INVESTIGATING THE EVOLUTION OF DIVERSITY AND COMPLEXITY OF PROKARYOTIC GENE REGULATORY NETWORKS USING *IN SILICO* MODELS

by

DAFYD JAMES JENKINS

A thesis submitted to

The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Biosciences

The University of Birmingham

August 2009

Abstract

There is much debate about the evolutionary origins of diversity and many complex gene regulatory network features, such as global regulation. Using novel *in silico* models the evolutionary origins of complex features, heterogeneity and the role of stochastic molecular processes in gene regulatory network evolution are investigated.

It is shown that:

i) Repression is essential, even in constant environments, due to energetic constraints.

ii) Stochastic basal gene expression forces shrinkage of genomes, whilst its absence leads to 'bloating'.

iii) Models evolved towards a biological goal have a very different network structure to non-adaptively evolved models.

iv) Unstable mRNA, stable protein and rapid but robust replication times are strongly selected properties of evolved networks.

v) Multiple network solutions within identical environmental conditions are observed with the presence of stochastic basal gene expression.

vi) Two attractor states, one with high, and one with low stochastic basal gene expression, are observed in networks which can evolve their levels of basal expression.

vii) Functional complexity of a gene regulatory network is dependent on environmental complexity.

viii) Functional complexity evolves in hierarchical stages, requiring 'core' energy regulation mechanisms before environmental responses and adaptations for growth can be sustained and fixed.

ix) Global gene regulation is strongly selected as an efficient energy regulation mechanism.

Acknowledgements

Many thanks to my family and friends for their support during the last three years. A special thanks goes to Helen for help, advice and general support on all matters.

Thanks to my supervisor Dov Stekel whose limitless enthusiasm, help, encouragement and guidance has led to this research. This research would also not have been possible without funding by the BBSRC.

Special mentions go to Patsy, Charles, Jon, Rafik, Peter, Jan, Tony and countless others who helped from the CSB and elsewhere in the university.

Contents

List of Figures

List of Tables

Part I

INTRODUCTION TO THESIS

This part introduces the thesis, its aims and objectives, and relevant background material. Chapter 1 provides background for the thesis and introduces the aims of the thesis, Chapter 2 presents a brief review of the biological processes which make up gene regulatory networks and Chapter 3 presents a review and discussion of existing *in silico* gene regulatory network models.

Chapter 1

INTRODUCTION

Biology is, at its core, the study of systems that evolve by natural selection. One such biological system which receives much interest and research is that of gene regulatory networks. Gene regulatory mechanisms consist of molecular processes which control the expression of genes and, as such, phenotypic responses. However, even in simple model organisms, such as the bacterium *Escherichia coli*, these networks consist of many thousands of genes and tens of thousands of regulatory interactions between them (as shown in the EcoCyc database of *E. coli* genes and interactions [69] or the RegulonDB database [37]). The evolution of such complex gene regulatory networks is subject to a number of forces: physical replication processes, and the inherent randomness associated with these, and adaptation for survival of the fittest. How these forces interact to form the complex networks we observe today is one problem evolutionary biologists and theoreticians have faced since Charles Darwin first proposed the theory of evolution 150 years ago. One such hurdle is the lack of a complete fossil record from ancestor to present-day organisms, complete with genomic DNA.

Without repeating evolution of life on Earth, it is difficult to resolve questions surrounding the evolution of gene regulatory networks. Notwithstanding this, four approaches to experimentally or theoretically 're-run' or analyse evolution have been developed: long-term laboratory evolution; large-scale genomic comparison; synthetic biology; and *in silico* evolution.

Several attempts at long-term laboratory evolution have been very successful. Lenski's 'long-term evolution experiment' (LTEE), which has monitored 20 years of evolutionary history of twelve populations of *E. coli*, has shown that novel metabolic responses to citrate can evolve in a glucose-limited environment [19] and parallel adaptations in gene expression quickly become

fixed in independent populations [28]. Shorter-scale evolutionary work by Palsson has shown that *de novo* mutations in different *E. coli* strains are rapidly fixed in a 44-day timescale providing improved fitness [51]. However, 20 years is short compared with the millions of years that evolution has had to tinker with organisms to obtain the impressive array of diversity and complexity observed in the present day. Moreover, environmental conditions in a laboratory may not be truly representative of conditions found in the wild.

Novel high-throughput sequencing technologies now provide new genome sequences on an almost daily basis [49]. The availability of large numbers of whole genomes allows unprecedented phylogenetic analysis to reconstruct evolutionary history [75, 26] as well as horizontal gene transfer events [128, 78]. However, the number of sequenced genomes is fractionally small compared with true biodiversity: it is estimated that as many as one million distinct genomes exist in only 10g of pristine soil [38], and so an almost impossible amount of sequencing is required to get adequate coverage of even bacterial species. The biodiversity and variety of functions and complexity observed in nature, even in identical conditions, is immense. However, how and why the Darwinian process can result in so much diversity remains an open question.

Large-scale genetics has allowed comprehensive perturbation experiments on whole organisms, from comprehensive gene knockout libraries [7] to 'network rewiring' [57], offering new insights into the properties of the networks, notably network robustness and responses to environmental stimuli. Synthetic biology [6] projects, such as BioBricks and the Registry of Standard Biological Parts [108], also provide a means of network-level experimentation and novel network synthesis. However, currently the size and complexity of the networks constructed in this way are just fractions of even simple model organisms such as *E. coli* [48].

In silico biology offers great potential to explore the consequences of the processes of Darwinian evolution by studying digital "organisms" on timescales not possible in the laboratory. Moreover, *in silico* biology allows us to investigate very specific questions, such as what environmental and evolutionary conditions must be present to facilitate diversity within an environment. All the scenarios described above can be simulated, including, long-term evolutionary experimentation and high-throughput single and multi-gene knockouts, network rewiring and other perturbation experiments. In order for *in silico* approaches to be truly helpful in understanding living systems, it is essential that models capture the important components that evolve, e.g. genes, proteins and regulators, and a suitable level of abstraction is used.

1.1 Aims and objectives

For more than fifty years, the experimental techniques of genetics and molecular biology have been highly successful in developing our understanding of gene regulation, elucidating the functions of regulatory mechanisms, and proposing hypotheses about their evolution. However, an intrinsic limitation of experimental techniques is the impossibility of running evolution under controlled and monitored conditions on the timescales for which observed organisms have evolved. Computational systems modelling biological evolution provide a suitable framework for such controlled and monitored evolution. Therefore, the aim of this thesis is to investigate the evolution of realistic and complex gene regulatory networks using *in silico* models based on prokaryotic biology. To this end, a number of objectives can be defined.

The first objective is to review current *in silico* modelling techniques and discuss their successes and flaws in aiding our understanding of gene regulatory network structure and function, and the evolutionary processes which have shaped them. To do this, a brief review of the biological processes and mechanisms which when grouped together form gene regulatory networks will be presented. The existing computational models will then be introduced and their suitability in realistic evolutionary modelling of gene regulatory networks will be discussed.

The second objective is to develop a number of computational models of gene regulatory networks and evolve these in simplistic environments in order to ascertain their suitability in capturing biological phenomena and behaviours. Determining which biological processes are essential to be modelled in order to generate realistic behaviours will be discussed.

The third objective is to investigate the effects of stochastic molecular processes on the evolution of realistic gene regulatory networks and diversity. Whilst the evolutionary process is necessarily subjected to randomness in order to produce novelty, is the randomness within gene regulatory mechanisms a fixed property of the processes, or can the randomness be tuned or suppressed to the organisms own ends? A specific stochastic effect investigated is that of basal gene expression levels.

The fourth objective is to investigate the evolution of complexity within gene regulatory networks. Complexity is itself difficult to define, and thus leads to ambiguity when attempting to classify complex networks. A concise definition of complexity, in context of the models and networks, is required. Using this definition and the evolutionary computational models

complexity and how it arises can be studied.

1.2 Thesis structure and contributions

This thesis is divided into a number of parts, which can be conceptually separated into two groups. Part I introduces the thesis and gene regulatory networks from biological and computational perspectives and is adapted from the publications, and a corresponding Part V concludes the thesis, summarising the results, their implications and future directions.

Parts II, III and IV present the original contributions, consisting of three novel computational models, detailed analyses of model parameters and in-depth evolutionary investigations. These three parts are based on four publications: Part II is based on a conference proceedings paper, from *ALife XI: The Eleventh International Conference on the Simulation and Synthesis of Living Systems* [60], in which the 'coarse-grained' model is presented in detail and evolutionary simulations in an idealised environment are analysed. Part III is based on a journal article, published in *Artificial Life* [61], in which the 'fine-grained' model is presented in depth with a thorough analysis of model parameters and evolutionary simulations in an idealised environment. Model limitations and future directions are also discussed. Part IV is based on two further journal articles, currently in submission, in which the 'extended coarse-grained' model is described in full and is used in two investigations; an analysis of the effects of stochastic molecular processes on network dynamics and on the evolution of structure and function, and an analysis of environmental complexity and its influence on network structure and function. Model limitations and future directions are also discussed.

As the majority of the work presented in this thesis is taken from the authors publications, a significant contribution was made by the co-author of the papers, and PhD supervisor, Dr Dov Stekel. These contributions range from minor editing of the manuscripts, such as typographical errors and sentence reorganisation, up to some text insertions. However, the ideas, models, simulations and analysis presented are the authors own contributions (unless explicitly noted).

Chapter 2

GENE REGULATORY NETWORKS

Biological cells have many interacting processes governing cell growth and division, stress response, metabolism of food to release energy and transcription and translation. The interaction between these processes and the cells environment produce the complex behaviours we observe. One process in particular, gene regulation, has an enormous impact on a cells ability to respond to changes in environment, such as food availability and starvation, or shock such as heat or acid, by the use of positive and negative feedback.

2.1 Gene regulation

Biological processes are not free, requiring 'energy' to fuel them. The energy can take many forms, such as ATP, nucleotides or amino acids. As such, it is favourable to only use specific processes when necessary. Some proteins are required under many or all conditions, and so their production may be less strongly regulated. However, other proteins may only be required in specific conditions, such as shock, meaning that much stronger and complex regulation is required. Transcription can be regulated in a number of ways, one of which is via transcription factors (TF). A gene may need to be 'turned on' (activated) or 'turned off' (repressed) by one or more TFs to affect when it is transcribed. TFs bind to specific sequences on the DNA, which act as regulatory sites for the associated genes, either helping the RNA polymerase (RNAP), the protein which transcribes the gene, to bind to the promoter site in the case of activator TFs,

or blocking the promoter site preventing RNAP binding.

