eukaryotes comprise multicellular animals, plants, and fungi as well as unicellular organisms. In contrast, *prokaryotes*, such as bacteria, lack nuclei and other complex cell structures. All cells have a membrane, which envelopes the cell and separates its interior from its environment. Inside the membrane, the salty *cytoplasm* takes up most of the cell volume. The most prominent structure inside the eukaryotic cell is the *nucleus* containing *DNA*, the carrier of genetic information. Deoxyribonucleic acid (DNA) is a *double-helix* formed by two anti-parallel complementary strands composed of the *nucleotides* adenine, guanine, cytosine, and thymine. The double-helix is packaged into a highly organized and compact nucleoprotein structure called *chromatin*. The fundamental dogma of molecular biology is that DNA produces *ribonucleic acid* (RNA) which in turn produces *proteins*. The functional units in the DNA that code for RNA or proteins are called *genes*. The

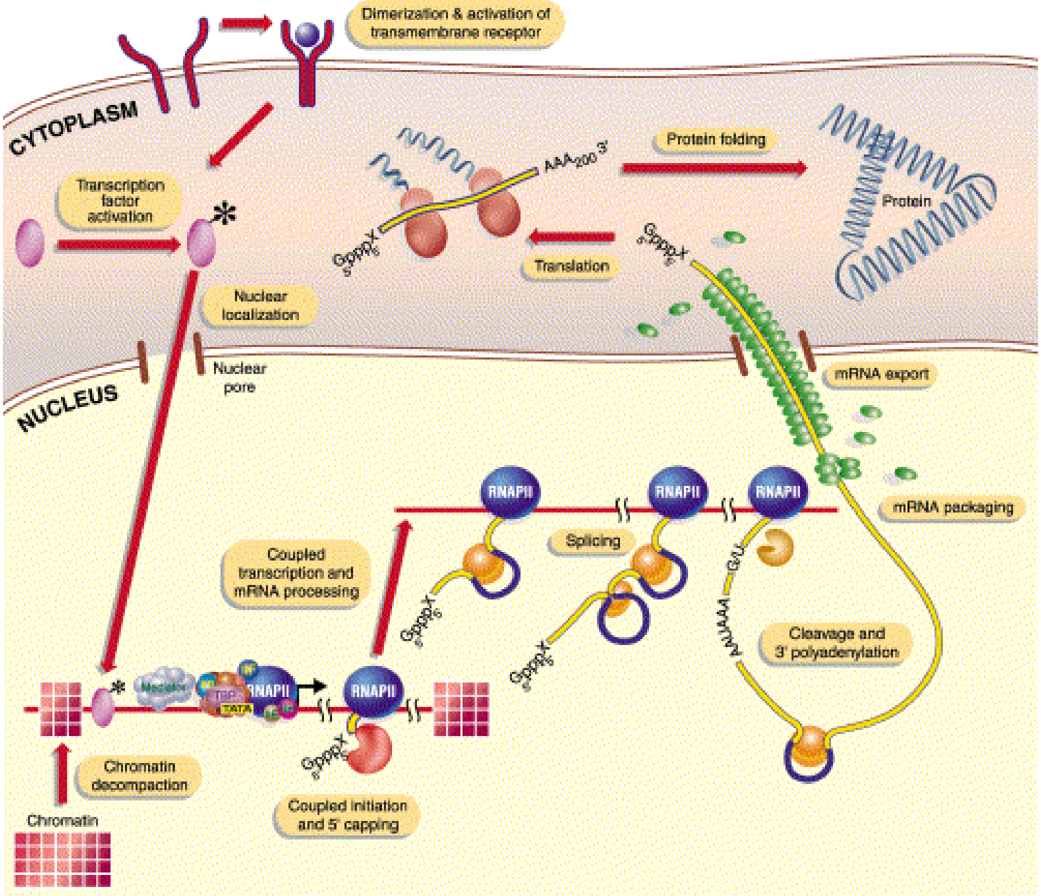

Figure 1.1: *Gene expression in a nutshell. A protein is produced in response to an external signal. See text for details. Reproduced from [94].*

DNA is the same in all cells, but the amount of gene products is not. The diversity of cell types and tissues in complex organisms like humans results from different genes being active.

