The symphony of gene regulation
van Dijk, D.

Citation for published version (APA):
Chapter 1 Introduction

1.1 A new age in biology
The exponentially decreasing cost of DNA sequencing has resulted in an exponentially increasing number of sequenced genomes1 (Figure 1.1). In parallel, new high-throughput experimental techniques can generate equally large amounts of functional data, e.g. expression levels of all genes in a cell, or the interaction-network of all proteins with all other proteins. However the huge amount of data that we are able to generate has created a new bottleneck. One of the great challenges in the post genomic era is to infer biological functioning from DNA sequence and other high-throughput experimental data. Our ability to generate quantitative and accurate predictions regarding basic things such as mechanisms of gene regulation and protein-protein interactions remain severely limited. The rapidly expanding field of systems biology combines mathematical models and high-throughput experimental data in order to understand these fundamental processes.

![Figure 1.1: The increasing number of sequenced organisms. The exponentially decreasing cost of DNA sequencing has resulted in an exponentially increasing number of sequenced genomes. Shown here are the number of fully (blue) and partially (red) sequenced genomes.](image)

1.1.1 Biology and information processing
This systems biology approach in recent years has put forward a picture of great biological complexity yet one that appears to have emerged with a simple set of rules much like the laws of physics. It is the goal of systems biologists to uncover these rules or design principles and understand how they evolved. In particular, the quantitative sciences have contributed to our ability to measure, predict and understand biology. Mathematics, physics and computer science have provided the tools to model biological functioning and enable us to answer questions on how
organisms respond to their environments and regulate reproduction given the constraints of biochemistry.

Diverse types of such regulatory systems lie at the heart of biology, most importantly the regulation of gene expression. In this thesis we therefore ask: “How is gene regulation encoded in the “wiring” of biological interactions?” Using two different biological systems and two different regulatory mechanisms – protein-protein interaction between the HIV virus and its host cell, and transcriptional regulation in *Saccharomyces cerevisiae* (budding yeast), we show that despite the complexity and diversity of biology there are clear design principles, employed by evolution, of how gene regulation and therefore information processing occurs in nature.

### 1.2 Biological Background

#### 1.2.1 Virus-host protein-protein interaction networks with HIV-1

The DNA encoded cellular circuitry that determines biological behavior is for the most part implemented in protein-protein interactions. This means that the behavior of a protein is largely determined by its interaction partners. Therefore, large-scale maps of protein interaction can be used to understand biological functioning. Usually protein-protein interaction networks of single organisms are studied, however the interaction between two different organisms or biological entities result in similar interaction networks. This is the case for virus infection.

One of the most studied viruses is HIV-1, which causes around 2 million deaths per year. Currently 33.3 million people are estimated to be living with it and each year around 2.7 million people are newly infected[2]. The HIV-1 positive strand RNA virus that, like all viruses, relies on the bio-molecular machinery of the host cell to reproduce. It is the physical interaction between the viral proteins and the proteins produced by the host cell that permits the virus to infect the host cell, replicate, and lyse the host cell in order to go on and infect more cells. Understanding this viral ‘life-cycle’ is essential for designing a cure for any viral disease. Thus, understanding the ways in which HIV interacts with it’s host cells, macrophages and CD4+ T-cells, has medical value, and has become a well-studied model system for the viral life-cycle.

During each stage of the virus life-cycle complex interaction occurs between virus and host. First, when virus particles enter the human body an immune response is triggered when immune cells recognize the virus exterior as a foreign entity. Next, infection occurs when the virus envelope recognizes receptors that are expressed on the surface of the host’s cells. Then, after the virus releases its contents into the host cell, reverse transcribes its genome to DNA and integrates it into the host, viral proteins regulate host transcription to prefer expression of the viral genome. In turn, intra-cellular immune response of the host tries to down-regulate virus expression and interfere with virion production[3]. These stages of infection
combined give rise to a complex gene and protein interaction network. In this thesis we seek to quantify this network using a set of computational techniques with the goal to confirm existing and uncover new biological mechanisms that occur between virus and host (see Chapter 2). In addition, we seek to uncover new interactions by predicting cellular surface proteins that interact with HIV (see Chapter 3).

1.2.2 Transcriptional regulation in budding yeast

Proper control of mRNA levels is critical in nearly all biological processes. Since much of this control is encrypted within non-protein-coding regulatory regions, deciphering the details of this mapping between DNA sequence and mRNA expression levels is key for understanding transcriptional control. Such an understanding could allow us to predict gene expression from DNA sequence, with far-reaching implications. Most notably, genetic studies in a broad range of human diseases found a substantial contribution of genetic variation in non-coding regions to phenotypic diversity, and many expression changes have in turn been linked to disease states. However, without a ’regulatory code’ (comparable to the ’genetic code’), we cannot tell which sequence changes cause the observed expression changes, and by what mechanism. Despite many studies of transcriptional control, it is surprising how little we know about the quantitative effect on expression of even the most basic organizational features of promoters.