As gene regulation consists mainly of gene products interacting with other genes, the genes and their interactions can be visualised as a network, with nodes (genes) and edges (regulatory interactions). Networks of gene regulation for responding to the environment can be both simple and complex. Many of the particularly well studied networks, at both experimental and theoretical levels, are in the model bacterium *E. coli*. These include the *lac* operon, which enables response to glucose or lactose in the environment [59, 129], the *trp* operon which controls production of the amino acid tryptophan using a repressor [3, 105] and the heat shock system [34, 77].

2.1.1 Transcription and translation

Transcription and translation are the two main processes involved in the production of protein from a gene. Transcription involves a protein, known as RNA polymerase (RNAP), binding to the DNA at a specific place, known as a promoter site. Once the RNAP has bound to the promoter site for a gene, transcription initiates causing the DNA helix to unwind immediately in front of the RNAP. The RNAP molecule then, using one of the strands of DNA as a template, produces a molecule of messenger RNA (mRNA). This mRNA transcript is then translated into one or more identical proteins by ribosomes. This thesis considers only prokaryotic biology, and as such uses the simple processes involved in prokaryotic transcription and translation, and does not include more complicated processes found in eukaryotes such as splicing.

2.1.2 Post-translational processes

Regulation of genes and protein is not limited just to the level of DNA. A multitude of post-translational processes, such as passive and active degradation, and phosphorylation can affect the function of a protein, up to effectively switching it off, or changing its function completely, further complicating the already complex gene regulation mechanism.

2.1.3 Stochasticity within biological processes

Experimental work on gene expression in single cells has demonstrated that this is a stochastic process [88, 35, 62, 21]. This further complicates gene regulation, as the level of gene expression will fluctuate, including random or basal rates of expression, and therefore protein expression

will also fluctuate widely. Molecular movement is mostly subject to 'random walks', and as such interactions between molecules will also be random. Therefore, even under identical environmental conditions, two cells could have vastly different levels of protein, with different molecular interactions taking place. How much evolution has 'tuned' or can control the level of stochasticity in the processes is unknown.

2.2 Analysis of gene regulatory networks

In discussing biological systems Lynch [81] suggests that there are five popular inter-connected properties of such systems which receive much attention:

1. Complexity, that can be defined in physical, structural or functional ways [1], typically used to describe a system with many parts.
2. Evolvability, the efficiency of a lineage in discovering beneficial mutants [30].
3. Modularity, a structural measure [95] splitting a network into 'related communities' [112, 73].
4. Redundancy, "occurs when the same function is performed by identical elements" [119].
5. Robustness, the insensitivity to mutations affecting a given phenotype [121].

Several of these properties have been applied to network theory in many different fields, and as such many techniques are readily available for gene regulatory network analysis.

2.2.1 Genetics approaches

Laboratory-based experimentation for analysing gene regulatory networks has been available for many decades. Experimentation on specific genes within the network can be performed using knock-outs and DNA mutations to analyse robustness and determine function or redundancy, whilst evolvability can be investigated using knock-ins to simulate additions to the network and its stability and fitness. The development of high-throughput 'omics' technologies has further increased the potential of laboratory-based experimentation and the wealth of knowledge to be obtained through these techniques. However, a reasonable level of precision is not always available in many of these techniques, and the vast amounts of data obtained generates a further problem - how to analyse and make sense of it all, requiring development of new bioinformatic

techniques. Moreover, a more fundamental problem of gene regulatory network analysis is still mostly unsolved by the high-throughput technologies - how to observe the evolution of such networks on a suitable timescale? A number of very successful laboratory-based evolutionary projects have revealed some very interesting results as to the evolution of gene regulatory networks (several stand-out examples are Richard Lenski's 'long-term evolution experiment', providing over 20 years of monitored evolution [19, 28, 100], Bernhard Palsson's 44-day experiment, during which fixation of mutations leading to increased fitness were observed [51] and Thomas Ferenci's 26-day experiment, in which five phenotypic combinations, with multiple variations in global regulation and other mechanisms, were observed from a uniform population in a constant environment [84]). However, the timescales available to laboratory experimentation are not sufficient to a allow a re-running of evolution and therefore address major questions, such as the adaptive or non-adaptive evolution of complex gene regulatory network mechanisms.

2.2.2 Computational and mathematical approaches

Many different computational or mathematical techniques have been applied to gene regulatory network analysis. Phylogenetic analysis, using tools such as BLAST [4], allows comparison between organisms to generate phylogenies and species relationships. This can also help determine possible function of newly sequenced genes and proteins, meaning laboratory experimentation is not always necessary. A number of mathematical analyses, using elements of network theory, can be applied to networks to elucidate different properties of the networks. Modularity can be investigated using a multitude of different measures [95], and network motif analysis reveals over-represented sub-graphs, which have been thought to represent complex 'building blocks' from which larger networks are built upon [90]. Moreover, computational approaches can be used to simulate experiments not possible in the laboratory.

2.2.3 *In silico* modelling and genetics

A different computational and mathematical approach is to model and simulate the biological processes. Whilst such an approach necessitates an abstraction of the actual biology, it often provides the only feasible approach to address certain questions. Moreover, the availability of large-scale computational power has allowed the development of realistic and quantitative *in silico* models of many biological systems [97]. One particular field which has benefited hugely

from computational and mathematical modelling is that of evolutionary biology. As previously discussed, it is unfeasible to perform evolutionary experiments in a laboratory over a sufficiently long period of time, whereas computational models allow the simulation of much longer periods of evolution in a practical timescale. This is because the lifespan of a simulated individual can take a fraction of the time when compared to its real-life equivalent. Further, computational models allow selection of mutants and phenotypes at a much more specific level than laboratory experiments, as much more detail about specific pathways, interactions, genotypes and behaviours can be more easily obtained. Indeed, the use of '*in silico* genetics' provides more than just a way of selecting or viewing molecules and organisms, it provides a new and powerful tool for investigating a model molecular system - allowing knock-outs to be instantly generated, accurate and specific mutations to be applied to any molecule or any kinetic rate to be modified.

2.3 Evolution of gene regulatory networks

The biological networks we observe are the result of millions of years of evolution, and are still constantly subject to the same evolutionary processes. Therefore, we are only capable of observing snapshots of evolutionary history. Further, the lack of a complete 'fossil record' with genomic data hinders our knowledge as to how such networks have evolved. As such, whilst the physical processes governing gene regulatory network evolution, such as gene duplication, genetic divergence and horizontal gene transfer are known, how these have interacted with adaptive selection is an open question and open to debate.

There are two extreme views which are held by researchers on the evolution of gene regulatory network structure:

1. Gene regulatory network structure is completely adaptive, that is every aspect of the network has been shaped solely on its function.
2. Gene regulatory network structure is completely non-adaptive, that is the structure of the networks is simply due to the physical processes, and complex network features are merely an evolutionary by-product.

Clearly, a sliding scale of views are also present, with the network evolving by a combination of adaptive and non-adaptive processes.

Recent evolutionary work using *E. coli* has shown the *de novo* evolution of global regulatory

networks governing DNA superhelicity and stringent response in multiple, independent populations, indicating adaptive selection [100]. Further key innovations such as the evolution of a population capable of using citrate as a food source in a glucose-limited environment indicate adaptive selection for such functionality [19]. One aspect debated is that of network motif abundance. Proponents of network motifs claim that the over-abundance of several motifs, such as the Feed-Forward Loop (FFL), is evidence of adaptive selection [90]. This argument is further strengthened by research indicating that FFLs have specific function with GRNs and are therefore adaptively selected for on this functionality [85, 31, 63] and specific motifs are more conserved in organisms sharing similar lifestyles [9]. Yet, evidence also suggests that network motif structure does not determine function [87, 56, 89]. Moreover, the over-abundance of specific motifs is based on a flawed argument using random graphs [90, 107]. When the realistic replication process of GRNs is taken into account the over-abundance of network motifs is easily explained as a result of non-adaptive processes [117, 8, 12, 22, 29]. Another aspect are global regulators [42, 86] and 'scale-free' network properties [13]. Cases and de Lorenzo describe a process of non-adaptive genome evolution that will result in the occurrence of global regulators through purely non-adaptive gene duplications [22]. Lynch takes this argument further [81, 82] showing that "many of the qualitative features of known transcriptional networks can arise readily through the non-adaptive processes of genetic drift, mutation and recombination".

Using *in silico* evolutionary models of gene regulatory networks it is possible to help shed some light on this issue, by effectively 're-running' evolution and obtaining a complete fossil record, from ancestor organisms onwards, which can be analysed not only using bioinformatic static analysis tools, but also with *in silico* genetics, allowing experimentation on any 'digital organism', past or present, in real-time.

Chapter 3

IN SILICO GENE REGULATORY NETWORKS

A number of different approaches and models for gene regulatory network modelling have been proposed over recent decades. Their success has proved they are a useful tool for studying evolutionary biology. Much work has focused on explaining network architectures that have been observed in biological networks, including both "global" architectures such as scale-free networks [13], and "local" structures such as network motifs [90].

The Artificial Genome (AG) model introduces binding template matching using nucleotide-like genetic sequences, allowing investigation into the role of so-called junk DNA [102]. Evolution of the AG model under non-adaptive conditions indicates a typical scale-free network topology is observed [101], similar to the frequently discussed scale-free topology of many biological networks [2]. The Artificial Regulatory Network (ARN) model, also using a binding template matching mechanism on sequences, has explored the impact of binding strength on network topology [12]. Evolution of the ARN, also under non-adaptive conditions, using the processes of duplication and divergence, resulted in not only a scale-free topology, but also the over-representation of network motifs [76, 79]. Work by Cordero and Hogeweg also show the emergence of feed-forward loops (FFLs), an over-represented network motif, as a non-adaptive by-product of the evolutionary process [29], in contrast with the adaptive, functional view of FFLs discussed by Mangan and Alon [85].

