The Construction of Transcription Factor Networks Through Natural Selection

Alexander James Stewart

January 2010

A thesis submitted submitted to University College London for the Degree of Doctor of Philosophy

CoMPLEX
UCL
Physics Building
Gower Street
London, WC1E 6BT

I, Alexander James Stewart, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

for Charlie and Vicky, who spent the last four years doing something way more difficult than writing a thesis

Acknowledgements

I'd like to thank my supervisors, Andrew Pomiankowski and Rob Seymour, for their support, guidance and, in several instances, patience. I'd also like to thank Max Reuter for his contributions to the work on negative autoregulation, and for the idea to work on it in the first place.

Thanks all the people in CoMPLEX who have contributed to this work, particularly by making me realise when I'm making no sense, which is something I find rather difficult to come to terms with on my own. In this respect I'd particularly like to thank Dave, Lisa, Little Sam and Big Sam. And also to all my friends in the outside world - your tolerance at times seems almost infinite. In particular this applies to Marina - I really do appreciate you, I promise. It also applies to Charles, but you don't seem to mind so much.

Finally I'd like to thank my parents, who have been neglected a little, particularly in the last year, but have never shown anything other than support and mild bafflement. You are as cool as you think you are. Possibly cooler.

Abstract

Transcription regulation plays a key role in determining cellular function, response to external stimuli and development. Regulatory proteins orchestrate gene expression through thousands of interactions resulting in large, complex networks. Understanding the principles on which these networks are constructed can provide insight into the way the expression patterns of different genes co-evolve.

One method by which this question can be addressed is to focus on the evolution of the structure of transcription factor networks (TFNs). In order to do this, a model for their evolution through *cis* mutation, *trans* mutation, gene duplication and gene deletion is constructed. This model is used to determine the circumstances under which the asymmetrical *in* and *out* degree distributions observed in real networks are reproduced. In this way it is possible to draw conclusions about the contributions of these different evolutionary processes to the evolution of TFNs. Conclusions are also drawn on the way rates of evolution vary with the position of gene in the network.

Following this, the contributions of *cis* mutations, which occur in the promoters of regulated genes, and *trans* mutations, which occur in the coding reign of transcription factors, to the evolution of TFNs are investigated. A space of neutral genotypes is constructed, and the evolution of TFNs through *cis* and *trans* mutations in this space is characterised. The results are then used to account for large scale rewiring observed in the yeast sex determination network.

Finally the principles governing the evolution of autoregulatory motifs are investigated. It is shown that negative autoregulation, which functions as a noise reduction mechanism in haploid TFNs, is not evolvable in diploid TFNs. This is attributed to the effects of dominance in diploid TFNs. The fate of duplicates of autoregulating genes in haploid networks is also investigated. It is shown that such duplicates are especially prone to loss of function mutations. This is used to account for the lack of observed autoregulatory duplicates participating in network motifs.

From this work, it is concluded that the relative rates of different evolutionary processes are

responsible for shaping the global statistical properties of TFN structure. However, the more detailed TFN structure, such as network motif distribution, is strongly influenced by the population genetic details of the system being considered. In addition, extensive neutral evolution is shown to be possible in TFNs. However, the effects of neutral evolution on network structure are shown to depend strongly on the structure of the space on neutral genotypes in which the TFN is evolving.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Transcriptional regulation lies at the heart of many of the most important questions currently facing Biology. Sets of regulatory interactions between genes can be characterized as transcription factor networks (TFNs), which determine how those genes are expressed over time to give rise to complex traits. This means that understanding how TFNs function and evolve is a vital step in linking genotype to phenotype. Construction of the genotype-phenotype map is, in turn, a key step in understanding how complex organisms function and evolve.

TFNs can be studied at different levels of detail, from coarse measures of global statistical properties, such as network degree distribution [35, 47, 83, 116, 119], the function of specific subnetworks, such as network motifs [2, 57, 68, 79, 88, 87, 110, 146], down to the molecular details of transcription factor binding [11, 18, 39, 40, 63, 89, 108, 118]. A complete understanding of the mechanisms of TFN evolution requires us to embrace all of these levels of detail. Some network properties, such as a broad tailed degree distribution, are common to a wide range of biological networks, not just TFNs, but also protein interaction networks, metabolic networks and even social networks. Such universal properties require a general explanation which does not depend on the details of any one system. Other properties are specific to TFNs, but are independent of the species being considered. For example an asymmetrical *in* and *out* degree distribution [47, 83, 119], modular network structure [58, 129], and certain network motifs [88, 87, 110] are all found in organisms as diverse as bacteria, yeast and *Drosophila*. Properties such as these require an explanation based on the general mechanisms of TFN evolution, but independent of the details of any one species' environment or specific evolutionary history. Yet other properties are specific to particular species, or even vary from individual to individual within a population.

For example the patterns of *cis* regulation involved in *Drosophila* wing patterning [43, 96, 97] or yeast sex determination [124, 125] vary even between closely related species. These properties require species specific explanations. When studying TFNs, an approximate rule of thumb is that the greater the level of detail used when analysing a network, the more specific the information obtained is to that particular network.

The research presented here seeks to use several different properties of TFNs to elucidate the factors that are important in their evolution. It is focused on the role played by different types of mutation and the role played by population genetic factors in shaping the structure of TFNs. In this introduction I describe the methods of modelling the function and evolution of TFNs which will be used to address these questions.

1.1 Modelling TFN Evolution

The function of TFNs can be studied both empirically and theoretically. Empirically, sets of genes which are coregulated or involved in the same biological process (e.g development, homeostasis, apoptosis) are identified. From this, subnetworks which perform a particular function can be constructed. Typically such subnetworks are inferred from a combination of gene expression data (e.g from microarry experiments), DNA binding data and binding motifs in the promoters of regulated genes [47, 83, 119]. Additionally, network motifs - subnetworks consisting of a small number of genes which occur at higher frequency in real biological networks than would be expected by chance - can be identified. Network motifs are thought to represent the "functional building blocks" of biological networks [88, 87, 110]. Once a network motif has been identified, it can be studied for its functional properties (e.g noise filtering, bi-stability, rapid response to external stimuli), allowing the function of larger subnetworks to be deconstructed.

Theoretically, the functional properties of TFNs can be studied by constructing models of transcription regulation. These relate the expression of one gene to the expression of another, given a regulatory interaction exists between them. Such models can be very abstract, for example models in which TFNs are represented as boolean networks, with genes either in an "on" or an "off" state [1, 61, 103, 111]. Alternatively they can be very detailed, for example when features such as the binding energies of regulatory binding sites, nucleosome occupancy or protein-protein interactions, are explicitly included in the models [108, 118, 125]. The functional properties of such "toy" networks can often be explored analytically, if the networks considered are simple enough, or else through computer simulation.

In order to study the evolution of TFNs, an understanding of TFN function must be combined with an understanding of the process through which evolution occurs. This can be done by determining the possible mutations which can occur and give rise to changes to the network. Empirically, these mutations can take four possible forms:

- *cis* mutations
- *trans* mutations
- gene duplication
- gene deletion

Mutations at *cis* are changes to transcription factor binding sites, which lie in the promoter region of a regulated gene. These mutations may result in the increase or decrease of the binding affinity of a binding site, which in turn may alter the function of the network. In addition, *cis* mutations may result in the complete loss or gain of a regulatory interaction, which as well as potentially changing the function of the network, also changes its architecture. Mutations at *trans* are changes to the transcription factor protein itself. These mutations typically occur in the coding region of the gene which codes for the transcription factor. A *trans* mutation may simultaneously affect some or all of the regulatory interactions in which the transcription factor participates, or give rise to multiple new interactions which did not exist previously. As such *trans* mutations can affect the functioning of a network by altering the strength of several existing regulatory interactions, as well as changing the network architecture by giving rise to the loss or gain of multiple regulatory interactions. Gene duplication is the copying of a gene and (potentially) all of its regulatory interactions. Gene duplication can occur through the independent copying of a single gene, or through sets of genes being duplicated together. In the most extreme case, a whole genome duplication occurs in which all of the genes and interactions in the TFN are copied. Duplication results in a change to the architecture in the network, both by copying a set of regulatory interactions and by increasing the size of the network. Gene deletion results in the loss of a gene and all of its regulatory interactions from the network. This may occur through the loss of a single gene, or of sets of genes simultaneously. Deletion results in a change to the architecture of the network through loss of multiple regulatory interactions and a decrease in the size of the network.

When considering the evolution of a TFN, we must consider those mutations which arise in an individual and then become fixed in the population. Therefore both the rate at which mutations occur and the probability of them becoming fixed in the population must be determined, in order

to understand the role that different types of mutation play in the evolution of TFNs. As well as their effect on the primary function of the network, a number of other factors determine whether a particular mutation will become fixed.

One of the most important of these factors is robustness. Robustness to intrinsic noise (due to the stochastic nature of transcription) or external noise (due to variation in the environment external to a cell) may be viewed as part of the function of a network - if a network is not sufficiently robust to these forms of noise it can reasonably be said not to function properly [2, 1, 79, 78, 101]. However, the role played by mutational robustness in the evolution of TFNs is more ambiguous. It has often been argued [20, 27, 31, 82, 112, 134, 133, 135] that robustness to mutations resulting in changes to the strength of regulatory interactions, loss or gain of an interaction or loss or gain of a gene, has played an important role in shaping the structure of TFNs. In addition to robustness, other population genetic factors may have a strong influence on the evolution of TFNs. Because TFNs are by their nature strongly interconnected, mutations may affect the functioning of the network in a complicated way. In particular, mutations which affect multiple regulatory interactions simultaneously may have strong pleiotropic effects. It is for this reason that it is often argued that evolution is dominated by *cis* mutations, as they tend to affect only a single regulatory interaction [43, 56, 97]. This also suggests that multiple mutations in a TFN will tend not to be additive in their affect on gene expression, but will interact epistatically. This is particularly important in sexual populations, in which recombination will tend to bring mutations at different loci together in the same individual. As a result, it has been suggested that recombinational robustness may play a significant role in shaping the evolution of TFNs in sexual populations [72]. Finally, the evolution of a TFN may be strongly influenced by whether the organism considered is haploid or diploid. This results from the tendency of mutations in diploid networks to give rise to dominance effects. In particular the different dominance effects arising from *cis* and *trans* mutations may influence the extent to which these different types of mutations become fixed in diploid as opposed to haploid organisms [70].

In addition to determining the mechanisms of evolution, other key questions concerning TFN evolution are the extent to which network structure is shaped by neutral or adaptive processes and the extent to which network structure constrains evolutionary change. These three questions are strongly interconnected. Where networks evolve through an adaptive process, mutations are fixed which improve the function of at least one of its subnetworks. Where evolution is neutral, mutations are fixed which leave the global function of the TFN unchanged. Whether different

types of mutations tend to be deleterious, neutral or adaptive will depend to a large extent on the population genetic factors outlined above, as well as on the structure of the network. Similarly, the extent to which network structure constrains evolutionary change is determined by the extent to which networks of a particular structure are robust to different types of mutation. In order to address these questions, it is common to construct "toy" models of transcription networks which can be subjected to *in silico* evolution [3, 12, 20, 30, 54, 61, 75, 74, 80, 112, 113, 131]. This has the advantage that populations of networks can be created, and the properties of these networks studied in order to determine general principles about the evolution of TFNs. These properties can then be compared to those observed in real networks, allowing predictions to be made about the principles governing TFN evolution.

1.2 Models of Transcription Regulation

In order to construct a model of a TFN it is necessary to understand the mechanism through which transcription regulation occurs. A TFN consists of nodes (genes) and edges (regulatory interactions between genes), and provides a description of how the expression levels of different genes affect one another. Depending on the level of detail included in a model of transcription regulation, a TFN can be used to address different questions about the evolution of regulatory interactions. Models which include little detail (for example boolean networks) are less computationally expensive to model *in silico* and may be easier to deal with analytically, than models which include a high level of detail. They are often used to construct large TFNs consisting of tens or hundreds of genes, which can be subjected to *in silico* evolution in order to address questions concerning the architectural properties of TFNs, such as the relationship between network structure and mutational robustness [12, 61, 112, 113, 131]. In contrast, models that include a high level of detail can be used to explore the properties of small networks (consisting of only a few genes), which are of special interest, such as the noise filtering properties of negative autoregulation or feed forward loop (FFL) motifs [2, 77]. These approaches are often complementary. In this section I discuss four of the most commonly used models of transcription regulation, and how they relate to one another. In addition, I discuss the physical mechanism of transcription regulation which these models attempt to capture.

1.2.1 Transcription Regulation

The basic mechanism of transcription regulation is the binding of transcription factors (TFs) to the promoter region of a regulated gene. TF binding occurs at specific binding sites within the promoter region of the regulated gene. Regulatory binding sites consist of a nucleotide sequence (a binding motif), usually with a length of between 6 and 12 bp [76, 150]. These are identified using a combination of DNA binding data and expression data, which allow a consensus sequence for the binding motif to be determined. A consensus sequence can then be used to search for other potential targets of a TF, which in turn allows the set of all regulatory targets of that TF - its regulon - to be identified. Pairs of TFs which regulate the same target may interact, giving rise to a more complicated regulatory process. In particular, different types of TF may compete for the same binding site, (competitive binding). Alternatively, one TF may aid the binding of a second TF to a target through a protein-protein interaction (cooperative binding), or by making its binding site available by ejecting a nucleosome which covers it [67, 125]. TFs may also bind non-specifically to regions of DNA, which can affect the ability of other TFs to bind to their specific binding sites [40, 18].

In eukaryotes the ability of a TF to bind to a specific binding site is limited by the chromatin structure at the region of the DNA in which the binding site lies [94, 100, 101]. When regulatory binding sites are not occupied by a nucleosome, TFs are able to bind to them. However, when binding sites are occupied by a nucleosome, TFs are unable to bind [48]. The changes in chromatin structure which result in binding sites moving in and out of a state in which they are occupied by a nucleosome is thought to be responsible for the bursting dynamics observed in the expression of eukaryotic genes [101]. In order for transcription of a gene to occur, the correct chromatin structure at the transcription start site and the presence of a preinitiation complex (PIC) are required [28]. Here I discuss the changes in chromatin structure which occur in the yeast *Saccharomyces cerevisiae*, as this is the eukaryote for which a TFN has be most widely studied, and the organism to which the results in later chapters will be compared. In *S. cerevisiae* it is found that, in general, chromatin change is causally preceded by TF binding [94]. Bound TFs recruit histone acetyl transferases if they activate transcription, or deacetylases if they repress transcription [48, 94]. This recruitment alters nucleosome acetylation, which alters nucleosome occupancy at a particular region of the promoter [48, 94]. Nucleosomes containing a variant histone H2A.Z, flank a nucleosome depleted region just upsteram of the transcription start site [48, 99]. In order to allow transcription initiation, H2A.Z nucleosomes are acetylated and ejected, making the transcription start site

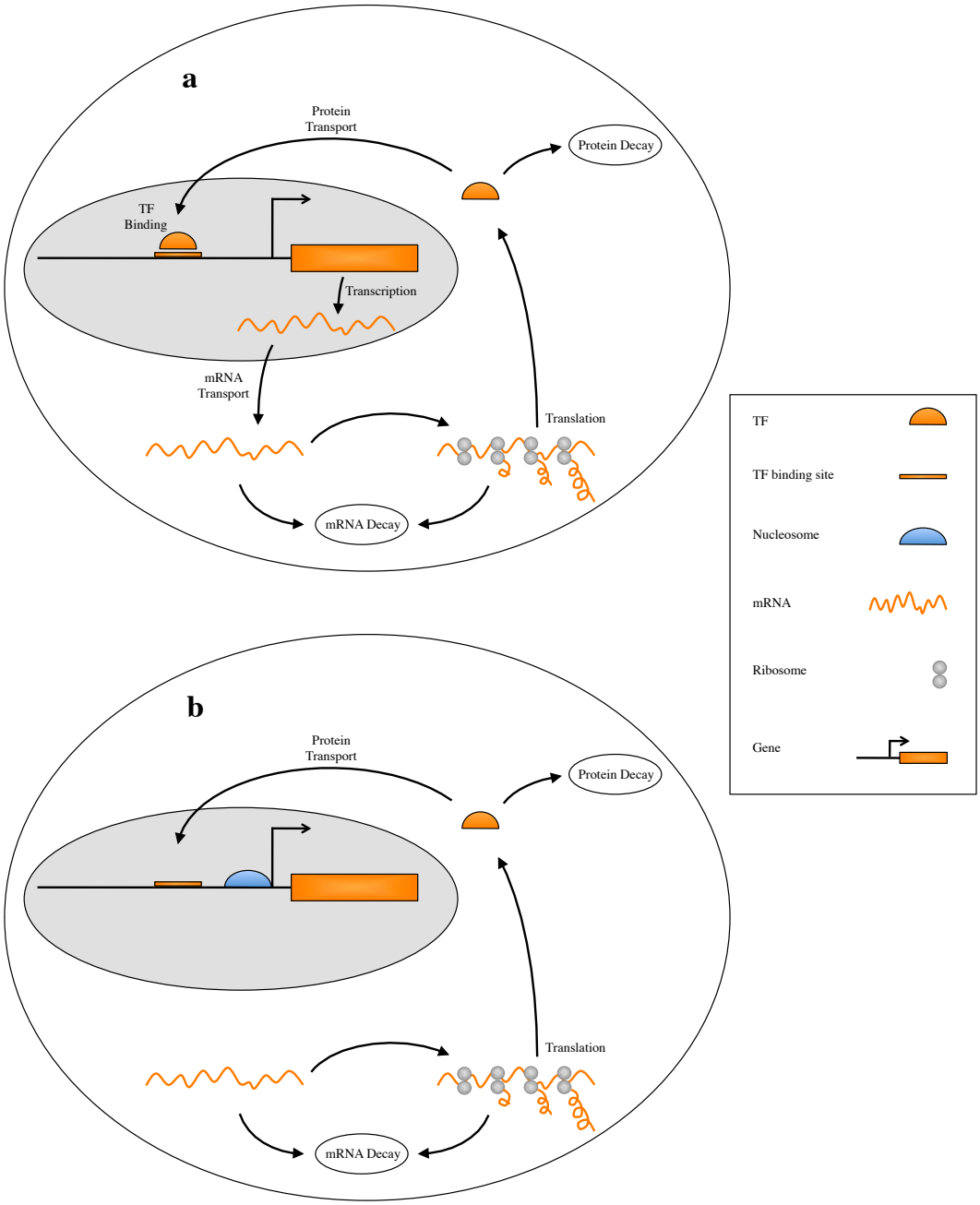

Figure 1.1: Mechanism of transcription regulation. For convenience, a gene whose protein product activates its own transcription is shown. a) Transcription occurs in the nucleus (grey region). When a TF is bound, transcription occurs and mRNA s produced. This is then transported out of the nucleus into the cell (white region). mRNA is then translated into protein at the ribosome. Protein is then transported back into the nucleus. b) When no TF is bound, a nucleosome occupies the transcription start site, and no transcription occurs. However, any mRNA remaining in the cell continues to be translated into protein, resulting in a time lag between transcription stopping and protein concentration decreasing.

accessible. Repression occurs when deacetylated nucleosomes return to the transcription start site, preventing further transcription [60]. Bound TFs can either preclude or allow transcription initiation by causing the transcription start site of the regulated gene to be bound or unbound by a nucleosome. Alternatively, bound TFs may change the nucleosome occupancy at other regulatory sites, in turn allowing other TFs to bind to those sites [67].

In addition to histone acetylation, it is necessary for a complete PIC to assemple at the transcription start site. The PIC consists of general TFs and accessory complexes [28]. Assembly of a PIC takes a long time (20-30 mins [143]) compared to the time required for ejection of the nucleosome from the transcription start site (approximately 5 mins [15]). However, a partially formed PIC may remain bound to the promoter following previous transcription events, speeding up formation of a new PIC in subsequent rounds of transcription [28, 144, 145]. In this case, reinitiation of transcription occurs as quickly as 3-5 mins [144, 145]. As a result the promoter may occupy a number of states - that in which transcription cannot occur, that in which transcription can occur but is slow to begin, or than in which transcription begins quickly [108].

Following transcription, mRNA is synthesised, which is translated at the ribosome into protein. Where the translated protein is a TF, it in turn will then bind to its target genes to regulate their transcription. The process of DNA binding, transcription and translation is illustrated in figure. 1.1. Although this qualitatively describes the process of transcription regulation as it occurs in *S. cerevisiae*, it is often convenient to model transcription regulation more simply when studying the evolution of TFNs *in silico*. In the following sections I describe four types of model which capture this process of transcription regulation at different levels of detail.

1.2.2 Boolean Network Models

One of the simplest ways of representing a TFN is as a boolean network [1, 61, 103, 111]. In a boolean network model, genes can only occupy one of two states - either “on” or “off”. Therefore genes are thought of as either expressed at their maximum level or else they are not expressed at all. The state of a boolean network consisting of N genes, at a time t, is given be the state vector $S(t) := (s_1(t), s_2(t).....s_N(t))$, representing the expression levels of each gene in the network. Time in boolean networks is taken to be discrete. Therefore the time evolution of a gene, i, may be written as

$$s_i(t+1) = f\left(\sum_{j=1}^{N} w_{ij}s_j(t) - \theta_i\right) \tag{1.1}$$

where $f(x)$ is a threshold function such that $f(x) = 1$ if $x > 0$ and $f(x) = 0$ otherwise. The term $\sum_{j=1}^{N} w_{ij}s_j(t) - \theta_i$ determines the input to gene i from all other genes in the network. w_{ij} is the strength of the input from gene j to gene i. If $w_{ij} = 0$ there is no regulation of i by j. If $w_{ij} > 0$, j activates i, whilst if $w_{ij} < 0$, j represses i. The matrix W of elements w_{ij} is the connectivity matrix for the network and defines the network structure. The term θ_i defines the activation threshold of gene i, so that when $\sum_{j=1}^{N} w_{ij}s_j(t) > \theta_i$, gene i is "on".

The evolution of boolean networks can be studied *in silico*, with mutations occurring which change the strength of the interactions w_{ij}, and selection occurring either on the final equilibrium state $S(t \to \infty)$ of the network, or else on the time evolution of the state of the network. The second case can be used to model a developmental process (e.g [112]).

A simple generalisation of the model in equation (1.1) can be made by replacing the threshold function $f(x)$ with the sigma function $\sigma(x)$, where

$$\sigma(x) = \frac{1}{1 + \exp[-hx]} \tag{1.2}$$

In this case genes are no longer either "on" or "off", but have an expression level which may vary continuously between 0 and 1. Here h defines how steep the threshold of the sigma function is, so that in the limit $h \to \infty$ the sigma function becomes the threshold function $f(x)$. Models of this type have been widely used to construct "toy" models of TFNs. They allow the dynamical properties of ensembles of networks with different structural properties to be investigated [61], as well as the evolution of properties such as mutational robustness [131], genetic canalization [112] and epistasis [3, 74] in TFNs. They have also been employed to investigate the evolution of real networks with a known function, such as the *Drosphila* sex determination network [75]. Such models are simple to construct and to simulate, and are therefore of use in the *in silico* study of TFN evolution. However, they lack some important features of transcription regulation as observed in real systems. In particular they do not model mRNA concentration, stochasticity in gene expression or time delays associated with transcription and translation, which are important factors in transcription regulation. For this reason it is sometimes difficult to draw clear conclusions about the evolution of real TFNs from *in silico* studies of boolean networks and their derivatives.

1.2.3 ODE Models

In order to capture the process of transcription regulation in more detail, many studies construct a system of ordinary differential equations (ODEs). ODE models describe the time evolution of gene expression,which is often referred to as protein concentration in this context. Perhaps the simplest ODE model describes the rate of production of protein i, $\frac{dP_i}{dt}$ as

$$\frac{dP_i(t)}{dt} = \beta_i f\left(\sum_{j=1}^{N} w_{ij} P_j(t) - \theta_i\right) - \gamma_i P_i(t) \tag{1.3}$$

where $f(x)$ is a threshold function and $\sum_{j=1}^{N} w_{ij} P_j(t) - \theta_i$ is the input to gene i from all the other N genes in the network, just as described for boolean networks above [1]. The term β_i is the maximum rate of production and γ_i the rate of degradation of protein i. There is a clear analogy between the ODE model in equation (1.3) and the discrete time model in equation (1.1), since at equilibrium equation (1.3) either has solution $P_i = \frac{\beta_i}{\gamma_i}$ (i.e. it is maximally expressed), or else $P_i = 0$ (i.e. it is not expressed at all). However, this model can be used to investigate properties such as the time taken for a gene whose expression is perturbed to return to equilibrium. This may be important if genes are faced with a noisy external environment, or noise in the expression of other genes in the network [1]. For example, this model has been used to show that negative autoregulation functions as a noise filter, allowing genes to return quickly to their equilibrium expression once they are perturbed. This in turn has been used to explain the abundance of negatively autoregulating genes found in the *Escherichia coli* TFN [11, 106]. This model has also been used to show that feed forward loop (FFL) motifs, which are found at high frequency in the TFNs of both *E. coli* and *S. cerevisiae* [88, 87, 110] can be used to distinguish random fluctuations in the external environment from persistent environmental signals, and therefore also act as a form of noise filter [79, 78]. Just as in the case of boolean networks, the threshold function $f(x)$ can be replaced with a sigma function $\sigma(x)$ in order to allow a greater range of equilibrium expression levels.

A more complex system of ODEs can also be used so that mRNA concentration is explicitly modelled. An example of such a system is given by

$$\begin{aligned} \frac{dm_i(t)}{dt} &= \beta_i^m f\left(\sum_{i=1}^{N} w_{ij} P_j(t) - \theta_i\right) - \gamma_i^m m_i(t) \\ \frac{dP_i(t)}{dt} &= \beta_i^P m_i(t) - \gamma_i^P P_i(t) \end{aligned} \tag{1.4}$$

where β_i^m is the maximum rate of production and γ_i^m is the rate of degradation of mRNA i, β_i^P is the rate of synthesis of protein from mRNA and γ_i^P is the rate of degradation of protein i. Models of this type, which capture the time evolution of both mRNA and protein concentration, have been used to study the properties of networks with complex dynamics, such as noise resistance in genetic oscillators [128]. As in the two previous cases, the threshold function $f(x)$ can be replaced with other functions, such as the sigma function $\sigma(x)$. equation (1.4) can be further modified to include time delays resulting from the time taken for mRNA transcription and translation of mRNA into protein. ODE models allow the dynamics of gene expression in TFNs to be studied in more detail than in boolean network models, as they capture the time evolution of these networks more accurately. However, modelling large TFNs using systems of ODEs can often be computationally expensive. In addition, modelling the evolution of TFNs using ODEs may be difficult, since there are a greater number of parameters (e.g rates of mRNA and protein production and degradation) which may undergo mutation. The rates at which such mutations occur is not always known from empirical studies, and the large parameter space associated with these models often makes an exhaustive exploration of the dynamics of networks modelled in this way unfeasible.

1.2.4 Stochastic Models

Although ODE models are able to capture the time evolution of mean gene expression, they do not capture the stochasticity in gene expression observed in both prokaryotes and eukaryotes. Selection to reduce intrinsic noise in gene expression has been shown to be important in determining the clustering of genes on chromosomes [10], as well as favouring certain network structures such as negative autoregulation [11, 106, 118]. In this section I focus on models which represent transcription regulation as a Markov process. Stochastic models allow the mean and variance in protein number to be calculated explicitly for simple networks, and provide a framework for constructing *in silico* models of larger networks, from which these properties can be measured. In this section I refer to the number, r, of mRNA molecules and number, p, of protein molecules present in the

system at time t (as opposed to the expression level or protein and mRNA concentration, as in the previous sections). The scheme for one of the simplest such models of gene regulation is described below. It is described by the probability distribution $n_{r,p}(t)$ that the system is in a state $\{r, p\}$ at time t. A single, unregulated gene evolves according to Scheme 1:

$$
\begin{aligned}
n_{r,p} &\xrightarrow{\beta_r} n_{r+1,p} \\
n_{r,p} &\xrightarrow{r\beta_P} n_{r,p+1} \\
n_{r,p} &\xrightarrow{r\gamma_r} n_{r-1,p} \\
n_{r,p} &\xrightarrow{p\gamma_P} n_{r,p-1}
\end{aligned}
$$

Scheme 1

where β_r is the probability per unit time at which mRNA molecules are transcribed from DNA, γ_r is the rate of mRNA degradation, β_p is the rate at which mRNA is translated into protein and γ_p is the rate of protein degradation. This scheme can be generalised to a TFN consisting of N genes, indexed by i, with the state of the network given by $\{r_i, p_i\}$ and the rate constants given by β_{r_i}, β_{p_i}, γ_{r_i} and γ_{p_i} [118]. The simplest way to do this is to assume that regulatory interactions in the TFN function such that the rate of production of mRNA at the ith gene depends linearly on the number of proteins of the jth gene present [118], such that

$$\beta_{r_i} = \beta_{r_i}^0 + \sum_{j=1}^{N} w_{ij} p_j \tag{1.5}$$

where $\beta_{r_i}^0$ defines the basic rate of transcription of gene i when no TFs are bound to its promoter, and w_{ij} defines the strength of regulation of gene i by gene j. The first and second moments for the steady state solution of this system can be calculated from simple linear equations, allowing the mean and variance of mRNA and protein concentrations to be calculated directly [118]. This in turn allows the noise in the number of proteins to be calculated for different network architectures.

The assumption that gene regulation effects the rate of mRNA transcription in a linear manner is not realistic in general. However, if the system reaches a static equilibrium, the case in which gene regulation effects the rate of mRNA transcription non-linearly can be dealt with. In this case, at equilibrium, the non-linear function can be well approximated by its lineraization about

the mean protein concentration [118]. Therefore the linear model described above can still be used to calculate the mean and variance in protein concentration for such networks. In cases where the system does not reach a static equilibrium, Scheme 1, and its generalisation to TFNs of many interacting genes, can be used as a framework for simulations of TFNs in which gene regulation is non-linear in the number of proteins present.

The stochastic model presented here considers promoters with only a single state - i.e. it does not consider changes in chromatin structure which lead to changes in the rate of transcription. However it is possible to construct a more general model along the same lines in which the promoter undergoes transitions between an arbitrary number of chromatin states [108]. Once again, the mean and variance in protein number can be calculated for this model, allowing the effects of changes in chromatin structure on gene expression to be explored theoretically.

1.2.5 Statistical Physics Models

The three models above describe the expression levels of genes in terms of the concentration of the TFs which regulate them. They all assume that gene expression follows a threshold function in the concentration of the regulating TFs. However, if we wish to construct models for the evolution of TF binding sites, we must understand how the probability of TF binding changes with the binding site sequence. That is, we must determine how the threshold function which is used to describe transcription regulation changes under *cis* mutations at the regulated gene. All TFs have a probability of binding non-specifically to a region of DNA. However, specific TF binding occurs at *cis* regulatory binding sites, typically of length between 6 and 12 bp, but with an information content equivalent to approximately 6bp (due to some sites being degenerate, such that different nucleotides at the same site give rise to the same binding strength at the binding site), recognised as an optimal binding sequence [150]. The binding affinity of a single species of TF at a target gene with a single specific binding site can be determined using a model derived from statistical physics [40, 18]. In this case the probability that a TF is bound to a binding site is $q(\mu)$ where μ is the chemical potential of TFs in the cytoplasm [40]. $q(\mu)$ is given by the equilibrium thermodynamic expression

$$q(\mu) = \frac{1}{1 + \exp\left[\frac{E-\mu}{k_B T}\right]} \tag{1.6}$$

where E is the binding affinity of a TF to the binding site, T is temperature and k_B is Boltzmann's

constant. The binding affinity of a TF to the binding site can be calculated from the number of mismatches between the binding motif and the optimal binding sequence [40]. In the simplest case, it is assumed that each mismatched nucleotide contributes an amount $k_BT\epsilon$, whilst each correctly matched nucleotide contributes 0 to the binding energy. Therefore we can write $E = k_BT\epsilon r$ where r is the number of mismatches between the real and the optimal binding sequences. The chemical potential μ is found to depend on the number of TFs, N_{TF}, in the cell according to $\mu = k_BT\epsilon_0 + k_BT\ln[N_{TF}]$, where $k_BT\epsilon_0$ is the binding free energy of a single TF to nonspecific binding sites (i.e to the rest of the genome). This allows us to write

$$q(N_{TF}) = \frac{N_{TF}}{N_{TF} + \exp[\epsilon r - \epsilon_0]} \tag{1.7}$$

which is a Michaelis-Menten function, and is frequently used to describe the dynamics of gene activation functions [18]. As the binding site moves closer to the optimal binding sequence (as r decreases), the probability of a TF being bound increases. For a fixed number of TFs in a cell, the probability of a TF being bound is a sigma function in the number of mismatches r. We may view the binding site as undergoing a transition from a non-specific to a specific binding site when the number of mismatches reaches the threshold $r = \frac{\epsilon_0}{\epsilon}$. The case considered here is for a TF regulating a single gene. In the case of global regulators, with many binding sites throughout the genome, the number of TFs, N_{TF}, available to bind to a specific binding site must be appropriately adjusted by the number of possible binding sites [18].

Interactions between TFs, either cooperative or antagonistic, alter the probability of a TF being bound to a promoter. As described above, models of gene regulation frequently use gene activation functions with sharp thresholds. However, the function described in equation (1.8) does not give rise to the type of "on-off" behaviour these models assume. It is possible to produce an activation function with a sharp threshold if, firstly, multiple TFs must be bound to the promoter to activate transcription, and secondly, if those TFs interact cooperatively when binding [18]. For example, if h TFs of the same species must be bound to h binding sites in order to activate transcription, and those TFs interact cooperatively, then the probability that transcription is activated is given by

$$q(N_{TF}) = \frac{N_{TF}^h}{N_{TF}^h + K^h} \tag{1.8}$$

which is a Hill function with coefficient h, where K determines the threshold of activation [18] and depends on binding site strength. Whilst many other activation functions are possible, depending on the degree of cooperatively between TFs and the number of possible binding sites in a promoter [18], equations (1.8) and (1.9) illustrate the general point that increasing the strength of a binding site changes the threshold of activation, whilst changing the number of binding sites alters the steepness of the threshold function.

The model presented here relates changes to single nucleotides in regulatory binding sites to changes in the probability of TF binding, and therefore to changes in the expression of the regulated gene. As such, it illustrates the possibility of modeling transcription networks at an extremely high level of detail. However, such models require a great deal of computational power when applied to large TFNs, in addition to consisting of a very large number of parameters. The models described in this and the preceding sections, capture the dynamics of gene regulation at different levels of detail. The conclusions that can be drawn about gene regulation at one level of detail can be used to inform the assumptions made when constructing less detailed models. In this way, a trade-off between the biological accuracy of a model and its tractability to theoretical or computational exploration can be achieved. In the following section, I describe some of the questions that such models of transcription regulation can be used to address.

1.3 TFN Architecture

The architectural properties of TFNs are of interest for a number of reasons. Firstly, examples of complex networks other than TFNs abound throughout nature. These include protein interaction, metabolic, ecological and social networks. Given the diverse nature of the systems in which these networks are found, it is striking that they nonetheless share certain global statistical properties. In particular such networks tend to be "small world", with a short distance between any two nodes, and highly clustered connections. They also tend to be "scale free" with a few nodes having many connections, and most nodes having only a few connections [8]. These global properties are found in a wide range of systems, which are constructed through different processes. For this reason it has been suggested that these global properties reflect something more general about the architecture of networks in nature. For example, it has been suggested that the scale free degree distribution of networks gives rise to robustness to removal of nodes from the network [7, 8, 55]. Such robustness, the argument goes, is necessary for the continued functioning of all networks in a noisy environment, and therefore is a general property of biological networks regardless of their

specific function.

The second way in which the architecture of TFNs is of interest is through the way they are clustered. TFNs are modular in structure and can be separated into units that function almost independently [53, 58, 102]. Families of TFs with related function can be identified experimentally. In addition modules can be identified from network structure, by identifying how interactions are clustered (e.g [104]). The correspondence between functional and structural modules provides a framework in which to study TFN evolution. A modular structure is a feature common to most biological networks, however the function of particular modules varies, both between different types of networks and between the TFNs of different species. The third way of studying the properties of TFN architecture is by identifying network motifs. Network motifs are recurring patterns of interconnections between small numbers of genes, and have been suggested as the functional building blocks of biological networks [88, 87]. The patterns of motifs observed varies between the type of biological network studied. However, motif patterns are conserved between some highly diverged species, for example between the yeast and *E. coli* TFNs. The functional properties of network motifs can be studied in isolation, and then used to deconstruct the function of larger modules. In the following sections I describe the architectural properties of the yeast TFN, which will be used as the basis for the study of TFN evolution in the following chapters.

1.3.1 Degree Distribution

Abstractly a TFN is a directed graph, consisting of genes with incoming edges, outgoing edges, or both. Where a gene has outgoing edges, it regulates the expression of other genes, and is therefore designated as a TF. Where a gene has only incoming edges, it does not regulate the expression of other genes, and is therefore designated a target gene. The *out* degree distribution of a TFN, $n_{out}(k)$, describes the probability that a TF has k outgoing edges. For both the yeast and *E. coli* TFN, the *out* degree distribution follows a broad tailed distribution that is best described by a power-law for large k:

$$n_{out}(k) \propto k^{-\gamma} \tag{1.9}$$

A broad-tailed distribution indicates that there are a small number of hub TFs that regulate a large number of genes [8]. Interpretation of power-law degree distributions, and the small world structure they confer, has been the focus of a great deal of attention [7, 8, 13, 19, 92, 93, 132]. In particular, it

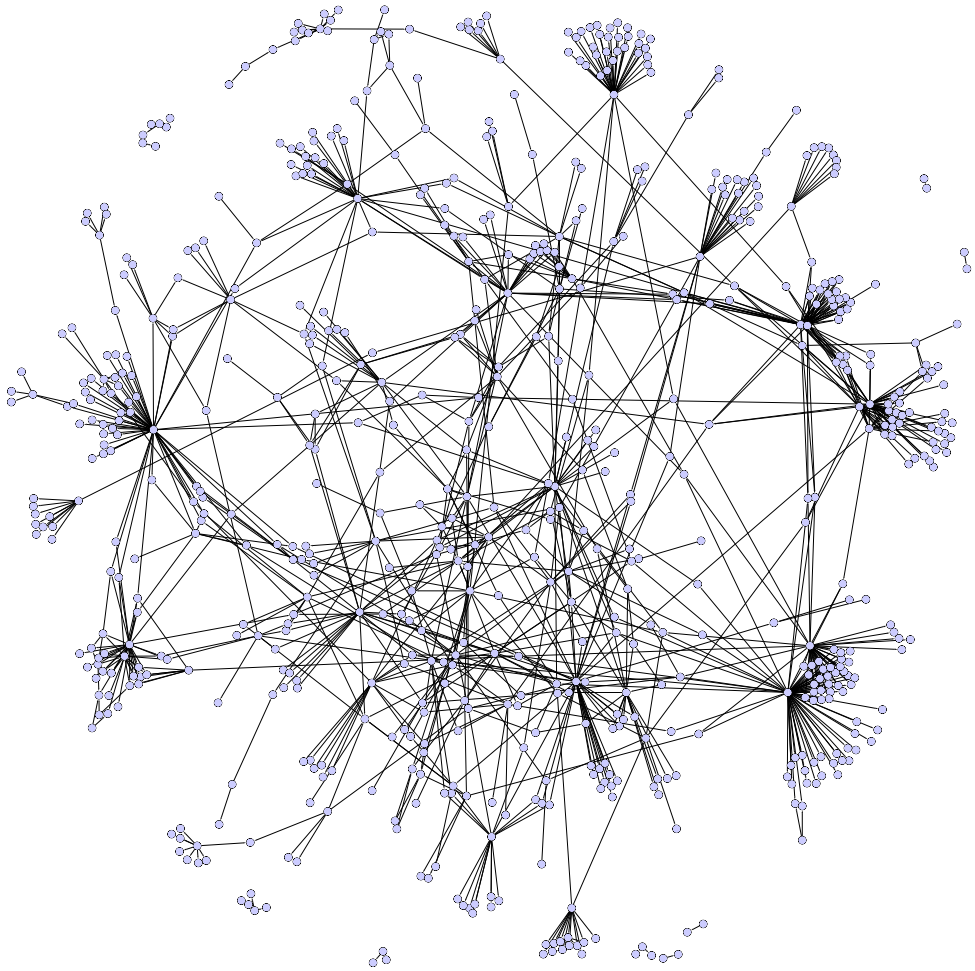

Figure 1.2: Yeast transcription network. Around the edges of the network hub TFs can be seen, in which many targets are co-regulted by a single TF. Data for regulatory interactions is taken from [88].

has been suggested that a power-law distribution may deliver an evolutionary advantage through increased mutational robustness and evolvability [8]. This mutational robustness is said to be conferred through the robustness of networks to the deletion of genes. Evolvability is conferred through the existence of hub TFs, which can simultaneously effect the expression of many genes.