Precise control of the average expression across a population of cells is not enough. Gene expression is noisy, and this limits the precision with which cells can regulate protein levels. Changes in transcription factor (TF) activity results in changes in the expression of target promoters, but the way in which cell-to-cell variability in expression (noise) changes as a function of TF activity is less well understood. This is in spite of observations that noise in gene expression has a substantial fitness cost and that the level of noise for each gene is under selective pressure. Furthermore, stochastic variation in gene expression has been implicated in diverse processes such as resistance to chemotherapy and antibiotics, stem cell reprogramming and penetrance of otherwise Mendelian traits. In order to fully understand how organisms and genetic networks function in-vivo it is absolutely essential that we understand the regulation of noise. Furthermore, because different mechanisms of regulation have different effects on noise, measurements of noise in gene expression provide insights into the molecular mechanisms of transcriptional regulation.

The model organism Saccharomyces cerevisiae (budding yeast) is uniquely suited for this work because its gene regulation is well characterized and because of the range of regulatory manipulation possible. Budding yeast has long been a model organism for genetics due to the ease with which it can be grown, mutated and mated in order to conduct experiments in classical genetics. Budding yeast was at the forefront of the revolution in molecular genetics due to the ease with which its genome can be manipulated. Thus it is well suited to continue to lead the functional-genomics revolution due to the ease with which large-scale libraries of
altered organisms can be both generated and measured. For these reasons we use baker’s yeast as a model organism for studying many different types of molecular mechanisms, such as gene regulation.

1.2.3 Modeling Background
The complex and quantitative nature of gene regulation and interaction makes mathematical and computational modeling ideal for studying and predicting regulatory patterns. Different modeling techniques and mathematical representations have been proposed for studying biological regulatory systems (reviewed in 4,5). In this thesis we describe the topology of virus-host regulation using a network representation and the dynamics of individual regulatory interactions in yeast using kinetic and stochastic modeling.

In Chapter 2 and 3 we model gene regulation and gene interaction using a network approach. We focus on the wiring of the network as a whole and use a set of mathematical tools to infer biological meaning from the network structure and topology. In Chapter 4 and 5 we present a kinetic model of gene regulation. Here we focus on individual genes and model transcriptional regulation as a function of promoter DNA sequence. We use both stochastic simulations and an analytical solution to predict gene expression level and cell-to-cell variability.

1.2.3.1 Properties of complex biological networks
Interaction networks are central to biological regulation. Even when we do not know the quantitative properties of individual interactions between components in a network, the wiring itself, especially in larger networks, can be used to understand system-wide behaviors.

Networks that are made up of many interacting components and have a non-trivial topology are often called complex networks6,7. In such networks relatively simple interactions can give rise to complex behavior and network wiring which cannot be explained from the simple interactions alone. A wide range of mathematical concepts and computational tools has been developed to find structure in these networks. Complex networks are often found in the real world such as social, biochemical4,5 and computer networks and share common features such as heavy tail degree distributions, clustering and a hierarchical or community structure. By studying properties of the topology we can learn about the behavior of the system. For example, the heavy tail degree distribution is often a sign of a process called preferential attachment in which new connections are more likely to form in already well-connected nodes. Many biological networks, such as transcriptional regulatory, protein-protein interaction and metabolic networks are “complex” and show the typical topological features. In this thesis we will use complex network analysis to study virus-host interaction (see Figure 1.2). We find that the HIV-human protein and gene interaction networks have a heavy-tail degree distribution and we use this topological property to find which HIV and which human proteins are important for viral infection.
A network motif is the simplest topological building block that a network is made of. By quantifying if certain three (see Figure 1.3) or four node patterns are present more than would be expected from chance, we can learn about how a network grew (evolved) and how it behaves. Network motifs have been found in a wide range of networks. For example the transcriptional regulatory network of yeast and E. coli showed network motifs that represent regulatory logic or logical gates. Interestingly, the same logical gates where found in the network motifs of electronic circuits. In this thesis we use network motifs to quantify virus-host interaction in terms of immune response and immune evasion.
Figure 1.3 Network Motifs. All possible three-node network motifs of a directed graph. (Figure is from12).