While these models show the potential of *in silico* models to capture the inherent structure

of biological networks, they are typically evolved without function or fitness, and so represent non-adaptive evolution. In biology, however, the networks we observe are selected for specific and complex functions. A number of evolutionary models use biologically realistic functions, such as gene expression profiles in development [110], biomass production [99], or chemotaxis to a given stimulus [41], and realistic network structures and dynamics are frequently observed. However, the functions and fitness chosen as evolutionary objectives in many models are often unbiological, such as mathematical or logic functions [127, 80, 36, 65]. Whilst certain biological network components, such as the *lac* operon or FFL motif, could, theoretically, act as Boolean logic gates [67], or perform other functions in electrical circuits [85], the networks have not evolved to act as 'adder' or 'cube-root' logic functions, but to metabolise food, survive stress, grow and replicate. Moreover, energetic costs associated with gene expression or replication are only rarely included [124, 113]. How the structure of the *in silico* networks will vary when evolved using more realistic fitness functions, such as growth and replication, compared to networks evolved to perform unrealistic specific logic functions, is not known. An important question therefore arises: what will be the impact of incorporating biologically realistic function, fitness and energy constraints into the dynamics of evolution of gene regulatory network models on the structure of the final networks?

Another approach to modelling evolution of cells often used within the Artificial Life literature, is Individual-based Models (IbM). In the IbM approach, each model is treated as 'an individual' or 'agent' each with its own set of specific components. These 'individuals' then interact and compete with each other within an environment for resources, much like any biological organism. A number of biologically focused IbMs, such as BacSim [72], COSMIC and COSMIC-Rules [44, 46] which aim to evolve bacterial function from the genetic level (transcription networks), up to the environment level (population dynamics) have met with relative success [71, 45], highlighting the importance of biological, ecological and environmental realism in models.

Other work has explored dynamical simulations of gene regulatory networks, and a number of different paradigms have been developed. The Boolean network model introduced by Kauffman in the late 1960s [66] is a deterministic paradigm that has been used for several decades, and was one of the first models developed specifically for modelling gene regulatory networks. The model consists of a number of genes and the regulatory interactions between them, forming a

graph or network. The networks are simulated using discrete time-steps, during which each gene is either on or off, depending on the state of its input genes in the previous time-step, and a Boolean logic function specific to each gene. Whilst this is a highly abstract model, it has been used to model the dynamics of many simple regulatory networks, including the *lac* and *ara* operons in *E. coli* and regulation of *bacteriophage lambda* genes [67]. Yet, this model has also been shown to be inappropriate for many systems [47]. Many other models have been developed, that are simulated using deterministic dynamics, including both Boolean networks and ordinary differential equations (ODEs), as they are very quick and simple to generate and simulate.

One of the most successful 'digital organisms' for evolutionary research to date is Avida [96]. Avida has a genomic structure that consists of CPU instructions and a number of registers simulating a 'metabolism'. Simulations using this system have revealed that complex features evolve by building on simpler features [80], much like the evolutionary chemotaxis mechanism discussed by Goldstein and Soyer [41], and in high-mutation environments genotypes which led to a more robust replication rate are favoured [127]. However, Avida, whilst clearly inspired by biological processes and mechanisms, abstracts them to an unbiological degree. The 'genome' consists of CPU instructions, which interact using Logic functions such as XOR, NAND and NOT. Whilst some biological interactions can be approximated to Logic functions, for example, the *lac* operon, biological systems process information through molecular mechanisms.

François and Hakim also proposed a simple *in silico* evolutionary procedure to evolve small, functional modules, such as bi-stable switches and oscillator circuits, which demonstrate the role of post-transcriptional interactions and also show similar structures and network principles observed in biological networks [36]. Kashtan *et al.* showed that using varying goals during the evolution of networks will speed up the evolution of efficient solutions [65]. However, deterministic formulations have one major draw-back: biological gene regulatory networks are not deterministic.

This leads to stochastic dynamics as a more realistic simulation paradigm that is able to describe the inherent randomness in processes such as transcription, translation and binding events. A number of stochastic formulations have been described, such as the Gillespie algorithm for simulating coupled chemical reactions [40], which can be used to simulate gene regulatory networks. This has been particularly successful in developing theoretical understanding of the intrinsic and extrinsic noise in gene regulation [35], and the extent and control of noise has been

explored for a number of simple networks, including constitutively expressed, self-activating and self-repressing genes [68, 116, 111, 114, 61, 60]. However, despite the success and realism of this paradigm, it too has a major draw-back: the computational simulation time is much larger than the deterministic paradigm, and so these models can only be efficiently simulated for relatively small networks.

3.1 Discussions

There are a number of very successful and realistic evolutionary *in silico* gene regulatory network models, many of which have been previously introduced. However, many models within the literature can be separated into two broad classes, which I term the 'Artificial Life approach', in which a more computational or engineering approach is taken to produce a large-scale, cellular model, often resulting in models with limited and highly abstracted representations of the biological processes, and the 'biological approach' in which a very detailed biologically-realistic model is produced of a specific system, which in turn is less flexible in modelling of generic systems and mechanisms. The main difference between these two approaches is therefore the level of detail employed within the models. Both approaches have their merits and provide valuable research platforms for molecular, systems and evolutionary biology, but a combination of the two approaches could provide an ideal research platform. Several models do provide a combination of the two approaches and required careful consideration of both computational and biological constraints, as in for instance, the Avida platform, and experiments using this system have yielded some very interesting insights into the evolution of complex systems. However, the underlying model is still largely computational in nature, with a 'genome' of CPU-like instructions.

In this thesis I will present a number of *in silico* models which also attempt to combine the two approaches. The aim of these models will be to accurately model biological processes to evolve realistic gene regulatory network mechanisms. These models will then ultimately be used to investigate the evolution of functional complexity within gene regulatory networks.

Part II

COARSE-GRAINED MODEL

This part contains the work on the 'coarse-grained' model. Chapter 4 introduces the model and describes it in detail, Chapter 5 introduces the results of evolutionary simulations and Chapter 6 presents a discussion of the model and results. These chapters are formed from a conference proceedings paper titled "Effects of Signalling on the Evolution of Gene Regulatory Networks" [60], and presented at *ALife XI: The Eleventh International Conference on the Simulation and Synthesis of Living Systems* in August 2008.

Chapter 4

'COARSE-GRAINED' MODEL INTRODUCTION AND DESCRIPTION

4.1 Model introduction

The 'coarse-grained' model was designed to be able to model prokaryotic gene regulation with the goal of evolving biologically realistic networks, and their dynamics and function, whilst having a minimal computational complexity footprint. The basis for this model was an existing evolutionary gene regulatory model [29]. A simple simulation paradigm was applied to the model, adding dynamics and function to the previously static networks. The model used abstracted implementations of biological processes associated with gene regulation, such as transcription, translation and protein-DNA interaction, maintaining a high-level of biological realism and complexity. As such, realistic and complex gene regulation mechanisms are able to evolve. Energetic costs are applied to transcriptional and translational processes, reflecting the energetic costs to biological cells. This is an important concept that is often overlooked in many gene regulatory network models, yet, energy regulation is a very important challenge to cells. The effects of stochasticity at the single cell level have been shown experimentally [88, 35, 62, 21], yet are also often neglected in computational systems. Therefore, one important aspect of the model design was the incorporation of stochasticity into all processes. Fitness of a model is

measured as the speed of growth and is analogous to the goal of biological organisms, survival and growth. Evolution to a biologically-realistic goal, rather than an artificial goal, is a fundamental requirement of the model, often ignored in computational models. Simulation is also considerably abstracted, using a stochastic formulation of the discrete Boolean network [66] model which has a very low computational complexity, providing a fast simulation of large networks. Consequently, simulated evolutionary timescales can be very large, providing an ideal framework for modelling gene regulatory network evolution.

4.2 Investigation aims

We investigate the effects of dynamics in the evolution of transcription network structure using models with and without energy signalling. The use of energy signals in biological regulatory networks is well studied. The transcriptional regulator complex CRP-cAMP is one of *E. coli*'s global regulators, known to regulate several hundred genes as listed in the EcoCyc database [69]. The large number of positive interactions by CRP-cAMP in biosynthesis pathways indicates that energy signals are used for growth by cells [130, 50]. A subunit of the CRP-cAMP complex, cAMP, is a signalling molecule derived from ATP; ATP concentration indicates 'energy' within the cell. When the concentrations of CRP and cAMP reach sufficient levels, the activated transcription factor complex forms. Whilst CRP-cAMP is a dual-regulator (it both activates and represses different genes), 142 of the 173 known and predicted interactions in the EcoCyc database are identified as activating interactions.

Organisms without energy signalling are also prevalent in nature. *Buchnera aphidicola* is a bacterium related to *E. coli*, having a common ancestor diverging 250 million years ago [91, 109]. *B. aphidicola* has a different lifestyle to *E. coli*; it has evolved an endosymbiotic relation with aphids, while *E. coli* exists as a free-living bacterium. *B. aphidicola* cells live in an environment of sufficient food, which is simpler than many other bacterial environments. *B. aphidicola* strains have lost most of their genome and regulatory network, retaining around 600 genes, representing a subset of *E. coli* genomes [109, 126]. This lack of regulation allows the over production of several amino acids, which are excreted and subsequently used by the aphid. The lack of an 'energy signal' observed in *B. aphidicola* is due to the absence of *crp* and *cycA*, the genes responsible for the CRP-cAMP transcription factor [109].

We introduce a model that evolves networks using realistic evolutionary operators and is simulated with simple inputs and output to determine fitness. Our model introduces regulation type for binding sites, new evolutionary operators, signalling mechanisms as inputs and biosynthesis as output. We simulate the networks using a stochastic Boolean network paradigm, representing simplified transcriptional network dynamics. The results of these evolutions are presented and analysed, and relevance to biological systems is discussed. Graph theoretic approaches are used to compare the adaptive evolution and networks that have evolved non-adaptively over the same time period, highlighting the effects of the adaptive evolution.

4.3 Model definition

The model has three distinct components: 1) network generation and static architecture 2) network simulation and 3) evolutionary framework.

4.4 Network generation

To generate the gene regulatory network, we use the model introduced by [122] and extended by [29]. This model produces a network with realistic connectivity and structure of specific protein-DNA binding interactions when evolved without a fitness function, 'non-adaptive evolution'. A genome initially consists of I regulatory genes, g, where each gene has a regulatory region with between 0 and J binding sites, bs, (with the exception of 'input genes') and a protein, p. Each binding site and protein has a specific shape, S, represented by an integer drawn from a discrete circular space $\{0, 1, 2, ..., S_{max} - 1\}$ (with $S_{max} - 1$ adjacent to 0). Each binding site, bs_j, can be either activating, $r_{ij} = 1$, or inhibitory, $r_{ij} = -1$, and an occupancy value, o_{ij}, which is 1 if bound and otherwise. Each gene therefore consists of a vector of binding site regulatory type, r, and a vector of binding site occupancy, o. Gene activation, a_i, uses the binding site regulatory type and occupancy vectors, and is defined as:

$$a_i = \sum_{j \in J} r_{ij} o_{ij} \tag{4.1}$$

where i is gene, i is binding site, r_{ij} is binding site regulation type and o_{ij} is binding site occupancy. As 'input' genes do not contain a regulatory region, their activation is determined

by specific equations (see later).