In contrast to the broad tailed *out*-degree distribution, the *in*-degree distribution, $n_{in}(k)$, observed in TFNs is much narrower than a power-law, and has no broad tail. It is best described by an exponential distribution:

$$n_{in}(k) \propto \exp[-\alpha k] \tag{1.10}$$

The exponential *in*-degree distribution reflects the fact that only a few transcription factors combinatorially regulate any one gene. There exist no hub target genes (figure 1.2). For example, in the yeast transcription network, 93 per cent of genes are regulated by less than five transcription factors [47]. Asymmetrical *in*- and *out*-degree distributions are a striking global architectural property of TFNs, and are not observed in other biological networks. It is therefore interesting to consider how the evolutionary process through which TFNs are constructed differs from that of other biological networks. Broad tailed degree distributions in networks are commonly attributed to a process of preferential attachment, in which nodes gain new edges at a rate proportional to the number of edges they already have. Exponential distributions are commonly contributed to random attachment of edges, independent of the degree of a node. The presence of a narrow *in*-degree distribution in TFNs suggests that the evolutionary processes which effect incoming and outgoing edges are different. We can therefore seek to use the degree distribution of TFNs to gain insight into the different ways that *cis* regulatory binding sites and TF function evolve [116].

1.3.2 Modularity

The modular structure of TFNs can be determined empirically either by grouping TFs by function, e.g by their involvement in the same process, such as sex determination, apoptosis or nutrient homeostasis [53]. Alternatively TFs can be clustered according to their position in the network, e.g by identifying TFs which regulate similar sets of targets. Modules identified in these different ways tend to have a high degree of overlap, demonstrating that modularity in network structure reflects functional modularity [53, 58, 102, 104]. Identification of modules based on involvement of TFs in the same cellular function, reveal between 50 and 100 distinct regulatory modules in

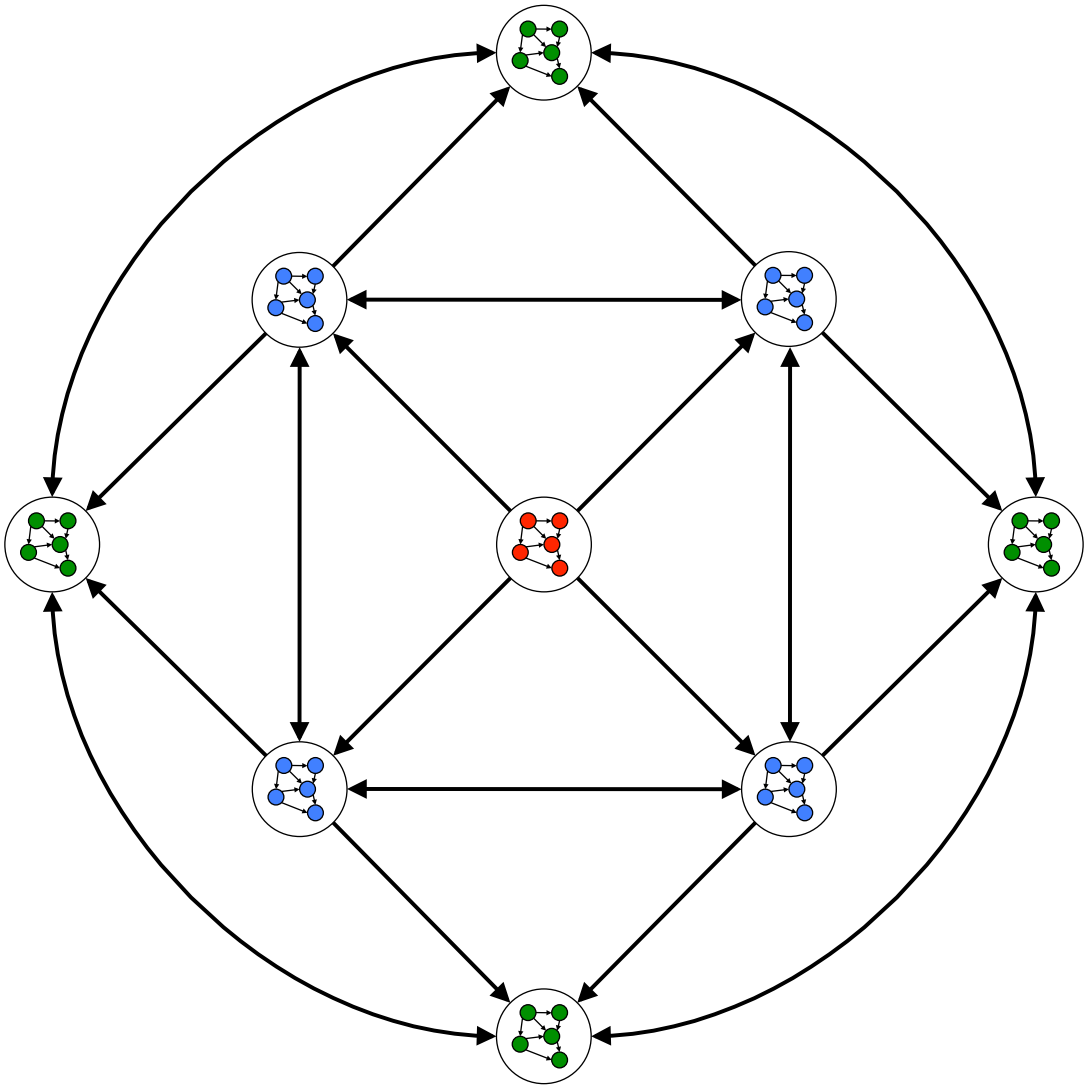

Figure 1.3: Illustration of a nested hierarchy of modules. Each black circle contains a module of TFs involved in related functions. Modules are coloured according to their position in the hierarchy. Modules can regulate other modules at the same position or lower than them in the hierarchy. The red module is at the top, blue modules in the middle and green modules at the bottom of the hierarchy.

the yeast TFN. In order to determine modules from network structure, measures such as the clustering coefficient, which measures the tendency of genes to form "cliques", are used [90]. The clustering coefficient measures the tendency of pairs of genes which both interact with a third gene to interact with each other. Measurements of the clustering coefficient in the yeast TFN are 5 times higher than would be expected from a random network, suggesting a high degree of modularity [47]. However, algorithms which separate TFNs into modules based on their structure reveal few modules which are entirely separate from the rest of the network. Rather, networks are structured as a nested hierarchy of modules, resulting in a highly interconnected network [8, 23]. The picture of the yeast TFN which emerges is of a large number of modules with a high degree of overlap between modules involved in different functions (figure 1.3) [23].

The conditions under which networks with a modular structure will evolve can also be investigated. For example, networks which are asked to perform two distinct signal processing tasks will evolve a non-modular, interconnected structure if the tasks remain fixed throughout evolution. However, if the signal processing task required of the network varies between generations, networks will evolve a modular structure [58]. As such, modularity in TFNs can be seen as reflecting a response to changing environments, as well as reflecting the different functions which sets of genes are involved in.

1.3.3 Network Motifs

Modules themselves can be further deconstructed into functional subunits which perform distinct signal processing tasks. These functional subunits, termed network motifs, are identified by searching for subnetworks consisting of a small number of interconnected genes, which are over represented in real networks, compared to what would be expected from a random network [88, 87]. The smallest network motif, identified in the *E.coli* TFN, is the negative feedback loop. In this network 42 out of 115 (37%) TFs are found to negatively autoregulate [110]. Negative autoregulation is thought to provide both a reduction in intrinsic noise, as well as allowing rapid response to external perturbation [1, 11, 106]. Larger network motifs have also been identified. If autoregulatory interactions are ignored, there are 13 possible interconnected subnetworks consisting of 3 genes (figure 1.4). In both yeast and *E. coli*, only one of these, the feed-foward loop (FFL - figure 1.4, number 5), is over represented. The functional properties of FFLs have been extensively investigated [1, 41, 77, 79, 78, 137].

A combination of modelling, simulation and experimental investigations [79, 78, 137] reveal

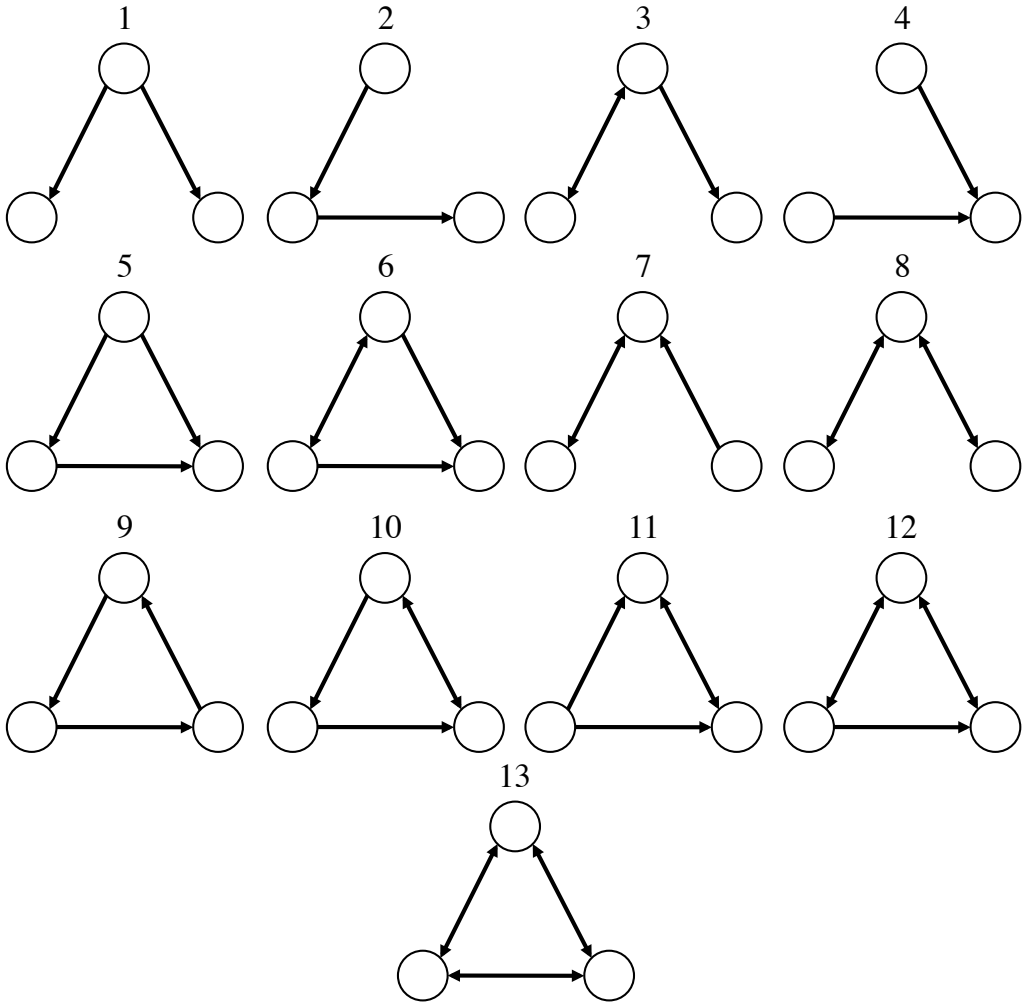

Figure 1.4: Three node motifs. Only the feed forward loop (number 5) is over represented in the TFNs of yeast and *E. coli*, suggesting it has functional importance

that FFLs may perform a number of functions, depending on the combination of activating and repressing regulatory interactions that make them up. Most significantly, certain kinds of FFL which occur commonly in observed TFNs have been shown to act as a filter for transient environmental signals. FFLs are found to be robust to brief changes , but to respond well to sustained changes in environmental signals. Similarly, larger motifs have been identified, and found to display a range of expression dynamics (e.g the four node bi-fan motif [88, 87, 139]). The evolution of network motifs is of interest, since there are several possible ways of explaining their over representation in observed TFNs. Firstly, they may recur because they represent optimal solutions to signal processing problems. Secondly, they may recur because they are easily evolvable from available network mutations. In particular the possibility that observed network motifs are particularly easy to evolve through gene duplication has been investigated. However, analysis of the genes involved in motifs shows that duplicate genes are no more likely to occur as part of a motif than would be expected due to chance. Finally, network motifs may be explained as a by-product of the mutational process. In this case, motifs do not arise because they perform any particular function, but arise as a product of neutral evolution or mutational robustness [22, 72]. For example, it has been shown that FFL motifs can arise at high frequency as a result of recombination [72].

In order to address the questions which arise concerning TFN architecture, an understanding of the mutational process through which TFNs evolve is necessary. In the next section I discuss the types of mutations that arise in TFNs, and the roles they play in shaping TFN architecture.

1.4 Mechanisms of TFN Evolution

There are four types of mutation which give rise to changes in the structure of TFNs. These are mutations at regulatory binding sites in the promoter regions of regulated genes (*cis* mutation), mutations which affect the function of a TF at the gene which codes for it (*trans* mutation), duplication and deletion of genes (either TFs or TGs). Each type of mutation alters the structure of the TFN in a different way. It is not sufficient to study only the rates at which different types of mutation occur to understand the role of these different mutations in the evolution of TFNs. Rather, it is the mutations which arise in a population and then become fixed which determine the course of TFN evolution. The relevance of this is most clearly illustrated by gene duplications which become fixed in a TFN. A duplicate gene will initially result in all the regulatory interactions of that gene (both incoming and outgoing edges) being copied. However, many duplicates will initially be redundant and rapidly fix new mutations. This in turn will result in many dupli-

cates becoming pseudogenes which are subsequently lost from the population. Duplicates which result in pseudogenes do not contribute to the evolution of TFN architecture, even if they occur frequently. Where duplicates remain in the population, they will often undergo neofunctionalization. Neofunctionalization is, in turn, often associated with changes to the regulatory interactions participated in by the gene following duplication. Therefore duplicates may not become fixed in a population along with copies of all the interactions of their parent, but undergo gain of new interactions and loss of existing interactions. Less drastically, *cis* mutations may frequently occur which result in the loss of a regulatory binding site. However, these may be compensated for by protein-protein interactions between TFs which result in cooperative binding to targets [125, 126]. Therefore, although *cis* mutations may be occurring frequently, they may not be changing the architecture of the network. This illustrates that measurements of sequence evolution at *cis* and *trans* or rates of duplication and deletion, do not provide an accurate picture of the contribution of these precesses to the evolution of TFN structure. In this section I discuss the types of mutations which are seen to arise and become fixed in the evolution of yeast TFNs.

1.4.1 Evolution of Regulatory Binding Sites

The evolution of regulatory binding sites occurs through mutations which alter the binding strength of a *cis* regulatory site for the transcription factor(s) which bind to it. A point mutation at a binding site which brings the site closer to or further from the optimal binding sequence will either increase or decrease its binding strength. As described above, the binding strength of a single site follows a sigma function in the number of mismatches between the real site and the optimal site (equation (1.8)). As such, a single point mutation can result in a site going from specific to non-specific or vice versa [40, 89]. In addition, insertions or deletions in the promoter region of a gene may result in a binding site being lost or gained. In both cases, it is reasonable to model regulatory binding sites being lost or gained through a single mutation. Turnover, in which loss of a binding site for one TF is compensated for by gain of a binding site for another TF, may also occur. In order to understand how regulatory binding sites evolve, it is necessary to asses the evolutionary constraint they are under, and the rate at which they are lost, gained or undergo turnover.

It is also possible to ask how the rate of evolution of regulatory binding sites varies from gene to gene. For example, neofunctionalization of duplicate genes suggests that they will undergo a faster rate of *cis* evolution than other genes. More generally, the rate of regulatory evolution at a gene may depend on its position in the network, for example genes with more regulatory interactions

may undergo a different rate of *cis* evolution than those with only a few interactions. In general, the rate of *cis* evolution may vary with *in* degree, *out* degree, or both.

Studies of the *S. cerevisiae* genome suggest that only $2-3\%$ of the combined promoter regions of genes is covered by TF binding sites under strong selection, whilst as much as 30% is covered by binding sites under weak selection [98]. Comparison of *S. cerevisiae* with two closely related yeast species, *S. paradoxus* and *S. mikatae*, reveals that, of approximately 20000 identified TF binding sites, 80% are conserved in all three species, 5% are semi-conserved, whilst the remaining 15% are lost in at least one species. In addition, approximately half of the observed loss events are the result of turnover [29]. Thus regulatory binding sites under strong selection cover only a small fraction of yeast promoter regions, but there is significant change to these binding sites even between closely related species, with turnover accounting for many of the observed changes.

The question of variation in the rate of *cis* evolution between genes has also been addressed in yeast [121, 45]. Following gene duplication the expression of a pair of duplicate genes diverges at a rate approximately 10 times the rate of divergence between ancient duplicate pairs [45]. This suggests an accelerated rate of *cis* evolution following a duplication event, and an increased number of evolutionary events changing regulatory interactions at duplicate pairs is indeed observed [45]. The question of variation in the rate of *cis* evolution with the position of a gene in the TFN is harder to address empirically. However, if turnover is common it is likely that genes with more interactions will gain new interactions and lose existing interactions at a faster rate than genes with fewer interactions [116].

The resulting picture of *cis* evolution which emerges is one in which regulatory binding sites are gained and lost frequently. The rate of *cis* evolution is seen to depend on the evolutionary history of a gene, particularly on whether it has undergone a recent duplication. In addition it is likely that the rate of *cis* evolution varies with the position of a gene in the network.

1.4.2 Evolution of TF Function

The relative contributions of changes to regulatory binding sites and changes to TF function in the evolution of TFNs has been the subject of growing debate [16, 43, 70, 73, 97, 96, 115, 126, 125, 136, 142]. It has been argued that, since changes to a TF will affect multiple regulatory interactions and therefore the expression of multiple genes simultaneously, they will tend to have pleiotropic effects. In contrast, changes to regulatory binding sites will affect only the expression of one or a few genes, and therefore will tend to minimize pleiotropic effects [73]. As a result, *cis* regulatory

changes are seen as less likely to be deleterious than changes to TF function (*trans* mutations). This leads to the expectation that the majority of regulatory evolution is through changes at *cis*.

This view has been challenged both on an empirical and a theoretical basis [73, 126, 125]. TFs function through protein-DNA and protein-protein interactions. These interactions are usually thought to be mediated by domain-domain contacts and the secondary structure motifs of the protein [73]. However, there is growing evidence that many protein-protein interactions are in fact mediated by short linear motifs (SLiMs), which consist of 3-10 amino acids, with an information content equivalent to only 2 or 3 amino acids (due to different amino acids at some sites giving rise to a functional SLiM) [73]. SLiMs also tend to lie in regions of the protein free from structural constraints, which allows them to evolve independently of the rest of the protein. Protein-protein interactions mediated by SLiMs allow the possibility of evolving each interaction independently, thus greatly reducing the negative pleiotropic effects associated with changes to domain architecture. This view is supported by patterns of conservation in TFs across eukaryotes, where it is found that, whilst domain architecture is strongly conserved, SLiMs are poorly conserved between lineages.

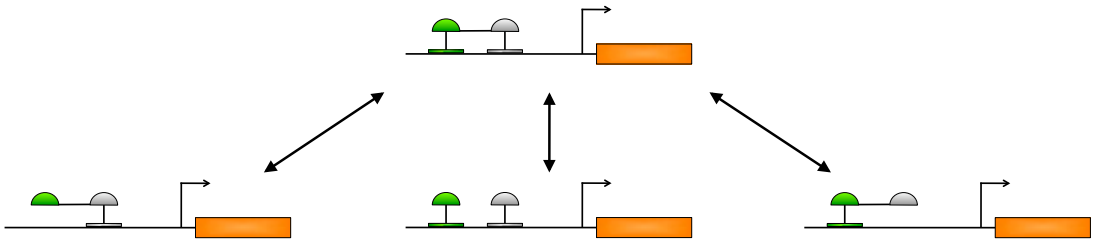

Figure 1.5: Interaction between *cis* and *trans* mutations. A black line connecting two TFs (green and white semicircles) indicates a protein protein interaction between them. Black arrows between genes indicate possible mutations. When a protein-protein interaction is present, cooperative binding can compensate for loss of a TF binding site. A *trans* mutation leading to the gain of a protein-protein interaction can allow a target gene to move between different patterns of *cis* regulation. Thus *cis* and *trans* mutations can be seen to interact with one another to produce the same pattern of regulation in different ways.

The view that TFs evolve primarily through changes to protein-protein interactions has interesting consequences. In particular it suggests that mutations at *cis* and at *trans* will tend to interact strongly. This is because if a TF requires a protein-protein interaction to bind to a target, changes to the binding site of one TF will affect the ability of the other to regulate the target. Similarly, gain of a protein-protein interaction resulting in cooperative binding between two TFs may relax the evolutionary constraint on their associated binding sites (figure 1.5). The resulting

interplay of *cis* and *trans* mutations can give rise to complex dynamics in the evolution of TFN architecture.

1.4.3 Gene Duplication

In addition to *cis* and *trans* mutations, gene duplication also plays an important role in the evolution of TFNs. Most significantly, gene duplication provides new genetic material on which natural selection can act. A duplicate provides a new gene which can be adapted to new functions, where the original copy of the gene may be constrained by its existing function. This process is known as neofunctionalization. In addition, duplicate genes may take on some of the functions of the parent gene. In this case, the functions of the original gene are split between a duplicate pair, which may allow further optimization of those functions, which was not possible with only one copy of the gene. This process is known as subfunctionalization. Both of these processes have been found to be important in the evolution of TFNs [120]. More generally, all gene families which make up the TFNs of existing organisms are thought to have expanded through duplication events [34, 62, 66]. As such, gene duplication plays a central role in the evolution of TFNs.

Gene duplication can occur in a number of forms. Most common are tandem duplications, in which some or all of a gene (its promoter and coding region) are duplicated along side the original gene. In such cases, the duplicate gene will tend to be under the same regulation as its parent. As a result, the effect on network structure is that each of the regulatory interactions of the parent gene will initially be copied. However, as described above, the expression of duplicate genes diverges rapidly as they undergo neo- or subfunctionalization, or become pseudo genes and are lost from the genome. Duplications can occur on a larger scale, with sets of genes being duplicated together. The largest possible duplication event is a whole genome duplication (WGD). WGD events have occured a number of times during the evolution of eukaryotes [34, 66]. A recent WGD in the evolution of *S. cerevisiae* has allowed the structure of TFNs before and after the event to be investigated. A number of studies [17, 26, 46, 120] have indicated that the fate of duplicates resulting from a WGD may be different from those resulting from small scale duplication events. In particular, it was found that duplicates resulting from the yeast WGD tended to undergo greater expression divergence than duplicates arising from small scale events [120]. This in turn suggests that regulatory divergence (i.e *cis* evolution) plays a more significant role in the fate of duplicates resulting from large scale events than those resulting from small scale events. A further hypothesis is that expression divergence between duplicate pairs indicates neofunctionalization

whilst conserved expression indicates subfunctionalization [120].

From the point of view of the evolution of network structure, gene duplication has two important effects. Firstly, it causes networks to grow. Secondly, it causes genes to gain new interactions at a rate proportional to the number of interactions they already have. Models of network evolution through duplication have been extensively studied and shown to be able to reproduce both the global statistical properties observed in biological networks, as well as certain network motifs [13, 19, 44, 93]. However, the rate of gene duplication observed in yeast is low, and as a result the effects of rewiring of the network through *cis* and *trans* mutation will overwhelm the effects of gene duplication in shaping network structure. In addition, analysis of observed network motifs shows that they do not, in general, contain pairs of duplicate genes. Therefore, although gene duplication is undoubtedly an important process in the evolution of TFNs, it does not in itself account for the structure of the networks observed in existing organisms.

1.4.4 Gene Deletion

The final process through which TFNs evolve is gene deletion. Loss of a gene from a network alters the network structure, since it reduces the size of the network and removes all the regulatory interactions which the lost gene participated in. As has been mentioned already , duplicate genes frequently lose all function to become pseudogenes, before being lost from the network altogether. In addition, a duplicate gene may render other genes in the network redundant, allowing them to be deleted. While estimates of the rate of gene deletion are difficult, it is thought to be similar to the rate of gene duplication [38]. Experiments in yeast indicate that a high proportion of genes are non-essential [84], indicating that gene deletion events may be fairly common. However, the apparent redundancy of genes in knockout experiments often does not take account of variation in environmental conditions, in which genes which are redundant in one environment, may be necessary in another.

Although there is a growing body of empirical evidence on the roles of different mutations in TFN evolution, accurate estimates of evolutionary rates, and variation in evolutionary rates between genes, are not available in most cases. However, the evolutionary mechanisms outlined in this section can be used to construct models of TFNs. In the next section I discuss the selection schemes which can be used with these models, and the role they play in shaping TFN structure.

1.5 Selection in TFNs

Natural selection shapes the evolution of TFNs in a number of ways. Firstly, networks are adapted to perform a particular function or set of functions. As such, selection on gene expression plays an important role. Secondly, networks are subject to several sources of noise - intrinsic noise within a cell, variation between cells and variation in the external environment. TFNs must be robust to all of these sources of noise in order to function properly. Mutational robustness may also play a role in TFN evolution, however the competing needs to be robust to deleterious mutations whilst being able to adapt to changing environments makes this form of robustness qualitatively different from the other forms described. Finally, neutral evolution may also play a significant role in shaping TFNs. Neutral changes to regulatory interactions result in network structure being strongly influenced by the relative rates at which different kinds of mutations occur. Determining the contributions of these different evolutionary processes is one of the most important challenges in the study of TFNs, and in the study of biological networks generally. In this section I describe the different ways in which they affect TFN structure.

1.5.1 Gene Expression

TFNs must be adapted so that the correct genes are expressed at the correct time in order to perform their function. This may take the form of different subsets of genes being activated under different environmental conditions [53], as in the case of genes which maintain nutrient homeostasis or trigger the haploid phase in *S. cerevisiae*. Alternatively it may require that gene expression follows a particular time course, as in the case of developmental processes or circadian oscillators. In general, not all genes in a TFN may be under strong selection for a particular expression pattern. Since it is the output of the network upon which selection acts, upstream TFs may have noisy expression provided this is compensated for by other genes in the network.

The study of TFNs *in silico* has often focused on selection for optimal gene expression (e.g [112]). One method of doing this is to generate an artificial network, and define the resulting expression patterns of the genes to be optimal. The network is then allowed to evolve with selection acting to maintain optimal gene expression. This serves as a way of removing adaptive evolution from the system, so that other forms of selection can be studied. Other studies require artificial networks to evolve to perform a particular function (e.g [58]). Under a sufficiently realistic mutational process, this allows the solutions adopted by the evolutionary process to be studied and compared to real networks. This allows the conditions under which network properties, such as modularity

or a particular pattern of network motifs, arise to be studied. An alternative approach is to construct an *in silico* model of real a network with known function, and compare the properties of that network to all other networks with the same function (e.g [75]). This gives insight into the particular properties of the network that has been evolved. For example, this method was used to show that the *Drosphila* sex determination network displays a high level of parsimony [75]. Finally *in silico* models of large networks with known function can be constructed in order to determine the relative importance of the various subnetworks (e.g via virtual gene knockout experiments). For example, the developmental network of the sea urchin has been studied extensively in this way [25].

A more general approach to the evolution of optimal gene expression is to determine subnetworks which act as circuit elements. These circuit elements perform particular functions, e.g AND gates, OR gates or bistable switches. By connecting together circuit elements a larger network with a particular function can be constructed. The identification of circuit elements has generally been through the investigation of the functional properties of network motifs.

In general the way in which selection for optimal gene expression affects network structure will depend on the function which the network is being selected for. As indicated above, this means that the study of optimal gene expression is often a task of reverse engineering - i.e given a network, can we deconstruct it to determine which subnetwork performs which function? However, where there is a large set different networks, all of which perform the same function, we can ask which network from within that set evolution will adopt. The question then becomes, how do networks evolve in a neutral space of functionally equivalent network? This depends strongly on the structure of that neutral space. As such, an important challenge is to define the structure of the neutral space of networks with particular functions of interest. This may allow us to define general rules for how the structures of TFNs evolve.

The role of noise in gene expression has become the focus of intense research in recent years [11, 51, 69, 100, 101, 106, 118, 128]. Noise in gene expression can be viewed as a hinderance, which disrupts the proper functioning of a gene. Alternatively it can be viewed as a source of variation which cells can exploit to there advantage. Where noise disrupts gene function, mechanisms that reduce the level of noise in gene expression are selected for. Examples of noise reduction mechanisms include negative autoregulation [11, 118, 106], feed forward loops [9, 41, 78] and the clustering of coexpressed genes together on the genome [10].

Noise can be advantageous to an organism in a number of ways. Underlying most cases in which

variability in gene expression can be seen as advantageous, is the ability of cells to stochastically switch between a number of stable gene expression states [51, 101]. This produces phenotypic variation in a population. In microbial populations such stochastic switching between gene expression states speeds the response time of an individual (and population) to changes in the environment, such as changes in the abundance of available nutrients [51, 101]. In multicellular organisms, stochastic gene expression can also be used in cell-fate decision during development. For example in *Drosophila* eye development, whether cells become a blue-sensitive or yellow-sensitive photoreceptor is determined stochastically. This results in the desired distribution of photoreceptors across the eye, without the need for a complex regulatory architecture to specify the fate of each individual cell [141]. By using naturally occurring stochastic variation in gene expression, cells can by-pass the need to evolve overly elaborate gene networks to perform complex functions.

1.5.2 Robustness

TFNs are found empirically to display a high level of robustness both to environmental fluctuations and to mutations. However, the extent to which these properties are the product of direct selection is not obvious. Environmental robustness can take several forms. As described above, the process of transcription and translation is inherently noisy. This noise can be reduced via a number of mechanisms, including through negative autoregulation by TFs [1, 11, 106, 118]. Similarly, noise in the expression of upstream TFs, or in environmental signals external to the cell, can be produced by adopting network structures which filter noise. Empirically, it seems that reduction of this type of environmental noise is selected for. It has also been shown that noise can be used to advantage, by producing variation in gene expression within a population [101]. Variation in gene expression is maintained in a population as it allows the population to adapt to changes in the environment. Similarly, noise in bistable systems which allows the network to switch stochastically between states, and can provide a mechanism for an individual cell to sense its external environment [51].

In many cases, robustness to environmental noise may be regarded a part of the function of a TFN. The mechanisms by which such noise is dealt with are interesting in themselves. However, it has also been suggested that there is a relationship between environmental and mutational robustness. This relationship, known as the congruence hypothesis [85, 135], states that robustness to one kind of perturbation also results in robustness to other kinds of perturbation. Thus, environmental robustness, which tends to be strongly selected for, also conveys mutational robustness, which is weakly selected for [127]. Whilst it is certainly true that the congruence hypothesis holds

in some cases, it also possible to find examples in which it does not [82]. Thus, where mutational robustness is observed, it still remains to be established whether this is the result of direct selection, or a by-product of selection for other kinds of robustness.

The issue of mutational robustness is further complicated because it seems to act in opposition to the idea that TFNs need to be able to adapt to new environments. If mutations do not produce new patterns of gene expression, this means the network cannot adapt to changes in the environment. This is clearly not desirable in many cases. Thus there is a trade-off between mutational robustness and adaptability. This issue has been addressed extensively by studying RNA [20, 27, 37, 133], where an explicit genotype-phenotype map is constructed, with RNA sequence providing the genotype and RNA secondary structure providing the phenotype. The neutral space of all genotypes that map onto the same phenotype can then be constructed. From this it is possible to show that the larger the neutral space, the greater the number of different phenotypes which border the neutral space. Since these neighbouring phenotypes are all accessible to a population through a series of neutral mutations, the tension between mutational robustness and adaptability can be resolved. It is highly plausible that a similar argument holds for the structure of TFNs. However, the number of phenotypes accessible from a set of neutral genotypes depends on the structure of the space of neutral genotypes. Therefore the necessity of exploring the neutral genotype space associated with TFNs becomes clear.

The causes and consequences of mutational robustness remain contentions. There are strong theoretical arguments that organisms will evolve genotypes such that the number of strongly deleterious, or lethal, mutations they are subject to is minimised [127]. Such mutational robustness will arise when the product of the population size M and mutation rate μ is large, $M\mu \gg 1$. However, it is less clear whether robustness to mutations which are weakly deleterious is directly selected for in the same way. The situation is further complicated when sexually reproducing organisms are considered. On the one hand recombination is capable of reinforcing selection for mutational robustness [72]. On the other hand, complex dominance effects arise in diploid gene networks which can reduce the effects of deleterious mutations and render mutational robustness less important [70, 91]. Unravelling the effects of selection strength, mutation rate, recombination and dominance on the evolution of mutational robustness remains an important challenge for the development of a proper understanding of the evolution of gene networks.

The interplay between mutational robustness and adaptation is also a developing field. As described above, mutational robustness can facilitate adaptation by opening up a wide range of

alternative phenotypes to a population. As such, mutational robustness can be seen as increasing the adaptability of a population when faced with new environments. However, it is not clear whether selection for adaptability occurs directly. In this respect, one particularly interesting potential mechanism for the evolution of adaptability is evolutionary capacitance. Evolutionary capacitance occurs when a build up of genetic variation occurs in a population, the effects of which are suppressed by an "evolutionary capacitor" [12, 30, 42, 71, 81, 109, 114, 123]. The effects of this cryptic genetic variation can be revealed either due to changes in the environment, or due to mutation at the gene which acts as an evolutionary capacitor. For example the Heat Shock Protein, Hsp90, has been shown to reveal cryptic genetic variation in *Drosophila* when they are subject to increased in temperatures [109, 114]. Similarly, the Yeast prion [Psi+] reveals genetic variation by allowing read through of stop codons [123]. In order for selection to favour evolutionary capacitors as a mechanism for facilitating adaptation, it is necessary that populations be subject to frequent environmental changes which require adaptation. For a given rate of environmental change, θ, an evolutionary capacitor can invade and be maintained in a population provided the population size, M, is greater then a minimum M_{min}. M_{min} is typically quite small, and grows weakly as $M_{min} \propto \theta^{-\frac{1}{2}}$ [81], suggesting that evolutionary capacitors can be selected for their ability to adapt to new environments.

1.5.3 Neutral Evolution

The possibility of neutral evolution in TFNs is well known. As discussed above, such neutral evolution takes place in a space of neutral genotypes, and the course of that evolution depends on the structure of the space. Neutral evolution in TFNs is likely to be a more complex process to understand than in other systems, such as junk DNA or in the case of RNA structure described above. In these cases neutral evolution through single nucleotide substitutions can be studied - i.e only one type of mutation need be considered. However, in TFNs several qualitatively different types of mutation can occur. These different mutations may interact, as illustrated in figure 1.5 for *cis* and *trans* mutations. As a result defining the neutral space of TFNs may be difficult.

Neutral evolution may also play a role in determining such TFN structural properties as motif distribution [72]. For example, given a neutral space in which different three node motifs can be adopted, the frequency with which FFLs occur depends on the relative rate of gain and loss of regulatory interactions, and on the rate of recombination [72]. In particular it is found that FFL frequency follows a peaked distribution in the rate of recombination. Studies such as this highlight

the importance of considering non-adaptive explanations of network structure. The process of mutation combined with the structure of the neutral genotype space can give rise to biases in network structure which have nothing to do with adaptation. Just because a particular network structure is common does not mean it is functionally important.

The different types of selection discussed here all play a role in shaping network structure. However, just as important are the population genetic details of the system being considered. The importance of recombination rate has already been mentioned. In the next section I discuss the role of this, and other population genetic factors, in more detail.

1.6 Population Genetics and TFN Evolution

Population genetic factors play an important role in shaping the structure of TFNs. Factors such as mutation rate, population size and the strength of selection determine the extent to which populations evolve mutational robustness. In addition, since TFNs encode regulatory interactions between genes, they also naturally encode epistatic interactions between loci. As a result, the effects of multiple mutations are rarely additive in TFNs. The roles of recombination and dominance in diploid, sexual populations are also significant, and tend to have very different effects with respect to *cis* and *trans* mutations. Therefore, population genetic factors can be seen to influence both the structure adopted by TFNs, and the types of mutations through which a TFN can evolve. In this section I discuss these effects in more detail.

1.6.1 Population Size

The effect of population size on evolution is significant. For example, small populations are heavily influenced by genetic drift, whilst very large populations are not. In the case of TFNs, population size is of particular interest for the role it plays in neutral evolution and mutational robustness. When evolving on a space of neutral genotypes, the most important factor is the product of population size, M, and mutation rate, μ. If $M\mu \ll 1$, the population at any point in time is monomorphic for a single genotype [127]. If a mutation arises, it is either completely lost, or fixed by all members of the population. Therefore the whole population may be thought of as moving from point to point in the space of neutral genotypes. If the population encounters a deleterious mutation, it will not become fixed, provided the product of the fitness penalty s and the population size is sufficiently large, $Ms \gg 1$. Thus the population may be thought of as moving from point to

point in neutral genotype space, but never leaving it. The probability that the population adopts any one genotype is simply the stationary distribution for the Markov process which describes a random walk on the space of neutral genotypes. This is in turn determined by the rates at which different kinds of network mutations occur. Mutational robustness does not influence the evolution of the network in this case, only the structure of the space of neutral genotypes and the relative rates of different network mutations determines TFN structure.

However, for large populations, in which $M\mu \gg 1$, mutational robustness plays a role in shaping network evolution. In this case the population is not located at a point in neutral genotype space, but is distributed over a number of genotypes. That is, the population contains a number of different genotypes. When the population is at the edge of neutral genotype space, a certain fraction of the population will suffer a deleterious mutation at each generation. This has the effect that the more deleterious neighbours a genotype has, the more it is disadvantaged. As a result, the population tends to move to genotypes which neighbour fewer deleterious genotypes, i.e which are mutationally robust. This effect is independent of the strength of selection against deleterious mutants (provided the selective disadvantage, s, satisfies $Ms \gg 1$), and of population size (provided $M\mu \gg 1$). In TFNs, where different types of mutation can occur at different rates, it is the mean rate of deleterious mutations which determines the degree of mutational robustness of a genotype. Therefore the structure adopted by the network is determined by the structure of neutral genotype space and the relative rates of different types of network mutation - It is the mutational robustness of the genotypes in a region of neutral space which determines whether the population lies in that space.

1.6.2 Sexual Reproduction

The structure and evolution of TFNs is influenced by whether or not the organism considered is sexual or asexual. From the point of view of TFNs, the most significant difference between these two cases is that sexual populations undergo recombination, whilst asexual populations do not. The role of recombination in shaping TFN structure has not been fully characterized. However, it is known that recombination can affect the degree to which populations develop mutational robustness by bringing together different mutations in the same genome [3, 72, 74]. In relation to this, it is interesting to consider how the structure of TFNs gives rise to epistatic interactions between mutations. Directional epistasis occurs when the effect of a mutations on fitness changes in the presence of other mutations in the genome. This can be synergistic, in which case the

average effects of successive mutations becomes more harmful, or else it can be antagonistic, in which case successive mutations become less harmful.

Epistasis is of particular interest when considering the evolution and maintenance of sexual reproduction [3, 74]. The mutational deterministic hypothesis states that sex enhances the ability of natural selection to purge deleterious mutations after they are brought together by recombination. This suggests that synergistic epistasis is required in order for sexual reproduction to be maintained. Studies of the *in silico* evolution of TFNs have shown that recombination in sexual populations favours network structures which show mutational robustness. This in turn results in synergistic epistasis. As a result, sexual reproduction favours network structures which display synergistic epistasis, and thereby favours its own maintenance [3]. However, these conclusions have been challenged by another *in silico* study of TFNs which finds that when reproductive mode and epistasis are allowed to co-evolve, asexual populations out-compete sexual populations [74].