Gene activity Gene expression is a highly regulated process by which a cell can answer to external signals and adapt to changes in the environment. Fig. 1.1 shows the basic principles of gene expression in eukaryotic cells. In the upper left part of the figure, a signal reaches the cell membrane and is recognized by a *transmembrane receptor*. Binding of a ligand to a receptor initiates an intracellular response. In this way receptors play a unique and important role in cellular communication and signal transduction. In our example, the signal activates a *transcription factor* protein in the cytoplasm. The activated transcription factor enters the cell nucleus and acts on the *promoter region* of a gene in the genome. The promoter region contains the information to turn the gene on or off. Depending on its function the bound transcription factor activates or inhibits gene expression. In the case of an activator, a process called *transcription* is started. A protein called *RNA polymerase II* (RNAP II) starts to copy the information contained in the gene into *messenger RNA* (mRNA).

The nuclear mRNA contains two kinds of regions: *exons*, which are exported from the nucleus as part of the mature mRNA, and *introns*, which are removed from the mature mRNA by a process called *splicing*. The spliced mRNA is transported from the nucleus into the cytoplasm. There it is *translated* into a protein poly-peptide sequence, which then folds into a three-dimensional protein structure.

Fig. 1.1 depicts the expression of a single gene and does not show the influence of other genes and proteins on the expression state. Regulation takes place at all levels, *e.g.*, in signal propagation, in transcription, in translation, and in protein degradation. At each single step many regulatory processes can concur. A transcription factor, for example, can be regulated transcriptionally and non-transcriptionally. Transcriptional regulation means control of the transcription factor mRNA level. Non-transcriptional regulation means controlling the activity level of the transcription factor protein by binding to a ligand, by dissociation of an inhibitor protein, by a protein modification like phosphorylation, or by cleavage of a larger precursor [71]. Of particular interest for this thesis are *transcriptional regulatory networks* and *signal transduction pathways*.

Transcriptional regulatory networks The process described in Fig. 1.1 can be iterated if the protein produced is again a transcription factor, which enters the nucleus and starts to activate or inhibit gene expression of other genes in the genome. Networks of transcription factors and their targets, which again could be transcription factors, are called *transcriptional regulatory networks* or *gene regulatory networks*. Reconstruction of regulatory networks is a prospering field in bioinformatics. This is mainly due to the availability of genome-wide measurements of gene-expression by microarrays, which provide a bird's eye view on gene activity in the cell and promise new insights into regulatory relationships [95, 118, 41].

Signal transduction pathways The second important process is indicated by a single arrow in the upper left corner of Fig. 1.1 leading from the receptor to the activation of a transcription factor. This arrow represents complex biochemical signal transduction pathways, which connect external signals to a transcriptional response. The main steps in signal propagation are protein interactions and modifications that do not act on a transcriptional level. We will explain essential parts of signaling pathways by the example of the *immune deficiency pathway* (Imd), which governs defense reactions against Gram-negative bacteria in *Drosophila melanogaster*. It is related to the mammalian tumor necrosis factor signaling pathways, as it uses structurally and functionally similar components [59]. The Imd pathway will play a central role in the application of the methodology developed in this thesis to a study of *Drosophila* immune response in chapter 4. Fig. 1.2 shows a schematic sketch of this pathway [111].

Immune induction of genes encoding antibacterial peptides like *Diptericins* relies on a transcription factor called Relish. In its inactive state Relish carries inhibitory sequences in the form of several ankyrin repeat domains. To activate Relish, it has to be phosphorylated and then cleaved from these inhibitory domains. Here we see a clear difference to gene regulatory networks. Relish is not regulated on a transcriptional level, it just changes from an inactive into an active form, while the total amount of protein stays the same. This principle is often found in biology and ensures a quick response of the cell to a stimulus. Many pathway components mediating between the receptor at the cell membrane and activation of Relish are known. The phosphorylation of Relish before proteolytic cleavage is mediated by the IKK complex, which can directly phosphorylate Relish *in vitro*. TAK1 is a candidate for activation of the signalosome-equivalent IKKβ-IKKγ. IMD is a partner of an extensive receptor-adaptor complex, which detects infection by Gram-negative pathogens [111]. However, the precise roles of pathway components are often unknown and the object of intense research at present. Fig 1.2 also shows that signaling cascades form cycles and forks, and that different pathways may be connected by sharing components. Boutros *et al.* [12] found a fork in the signaling pathway below TAK1 leading to a Relish-independent response of cytoskeletal regulators via the JNK-pathway.