4.4.1 Protein-DNA interaction

The binding strength, B_{ij}, between two shapes, S_i and S_j is defined as:

$$B_{ij} = \begin{cases} 1/(D_{ij}+1) & if\ D_{ij} \leq D_{max} \\ 0 & otherwise \end{cases} \tag{4.2}$$

where D_{ij} is the shortest integer distance between the shape of the protein, S_i, and the binding site, S_j. A binding distance, D_{max}, is defined as the maximum distance between two shapes that will interact. A matrix, M, is created where M_{ij} is the strength of binding B between protein i and binding site j.

To determine whether a binding event occurs, the binding strength, B_{ij}, and gene activation, a_i, are used in the following equation:

$$o_{kj} = (B_{ij}a_i) > R \tag{4.3}$$

where o_{kj} is the occupancy of binding site j of gene k, B_{ij} is the binding strength between the protein of gene i and the jth binding site of gene k, a_i is the activation state of gene i and R is a random number between 0 (inclusive) and 1 (exclusive). Figure 4.1 shows an example network and interactions.

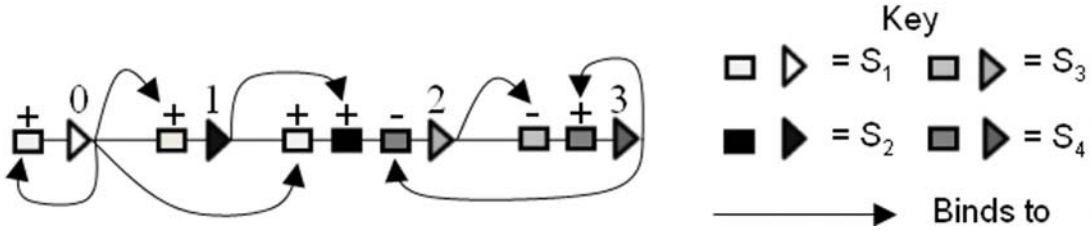

Figure 4.1: An example 4-gene network showing protein-DNA interactions. Genes 0,1 and 2 form a type-1 coherent Feed-Forward Loop (FFL). Additionally, gene 0 has an activating self-regulating connection. A fourth gene in the circuit acts as an AND gate in the FFL, by negatively regulating gene 2. If gene 3 is transcribed, it negatively regulates the FFL, and causes the FFL to be an AND gate. If gene 3 is not present, then then the FFL will be OR gate.

4.4.2 Specialised genes

In addition to the regulatory genes in the original models, we introduce three new types of genes, which are divided into two classes, 'input' and 'output':

- **Energy signal genes**: these 'input' genes have a protein product, but no regulatory region. The expression status is based on the amount of energy within the cell. Energy in the model abstractly represents the ATP, amino acids and other molecules a biological cell requires to grow, transcribe mRNA molecules and translate them into protein molecules and other processes.
- **Food signal genes**: these 'input' genes represent the food available to the cell and are used as the input into the model when it is simulated. they have a protein product, but no regulatory region. The energy level of the model increases whenever a *food signal gene* is activated. Each *food signal gene* has an energy value associated with it, which is the amount of energy added to the model when the gene is activated.
- **Biomass pathway genes**: these 'output' genes have a regulatory region, and generate *biomass* when expressed. They are used as the output of the model when it is simulated, represent cell growth, and have both an energy consumption (amount of energy used when gene is activated) and biomass production (amount of biomass added when activated) value associated with them.

4.5 Network simulation

In order to further investigate the structure of the networks evolved using realistic evolutionary operators, we introduce a simulation system for examining the dynamics of the networks. We use a Boolean network model [66] to simulate the dynamics of the network over a number of discrete time-steps. Stochasticity is added to the simulation with random, basal levels of transcription and probabilistic binding events. At each time-step a number of sub-steps takes place in order:

1. Determine 'input' gene status: Energy signal genes (ON if energy threshold is exceeded, OFF otherwise) and food signal genes(ON if food available this time-step, OFF otherwise).
2. Determine protein-DNA interactions (o_{kj}) for all ON 'input' genes ($a_i > 0$).
3. Determine gene activation status of non-input genes (a_i).

4. Determine protein-DNA interactions (o_{kj}) for all ON non-input genes ($a_i > 0$).
5. Update energy and biomass levels.
6. All bound binding sites unbind ($o_{ij} = 0$), all genes deactivate ($a_i = 0$) and all proteins removed.
7. Check model has energy remaining - if the energy level is ≤ 0 then the model 'dies' due to lack of energy, and simulation terminates.

where ON = 1 and OFF = 0. All genes are OFF initially.

4.5.1 Molecular production costs

Transcription and translation are not free processes: energy is used whenever they take place. To approximate energy consumption within the model, the energetic cost of the transcription and translation events are only applied after all possible binding events have occurred. The energy level is decreased by the number of occupied binding sites within the genome representing the cost of producing those proteins whose expression is under control of the bound sites.

Biomass production also requires energy. Whenever a *biomass gene* is activated, the energy level decreases by the gene's 'energy consumption' value, and the biomass level increases by the gene's 'biomass production' value.

4.5.2 Deterministic simulation

The simulation can be turned into a deterministic Boolean network, by replacing the DNA-protein interaction steps (2,4) with a binding threshold:

$$o_{kj} = (B'_{ij} \times a_i) \tag{4.4}$$

$$B'_{ij} = \begin{cases} 1 & if\ B_{ij} \geq T_{bind} \\ 0 & otherwise \end{cases} \tag{4.5}$$

Basal transcription, K_{basal}, is also set to 0, meaning that a gene must be bound by an activator to transcribe.

4.6 Evolution framework

The evolution framework used in the model is a standard genetic algorithm, with a fixed population size, and a purely elitist strategy that emulates the spatial constraints on a bacterial population, in, for example, a chemostat, where the fittest cells are ones that replicate fastest. A daughter cell is generated at each generation representing a simplified bacterial asexual replication.

4.6.1 Non-adaptive evolution and evolutionary operators

Once the network has been initialised, it is 'non-adaptively' evolved for a given number of steps by randomly selecting a gene from the genome and applying a mutation operator. This aims to generate a more natural gene regulatory network architecture. These same mutation operators are used during the 'adaptive' evolution. Cordero and Hogeweg define six mutation operators which operate at either the gene or binding site level:

1. gene duplication: the entire gene (protein product and regulatory region) is copied and added to the genome, producing an exact replica of the original gene
2. gene loss: the entire gene is removed from the genome
3. protein mutation: the protein shape is changed
4. binding site duplication: a binding site from another gene is randomly copied into the regulatory region
5. binding site loss: the binding site is removed from the regulatory region
6. binding site mutation: the binding site shape is changed.

Shape mutation (protein and binding site) in the Cordero and Hogeweg model consists of either incrementing or decrementing the shape, S, by 1 with equal probability. We use a more realistic mutation operator allowing the shape to make larger jumps around the shape space, using the integer part of a normal random variable with $\mu = 0$ and $\sigma = log_{10}S_{max}$.

We define two new evolutionary operators:

7. binding site regulation 'flip': the binding site 'flips' its regulation type from positive to negative or vice versa. A possible biological example of this process could be transposition

of an activating binding site through insertion or deletion of DNA, therefore potentially blocking binding if the relocated binding site interacts with the promoter site.

8. horizontal gene transfer (HGT): a portion of another genome is horizontally transferred and copied into the genome (corresponding to DNA-uptake or plasmid transfer). This operator is applied at the genome level only.

Figure 4.2 gives a diagrammatic representation of the evolutionary operators.

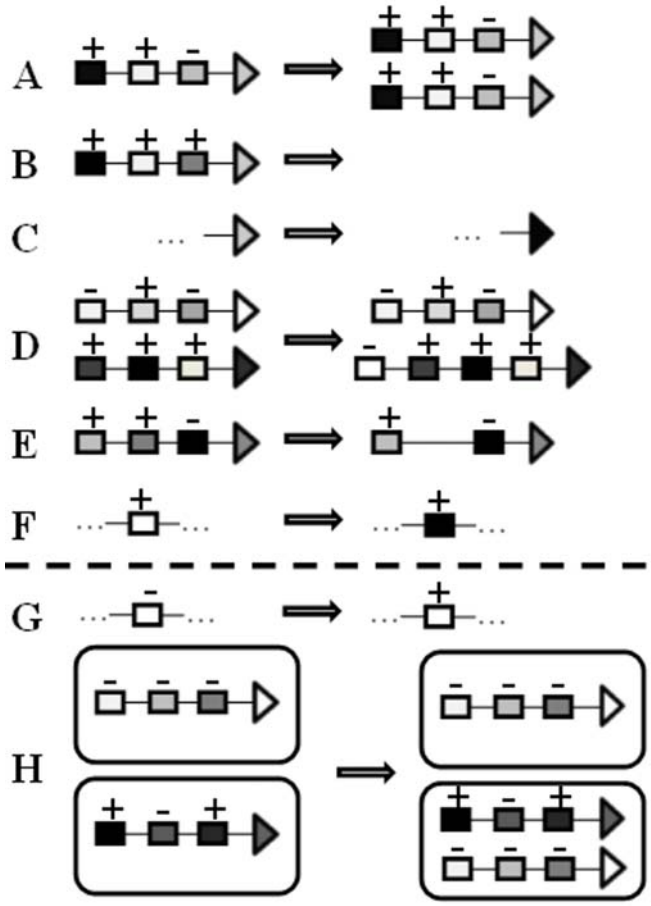

Figure 4.2: Evolutionary operators. Rectangles represent binding sites (+ activating; - repressing), triangles represent gene/protein product. Shape is represented by greyscale colour. Original operators defined in [29] are shown in parts A - F. A) gene duplication, B) gene loss, C) protein mutation, D) binding site duplication, E) binding site loss and F) binding site mutation. The two evolutionary operators introduced in this work are, G) binding site regulation 'flip' and H) horizontal gene transfer. Operators A - G apply to each gene or regulatory region within a genome, whereas H only applies to the whole genome level.