In the context of TFNs, it is easy to see that epistatic interactions between mutations are likely to be the rule. For example the prevalence of multiple pathways linking TFs at the top of a regulatory cascade to target genes at the bottom [134] is likely to result in synergistic epistasis. This is because the redundancy resulting from the existence of multiple pathways means that breaking one of those pathways will have little deleterious effect on the function of the network. However, breaking of subsequent pathways is likely to be increasingly deleterious, as the number of intact pathways completing the regulatory cascade decreases. A similar effect can be seen for deletion of genes in the network. As has been observed previously, a scale free network structure tends to make networks robust to random deletion of nodes. However, as more and more nodes are deleted, the network will eventually fracture into disconnected subnetworks. Therefore deletion of a single gene from a TFN may have only a small deleterious effect on its function. As more and more genes are deleted, the deleterious effect is likely to increase rapidly.

In a more general sense, it can be seen that mutational robustness will tend to result in negative epistasis. This is because mutational robustness reduces the deleterious effect of mutations. However, if multiple mutations occur, that robustness will eventually be lost and, and they will become increasingly deleterious - the previously hidden intrinsic genetic load will become increasingly exposed.. The hypothesis that recombination favours mutational robustness and synergistic epistasis in TFNs leads to the expectation of clear structural differences between the TFNs present in sexual and asexual populations. As a result, sex is likely to significantly influence the structure of TFNs

1.6.3 Dominance

The final factor which may influence the structure of TFNs is whether the network considered is haploid or diploid. This is because of the dominance effects which may arise in diploid organisms, and affect the ability of the network to fix certain kinds of mutations. The clearest example of this is provided by the different effects of *cis* and *trans* mutations on gene expression. In a diploid organism, if a gene suffers a *cis* mutation, resulting in the loss of a TF binding site, it becomes heterozygous at the promoter region of the gene. Thus, instead of two copies of the gene being regulated by a TF, only one copy is regulated. It is easy to see the the effects of this on gene expression will tend to be additive. Indeed, *cis* mutations are found to show a high degree of additivity in their effect on gene expression in *Drosohila* [70]. If a TF suffers a *trans* mutation in a diploid organism, this results in one copy of the TF having one set of targets, and the other TF having a different set of targets. As a result, both sets of targets are regulated by one copy of the TF. This tends to result in deviations from additivity in gene expression, i.e dominance. Again, dominance in the effects of *trans* mutations is observed in *Drosophila* populations [70].

Dominance may slow or accelerate adaptive evolution. In the case of over-dominace it may favour hetrozygotes, and therefore result in heterogeneity in the structure of TFNs in a population. In the case of under-dominace it may provide a barrier to the fixation of adaptive mutations. The study of the effects of dominance in *Drosophila* TFNs suggests that additivity in *cis* mutations tends to favour them as a mechanism for adaptive evolution. As a result variation in TFNs between closely related species tends to lie at *cis*. In contrast, the greater scope for the maintenance of recessive deleterious mutations in a population means that variation within a population tends to be at *trans*. The roles of *cis* and *trans* mutations are therefore further differentiated in diploid populations as compared to haploids.

1.7 Research Aims

The aims of the research presented in this thesis are to elucidate the principles by which TFNs are constructed by natural selection. The approach taken is to construct models of TFN evolution which include an accurate representation of the mutational processes through which they evolve. Different models of TF binding to regulatory binding sites are used to model mean gene expression and noise in gene expression at different levels of detail. In addition a variety of population genetic scenarios are considered in order to elucidate their role in shaping TFN structure. A number

of different structural properties of TFNs are considered. The asymmetrical degree distribution observed in yeast and *E. coli* is used as a starting point, to elucidate the roles of the different mutational processes in TFN evolution. The interactions between *cis* and *trans* mutations are then investigated in more detail. In particular, the influence of these mutations on the structure of the neutral genotype space associated with TFNs is investigated. This is used to draw a number of general conclusions about the neutral evolution of TFNs. Finally the evolution of the most basic network motif - autoregulation - is investigated. This is used to illustrate how the population genetic details of the organism considered strongly influence the evolvability of even this most basic network motif. An investigation of the fate of duplicates of autoregulatory genes is also made. This is used to elucidate why duplicate pairs of genes do not tend to form network motifs.

In all cases the models that are constructed are based on observations about the yeast and *E. coli* TFNs. The conclusions which are drawn from these models are used to explain both the similarities and differences in the structural properties of these networks. A combination of analytical results and evolutionary simulations are used to draw conclusions about the mechanisms through which TFNs evolve.

1.8 Thesis Outline

Chapter 2 is an investigation into the evolutionary process that gives rise to the asymmetrical degree distributions observed in TFNs. Conclusions are drawn about the relative contributions of different types of mutation to TFN evolution, as well as the way that evolutionary rates vary with the position of a gene in the network.

Chapter 3 concerns the interaction of *cis* and *trans* mutations in the neutral evolution of gene regulation. A space of neutral genotypes is explicitly constructed and the evolution of a population in this space is described. The results are then applied to explain the observed large scale neutral rewiring of the yeast sex determination network.

Chapter 4 concerns the evolution of negative autoregulation in diploids. Whilst negative autoregulation is frequent in *E. coli*, it is relatively rare in *S. cerevisiae*. This is explained through an increase in the noise in gene expression in diploids which are heterozygous in the binding site strength for their own gene product. This results in under-dominace and a barrier to the evolution of negative autoregulation. The results are also used to consider the fate of duplicates of autoregulating genes in haploids.

Chapter 5 discusses the conclusions which can be drawn from this work about the mechanisms

of TFN evolution, and describe the directions of further work in this area.

Chapter 2

Evolution of TFN Degree Distribution

Transcription networks have an unusual structure. In both prokaryotes and eukaryotes, the number of target genes regulated by each transcription factor, its *out*-degree, follows a broad tailed distribution. By contrast, the number of transcription factors regulating a target gene, its *in*-degree, follows a much narrower distribution, which has no broad tail. We constructed a model of transcription network evolution through *trans*- and *cis*-mutations, gene duplication and deletion. The effects of these different evolutionary processes on the network structure are enough to produce an asymmetrical *in*- and *out*-degree distribution. However, the parameter values required to replicate known *in*- and *out*-degree distributions are unrealistic. We then considered variation in the rate of evolution of a gene dependent upon its position in the network. When transcription factors with many regulatory interactions are constrained to evolve more slowly than those with few interactions, the details of the *in*- and *out*-degree distributions of transcription networks can be fully reproduced over a range of plausible parameter values. The networks produced by our model depend on the relative rates of the different evolutionary processes. By determining the circumstances under which the networks with the correct degree distributions are produced, we are able to assess the relative importance of the different evolutionary processes in our model during evolution.

2.1 Background

Transcription regulation plays a key role in determining cellular function, response to external stimuli and development. Regulatory proteins orchestrate gene expression through thousands of interactions resulting in a system too complex to be easily understood in detail. This makes elucidation of gene regulation from a global perspectivethat of the transcription network as a wholean important challenge.

Genes in a transcription network either have outgoing edges, incoming edges or both. Outgoing edges from a gene represent the different targets that it regulates, while incoming edges at a gene represent the different transcription factors that regulate it. A number of studies [47, 83, 119] have established that, in both prokaryotes and eukaryotes, the degree distributions for outgoing and incoming edges are very different. The *out*-degree distribution, $n_{out}(k)$, follows a broad tailed distribution that is best described by a power-law: $n_{out}(k) \propto k^{-\gamma}$. The exponent γ is observed to be in the range $1 < \gamma < 2$ [47, 83]. A power-law distribution indicates that there are a small number of hub transcription factors that regulate a large number of genes [8]. Interpretation of power-law degree distributions, and the small world structure they confer, has been the focus of a great deal of attention [7, 8, 13, 19, 92, 93, 132]. In particular, it has been suggested that a power-law distribution may deliver an evolutionary advantage through increased mutational robustness and evolvability [8].

However, the *in*-degree distribution of transcription networks is much narrower than a power-law and has no broad tail [47, 83, 119]. It is best described by an exponential distribution $n_{in}(k) \propto \exp[-\alpha k]$. The exponential *in*-degree distribution reflects the fact that only a few transcription factors combinatorially regulate any one gene. There exist no hub target genes. For example, in the yeast transcription network, 93 per cent of target genes are regulated by fewer than five transcription factors [47].

The extent to which the *in*- and *out*-degree distributions of transcription networks are different is intriguing, and the cause unknown. In this chapter, I develop a model to explain the evolution of the asymmetrical transcription network degree distribution observed in yeast and other organisms. I focus on the different types of mutation through which the network evolves. Changes to the outgoing and incoming edges at a gene may occur as the result of mutation to a regulatory protein (*trans*-mutation) or as the result of mutation to transcription factor-binding sites (*cis*-mutation). These two processes change the network structure in different ways, but both result in either the loss or gain of regulatory interactions between existing genes. In addition, genes themselves may

be lost or gained in the network through deletion and duplication.

The rates at which a gene evolves may vary according to its connectivity in the transcription network [84, 134]. We investigate two types of connectivity-dependent evolution. It is often argued [8] that hub genes, which participate in many regulatory interactions, are particularly important for the proper functioning of the network, and are therefore constrained to evolve more slowly. This leads to the expectation of a slower rate of evolution among genes that regulate many downstream targets and a faster rate of evolution among genes that regulate only a few targets. It has also been suggested that a process of preferential attachment may occur in biological networks [8]. Under preferential attachment, new interactions are gained in proportion to the number of interactions a node already participates in. Such a process has been shown to occur in proteinprotein interactions networks [92, 132].

I construct a model incorporating evolution through *trans*- and *cis*-mutations, gene duplication and deletion along with variation in evolutionary rates depending on the connectivity of a gene. We use our model to unravel the relationship between the rates of evolution of genes through different processes in relation to the network structure.

2.2 Model

There are four types of network mutation in our model - gene deletion and duplication, plus *cis*- and *trans*-mutation. The *in*- and *out*-degree distributions of the network are determined by the rates at which these different types of mutation become fixed in the transcription network of a population. Since there is a clear functional difference between genes that code for transcription factors and those that code for other types of protein, we separate genes into two groups. Those with regulatory functions are labelled transcription factors (TFs) and those that are only regulated are labelled target genes (TGs). TGs have only incoming edges, while TFs may have either outgoing or incoming edges. We establish the equilibrium *in*- and *out*-degree distributions for four different versions of our model. In the first version, the rates of evolution are independent of a genes connectivity. We then consider two types of connectivity dependence in TF evolution. In the second version of our model, there is connectivity dependence such that the TFs with a large number of interactions undergo *trans*-evolution more slowly than those with few interactions. This is referred to as degree dependence in the rate of *trans*-evolution. In the third version of our model, there is connectivity dependence such that TFs gain new targets at a rate proportional to the number of targets they regulate. This is referred to as preferential attachment. The final

version of our model includes both degree dependence in the rate of *trans*-evolution and preferential attachment.

2.2.1 Gene deletion and duplication

We assume that when genes are duplicated they inherit all the regulatory interactions of their parent. Evolution through duplication occurs at rate D_+ and deletion occurs at rate D_- per gene (figure 2.1). A TF of *out*-degree k gains outgoing edges due to duplication of its targets at rate kD_+, and loses outgoing edges due to deletion of its targets at rate kD_-. Similarly, a gene of *in*-degree j gains incoming edges due to duplication of TFs at rate jD_+, and loses incoming edges due to deletion of TFs at rate jD_-. If the rates of gene deletion and duplication are different, this will result in either growth (if the rate of duplication is greater than the rate of deletion), or decline (if the rate of deletion is greater than the rate of duplication) in the size of the network. We assume that the rate of growth (or decline) of the network is small compared to the rate of rewiring of regulatory interactions through *trans*- and *cis*-mutation [29, 38, 139]. Thus, we consider only networks of constant size, and therefore assume that $D_+ = D_- = D$.

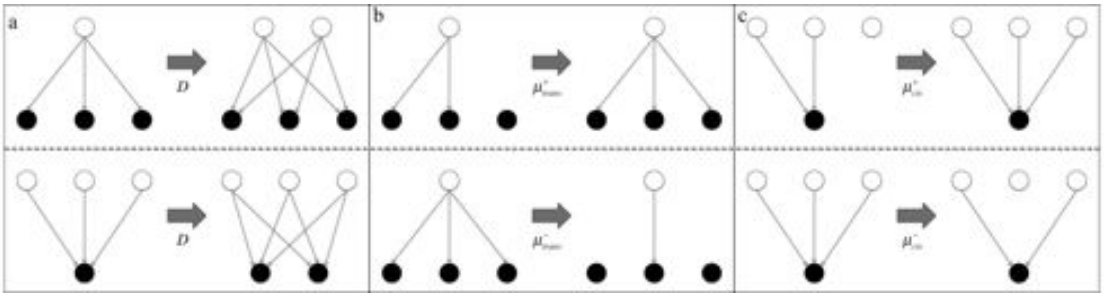

Figure 2.1: Mutations in a TFN. (a)(i) Duplication of a TFand all its outgoing edges, (ii) duplication of aTGand all its incoming edges. (b) Evolution via *trans*-mutation: (i) gain of an interaction through *trans*-evolution; (ii) loss of interactions through *trans*-evolution. (c) Evolution via *cis*-mutation: (i) gain of an interaction through *cis*-evolution; (ii) loss of an interaction through *cis*-evolution.

2.2.2 Evolution of regulatory-binding sites and transcription factors

A *trans*-mutation results in a change in the ability of TFs to bind to the promoter region of a gene. This may occur through a change in the binding affinity of a TF for a regulatory site. Alternatively, it may be the result of a TF gaining or losing an interaction with another TF, which helps it bind to the promoter region of a target [126]. Therefore, a *trans*-mutation in our model refers to a

mutation affecting a transcription factor protein only. It does not refer to mutations affecting the *cis*-regulatory regions of *trans*-acting genes. Following fixation of such a trans-mutation, a TF can cease to control some of the genes it currently regulates and can gain control over new genes. We assume that *trans*-evolution resulting in a TF potentially losing targets occurs at a constant rate μ^-_{trans}. In this process, an existing target is lost with probability m. The probability, $P^-_{k,\Delta k}$, that a TF with k *out*-edges loses Δk of its targets following a *trans*-mutation is given by

$$P^-_{k,\Delta k} = \frac{k!}{\Delta k!(k-\Delta k)!} m^{\Delta k}(1-m)^{k-\Delta k} \tag{2.1}$$

Similarly, we assume that trans-evolution resulting in the gain of new targets by a TF occurs at a constant rate μ^+_{trans}, which is independent of the *out*-degree of the TF. Overall, *trans*-evolution results in a gene losing incoming edges at rate $m\mu^-_{trans}$ (per edge) and gaining a new incoming edge at rate $\mu^+_{trans}(1-\frac{k}{N})$ (figure 2.1b). The factor$(1-\frac{k}{N})$ gives the probability that the gene gaining the new incoming edge is not one of the k genes currently regulated by the mutated TF. A *cis*-mutation results in the gain of a new binding site or the loss of an existing binding site in the promoter region of a gene. The rate at which binding sites are lost is μ^-_{cis}. The probability that a gene, which is regulated by k TFs, loses an interaction through loss of a TF binding site is $k\mu^-_{cis}$. A gene may also gain a new regulatory binding site for any TF in the network to which it is not currently connected, at rate μ^+_{cis} (figure 2.1c). Therefore, a gene currently regulated by k TFs gains an incoming edge through *cis*-evolution at a rate $\mu^+_{cis}(1-\frac{k}{N})$. Throughout, we assume that the size of the network N is large compared to any realistic *in*- or *out*-degree k, so that the terms $\frac{k}{N}$ may be neglected. Thus, new incoming edges are gained at constant rates μ^+_{trans} through *trans*-evolution and μ^+_{cis} through *cis*-evolution. We also develop a model in which degree dependence in the rate of *trans*-evolution occurs. In this model, a *trans* mutation, which results in TF-losing interactions, is fixed with a probability that depends on its *out*-degree. Since a trans-mutation affects the functioning of the transcription factor itself, it potentially alters all of the interactions in which a TF takes part. We assume that a *trans*-mutation at a TF with k targets has a deleterious effect on the functioning of the network that is proportional to k. We assume that a *trans*-mutation resulting in the loss of edges from the network is fixed with probability proportional to $\frac{1}{k}$. This has the effect that the mean and variance in the number of outgoing edges that are lost by a TF due to *trans* mutation is independent of k. In this way, the rates of evolution of TFs are degree dependent. We consider the possibility of other forms of degree dependence in the discussion. A

k	number of regulatory interactions
N_{TG}	expected number of TGs
N_{TF}	expected number of TFs
N	expected size of the network ($N = N_{TF} + N_{TG}$)
μ^{+}_{trans}	rate of gain of interactions due to *trans*-evolution
μ^{P}_{trans}	in preferential attachment modelthe rate at which new edges produced by *trans*-mutation undergo preferential attachment based on the *in*-degree of genes
μ^{R}_{trans}	in preferential attachment modelthe rate at which new edges produced by *trans*-mutation undergo random attachment to genes
μ^{-}_{trans}	rate of loss of interactions due to *trans*-evolution
m	probability a TF loses an existing target immediately following a *trans*-mutation
μ^{+}_{cis}	rate of gain of TF-binding sites through *cis*-evolution
μ^{P}_{cis}	in preferential attachment modelthe rate at which new edges produced by *cis*-mutation undergo preferential attachment based on the *out*-degree of TFs
μ^{R}_{cis}	in preferential attachment modelthe rate at which new edges produced by *cis*-mutation undergo random attachment to TFs
μ^{-}_{cis}	rate of loss of TF-binding sites through *cis*-evolution
D	rate of duplication and deletion
$P^{-}_{k,\Delta k}$	probability that a TF of *out*-degree k loses Δk edges as a result of a *trans*-mutation

Table 2.1: Model parameters

summary of all the parameters used in the model is given in Table 1.

2.2.3 Network Evolution

We allow evolution of the network by updating it at time intervals Δt, taken so that at most one mutation occurs and goes to fixation within each interval. Hence, the mean field equation for the expected number of genes with *in*-degree k at time t, changes in the time interval Δt by,

$$\begin{aligned}\Delta n_{in} &= \left(\Pi^{in}_{TG+}(k-1) + \Pi^{in}_{TF+}(k-1)\right) n_{in}(k-1,t) \\ &+ \left(\Pi^{in}_{TG-}(k+1) + \Pi^{in}_{TF-}(k+1)\right) n_{in}(k+1,t) \\ &- \left(\Pi^{in}_{TG+}(k) + \Pi^{in}_{TF+}(k) + \Pi^{in}_{TG-}(k) + \Pi^{in}_{TF-}(k)\right) n_{in}(k,t) \end{aligned} \tag{2.2}$$

where $\Pi^{in}_{TG+}(k)$ and $\Pi^{in}_{TG-}(k)$ are the probabilities of a gene with *in*-degree k gaining or losing an edge through mutation at the regulated gene; and $\Pi^{in}_{TF+}(k)$ and $\Pi^{in}_{TF-}(k)$ are the probabilities of a gene with *in*-degree k gaining or losing an edge through mutation at a TF regulating it, in the time interval Δt. Similarly, the expected number of genes with *out*-degree k at time t changes in the time interval Δt by

$$\begin{aligned}\Delta n_{out} &= \left(\Pi^{out}_{TG+}(k-1) + \Pi^{out}_{TF+}(k-1)\right) n_{out}(k-1,t) \\ &+ \Pi^{out}_{TG-}(k+1) n_{out}(k+1,t) \\ &- \left(\Pi^{out}_{TG+}(k) + \Pi^{out}_{TG-}(k) + \Pi^{out}_{TF+}(k)\right) n_{out}(k,t) \\ &+ \sum_{j=k}^{N} \Pi^{out}_{TF-}(j,k) n_{out}(j,t) - \sum_{j=0}^{k} \Pi^{out}_{TF-}(k,j) n_{out}(k,t) \end{aligned} \tag{2.3}$$

where N is the number of genes (TFs and TGs) in the network; $\Pi^{out}_{TG+}(k)$ and $\Pi^{out}_{TG-}(k)$ are the probabilities of a gene with *out*-degree k gaining or losing an edge through mutation at one of its targets; $\Pi^{out}_{TF+}(k)$ is the probability that a TF with *out*-degree k gains a target through mutation at the TF; and $\Pi^{out}_{TF-}(j,k)$ is the probability that a TF with *out*-degree $j \geq k$ loses interactions to become a TF with *out*-degree k due to mutation at the TF.

The equilibrium *in*-and *out*-degree distributions for the model can be found from equations (2.2) and (2.3), by setting the left-hand sides of both equations to 0. (see Appendix A). The equilibrium *in*-degree distribution satisfies

$$\begin{aligned}&\left(\Pi^{in}_{TF-}(k+1) + \Pi^{in}_{TG-}(k+1)\right) n_{in}(k+1) \\ &\qquad = \left(\Pi^{in}_{TF+}(k) + \Pi^{in}_{TG+}(k)\right) n_{in}(k) \end{aligned} \tag{2.4}$$

After making a number of approximations (see Appendix A), the equilibrium *out*-degree distribution satisfies

$$\begin{aligned}&\left(\Pi^{out}_{TG-}(k+1)+(k+1)\mu^{-}_{trans}K(\gamma,m)\right)n_{out}(k+1)\\&=\left(\Pi^{out}_{TG+}(k)+\Pi^{out}_{TF+}(k)-k\mu^{-}_{trans}K(\gamma,m)\right)n_{out}(k)\end{aligned}\tag{2.5}$$

for the model excluding degree dependence in the rate of *trans*-evolution and

$$\begin{aligned}&\left(\Pi^{out}_{TG-}(k+1)+\mu^{-}_{trans}K(\gamma,m)\right)n_{out}(k+1)\\&=\left(\Pi^{out}_{TG+}(k)+\Pi^{out}_{TF+}(k)-\mu^{-}_{trans}K(\gamma,m)\right)n_{out}(k)\end{aligned}\tag{2.6}$$

for the model including degree dependence in the rate of *trans*-evolution. The positive parameter γ arises from the approximations used to obtain equations (2.5) and (2.6) (see Appendix A), and the functions $K(\gamma,m)$ are specific to each of the models we consider and will be described below. We now solve equations (2.4), (2.5) and (2.6) for the *in*- and *out*-degree distributions for four specific models of transcription network evolution. We start using a simple model and then investigate different models including degree dependence in the rate of *trans*-evolution and preferential attachment, to ask what conditions are required to explain the observed difference between the *in*- and *out*-degree distributions of transcription networks.

2.2.4 Simulations of Network Evolution

Simulations were carried out using ensembles of 1000 networks, each with an expected size of 100 TFs and 100 TGs. Networks were subject to 10^6 mutations after which the average degree distributions were taken over the ensemble, and the mean degree distributions determined. The evolutionary algorithm used allowed networks to vary in size between a lower and upper boundary of 50 and 150 nodes, for both TFs and TGs. Loss of interactions through trans-mutation was executed by deleting each of a TFs outgoing edges with probability m. For gain of new interactions, random attachment was executed by selecting a gene and a TF at random and adding an edge

	Model 1	Model 2	Model 3	Model 4
$\Pi^{in}_{TG+}(k)$	μ^+_{cis}	μ^+_{cis}	μ^+_{cis}	μ^+_{cis}
$\Pi^{in}_{TG-}(k)$	$k\mu^-_{cis}$	$k\mu^-_{cis}$	$k\mu^-_{cis}$	$k\mu^-_{cis}$
$\Pi^{in}_{TF+}(k)$	$kD + \mu^+_{trans}$	$kD + \mu^+_{trans}$	$kD + \mu^R_{trans} + k\mu^P_{trans}$	$kD + \mu^R_{trans} + k\mu^P_{trans}$
$\Pi^{in}_{TF-}(k)$	$kD + k\mu^-_{trans}m$	$kD + k\left\langle \frac{1}{k} \right\rangle \mu^-_{trans}m$	$kD + k\mu^-_{trans}m$	$kD + k\left\langle \frac{1}{k} \right\rangle \mu^-_{trans}m$

Table 2.2: Incoming edge event probabilities

	Model 1	Model 2	Model 3	Model 4
$\Pi^{out}_{TG+}(k)$	$kD + \mu^+_{cis}$	$kD + \mu^+_{cis}$	$kD + \mu^R_{cis} + k\mu^P_{cis}$	$kD + \mu^R_{cis} + k\mu^P_{cis}$
$\Pi^{out}_{TG-}(k)$	$kD + k\mu^-_{cis}$	$kD + k\mu^-_{cis}$	$kD + k\mu^-_{cis}$	$kD + k\mu^-_{cis}$
$\Pi^{out}_{TF+}(k)$	μ^+_{trans}	μ^+_{trans}	μ^+_{trans}	μ^+_{trans}
$\Pi^{out}_{TF-}(k)$	$\mu^-_{trans}P^-_{j,j-k}$	$\mu^-_{trans}\frac{P^-_{j,j-k}}{j}$	$\mu^-_{trans}P^-_{j,j-k}$	$\mu^-_{trans}\frac{P^-_{j,j-k}}{j}$
$K(\gamma, m)$	$\frac{1}{2(\gamma-1)}(1-(1-m)^{\gamma-1})$	$\frac{1}{2\gamma}(1-(1-m)^{\gamma})$	$\frac{1}{2(\gamma-1)}(1-(1-m)^{\gamma-1})$	$\frac{1}{2\gamma}(1-(1-m)^{\gamma})$

Table 2.3: Outcoming edge event probabilities

between them. Preferential attachment of incoming edges was executed by selecting a gene with a probability proportional to its in-degree and a TF at random. A new edge was then added between them. Similarly for preferential attachment of outgoing edges, a TF was selected with probability proportional to its *out*-degree, and another gene was selected at random. An interaction was then added between them. Simulations were run for a range of parameter values. Data shown are for $m = 0.01$, corresponding to the case $m \to 0$ (equation (2.15)). The rate of duplication used is $D = 0.26$, the rate of gain of interactions through *trans*-evolution is $\mu^+_{trans} = 0.04$, and through *cis*-evolution is $\mu^+_{cis} = 0.31$. The rate of loss of interactions through *trans*-evolution is $m\mu^-_{trans} = 0.25$ and through *cis*-evolution is $\mu^-_{cis} = 0.14$.

2.3 Results

2.3.1 Model 1: no connectivity dependence

In the first model, we assume there is neither any degree dependence nor any preferential attachment in the rate of *trans*-evolution. The event probabilities for the *in*- and *out*-degree distributions in this model are given in Table 2a. Substituting these in equations (2.4) and (2.5), we find for the

in-degree distribution (see Appendix A)

$$n_{in}(k) \propto k^{-\lambda} \exp[-\alpha k]$$

where

$$\alpha = \ln\left(1 + \frac{\mu^-_{cis} + m\mu^-_{trans}}{D}\right),$$
$$\lambda = 1 - \frac{\mu^+_{cis} + \mu^+_{trans}}{D} \tag{2.7}$$

This is approximately an exponential distribution, characterized by α, unless α is small, which occurs if $D \gg \mu^-_{cis} + \mu^-_{trans}m$, or λ is large and negative, which occurs if $D \ll \mu^+_{cis} + \mu^+_{trans}$. The equilibrium *out*-degree distribution for this model obtained from equation (2.5) is

$$n_{out}(k) \propto k^{-\gamma} \exp[-\beta k]$$

where

$$\beta = \ln\left(\frac{\mu^-_{cis} + D + \mu^-_{trans}K(\gamma, m)}{D - \mu^-_{trans}K(\gamma, m)}\right),$$
$$\gamma = 1 - \frac{\mu^+_{cis} + \mu^+_{trans}}{D - \mu^-_{trans}K(\gamma, m)} \tag{2.8}$$

and $K(\gamma, m)$ is as in Table 2b. This distribution is a power-law characterized by γ only if β is 0. This occurs if $\mu^-_{cis} = -2\mu^-_{trans}K(\gamma, m)$. However, as the rates, μ^-_{cis} and μ^-_{trans}, are both positive constants, and $K(\gamma, m) > 0$ (Table 2b), this condition cannot be met. Therefore, this model cannot produce a power-law *out*-degree distribution.

2.3.2 Model 2: degree dependence in the rate of *trans*-evolution

In this model, we allow degree dependence in the rate of *trans*-evolution. Substituting the event probabilities for this model (Table 2a) into equations (2.4) and (2.6), we find for the *in*-degree distribution

$$n_{in}(k) \propto k^{-\lambda} \exp[-\alpha k]$$

where

$$\alpha = \ln\left(1 + \frac{\mu^-_{cis} + m\left\langle\frac{1}{k}\right\rangle \mu^-_{trans}}{D}\right),$$

$$\lambda = 1 - \frac{\mu^+_{cis} + \mu^+_{trans}}{D} \tag{2.9}$$

and $\left\langle\frac{1}{k}\right\rangle = \sum_{j=1}^{N} \frac{n_{out}(j)}{j}$ determines the mean rate of *trans*-evolution across the network. Following the same procedure as for Model 1, this distribution will be approximately exponential unless α is small or λ is large and negative, which occurs when $D \gg \mu^-_{cis} + \left\langle\frac{1}{k}\right\rangle \mu^-_{trans} m$ or $D \ll \mu^+_{cis} + \mu^+_{trans}$, respectively.

The equilibrium *out*-degree distribution for this model is

$$n_{out}(k) \propto k^{-\gamma} \exp[-\beta k]$$

where

$$\beta = \ln\left(1 + \frac{\mu^-_{cis}}{D}\right),$$

$$D(\gamma - 1) + \mu^+_{cis} + \mu^+_{trans} = \mu^-_{trans} K(\gamma, m)\left(1 + \frac{D}{D + \mu^-_{cis}}\right), \tag{2.10}$$

and $K(\gamma, m)$ is as in Table 2b. This distribution is a power-law characterized by γ only if β is 0. This occurs if $D \gg \mu^-_{cis}$. Under this condition, equation (2.10) has solutions with $\gamma > 1$ provided $m\mu^-_{trans} > \mu^+_{cis} + \mu^+_{trans}$.

2.3.3 Model 3: preferential attachment

This model includes preferential attachment, but excludes degree dependence in the rate of *trans*-evolution (considered in Model 2). In preferential attachment models, the rate at which nodes gain new edges is proportional to the number of edges already attaching to them. Preferential attachment has been discussed widely in the study of other biological networks [8, 92, 132], including in the proteinprotein interaction network of yeast [132].

We model preferential attachment of incoming and outgoing edges separately. For incoming edges our model is as follows: new edges arise due to *trans*-evolution at rate μ^+_{trans}. When such a

new edge arises, it may be either through preferential attachment or through random attachment (i.e. the new edge attaches to each gene with equal probability) at the gene that is regulated. In the case of preferential attachment, the probability that a gene gains a new incoming edge is proportional to its *in*-degree. In the case of random attachment, the probability that a gene gains a new incoming edge is independent of its *in*-degree. We assume that such new edges undergo preferential attachment to a gene at rate μ^P_{trans}, and undergo random attachment at a rate μ^R_{trans}. The rate at which a gene of *in*-degree k gains a new edge due to preferential attachment is $k\mu^P_{trans}$, and the rate at which it gains a new edge due to random attachment is μ^R_{trans}. The total rate at which TFs gain new outgoing edges is then $\mu^+_{trans} = \frac{E}{N_{TF}}\mu^P_{trans} + \frac{N}{N_{TF}}\mu^R_{trans}$, where E is the total number of edges in the network.

Our model of preferential attachment for outgoing edges is of the same form: new edges arise due to *cis*-evolution at rate μ^+_{cis}. The rate at which a TF of *out*-degree k gains new outgoing edges due to preferential attachment is then $k\mu^P_{cis}$, and the rate at which it gains new edges due to random attachment is μ^R_{cis}. The total rate at which genes gain new incoming edges is then $\mu^+_{cis} = \frac{E}{N}\mu^P_{cis} + \frac{N_{TF}}{N}\mu^R_{cis}$. Our model of preferential attachment is illustrated in figure 2.2a.

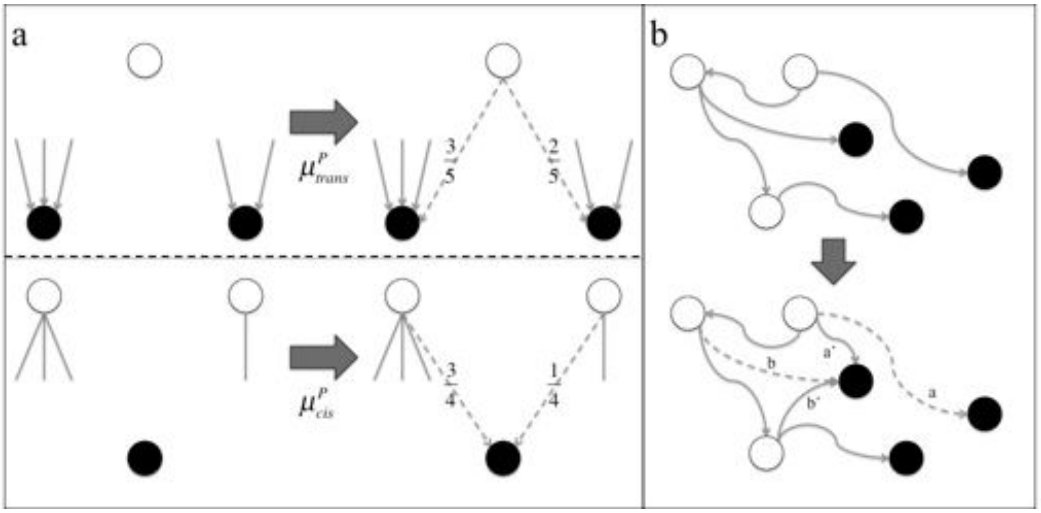

Figure 2.2: Preferential attachment and rewiring. (a) Preferential attachment: (i) preferential attachment of incoming edges. A TF choosing between a TG with three incoming edges and a TG with two incoming edges gains an interaction with the first with probability 0.6 and with the second with probability 0.4 due to preferential attachment; (ii) preferential attachment of outgoing edges. When choosing between a TF with three outgoing edges and a TF with one outgoing edge, a TG gains an interaction with the TF with three edges with probability 0.75 and with the TF with one edge with probability 0.25 due to preferential attachment. (b) Rewiring: (i) network prior to rewiring; (ii) edge **a** is rewired to edge **a'**. This results in a change in the *in*-degree of two TGs but leaves the *out*-degree of the TF unchanged. Edge **b** is rewired to edge **b**, changing the *out*-degree of two TFs but leaving the *in*-degree of the TG unchanged.

The event probabilities for the *in*- and *out*-degree distributions in this model are given in Table 2a. Substituting these into equations (2.4) and (2.5), the *in*-degree distribution is

$$n_{in}(k) \propto k^{-\lambda} \exp[-\alpha k]$$

where

$$\alpha = \ln\left(\frac{D + \mu^-_{cis} + m\mu^-_{trans}}{D + \mu^P_{trans}}\right),$$
$$\lambda = 1 - \frac{\mu^+_{cis} + \mu^R_{trans}}{D + \mu^P_{trans}} \tag{2.11}$$

This distribution will be approximately exponential unless α is small or λ is large and negative. That is, unless $D + \mu^P_{trans} \gg \mu^-_{cis} + \mu^-_{trans} m$, or $D + \mu^P_{trans} \ll \mu^+_{cis} + \mu^R_{trans}$. Therefore, we require $\mu^P_{trans} \sim \mu^-_{cis} + \mu^-_{trans} m$. The equilibrium *out*-degree distribution for this model is

$$n_{out}(k) \propto k^{-\gamma} \exp[-\beta k]$$

where

$$\beta = \ln\left(\frac{D + \mu^-_{cis} + \mu^-_{trans} K(\gamma, m)}{D + \mu^P_{cis} - \mu^-_{trans} K(\gamma, m)}\right),$$
$$\gamma = 1 - \frac{\mu^R_{cis} + \mu^+_{trans}}{D + \mu^P_{cis} - \mu^-_{trans} K(\gamma, m)}, \tag{2.12}$$

and $K(\gamma, m)$ is as in Table 2b. This distribution is a power-law characterized by γ only if β is 0. This occurs if $\mu^-_{cis} = \mu^P_{cis} - 2\mu^-_{trans} K(\gamma, m)$. Equation (2.12) then gives $\gamma = 1 - \frac{\mu^R_{cis} + \mu^+_{trans}}{D + \frac{1}{2}(\mu^-_{cis} + \mu^P_{cis})}$, and the only solutions have $\gamma < 1$.

2.3.4 Model 4: degree dependence and preferential attachment

In the final model, we include degree dependence (as described in Model 2) and preferential attachment (as described in Model 3). The event probabilities for this model are given in Table 2a. Using these with equations (2.4) and (2.6), we find for the *in*-degree distribution

$$n_{in}(k) \propto k^{-\lambda} \exp[-\alpha k]$$

where

$$\alpha = \ln\left(\frac{D + \mu^-_{cis} + m\left\langle\frac{1}{k}\right\rangle \mu^-_{trans}}{D + \mu^P_{trans}}\right),$$

$$\lambda = 1 - \frac{\mu^+_{cis} + \mu^R_{trans}}{D + \mu^P_{trans}} \tag{2.13}$$

This distribution is approximately exponential unless α is small or λ is large and negative. That is, unless $D + \mu^P_{trans} \gg \mu^-_{cis} + \left\langle\frac{1}{k}\right\rangle \mu^-_{trans} m$, or $D + \mu^P_{trans} \ll \mu^+_{cis} + \mu^R_{trans}$. Therefore, we require $\mu^P_{trans} \sim \mu^-_{cis} + \left\langle\frac{1}{k}\right\rangle \mu^-_{trans} m$. The equilibrium *out*-degree distribution for this model is

$$n_{out}(k) \propto k^{-\gamma} \exp[-\beta k]$$

where

$$\beta = \ln\left(\frac{D + \mu^-_{cis}}{D + \mu^P_{cis}}\right),$$

$$(D + \mu^P_{cis})(\gamma - 1) + \mu^R_{cis} + \mu^+_{trans} = \mu^-_{trans} K(\gamma, m)\left(1 + \frac{D + \mu^P_{cis}}{D + \mu^-_{cis}}\right), \tag{2.14}$$

and $K(\gamma, m)$ is as in Table 2b. This distribution is a power-law characterized by γ only if β is 0. This requires $\mu^-_{cis} = \mu^P_{cis}$. Under this condition, the third term in equation (2.14) has solutions with $\gamma > 1$ provided $m\mu^-_{trans} > \mu^R_{cis} + \mu^+_{trans}$.

2.4 Discussion

To assess the four models we have presented, we compare their results to empirical observations from the yeast transcription network. The *out*-degree distribution of the *Saccharomyces cerevisiae* transcription network is best described by a power-law distribution with an exponent $\gamma = 1.5$, while the *in*-degree distribution is best described by an exponential distribution with exponent $\alpha = 0.4$ [83].

Since the exponent of the *out*-degree distribution for yeast is greater than 1, we conclude that Models 1 and 3, which do not include degree dependence in the rate of *trans*-evolution, cannot account for the observed *out*-degree distribution of the *S. cerevisiae* transcription network.

However, Models 2 and 4, which include degree dependence in the rate of transcription factor evolution, can both produce networks with power-law out-degree distributions whose exponent is $\gamma > 1$. Therefore, we conclude that degree dependence in the rate of transcription factor evolution could be an important factor in producing the structure of the yeast transcription network.

2.4.1 Empirical rates of evolution

We can further distinguish between Models 2 and 4 by referring to empirical data on the rates of evolution in the yeast transcription network. The rate of gene duplication in yeast is found to be in the range 1×10^5 - 6×10^5 per Myr [38]. The rate of evolution (gain or loss) of regulatory interactions is an order of magnitude higher, approximately 36×10^5 per Myr [45]. Evolution of regulatory interactions may occur due to changes in regulatory proteins (*trans*-mutations in our model) or due to changes in *cis*-regulatory elements. A trans-mutation in our model refers to a mutation affecting a transcription factor protein only. It does not refer to mutations affecting the *cis*-regulatory regions of trans-acting genes. In practice, it is difficult to distinguish between the effects of the *trans*- and *cis*-mutations of our model without much more detailed comparative data. Studies on the contribution of the evolution of *cis*-regulatory elements and of *trans*-acting proteins to the evolution of gene expression have mixed findings. Variation between yeast strains have been found to be mainly due to variation in *trans*-acting proteins by some studies [138, 147, 148], while this has been contradicted by others [105].

In Model 2, a power-law *out*-degree distribution is only produced if $D \gg \mu_{cis}^-$. If we consider the case in which *trans*-evolution is more rapid than *cis*-evolution, then, given a rate of evolution of regulatory interactions of 36×10^5 per Myr [45] and a rate of gene duplication of range 1×10^5 - 6×10^5 per Myr [38], Model 2 suggests that the loss of regulatory interactions must be approximately 99 per cent due to *trans*-evolution. Such a disproportionate rate is not consistent with empirical data on the relative contributions of trans- and cis-change to the evolution of gene expression in yeast [105, 138, 147, 148]. Therefore, we can reject Model 2, as inadequate to explain the structure of the yeast transcription network.