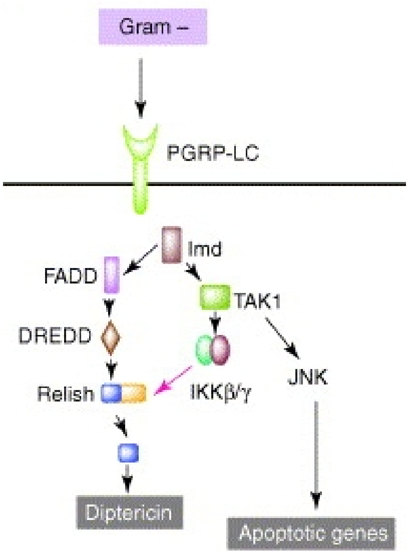

Figure 1.2: *The Imd pathway in Drosophila. Reproduced from [111]*

Cellular signaling pathways regulate essential processes in living cells. In many cases, alterations of these molecular mechanisms cause serious diseases including cancer. Understanding the organization of signaling pathways is hence a principal problem in modern biology. The next section describes RNA interference, which can be used in genome-wide screens to identify new pathway components and to order pathways in regulatory hierarchies.

1.2 Gene silencing by RNA interference

Physicist Richard Feynman once said: “What I cannot create, I do not understand”. This quote stresses the importance of action for understanding. A complex system is not understood solely by passive contemplation, it needs active manipulation by the researcher. In biology this fact is long known. Functional genomics has a long tradition of inferring the inner working of a cell—by breaking it. “What I cannot break, I do not understand” is the credo of functional genomics research.

Until recently external interventions have been labor intensive and time consuming. With methods making use of RNA interference (RNAi), this situation has changed. RNAi [38] is a cellular mechanism of post-transcriptional gene silencing. It is prominent in functional genomics research for two reasons. The first one is the physiological role it plays in gene regulation. The traditional role of RNA was a passive intermediate in the translation of information from genes to proteins. Discovering its regulatory function is arguably one of the most important advances in molecular biology in decades. The second reason is that screens triggering RNAi of target genes can be applied on a genomic scale and allow rapid identification of genes contributing to cellular processes and pathways [19].

The RNAi mechanism RNAi is the disruption of a gene's expression by a double stranded RNA (dsRNA) in which one strand is complementary to a section of the gene's mRNA. It is described in detail in several recent reviews [85, 92, 15]. Fig. 1.3 gives an overview over the RNAi pathway. In an RNAi assay dsRNAs get introduced into the cell. In the cytoplasm they are processed by an enzyme of the Dicer family into small interfering RNAs (siRNAs). In mammals dsRNA molecules longer than 30 bp provoke interferon response, an antiviral defense mechanism, which results in the global shutdown of protein synthesis. RNAi can still be started by introducing siRNA molecules directly. Next, siRNA is assembled into an RNA-induced silencing complex (RISC). In fruitflies and mammals, the antisense strand is directly incorporated into RISC and activates it. In worms and plants the antisense strand might first be used in an amplification process, in which new long dsRNAs are synthesized, which are again cleaved by Dicer. Finally, antisense siRNA strands guide the RISCs to complementary RNA molecules, where they cleave and destroy the cognate RNA. This process leaves the genomic DNA intact but suppresses gene expression by RNA degradation.