Due to the specific function of the *energy*, *food* and *biomass* genes, not all evolutionary operators are applied to them. The evolutionary operators applied to *energy signal genes* and *food signal genes* are 1) gene duplication (however, the duplicated gene loses the specialised

functionality of the original) and 3) protein mutation. The evolutionary operators applied to *biomass pathway genes* are 1) gene duplication (functionality not duplicated), 3) protein mutation, 4) binding site duplication, 5) binding site loss, 6) binding site mutation and 7) binding site regulation 'flip'.

4.6.2 Replication and fitness

Due to the nature of DNA replication both the daughter and parent cells are subject to possible mutation. During replication, each gene in the genome can be affected by one of the evolutionary operators (#1-7). HGT (#8) is applied after genome replication and mutation. If HGT takes place, a donor genome from the population is selected at random, and a randomly selected number of genes are copied from the donor genome.

Fitness of an individual model is based solely on the level of biomass production after the defined number of time-steps. If the simulation terminates due to lack of energy, the model has died and has a fitness of -1. If not enough 'living' models exist at the end of a generation, new random models are added to the population. In the non-adaptively evolved populations, the fitness function is a random number between 0 and 1, implying no selection pressure.

Model lineages are defined as a group of models with a common ancestor and are determined after evolution.

4.7 Model parameters

All model and evolutionary parameters are given in Table 4.1.

Parameter	Value	Note
S_{max}	128	
D_{max}	3	
Starting genome size	32, 256	
Max. starting binding sites/gene	3	
Initial mutations	2000	
Gene duplication	1×10^{-3}	Value taken from [29]
Gene loss	1×10^{-3}	Value taken from [29]
Protein mutation	5×10^{-3}	Value taken from [29]
Binding site duplication	8×10^{-3}	Value taken from [29]
Binding site loss	8×10^{-3}	Value taken from [29]
Binding site mutation	8×10^{-4}	Value taken from [29]
Binding site 'flip'	8×10^{-4}	
Horizontal gene transfer	5×10^{-5}	
Max. genes horizontally transferred	10	
Basal transcription rate, K_{basal}	1×10^{-2}	
Binding threshold, T_{bind}	0.5	
Population size	1000	
Generations	100	
Simulation time steps	1000	
Starting energy	500	
Energy signal gene threshold	250	
Food gene energy generated	5	
Biomass gene energy consumed	50	
Biomass gene biomass produced	50	
Biomass genes in genome	2	

Table 4.1: Model and evolution parameters

Chapter 5

EVOLUTIONARY SIMULATIONS

5.1 Model and environment regimes

In a simple environment, where the model has a constant supply of food, we evolved four types of models: 1) Energy signal gene present in a small genome, 2) Energy signal gene present in a large genome, 3) Energy signal gene not present in a small genome, 4) Energy signal gene not present in a large genome.

5.1.1 Small genome with energy signal

With an energy signal gene and a small genome, a final population evolves with a very simple regulatory network (Table 5.1). The main component of this network is a strong positive regulation of one of the *biomass genes* from the *energy signal gene*, but also has some residual connectivity between regulatory genes (Figure 5.1A). However, no regulation (positive or negative) due to the input food genes was evolved. This is to be expected, as the environment remains constant, and so provides no useful information to be exploited. This regulation network is a simple, but effective system; whenever the model has sufficient energy, the energy signal is present, and it strongly activates the biosynthesis pathway gene; when the energy drops below this level activation of the biosynthesis pathway ceases. Only one of the biomass genes is activated, so whilst the system may not be maximally efficient at generating biomass, the model is far less likely to over-express genes, in particular the energy-expensive biomass genes, and so is far more likely to survive to the end of the simulation. This network also allows a far more robust regulation of biosynthesis, as the energy signal gene is not affected by noise. The use of an energy signal

for activating growth parallels many organisms such as *E. coli*, with its use of CRP-cAMP.

	Regulator type	Population A	N	R
ES	Activator	6.17	43.45	47.41
	Repressor	1.66	51.64	47.70
	Dual	0.05	3.48	1.75
No ES	Activator	31.85	43.08	45.92
	Repressor	17.27	43.58	46.05
	Dual	1.73	3.02	1.77

Table 5.1: Mean number of different connection types per model in energy signal (ES) and no energy signal (NoES) populations. A is adaptive, N is non-adaptive, and R is random population. Dual are regulators that function as both activators and repressors either on the same gene or over multiple genes

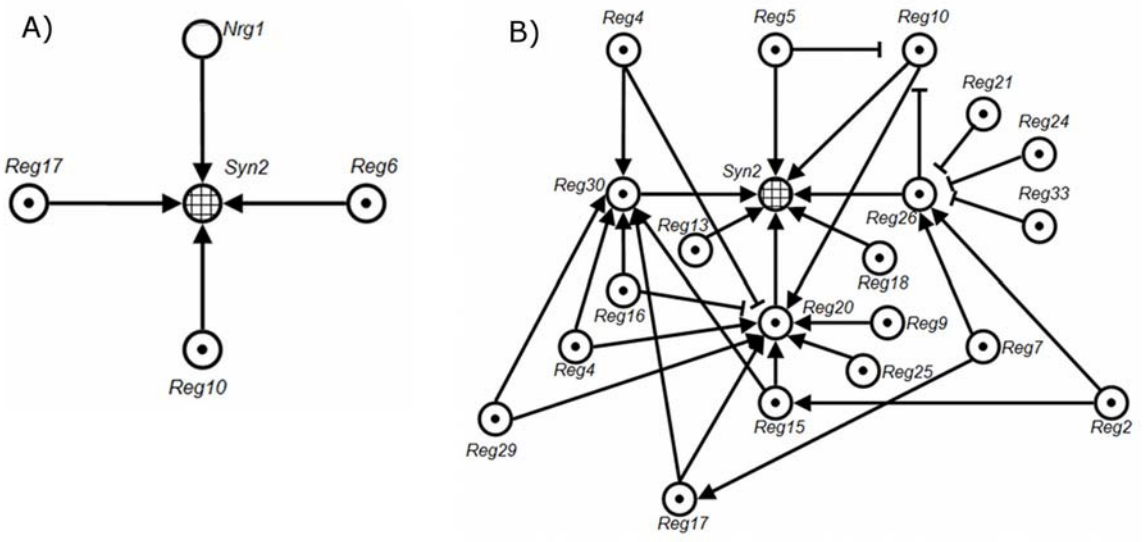

Figure 5.1: Example networks evolved adaptively with and without energy signal in a small genome. A) energy signal network solution, with activation of the biosynthesis pathway by the energy signal, and several other regulatory genes. B) no energy signal network solution, with activation from many regulatory genes, which may also be activated by other regulatory genes forming regulatory cascades.

5.1.2 Small genome without energy signal

With no energy signal gene and a small genome, a very different regulation network is evolved. In this population, the most successful models again consisted of no regulation due to the input food genes, and so no input stimuli at all were available (as the energy threshold gene is regulated by the model itself, it can be classed as an input). Thus the models rely solely on stochasticity for transcription and translation of random genes (Figure 5.1B). The model did

however evolve some positive regulation from a small number of standard genes to the biomass genes (Table 5.1); this increases the probability that the biomass genes will be activated at a given time-step, and so the efficiency of generating biomass. Whilst this network is not as efficient at generating biomass, or robust to noise due to the reliance on stochasticity, it is well adapted for survival. An alternative strategy, not possible within the model, would be a fixed constant low level of basal expression. The lack of energy signalling used in the evolution of this population of models shares several parallels with the lack of signalling in *B. aphidicola* cells, and a similar, simple regulatory network in observed in both. The exploitation of stochastic gene expression seems to be a robust sub-optimal solution for survival without environmental information. Modification of the basal expression rate (such as that assumed in *B. aphidicola*) may allow a simpler, more efficient network, without the need for an energy-sensing mechanism. This solution may also provide a mechanism for survival in early gene regulatory networks, until more precise signalling networks evolve, or could itself be the basis for a signalling network. Figure 5.2 shows that although the number of regulatory interactions rapidly decreases in both populations, the populations without an energy signal lose connections slower. However, the fittest individuals in the populations without an energy signal are within 20% of the biomass production of the populations with an energy signal (5500 with energy signal, 4400 without), indicating that the solution reliant on stochasticity can still be quite efficient.

5.1.3 Large genome

With a much larger genome, with or without an energy signal, we observe very different results. Network connectivity is necessarily high because of the number of genes and small shape space. Models are unable to survive because they very quickly over-express many genes and use up all energy. Even under more energetically favourable conditions (energy from food = 40; starting energy = 4000) the models are still unable to survive. As the evolutionary framework removes 'dying' models from the population, they are unable to utilise the evolutionary operators to remove genes and connections. The rate of mutation is also therefore very important. This indicates the importance of repressors within biological networks to tightly regulate the processes of transcription and translation, as are not 'free' (they require energy sources e.g. ATP), even as a temporary measure until the gene is lost. Other computational models have also obtained the evolution of repressor systems, even in constant environments, due to energetic constraints [61].