2.4.2 Preferential attachment

Model 4 can produce a power-law *out*-degree distribution provided $\mu_{cis}^- = \mu_{cis}^P$. This requirement means that the rate at which transcription factors lose connections to target genes through *cis*-mutations must be balanced by the rate at which they gain new targets through preferential

attachment. From this we also conclude that preferential attachment for outgoing edges is a likely factor in producing the observed yeast transcription network. The condition $\mu_{cis}^{-} = \mu_{cis}^{P}$ is identical to a model in which transcription factors undergo rewiring (figure 2.2b), and suggests that transcription factors undergo a constant turnover of targets, without net gain or loss. In order to determine whether preferential attachment among incoming edges occurs, we must consider the in-degree distribution of Model 4. This is given by equation (2.13), with an exponential exponent, a, of approximately 0.4. Given a low rate of duplication, equation (2.13) suggests that preferential attachment of incoming edges at target genes is also likely to be a factor in producing the structure of the yeast transcription network. Figure 2.3 shows the result of simulations using Model 4, which confirm that this model can reproduce the observed structure of the yeast transcription network.

There are several mechanisms by which preferential attachment in transcription networks may occur. One possibility is that different TFs have different "stickiness", such that those which are more sticky gain new targets at a higher rate than those which are less sticky. Stickiness is considered to be an intrinsic property of a TF, resulting from its structure. A distribution of stickiness amongst different TFs is able to give rise to network evolution identical to that which results from preferential attachment [92]. Alternatively, preferential attachment may result from turnover of TF binding sites. Turnover of binding sites occurs when a binding site for one TF mutates to become a binding site for another TF. Thus there is no net loss or gain of TF binding sites from the network. However, the rate at which a TF undergoes turnover of binding sites is proportional to the number of binding sites it has. Thus a form of preferential attachment occurs. As discussed above, our model suggests that turnover is a likely mechanism driving the evolution of transcription networks, as our model requires $\mu_{cis}^{-} = \mu_{cis}^{P}$ in order to reproduce the observed *out*-degree distribution of the Yeast transcription network.

2.4.3 Evolution via *trans*-mutation

Our model for loss of interactions through *trans*-evolution includes two parameters, μ_{trans}^{-}, the rate at which trans-mutations are fixed, and m, the probability each interaction is lost given that a *trans*-mutation is fixed. This means that following a *trans*-mutation a transcription factor will retain, on average, a fraction $1 - m$ of its interactions. As it is difficult to estimate m, we consider two important cases: $m \to 0$ and $m = 1$. In the first case, transcription factors evolve by small changes, one interaction at a time. In the second case, transcription factors lose all their existing interactions, and subsequently gain new ones through both *cis*- and *trans*-evolution. In this case,

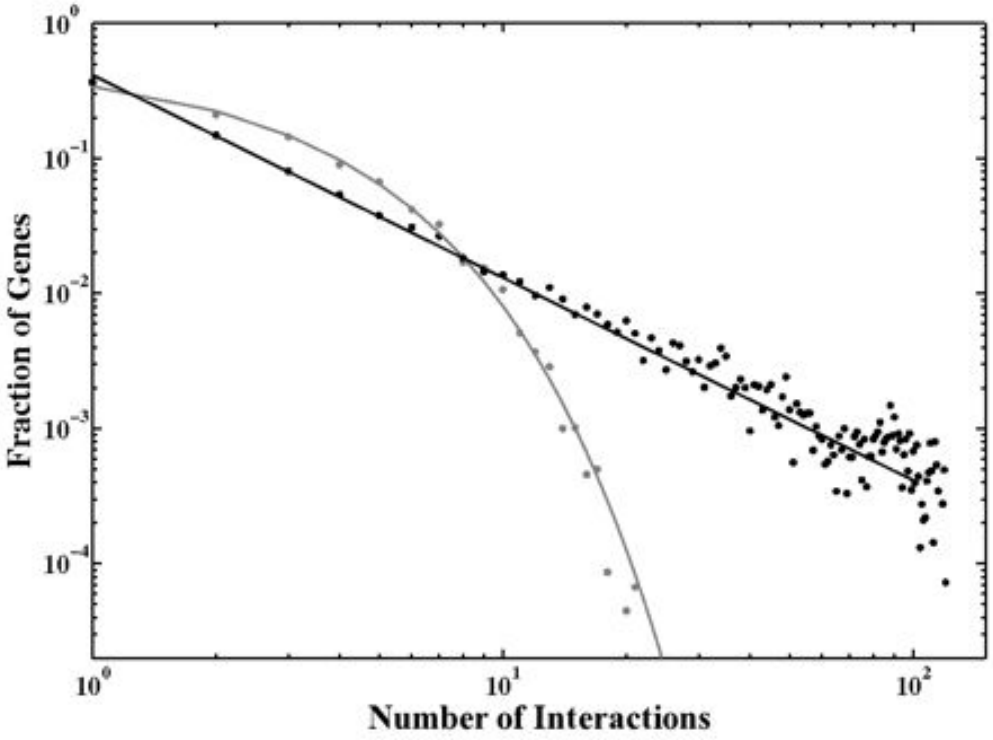

Figure 2.3: Simulation results. Simulated networks with an expected size of 100 TFs and 100 TGs. Networks have an exponential in-degree and power-law *out*-degree with $\alpha = 0.4$ and $\gamma = 1.5$. Simulations consist of ensembles of 1000 networks evolved for 106 mutations. Plot is on a loglog scale. Simulations were run for a range of parameter values. A typical example is shown. Data points show the degree distributions for simulated networks, solid lines are the predicted distribution. In-degree is shown in grey. *Out*-degree is shown in black. The networks were produced using model 4, including degree dependence in the rate of *trans*-evolution and preferential attachment. Here, the rate of duplication is $D = 0.26$, the rate of gain of interactions through *trans*-evolution is $\mu^+_{trans} = 0.04$, and through *cis*-evolution is $\mu^+_{cis} = 0.31$. The rate of loss of interactions through *trans*-evolution is $m\mu^-_{trans} = 0.25$ and through *cis*-evolution is $\mu^-_{cis} = 0.14$.

the TF may be seen as completely losing its old function before acquiring a new function. When $m \to 0$, equation (2.14) for the *out*-degree distribution in Model 4 may be used to obtain the approximation

$$\gamma = 1 + \frac{m\mu^-_{trans} - \mu^R_{cis} - \mu^+_{trans}}{D + \mu^P_{cis}} \tag{2.15}$$

Similarly, if $m = 1$, equation (2.14) may be used to obtain

$$\gamma = \frac{1}{2}\left(1 - \frac{\mu_{cis}^{R} + \mu_{trans}^{+}}{D + \mu_{cis}^{P}} + \sqrt{\left(1 - \frac{\mu_{cis}^{R} + \mu_{trans}^{+}}{D + \mu_{cis}^{P}}\right)^{2} + \frac{4\mu_{trans}^{-}}{D + \mu_{cis}^{P}}}\right) \tag{2.16}$$

Therefore, given measurements of the relative rates of *cis*- and *trans*-evolution, it would be possible to distinguish between these two cases.

Given values for the other parameters, the value of μ_{trans}^{-} required in equation (2.16) to produce $\gamma = 1.5$ will be greater than the value of μ_{trans} required in equation (2.15) to produce the same distribution. The rate at which interactions are lost through *trans*-evolution is proportional to $m\mu_{trans}^{-}$. Therefore, the case $m \to 0$ is consistent with a slower rate of loss of interactions though *trans*-evolution than the case $m = 1$. This can be compared with recent work [44], suggesting that gene network evolution may be characterized by a 2-2-1 pattern (net gain of two genes and two edges along with loss of one edge). This suggests that loss of edges occur less frequently than gains. In our model, the ratio of gain of two edges to loss of one edge is more consistent with the case $m \to 0$ than with the case $m = 1$, since when $m \to 0$ edges are lost more slowly through *trans* mutation as compared to gains.

2.4.4 Alternative forms of degree dependence

We have chosen to consider a form of degree dependence in the rate of *trans* mutation such that the mean and variance in the number of outgoing edges lost by a TF following *trans* mutation is independent of k. This is in contrast to the case without degree dependence, in which the mean and variance in the number of edges lost is proportional to k. Whilst it is natural to contrast these two cases, it is clear that other forms of degree dependence are possible. One case of particular interest is to apply Kimura's formula for the probability, u, of fixation of a mutant gene in a population [64]. This is given by

$$u = \frac{s}{1 - \exp[-Ms]} \tag{2.17}$$

for haploid populations with small s. Here s is the selection coefficient and M is the effective population size. Effective population size is the size of an ideal population that would show the same amount of dispersion of allele frequencies as that observed in the population being considered. It is generally much smaller than the number of individuals in a population. If we assume that

loss of each outgoing edge due *trans* mutation has a small deleterious effect σ, then the total deleterious effect of a *trans* mutation which results in the loss of i outgoing edges is given by $s = -\sigma i$. Combining equation (2.17) with equation (2.1) for the probability that a TF with k outgoing edges loses i edges, we have that the mean number of edges, $\bar{k}$, lost by a TF following *trans* mutation is

$$\bar{k} = \frac{\sum_{i=0}^{i=k} \frac{k!}{i!(k-i)!} m^i (1-m)^{k-i} \frac{\sigma i}{\exp[M\sigma i]-1} i}{\sum_{i=0}^{i=k} \frac{k!}{i!(k-i)!} m^i (1-m)^{k-i} \frac{\sigma i}{\exp[M\sigma i]-1}} \tag{2.18}$$

In general this cannot be solved. However, if we take $M\sigma \gg 1$ we can write $\frac{\sigma i}{\exp[M\sigma i]-1} \approx \sigma i \exp[-M\sigma i]$. Using this, equation (2.18) can be solved to give

$$\bar{k} = 1 + \frac{(k-1)m}{\exp[M\sigma](1-m) + m} \tag{2.19}$$

Since $M\sigma \gg 1$, this gives $\bar{k} \approx 1$, and is approximately independent of k. Under these assumptions all TFs will tend to undergo *trans* mutations which result in the loss of a single outgoing edge at a time, independent of their *out*-degree, k. This has the same qualitatively effect on how *trans* mutations affect the network, as does assuming degree dependence of the form used in models 2 and 4. It is the subject of further work to investigate the effects of degree dependence that arise from Kimura's formula in the more general case, when effective populations are small or selective coefficients are large.

2.4.5 Growing and shrinking networks

We have considered networks in which the rates of gene duplication and deletion balance. However, it is well known that duplication growth models of networks can produce power-law distributions [13, 19, 44, 93]. We have not considered growing networks for two reasons. First, the observed low rate of gene duplication in yeast means that genes will undergo rewiring events at a rate that is 10-fold greater than the rate of duplication events. Second, the observed rates of gene duplication and deletion are comparable [38] and suggests that the yeast transcription network is not undergoing constant growth. Therefore, any model that relies on network growth by duplication to reproduce the observed degree distributions in the yeast transcription network is not consistent with the data.

We have also investigated the case of shrinking networks. Although it is obvious that real

networks cannot be continuously shrinking, the recent whole genome duplication in yeast [62] means that there have been a great many redundant genes that have been lost resulting in an increased rate of gene deletion. Thus, the network has recently been undergoing a period of evolution in which it has been shrinking. We have considered a model in which a network is shrinking (see Appendix A). We show that this model is not able to reproduce the observed structure of the yeast transcription network without both degree dependence in the rate of *trans*-evolution and preferential attachment. Therefore, this model does not alter our conclusions.

2.4.6 Autoregulation

In the analysis above, we neglected autoregulation of transcription factors. Autoregulation alters the consequences of transcription factor duplication. When an autoregulating transcription factor with k outgoing edges is duplicated, it gains an edge and becomes a transcription factor with $k+1$ outgoing edges. In our model, we assume that the transcription factors regulate each of the possible N targets with equal probability. Therefore, the probability that a transcription factor of *out*-degree k autoregulates is k/N. So the rate at which new transcription factors with *out*-degree k are produced due to duplication of autoregulators is $(k-1)\frac{D}{N}n_{out}(k-1)$, and the rate at which transcription factors with *out*-degree k are lost due to duplication of autoregulators is $k\frac{D}{N}n_{out}(k)$. Therefore, duplication of autoregulators provides a mechanism for a form of preferential attachment, since it results in transcription factors gaining new outgoing edges at a rate proportional to their *out*-degree. However, the rate at which this preferential attachment occurs is $\sim (1/N)$ times the rate of gene duplication, D. Since N is large, duplication of autoregulating transcription factors is therefore expected to have little impact on the equilibrium degree distributions produced by our models. To verify these arguments, we carried out simulations in which autoregulation was permitted in each of the four models (data not shown). The results showed that autoregulation had only a minor quantitative effect on the outcome of the models provided the rate of duplication D was not high. We also note empirical findings in the yeast transcription network, which show that only 12 out of 131 (9 per cent) of transcription factors admit autoregulation [87]. Given this, the rate at which new edges are produced through duplication of autoregulating transcription factors is approximately an order of magnitude less than the rate of gene duplication. Even at this rate of autoregulation, duplication of autoregulating transcription factors will not have a significant impact on the degree distribution of the network.

2.5 Conclusion

We have compared four simple models for the evolution of transcription networks. Genes are separated into regulatory transcription factors and non-regulatory target genes, which evolve through mutation of *trans*- and *cis*-elements, as well as through deletion and duplication. When rates of evolution are constant across the network, our model can reproduce the exponential *in*-degree and power-law *out*-degree distributions characteristic of transcription networks. However, this model cannot produce networks with the power-law exponent observed *in*- the *out*-degree of the yeast transcription network. It is only when the effects of variation in the rate of protein evolution are taken into account that the correct degree distributions are fully reproduced. This variation takes two forms. First, degree dependence in the rate of *trans*-evolution, meaning that the more regulatory interactions a transcription factor participates in, the more slowly it undergoes transevolution. Second, preferential attachment, meaning that genes gain new interactions at a rate proportional to the number of interactions they already participate in. The requirement for preferential attachment can be relaxed if the rate of evolution through gene duplication and deletion is high compared to the rate of *cis*-evolution. We have proposed a model in which the rate of *trans*-evolution among transcription factors varies in inverse proportion to the number of targets they regulate. The true rate of trans-evolution depends on the rate of evolution of gene sequence and gene expression [122]. The relationship between the evolution of gene expression, gene sequence and position in the transcription network is likely to be complex and is not fully understood [132]. Our model suggests that variation in the rate of *trans*-evolution with the position of a gene in an interaction network significantly affects the structure of that network. We have considered these effects in relation to the structure of transcription networks, although they may also play a role in shaping the structure of protein interaction networks and metabolic networks.

2.6 Appendix A

2.6.1 Derivation of Equilibrium Degree Distributions

In order to find the equilibrium *in*- and *out*-degree distributions we must set the left hand side of equations (2.2) and (2.3) to zero. Equation (2.2) is straightforward to solve. It's solutions satisfy equation (2.4)

In order to solve equation (2.3) we make an approximation for the term $\sum_{j=k}^{N} \Pi_{TF-}^{out}(j,k) n_{out}(j,t) - \sum_{j=0}^{k} \Pi_{TF-}^{out}(k,j) n_{out}(k,t)$, which describes loss of interactions through trans evolution. For the

model without degree dependence in the rate of *trans* evolution (Model 1 and Model 3), $\Pi^{out}_{TF-}(j,k)$ is given by

$$\Pi^{out}_{TF-}(j,k) = \mu^{-}_{trans}\frac{j!}{k!(j-k)!}m^{j-k}(1-m)^k. \tag{2.20}$$

By assuming a solution of the form $n_{out}(k) = A_{out}k^{-\gamma}$, where A_{out} is a normalization constant for the *out*-degree distribution, we can use the approximation

$$\begin{aligned}
&\sum_{j=k}^{N}\Pi^{out}_{TF-}(j,k)n_{out}(j) - \sum_{j=0}^{k}\Pi^{out}_{TF-}(k,j)n_{out}(k) \\
&= \mu^{-}_{trans}n_{out}(k)(1-m)^{\gamma-1} - \mu^{-}_{trans}n_{out}(k) + O(k^{-(\gamma+1)}) \\
&= \frac{\mu^{-}_{trans}}{2(\gamma-1)}\left(1-(1-m)^{\gamma-1}\right)\left[(k+1)n_{out}(k+1) - (k-1)n_{out}(k-1)\right] + O(k^{-(\gamma+1)})
\end{aligned} \tag{2.21}$$

Observe that, when $\gamma = 1$ the right hand side of equation (2.19) is zero. To derive this, first note from equation (2.18) that $\sum_{j=0}^{k}\Pi^{out}_{TF-}(k,j)n_{out}(k) = \mu^{-}_{trans}n_{out}(k)$. We use Lemma 2 of [19] to show that $\sum_{j=k}^{N}\Pi^{out}_{TF-}(j,k)n_{out}(j) = \mu^{-}_{trans}n_{out}(k)(1-m)^{\gamma-1} + O(k^{-(1+\gamma)}$. To see this, we have

$$\begin{aligned}
\sum_{j=k}^{N}\Pi^{out}_{TF-}(j,k)n_{out}(j) &= A_{out}\mu^{-}_{trans}\sum_{j=k}^{N}\binom{j}{j-k}m^{j-k}(1-m)^k j^{-\gamma} \\
&= A_{out}\mu^{-}_{trans}k^{-\gamma}(1-m)^k\sum_{j=k}^{N}\binom{j}{j-k}m^{j-k}\left(\frac{k}{j}\right)^{\gamma} \\
&= A_{out}\mu^{-}_{trans}k^{-\gamma}(1-m)^k\left(1+O(k^{-1})\right)\sum_{j=k}^{N}\binom{j-\gamma}{j-k}m^{j-k} \\
&= A_{out}\mu^{-}_{trans}k^{-\gamma}(1-m)^k\left(1+O(k^{-1})\right)\sum_{i=0}^{N-k}\binom{i+k-\gamma}{i}m^{i} \\
&= A_{out}\mu^{-}_{trans}k^{-\gamma}(1-m)^k\left(1+O(k^{-1})\right)(1-m)^{\gamma-k-1} \\
&= \mu^{-}_{trans}n_{out}(k)(1-m)^{\gamma-1}\left(1+O(k^{-1})\right)
\end{aligned} \tag{2.22}$$

valid for $N \gg k$ (in fact, exactly valid in the limit $N \to \infty$). We now have

$$\sum_{j=k}^{N} \Pi^{out}_{TF-}(j,k) n_{out}(j) - \sum_{j=0}^{k} \Pi^{out}_{TF-}(k,j) n_{out}(k)$$
$$= \mu^{-}_{trans} n_{out}(k)(1-m)^{\gamma-1} - \mu^{-}_{trans} n_{out}(k) + O(k^{-(\gamma+1)}). \tag{2.23}$$

Using our assumed solution form we can also write

$$(k+1)n_{out}(k+1) - kn_{out}(k) + kn_{out}(k) - (k-1)n_{out}(k-1)$$
$$= 2(1-\gamma)n_{out}(k) + O(k^{-(\gamma+1)}) \tag{2.24}$$

Equations (2.20) and (2.21) combine to give equation (2.18). For large k we can neglect terms $O(k^{-(\gamma+1)})$ and define $K(\gamma, m) = \frac{1}{2(\gamma-1)}\left(1-(1-m)^{\gamma-1}\right)$. This allows us to write

$$\sum_{j=k}^{N} \Pi^{out}_{TF-}(j,k) n_{out}(j) - \sum_{j=0}^{k} \Pi^{out}_{TF-}(k,j) n_{out}(k)$$
$$= \mu^{-}_{trans} K(\gamma, m)\left[(k+1)n_{out}(k+1) - (k-1)n_{out}(k-1)\right], \tag{2.25}$$

which is the form used in equation (2.5).

For the model including degree dependence in the rate of *trans* evolution (Models 2 and 4), $\Pi^{out}_{TF-}(j,k)$ is given by

$$\Pi^{out}_{TF-}(j,k) = \mu^{-}_{trans} \frac{j!}{k!(j-k)!} \frac{m^{j-k}(1-m)^{k}}{j} \tag{2.26}$$

Once again assuming a solution of the form $n_{out} = A_{out}k^{\gamma}$, we can use the approximation

$$\begin{aligned}
&\sum_{j=k}^{N} \Pi^{out}_{TF-}(j,k) n_{out}(j) - \sum_{j=0}^{k} \Pi^{out}_{TF-}(k,j) n_{out}(k) \\
&= \mu^{-}_{trans} \frac{n_{out}(k)}{k} (1-m)^{\gamma} - \mu^{-}_{trans} \frac{n_{out}(k)}{k} + O(k^{-(\gamma+2)}) \\
&= \frac{\mu^{-}_{trans}}{2\gamma} \left(1-(1-m)^{\gamma}\right) \left[n_{out}(k+1) - n_{out}(k-1)\right] + O(k^{-(\gamma+2)})
\end{aligned} \tag{2.27}$$

To derive this, first note that in this case $\sum_{j=0}^{k} \Pi^{out}_{TF-}(k,j) n_{out}(k) = \mu^{-}_{trans} \frac{n_{out}(k)}{k}$. Again we use Lemma 2 of [19] to show that $\sum_{j=k}^{N} \Pi^{out}_{TF-}(j,k) n_{out}(j) = \mu^{-}_{trans} \frac{n_{out}(k)}{k} (1-m)^{\gamma} + O(k^{-(2+\gamma)}$. To see this, we have

$$\begin{aligned}
\sum_{j=k}^{N} \Pi^{out}_{TF-}(j,k) n_{out}(j) &= A_{out} \mu^{-}_{trans} \sum_{j=k}^{N} \binom{j}{j-k} m^{j-k} (1-m)^{k} j^{-(1+\gamma)} \\
&= A_{out} \mu^{-}_{trans} k^{-(1+\gamma)} (1-m)^{k} \sum_{j=k}^{N} \binom{j}{j-k} m^{j-k} \left(\frac{k}{j}\right)^{1+\gamma} \\
&= A_{out} \mu^{-}_{trans} k^{-(1+\gamma)} (1-m)^{k} \left(1+O(k^{-1})\right) \sum_{j=k}^{N} \binom{j-\gamma-1}{j-k} m^{j-k} \\
&= A_{out} \mu^{-}_{trans} k^{-(1+\gamma)} (1-m)^{k} \left(1+O(k^{-1})\right) \sum_{i=0}^{N-k} \binom{i+k-\gamma-1}{i} m^{i} \\
&= A_{out} \mu^{-}_{trans} k^{-(1+\gamma)} (1-m)^{k} \left(1+O(k^{-1})\right) (1-m)^{\gamma-k} \\
&= \mu^{-}_{trans} \frac{n_{out}(k)}{k} (1-m)^{\gamma} \left(1+O(k^{-1})\right)
\end{aligned} \tag{2.28}$$

valid for $N \gg k$ (in fact, exactly valid in the limit $N \to \infty$) We now have

$$\begin{aligned}
&\sum_{j=k}^{N} \Pi^{out}_{TF-}(j,k) n_{out}(j) - \sum_{j=0}^{k} \Pi^{out}_{TF-}(k,j) n_{out}(k) \\
&= \mu^{-}_{trans} \frac{n_{out}(k)}{k} (1-m)^{\gamma} - \mu^{-}_{trans} \frac{n_{out}(k)}{k} + O(k^{-(\gamma+2)}).
\end{aligned} \tag{2.29}$$

Using our assumed solution form, we can also write

$$\begin{aligned} n_{out}(k+1) - n_{out}(k) + n_{out}(k) - n_{out}(k-1) \\ = 2\gamma \frac{n_{out}(k)}{k} + O(k^{-(\gamma+2)}) \end{aligned} \tag{2.30}$$

Equations (2.26) and (2.27) combine to give equation (2.24). For large k we can neglect terms $O(k^{-2+\gamma})$ and higher, and define $K(\gamma, m) = \frac{1}{2\gamma}\left(1 - (1-m)^{\gamma}\right)$. This allows us to write

$$\begin{aligned} \sum_{j=k}^{N} \Pi^{out}_{TF-}(j,k) n_{out}(j) - \sum_{j=0}^{k} \Pi^{out}_{TF-}(k,j) n_{out}(k) \\ = \mu^{-}_{trans} K(\gamma, m) \left[n_{out}(k+1) - n_{out}(k-1) \right], \end{aligned} \tag{2.31}$$

which is the form used in equation (2.6).

2.6.2 Solution for *in*- and *out*-degree Distributions

The procedure used to calculate the equilibrium degree distributions is the same for all four models. The method is illustrated for Model 1. The solution to equation (2.2) for the *in*-degree distribution of this model, using the incoming edge event probabilities from Table 2, gives an equilibrium degree distribution

$$n_{in}(k) = A_{in} \frac{\Gamma\left(\frac{\mu^{+}_{cis} + \mu^{+}_{trans}}{D} + k\right)}{\Gamma(1+k)} \left(\frac{D}{D + \mu^{-}_{cis} + m\mu^{-}_{trans}} \right)^{k} \tag{2.32}$$

where A_{in} is a normalization constant. Following [19] we can write

$$\frac{\Gamma(k-c)}{\Gamma(k)} = \left(1 + O(k^{-1})\right) k^{-c} \tag{2.33}$$

For large k, terms $O(k^{-1}$ can be neglected, and equation(2.29) can be written in the form given in equation (2.7).

The solution to equation (2.5) for the *out* degree distribution of this model, using the outgoing

edge event probabilities from Table 2, gives an equilibrium degree distribution

$$n_{out}(k) = A_{out} \frac{\Gamma\left(\frac{\mu^+_{cis}+\mu^+_{trans}}{D-\mu^-_{trans}K(\gamma,m)} + k\right)}{\Gamma(1+k)} \left(\frac{D-\mu^-_{trans}K(\gamma,m)}{D+\mu^-_{cis}+\mu^-_{trans}K(\gamma,m)}\right)^k \tag{2.34}$$

Using the approximation given in equation (2.30), this can be written in the form given in equation (2.8). The same procedure gives equations (2.9)-(2.14), the solutions to the remaining three models.

2.6.3 Shrinking Networks

We now consider a model of a shrinking network, in which the rate of gene deletion is greater than the rate of gene duplication. This model is appropriate as a model of transcription network evolution immediately following a whole genome duplication, such as that which occurred in yeast around 100 million years ago [62]. We use a rate of gene duplication D^+, and gene deletion D^-, such that

$$D^- = D^+ + \Delta D \tag{2.35}$$

where $\Delta D > 0$. Firstly note that, the rate at which genes gain new edges through duplication of other genes is kD^+, and the rate at which they lose edges through deletion of other genes is $kD^- = kD^+ + k\Delta D$. The rate at which new TFs of *out*-degree k are produced by this model is $D^+ n_{out}(k)$, and the rate at which they are lost is $D^- n_{out}(k)$. Therefore TFs with *out*-degree k are lost at a rate $\Delta D n_{out}(k)$. Similarly, TGs with *in*-degree k are lost at a rate $\Delta D n_{in}(k)$.

To see that this term is not sufficient produce an *out*-degree distribution with exponent $\gamma > 1$, we make the following approximation. Assuming an *out*-degree of the form $n_{out}(k) = A_{out}k^{-\gamma}$ we can write using equation (2.21)

$$\Delta D n_{in}(k) = \frac{\Delta D}{2(\gamma-1)} \left[(k+1)n_{out}(k+1) - (k-1)n_{out}(k-1)\right] + O(k^{-(1+\gamma)}) \tag{2.36}$$

Using this with Model 1, we now define $K(\gamma,m) = \frac{1}{2(\gamma-1)}\left(1 + \frac{\Delta D}{\mu^-_{trans}} - (1-m)^{\gamma-1}\right)$. Then the solution for the *out*-degree distribution of this model can be written as

$$n_{out}(k) = A_{out} \frac{\Gamma\left(\frac{\mu^+_{cis}+\mu^+_{trans}}{D^+-\mu^-_{trans}K(\gamma,m)} + k\right)}{\Gamma(1+k)} \left(\frac{D^- - \mu^-_{trans}K(\gamma,m)}{D^+ + \mu^-_{cis} + \mu^-_{trans}K(\gamma,m)}\right)^k \tag{2.37}$$

which can be approximated to

$$n_{out}(k) \propto k^{-\gamma} \exp[-\beta k]$$

where

$$\beta = \ln\left(\frac{\mu^-_{cis} + D^- + \mu^-_{trans}K(\gamma,m)}{D^+ - \mu^-_{trans}K(\gamma,m)}\right),$$

$$\gamma = 1 - \frac{\mu^+_{cis} + \mu^+_{trans}}{D^+ - \mu^-_{trans}K(\gamma,m)} \tag{2.38}$$

A power-law degree distribution requires $\beta \approx 0$. This occurs if $\mu^-_{cis} + \Delta D = -2\mu^-_{trans}K(\gamma,m)$. If $D^+ > D^-$, $K(\gamma,m)$ is always positive and this equality cannot be satisfied. Therefore shrinking networks cannot produce networks with a power-law degree distribution.

2.6.4 Growing Networks

For growing networks we may use the same model developed in the previous section, with $\Delta D < 0$. In this case we can have $K(\gamma,m) < 0$ and networks with a power-law *out*-degree distribution can be produced.

Moreover, from equation (2.35), we can see that if $-\mu^-_{trans}K(\gamma,m) > \mu^-_{cis} + D^-$, networks with a power-law degree distribution with $\gamma > 1$ can be produced. However, such a model requires continuous network growth. Once the network stops growing, the equilibrium degree distribution will move away from a power-law, to that given by Model 1. In order to sustain a power-law *out*-degree distribution with exponent $\gamma > 1$ throughout evolution, degree dependence in the rate of TF evolution is required.

Chapter 3

Neutral Evolution of Cooperative TF Binding

Transcription regulation can occur in a number of ways. The most basic mechanism is the binding of a single transcription factor to a specific binding site in the promoter region of a regulated gene. In addition to this, transcription factor proteins may interact to facilitate or prevent each other binding to regulated genes. Observations in the yeast transcription network have revealed that the evolution of such pairs of co-regulating transcription factors can have complex dynamics. In particular, the yeast sex determination network appears to have undergone a significant degree of neutral rewiring. This consists of the gain of a protein-protein interaction between co-regulating transcription factors, accompanied by changes to the binding sites present at multiple target genes. Despite these changes, the function of the network has remained unchanged. We constructed a model for the neutral evolution of pairs of transcription factors which co-regulate sets of target genes. We assumed transcription factors were able to to gain a protein-protein interaction, which allowed them to co-operatively bind to their targets. This was assumed to occur through a *trans* mutation at one of the transcription factors. In addition, we assumed that *cis* mutations, which changed the strength of specific binding sites for the transcription factors at each of the regulated genes, were able to occur. We showed that the probability of a protein-protein interaction becoming fixed in a population follows a (soft) threshold function in the number of regulated genes. When the number of regulated genes is below the threshold, a protein-protein interaction is almost entirely absent from the population. When it is greater than the threshold, a protein-protein interaction is close to fixation. The position of the threshold is determined by the rate of *cis* and *trans* mutations,

as well as the size of the population being considered. These results are used to account for the observed neutral rewiring of the yeast transcription network.

3.1 Background

Transcription regulation lies at the heart of many of the most interesting and important evolutionary questions currently facing biologists. It is key to determining the expression levels of individual genes, and the co-expression of sets of genes. Changes to regulatory interactions are capable of producing changes to gene expression on the scale of a single gene or of a large fraction of the genome [52]. Developing an understanding of the mechanisms through which transcription networks evolve is therefore an important challenge.

Evolution of transcription regulation occurs through a variety of mechanisms. A great deal of debate has focused on the relative contributions of evolutionary change of regulatory binding sites in promoter regions (*cis* evolution) [96, 97] and of the regulatory proteins themselves (*trans* evolution) [73, 126, 125, 136]. Some authors have claimed a predominance of *cis*-regulatory changes because of the expectation that trans mutations will tend to have negative pleiotropic effects, whereas *cis* mutations do not [16, 96, 115, 136, 142]. However, a number of recent studies have challenged this position and reported many cases in which trans evolution, along with cis evolution, plays an important role, [70, 73, 125].

Some of the most striking evidence for the role played by changes in trans has been provided by studies of single-celled yeasts [124, 126, 125]. These studies have focused on the evolution of combinatorial gene regulation, in which pairs of transcription factors co-bind to sets of target genes. Co-regulation is found to occur either (i) through the presence of binding sites for both transcription factors in the promoter regions of target genes, or (ii) through trans interactions between the two transcription factors which allows one to facilitate the binding of the other at the target genes. Comparison of regulatory circuits in the ascomycete yeast species *Saccharomyces cerevisiae*, *Kluyveromyces lactis* and *Candida albicans* reveal substantial changes to the transcription factors involved in co-regulation, as well as to the target genes they regulate [125]. These changes have involved the loss and gain of transcription factor binding sites (*cis* evolution), as well as the loss and gain of trans interactions between transcription factors (*trans* evolution). The changes observed in yeast transcription circuits are not necessarily correlated with changes to the regulatory logic of those circuits rewiring of some regulatory interactions has occurred, but the input and output of the network has remained the same. This leads to the suggestion that such rewiring of transcription

networks may occur neutrally [124, 126].

In this chapter, I consider how neutral evolution of transcription circuits occurs when sets of target genes are co-regulated by a pair of transcription factors. In contrast to most models of transcription network evolution, we include population genetic details along with details of network structure in our model. Such details are necessary if the evolution of gene networks through non-adaptive processes is to be properly characterized and understood [72, 116]. I use this model to investigate the conditions under which a trans interaction between two co-regulating transcription factors is maintained in a population and how this depends on the number of target genes co-regulated. We then investigate how the presence or absence of a trans interaction between co-regulating transcription factors alters the level of genetic variation in a population and the ability of a species to adapt to changing environments. We determine how population genetic details, such as population size and rates of deleterious mutations influence the evolution of co-regulated transcription networks. I characterize the dynamics of neutral evolution in such networks, that is how a change to one part of a network can have knock on effects, resulting in changes to other parts. We also investigate how mutations in cis and in trans interact, and determine whether changes in cis can drive changes in trans and vice versa. I finally apply this analysis to account for differences in the way genes are co-regulated in related yeast species.

3.2 Model

3.2.1 Regulation of a Single Target

We model the evolution of cooperative binding between a pair of transcription factors, A and B in a haploid organism. We assume that the *trans* interaction allows A to cooperatively bind B at the target genes of A. Therefore when a *trans* interaction is present both A and B bind to target genes which have an unmutated binding site for A, even if the binding site for B has a mutation. We assume that the *trans* interaction is asymmetrical. Therefore B is unable to bind A to genes with an unmutated binding site for B, even when a *trans* interaction between the two is present (figure 3.1).

To avoid complications of modelling the separate evolution of A and B, we treat the interaction between A and B as a single locus, referred to as the *trans* locus. Two alleles are associated with the *trans* locus: t^+ when the *trans* interaction between A and B is present and t^- when it is absent. Initially we discuss a system in which only a single target gene is regulated (figure 3.1).

We are interested in a system in which both transcription factors A and B are required to regulate the target gene. This may occur in two ways. i) When a *trans* interaction is absent, A and B must bind to the target independently. This means that the target gene must have unmutated binding sites for both A and B. ii) When a *trans* interaction is present, A and B act cooperatively. This means that the target gene only requires an unmutated binding site for A.

We treat the structure of binding sites in the promoter region of the target gene as a single locus. This is referred to as the *cis* locus. The *cis* locus has four possible alleles associated with the presence or absence of a mutation at a binding site for A and/or B. a^+ corresponds to a binding site for A which does not have a mutation, while a^- corresponds to a binding site with a mutation. Similarly b^+ corresponds to an unmutated binding site for B being present, while b^- corresponds to a mutated binding site. The four alleles associated with the *cis* locus are denoted a^+b^+, a^+b^-, a^-b^+ and a^-b^-. Our model for a single gene therefore consists of two loci.

3.2.2 Mutation and Selection

Mutations may occur in our model at both the *trans* and the *cis* loci. Mutations resulting in the gain of a *trans* interaction (from allele t^- to allele t^+) occur at rate $\mu)trans^+$, and loss of this interaction (from allele t^+ to allele t^-) occur at rate $\mu)trans^-$. At the *cis* locus, deleterious mutations at a binding site for A (from a^+ to a^-) occur at rate μ_a^-, whilst back mutations at a binding site for A (from a^- to a^+) occur at rate μ_a^+. Similarly, deleterious mutations and back mutations at a binding site for B occur at rate μ_b^- and μ_b^+ respectively.

The fitness of different genotypes is given in Table 3.1. We assume that when a *trans* interaction is absent, allele a^+b^+ has fitness $w = 1$. When a binding site for either A or B (or both) has a mutation, a^+b^-, a^-b^+ or a^-b^- we assume a fitness reduction of s. When a *trans* interaction is present, we also assume that allele a^+b^+ has fitness $w = 1$. When only a binding site for B has a mutation, a^+b^-, we assume that cooperative binding between A and B prevents any loss of fitness. When a binding site for A has a mutation, a^-b^+ or a^-b^- we assume a fitness reduction of s. This fitness scheme is laid out in Table 3.1. It is clear from this that the presence of a *trans* interaction buffers against mutations to a binding site for B.

3.2.3 Regulation of Multiple Targets

We wish to consider situations in which multiple target genes are co-regulated by the same pair of transcription factors. We assume that A and B co-regulate N target genes. Each target gene

	t^+	t^-
a^+b^+	1	1
a^+b^-	1	$1-s$
a^-b^+	$1-s$	$1-s$
a^-b^-	$1-s$	$1-s$

Table 3.1: Fitness scheme for regulation of a single target gene in a population of asexual, haploid organisms

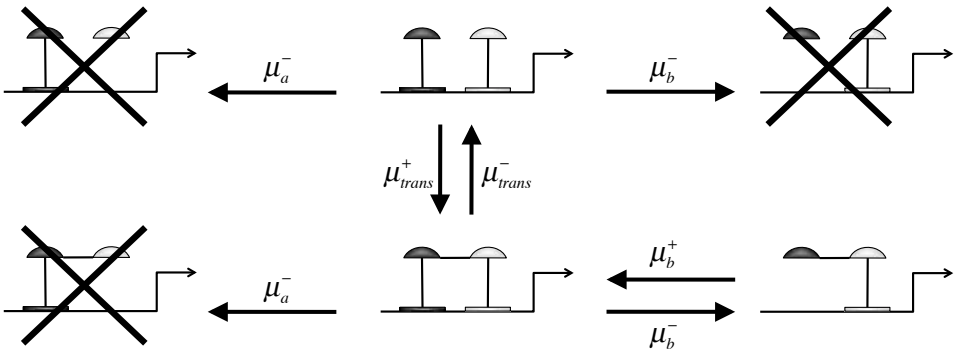

Figure 3.1: Selection scheme used in this model for regulation of a single target. Addition of a *trans* interaction between TF A (white semicircle) and B (black semicircle), buffers against changes in *cis* to the binding site of B. The possible genotypes as depicted here are (from top left, clockwise) - $t^-a^-b^+$, $t^-a^+b^+$, $t^-a^+b^-$, $t^+a^+b^-$, $t^+a^+b^+$, $t^+a^-b^+$. Unfit genotypes are indicated with a cross. Possible mutations, and the rates at which they occur, are indicted with black arrows. Back mutations form unfit genotypes to fit genotypes are not indicated, since they are neglected in our analysis.

has binding sites for A or B that are independent of the other $N-1$ target genes. If a *trans* interaction is present between A and B, each target gene is free to suffer a mutation at a binding site for B without a reduction in fitness. Each of the target genes may have four different genotypes associated with its promoter region (corresponding to the presence or absence of a mmutation at binding sites for A and B). Since these are assumed to be independent, this gives 2^{2N} possible genotypes associated with the promoter regions of the target genes. In addition, there are two possible genotypes associated with the *trans* locus. The space of possible genotypes, G, therefore has size $|G| = 2^{2N+1}$.

We define the network of possible genotypes, with vertices corresponding to a genotype and edges corresponding to possible mutations between genotypes. We may also define the subset $g \in G$

of genotypes which have maximum fitness. This is referred to as the neutral genotype space, since each genotype in the space can be adopted without loss of fitness. In our model, this corresponds to genotypes in which all N target genes are regulated properly by both transcription factors A and B. The network of neutral genotypes then has vertices corresponding to fit genotypes and edges corresponding to possible mutations between fit genotypes. When a *trans* interaction is present, there are 2^N possible fit genotypes, since there are two possible alleles (a^+b^+ and a^+b^-) at each of the N target genes. When a trans interaction is absent there is only one fit allele (a^+b^+) at each of the N target genes. The size of the neutral genotype space, g, is therefore $|g| = 2^N + 1$.