1.2.3.2 Kinetic models of transcriptional regulation

If the connections in the network describe reactions, such as transcriptional regulation, we can describe not just the topology, or static behavior, but also the dynamics of the network.

A major challenge in biology is to predict the expression level for a set of interacting genes, given the cellular condition. This prediction requires knowledge of the regulatory circuit in which gene products, such as transcription factors, regulate other genes through promoter binding. While often the qualitative network, or "wiring", is known, a quantitative prediction of gene expression level is much harder to achieve since it relies on knowledge of the specific stoichiometry of many (unknown) reactions.

In order to model these biochemical reactions we can setup a kinetic model that describes the different reactions of the system. In transcriptional regulation we can model gene expression as a function of transcription factor (TF) binding to the promoter region where TF binding alters the promoter configuration and as a consequence change the transcriptional efficiency of the gene15. Such a system consists of a set of reactions describing the switching between different promoter states as a result of binding and unbinding of TFs. In addition there are transcription, translation and mRNA and protein degradation reactions. Mathematical representation of these reactions exist in the form of differential equations which, in order to get the steady state solutions, can be solved analytically or numerically.
Gene expression is a stochastic process. In order to model the distribution of mRNA or protein abundance, i.e. cell-to-cell variability, several techniques exist. Numerical simulation using the Gillespie algorithm\textsuperscript{18} is a useful method in particular for systems that do not have analytical solutions or are difficult to solve analytically. Preferably, the system is solved analytically, for example using the Master equation\textsuperscript{19}.

1.3 Research questions
Gene regulation is one of the main mechanisms with which biological systems process information. In this thesis our goal is to understand, using both a computational and experimental approach, how biology is designed to process information. Our goal is to find the principles of biological regulation and therefore ask: "How is information processing encoded in biological functioning?"

Using different mathematical representations and modeling techniques we set out to uncover the different layers of biological regulation. We start out at the network level where we quantify the topological wiring of regulation. For this we use the case study of HIV virus-host interaction. Here we ask: "How is gene regulation encoded in the topology of interaction networks?".

We then zoom in on the bio-molecular mechanisms of gene expression and investigate the kinetics of individual gene regulatory interactions. For this we use baker’s yeast as a model organism. We ask: "How is transcriptional regulation encoded in the DNA sequence?" In specific we study the kinetic properties of transcription and translation that result in stochastic (noisy) gene expression, and ask to what extent the noise of gene expression reflects the underlying mechanisms of gene regulation.

1.4 Thesis overview
In chapter 2 we investigate how the HIV-1 virus infects and survives given the human immune system. We model HIV-1 infection using a virus-host interaction network and quantify infection dynamics and host immune response using complex network theory. We find that the central connectivity of the HIV-1 virus and it’s
involvement in specific interaction patterns enable it to evade the human immune response.

In chapter 3 we predict new surface membrane receptors that interact with HIV-1 using a network analysis approach. We present a bioinformatics algorithm that predicts existing and new receptors. We confirm existing and find a set of putative surface receptors.

In chapter 4 we investigate the different regulatory mechanisms of transcription factors using the stochastic nature of gene expression. We measure the protein abundance of native and mutant promoters that are all regulated by the same TF. We perform these measurements in a dose response, where we incrementally change the concentration of the input TF and measure gene expression output. We model the behavior with a stochastic model that describes mean gene expression and cell-to-cell variability (noise) as a function of TF concentration. We find that the relationship between mean expression level and noise is unique to each promoter and reflects the underlying mechanisms of transcriptional regulation. We conclude that a single TF can regulate gene expression in multiple different ways, each resulting in a unique noise profile.

In chapter 5 we used single cell time-lapse microscopy to compare the effect on transcriptional dynamics of two distinct types of sequence changes in the promoter that can each increase the mean expression of a cell population by similar amounts but through different mechanisms. We show that increasing expression by strengthening a transcription factor binding site results in slower promoter dynamics and higher noise as compared to increasing expression by adding nucleosome-disfavoring sequences. The latter strategy likely reduces the expression variability of the cell population by increasing the frequency of transcription events. Our findings are explained well by a model in which the promoter stochastically switches between a transcriptionally active and inactive state, and where expression is modulated by changing either the frequency or period of activity. Taken together, our study demonstrates the effect of cis-regulatory elements on expression variability and points to concrete types of sequence changes that may allow partial decoupling of expression level and noise.

Finally, in chapter 6, we summarize our findings and state the contributions of this dissertation to the scientific community, and end with perspectives and future directions of this work.
1.5 References


14 | Introduction

57. Sanchez, A. & Kondev, J. Transcriptional control of noise in gene expression.