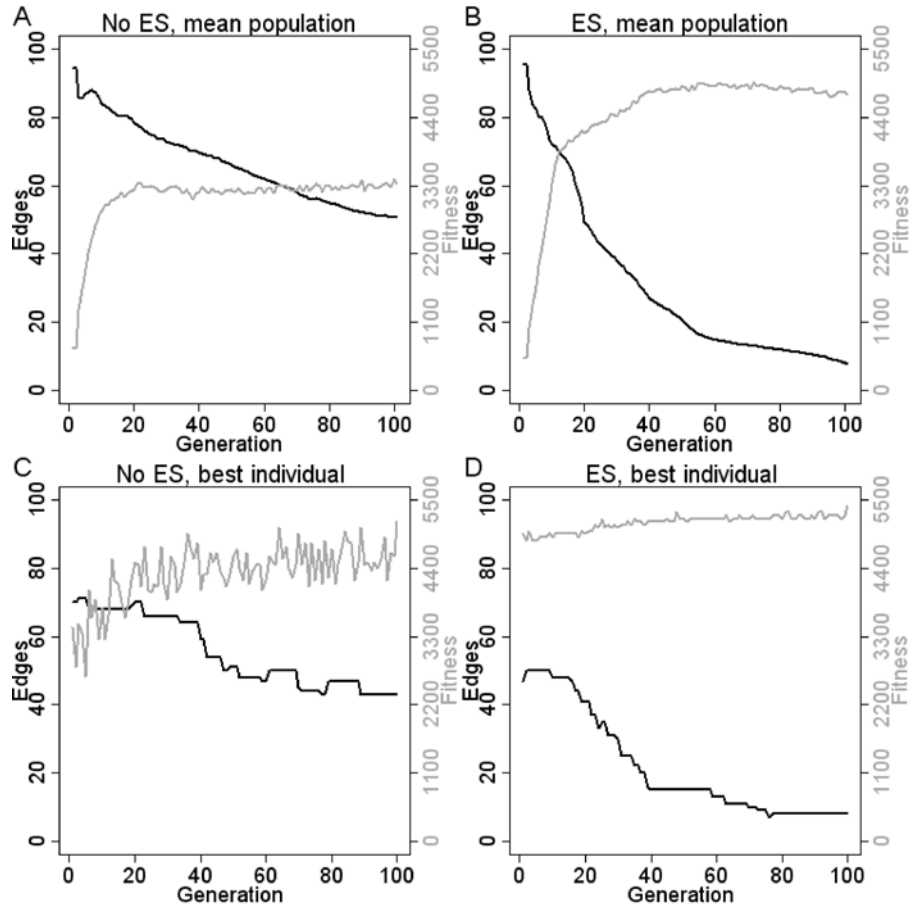

Figure 5.2: Evolutionary history of mean population fitness and number of regulatory connections in a small genome. A) mean fitness and number of regulatory connections in the population with no energy signal. B) mean fitness and number of regulatory connections in the population with an energy signal. C) best individual in the final population without an energy signal. D) best individual in the final population with an energy signal. The decrease in network connectivity and increased fitness can be seen in all plots. Lack of convergence between 'best' and mean fitness may be due to the short evolutionary time-scale.

The regulatory network of *E. coli* displays a preference for negative regulation by transcription factors in many different systems [69]. This may indicate a further use of negative regulation as an adaption for efficiency, as well as enabling large scale switching of regulatory systems, fast responses and maintaining homeostasis. Whilst removal of activating connections may also be an important process to increase efficiency, the removal of activators does not prevent basal expression, and thus does not remove the need for repressors. Indeed, strong negative self-regulation has been shown to decrease the amount of mRNA needed to express a protein at a set level, thus reducing the use of energy expensive processes as shown in [114]. One possible explanation for the lack of large global repressors evolving in the current implementation of

the model is the energy cost of maintaining sufficient numbers of repressor proteins. Protein stability is fixed to one time-step, so proteins must be produced each time-step, using up large amounts of energy. In biological systems, protein stabilities ranging from minutes to many hours are observed [93]. The stability of a protein is often associated with function: signalling proteins are typically short-lived; metabolic proteins are often more stable. Modifying the model to allow proteins to evolve their stability may allow the evolution of global regulators. In addition, real biological molecules have a large shape space, due to the very high dimensionality of protein shape. Increasing the shape space in the model could help alleviate the high network connectivity.

5.2 Effects of stochasticity

Removing stochasticity dramatically alters the networks evolved. Whilst a similar regulation mechanism is observed in populations with an energy signal, the number of connections does not rapidly decrease. Network connectivity remains high, with the exception of input genes. This occurs as the regulatory genes will only be transcribed if activated by another gene, leading to large parts of the network which are highly intra-connected with no external inputs. There is no pressure to reduce this connectivity, provided no input genes connect into the large highly connected parts. The high connectivity may appear to indicate a complex solution, however, the increased connectivity may merely mask the underlying core functionality of the model. The similar functional network topologies and fitness values, of around 5500, observed between the deterministically simulated populations (Figure 5.3) and the stochastically simulated populations (Figure 5.2) indicates that the core solutions evolved in either population are in fact the same.

Populations without an energy signal were unable to produce any surviving models, indicating that the exploited stochasticity of the original solution is indeed essential.

5.3 Non-adaptive evolution and comparison

To compare the effects of adaptive evolution to a biological fitness function, we evolved populations under the same four conditions but used a random fitness function to simulate 'non-adaptive' evolution. We examined the model networks at two points during the evolution: 1)

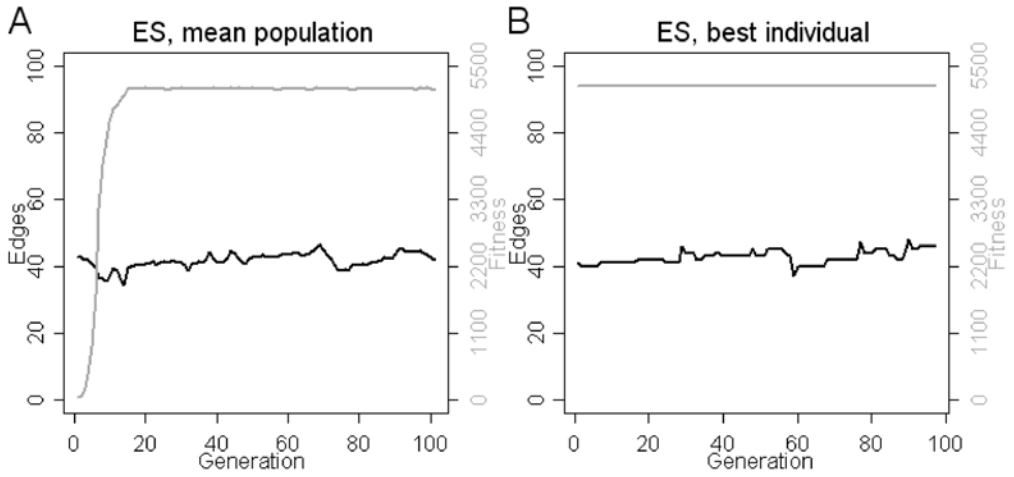

Figure 5.3: Evolutionary history of deterministically simulated small genome population with energy signal. A) mean fitness and number of regulatory connections in the population. B) best individual in the final generation of the population. Network connectivity remains high in both plots, unlike in the stochastic simulation populations.

after random model initialisation (R) and 3) after a given number of generations (N).

Several network properties are extracted from the networks: binding site distribution, binding site regulation type ratio, gene 'out' degree (number of genes the transcription factor interacts with), gene 'in' degree (number of transcription factors which regulate the gene) and number and type of self-regulating connections.

5.3.1 Binding sites

A general trend for loss of binding sites can be see in Figure 5.4. In the adaptively evolution populations, a larger number of genes in both populations have no binding sites, and have a much smaller distribution of maximum binding sites per gene. This shows how the model has evolved its regulatory network, by reducing it. There was no significant bias to binding site regulation type in each population (Table 5.2), however, a clear trend for activating connections in the evolved populations is shown in Table 5.1. This may be linked to the lack of the evolution of global repressors as discussed above. Without a global regulatory mechanism, the model is unable to effectively regulate the expression of the genes, and so the alternative solution is to reduce the probability of transcription factor activity by losing binding sites. Whilst this solution does not prevent transcription, it does reduce it. In fact this mechanism is exploited in the populations without an energy signal.

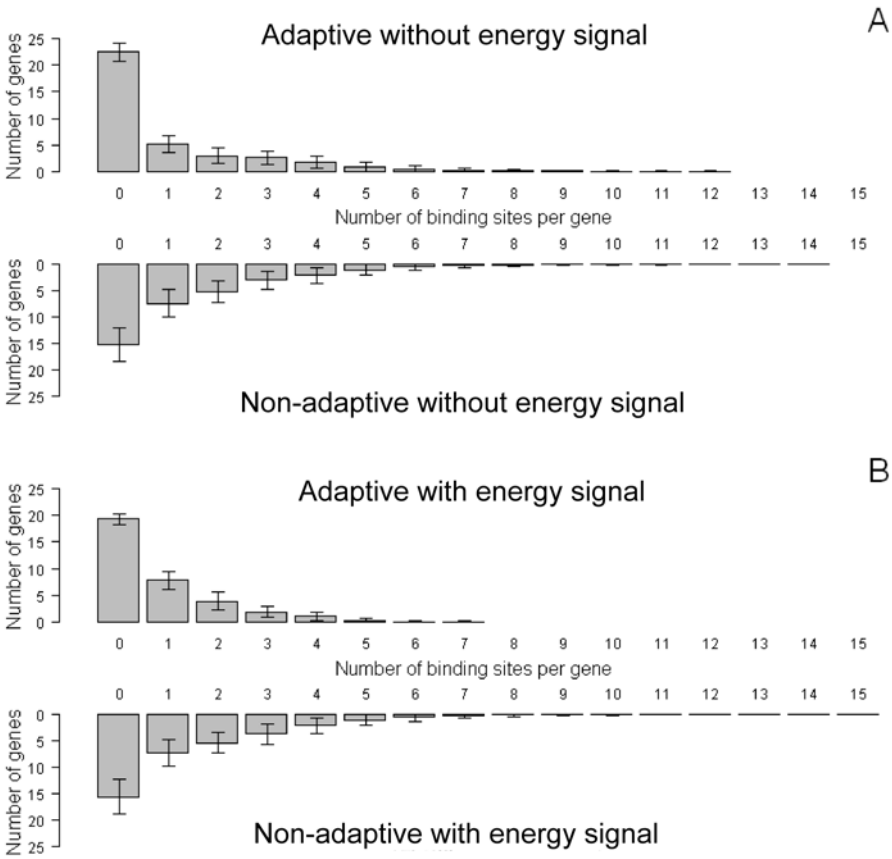

Figure 5.4: Mean number of binding site for each gene in each model in small genome populations. Error bars are 1 s.d. of the population. A) adaptively evolved and non-adaptive populations without energy signal, B) adaptively evolved and non-adaptive populations with energy signal. The smaller number of binding sites in the adaptive populations can be seen in both panels; the adaptively evolved populations have a larger number of genes without any binding sites, and have a lower maximum number of binding sites per gene. It can also be seen that there is little difference between the number of activating and repressing binding sites for each population.

5.3.2 Transcription factor activity

The loss of large amounts of regulation can be seen in the interactions between transcription factors (TFs) and genes. This is indicated by the 'out' degree for each transcription factor (Figure 5.5), and the 'in' degree for each gene (Figure 5.6). We observe an increase in the number of proteins that do not act as TFs and the number of genes which are not regulated by any TFs when evolved with the biological fitness function. The maximum number of genes regulated by a TF is also significantly reduced in the adaptive populations, in particular the population with an energy signal. The maximum number of TF's regulating a gene is also significantly reduced in the adaptive populations.