Rather than following all genotypes, we simplify by calculating the frequency of genotype classes in which k target genes have a mutation at a binding site for B. This is justified by the assumption that *cis* mutation rates are the same for all N regulated target genes, and the fitness effect of a mutation is the same for any locus. A genotype with k mutations, mutates to a genotype with $k+1$ mutations at rate $(N-k)\mu_b^-$ and to a genotype with $k1$ mutations at rate $k\mu_B^+$. The number of genotypes belonging to the neutral space is thus reduced to $g = N + 2$. The neutral space consists of $N + 1$ genotypes in which a *trans* interaction is present and between 0 and N target genes have a mutation, and one genotype in which a *trans* interaction is absent. We now use this model to determine the circumstances under which a *trans* interaction will become fixed in a population.

3.3 Results

3.3.1 Infinite Population Model

We determine the distribution of genotypes over g in an infinite population of haploid, asexual organisms. Let the fraction of the population lying on g at equilibrium be P, and the mean fitness of the population be $\bar{w}$. At equilibrium, we have

$$P = \frac{\langle \nu \rangle}{\bar{w}} P + Q \tag{3.1}$$

where $\langle \nu \rangle$ is the fraction of P, that, under mutation, remains on g between successive generations (i.e the fraction of P which do *not* fall off g through mutation). Q is the rate at which individuals outside of g mutate onto g [127]. We assume that any genotype lying outside g has markedly lower fitness than those belonging to g, and that the mutation rates are small enough that the majority of the population lies on g [127]. Therefore we may assume that the contribution of Q to equation

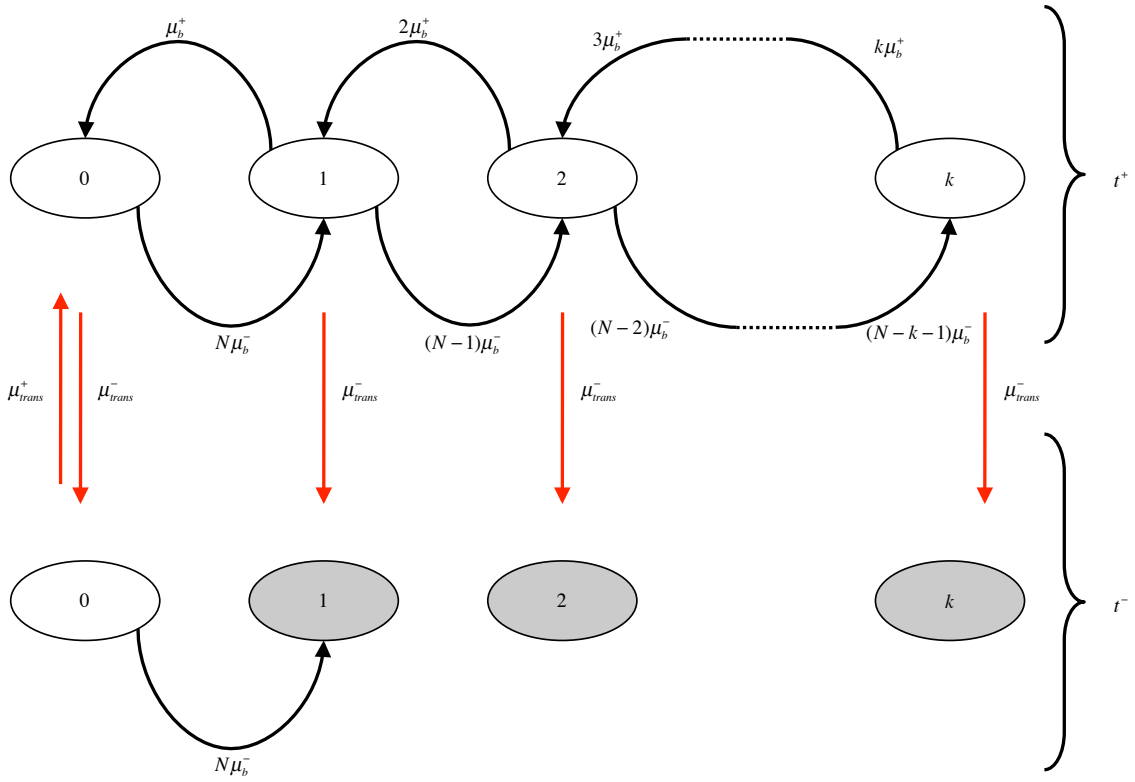

Figure 3.2: Regulation of multiple targets. Ovals indicate possible genotypes. Top row indicates genotypes with a *trans* interaction, bottom row indicates genotypes lacking a *trans* interaction. Numbers indicate the number of target genes with a mutation in *cis*. White ocals are fit genotypes (those which lie on g), grey ovals indicate unfit genotypes. Black arrows give possible *cis* mutations and the rates at which they occur. Red arrows indicate possible *trans* mutations and the rates at which they occur. Back mutations form unfit genotypes to fit genotypes are not indicated, since they are neglected in our analysis.

(3.1) is sufficiently small that it can be neglected. In this case, the mean fitness of the population is $\bar{w} = \langle \nu \rangle$. For the neutral network we are considering, $\langle \nu \rangle$ can be calculated simply as follows. Let P_- be the fraction of P that lacks a *trans* interaction, and $P_+(k)$ be the fraction of P that has a *trans* interaction and in which k targets have a mutation at a binding site for B. We then have

$$\langle \nu \rangle = 1 - N\mu_a^- - N\mu_b^- P_- - \mu_{trans}^- \sum_{k=1}^{N} P_+(k) \tag{3.2}$$

Where $N\mu_a^- + N\mu_b^- P_- + \mu_{trans}^- \sum_{k=1}^{N} P_+(k)$ is the fraction of P which mutates off g. The first term comes from mutations at binding sites for A. The second term comes from individuals which lack a *trans* interaction, undergoing mutations at a binding site for B. The third term comes from

individuals which have a *trans* interaction and in which at least one target has a mutation at a binding site for B, undergoing a mutation resulting in the loss of the *trans* interaction between A and B.

The equation for the evolution of P_- can then be written as follows

$$\bar{w}P_-^{'} = P_-(1 - N\mu_a^- - N\mu_b^- - \mu_{trans}^+) + \mu_{trans}^- P_+(0) \tag{3.3}$$

Similarly, the equation for the evolution of and $P^+(k)$ can be written as

$$\bar{w}P_+^{'}(0) = P_+(0)(1 - N\mu_a^- - N\mu_b^- - \mu_{trans}^-) + \mu_b^+ P_+(1) + \mu_{trans}^+ P_- \tag{3.4}$$

for $k = 0$ and

$$\begin{aligned}\bar{w}P_+^{'}(k) = P_+(k)(1 - N\mu_a^- - (N-k)\mu_b^- - k\mu_b^+ - \mu_{trans}^-) \\ +(k+1)\mu_b^+ P_+(k+1) + (N-k+1)\mu_b^- P_+(k-1)\end{aligned} \tag{3.5}$$

for $k > 0$.

In order to solve this, we write $P_+(k > 0) = \sum_{k=1}^{N} P_+(k)$, and take the sum of both sides of equation (3.5) to give

$$\bar{w}P_+^{'}(k > 0) = P_+(k > 0)(1 - N\mu_a^- - \mu_{trans}^-) - \mu_b^+ P_+(1) + N\mu_b^- P_+(0) \tag{3.6}$$

We now make the simplifying assumption that $\mu_b^+ P_+(1) = 0$. This is valid provided that only a small fraction of the population lies at $P_+(1)$ (see Appendix B). Therefore equations (3.4) and (3.6) can be written as

$$\bar{w}P_+^{'}(0) = P_+(0)(1 - N\mu_a^- - N\mu_b^- - \mu_{trans}^-) + \mu_{trans}^+ P_- \tag{3.7}$$

and

$$\bar{w}P_{+}^{'}(k>0) = P_{+}(k>0)(1 - N\mu_{a}^{-} - \mu_{trans}^{-}) + N\mu_{b}^{-}P_{+}(0) \tag{3.8}$$

Equations (3.3), (3.7) and (3.8) can now be solved explicitly to find the equilibrium distribution of the population on g. We calculate the frequency of a *trans* interaction between A and B in the population, $P_{+} = \sum_{k=0}^{N} P_{+}(k)$. This gives

$$P_{+} = \frac{\mu_{trans}^{+}}{\mu_{trans}^{+} + \mu_{trans}^{-} - N\mu_{b}^{-}}$$

for $\mu_{trans}^{-} \geq N\mu_{b}^{-}$ and

$$P_{+} = 1 \tag{3.9}$$

otherwise. Equation (3.9) gives a good approximation for the frequency of the *trans* interaction in the population, for values of $\mu_{cis}^{+} \leq \mu_{cis}^{-}$ (figure 3.1).

Equation (3.9) says that the frequency of the *trans* interaction between the two transcription factors A and B follows a threshold function in N. If the number of target genes (N) is greater than the threshold, $N > \frac{\mu_{trans}^{-}}{\mu_{b}^{-}}$, then the *trans* interaction between A and B is fixed in the population. If the number of target genes is less than the threshold, the *trans* interaction is lost from the population. When the number of target genes is greater than the threshold, such that the *trans* interaction is fixed in the population, the equilibrium genotype distribution is given by

$$P_{+}(k) = \binom{N}{k} \left(\frac{\mu_{b}^{-}}{\mu_{b}^{+} + \mu_{b}^{-}}\right)^{k} \left(\frac{\mu_{b}^{+}}{\mu_{b}^{+} + \mu_{b}^{-}}\right)^{N-k} \tag{3.10}$$

This is a binomial distribution. Therefore the mean number of target genes which have a mutation at a binding site for B in the population is given by the mean of the distribution, $\frac{\mu_{b}^{-}}{\mu_{b}^{+}+\mu_{b}^{-}}$.

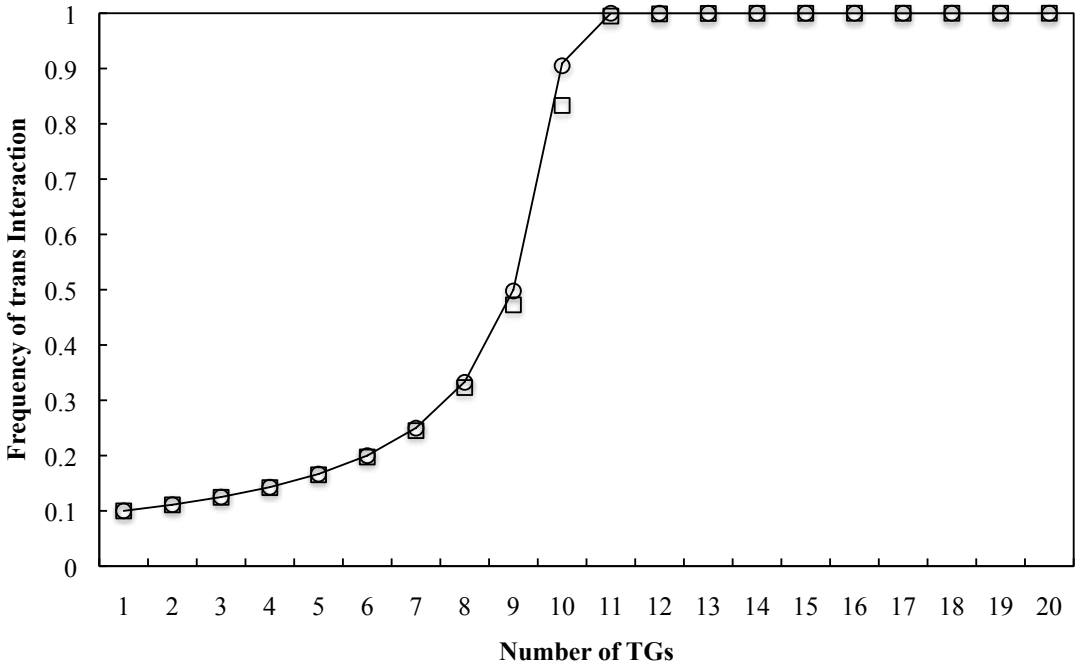

Figure 3.3: Frequency of *trans* interaction in an infinite population. Figure shows the variation in the frequency of the *trans* interaction with the number of target genes N. The relative mutation rates used are $\mu_b^- = 0.1\mu_{trans}^-$ (resulting in a threshold at $N = 10$) and $\mu_{trans}^+ = 0.1\mu_{trans}^-$. The solid line shows the frequecy of a *trans* interaction given by equation (3.9). Points show numerical solutions to equations (3.3)-(3.5). Squares show a mutation rate of $\mu_b^+ = \mu_b^-$ and circles show a mutation rate of $\mu_b^+ = 0.1\mu_b^-$.

3.3.2 Small Population Model

The results above apply to infinite populations, but also hold for large populations in which the product of the population size and the mutation rate is much greater than one; i.e. $M\mu_T \gg 1$ [127], where μ_T is the total rate at which a particular genotype undergoes a mutation. In this model we may take $\mu_T = N\mu_a^- + N\mu_b^- + \mu_{trans}^-$ as the maximum rate at which any genotype undergoes mutation. This follows, since in general the rate of mutations resulting in wakening of a binding site or loss of a *trans* interaction, are greater than the rate of mutations resulting in strengthening of a binding site of gain of a *trans* interaction; i.e. $\mu_a^- > \mu_a^+$, $\mu_b^- > \mu_b^+$ and $\mu_{trans}^- > \mu_{trans}^+$.

We now consider the evolution of small populations, in which $M\mu_T \ll 1$. When this condition is satisfied, the entire population converges onto a single genotype [127]. In this case we can model neutral evolution as the probability that the entire population moves from its current genotype to a neighbouring genotype on g, with a probability determined by the rate of mutation between those two genotypes. Deleterious mutations, resulting in the population moving off g, occur with probability zero. This assumes that deleterious mutations never become fixed in the population.

Let π_- be the probability that the population has a genotype that lacks a *trans* interaction, and $\pi_+(k)$ be the probability that the population has a genotype with a *trans* interaction, and in which k of N target genes have lost a binding site for B. Since we assume only mutations between genotypes belonging to g occur, the genotype of the population evolves according to

$$\pi_-^{'} = \pi_- \left(1 - \mu_{trans}^{+}\right) + \mu_{trans}^{-}\pi_+(0) \tag{3.11}$$

for genotypes lacking a *trans* interaction and

$$\pi_+(0)^{'} = \pi_+(0)\left(1 - N\mu_b^- - \mu_{trans}^-\right) + \mu_b^+\pi_+(1) + \mu_{trans}^+\pi_- \tag{3.12}$$

for genotypes with a *trans* interaction and $k = 0$, and

$$\pi_+(k)^{'} = \pi_+(k)\left(1 - k\mu_b^+ - (N-k)\mu_b^-\right) + (k+1)\mu_b^+\pi_+(k+1) + (N-k+1)\mu_b^-\pi_+(k-1) \tag{3.13}$$

for genotypes with a *trans* interaction and $k > 0$.

Equations (3.11)-(3.13) can be solved explicitly to find the equilibrium probability distribution (see Appendix B). The probability, $\pi_+ = \sum_{k=0}^{N}\pi_+(k)$, that the population contains a *trans* interaction between A and B is

$$\pi_+ = \frac{\frac{\mu_{trans}^+}{\mu_{trans}^-}\left(1 + \frac{\mu_b^-}{\mu_b^+}\right)^N}{1 + \frac{\mu_{trans}^+}{\mu_{trans}^-}\left(1 + \frac{\mu_b^-}{\mu_b^+}\right)^N} \tag{3.14}$$

Equation (3.14) is a sigma function in the number of TGs, N. This function is characterized by threshold-like behaviour (figure 3.2).

The threshold occurs at the value of N, N_{thresh}, for which $\pi_+ = 0.5$. N_{thresh}, is given by

$$N_{thresh} = \frac{\ln\left(\frac{\mu_{trans}^-}{\mu_{trans}^+}\right)}{\ln\left(1 + \frac{\mu_b^-}{\mu_b^+}\right)} \tag{3.15}$$

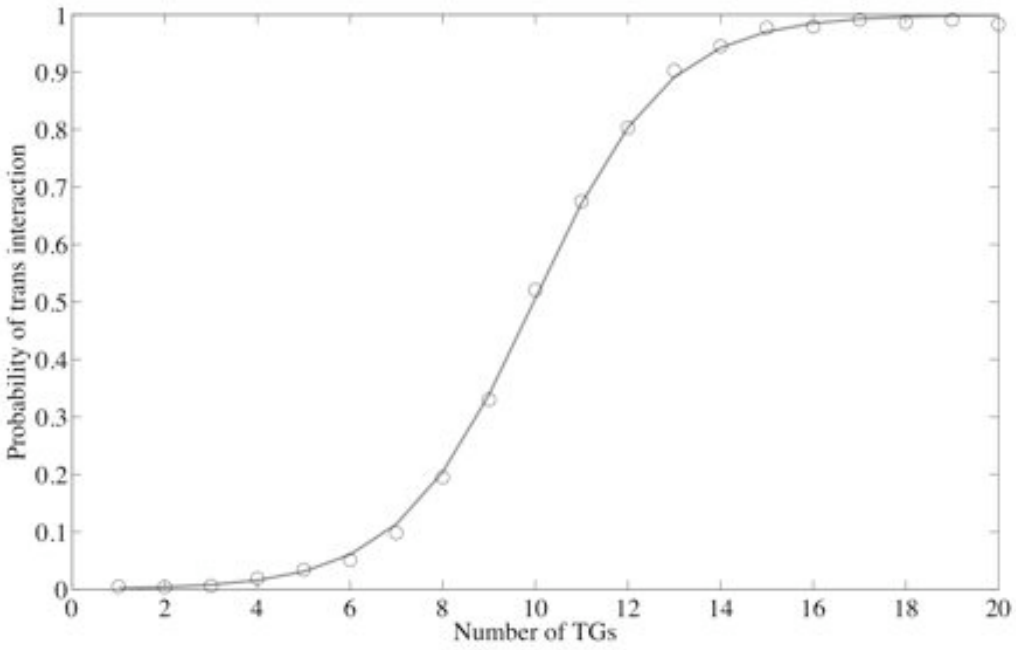

Figure 3.4: Frequency of *trans* interaction in a small population. Figure shows the variation in the frequency of the *trans* interaction with the number of target genes N. Solid line indicates the stationary distribution, given in equation (3.14). Circles indicate the results of simulations, using 1000 populations of 10000 individuals, with $M\mu_T \approx 0.1$. Values of $\mu^-_{trans} = 1000\mu^+_{trans}$ and $\mu^-_b = \mu^+_b$, to give a value of $N_{thresh} = 10$ and $\Delta N = 3$.

The steepness of the sigma function, and therefore the sharpness of the threshold behaviour, can be found by determining the slope of equation (3.14) at $N = N_{thresh}$. This yields

$$\left|\frac{d\pi_+}{dN}\right|_{N=N_{thresh}} = \frac{1}{4}\ln\left(1+\frac{\mu^-_b}{\mu^+_b}\right) \tag{3.16}$$

The range of N, ΔN, over which the sigma function moves from being close to zero, to being close to 1, is given by $\Delta N = \frac{4}{\ln\left(1+\frac{\mu^-_b}{\mu^+_b}\right)}$. When $N < N_{thresh}$ the probability of a *trans* interaction being present in the population is low, and tends to zero when $N < N_{thresh} - \frac{1}{2}\Delta N$. When $N > N_{thresh}$ the probability that a *trans* interaction is present in the population is high, and tends to 1 when $N > N_{thresh} + \frac{1}{2}\Delta N$.

We have shown that the probability of a *trans* interaction being present is a sigma function in the number of TGs, N. The threshold above which a *trans* interaction is likely to be present depends both on the ratio of forward and back mutations in *cis*, and on the ratio of the rates of

gain and loss of a *trans* interaction between A and B. The steepness of the threshold depends only on the ratio of forward and back mutations in *cis*. In the case where gain of *trans* interactions and strengthening of binding sites for B occur very rarely, $\mu^{+}_{trans} \to 0$ and $\mu^{+}_{b} \to 0$, the threshold, equation (3.15), is given by $N_{thresh} \to 1$ and the steepness of the threshold, equation (3.16) tends to infinity. In this case the frequency of *trans* interactions follows a step function, such that, if more than a single TG is co-regulated by A and B, a *trans* interaction will be completely fixed in the population.

3.3.3 Permanent fixation of a *trans* interaction

We now consider the probability that a *trans* interaction becomes permanently fixed in the population. In order to do this we assume that the system contains two absorbing states; that in which a *trans* interaction and none of the N target genes have a mutation at the binding site for B, π_{-}, and that in which a *trans* interaction is present and all N target genes have suffered a mutation to the binding site for B, $\pi_{+}(N)$. We assume that the system begins in a state in which a *trans* interaction is present, and none of the N target genes have suffered a mutation at the binding site for B, $\pi_{+}(0)$. We calculate the probability, ρ_{+}, that system reaches the sate $\pi_{+}(N)$

$$\rho_{+} = \frac{1}{1 + \frac{\mu^{-}_{trans}}{\mu^{-}_{b}} \sum_{k=1}^{N-1} \frac{k!(N-k-1)!}{N!} \left(\frac{\mu^{+}_{b}}{\mu^{-}_{b}}\right)^{k}} \tag{3.17}$$

(see Appendix B).

When $\mu^{-}_{b} \geq \mu^{+}_{b}$, ρ_{+} is an increasing function of N. When $\mu^{-}_{b} \approx \mu^{+}_{b}$, ρ_{+} is increasing for small N and decreasing for large N - there is a finite value of N for which ρ_{+} is maximum. When $\mu^{-}_{b} \gg \mu^{+}_{b}$, ρ_{+} is a decreasing function of N figure 3.3. Therefore whether a *trans* interaction becomes fixed when is an absorbing state depends on the number of target genes regulated and the ratio of loss and gain of binding sites for B. It is therefore clear that if a pair of transcription factors undergo a change to the number of target genes they regulate this will change the probability that they evolve a *trans* interaction between them. In the case that gain is much less frequent than loss, increase in the number of targets increases the probability of a *trans* interaction becoming fixed. However, when gain and loss occur at similar rates (as may occur if, for example, change to a single nucleotide is sufficient to constitute loss of a binding site), the probability of fixing a *trans* interaction is maximum for a finite number of target genes.

3.3.4 Recombination

The low rate of out-crossing in *S. cerevisiae* [107] has led us to consider asexual populations in which recombination does not occur. However, recombination occurs at much greater rates in other organisms and has the potential to significantly alter our results. A qualitative understanding of how greater rates of recombination influence evolution in our model can be gained as follows: For simplicity, consider mating between haploid organisms. We assume the N target genes are in linkage equilibrium. Suppose an organism, which has a *trans* interaction and in which k of the N target genes have a mutation at a binding site for B mates with another, which has a *trans* interaction and in which j of the N target genes have a mutation at a binding site for B. Provided no mutation occurs, the offspring of this mating will have a *trans* interaction. If we assume the mutated binding sites for B are distributed randomly between the N target genes, the probability that i of the N target genes lack a binding site for B in the offspring will be $\binom{N}{k}\left(\frac{1}{2}\frac{k}{N}+\frac{1}{2}\frac{j}{N}\right)^{i}\left(1-\frac{1}{2}\frac{k}{N}-\frac{1}{2}\frac{j}{N}\right)^{N-i}$, with the expected number of targets lacking a binding site given by $\langle i\rangle = \frac{k+j}{2}$. Therefore the introduction of recombination means that mating between organisms with a *trans* interaction can substantially change the number of target genes that lack a binding site for B between parent and offspring.

Now consider the case in which an organism, which has a *trans* interaction and in which k of the N target genes have lost a binding site for B mates with an organism which lacks a *trans* interaction, and therefore does not lack a binding site for B at any of its target genes. In this case the probability that the offspring lacks binding sites for B at i of its target genes is $\binom{N}{i}\left(\frac{1}{2}\frac{k}{N}\right)^{i}\left(1-\frac{1}{2}\frac{k}{N}\right)^{N-i}$ and the expected number of targets lacking a binding site is $\langle i\rangle\frac{k}{2}$. However, there is now a probability of a half that the offspring lacks a *trans* interaction. Therefore the offspring will have reduced fitness unless $i=0$, which occurs with probability $\left(1-\frac{1}{2}\frac{k}{N}\right)^{N}$. As N increases, this probability declines exponentially, for a given value of k. Thus as N increases, matings between organisms which lack a trans interaction and those which have a *trans* interaction will result in unfit offspring 50% of the time. In the cases where the offspring of the mating are fit, a *trans* interaction is present. Therefore recombination will tend to favour the presence of a *trans* interaction as N increases.

This argument does not hold if we consider sexual reproduction in diploids. In diploids when an organism is heterozygous for the presence of a *trans* interaction, we may assume that this is insufficient to buffer against all mutations at *cis*. Thus heterozygotes will tend to have reduced fitness following a *cis* mutation, whilst homozygotes which have a *trans* interaction will not. This

is an example of under-dominance, in which the fitness of heterozygotes is reduced compared to that of homozygotes. Under-dominance results in a barrier to evolution, which will tend to prevent *trans* mutations becoming fixed in a population, even when they result in increased mutational robustness. As a result we may expect recombination in diploids to act to prevent fixation of a *trans* mutation in a population.

3.4 Discussion

We have constructed a simple model for the neutral evolution of co-operative binding in transcription factor networks. We constructed our model based on observed changes in the yeast transcription network [124, 126, 125], and have presented our results for a population of haploid, asexual organisms. Our results show that *trans* interactions between a pair of co-operatively binding transcriptions factors will be fixed in a large population when it leads to an increase in the mutational robustness of the network. A large population is defined as one in which the product of the population size M and the total rate of mutation μ_T, is much greater than 1. When this condition is satisfied, the population maintains genetic variation, with more than one member of the neutral genotype space g present in the population.

When a *trans* interaction is absent, the frequency of deleterious mutations is determined by the number of target genes multiplied by the rate at which binding sites for TF B suffer mutations. When a *trans* interaction is present, mutations at binding sites for B are buffered against, but loss of the *trans* interaction becomes deleterious (since target genes with a mutation at a binding site for B become incorrectly regulated). When the rate of mutations at binding sites for B is greater than the rate of loss of the trans interaction between A and B, $N\mu_b^- > \mu_{trans}^-$, the presence of a *trans* interaction leads to an increase in the mutational robustness of the network. Therefore as the number of target genes being regulated increases, there is a threshold at $N_{thresh} = \frac{\mu_{trans}^-}{\mu_b^-}$ above which a *trans* interaction becomes fixed.

For small populations, mutational robustness has little effect on the evolution of the network [127]. A small population is defined as one in which the product of the population size and the total mutation rate is less than 1 ($M\mu_T < 1$) [127]. When the population is small, it maintains little genetic variation, and will in general consist of only a single member of the neutral genotype space g. Deleterious mutations are assumed to never become fixed, and are not maintained in the population over time. The network adopted by a small population is determined only by the rates of mutation and the structure of the neutral genotype space, g. Intriguingly, our results show

that in this case the *trans* interaction will also become fixed in the population when the number of target genes is greater than a threshold, and be absent otherwise. However, unlike for large populations, in this case the threshold is not driven by mutational robustness. The threshold in this case is given by equation (3.15). It can be seen the position of the threshold depends in a more complicated way on the different mutation rates. In addition, the threshold is less steep than in the large population case. However, the same qualitative effect is observed in both the large and small population cases - a threshold in the number of target genes, N, above which a *trans* interaction is fixed and below which it is absent. In both cases we find that changes in regulon size (the number of TGs co-regulated by A and B), is sufficient to drive the gain or loss of a *trans* interaction in a population.

3.4.1 Accumulation of Genetic Variation

Changes to regulon size necessarily require some of the targets currently regulated by a pair of transcription factors to lose their binding sites, whilst others, which are not regulated, must gain new binding sites. In our model, when a *trans* interaction is absent from the population, all N target genes must have fully functional binding sites for A and B. However, when a *trans* interaction is present, this allows variation in *cis*, since the binding site for B may suffer mutation at each target gene. If a *trans* interaction is present in the population, the probability that an individual picked at random from the population has k mutations in *cis* is given by equation (3.10), (this holds for both large and small populations - see Appendix B).

In a large population, if an organism has a mutation at a binding site for B at k of its N target genes, this corresponds to $\binom{N}{k}$ possible genotypes, each occurring at equal frequency in the population. The presence of a *trans* interaction therefore masks a great deal of variation in *cis*. Loss of a *trans* interaction between A and B reveals this variation. Different combinations of target genes that have a mutation at their binding site for B are revealed. Thus the presence of a *trans* interaction allows the build up of genetic variation in *cis*, with loss of the *trans* interaction revealing this variation. As a result, when a *trans* interaction is present in the population, the number of alternative phenotypes accessible to the population through a single mutation is of order 2^N. In contrast, when the *trans* interaction is absent from the population, far fewer alternative patterns of regulation are accessible through a single mutation. Each target gene may suffer a mutation in *cis* to reveal a different regulatory pattern. As a result, when a *trans* interaction is absent from the population, the number of alternative phenotypes accessible to the population through a single

mutation is of order N.

We may hypothesize that in rare cases when a *trans* interaction is lost, the set of targets with a mutation at B may represent an improvement on the previous regulatory scheme. For example, this may occur when a population is subject to a change in environment resulting in new regulatory schemes becoming advantageous [81]. In such cases the presence of a *trans* interaction allows new, advantageous genotypes to evolve which would be inaccessible if the *trans* interaction were absent due to the large number of (possibly deleterious) mutational steps required to reach the new genotype. The ability to reveal new patterns of regulation amongst target genes may be particularly advantageous when changes to regulon size occur.

In a small population, only a single genotype is present. Therefore when a *trans* interaction is present in the population, although mutations may accumulate in *cis*, there will be little variation between individuals. As a result loss of a *trans* interaction will only reveal a single alternative pattern of regulation. Therefore the number of alternative phenotypes accessible to the population through a single mutation is of order N. When the *trans* interaction is absent, the number of alternative phenotypes accessible to the population through a single mutation is also of order N. Therefore presence of a *trans* interaction does not provide greater accessibility to alternative phenotypes in small populations.

3.4.2 Yeast Mating System

We now discuss our results in light of observed changes to mating type regulation in yeast. Yeast has two mating types, **a** and α. The mating type adopted by a particular cell is controlled by the MAT locus. This lies at the top of a regulatory cascade in which sets of mating-type specific genes are activated. In **a** cells, **a**-specific genes are activated, whilst in α cells, α-specific genes are activated. It is the manner in which **a**- and α- specific genes are regulated by the MAT locus that we focus on in this section.

During the evolution of *S. cerevisiae*, the regulation of **a**-specific genes, which are expressed in **a**-cells but not in α-cells, has undergone a neutral change of the type considered in our model. In the ancestral mating-type network, **a**-specific genes are in a default off state. They are then activated by a pair of transcription factors, MATa2 and Mcm1, in **a**-cells. This regulatory logic has changed during the evolution *S. cerevisiae*. In the evolved network **a**-specific genes are up-regulated in **a**-cells by Mcm1 alone, whilst in α-cells MATα2 interacts with Mcm1 to prevent the activation of **a**-specific genes by Mcm1. These changes have taken place through the evolution of

a *trans* interaction between Mcm1 and MATα2, in order to prevent activation of **a**-specific genes in α-cells, along with an increase in the strength of the Mcm1 binding site already present at **a**-specific genes such that Mcm1 alone can activate **a**-specific genes in **a**-cells [124].

This system can be compared to our model as follows: Strengthening of the Mcm1 binding site is deleterious if a *trans* interaction is absent, since this leads to up-regulation of **a**-specific genes in α-cells. Therefore a *trans* interaction buffers against increase in the binding strength of the Mcm1 binding site. Therefore the Mcm1 binding site, is the binding site for TF B in our model. The binding site for TF A in our model may be thought of as the binding site for MATα2, present at all **a**-specific genes [124]. Our model therefore remains qualitatively the same, but with the following changes: The parameter k refers to the number of **a**-specific genes with strengthened binding sites for Mcm1. The rate of mutations resulting in mutation at a binding sites for B, in this context has a specific interpretation as the rate of mutations leading to strengthening of the Mcm1 binding site. This strengthening of the Mcm1 binding site is observed to occur through an increase in the AT content at regions flanking the binding site [124].

In a large population, a *trans* interaction will become fixed when the number of **a**-specific genes is larger enough, such that the rate of mutations leading to strengthening of Mcm1 binding sites is greater than the rate of mutations leading to loss of a *trans* interaction between Mcm1 and MATα2. Similarly, in small populations, a *trans* interaction will tend to become fixed when the number of **a**-specific genes is greater than the threshold given in equation (3.15). The threshold in N above which a *trans* interaction becomes fixed differs between the large and small population cases.

Thus leads us to two possible explanations for the neutral evolution of a *trans* interaction between Mcm1 and MATα2 in the yeast sex determination network. Firstly, the neutral evolution may be driven by changes in the number of **a**-specific genes regulated by the MAT locus. Secondly, the neutral evolution may be driven by changes in population size between different yeast species, resulting from adaptation to different environments.

We can consider the time to reach an absorbing state in the case of the yeast transcription network.In this case the absorbing state is reached when a *trans* interaction is present between Mcm1 and MATα2, and all **a**-specific genes have strengthened Mcm1 binding sites. When this occurs, MATa2 is no longer required for the up-regulation of **a**-specific genes and is redundant (figure 3.3). This has occurred in *S. cerevisiae*, where MATa2 has been completely lost. We may regard this state as an absorbing state, since once MATa2 is lost, mutations resulting in the

weakening of the Mcm1 binding site will become deleterious. Equation (3.17) indicates that when the rate of mutations strengthening the Mcm1 binding site are greater than the rate of mutations weakening the Mcm1 binding site, ($\mu_{cis}^{-} > \mu_{cis}^{+}$ in the notation of our model) the probability of the population reaching the absorbing state in which MATa2 is lost, is an increasing function of N. When the rate of mutations strengthening the binding site is much less than the rate of those weakening it ($\mu_{cis}^{-} \ll \mu_{cis}^{+}$), the probability of the population reaching the absorbing state is a decreasing function of N. Finally, when the rates of strengthening and weakening of the binding site are similar, ($\mu_{cis}^{-} \approx \mu_{cis}^{+}$), the probability of reaching the absorbing state is a peaked distribution.

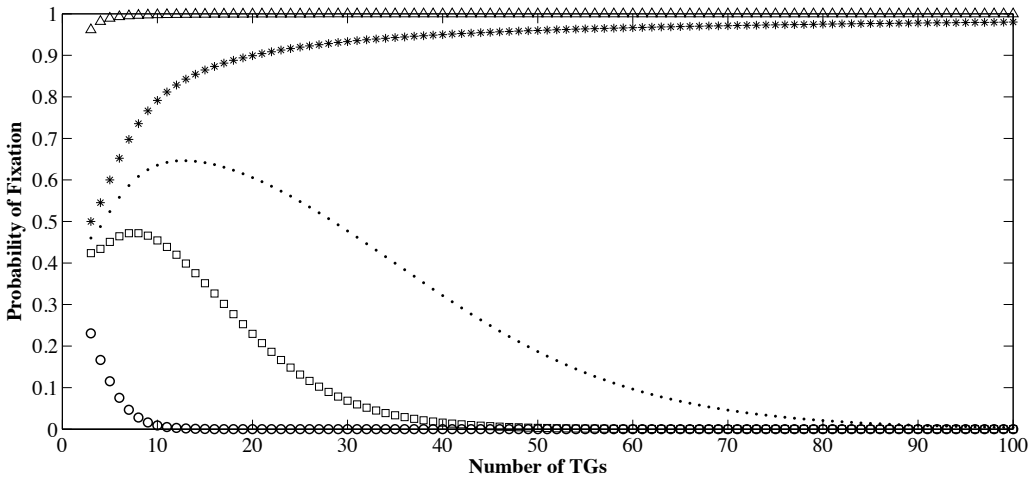

Figure 3.5: Probability of fixation of a *trans* interaction. Variation of ρ_+ wth N is shown for different values of $\frac{\mu_b^-}{\mu_b^+}$. From top to bottom the values used are $\frac{\mu_b^-}{\mu_b^+} = 2$ (triangles), $\frac{\mu_b^-}{\mu_b^+} = 1$ (asterisks), $\frac{\mu_b^-}{\mu_b^+} = 0.9$ (dots), $\frac{\mu_b^-}{\mu_b^+} = 0.8$ (squares) and $\frac{\mu_b^-}{\mu_b^+} = 0.5$ (circles).

Since strengthening and weakening of the Mcm1 binding site occurs through changes to the AT content flanking the binding site, we suggest that $\mu_{cis}^{-} \approx \mu_{cis}^{+}$ is the most likely scenario, since these rates are determined by the rates of single nucleotide substitutions, each of which either increase or decrease the binding strength of the site.

3.4.3 Changes in Regulon Size

Our results indicate that neutral gain and loss of *trans* interactions depends critically on the number of target genes co-regulated by a pair of transcription factors, as well as on population genetic details, particularly population size and mutation rate. We therefore expect any observed

neutral change of the type considered here to be either correlated with changes in the number of target genes co-regulated by pairs of transcription factors, or with changes in the lifestyle of the species being considered. In the case of the yeast mating type network, there are a number of significant life-style differences between *C. albicans*, which contains the ancestral mating type determination network and *S. cerevisiae* which contains the newly evolved network [124]. For example, *C. albicans* has no haploid phase, with mating occurring between diploids, whilst in *S. cerevisiae* mating occurs between haploids [86]. In addition, the environments inhabited by the two species are very different, as might be expected from a pair of species that have undergone the same degree of protein divergence as humans and sea squirts [33, 32, 125]. It is therefore difficult to draw quantitative conclusions about the neutral evolution of the species based on our model. However, if neutral evolution of co-regulation has occurred as described in our model, we may expect the number of target genes co-regulated by pairs of transcription factors, the regulon size, to have undergone substantial changes.

In the yeast sex mating type determination network, the number of **a**-specific genes is 13 in *S. cerevisiae*, 8 in *K. lactis* and 9 in *C. albicans*. However only 4 **a**-specific genes are conserved between all three species, suggesting substantial gain and loss of targets throughout the evolution of the *S. cerevisiae* network. This pattern, with only a small fraction of target genes conserved between all three yeast species, is repeated in the Mcm1-Yox1 and Mcm1-Fkh2 regulons, [125]. It is therefore possible conclude that significant changes to regulon size do occur during the evolution of yeast. This in turn suggests that neutral evolution of cooperative binding, as described here, may play a significant role in determining the way genes are regulated by pairs of transcription factors.

3.4.4 Co-regulation in the Yeast TFN

Our results may be compared to those of recent studies investigating the extent of co-regulation of targets by transcription factors [5, 6, 4]. These studies reveal a number of trends. In the Yeast transcription network, in many cases, TFs which regulate a large number of targets (network hubs), have a larger than expected number of co-regulatory partners [5]. This is supported by the observation that regulatory hubs tend not to be essential, and can suffer mutation or even complete loss from a transcription network without strongly deleterious effects. This lack of essentiality is suggested to be due to mutational robustness, provided by the presence of a large number of co-regulatory partners [6]. These results suggest that co-regulation in the Yeast transcription network

provides mutational robustness, which is in line with the results presented here. In addition the finding that hub transcription factors tend to have more co-regulatory partners corresponds with the intuition developed from our results, that the more targets a pair of TFs co-regulate, the more likely they are to have a *trans* interaction between them. It would therefore be interesting to reanalyse the full set of co-regulatory interactions in the Yeast transcription network in light of our results, to see how number of targets co-regulated by a pair of TFs correlates with the probability of them having a *trans* interaction between them. A further observation of these studies, states that, in both Yeast and *E. coli*, the number of target genes co-regulated by a pair of TFs follows a broad tailed distribution, best described by a power-law [5, 4]. Another possible extension of our model is to analyse the evolution of the degree-distribution of the co-regulatory network, to determine to what extent this is driven by mutational robustness.