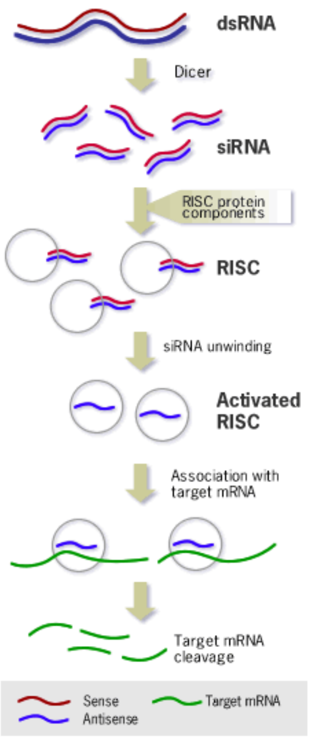

Figure 1.3: *The RNAi pathway. Reproduced from* `www.ambion.com`.

Bioinformatic challenges of RNAi RNA interference poses many challenges to research in computational biology. The first one is a better understanding of the RNAi mechanism by mathematical modeling and simulations [51]. Other challenges are specific to analyzing large-scale RNAi screens and include (i.) storage and preprocessing of data from RNAi experiments [113], (ii.) sequence analysis to identify unique siRNA targets and guard against off-target effects [91], and (iii.) ordering pathway components into regulatory hierarchies from phenotypic effects in RNAi silencing assays. This thesis contributes to the latter challenge. It proposes probabilistic models to infer pathway topologies from RNAi gene silencing data. Experimental techniques using the RNAi mechanism have drastically reduced the time required for testing downstream effects of gene silencing [19], but no work has been published so far to adapt statistical methodology to the specific needs of RNAi data. We will focus on two problems peculiar to RNAi. The first becomes apparent when comparing RNAi knockdowns to DNA knockouts, the second when deciding which phenotypes to observe.

Knockouts and knockdowns Genetic studies can be divided into forward or reverse screens [122]. In a *forward screen*, genes are mutated at random. To attribute a phenotype to a specific gene, the mutation must first be identified. This process is time-consuming and not easily applicable for all species. Additionally, some genes may always be missed by random sampling [19]. In contrast to random mutagenesis, *reverse screens* target specifically chosen genes for down regulation. This is what we will be concerned with in this thesis. The most direct way to silence a gene is by a gene knockout at the DNA level. Gene knockouts create animals or cell lines in which the target gene is non-functional [61]. It is difficult to interpret data from knockout mutants and to decide whether the phenotype is a direct effect of the non-functional gene or whether it is the result of the cell trying to compensate for the gene-loss. The danger of compensatory effects is less prominent for intervention techniques which allow faster down-regulation of target genes. In most cases, silencing genes by RNAi results in almost complete protein depletion after only a few days. Compared to gene knockouts, this makes silencing by RNAi more applicable in genome-wide screens and reduces compensatory effects at the same time. Two features make RNAi kockdowns "softer" than DNA knockouts. First, in an RNAi experiment the protein is not necessarily eliminated from the cells completely. A small amount of mRNA might escape degradation and protein can last a long time in the cell, if protein turnover is slow. This may mask or weaken phenotypes. On the other hand, this phenomenon may be useful in cases where a fully silenced gene would be lethal. Then the softer silencing by RNAi may still allow observations of phenotypes of the living cell. Second, even though transfection efficiency is typically high in RNAi experiments, transfection of cultured cells often results in a mixed population of cells, where some escape the RNAi effect. The observed phenotype is then an average over affected and not-affected cells.

In summary, all perturbation experiments *push* a gene's expression level towards a "no expression" state. Only in knockouts, however, the intervention leads to a completely non-functional gene. In RNAi experiments the gene is still active, but silenced. It is less active than normal due to human intervention. Hence, we do not fix the state of the gene, but push it towards lower activities. In addition this pushing is randomized to some extent: the experimentalist knows that he has silenced the gene, but in large-scale screens he cannot quantify the effect. This is the first problem approached in this thesis.

Phenotypic readout The term "phenotype" can refer to any morphologic, biochemical, physiological or behavioral characteristic of an organism. A number of phenotypes can be observed as results of perturbations [19]. Many genetic studies use *cell proliferation* versus *cell death* as a binary phenotype to screen for essential genes. Recently, large-scale identification of "synthetic lethal" phenotypes among nonessential genes, in which the combination of mutations in two genes causes cell death, provided a means for mapping genetic interactions [26]. To find genes essential for a pathway of interest, *reporter genes* or fluorescent markers are used to monitor activity of a signaling pathway [50]. Alternatively, visible phenotypes like *cell growth and viability* are screened for [13]. A global view of intervention effects can be achieved by transcriptional phenotypes measured on microarrays. These can either be global time courses in development [31] or differential expression of single genes [61, 12]. Also other cellular features like activation or modification states of proteins could be used as phenotypes of interventions. What singles out the phenotypes described above is that they are accessible to large scale screens by high-throughput techniques.