	Binding site type	Population A	N	R
ES	Activator	12.38	24.70	25.54
	Repressor	13.69	26.56	25.64
NoES	Activator	17.06	24.37	25.42
	Repressor	18.55	25.27	25.51

Table 5.2: Mean number of binding site regulation types per model in energy signal (ES) and no energy signal (NoES) populations. A is adaptive, N is non-adaptive, and R is random population

The number of self-regulating genes were separated into: activating only, repressing only, and dual regulation. Again, a clear trend can be observed from the adaptive populations from Table 5.3. The two adaptively evolved populations have lost nearly all of their self-activating connections, and a large proportion of their self-repressing connections. Whilst more activating connections in total are conserved (Table 5.1), a larger number of negatively self-regulating connections are conserved, indicating the importance of negative self-regulation in transcription networks.

	Regulator type	Population A	N	R
ES	Activator	0.0005	1.0070	1.3205
	Repressor	0.2830	1.4440	1.3395
	Dual	0	0.1060	0.0520
No ES	Activator	0.0165	1.0535	1.3295
	Repressor	0.4295	1.1645	1.3060
	Dual	0	0.0785	0.0435

Table 5.3: Mean number of activating, repressing and dual-regulating self-regulating connections per model within the energy signal (ES) and no energy signal (NoES) populations. The loss of connectivity can be seen in the evolved populations. The adaptively evolved populations show a significantly smaller number of activating and repressing and no dual interactions compared with the non-adaptive and random populations. A is adaptive, N is non-adaptive and R is random population. Dual are regulators that function as both activators and repressors either on the same gene or over multiple genes

These results indicate the loss of interaction within the network, and highlight that complex regulatory networks are unnecessary to survive within a stable environment. The preference for preserving self-repressing connections and overall reduced network connectivity indicates that the network regulates its energy usage by preventing transcription of unrequired genes.

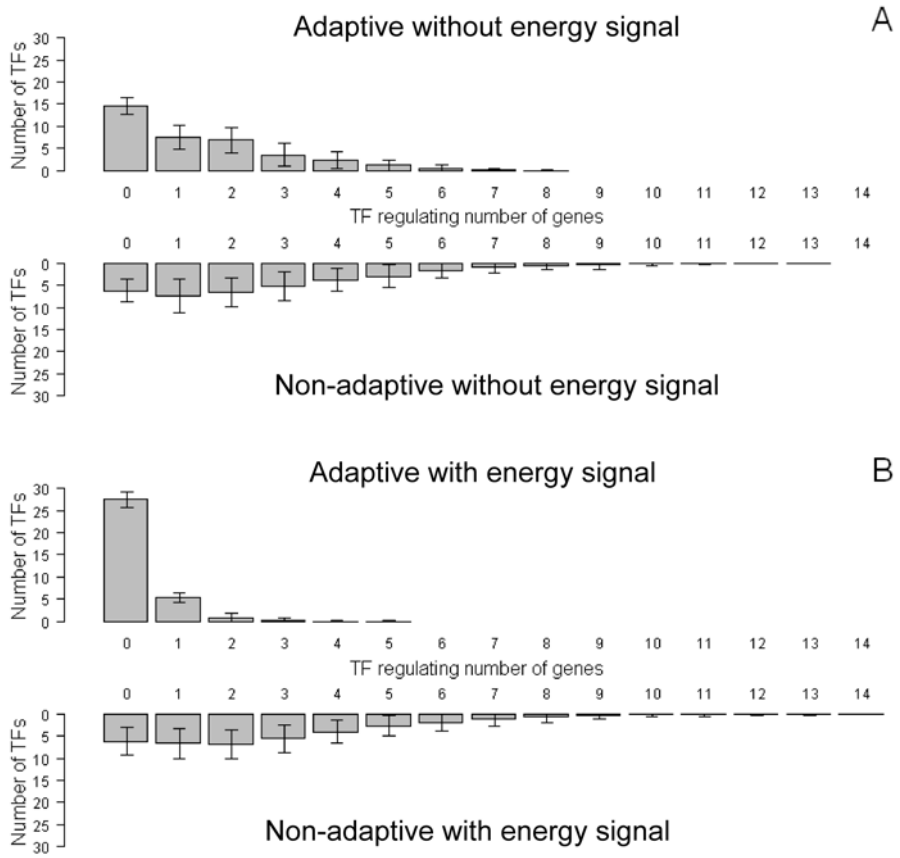

Figure 5.5: Mean gene 'out' degrees for each gene in each model in small genome population. Error bars are 1 s.d. of the population. A) adaptively evolved and non-adaptive populations without energy signal, B) adaptively evolved and non-adaptive populations with energy signal. The small amount of connectivity in the adaptively evolved populations is indicated by a large number of genes which do not act as transcription factors and a smaller number of 'global' transcription factors.

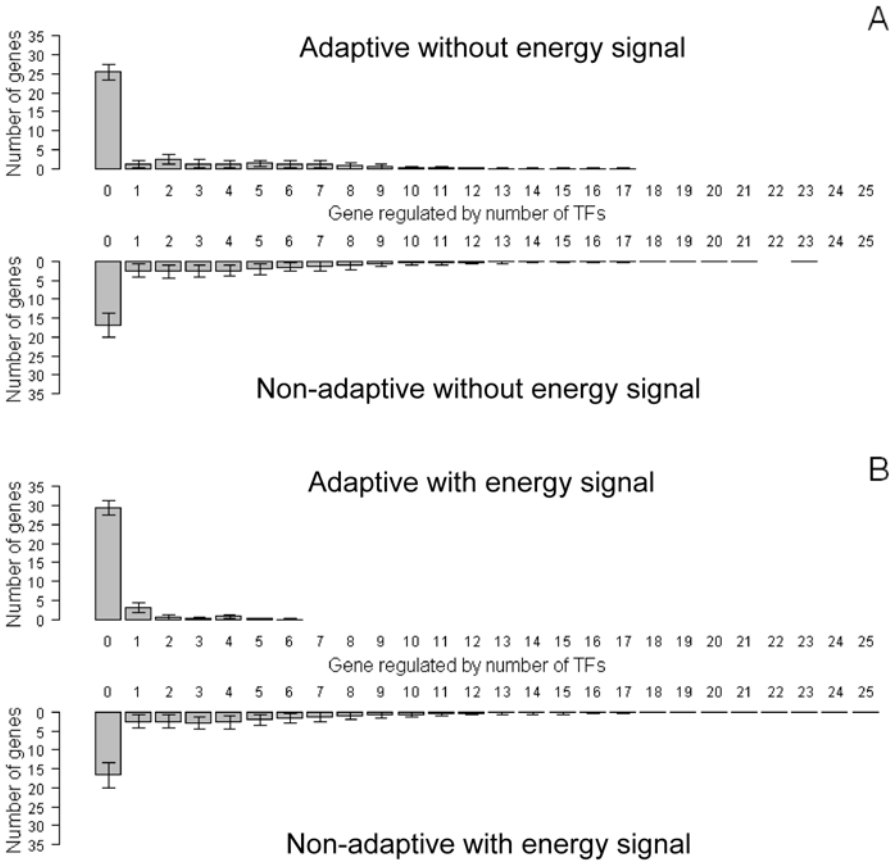

Figure 5.6: Mean gene 'in' degrees for each gene in each model in small genome populations. Error bars are 1 s.d. of the population. A) adaptively evolved and non-adaptive populations without energy signal, B) adaptively evolved and non-adaptive populations with energy signal. The smaller amount of connectivity in the evolved populations is indicated by a large number of genes without any transcription factor interaction and a smaller distribution of interactions.

Chapter 6

MODEL DISCUSSIONS

The results obtained from the evolutions described in the previous chapter have shown two very different, but realistic regulation mechanisms have been selected and evolved. When no energy signal gene is present in the genome, the population has evolved to exploit the stochasticity within the transcription and translation processes. Whilst the biomass genes are seemingly not activated by the food inputs, they have evolved a large number of activating connections from many other genes. This strategy allows the model to exploit the stochastic gene expression, potentially tuning the number of activating connections to ensure that enough genes will randomly activate the biosynthesis pathways, and ensuring that these pathways are not over-expressed.

The other regulatory mechanism evolved, whilst being less complex, is one that is observed in many biological regulatory systems. The energy signal is used as input for the biosynthesis pathways, and regulation of other genes is much more tightly controlled through loss of connectivity.

These results highlight the potential complexity of regulation networks even in the most simple of environments. They also show the mechanisms which natural selection, and the evolutionary operators it uses, have discovered and optimised in both the model networks presented here and the real biological systems. Further, the results indicate the ability of the model to capture biological phenomena and potential for large-scale realistic evolutionary modelling.

6.1 Summary and conclusions

This work expanded an existing model for genome evolution and added a simulation method, developed from Boolean network models. Models are evolved in populations with and without energy signalling genes, and the adaptively evolved models are compared with models evolved non-adaptively, and random models.

Results from the evolutions indicate a decrease in the number of regulatory connections within the networks, and a preference towards preserving activating and negative self-regulatory interactions. A number of parallels are drawn between the evolved models and biological systems, including: regulation by the global regulator CRP-cAMP in *E. coli*; a regulation mechanism similar to the endosymbiont *B. aphidicola*; the use of negative regulation as a mechanism for efficiency; and the need for differing protein stabilities dependent on function.

The model, whilst capable of reproducing biological behaviours and mechanisms, has several flaws. The protein production mechanism is inappropriate, as proteins are synthesised 'on-demand' when a potential DNA-binding interaction takes place. Additionally, proteins need to be synthesised each time-step, assuming a fast degradation rate. Whilst some protein will have a rapid turnover, many are very stable. Protein turnover is therefore an important biological mechanism which is not accurately modelled. Further, protein-protein interaction, an important mechanism in gene regulation and other systems such as signalling, is not modelled.

Despite the limitations of the model, it is able to capture several important mechanisms. The evolution of negative regulation of genes to regulate energy usage, and exploitation of basal gene expression, indicates the importance of including 'energy' and stochasticity within a gene regulation network model. The model is very efficient computationally, and several thousand generations can be simulated in a matter of hours.