3.4.5 Diploids

Our model can also be extended to account for diploids in the following way. In a diploid organism we now have $2N$ target genes, which must be regulated properly. At the *trans* locus we now have three possible cases: (i) when the *trans* locus is homozygous and lacks a *trans* interaction, mutations at the binding sites for B are deleterious, (ii) when the *trans* locus is homozygous and has a *trans* interaction, mutations at binding sites for B are buffered against and (iii) when the *trans* locus is heterozygous. In the third cases two scenarios are possible. Either the heterozygous *trans* interaction provides buffering against mutations at binding sites for B or it does not. In the first case we have increased the size of the neutral genotype space, g, from $N + 2$ to $4N + 3$ genotypes - $2N + 1$ genotypes associated with the heterozygous case, $2N + 1$ genotypes associated with the homozygous case in which a *trans* interaction is present , and 1 genotype associated with homozygous case in which the *trans* interaction is absent. In the second case we have increased the size of g to $2N + 3$ genotypes - 1 genotype associated with the hetwerozygous case, $2N + 1$ genotypes associated with homozygous presence of a *trans* interaction, and 1 genotype associated with homozygous lack of a *trans* interaction. However, the structure of the genotype space remains qualitatively the same as in the haploid case; a small number of genotypes are available when a *trans* interaction is absent, whilst a large number of genotypes are available when it is present. Analysis of the diploid system described above gives the same qualitative behaviour as in the haploid case a threshold in the number of target genes above which a *trans* interaction becomes fixed.

The above only holds true if no recombination occurs. As discussed previously, the effects of recombination in diploids are likely to give rise to under-dominance, such that evolution of a *trans* interaction is not favoured. A diploid model including recombination is complex to analyse but such a model may be investigated computationally, and provides a possible direction for further work. It will be interesting to compare the effects of different population genetic scenarios on the evolution of *trans* interactions of the type described here. Such an analysis would allow us to make predictions about the nature and frequency of combinatorial gene regulation in *E. coli*, Yeast and higher Eukaryotes.

3.4.6 Alternative Selection Schemes

We have presented a simple selection scheme, in which we treat the *trans* interaction between A and B as a single locus, and in which A is able to cooperatively bind B but not vice versa. Other schemes, for example, requiring a mutation at both A and B for a *trans* interaction to be present, or allowing B to cooperatively bind A, are possible. In addition, the introduction of diploidy allows for more complex selection schemes, as described above. However, alternative selection schemes give rise to the same qualitative structure of neutral genotype space, and our results are therefore qualitatively unchanged by considering different selection schemes there is still a threshold value for the number of target genes above which a *trans* interaction is fixed and below which it is absent.

3.5 Conclusion

We have presented a simple model for the neutral evolution of co-operative binding between pairs of transcription factors. Our model is based on observed neutral changes in the yeast transcription network, in which a new *trans* interaction has evolved between transcription factors which co-regulate sets of target genes. This has occurred without apparent change to the logic of the network, suggesting the evolution is neutral. We have shown that such a neutral change can occur. It can be driven by changes to the life-style of a species (particularly changes in population size). Alternatively, it can be driven by changes to the number of target genes co-regulated by a pair of transcription factors.

We have shown that the probability that a *trans* interaction between two transcription factors is present in a population, follows a threshold function in the number of target genes regulated. Above the threshold, a *trans* interaction will be present in the population, whilst below the threshold it

will be absent. When a trans interaction becomes fixed, this in turn allows genetic variation at cis amongst the regulated target genes. As a result fixation of a *trans* interaction will tend to be accompanied by significant changes to the *cis* regulatory regions of regulated target genes. This neutral co-evolution of *cis* and *trans* is precisely what is observed in the yeast mating type determination network.

Our model suggests that the neutral evolution of transcription networks may have complicated dynamics. We have shown that changes to one part of a network, the number of target genes regulated by a pair of transcription factors, can have knock on effects at both *cis* and *trans*. Far from being small, these knock on effects represent significant changes to the way a set of target genes are regulated. However, our results relate to neutral evolution in the case of a particularly simple network. The evolutionary dynamics predicted or this network, suggest that the neutral evolution of gene networks may be extremely complex when larger networks are considered. It must therefore be concluded that in order to properly understand the evolution of gene networks, a better understanding of their neutral evolution must first be gained.

3.6 Appendix B

3.6.1 Equilibrium Genotype Distribution for an Infinite Population

We now find approximate solutions to equations (3.3)-(3.5) in order to find the equilibrium frequency of *trans* interactions in the population. In order to do this, equation (3.8) was derived by assuming that the term $\mu_b^+ P_+(1)$ in equation (3.6) is sufficiently small that it can be neglected. To show that this assumption is valid, we show that the mean time taken for an organism with genotype $P_+(1)$ to reach genotype $P_+(0)$ via mutation, increases exponentially with N. As a result, the rate at which organisms mutate from genotype $P_+(k > 0)$ to genotype $P_+(0)$ declines exponentially with N, and can be neglected. To calculate the time taken reach genotype $P_+(0)$ from genotype $P_+(1)$ via mutation, we treat the set of genotypes $P_+(k)$ as a markov chain, with absorbing state $P_+(0)$. An organism will continue to mutate between different genotypes $P_+(k > 0)$, unless it suffers a deleterious mutation (the *trans* interaction is lost) or it reaches the absorbing state $P_+(0)$. If a deleterious mutation occurs, the organism is lost from the population. Therefore we calculate the time taken for an organism to reach the absorbing state, provided it does not suffer a deleterious mutation. Since all genotypes have the same fitness, this time will depend only on the rates of mutation between the different genotypes belonging to $P_+(k)$.

The mean time, $\bar{t}_i$, taken to reach absorbing state $P_+(0)$ from initial state $P_+(i)$ is given by

$$\bar{t}_i = \sum_{j=1}^{N} \bar{t}_{ij} \tag{3.18}$$

where $\bar{t}_{ij}$ is the mean time spent in state $P_+(j)$ given the initial state is $P_+(i)$. The term $\bar{t}_{ij}$ is given by

$$\bar{t}_{ij} = \frac{1}{\beta_j}\left\{1 + \frac{\alpha_{j-1}}{\beta_{j-1}} + \frac{\alpha_{j-1}\alpha_{j-2}}{\beta_{j-1}\beta_{j-2}} + \ldots\ldots + \frac{\alpha_{j-1}\alpha_{j-2}\ldots\alpha_1}{\beta_{j-1}\beta_{j-2}\ldots\beta_1}\right\}$$

for $j = 1, 2, \ldots\ldots, i$, and

$$\bar{t}_{ij} = \bar{t}_{ii}\left(\frac{\alpha_i\alpha_{i+1}\ldots\alpha_{j-1}}{\beta_{i+1}\beta_{i+2}\ldots\beta_j}\right) \tag{3.19}$$

for $j = i+1, \ldots\ldots, N$ [36].

Where α_j is the probability of moving from genotype $P_+(j)$ to genotype $P_+(j+1)$, which for this model is given by $\alpha_j = (N-j)\mu_b^-$. Similarly, β_j is the probability of moving from genotype $P_+(j)$ to genotype $P_+(j-1)$, which for this model is given by $\beta_j = j\mu_b^+$.

We wish to calculate the time $\bar{t}_1$ to reach $P_+(0)$ from $P_+(1)$. Using equation (3.19) with (3.18) this gives

$$\begin{aligned}
\bar{t}_1 &= \frac{1}{\mu_b^+}\sum_{j=1}^{N}\frac{(N-1)!}{j!(N-j)!}\left(\frac{\mu_b^-}{\mu_b^+}\right)^{j-1} \\
&= \frac{1}{N\mu_b^-}\sum_{j=1}^{N}\binom{N}{j}\left(\frac{\mu_b^-}{\mu_b^+}\right)^{j} \\
&= \frac{\left(1+\frac{\mu_b^-}{\mu_b^+}\right)^N - 1}{N\mu_b^-}
\end{aligned} \tag{3.20}$$

Therefore the mean time taken for an organism to reach genotype $P_+(0)$ from genotype $P_+(1)$ increases approximately exponentially with N. Organisms with a *trans* interaction and at least one mutation in *cis*, suffer deleterious mutations (through loss of the *trans* interaction) at rate μ^-_{trans}. Therefore the mean time taken for such an organism to suffer a deleterious mutation is $\frac{1}{\mu^-_{trans}}$. If the mean time taken to suffer a deleterious mutation is greater than the mean time take for an organsim with genotype $P_+(1)$ to reach genotype $P_+(0)$, $\bar{t}_1 > \frac{1}{\mu^-_{trans}}$, then most organism with genotype $P_+(1)$ will be lost from the population due to deleterious mutations. Therefore when N satisfies

$$\frac{\left(1+\frac{\mu^-_b}{\mu^+_b}\right)^N - 1}{N} > \frac{\mu^-_b}{\mu^-_{trans}} \tag{3.21}$$

The value $N_{thresh} = \frac{\mu^-_{trans}}{\mu^-_b}$ is the value of N above which organisms without a *trans* interaction suffer deleterious mutations more frequently than organisms with a *trans* interaction. It is plausible to assume that $\mu^-_b \leq \mu^+_b$. Therefore taking the upper limit, $\mu^-_b = \mu^+_b$, we can use equation (3.21) to write

$$N_{thresh} > \frac{N}{2^N - 1} \tag{3.22}$$

For values of $N_{thresh} > 1$, equation (3.22) is always satisfied for all $N \geq 1$. In addition $\bar{t}_1$ increases exponentially with N. As such, we are justified in assuming $\bar{t}_1 \gg \frac{1}{\mu^-_{trans}}$, even for relatively small values of N. For example, taking $N_{thresh} = 10$ and $\mu^-_b = \mu^+_b$, it takes on average 10 times longer for an organism with genotype $P_+(1)$ to reach $P_+(0)$ than it does for a deleterious mutation to occur when $N = 1$ and 10^3 times longer when $N = 10$. For $\mu^-_b = 10\mu^+_b$, it takes 10^2 times longer when $N = 1$ and 10^{10} times longer when $N = 10$. Therefore we can assume than organisms with a genotype lying on $P_+(k > 0)$ will tend to suffer a deleterious mutation before they return to $P_+(0)$.

This, along with the numerical results present in figure (3.3), justify neglecting the term $\mu^+_b P_+(1)$ in equation (3.6). As a result, equations (3.3), (3.7) and (3.8) form a straightforward set of simultaneous equations which can be solved explicitly to give equation (3.9), for the approximate equilibrium frequency of a *trans* interaction in the population,

3.6.2 Equilibrium Genotype Distribution for a Small Population

Equations (3.11)-(3.13) for the probability that a small population has a particular genotype on g can be solved explicitly, without recourse to approximation. At equilibrium equation (3.11) has solution

$$\pi_+(0) = \frac{\mu^+_{trans}}{\mu^-_{trans}}\pi_- \tag{3.23}$$

substituting this into equation (3.12), we have at equilibrium

$$\pi_+(1) = \frac{N\mu^-_b}{\mu^+_b}\frac{\mu^+_{trans}}{\mu^-_{trans}}\pi_- \tag{3.24}$$

At equilibrium, equation (3.13) has solution

$$(k+1)\mu^+_b\pi_+(k+1) = (N-k)\mu^-_b\pi_-(k) \tag{3.25}$$

Substituting equation (3.24) into equation (3.25) results in the solution

$$\pi_+(k) = \binom{N}{k}\left(\frac{\mu^-_b}{\mu^+_b}\right)^k\frac{\mu^+_{trans}}{\mu^-_{trans}}\pi_- \tag{3.26}$$

Therefore the probability that an organism with a *trans* interaction has k mutations follows a binomial distribution with mean $\frac{\mu^-_b}{\mu^-_b+\mu^+_b}$. Since we must have $\pi_- + \sum_{k=0}^N \pi_+(k) = 1$, we have

$$\pi_- + \sum_{k=0}^{N}\binom{N}{k}\left(\frac{\mu^-_b}{\mu^+_b}\right)^k\frac{\mu^+_{trans}}{\mu^-_{trans}}\pi_- = 1 \tag{3.27}$$

which gives

$$\pi_- = \frac{1}{1+\left(1+\frac{\mu^-_b}{\mu^+_b}\right)^N\frac{\mu^+_{trans}}{\mu^-_{trans}}} \tag{3.28}$$

and taking $\pi_+ = 1-\pi_-$ gives equation (3.14) for the probability that a *trans* interaction is present

in the population.

3.6.3 Probability of Reaching an Absorbing State

Equation (3.17) for the probability, ρ_+ that an organism (or population) with genotype $\pi_+(0)$ reaches the absorbing state $\pi_+(N)$ as opposed to the absorbing state π_-. This probability ρ_i of reaching $\pi_+(N)$ given that the starting state is $\pi_+(i)$, is given by

$$\rho_i = \frac{1 + \sum_{k=1}^{i-1} \prod_{j=1}^{k} \frac{\beta_j}{\alpha j}}{1 + \sum_{k=1}^{N} \prod_{j=1}^{k} \frac{\beta_j}{\alpha j}} \tag{3.29}$$

The term ρ_+ is given by ρ_1 in equation (3.29). Substituting values $\beta_1 = \mu_{trans}^-$, $\beta_k = (k-1)\mu_b^+$ (for $k > 1$), and $\alpha_k = (N - k + 1)\mu_b^-$ into equation (3.29) gives equation (3.17).

To see how ρ_+ varies with N, we look at the change that results when N is increased by 1 - $\Delta\rho_+ = \rho_+(N+1) - \rho_+(N)$. In order to do this, we define

$$f(N) = \sum_{k=1}^{N-1} \frac{k!(N-k-1)!}{N!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^k \tag{3.30}$$

If $f(N)$ is increasing with increasing N, $\Delta\rho_+ < 0$, and if $f(N)$ is decreasing with increasing N, $\Delta\rho_+ > 0$. We consider how $f(N)$ changes with increasing N as follows

$$\begin{aligned} f(N+1) - f(N) &= \sum_{k=1}^{N} \frac{k!(N-k)!}{(N+1)!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^k - \sum_{k=1}^{N-1} \frac{k!(N-k-1)!}{N!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^k \\ &= \frac{1}{N+1} \left(\frac{\mu_b^+}{\mu_b^-}\right)^N + \sum_{k=1}^{N-1} \left(\frac{N-k}{N+1} - 1\right) \frac{k!(N-k-1)!}{N!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^k \\ &= \frac{1}{N+1} \left(\left(\frac{\mu_b^+}{\mu_b^-}\right)^N - \sum_{k=1}^{N-1} \frac{(k+1)!(N-k-1)!}{N!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^k \right) \end{aligned} \tag{3.31}$$

If $\mu_b^- \geq \mu_b^+$, $f(N+1) - f(N) < 0$ for all N. To see this write

$$\left(\frac{\mu_b^+}{\mu_b^-}\right)^N - \sum_{k=1}^{N-1} \frac{(k+1)!(N-k-1)!}{N!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^k$$
$$= \left(\frac{\mu_b^+}{\mu_b^-}\right)^N - \left(\frac{\mu_b^+}{\mu_b^-}\right)^{N-1} - \sum_{k=1}^{N-2} \frac{(k+1)!(N-k-1)!}{N!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^k \tag{3.32}$$

Since for this case $\left(\frac{\mu_b^+}{\mu_b^-}\right)^N \leq \left(\frac{\mu_b^+}{\mu_b^-}\right)^{N-1}$ for all N, $f(N+1) - f(N) < 0$ for all N and ρ_+ is a monotonically increasing function of N.

For the case $\mu_b^- < \mu_b^+$, $f(N+1) - f(N)$ may be increasing or decreasing, depending on N. Therefore we explore the variation of ρ_+ with N for this case numerically (figure 3.5). To see that $f(N+1) - f(N) > 0$ in the limit $N \to \infty$, write

$$\left(\frac{\mu_b^+}{\mu_b^-}\right)^N - \left(\frac{\mu_b^+}{\mu_b^-}\right)^{N-1} - \sum_{k=1}^{N-2} \frac{(k+1)!(N-k-1)!}{N!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^k$$
$$= \left(\frac{\mu_b^+}{\mu_b^-}\right)^N \left(1 - \left(\frac{\mu_b^+}{\mu_b^-}\right)^{-1} - \sum_{k=1}^{N-2} \frac{(k+1)!(N-k-1)!}{N!} \left(\frac{\mu_b^+}{\mu_b^-}\right)^{k-N}\right) \tag{3.33}$$

Since all terms in the sum are of order N^{-1} or higher, in the limit $N \to \infty$, the sum tends to zero and we are left with

$$f(N+1) - f(N) = \left(\frac{\mu_b^+}{\mu_b^-}\right)^N \left(1 - \left(\frac{\mu_b^+}{\mu_b^-}\right)^{-1}\right) \tag{3.34}$$

and since $\mu_b^- < \mu_b^+$, this is positive, meaning that ρ_+ is a decreasing function of N as $N \to \infty$.

Chapter 4

Evolution of Autoregulatory Motifs in Diploid Organisms

Transcription networks are subject to many forms of noise. They result both from changes in the environment external to a cell, as well as from intrinsic noise in gene expression resulting from the stochastic nature of transcription. In *Escherichia coli*, one of the most common strategies to reduce intrinsic noise is through negative autoregulation. Negative autoregulation has been shown, both theoretically and experimentally, to reduce the variance in gene expression compared to genes which do not autoregulate. It has also been shown to speed the response times of genes to perturbations. This is reflected in the fact that negative autoregulation is found in 37% of known transcription factors in *E. coli*. We investigated the behaviour of negatively autoregulating genes in diploids. This is of interest, since in diploids a pair of negatively autoregulating alleles form a network of four interactions and three feedback loops. This is in contrast to haploids, where negative autoregulation consists of a single interaction and a single feedback loop. We considered heterozygous diploids, in which the strength of negative autoregulation differs between two alleles. We showed that in such cases the contributions of the two alleles to total gene expression differs considerably. In particular, when negative autoregulation is strong, we showed that one allele will be almost completely unexpressed, with total gene expression accounted for by the other allele. We also showed that the noise in the total expression of heterozygotes is often greater than the noise in homozygous case. As a result, if noise reduction is selected for, this results in a barrier to the evolution of negative autoregulation. This is reflected in the frequency of negative autoregulation in the *Saccharomyces cerevisiae* transcription network, where between 2% and 4% of

transcription factors negatively autoregulate. We also applied our results to duplicates of negatively autoregulating genes in haploids. Our results suggest that negative autoregulation can mitigate the effects of increased dosage from duplicate genes. However, once one of the duplicates suffers *cis* mutations, it will tend to become under-expressed and behave as a pseudogene. This may explain why duplicates of negatively autoregulating genes are no more abundant than duplicates of other genes in the *E. coli* transcription network.

4.1 Background

Transcription factor networks (TFNs) consist of sets of genes and regulatory interactions between those genes. The pattern of interconnections that make up a TFN encodes information about how various sets of genes are co-expressed. As such, the regulatory interactions that make up a TFN are under direct selection to optimise mean gene expression. However, regulatory interactions, which allow sets of genes to be coexpressed, also allow perturbations which affect the expression of one gene to affect the expression of other genes in the network. As a result, TFNs are also under selection to minimize the effects of such perturbations - i.e TFNs are under selection for robustness [20, 24, 31, 49, 65, 71, 82, 101, 130, 133]. Determining how robust a network is requires an understanding of the expression dynamics of individual genes, how the expression levels of different sets of genes covary, and the nature of the perturbations that the network is subject to. The perturbations which affect a TFN may result from changes in the external environment of a cell. Alternatively, they may result from the stochastic nature of transcription, which gives rise to intrinsic variation in the expression of individual genes. Finally, they may result from mutations which change the strength of regulatory interactions in the network [82, 101].

Perturbations which result from changes in the environment external to a cell may have a global effect, such that they directly effect the expression of all genes in a similar manner, for example, changes in temperature. In other cases, they have a local effect, and directly effect only a subset of genes, for example, changes in the level of nutrients in the external environment. In general the response of the organism to such environmental changes is to change its pattern of gene expression, down-regulating some genes, up-regulating some genes, and keeping the expression of some genes constant. For example, the transition between the haploid and diploid phase, in response to changes in the availability of nutrients in the environment, seen in the yeast *Saccharomyces cerevisiae*, requires down-regulation of haploid specific genes and up-regulation of diploid specific genes [124]. In contrast, perturbations resulting from noise in gene expression frequently requires the evolution

of noise filtering mechanisms [1, 11, 51, 79, 78, 106, 110, 118]. These ensure that noise in the expression of upstream genes is not passed on and amplified in the expression of the downstream genes they regulate. Finally, changes to regulatory interactions resulting from mutations are often deleterious, resulting in sub-optimal patterns of gene expression. As a result, networks are required to evolve a degree of mutational robustness, such that the deleterious effects of such mutations are minimized [31, 127, 133].

One approach to determining how a set of genes will respond to different types of perturbation, is to look for patterns of regulatory interactions which perform particular functions. For example, feedback loops may give rise to switch-like behaviour, to oscillation in gene expression, or to a reduction in the level of noise in gene expression [14, 51, 106, 118, 128]. Feed forward loops may act as noise filters, maintaining constant gene expression in downstream genes when faced with transient environmental changes, but allowing changes in the expression of downstream genes when environmental changes are sustained for long periods [79, 78]. In order to determine that patterns of interactions which evolve to perform particular functions, network motifs have been defined [88, 87]. Network motifs are patterns of regulatory interactions which occur at high frequency in real TFNs, compared to the frequency that would be expected due to chance. The simplest network motif that has been found in this way is negative autoregulation. This motif consists of a single gene which suppresses its own transcription. Negative autoregulation occurs in 42 out of 115 transcription factors (37%) in *Escherichia coli* [110]. The expected number of negatively autoregulating genes that would occur due to chance in this network is 1.2 ± 1.1 [1]. The observed frequency of negative autoregulators is 37 standard deviations from the mean number that would be expected due to chance, and as such is highly significant [1].

The expression dynamics of negatively autoregulating genes has been extensively investigated. It has been shown that negative autoregulation results in a decrease in the noise of gene expression, compared to genes with the same mean expression level which do not negatively autoregulate [1, 106, 110, 118]. Similarly, the response time - the time for the expression level of a gene to return to the mean under perturbation - for a negatively auto-regulating gene is less than the response time in genes which do not negatively autoregulate [1, 106]. As such it has been suggested that negative autoregulation has evolved as a mechanism for noise reduction. However, *E. coli* is a haploid organism. The dynamics of negatively autoregulating genes in diploid organisms has not been investigated. Yet the behaviour of negatively autoregulating genes in diploids is likely to be different from that seen in haploids. Whereas in haploids negative autoregulation consists

of a single gene and a single regulatory interaction, in diploids it consists of a pair of genes and four regulatory interactions (figure 4.1). In this chapter I consider the dynamics of expression of negatively autoregulating genes in diploids. In particular I consider whether negative autoregulation can evolve as a mechanism for noise reduction in diploid organisms. The results obtained are interpreted in the light of data on the frequency of negative autoregulation in *S. cerevisiae*. The differences observed between the frequency of negative autoregulation in *S. cerevisiae* and *E. coli* are explained as due to the requirement for *S. cerevisiae* to function as a diploid.

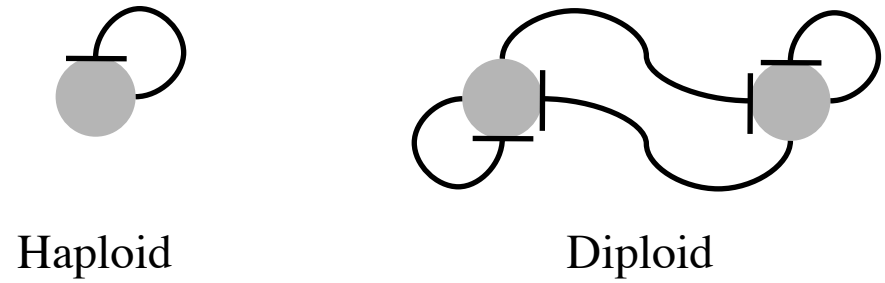

Figure 4.1: In haploids negative autoregulation consists of a single interaction and a single feedback loop. In diploids (or following duplication of an autoregulating gene in haploids) negative autoregulation consists of four interactions and three feedback loops.

4.2 Model

We develop two models for negative autoregulation. This is done by extending previous models of negative autoregulation in haploids to the diploid case, and exploring their behaviour. The first model is a simple system of ODEs, which allows us to explore the expression level of genes which negatively autoregulate. This is used to investigate the way gene expression changes under changes in the strength of negative autoregulation. The second model captures the noise in gene expression which results from the stochastic nature of transcription, by modelling transcription as a Markov process. This is used to investigate the way the noise in gene expression changes under changes in the strength of negative autoregulation.

4.2.1 ODE model of Haploid Autoregulation

In order to study the expression level of negatively autoregualting genes, we employ a simple ODE model which relates the rate of change of gene expression p, to the rate of production $f(p)$ and the rate of protein degradation γ_p [1]:

$$\frac{dp}{dt} = f(p) - \gamma_p p \tag{4.1}$$

where $f(p)$ is taken to be a Hill function [1]

$$f(p) = \frac{\beta_p}{1 + \left(\frac{p}{K}\right)^n} \tag{4.2}$$

β_p is the maximum rate of protein production which would occur in the absence of autoregulation. K determines the threshold of the Hill function, and is referred to as the repression coefficient for negative autoregulation [1]. K is determined by the binding strength of the protein for its binding sites in the promoter region of the regulated gene [18, 40]. The Hill coefficient n determines the steepness of the repression function. The equilibrium gene expression $\bar{p}$, occurs when the left hand side of equation (4.1) is zero, giving

$$\bar{p} = \frac{\beta_p}{\gamma_p} \frac{1}{1 + \left(\frac{\bar{p}}{K}\right)^n} \tag{4.3}$$

In general equation (4.3) cannot be solved explicitly. However, we can find the equilibrium expression level if we approximate the Hill function, $f(p)$, by a threshold function such, $\theta(p)$, such that $\theta(p) = 0$ for $p \geq K$ and $\theta(p) = 1$ otherwise [1]. This is exactly valid in the limit $n \to \infty$, and provides an increasingly accurate approximation as n increases (figure 4.2).

When $n \to \infty$ the equilibrium gene expression $\bar{p}$ is given by $\bar{p} = K$ - it is determined solely by the repression coefficient. Therefore, in haploid organisms, the expression of negatively autoregulating genes is strongly determined by the repression coefficient K.

4.2.2 ODE Model of Diploid Autoregulation

In order to extend the model described in equation (4.1)-(4.3) to diploids we also make a number of further assumptions. We refer to a pair of alleles, labelled 1 and 2. We assume that the proteins produced by the alleles are identical, such that both alleles have the same degradation rate γ_p and the same maximum production rate β_p. We refer to the expression level of alleles 1 and 2 as p_1 and p_2 respectively. Similarly we refer to the repression coefficients of alleles 1 and 2 as K_1 and K_2 and to the Hill coefficients as n_1 and n_2 respectively. The rate of change of protein produced

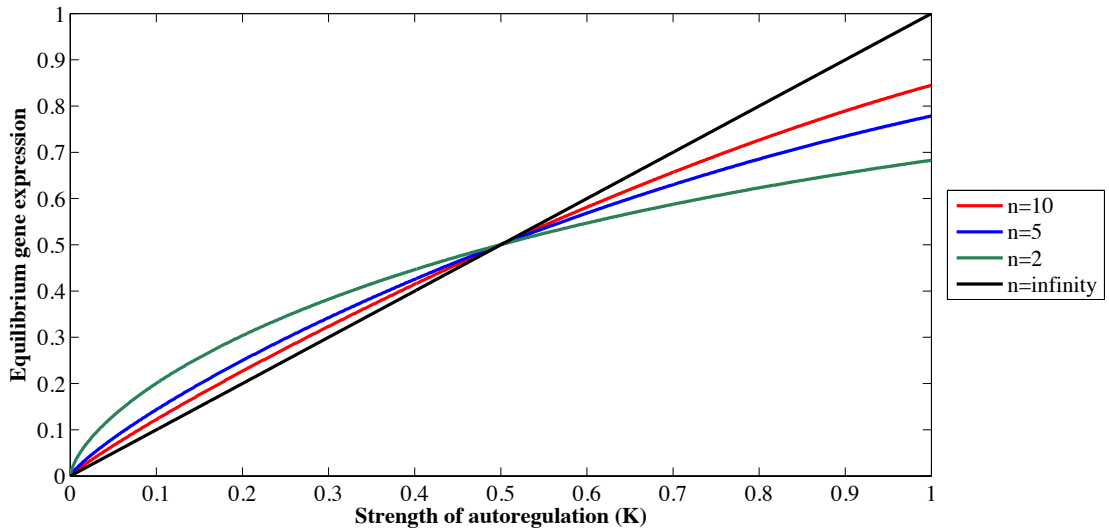

Figure 4.2: Equilibrium gene expression in a haploid. As K increases (binding strength decreases), equilibrium gene expression increases. For $n \to \infty$ the relationship is linear. The relationship deviates increasingly from linearity as the Hill coefficient n decreases

by each allele is then described by

$$\begin{aligned} \frac{dp_1}{dt} &= f_1(p_1 + p_2) - \gamma_p p_1 \\ \frac{dp_2}{dt} &= f_2(p_1 + p_2) - \gamma_p p_2 \end{aligned} \tag{4.4}$$

where f_1 and f_2 refer to the Hill functions with parameters K_1 and n_1, K_2 and n_2, respectively. The total expression of the gene is given by $p = p_1 + p_2$. By summing the pair of equations (4.4) we are given

$$\frac{dp}{dt} = f_1(p) + f_2(p) - \gamma_p p \tag{4.5}$$

which results in equilibrium gene expression

$$\bar{p} = \frac{\beta_p}{\gamma_p}\left(\frac{1}{1 + \left(\frac{\bar{p}}{K_1}\right)^{n_1}} + \frac{1}{1 + \left(\frac{\bar{p}}{K_2}\right)^{n_2}}\right) \tag{4.6}$$

In homozygotes, we assume that $K_1 = K_2 = K$ and $n_1 = n_2 = n$. As a result, equations (4.5) and (4.6) become

$$\frac{dp}{dt} = 2f(p) - \gamma_p p \tag{4.7}$$

which results in equilibrium gene expression

$$\bar{p} = \frac{2\beta_p}{\gamma_p} \frac{1}{1 + \left(\frac{\bar{p}}{K}\right)^n} \tag{4.8}$$

and the diploid case is identical to the haploid case, with a maximum protein production rate of $2\beta_p$. However, in the heterozygote case, where $f_1(p) \neq f_2(p)$, equations (4.5) and (4.6) may give rise to different expression dynamics and equilibrium expression levels from those which occur in haploids.

4.2.3 Stochastic Model of Haploid Autoregulation

We now employ a stochastic model in order to calculate the noise in the expression of a negatively autoregulating gene. The model is described by the number of mRNA molecules r, and the number of proteins p present in a cell. The probability $n_{r,p}(t)$ that the system has r mRNA molecules and p proteins at time t evolves according to Scheme 1:

$$\begin{aligned}
n_{r,p} &\xrightarrow{\beta_r} n_{r+1,p} \\
n_{r,p} &\xrightarrow{r\beta_P} n_{r,p+1} \\
n_{r,p} &\xrightarrow{r\gamma_r} n_{r-1,p} \\
n_{r,p} &\xrightarrow{p\gamma_P} n_{r,p-1}
\end{aligned}$$

Scheme 1

where β_r is the rate at which mRNA molecules are transcribed from DNA, γ_r is the rate of mRNA

degradation, β_p is the rate at which mRNA is translated into protein and γ_p is the rate of protein degradation. In order to include negative autoregulation in this model we assume that the rate or transcription of mRNA, β_r, is a function of the number of proteins p present in the cell, such that

$$\beta_r(p) = \frac{\beta_r^0}{1 + \left(\frac{p}{K}\right)^n} \tag{4.9}$$

where β_r^0 is the maximum rate of mRNA transcription [118], and the parameters K and n of the Hill function are as described in the previous section.

At equilibrium the repression function of equation (4.9) is well approximated by its linearization about the mean value of p [118]. Therefore we can write

$$\beta_r(p) \cong \beta_r(\langle p \rangle) + \frac{n}{\beta_r^0}\left[\beta_r^0 - \beta_r(\langle p \rangle)\right]\beta_r(\langle p \rangle) - \frac{n}{\beta_r^0}\left[\beta_r^0 - \beta_r(\langle p \rangle)\right]\beta_r(\langle p \rangle)\frac{p}{\langle p \rangle} \tag{4.10}$$

for simplicity we write

$$\begin{aligned} \beta_r^+ &= \beta_r(\langle p \rangle) + \frac{n}{\beta_r^0}\left[\beta_r^0 - \beta_r(\langle p \rangle)\right]\beta_r(\langle p \rangle) \\ \beta_r^- &= \frac{n}{\langle p \rangle\, \beta_r^0}\left[\beta_r^0 - \beta_r(\langle p \rangle)\right]\beta_r(\langle p \rangle) \end{aligned} \tag{4.11}$$

where β_r^+ describes the rate of transcription of mRNA at equilibrium, and β_r^- gives the strength of repression as a result of negative autoregulation [118]. This model can be solved explicitly (Appendix C) to give the mean protein number $\langle p \rangle$ and the noise in protein number, expressed as the ratio of the variance to the mean $\frac{\delta p^2}{\langle p \rangle}$ at equilibrium [118]

$$\langle p \rangle = \frac{b\beta_r^+}{\gamma_p + b\beta_r^-} \tag{4.12}$$

where $b = \frac{\beta_p}{\gamma_r}$ is the mean number of proteins produced per transcript. The noise in protein number is given by [118]

$$\frac{\delta p^2}{\langle p \rangle} = \left(\frac{\gamma_p - \beta_r^-}{\gamma_p + b\beta_r^-}\right)\left(\frac{b}{1+\eta}\right) + 1 \tag{4.13}$$

where $\eta = \frac{\gamma_p}{\gamma_r}$ is the ratio of mRNA to protein lifetimes. By replacing equations (4.9) and (4.11) into equation (4.12) we find

$$\langle p \rangle = \frac{\beta_r^0 b}{\gamma_p} \frac{1}{1 + \left(\frac{\langle p \rangle}{K}\right)^n} \tag{4.14}$$

(Appendix C) which is of the same form as equation (4.3) for the equilibrium protein concentration in the haploid ODE model. For large n we then have $\langle p \rangle \approx K$, and $\beta_r^- = \frac{\beta_r^0 n}{4K}$. Therefore the noise in gene expression, given by equation (4.13) can be decreased by increasing n or decreasing K.

4.2.4 Stochastic Model of Diploid Autoregulation

The stochastic model for negative autoregulation in haploids presented above can be extended to diploids. We assume once again that there are two alleles, 1 and 2, which are identical in all respects except for the strength of negative autoregulation. Thus the rate of transcription of mRNA from proteins is described by the function $\beta_r(p) = \beta_r^1(p) + \beta_r^2(p)$, such that

$$\begin{aligned} \beta_r^1(p) &= \frac{\beta_r^0}{1 + \left(\frac{p}{K_1}\right)^{n_1}} \\ \beta_r^2(p) &= \frac{\beta_r^0}{1 + \left(\frac{p}{K_2}\right)^{n_2}} \end{aligned} \tag{4.15}$$

where K_1 and n_1 refer to allele 1 and K_2 and n_2 to allele 2. At equilibrium this can be approximated by the linearization

$$\begin{aligned} \beta_r(p) \cong \beta_r(\langle p \rangle) + \frac{n_1}{\beta_r^0}\left[\beta_r^0 - \beta_r^1(\langle p \rangle)\right]\beta_r^1(\langle p \rangle) + \frac{n_2}{\beta_r^0}\left[\beta_r^0 - \beta_r^2(\langle p \rangle)\right]\beta_r^2(\langle p \rangle) \\ - \left(\frac{n_1}{\beta_r^0}\left[\beta_r^0 - \beta_r^1(\langle p \rangle)\right]\beta_r^1(\langle p \rangle) + \frac{n_2}{\beta_r^0}\left[\beta_r^0 - \beta_r^2(\langle p \rangle)\right]\beta_r^2(\langle p \rangle)\right)\frac{p}{\langle p \rangle} \end{aligned} \tag{4.16}$$

which gives

$$\begin{aligned}\beta_r^+ &= \beta_r(\langle p\rangle) + \frac{n_1}{\beta_r^0}\left[\beta_r^0 - \beta_r^1(\langle p\rangle)\right]\beta_r^1(\langle p\rangle) + \frac{n_2}{\beta_r^0}\left[\beta_r^0 - \beta_r^2(\langle p\rangle)\right]\beta_r^2(\langle p\rangle) \\ \beta_r^- &= \frac{n_1}{\beta_r^0\langle p\rangle}\left[\beta_r^0 - \beta_r^1(\langle p\rangle)\right]\beta_r^1(\langle p\rangle) + \frac{n_2}{\beta_r^0\langle p\rangle}\left[\beta_r^0 - \beta_r^2(\langle p\rangle)\right]\beta_r^2(\langle p\rangle \end{aligned} \tag{4.17}$$

replacing equations (4.15) and (4.17) into equation (4.12) gives

$$\langle p\rangle = \frac{\beta_r^0 b}{\gamma_p}\left(\frac{1}{1+\left(\frac{\langle p\rangle}{K_1}\right)^{n_1}} + \frac{1}{1+\left(\frac{\langle p\rangle}{K_2}\right)^{n_2}}\right) \tag{4.18}$$

which is of the same form as equation (4.6) for the equilibrium protein concentration in the diploid ODE model. The noise in the gene expression given be equation (4.13) is a function of β_r^- in equation 4.17. When $\beta_r^1(p) = \beta_r^2(p)$ (the homozygous case), this is of the same form as β_r^- in the haploid case (equation (4.11)) and the noise in gene expression decrases with increasing n and decreasing K. However in the heterozygous case, when $\beta_r^1(p) \neq \beta_r^2(p)$, the variation of the noise in gene expression with the parameters K_1, K_2, n_1 and n_2 may be more complex.

4.2.5 Mutation

We now consider how heterozygotes of the type described above may arise. We have assumed that the parameters of the Hill function K_1, K_2, n_1 and n_2 differ between the two alleles. Negative autoregulation of the type considered here requires the presence of regulatory binding sites in the promotor of each gene for its own protein product. A mutation at these binding sites will change the strength with which they bind their own protein product, whilst leaving the binding strength at the other allele unaffected. In contrast, a mutation which affects the protein produced by one of the alleles will affect the probability of it binding to both alleles equally (see Appendix C). Thus we focus on mutations in *cis* that affect the strength of regulatory binding sites at one of the alleles only.

The effects of point mutations on the strength of binding sites has been investigated in a number of studies [18, 40]. For a single TF binding site, the probability that a TF is bound to the promoter is described by

$$q(p) = \frac{p}{p + \exp[\epsilon r - \epsilon_0]} \tag{4.19}$$

where each mismatched nucleotide contributes an amount $k_B T\epsilon$ to the binding energy of TFs for the binding site, whilst each correctly matched nucleotide contributes 0 to the binding energy [40]. The parameter r is the number of mismatched sites between the real and optimal binding sites and $k_B T\epsilon_0$ is the binding energy of a TF to an arbitrary sequence of DNA (i.e a non-specific binding site).

Equation (4.19) describes the probability of a TF being bound to a single specific binding site. This is a Hill function with Hill coefficient $n = 1$. The models presented above assume a binding probability which follows a Hill function with Hill coefficient which may in general be greater than 1. In order to produce such a binding probability it is necessary to consider a case in which multiple binding sites must be bound in order for transcription to be activated or deactivated [18]. The probability that n binding sites are simultaneously bound by TFs is approximately

$$q(p) \approx \frac{p^n}{p^n + K^n} \tag{4.20}$$

where K is the rate of dissociation of a TF from one of the binding sites (i.e the strength of the binding sites themselves). Equation (4.20) is only valid if TFs bind with infinite cooperativity [18]. However, in the general case of an arbitrary number of binding sites and an intermediate degree of cooperativity, the probability of TF binding follows a curve in which the steepness is determined by the number of binding sites which must be occupied, n, and the threshold value of p for which $q(p) = 0.5$ is determined by the strength of the binding sites, K [18]. Therefore we focus on a binding probability which is described by a Hill function, as in equation (4.20), in order to gain a qualitative understanding of the behaviour of negatively autoregulating genes in diploids.

4.3 Results

We now investigate the evolution of negative autoregulators in diploids. We focus on the way mutations to autoregulatory binding site affect the expression level and noise in autoregulating genes. As described in the previous section, mutations to autoregulatory binding sites alter K, which determines the threshold value of p at which the probability of TF binding, $q(p)$, is $q(p) = 0.5$.