Primary and secondary effects To describe the second problem tackled in this thesis, we need to distinguish between primary and secondary effects of interventions. We speak of a *primary effect*, if perturbing a pathway component results in an observable change at another pathway component. To achieve this change a complex machinery could have been involved. Thus, primary effects are not indicators of direct interactions between molecules. They are primary in the sense that they only involve pathway components and allow direct observations of information flow in the network. A primary effect can, *e.g.*, be observed in a transcriptional regulatory network when silencing a transcription factor leads to an expression change at its target genes. Unfortunately, in the case of signaling pathways primary effects will mostly not be visible in large-scale datasets. For example, when silencing a kinase we might not be able to observe changes in the activation states of other proteins involved in the pathway. The only information we may get is that genes downstream of the pathway show expression changes, or that cell proliferation or growth changed. Effects, which are not observable at other pathway components, but only as phenotypical features downstream the pathway, will be called *secondary effects*. Secondary effects provide only indirect information about information flow and pathway structure. Reconstructing features of signaling pathways from secondary effects is the second problem addressed in this thesis.

Why probabilistic models? There are several reasons to use probabilistic models for regulatory networks and signaling pathways. First of all, the measurement noise in todays experimental techniques is notoriously high. Second, gene perturbation experiments always entail uncertainty of experimental effects. The most important reason for probabilistic models comes from the biological system itself. Signal transduction, gene expression and its regulation are a stochastic processes [106, 110, 98]. There are two types of noise: *intrinsic* noise due to stochastic events during gene expression, and *extrensic* noise due to cellular heterogeneity [106]. Intrinsic noise is responsible for differences between identical reporters in *the same* cell, and extrensic noise for differences between identical reporters in *different* cells. Probabilistic models take care of all these kinds of noise.

1.3 Thesis organization

In summary, there are two problems to be addressed when modelling data from RNAi experiments. First, how to account for the uncertainty of intervention effects in a noisy environment. Second, how to infer signaling pathways if direct observations of gene silencing effects on other network components may not be visible in the data. This thesis proposes novel methodology to address both questions. It is organized as follows.

Statistical models of cellular networks Chapter 2 gives an overview of recent approaches to visualize the dependency structure between genes. Even though reverse engineering is a fast developing area of research, the methods used can be organized by a few basic concepts. Statistical network methods encode statements of *conditional independence*: can the correlation observed between two genes be attributed to other genes in the model? Methods implementing this idea include graphical Gaussian models and Bayesian networks. Bayesian networks are the most powerful and flexible statistical model encoding the highest resolution of dependency structure. The methodology described here will be the basis for building models for interventional data in the following chapters.

Inferring transcriptional regulatory networks In chapter 3, we develop a theory of learning from gene perturbations in the framework of conditional Gaussian networks. The basic assumption is that effects of silencing genes in the model can be observed at other genes in the model. To model the uncertainty involved in real biological experiments, perturbations are modelled stochastically—and not deterministically as in classical theory. This answers the first question raised by RNAi data.

Inferring signal transduction pathways The methods described so far elucidate the dependence structure between observed mRNA quantities. Chapter 4 goes one step further. It shows that expression data from perturbation experiments allows to infer even features of signaling pathways acting by non-transcriptional control. The signaling pathway is reconstructed from indirect observations. This answers the second question raised by RNAi data. The proposed algorithm reconstructs pathway features from the nested structure of affected downstream genes. Pathway features are encoded as silencing schemes. They contain all information to predict a cell's behaviour to an external intervention. Simulation studies confirm small sample size requirements and high accuracy. Limits of pathway reconstruction only result from the information content of indirect observations. The practical use is exemplified by analyzing an RNAi data set investigating the response to microbial challenge in *Drosophila melanogaster*.