In contrast, the steepness of the function $q(p)$ is determined by the number of binding sites which must be bound in order for negative autoregulation to suppress transcription [18]. Therefore we assume that mutations at *cis* alter K but not n. This means that in the cases considered here, different alleles may have $K_1 \neq K_2$, but always have $n_1 = n_2$. Initially we present results in which $n \to \infty$, such that $q(p)$ is a step function with threshold K. This biologically unrealistic assumption is then relaxed and it is shown that the results obtained by employing this assumption hold for a wide range of cases.

4.3.1 Evolution of Gene Expression

We first use the diploid ODE model presented above to characterise the evolution of the mean gene expression level in negatively autoregulating genes. Assuming $n \to \infty$, the mean equilibrium gene expression is given by

$$\bar{p} = \frac{\beta_p}{\gamma_p} \left[\theta(\bar{p} < K_1) + \theta(\bar{p} < K_2)\right] \tag{4.21}$$

where $\theta(p < K)$ is a step function which has value 1 if $p < K$ and value 0 otherwise. We can also use equation (4.4) to determine the equilibrium expression levels $\bar{p}_1$ and $\bar{p}_2$ of alleles 1 and 2

$$\begin{aligned} \bar{p}_1 &= \frac{\beta_p}{\gamma_p}\theta(\bar{p}_1 + \bar{p}_2 < K_1) \\ \bar{p}_2 &= \frac{\beta_p}{\gamma_p}\theta(\bar{p}_1 + \bar{p}_2 < K_2) \end{aligned} \tag{4.22}$$

In solving equations (4.21) and (4.22) we assume that in all cases $K_2 \leq K_1$, such that allele 2 has either the same or stronger negative autoregulation than allele 1. Equation (4.21), for the total expression level of both alleles has three possible solution forms, depending on the values of K_1 and K_2. These are given below. We do not consider the cases in which $K > 2\frac{\beta_p}{\gamma_p}$, since in this case the genes behave as though there is no negative autoregulation, and genes are expressed at their maximum level.

Solution for $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 \geq \frac{\beta_p}{\gamma_p}$

The solution to equation (4.21) lies at the intersection of the functions $y(p) = p$ and $z(p) = \frac{\beta_p}{\gamma_p}(\theta(\bar{p} < K_1) + \theta(\bar{p} < K_2))$. For the case $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 > \frac{\beta_p}{\gamma_p}$ these two functions are plotted in figure 4.3. From this it is clear that when $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 \geq \frac{\beta_p}{\gamma_p}$, the line $y(p)$ will always intersect the line $z(p)$ when $p = K_2$. Therefore if K_1 and K_2 differ, and the organism is heterozygous, the total expression of both alleles together will be the same as for an organism which is homozygous, with both alleles having binding sites of strength K_2. In this sense, allele 2 shows complete dominance over allele 1 for gene expression.

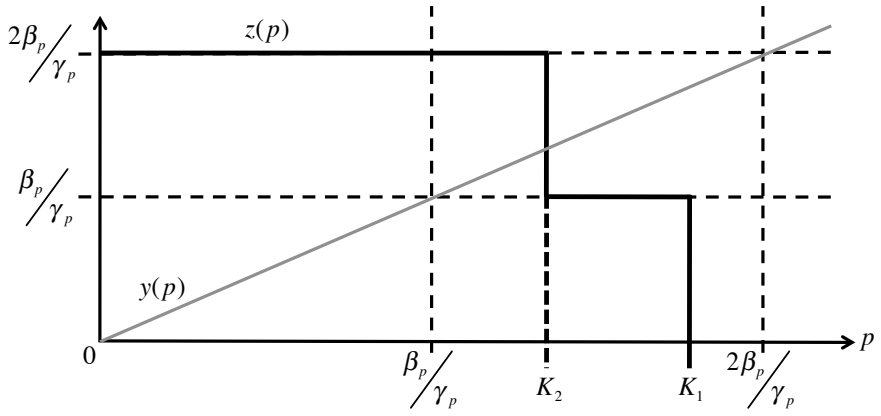

Figure 4.3: Solution for $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 \geq \frac{\beta_p}{\gamma_p}$. When the function $y(p) = p$ intersects the function $z(p)$. From this it is clear that the intersection always occurs when $p = K_2$

In order to calculate the expression levels of each allele, we take $\bar{p}_1 + \bar{p}_2 = K_2$ in equation 4.22. Since $K_2 < K_1$, this immediately gives $\bar{p}_1 = \frac{\beta_p}{\gamma_p}$. That is, allele 1 is expressed at its maximum level - the level at which it would be expressed in the absence of any negative autoregulation. The expression level of allele 2 is then $\bar{p}_2 = K_2 - \frac{\beta_p}{\gamma_p}$. In the homozygous case, in which $K_1 = K_2$, both alleles are expressed at the same level $\bar{p}_1 = \bar{p}_2 = \frac{K_2}{2}$. Since $2\frac{\beta_p}{\gamma_p} > K_2$, this means that allele 2 is under-expressed in the heterozygous case compared to the homozygous case.

Therefore in the case $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 \geq \frac{\beta_p}{\gamma_p}$, smaller values of K (stronger binding sites) are dominant over larger values of K in terms of gene expression. When an organism is heterozygous for K, the allele with the larger value of K (weaker binding site) will be maximally expressed and will contribute more to the total expression of the two genes than the allele with the smaller value of K (stronger binding sites).

Solution for $K_1 \leq \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$

For the case $K_1 \leq \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$ the intersection of the functions $y(p) = p$ and $z(p) = \frac{\beta_p}{\gamma_p}(\theta(\bar{p} < K_1) + \theta(\bar{p} < K_2))$ are plotted in figure 4.4. From this it is clear that when $K_1 \leq \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$, the line $y(p)$ will always intersect the line $z(p)$ when $p = K_1$. Therefore if K_1 and K_2 differ, and the organism is heterozygous, the total expression of both alleles together will be the same as for an organism which is homozygous, with both alleles having binding sites of strength K_1. In this sense, allele 1 shows complete dominance over allele 2 for gene expression. This is the opposite to the previous case. Therefore when $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 \geq \frac{\beta_p}{\gamma_p}$ stronger binding sites (smaller K) are dominant over weaker binding sites, and when $K_1 \leq \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$, stronger binding sites are recessive to weaker binding sites in terms of gene expression.

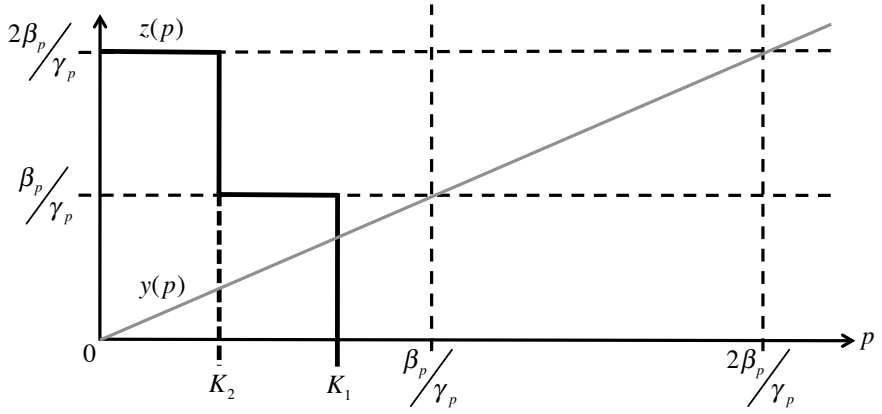

Figure 4.4: Solution for $K_1 \leq \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$. When the function $y(p) = p$ intersects the function $z(p)$. From this it is clear that the intersection always occurs when $p = K_1$

To calculate the expression levels of each allele, take $\bar{p}_1 + \bar{p}_2 = K_1$ in equation 4.22. Since $K_2 < K_1$, this immediately gives $\bar{p}_2 = 0$. That is, the expression of allele 2 is completely suppressed. The expression level of allele 1 is then $\bar{p}_1 = K_1$. In the homozygous case in which $K_1 = K_2$, both alleles are expressed at the same level $\bar{p}_1 = \bar{p}_2 = \frac{K_1}{2}$. This means that allele 1 is expressed at twice the level at which it is expressed in the homozygous case.

Therefore in the case $K_1 \leq \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$, larger values of K (weaker binding sites) are dominant over smaller values of K (stronger binding sites) in terms of gene expression. When an organism is heterozygous for K, the allele with the larger value of K (weaker binding site) will be expressed at twice the level at which it is expressed in a homozygote whilst the allele with the smaller value of K (stronger binding site) will not be expressed at all.

Solution for $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$

For the case $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$ the intersection of the functions $y(p) = p$ and $z(p) = \frac{\beta_p}{\gamma_p}(\theta(\bar{p} < K_1) + \theta(\bar{p} < K_2))$ is plotted in figure 4.5. From this it is clear that when $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$, the line $y(p)$ will always intersect the line $z(p)$ when $p = \frac{\beta_p}{\gamma_p}$. This is independent of K_1 and K_2, and means that the heterozygous case will in general have different expression to either of the two homozygous cases.

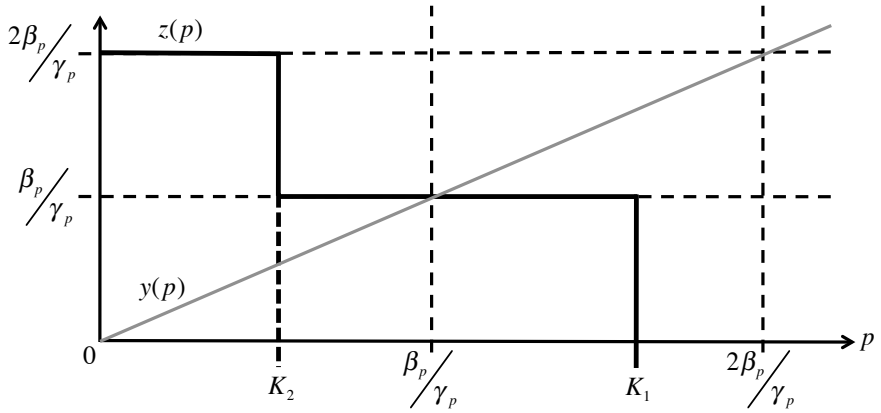

Figure 4.5: Solution for $K_1 > \frac{\beta_p}{\gamma_p}$ and $K_2 < \frac{\beta_p}{\gamma_p}$. When the function $y(p) = p$ intersects the function $z(p)$. From this it is clear that the intersection always occurs when $p = \frac{\beta_P}{\gamma_P}$

To calculate the expression levels of each allele, take $\bar{p}_1 + \bar{p}_2 = \frac{\beta_p}{\gamma_p}$ in equation 4.22. Since $K_2 < \frac{\beta_p}{\gamma_p}$, this immediately gives $\bar{p}_2 = 0$. That is, the expression of allele 2 is completely suppressed. The expression of allele 1 is then $\bar{p}_2 = \frac{\beta_p}{\gamma_p}$. Allele 2 is expressed at its maximum possible level.

Smaller Hill Coefficients

The results we have presented are for a Hill coefficient $n \to \infty$. However this is not biologically realistic. In most cases the repression function, which is determined by the probability of a TF being bound to its binding site, will have a much shallower gradient. It is not possible to solve equations (4.21) and (4.22) for an arbitrary Hill coefficient, n. Therefore we use computation to compare the effects of small Hill coefficients on gene expression with the case in which $n \to \infty$. To do this we employ a measure of the degree of dominance in gene expression when K_1 and K_2 differ. The degree of dominance, d_p, is given by [91]

$$d_p = \frac{\bar{p}_{12} - \frac{1}{2}(\bar{p}_{11} + \bar{p}_{22})}{\bar{p}_{22} - \frac{1}{2}(\bar{p}_{11} + \bar{p}_{22})} \tag{4.23}$$

where $\bar{p}_{12}$ is the equilibrium expression level in the heterozygote case, and $\bar{p}_{11}$ and $\bar{p}_{22}$ are the equilibrium expression levels in the two homozygote case, where both alleles have repression coefficient of either K_1 or K_2 respectively.

When $\bar{p}_{12} = \bar{p}_{11}$, $d_p = -1$, indicating that allele 2 is completely recessive to allele 1. In our model, where we assume $K_2 < K_1$, this means that stronger autoregulatory binding sites are completely recessive to weaker binding sites. Similarly, if $\bar{p}_{12} = \bar{p}_{22}$, $d_p = 1$, indicating allele 2 is completely dominant over allele 1. If $\bar{p}_{12} = \frac{1}{2}(\bar{p}_{11} + \bar{p}_{22})$, $d_p = 0$, indicating expression is additive. For intermediate values $\bar{p}_{22} < \bar{p}_{12} < \bar{p}_{22}$, d_p is in the range $-1 < d_p < 1$. The value of d_p in this case indicates the degree of partial recessiveness (if $d_p < 0$) or dominance (if $d_p > 0$) of allele 2 to allele 1.

The degree of dominance in gene expression, for $n = 2$, $n = 5$ and $n = 10$ are compared to the case $n \to \infty$ in figure 4.6. Here we also show the expression level of alleles 1 in heterozygotes. The results show that decreasing the Hill coefficient decreases the degree of dominance in gene expression which occurs when K_1 and K_2 differ. When $n = 2$, dominance only occurs when $K \to 0$, i.e when negative autoregulation is very strong. In contrast, the differential expression of alleles continues even for small values of n. In all cases there are large regions in which one allele is close to maximally expressed, and large regions where the expression of one allele is close to zero.

4.3.2 Evolution of Noise in Gene Expression

We now use the stochastic model of gene expression to determine the level of noise autoregulating diploid genes. Once again, we employ the assumption that K_1 and K_2 differ between alleles. We also assume that the Hill coefficient is very large, $n \gg 1$, such that negative autoregulation can be approximated by a threshold function $\theta(p < K)$. Therefore equation (4.17) for the rate of transcription in a diploid organism can be written as

$$\begin{aligned} \beta_r^+ &= \beta_r(\langle p \rangle) + n\beta_r^0 \left[1 - \theta(\langle p \rangle < K_1)\right] \theta(\langle p \rangle < K_1) + n\beta_r^0 \left[1 - \theta(\langle p \rangle < K_2)\right] \theta(\langle p \rangle < K_2) \\ \beta_r^- &= \frac{n\beta_r^0}{\langle p \rangle} \left[1 - \theta(\langle p \rangle < K_1)\right] \theta(\langle p \rangle < K_1) + \frac{n\beta_r^0}{\langle p \rangle} \left[1 - \theta(\langle p \rangle < K_2)\right] \theta(\langle p \rangle < K_2) \end{aligned} \tag{4.24}$$

and equation (4.18) for the mean expression as

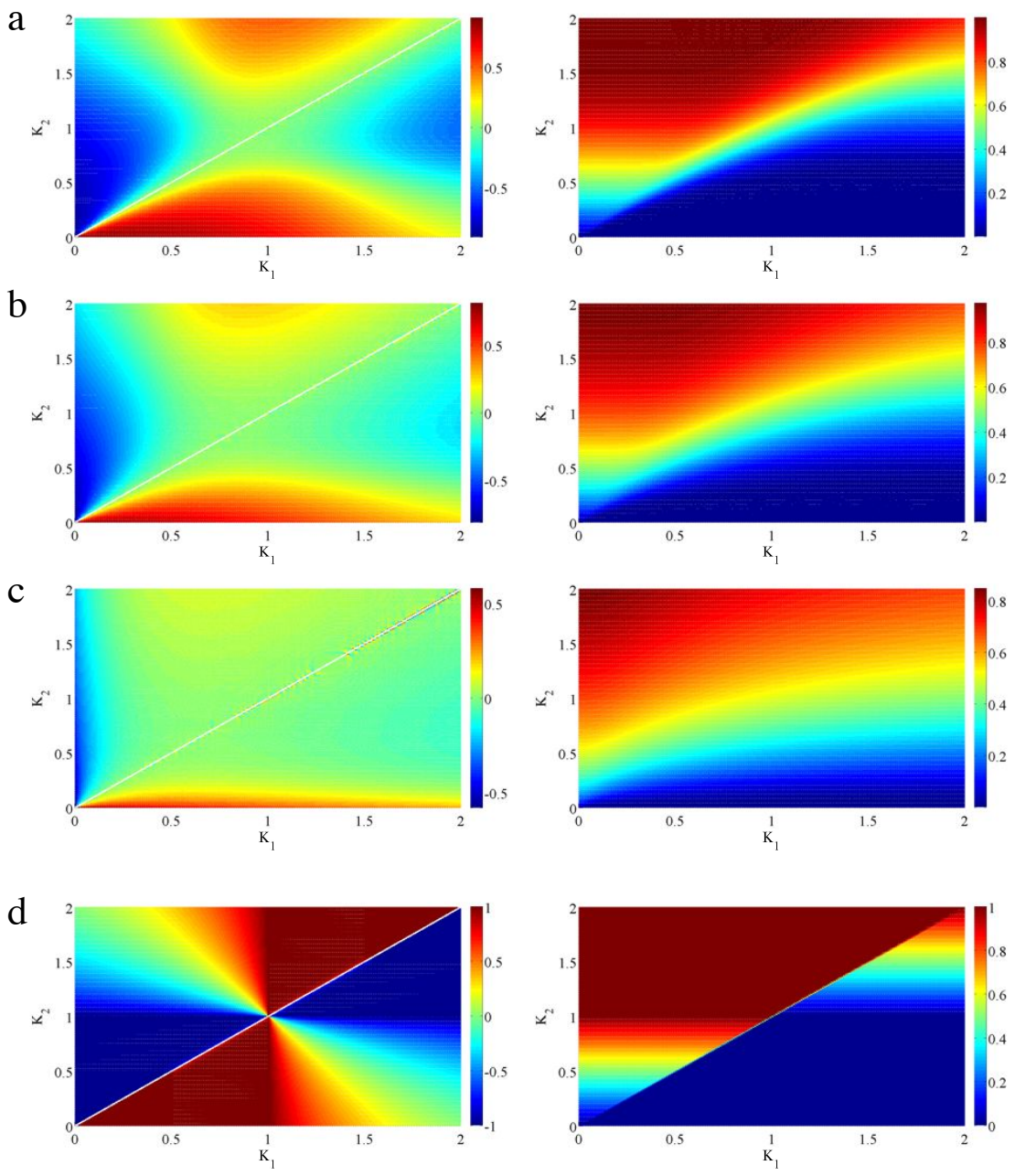

Figure 4.6: Dominance and allele expression for different Hill coefficients. The figure shows different values of K_1 and K_2, with the colour indicating the degree of dominance (left) and the level of expression of allele 1 (right). The diagonal line $K_1 = K_2$ corresponds to the heterozygous case. a) Dominance and allele 1 expression for $n = 10$. b) Dominance and allele 1 expression for $n = 5$. c) Dominance and allele 1 expression for $n = 2$. d) Dominance and allele 1 expression for $n \to \infty$.

$$\langle p \rangle = \frac{\beta_r^0 b}{\gamma_p} \left[\theta(\langle p \rangle < K_1) + \theta(\langle p \rangle < K_2) \right] \tag{4.25}$$

We employ the results of the previous section to solve equation (4.25), and then calculate β_r^- in the case $K_2 < K_1$. This is then used with equation (4.13) to calculate the noise in gene expression for different cases.

Solution for $K_1 > \frac{\beta_r^0 b}{\gamma_p}$ and $K_2 \geq \frac{\beta_r^0 b}{\gamma_p}$

In this case $\langle p \rangle = K_2$. This means that $\theta(\langle p \rangle < K_2) = 0.5$ and $\theta(\langle p \rangle < K_2) = 1$ which gives $\beta_r^- = \frac{n\beta_r^0}{4K_2}$ from equation (4.24). In contrast, for the homozygotes, we have $\beta_r^- = \frac{n\beta_r^0}{2K_1}$ for allele 1 and $\beta_r^- = \frac{n\beta_r^0}{2K_2}$ for allele 2. This means that β_r^- in the heterozygous case is always less than in the homozygous case in which both alleles have repression constant K_2. The noise is also greater in the in the heterozygous case than the homozygous case in which both alleles have repression constant K_1, unless $K_2 < 0.5K_1$. However, since K_1 lies in the range $\frac{\beta_r^0 b}{\gamma_p} < K_1 < 2\frac{\beta_r^0 b}{\gamma_p}$, in order for $K_2 < 0.5K_1$ we must have $K_2 < \frac{\beta_r^0 b}{\gamma_p}$. Since this violates our assumption that $K_1 > \frac{\beta_r^0 b}{\gamma_p}$ and $K_2 \geq \frac{\beta_r^0 b}{\gamma_p}$, we conclude that in the heterozygous case β_r^- is always smaller than in either of the two homozygous cases. Since the noise in gene expression, given by equation (4.13), is a monotonically decreasing function of β_r^-, this means that the expression of heterozygotes is always more noisy than either of the two homozygotes.

Solution for $K_1 \leq \frac{\beta_r^0 b}{\gamma_p}$ and $K_2 < \frac{\beta_r^0 b}{\gamma_p}$

In this case $\langle p \rangle = K_1$. This means that $\theta(\langle p \rangle < K_1) = 0.5$ and $\theta(\langle p \rangle < K_2) = 0$ which gives $\beta_r^- = \frac{n\beta_r^0}{4K_1}$ from equation (4.24). In contrast, for the homozygotes, we have $\beta_r^- = \frac{n\beta_r^0}{2K_1}$ for allele 1 and $\beta_r^- = \frac{n\beta_r^0}{2K_2}$ for allele 2. This means that β_r^- in the heterozygous case is always less than in either of the two homozygous cases. As a result the expression of heterozygotes is always more noisy than either of the two homozygotes.

Solution for $K_1 > \frac{\beta_r^0 b}{\gamma_p}$ and $K_2 < \frac{\beta_r^0 b}{\gamma_p}$

In this case $\langle p \rangle = \frac{\beta_r^0 b}{\gamma_p}$. This means that $\theta(\langle p \rangle < K_1) = 1$ and $\theta(\langle p \rangle < K_2) = 0$ which gives $\beta_r^- = 0$ from equation (4.24). In contrast, for the homozygotes, we have $\beta_r^- = \frac{n\beta_r^0}{2K_1}$ for allele 1 and $\beta_r^- = \frac{n\beta_r^0}{2K_2}$ for allele 2. This means that β_r^- in the heterozygous case is always less than in either of

the two homozygous cases. As a result the expression of heterozygotes is always more noisy than either of the two homozygotes.

Therefore we find that heterozygotes are always more noisy than homozygotes when negative autoregulation occurs in diploids. If selection occurs such that less noisy gene expression (smaller values of K) are favoured, this will result in under-dominance, and a barrier to the evolution of stronger binding sites.

Smaller Hill Coefficients

The results we have presented above are for a Hill coefficient $n \gg 1$. However, when smaller Hill coefficients are used it is not possible to solve equations (4.17) and (4.18). Therefore we use computation to determine the effects of small Hill coefficients on noise in gene expression. To do this we one again employ a measure of the degree of dominance in gene expression when K_1 and K_2 differ. The degree of dominance in this case, d_β, is given by [91]

$$d_\beta = \frac{\beta_{12}^- - \frac{1}{2}\left(\beta_{11}^- + \beta_{22}^-\right)}{\beta_{22}^- - \frac{1}{2}\left(\beta_{11}^- + \beta_{22}^-\right)} \tag{4.26}$$

where β_{12}^- refers to the value of β_r^- in heterozygotes, and β_{11}^- and β_{22}^- refer to the values of β_r^- in the two homozygotes, with repression coefficients K_1 and K_2 respectively.

The degree of dominance for different values of K_1 and K_2 are presented if figure 4.7, for Hill coefficients $n = 2$, $n = 5$ and $n = 10$. In all cases $d_\beta \leq 0$, indicating stronger autoregulatory binding sites (smaller values of K) are always recessive to weaker ones (larger values of K), in terms of noise in gene expression. When $d_\beta < -1$ this indicates that $\beta_{12}^- < \beta_{11}^-$ and the heterozygote has greater noise than either of the two homozygotes. From figure 4.7 it is clear that $d_\beta < -1$ occurs for a wide range of values of K_1 and K_2, even when the Hill coefficient is small. Thus, even when Hill coefficients are small, mutations which give rise to stronger autoregulation result in under-dominance in the degree of noise in gene expression. In particular under-dominance tends to arise when $K_1 < \frac{\beta_r^0 b}{\gamma_p}$ and $K_2 < \frac{\beta_r^0 b}{\gamma_p}$.

Evolution of a Single Binding Site

We can use our model to consider the evolution of a single autoregulatory binding site via point mutations. In this case the probability a TF is bound to its own promoter is given by $q(p)$ in equation (4.19). Thus, the repression function, $\beta_r(p)$, in equation (4.16) is given by $\beta_r(p) =$

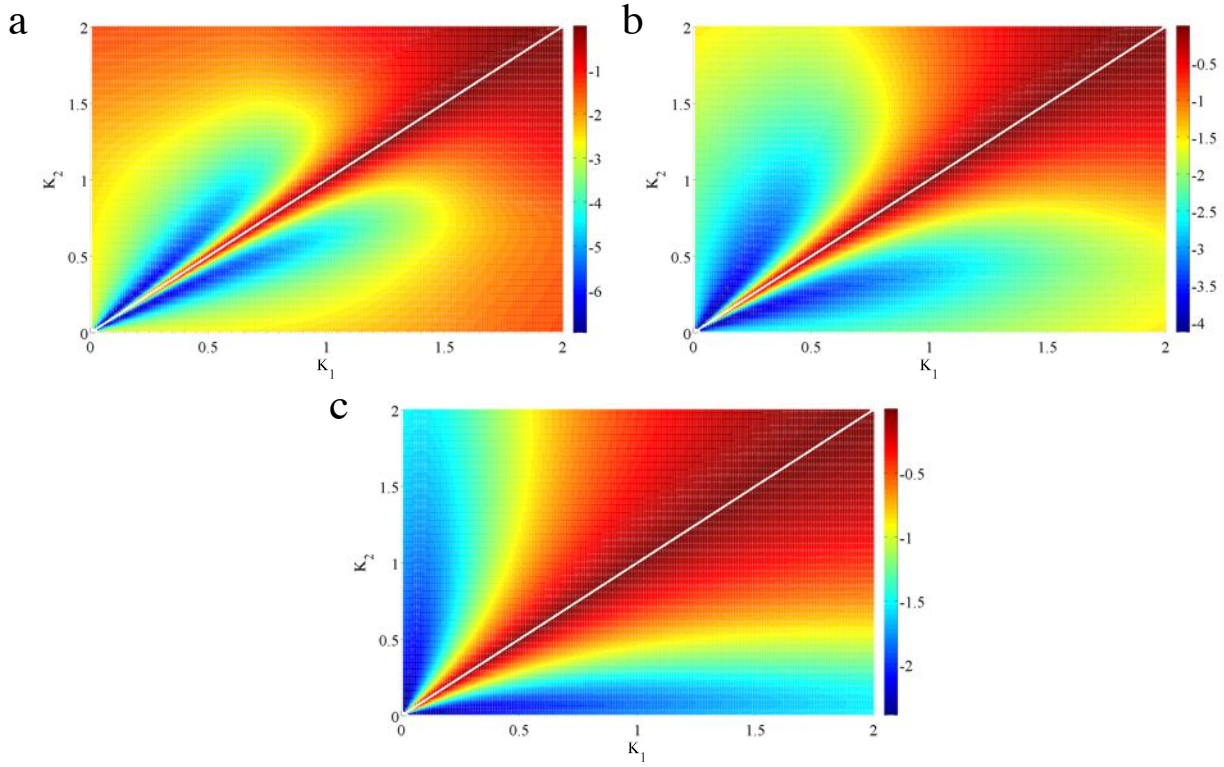

Figure 4.7: Dominance in noise for different Hill coefficients. Colours indicate the degree of dominance in the noise in gene expression for different values of K_1 and K_2. These are always negative. When dominance is less than -1, heterozygotes are more noisy than either homozygote. a) Dominance in noise for $n = 10$. b) Dominance in noise for $n = 5$. c) Dominance in noise for $n = 2$.

$\beta_r^0(1 - q(p))$, i.e by the probability that a TF is not bound to its own promoter, which is

$$\beta_r(p) = \frac{\beta_r^0}{1 + p\exp[\epsilon_0 - \epsilon r]}. \tag{4.27}$$

Therefore we have $K = \exp[\epsilon r - \epsilon_0]$. Empirically, the values of ϵ and ϵ_0 can be reasonably approximated by $\epsilon_0 \approx 0$ and $\epsilon = \frac{1}{k_B T} \approx 1.2$ [40], which gives

$$K(r) \approx \exp[1.2r] \tag{4.28}$$

and allows us to write the strength of a binding site K in terms of the number of mismatches, r, between the real and optimal binding sites. We can then consider the evolution of a single binding site through point mutations. We take $K_1 = K(r)$ and $K_2 = K(r - 1)$, and examine the noise in

all heterozygotes as a binding site evolves (moves closer to the optimal binding sequence). We plot the value of d_β against the number of correctly matched nucleotides in figure 4.8. This is done for different values of the maximum expression, $\frac{\beta_r^0 b}{\gamma_p}$, of a single allele, for a binding site 10 nucleotides in length. When the degree of dominance, $|d_\beta| > 1$, then heterozygotes have more noisy expression than either homozygote. If noise reduction is selected for, this will result in under-dominance. From figure 4.8 we see that under-dominance will occur for a single binding site with maximum expression $\frac{\beta_r^0 b}{\gamma_p} = 10000$, when 5 out of 10 nucleotides are correctly matched to the optimal binding site. Similarly, for $\frac{\beta_r^0 b}{\gamma_p} = 1000$, under-dominance occurs when 7 out of 10, and for $\frac{\beta_r^0 b}{\gamma_p} = 100$ when 9 out of 10 nucleotides are correctly matched to the optimal binding site.

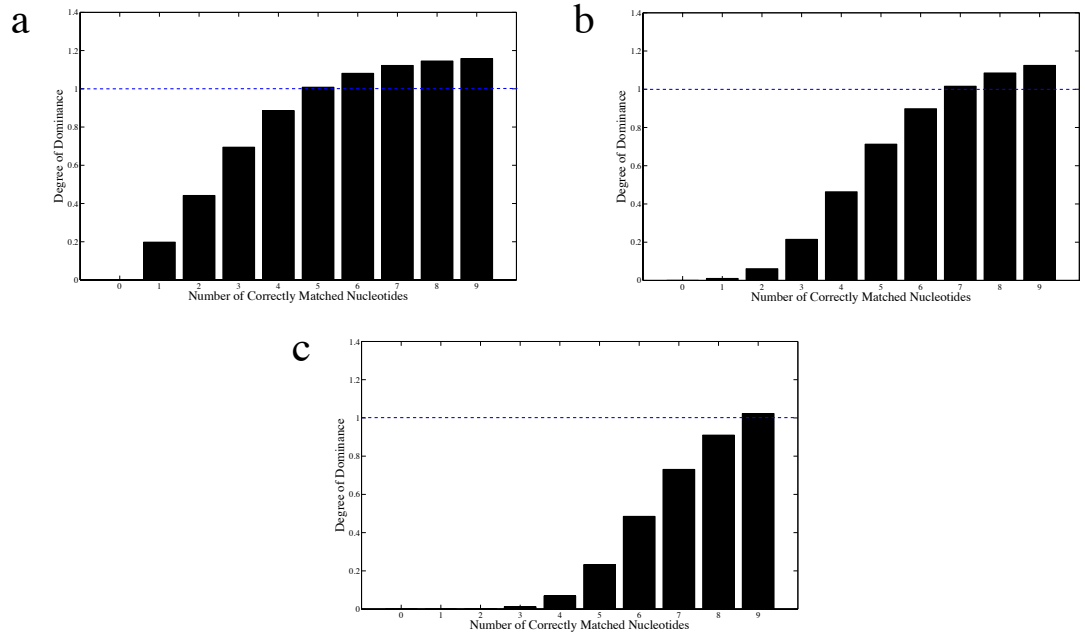

Figure 4.8: Degree of dominance for a single binding site. Horizontal axis plots the number of nucleotides correctly matched to the optimal binding site. Horizontal axis indicates the magnitude of dominance which occurs in heterozygotes in which one more nucleotide is matched to the optimal binding site. When the degree of dominance is greater than 1 (dashed blue line), under-dominance occurs, and heterozygotes are more noisy than either homozygote. a) Dominance for maximum gene expression of 10000, under dominance when 5 out of 10 nucleotides are matched to the optimal binding sequence. b) Dominance for maximum gene expression of 1000, under dominance when 7 out of 10 nucleotides are matched to the optimal binding sequence. c) Dominance for maximum gene expression of 100, under dominance when 9 out of 10 nucleotides are matched to the optimal binding sequence.

This shows that under-dominance can occur when selecting for noise reduction, even when only a single auto-regulatory binding site is present in the promoter of the gene.

4.4 Discussion

We have investigated the behaviour of negatively autoregulating genes in cells where two copies of the gene are present. This is of interest for two reasons. Firstly, although negative autoregulation is one of the most abundant network motifs in prokaryotes [110], the behaviour of such motifs in diploids has not been considered in depth. Secondly, the behaviour of duplicates of autoregulating genes is of interest in the evolution of other network motifs [117, 140]. We have constructed a model which considers a pair of alleles, 1 and 2, both of which undergo negative autoregulation. The alleles code for identical proteins, and therefore suppress the expression of both themselves and each other. However, the strength of negative autoregulation is allowed to differ between the alleles, such that they have different repression coefficients, K_1 and K_2. When this is the case, smaller values of K indicate stronger negative autoregulation. We have determined both the total expression of the pair of alleles, and the contribution of each allele to the total expression in the case where K_1 and K_2 differ. This reveals firstly that dominance tends to occur in the total expression of the alleles, and secondly that one allele will tend to be expressed at a higher level than the other. In particular, when negative autoregulation is strong, one of the pair of alleles will be almost completely off, and contribute little to the total expression of the pair.

We then investigated the noise in the total expression of a pair of alleles. This is of interest as auotregulation is thought to serve as a mechanism of noise reduction in prokaryotes [1, 11, 118, 110]. This reveals that, in a wide range of cases, heterozygotes will be more noisy than either of the two possible homozygote cases. Thus, under-dominance will occur in terms of the noise in the expression of paris of alleles.

4.4.1 Barrier to the Evolution of Autoregulation in Diploids

The occurrence of under-dominance in the noise levels of negatively autoregulating genes has interesting implications. In particular, if noise reduction is selected for, it suggests that there exists a barrier to the evolution of negative autoregulation. To further characterize this effect we have investigated the evolution of a single binding site through single nucleotide substitutions. This suggests that, for genes with a high maximum expression [118] (figure 4.7a), a binding site of 10 nucleotides, will encounter under-dominance when 5 out of the 10 nucleotides are correctly matched to the optimal binding site. Thus, binding sites which are strongly matched to the optimal binding sequence cannot evolve. Form this we may conclude that, where negative autoregulation has evolved as a mechanism for noise reduction in diploids, the binding may tend to be weak. In

contrast, for haploid organisms there is no barrier to the evolution of negatively autoregulating binding sites, and as such we may expect stronger binding sites to evolve.

The effectiveness of negative autoregulation as a mechanism for noise reduction is increased when multiple TF binding sites must be bound in order for repression to occur [18]. However, our results indicate that, in such cases, the tendency for strengthening of TF binding sites to result in under-dominance is increased, and thus results in an even stronger barrier to the evolution of negative autoregulation (figure 4.7). Therefore we conclude that the evolution of negative autoregulation as a mechanism for noise reduction faces significant difficulties in the diploid case as compared to the haploid case. We therefore expect that the frequency of negative autoregulation in diploids is likely to be less than observed in haploids.

In order to develop an intuition as to why under-dominance in the level of noise reduction occurs in diploids, we consider the expression levels of individual alleles in heterozygotes. As described above, in heterozygotes, the allele with the stronger autoregulatory binding site tends to be under-expressed, and the allele with the weaker autoregulatory binding site tends to be over-expressed, compared to the case in homozygotes. Noise level is given by the ratio of the variance in gene expression to the mean gene expression. The variance is determined by the strength with which a gene's expression level is returned to equilibrium following a perturbation, i.e by the strength of negative autoregulation. In a homozygote, each allele has mean expression $\frac{K}{2}$ and the strength of negative autoregulation is K. However, in heterozygotes the allele with the weaker autoregulatory binding site has increases expression as compared to the expression level in the homozygous case. As a result, it has an increased noise level as compared to the homozygous case. The other allele has reduced gene expression as well as an increase in the strength of negative autoregulation. Therefore it has reduced noise level compared to the homozygous case. However, the allele with reduced noise contributes little to total gene expression and in many cases may be completely silent. Therefore total gene expression is dominated by the allele with increased noise, and the total noise in the expression of heterozygotes is increased as compared to homozygotes.

4.4.2 Frequency of Autoregulation in Yeast and *E. coli*

In order to test this hypothesis we compare the frequency of negative autoregulation in *E. coli* with that observed in *S. cerevisiae* [47, 88, 110]. The results are shown in table 4.1.

From this we see that, whilst 37% of TFs undergo negative autoregulation, comprising 7% of all interactions in *E. coli*, in *S. cerevisiae* between 2% [88] and 4% [47] of TFs undergo negative

	E. coli [110]	*S. cerevisiae* [88]	*S. cerevisiae* [47]
TFs	115	131	124
Interactions	578	1094	909
Negative Auto	42	3	5
Positive Auto	14	9	5
Dual Auto	4	0	0

Table 4.1: Frequency of autoregulation in *E. coli* and *S. cerevisiae*

autoregulation, comprising between 0.3% [88] and 0.6% [47] of all interactions. There is an order of magnitude difference in the frequency of negative autoregulation between *E. coli* and *S. cerevisiae*. This is precisely in line with the results of our model. It is believed that negative autoregulation has evolved as a mechanism of noise reduction in *E. coli*, which is haploid. In *S. cerevisiae*, which exists most frequently as a diploid, the difficulty of evolving negative autoregulation due to under-dominance would be expected to lead to a dramatic reduction in the frequency of negative autoregulation. This is precisely what is observed.

4.4.3 Dominance Arising from *cis* Mutations

An intriguing insight from our model is that, when genes negatively autoregulate, mutations in *cis* will tend to give rise to dominance in their effects on gene expression. This is in contrast to *cis* mutations in most cases, which tend to be additive, with *trans* mutations tending to result in dominance [70, 73]. In diploids mutations to a TF binding site will tend to affect the expression of the allele at which it occurs, whilst leaving the expression of the other allele unchanged. This makes evolution of *cis* regulatory regions easier than *trans* evolution.

This conclusion is supported by observations in *Drosophila melanogaster*, where it is observed that genetic divergence between different lineages is associated with variation in *cis* [70]. This suggests that alleles with *cis* mutations may be preferentially fixed by positive natural selection because they tend to be additive more frequently [70]. In contrast the disruption of gene expression by recessive variation resulting from alleles with *trans* mutations suggests that these may be important in understanding variation within populations [70]. The pervasive dominance effects associated with negative autoregulation suggests that *cis* mutations at these genes may evolve differently to *cis* regulatory regions at other genes. If mutations with higher additivity are favoured by positive natural selection, this may provide a further explanation of the relative lack of negative

autoregulation in *S. cerevisiae* as compared to *E. coli*.

4.4.4 The Fate of Silent Alleles

Our investigation of the contributions of different alleles to total gene expression reveals that the allele with the weaker negative autoregulation tends to be over-expressed whilst the allele with the stronger negative autoregulation is under-expressed. In particular, where autoregulatory binding sites are strong, the allele with the stronger binding site will have its expression almost completely suppressed. We refer to such suppressed alleles as silent.

This conclusion applies both to heterozygous diploids, and to duplicates of negatively autoregulating genes in haploids which differ in the strength of their negative autoregulation. The tendency for alleles to be silent applies quite generally to pairs of identical negatively autoregulating genes. This can be seen from figure 4.6, in which allele expression can be seen to be low in one allele for a wide range of values of binding site strength. Where an allele is silent, any further mutations it suffers, either in *cis* or in *trans*, will be unexpressed.

In a diploid population, silent alleles will only be expressed in individuals which are heterozygous for the allele with stronger negative autoregulation. If the frequency of such alleles is low, copies of the allele may be silent for a number of generations. If an allele goes unexpressed for a number of generations, there is an increased probability that it will have suffered a mutation when it is next expressed. Thus silent alleles provide a potential store for genetic variation in diploid populations. The tendency of alleles with stronger negative autoregulation to accumulate mutations may provide another reason for the relative absence of negative autoregulation in *S. cerevisiae* as compared to *E. coli*.

4.4.5 Duplication of Autoregulators in Haploids

Duplication of negatively autoregulating genes in haploids is of interest for two reasons. Firstly, such duplications provide a potential mechanism through which other, larger network motifs may evolve [117]. Secondly, the expression of a pair of duplicate negatively autoregulating genes with identical binding site strength is the same (or close to) that of the unduplicated gene [140]. As such, the possible deleterious effects associated with gene duplication which result from increased dosage will be mitigated in genes which negatively autoregulate. This has led to two hypotheses. Firstly, that duplicate pairs of autoregulating genes will tend to participate in larger network motifs, and secondly, that negatively autoregulating genes will tend to have more duplicate copies fixed in

a population, compared to genes which do not negatively autoregulate [117, 140]. Interestingly, analysis of the *E. coli* transcription network reveals that neither of these hypotheses hold true [117, 140]. The analysis of negatively autorgulating genes in diploids presented here provides a possible explanation for this.

In particular, our analysis shows that, in order for a negatively autoregulating gene to avoid becoming silent it must maintain a level of negative autoregulation close to that of its duplicate pair. Any change to its negatively autoregulatory binding site, or any other mutation in *cis* that alters the strength of negative autoregulation, will result in the expression of one of the pair being suppressed. As discussed above, this will lead to the suppressed gene being released from selection, and free to fix further mutations. In effect it will become a pseudogene, and will tend to be lost from the population. In addition, the constraint placed by negative autoregulation on the *cis* regions of duplicate pairs reduces the ability of the pair to evolve divergent expression patterns and acquire new functions. It is only if one of the pair undergoes a *trans* mutation which affects the ability of the gene to autoregulate that a duplicate can avoid these constraints (see Appendix C). This removes the expectation that duplicates of negative autoregulators may be favoured over duplicates of other genes, and is therefore in line with observations in the *E. coli* transcription network.

4.4.6 Transcriptional Bursting in Eukaryotes

The stochastic model for gene expression in diploid organisms presented here allows a comparison of the expression dynamics of negatively autoregulating genes with that observed in haploids. However, it does not take account of the fact that in many cases haploid organisms will be prokaryotes whilst diploids will be eukaryotes. The model presented for haploids provides a good description of transcription in prokaryotes. Gene expression in eukaryotes has somewhat different characteristics to that of prokaryotes [101]. In particular, eukaryotic genes undergo transcriptional "bursts" in which the gene itself randomly switches between a state of transcriptional activity and inactivity [101]. As a result eukaryotes tend to have more noisy gene expression than prokaryotes.

The origins of increased noise in eukaryote gene expressions still the subject of debate, however it is thought to result from changes in nucleosome occupancy at the transcription start site [101]. However, the increased noise observed in eukaryotic cells provides an additional explaination for the relative lack of negative autoregulation in *S. cerevisiae* - that noise reduction of this type is not necessary in eukaryotes. Whilst this does not contradict the results presented here, it suggests

that a combination of factors may be responsible for these observations. Both the difficulty of evolving negative autoregulation in diploids, as well as the lack of a need to do so may explain why negative autoregulation is less prevalent in *S. cerevisiae* as compared to *E. coli*.

4.5 Conclusion

We have presented a model for the evolution of negative autoregulation in diploids. Negative autoregulation is of interest as it is one of the most prevalent network motifs in prokaryotes, where it is thought to function as a mechanism for the reduction of noise in gene expression. Our results show that in diploids, when a pair of alleles which differ in the strength of their negative autoregulation, this leads to dominance in the level of gene expression. In addition, the contributions of the different alleles to total gene expression is often highly assymetrical. In particular, when negative autoregulation is strong, the allele with the stronger negative autoregulation will be almost completely suppressed. It is suggested that such silent alleles will be more prone to the accumulation of mutations.

Dominance also occurs in the degree of noise in gene expression. In particular, in a wide range of cases, heterozygotes are found to be more noisy than either of the two homozygote cases. Therefore, in negative autoregulation is selected for as a mechanism of noise reduction, this would lead to a barrier to the evolution of stronger negative autoregulation. Comparison of the frequency of negative autoregulation in *S. cerevisiae* and *E. coli*, reveals an order of magnitude difference between the two cases. This observation supports our conclusion that there is a barrier to the evolution of negative autoregulation in diploids.

4.6 Appendix C

4.6.1 Stochastic Model of Negative Autoregulation

We have employed a model from the literature [118] in order to determine the noise in gene expression when diploid negative autoregulation occurs. Therefore we quote the results for the mean gene expression and noise in gene expression from this model without proof. Since we are assuming that alleles in the diploid model are identical except for the strength of negative autoregulation, we use results for the haploid model with the rate of mRNA production $\beta_r(p)$ given by equation 4.16. In order to obtain equation (4.14) for the mean protein expression, we

combine equations (4.12) and (4.13) to give

$$\begin{aligned} \langle p\rangle \gamma_p + \frac{bn}{\beta_r^0}\left[\beta_r^0 - \beta_r(\langle p\rangle)\right]\beta_r(\langle p\rangle) &= b\beta_r(\langle p\rangle) + \frac{bn}{\beta_r^0}\left[\beta_r^0 - \beta_r(\langle p\rangle)\right]\beta_r(\langle p\rangle) \\ \langle p\rangle &= \frac{b}{\gamma_p}\beta_r(\langle p\rangle) \end{aligned} \tag{4.29}$$

replacing β_r^p from equation (4.9) then gives equation (4.14).

4.6.2 Mutations in *trans*

Mutations in *trans* affect the function of a transcription factor itself. Therefore we assume that it affects the ability of the TF to bind to both copies of the allele equally. When the two alleles differ as the result of a *trans* mutation, our ODE model for gene expression in diploids becomes

$$\begin{aligned} \frac{dp_1}{dt} &= \beta_p\left(\frac{1}{1+\left(\frac{p_1}{K_1}+\frac{p_2}{K_2}\right)^n}\right) - \gamma_p p_1 \\ \frac{dp_2}{dt} &= \beta_p\left(\frac{1}{1+\left(\frac{p_1}{K_1}+\frac{p_2}{K_2}\right)^n}\right) - \gamma_p p_2 \end{aligned} \tag{4.30}$$

from this it is immediately obvious that the equilibrium expressions of the two alleles are identical $\bar{p}_1 = \bar{p}_2$. Since the total gene expression is given by $\bar{p} = \bar{p}_1 + \bar{p}_2$ this gives $\bar{p}_1 = \bar{p}_2 = \frac{\bar{p}}{2}$. Combining equations (4.30) and solving for the equilibrium protein concentration then gives

$$\bar{p} = \frac{2\beta_p}{\gamma_p}\frac{1}{1+\left(\bar{p}\left(\frac{1}{2K_1}+\frac{1}{2K_2}\right)\right)^n} \tag{4.31}$$

which is a Hill function with repression coefficient $K = \frac{2K_1K_2}{K_1+K_2}$. If a *trans* mutation occurs such that one allele loses its ability to regulate the other, $K_2 \to \infty$, then the effective binding coefficient becomes $K \to 2K_1$. Therefore a *trans* mutation allows autoregulation to be lost in one allele without the other allele becoming silent.

Chapter 5

Conclusion and Further Work

5.1 Conclusion

The aim of this thesis is to investigate the construction of transcription factor networks (TFNs) through natural selection. TFNs are central to many important evolutionary questions. They capture the way sets of genes interact with one other to produce complex patterns of coordinated gene expression. Patterns of gene expression are central to determining the function and behaviour of a cell which in turn goes to determine the phenotype of an organism.

It is beyond the scope of any single piece of work to fully characterise the evolution of TFNs. Rather than attempt to do this, I have focused on three important aspects of TFN evolution, which are related to three different aspects of TFN organisation. It now remains to tie these three aspects together, in order to construct a picture of how TFNs as a whole are constructed through natural selection.

The first property of TFNs that was considered was the degree distribution. The degree distribution of a network is one of the most general ways of characterising its structure. It is described by $n(k)$, which is the frequency distribution of nodes in the network that have k edges - i.e which have degree k. A TFN is a directed network, such that genes may have incoming edges - indicating they are regulated by another gene - or outgoing edges - indicating that they regulate another gene. As such, a TFN has two degree distributions, one for the frequency distribution of incoming edges, $n_{in}(k)$, and one for the frequency distribution of outgoing edges, $n_{out}(k)$. In contrast to most other biological networks, the *in*- and *out*-degree distributions of a TFN are very different. The *in*-degree distribution is best described by an exponential distribution, whilst the *out*-degree

distribution has a broad tail, which can be described by a power-law distribution [47, 83, 119].

We constructed a model for the evolution of the TFN degree distribution through *cis* and *trans* mutation, gene duplication and deletion. We allowed degree dependence in the rate of fixation of mutations. This included preferential attachment, such that genes gain new interactions at a rate proportional to the number of interactions they already participate in. This occurred for both incoming and outgoing edges. We also allowed degree dependence such that TFs fixed *trans* mutations more slowly, the more interactions they participate in. We showed that the *in*- and *out* degree distributions observed in yeast and *E. coli* can only be reproduced if degree dependence in the rate of *trans* evolution is included in the model. In addition the observed rates of mutation in these organisms suggest that preferential attachment in both incoming and outgoing edges occurs. From this model we are able to draw conclusions about the evolution of global TFN structure. In particular, we conclude that the rate at which mutations are fixed at different genes depends on their position in the network.

The second property of TFNs considered was the evolution of cooperative binding between different TFs. Cooperative binding of the type considered occurs when a pair of TFs co-regulate a set of genes. They are able to bind to the promoter regions of regulated genes independently, through their specific binding sites. This can be aided by a protein-protein interaction between the pair of TFs. This allows one TF to increase the strength of binding of the other TF to regulated genes. The presence of a protein-protein interaction therefore decreases the constraint on the specific binding sites of one of the TFs, since it is able to compensate for changes to the binding strength of the specific site.

This model considers neutral evolution, since we assume that both TFs must always bind to all the targets that they co-regulate, such that the expression of the regulated genes remains unchanged. However, the manner in which the genes are regulated may vary, depending on whether a protein-protein interaction is present or absent. Such a model describes observed changes in the yeast sex determination network, in which neutral evolution of the type considered in this model has been observed [124]. Our results showed that the probability of a protein-protein interaction between the two TFs becoming fixed in a population follows a threshold function in the number of target genes regulated. The threshold number of regulated genes is determined by the rate of mutations in *cis*, which affect specific binding sites, and in *trans* which affect the protein-protein interaction. In large populations a protein-protein interaction will become fixed when doing so increases the robustness of the network to deleterious mutations. In small populations,

where mutational robustness is less likely to evolve [127], a threshold still occurs. In this case the threshold is softer than in the large population case, but once again depends on the rates of *cis* and *trans* mutations. However the position of the threshold in the large and small population cases is different in general.

From this work we conclude that fixation of a protein-protein interaction in a population is likely to occur in response to a change in the number of genes co-regulated by a pair of TFs (regulon size). This is corroborated by observations in the yeast transcription network, in which it is found that significant changes in the sets of genes co-regulated by pairs of TFs have occurred [125]. We also conclude that changes in population size may be able to drive loss or gain of protein-protein interactions. As such, we conclude that the changes observed in the yeast sex determination network can indeed occur as a result of neutral evolution.

The final property of TFNs that is considered is the frequency of negative autoregulation observed in different networks. Negative autoregulation is one of the most abundant network motifs found in *E. coli* [110]. It has been shown to function as a mechanism of noise reduction in gene expression, both experimentally and theoretically [1, 106, 118]. We extended the theoretical analysis carried out in haploid organisms, to the case of diploids. The expression dynamics of genes which undergo autoregulation was considered. This was expected to differ from the haploid case since a pair of identical autoregulating genes form a network of three feedback loops and four regulatory interactions. In contrast, a haploid autoregulating gene consists of a single interaction and a single feedback loop.

We investigated both the mean expression and noise in gene expression of negatively autoregulating genes in diploids. We showed that when the two genes differ in the strength of negative autoregulation (the heterozygous case), the total expression of the pair of genes will show dominance, such that it is similar to one of the two possible homozygous cases. We also showed that the contribution of the different alleles to the total gene expression shows significant differences. In particular, the allele with weaker negative autoregulation will always be over expressed, and the allele with stronger negative autoregulation under expressed, relative to their expression levels in the homozygous cases. We also showed that when negative autoregulation is strong, in the heterozygous case one of the alleles will be almost completely unexpressed, with the total expression accounted for entirely by the allele with weaker negative autoregulation.

Similarly, the noise in gene expression in heterozygotes shows significant dominance. Importantly, in this case heterozygotes will always show a noise level closer to the homozygous case with

the weaker level of negative autoregulation. Further, in a wide range of cases, heterozygotes show a noise level which is greater then either of the two homozygous cases. Thus there is under-dominance in the level of noise in gene expression. This suggests that, if negative autoregulation is selected for as a form of noise reduction, it will be much more difficult to evolve in diploids than in haploids. This is corroborated by observations in the yeast transcription network, in which the frequency of negatively autoregulation is an order of magnitude less than in the *E. coli* transcription network [47, 88, 110].

We also conclude that the the existence of unexpressed alleles, described for diploid autoregulation, may be used to explain observations concerning the evolution of duplicates of autoregulating genes in haploid organisms. In particular, it has been suggested that genes which autoregulate may have duplicates fixed more frequently. However this is not observed [117, 140]. We suggest that this is explained by the tendency of one of the duplicate pair to become unexpressed following mutations in *cis*.

The picture of TFN evolution that emerges from these three pieces of work is as follows. Firstly, the position of a gene in the network (i.e the number and type of interactions it participates in) has a significant impact on the way that gene evolves. This is seen in the degree distribution of the network, in which degree dependence in rates of evolution is required to reproduce observations. It is seen in the evolution of cooperative binding, in which protein-protein interactions become fixed or lost in response to the number of target genes co-regulated. It is seen in the evolution of autoregulation, which is heavily favoured in *E. coli* but is relatively rare in yeast. This suggests that insight into the evolutionary path of a gene may be gained by characterising its position in a TFN - through it's degree, through the number of genes it co-regulates with other TFs and through the presence or absence of autoregulation. I suggest that this is likely to hold true for other measures of a gene's position in a TFN, such as participation in larger network motifs (e.g feed forward loops). Therefore, by determining general rules about how genes which occupy different positions in a TFN evolve, we can construct a picture of how TFNs as a whole evolve. For example we can determine which TFs are likely to undergo the highest rate of *trans* or *cis* evolution, or the likely fates of duplicates of genes which occupy different positions in the network.

The second conclusion which emerges, is that characterising the function of a particular network is not in itself sufficient to characterise the evolution of that network. In particular, the network structure which is adopted by evolution has been shown here to depend heavily on population genetic factors. These include the relative rates of different types of mutation, population size,

whether recombination occurs in a population, and whether an organism is haploid or diploid. In general there may be many different network structures which are capable of performing a particular function. When this is the case, the network structure that is adopted is likely to be both the result of functional optimisation, and a result of population genetic factors. These in turn may be the result of the environment in which a population exists or its evolutionary history. This suggests that we must be cautious when attempting to determine the function of networks observed in living organisms. Studying the function of such networks abstracted from the population genetic details of the organism's lifestyle may lead to false conclusions about that function.

The final conclusion which emerges concerns the role of *trans* mutations in TFN evolution. It has often been suggested that the majority of evolutionary change which occurs in TFNs occurs through changes in *cis* [43, 70, 97, 96, 115, 142]. In contrast, *trans* evolution has been said to occur rarely due to the pleiotropic effects of changes in *trans*. Whilst this may be true, the work presented here suggests that the role of *trans* evolution should not be neglected. Our results suggest that some degree of *trans* evolution is required to explain the observed degree distribution of TFNs. More significantly, our study of the evolution of co-regulation between pairs of TFNs suggests that changes in *trans* may fascilitate changes in *cis*, such that the two processes cannot be considered in isolation from one another. Thus we conclude that changes to protein-protein interactions between TFs may be an important factor in driving evolutionary changes in *cis* and the evolution of TFNs as a whole.

5.2 Further Work

The work presented here suggests three important directions for further work. The first direction consists of a more general analysis of how diploid TFNs evolve as compared to haploids. This was carried out for the particular case of negative autoregulation. A similar analysis could be carried out for the full range of network motifs observed in *E. coli*, yeast and higher eukaryotes. Such an analysis would require the effects of mutations on the function of these networks in haploids and diploids to be determined. For motifs of more than one gene in higher eukaryotes, it would also require the inclusion of recombination. Once this was achieved we would be able, for example, to investigate the differences observed in the motif distributions in *E. coli*, *S. cerevisiae* and *Drosophila*. This in turn would allow us to determine to what extent the differences in network structure between these organisms is determined by being haploid vs. diploid, or by a low vs. high rate of recombination, and to what extent the differences reflect intrinsic differences in the functions

of these networks.

The second direction consists of a more general analysis of the co-evolution of the protein-protein interaction network and the transcription factor network. This can be approached both theoretically, through computational modelling, and through the use of bioinformatic data on the structure of these networks. As a starting point, the results of chapter 3 could be tested more generally using data on the protein-protein interaction network of yeast. The results of this chapter suggest that gain of protein-protein interactions should be correlated with an increase in the number of genes co-regulated by the interacting TFs, whilst loss of protein-protein interactions should be correlated with a decrease in the number of genes co-regulated by the interacting TFs. This hypothesis could be tested directly, and provide a starting point for further work into the coevolution of these two networks.

The third direction concerns the impact of whole genome duplication on the evolution of transcription factor networks. There have been a number of recent studies on the impact of whole genome duplication on the evolution of gene expression [21, 50, 95, 59, 149]. These provide empirical evidence against which models of the evolution of transcription networks through whole genome duplication can be tested. Whole genome duplications provides a wealth of raw genetic material on which natural selection can act. However the constraints on how they can occur and subsequently be integrated into the genetic architecture of an organism are poorly understood. The methodologies developed here could be employed to elucidate these questions. In particular, the impact of a whole genome duplication on the global architecture of transcription and protein networks could be approached both theoretically and computationally, using the methods developed in Chapter 2 of this thesis.

Bibliography

[1] U. Alon. *An introduction to systems biology: design principles of biological circuits*, volume 10 of *Chapman and Hall/CRC mathematical and computational biology series*. Chapman and Hall/CRC, Boca Raton, FL, 2007.

[2] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007. ISI Document Delivery No.: 169PF.

[3] R. B. Azevedo, R. Lohaus, S. Srinivasan, K. K. Dang, and C. L. Burch. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature*, 440(7080):87–90, 2006.

[4] S. Balaji, M. M. Babu, and L. Aravind. Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of e. coli. *J Mol Biol*, 372:1108–1122, 2007.

[5] S. Balaji, M. M. Babu, M. I. Lakshminarayan, N. M. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol*, 360:213–227, 2006.

[6] S. Balaji, M. I. Lakshminarayan, L. Aravind, and M. M. Babu. Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. *J Mol Biol*, 360:204–212, 2006.

[7] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[8] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–13, 2004.

[9] S. Basu, R. Mehreja, S. Thiberge, M.-T. Chen, and R. Weiss. Spatiotemporal control of gene expression with pulse-generating networks. *Proceedings of the National Academy of Sciences*, 101(17):6355–6360, 2004.

[10] N. N. Batada and L. D. Hurst. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet*, 39(8):945–9, Aug 2007.

[11] A. Becskei and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–3, Jun 2000.

[12] A. Bergman and M. L. Siegal. Evolutionary capacitance as a general feature of complex gene networks. *Nature*, 424(6948):549–552, 2003. Article NATURE.

[13] A. Bhan, D. J. Galas, and T. G. Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493, 2002.

[14] O. Brandman, J. Ferrell, James E., R. Li, and T. Meyer. Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science*, 310(5747):496–498, 2005.

[15] G. O. Bryant, V. Prabhu, M. Floer, X. Wang, D. Spagna, D. Schreiber, and M. Ptashne. Activator control of nucleosome occupancy in activation and repression of transcription. *PLoS Biol*, 6(12):e317, 2008.

[16] S. Carroll. Evolution at two levels: On genes and form. *Plos Biology*, 3(7):1159–1166, July 2005.

[17] T. Casneuf, S. De Bodt, J. Raes, S. Maere, and Y. Van de Peer. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant arabidopsis thaliana. *Genome Biol*, 7(2):R13, 2006.

[18] D. Chu, N. R. Zabet, and B. Mitavskiy. Models of transcription factor binding: Sensitivity of activation functions to model assumptions. *Journal of Theoretical Biology*, 257(3):419–429, Apr. 2009.

[19] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas. Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–687, 2003.

[20] S. Ciliberti, O. C. Martin, and A. Wagner. Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput Biol*, 3(2):e15, 2007.

[21] G. Conant. Rapid reorganization of the transcriptional regulatory network after genome duplication in yeast. *Proceedings of the Royal Society B: Biological Sciences*, 277(1683):869–876, 2010.

[22] O. X. Cordero and P. Hogeweg. Feed-forward loop circuits as a side effect of genome evolution. *Molecular Biology and Evolution*, 23(10):1931–1936, 2006.

[23] M. Cosentino Lagomarsino, P. Jona, B. Bassetti, and H. Isambert. Hierarchy and feedback in the evolution of the escherichia coli transcription network. *Proc Natl Acad Sci U S A*, 104(13):5516–20, Mar 2007.

[24] B. C. Daniels, Y.-J. Chen, J. P. Sethna, R. N. Gutenkunst, and C. R. Myers. Sloppiness, robustness, and evolvability in systems biology. *Current Opinion in Biotechnology*, 19(4):389–395, 2008. 0958-1669 doi: DOI: 10.1016/j.copbio.2008.06.008.

[25] E. Davidson, J. Rast, P. Oliveri, A. Ransick, C. Calestani, C. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. Brown, C. Livi, P. Lee, R. Revilla, A. Rust, Z. Pan, M. Schilstra, P. Clarke, M. Arnone, L. Rowen, R. Cameron, D. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–1678, Mar. 2002.

[26] J. C. Davis and D. A. Petrov. Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet*, 21(10):548–51, Oct 2005.

[27] J. A. de Visser, J. Hermisson, G. P. Wagner, L. Ancel Meyers, H. Bagheri-Chaichian, J. L. Blanchard, L. Chao, J. M. Cheverud, S. F. Elena, W. Fontana, G. Gibson, T. F. Hansen, D. Krakauer, R. C. Lewontin, C. Ofria, S. H. Rice, G. von Dassow, A. Wagner, and M. C. Whitlock. Perspective: Evolution and detection of genetic robustness. *Evolution*, 57(9):1959–72, 2003.

[28] G. Dieci and A. Sentenac. Detours and shortcuts to transcription reinitiation. *Trends in Biochemical Sciences*, 28(4):202, 2003.

[29] S. W. Doniger and J. C. Fay. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol*, 3(5):e99, 2007.

[30] J. Draghi and G. P. Wagner. The evolutionary dynamics of evolvability in a gene network model. *Journal of Evolutionary Biology*, 22(3):599–611, 2009. 10.1111/j.1420-9101.2008.01663.x.

[31] W. G. Draghi J, Parsons T and P. JB. Mutational robustness can facilitate adaptation. *Nature*, in press, 2010.

[32] B. Dujon. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet*, 22(7):375–387, 2006 Jul.

[33] B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. de Montigny, C. Marck, C. Neuveglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. de Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G. Richard, M. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J. Souciet. Genome evolution in yeasts. *Nature*, 430(6995):35–44, JUL 1 2004.

[34] H. Escriva, L. Manzon, J. Youson, and V. Laudet. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol Biol Evol*, 19(9):1440–50, Sep 2002.

[35] A. M. Evangelisti and A. Wagner. Molecular evolution in the yeast transcriptional regulation network. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 302B(4):392–411, 2004. 10.1002/jez.b.20027.

[36] W. J. Ewens. *Mathematical population genetics*, volume v. 27. Springer, New York, 2nd ed edition, 2004.

[37] W. Fontana. Modelling 'evo-devo' with rna. *BioEssays*, 24(12):1164–1177, 2002. 10.1002/bies.10190.

[38] L.-Z. Gao and H. Innan. Very low gene duplication rate in the yeast genome. *Science*, 306(5700):1367–70, Nov 2004.

[39] U. Gerland and T. Hwa. Evolutionary selection between alternative modes of gene regulation. *Proc Natl Acad Sci U S A*, 106(22):8841–6, Jun 2009.

[40] U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-dna interaction. *Proc Natl Acad Sci U S A*, 99(19):12015–20, Sep 2002.

[41] B. Ghosh, R. Karmakar, and I. Bose. Noise characteristics of feed forward loops. *Phys Biol*, 2(1-2):36–45, 2005. 14783967.

[42] G. Gibson and I. Dworkin. Uncovering cryptic genetic variation. *Nat Rev Genet*, 5(9):681–90, 2004. 1471-0056 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review.

[43] N. Gompel, B. Prud'homme, P. J. Wittkopp, V. A. Kassner, and S. B. Carroll. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in drosophila. *Nature*, 433(7025):481–7, Feb 2005.

[44] X. Gu. An evolutionary model for the origin of modularity in a complex gene network. *J Exp Zool B Mol Dev Evol*, 312(2):75–82, Mar 2009.

[45] X. Gu, Z. Zhang, and W. Huang. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A*, 102(3):707–12, Jan 2005.

[46] Y. Guan, M. J. Dunham, and O. G. Troyanskaya. Functional analysis of gene duplications in saccharomyces cerevisiae. *Genetics*, 175(2):933–43, Feb 2007.

[47] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–3, 2002.

[48] B. Guillemette and L. Gaudreau. Reuniting the contrasting functions of h2a.z. *Biochem Cell Biol*, 84(4):528–35, 2006.

[49] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10):e189, 2007.

[50] C. Hittinger, P. Goncalves, J. Sampaio, J. Dover, M. Johnston, and A. Rokas. Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature*, 464(54-60), 2010.

[51] G. Hornung and N. Barkai. Noise propagation and signaling sensitivity in biological networks: A role for positive feedback. *PLoS Comput Biol*, 4(1):e8, 2008.

[52] J. Ihmels, S. Bergmann, M. Gerami-Nejad, I. Yanai, M. McClellan, J. Berman, and N. Barkai. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, 309(5736):938–940, 2005 Aug 5.

[53] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31(4):370–7, Aug 2002.

[54] D. J. Jenkins and D. J. Stekel. A new model for investigating the evolution of transcription control networks. *Artificial Life*, 15(3):259–291, 2009.

[55] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, 2001. 0028-0836 (Print) Journal Article.

[56] S. Jeong, A. Rokas, and S. B. Carroll. Regulation of body pigmentation by the abdominal-b hox protein and its gain and loss in drosophila evolution. *Cell*, 125(7):1387–99, Jun 2006.

[57] S. Kalir, S. Mangan, and U. Alon. A coherent feed-forward loop with a sum input function prolongs flagella expression in escherichia coli. *Molecular Systems Biology*, 2005.

[58] N. Kashtan and U. Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13773–13778, 2005. ISI Document Delivery No.: 969KB.

[59] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban, M.-Y. Hong, K. J. Karczewski, W. Huber, S. M. Weissman, M. B. Gerstein, J. O. Korbel, and M. Snyder. Variation in transcription factor binding among humans. *Science*, page science.1183621, 2010.

[60] Y. Katan-Khaykovich and K. Struhl. Dynamics of global histone acetylation and deacetylation in vivo: rapid restoration of normal histone acetylation status upon removal of activators and repressors. *Genes Dev*, 16(6):743–52, 2002.

[61] S. A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, Oxford, 1993.

[62] M. Kellis, B. W. Birren, and E. S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–24, Apr 2004.

[63] T. Kepler and T. Elston. Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophysical Journal*, 81(6):3116–3136, Dec. 2001.

[64] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719, 1962.

[65] H. Kitano. Biological robustness. *Nat Rev Genet*, 5(11):826–37, 2004.

[66] S. Kuraku, A. Meyer, and S. Kuratani. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? *Mol Biol Evol*, 26(1):47–59, Jan 2009.

[67] F. H. Lam, D. J. Steger, and E. K. O'Shea. Chromatin decouples promoter threshold from dynamic range. *Nature*, 453(7192):246–50, May 2008.

[68] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[69] B. Lehner. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol*, 4, 2008. 10.1038/msb.2008.11 10.1038/msb.2008.11.

[70] B. Lemos, L. O. Araripe, P. Fontanillas, and D. L. Hartl. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *Proc Natl Acad Sci U S A*, 105(38):14471–6, Sep 2008.

[71] S. F. Levy and M. L. Siegal. Network hubs buffer environmental variation in saccharomyces cerevisiae. *PLoS Biology*, 6(11):e264, 2008.

[72] M. Lynch. The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics*, 8(10):803–813, 2007.

[73] V. J. Lynch and G. P. Wagner. Resurrecting the role of transcription factor change in developmental evolution. *Evolution*, 62(9):2131–54, Sep 2008.

[74] T. MacCarthy and A. Bergman. Coevolution of robustness, epistasis, and recombination favors asexual reproduction. *Proc Natl Acad Sci U S A*, 104(31):12801–6, 2007. 0027-8424 (Print) Journal Article.

[75] T. MacCarthy, R. Seymour, and A. Pomiankowski. The evolutionary potential of the drosophila sex determination gene network. *J Theor Biol*, 225(4):461–8, Dec 2003.

[76] S. J. Maerkl and S. R. Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.

[77] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[78] S. Mangan, S. Itzkovitz, A. Zaslaver, and U. Alon. The incoherent feed-forward loop accelerates the response-time of the gal system of escherichia coli. *Journal of Molecular Biology*, 356(5):1073–1081, 2006.

[79] S. Mangan, A. Zaslaver, and U. Alon. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of Molecular Biology*, 334(2):197–204, 2003.

[80] J. Masel. Genetic assimilation can occur in the absence of selection for the assimilating phenotype, suggesting a role for the canalization heuristic. *J Evol Biol*, 17(5):1106–10, 2004. 1010-061X (Print) Journal Article Research Support, U.S. Gov't, P.H.S.

[81] J. Masel. Evolutionary capacitance may be favored by natural selection. *Genetics*, 170(3):1359–71, 2005.

[82] J. Masel and M. L. Siegal. Robustness: mechanisms and consequences. *Trends Genet*, in press, 2009.

[83] S. Maslov and K. Sneppen. Computational architecture of the yeast regulatory network. *Phys Biol*, 2(4):S94–100, Nov 2005.

[84] S. Maslov, K. Sneppen, K. A. Eriksen, and K.-K. Yan. Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol*, 4:9, Mar 2004.

[85] C. D. Meiklejohn and D. L. Hartl. A single mode of canalization. *Trends in Ecology and Evolution*, 17(10):468–473, 2002. Editorial Material 0169-5347.

[86] M. G. Miller and A. D. Johnson. White-opaque switching in candida albicans is controlled by mating-type locus homeodomain proteins and allows efficient mating. *Cell*, 110(3):293–302, 2002 Aug 9.

[87] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, Mar. 2004.

[88] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[89] V. Mustonen, J. Kinney, C. G. Callan, Jr, and M. Lässig. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci U S A*, 105(34):12376–81, Aug 2008.

[90] M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(2 Pt 2):026118, Aug 2001.

[91] S. W. Omholt, E. Plahte, L. Oyehaug, and K. Xiang. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics*, 155(2):969–980, 2000 Jun.

[92] M. Pagel, A. Meade, and D. Scott. Assembly rules for protein networks derived from phylogenetic-statistical analysis of whole genomes. *BMC Evol Biol*, 7 Suppl 1:S16, 2007.

[93] R. Pastor-Satorras, E. Smith, and R. Sole. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222(2):199–210, May 2003.

[94] H. Pham, R. Ferrari, S. J. Cokus, S. K. Kurdistani, and M. Pellegrini. Modeling the regulatory network of histone acetylation in saccharomyces cerevisiae. *Mol Syst Biol*, 3:153, 2007.

[95] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[96] B. Prud'homme, N. Gompel, and S. B. Carroll. Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104:8605–8612, May 2007.

[97] B. Prud'homme, N. Gompel, A. Rokas, V. Kassner, T. Williams, S. Yeh, J. True, and S. Carroll. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, 440(7087):1050–1053, Apr. 2006.

[98] D. Raijman, R. Shamir, and A. Tanay. Evolution and selection in yeast promoters: Analyzing the combined effect of diverse transcription factor binding sites. *PLoS Comput Biol*, 4(1):e7, 2008.

[99] R. M. Raisner and H. D. Madhani. Patterning chromatin: form and function for h2a.z variant nucleosomes. *Curr Opin Genet Dev*, 16(2):119–24, 2006.

[100] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, 2006.

[101] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–26, Oct 2008.

[102] O. Resendis-Antonio, J. A. Freyre-González, R. Menchaca-Méndez, R. M. Gutiérrez-Ríos, A. Martínez-Antonio, C. Avila-Sánchez, and J. Collado-Vides. Modular analysis of the transcriptional regulatory network of e. coli. *Trends Genet*, 21(1):16–20, Jan 2005.

[103] A. S. Ribeiro and S. A. Kauffman. Noisy attractors and ergodic sets in models of gene regulatory networks. *J Theor Biol*, 247(4):743–55, 2007.

[104] A. Rives and T. Galitski. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):1128–1133, Feb. 2003.

[105] J. Ronald, R. B. Brem, J. Whittle, and L. Kruglyak. Local regulatory variation in saccharomyces cerevisiae. *PLoS Genet*, 1(2):e25, 2005 Aug.

[106] N. Rosenfeld, M. B. Elowitz, and U. Alon. Negative autoregulation speeds the response times of transcription networks. *J Mol Biol*, 323(5):785–93, Nov 2002.

[107] D. M. Ruderfer, S. C. Pratt, H. S. Seidel, and L. Kruglyak. Population genomic analysis of outcrossing and recombination in yeast. *Nat Genet*, 38(9):1077–1081, 2006 Sep.

[108] A. Sánchez and J. Kondev. Transcriptional control of noise in gene expression. *Proc Natl Acad Sci U S A*, 105(13):5081–6, Apr 2008.

[109] T. A. Sangster, S. Lindquist, and C. Queitsch. Under cover: causes, effects and implications of hsp90-mediated genetic capacitance. *BioEssays*, 26(4):348–62, 2004. 0265-9247 (Print) Journal Article Review.

[110] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–8, 2002.

[111] I. Shmulevich, S. A. Kauffman, and M. Aldana. Eukaryotic cells are dynamically ordered or critical but not chaotic. *Proc Natl Acad Sci U S A*, 102(38):13439–44, 2005.

[112] M. L. Siegal and A. Bergman. Waddington's canalization revisited: Developmental stability and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10528–10532, 2002.

[113] M. L. Siegal, D. E. Promislow, and A. Bergman. Functional and evolutionary inference in gene networks: does topology matter? *Genetica*, 129(1):83–103, 2007.

[114] V. Sollars, X. Lu, L. Xiao, X. Wang, M. D. Garfinkel, and D. M. Ruden. Evidence for an epigenetic mechanism by which hsp90 acts as a capacitor for morphological evolution. *Nat Genet*, 33(1):70–4, 2003. 1061-4036 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[115] D. Stern. Perspective: Evolutionary developmental biology and the problem of variation. *Evolution*, 54(4):1079–1091, Aug. 2000.

[116] A. J. Stewart, R. M. Seymour, and A. Pomiankowski. Degree dependence in rates of transcription factor evolution explains the unusual structure of transcription networks. *Proceedings of the Royal Society B: Biological Sciences*, 276(1666):2493–2501, 2009. 10.1098/rspb.2009.0210.

[117] S. A. Teichmann and M. M. Babu. Gene regulatory network growth by duplication. *Nat Genet*, 36(5):492–496, 2004 May.

[118] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A*, 98(15):8614–9, Jul 2001.

[119] D. Thieffry, A. M. Huerta, E. Pérez-Rueda, and J. Collado-Vides. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in escherichia coli. *Bioessays*, 20(5):433–40, May 1998.

[120] I. Tirosh and N. Barkai. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol*, 8(4):R50, 2007.

[121] I. Tirosh and N. Barkai. Two strategies for gene regulation by promoter nucleosomes. *Genome Res*, 18(7):1084–91, Jul 2008.

[122] I. Tirosh and N. Barkai. Evolution of gene sequence and gene expression are not correlated in yeast. *Trends Genet*, 24(3):109–113, 2008 Mar.

[123] H. L. True and S. L. Lindquist. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*, 407(6803):477–483, 2000. English Article SEP 28 NATURE.

[124] A. E. Tsong, B. B. Tuch, H. Li, and A. D. Johnson. Evolution of alternative transcriptional circuits with identical logic. *Nature*, 443(7110):415–420, 2006. 0028-0836 10.1038/nature05099 10.1038/nature05099.

[125] B. B. Tuch, D. J. Galgoczy, A. D. Hernday, H. Li, and A. D. Johnson. The evolution of combinatorial gene regulation in fungi. *PLoS Biol*, 6(2):e38, 2008.

[126] B. B. Tuch, H. Li, and A. D. Johnson. Evolution of eukaryotic transcription circuits. *Science*, 319(5871):1797–9, Mar 2008.

[127] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 96(17):9716–9720, 1999. Article 0027-8424.

[128] J. M. G. Vilar, H. Y. Kueh, N. Barkai, and S. Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proc Natl Acad Sci U S A*, 99(9):5988–92, Apr 2002.

[129] G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406(6792):188–92, 2000.

[130] C. Waddington. *The strategy of the genes*. George Allen and Unwin, London, 1957.

[131] A. Wagner. Does evolutionary plasticity evolve? *Evolution*, 50(3):1008–1023, 1996.

[132] A. Wagner. How the global structure of protein interaction networks evolves. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 270(1514):457–466, 2003.

[133] A. Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B-Biological Sciences*, 275(1630):91–100, Jan. 2008.

[134] A. Wagner and J. Wright. Alternative routes and mutational robustness in complex regulatory networks. *Biosystems*, 88(1-2):163–172, 2007.

[135] G. P. Wagner, G. Booth, and H. Bagheri-Chaichian. A population genetic theory of canalization. *Evolution*, 51(2):329–347, 1997.

[136] G. P. Wagner and V. J. Lynch. The gene regulatory logic of transcription factor evolution. *Trends Ecol Evol*, 23(7):377–85, Jul 2008.

[137] M. E. Wall, M. J. Dunlop, and W. S. Hlavacek. Multiple functions of a feed-forward-loop gene circuit. *Journal of Molecular Biology*, 349(3):501–514, 2005.

[138] D. Wang, H.-M. Sung, T.-Y. Wang, C.-J. Huang, P. Yang, T. Chang, Y.-C. Wang, D.-L. Tseng, J.-P. Wu, T.-C. Lee, M.-C. Shih, and W.-H. Li. Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. *Genome Res*, 17(8):1161–1169, 2007 Aug.

[139] J. J. Ward and J. M. Thornton. Evolutionary models for formation of network motifs and modularity in the saccharomyces transcription factor network. *PLoS Computational Biology*, 3(10):e198, 2007.

[140] T. Warnecke, G.-Z. Wang, M. J. Lercher, and L. D. Hurst. Does negative auto-regulation increase gene duplicability? *BMC Evol Biol*, 9:193, 2009.

[141] M. Wernet, E. Mazzoni, A. Celik, D. Duncan, and C. Desplan. Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature*, 440:174–180, 2006.

[142] G. A. Wray. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*, 8(3):206–16, Mar 2007.

[143] D. Yean and J. Gralla. Transcription activation by gc-boxes: evaluation of kinetic and equilibrium contributions. *Nucleic Acids Res*, 24(14):2723–9, 1996.

[144] D. Yean and J. Gralla. Transcription reinitiation rate: a special role for the tata box. *Mol Cell Biol*, 17(7):3809–16, 1997.

[145] D. Yean and J. D. Gralla. Transcription reinitiation rate: a potential role for tata box stabilization of the tfiid:tfiia:dna complex. *Nucleic Acids Res*, 27(3):831–8, 1999.

[146] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):5934–5939, 2004. ISI Document Delivery No.: 814ME.

[147] G. Yvert, R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, R. Mackelprang, and L. Kruglyak. Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, 35(1):57–64, 2003 Sep.

[148] Z. Zhang, J. Gu, and X. Gu. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet*, 20(9):403–407, 2004 Sep.

[149] W. Zheng, H. Zhao, E. Mancera, L. M. Steinmetz, and M. Snyder. Genetic analysis of variation in transcription factor binding in yeast. *Nature*, advance online publication, 2010.

[150] J. Zhu and M. Q. Zhang. Scpd: a promoter database of the yeast saccharomyces cerevisiae. *Bioinformatics*, 15(7):607–611, 1999.