Part III

FINE-GRAINED MODEL

This part contains the work on the 'fine-grained' model. Chapter 7 introduces and describes the model, simulation paradigm and evolutionary environment in detail, Chapter 8 presents an analysis of random model dynamics over a range of parameters, Chapter 9 presents the results of evolutionary simulations and Chapter 10 presents a discussion of the model, its limitations and future directions. These chapters are formed from a journal article titled "A new model for investigating the evolution of transcription control networks" and published in *Artificial Life* [61].

Chapter 7

'FINE-GRAINED' MODEL INTRODUCTION AND DESCRIPTION

7.1 Model introduction

The 'fine-grained' model was designed to model prokaryotic gene regulation to a high-level of detail not typically used in biological and evolutionary *in silico* modelling, including the previously introduced 'coarse-grained' model. Additionally, the model attempts to incorporate a large number of cellular processes, as well as transcriptional regulation processes, such as metabolism and polymerisation. Therefore, the model represents a more complete 'cell' to a high-level of detail, and as such produces very biological behaviours and network structures. Molecular interactions and binding affinities are determined using a 2D shape-space and an affinity model using Euclidean distance. Allosteric or conformational changes caused by molecule binding are implemented, allowing simulation of cell-signalling pathways and transfer of information. Protein function is not predetermined, as in many models, but is based on binding affinity and complex stability, for instance to metabolise food, and thus is free to evolve. Additionally, protein and mRNA stability are free to evolve. Transcriptional activity is determined by 'transcriptional logic', a system using the occupancy of binding sites to modify binding affinity of RNA polymerase molecules to promoter sites. Energetic costs are again applied to transcriptional and

translational processes, reflecting the energetic costs to real cells. Model fitness, in contrast with many models, is simply survival and replication, much like any biological organism. This fitness function is a more biologically realistic goal, than evolution to, for instance, specific logic circuits, and therefore promotes evolution of more realistic network structures and regulation mechanisms. Models are simulated using an exact stochastic simulation algorithm, providing an exact trajectory of the molecular system, providing a realistic simulation environment. Evolution is simulated using a genetic algorithm producing a 'generational' structure of models, in which only the most fit models are replicated using biological mutational operators.

7.2 Investigation aims

The model presented in this study aims not only to more accurately model a transcription regulatory network, its processes and components (biological approach), but also the encapsulating cell and associated functions and systems within it (Artificial Life approach). The sole objective of this cell is that of all organisms: to survive in its environment and propagate. Whilst this single objective approach may seem to contrast with the arguments presented by Kashtan and Alon, that multiple, changing objectives during evolution result in more efficient solutions in a smaller time-scale [64], the objective is in fact a complex combination of many smaller, possibly conflicting, objectives. Unlike many other models presented previously, the model is simulated stochastically, as previous studies have shown the stochastic nature of intrinsic and extrinsic noise found within any biological system [88, 62].

This study presents the new model and methods in depth, along with results of comprehensive analysis of the model over parameter ranges and the behaviours observed. An introduction to the evolution methodology is also presented, supported by an analysis of the resultant evolution in an idealistic environment.

The study highlights the power and importance of using '*in silico* genetics' tools to investigate models and analyse their behaviour, and we make hypotheses about the model's behaviour in more complex environments.

7.3 Model definition

The model we present in this study is a novel transcription regulation network and cellular model for evolving bacteria within a range of environments. Like other models such as COSMIC, AG and ARN, our model can be viewed on a number of different levels: (i) molecular (ii) interaction networks (iii) cellular and population.

Each level provides different challenges which must be solved through evolution and natural selection.

7.3.1 Molecular level

At the lowest level, the model consists purely of molecules. Molecules can be divided into two types, '*mobile molecules*' which are molecules which can move freely within the cell cytoplasm, such as proteins, and '*DNA-based molecules*' which are portions of the DNA which perform specific functions, such as gene regulation.

Mobile molecules

In a single cell there are thousands of different types of molecules, ranging from individual ions, to sugars to larger macromolecules such as proteins [3]. Our model substantially reduces the types of molecules into five broad classes:

1. Protein - proteins are the 'workhorses' of the model, as they can potentially perform a number of functions: transcription factor, metabolic enzyme or signalling. Proteins are not assigned any function, instead the binding affinity with other molecules determines their functions.
2. RNA polymerase (RNAP) - this is a protein that performs the specific function of initiating transcription when bound to a gene promoter site, and transcribes the gene forming a molecule of messenger RNA. The level of RNA polymerase is determined at the start of simulation, and no more can be created, nor can any be degraded (it is assumed that this intrinsic machinery would be managed elsewhere by the model). This is the only protein with a prespecified function.
3. mRNA - messenger RNA molecules act as 'templates' for proteins.

4. Energy - energy is the global term used for any molecule which is used up to perform or fuel a function (such as in transcription or translation), and is analogous to ATP. Energy is used to determine cell state. The model has the capacity to include further types of energy that could be used in specific reactions.
5. Food - food provides energy to the model cell. Food molecules are broken down by a protein binding to it. Each food type has a number of parameters:
   (a) Time to be broken down
   (b) Molecule type yielded (either a different type of food, or energy)
   (c) Amount of molecules yielded.

Whilst the model abstracts an actual cell considerably, it still has an enormous and varying amount of complexity. For instance, a pathway such as the glycolytic cycle could be modelled completely, introducing numerous types of food as each individual metabolite is included also requiring many different protein enzymes; alternatively, a single food type could represent the entire pathway.

DNA-based molecules

In prokaryotic cells, the DNA typically has the following four types of region; *encoding gene* which contains the genetic information used to produce mRNA molecules, *cis-activating elements* and *cis-repressing elements* which when occupied by a transcription factor up regulate or down regulate transcription of its associated gene, and *promoter elements* which are used by RNA polymerase molecules as an indicator for the beginning of an encoding gene which can then be transcribed. Our model implements these types of regions by assigning a *regulatory region* consisting of a number of *cis-activating, cis-repressing* and *promoter elements* to an *encoding gene*, and also associated to the gene is an *mRNA* and *protein*. The encoding gene itself does not have a representation other than the transcribed product.

7.3.2 Molecule shape and binding domains

Each molecule within the model has a specific shape, which is used to determine binding affinity with other molecules. Molecule 'shape' is represented by a number of *binding domains/sites*, therefore, the number of binding sites a molecule has determines the number of molecules to

which it can bind at any time, and also determines dynamically what functions it could perform. The shape of real molecules depends on atomic and charge configuration, which would require a very high dimensional space to be accurately represented. In our model, we represent the shape of a binding domain with just two dimensions so that the shape is modelled by a point on the surface of a unit sphere. The two spherical polar coordinates (θ, ϕ) corresponding to the point on the sphere are the 'genetic' information of the binding domain, and thus are free to mutate. The polar coordinates transformed into the Cartesian coordinate system (x, y, z) are then used in the function to determine binding affinity with another shape, and so are the corresponding 'phenotype'.

Binding affinity

The binding affinity between two binding sites is a function of the Euclidean distance metric between one site and the antipode of the other site (denoted as Δ). This metric is the simplest calculation that adequately describes the required distance relationship between two shapes. In this way the strongest binding would be from two complementary, opposite shapes. Because association is diffusion limited, different binding strengths are implemented as dissociation rates, which are given by Equation 7.1:

$$K_{off} = \frac{\sigma\Delta}{1 - (\frac{\Delta}{2r})^{\alpha}} \tag{7.1}$$

where σ is a scaling factor, r is the radius of the sphere (in this case 1) and α is a Hill-like coefficient for modifying the affinity curve saturation.

This binding affinity function is used to calculate the stability of all complexes. An exception to this is the RNAP-promoter complex. Our current model implementation uses a fixed complex dissociation rate, which is dependent only on the occupancy of the associated activator and repressor sites, and not on the shape of the promoter or RNAP molecule. This is to ensure that regardless of mutation to the promoter site, the RNAP is still able to function.

Allosteric effects

In the cases where a molecule has multiple binding domains, it is possible for it to be bound to several other molecules simultaneously. The effect of a binding domain being occupied has been shown to potentially cause conformational changes to other domains of the molecule [3].

Our model introduces such a concept, such that each binding domain has two shapes: 'natural shape' in which the domain exists when the parent molecule is a monomer; and 'allosteric shape' in which the domain exists when another domain of the parent molecule is part of a larger, multi-molecule complex.

7.3.3 Molecular processes

Molecule interaction

Molecules are assumed to exist in a well-stirred system. This means that all molecules will have the same interaction rate. The diffusion-limited interaction rate of protein-protein molecule interactions is slower than DNA-protein molecule interactions [70, 27], and is reflected in the model.

Polymerisation

Polymerisation between molecules to form large complexes is an integral component of many cellular processes, such as in signalling networks, increasing molecular stability, or the formation of physical structures in a cell, such as the actin cytoskeleton or a flagellum. Our model allows polymer chains to dynamically form and break. This allows signalling mechanisms and transfer of information, and prevents protein and mRNA molecules from being degraded. Due to computational constraints, complexes are only permitted to consist of up to three molecules. Because we do not model physical structures this constraint does not negatively affect the model.

Metabolism

Metabolism is a core function of any cell. Metabolic pathways within the model are any reactions involving a food molecule and any protein. Pathways can be implemented with various levels of realism. For instance glycolysis could be included in a model by adding each metabolite from the cycle, each with its own catabolism time and product, or a single food could be used to represent the entire pathway.

Degradation

Molecule degradation can occur actively or passively. Active degradation involves another molecule binding to the molecule and changing its structure causing it to degrade, whereas passive degradation does not require any interaction from other molecules, breaking down either spontaneously, or due to environmental conditions. The model currently only implements passive degradation, and so molecules will break down spontaneously based on a 'stability' or rate. Only those that are produced by genes within the model (i.e. generic proteins and mRNAs) degrade; other molecules are treated as stable.

Transcription and translation

Transcription and translation are two of the fundamental processes represented within the model. Transcription initiation occurs after a promoter site has been bound by an RNA polymerase for a given period of time. Once transcription initiation has occurred, and the cell has enough free energy, the polymerase transcribes the gene in a single reaction. Each gene will have a specific length of nucleotides, which is used, together with a rate of elongation, to generate the reaction rates for transcribing a gene. Equation 7.4 shows the generic transcription reactions within the model:

$$RNAP + promoter \rightleftharpoons RNAP\_promoter \tag{7.2}$$

$$RNAP\_promoter + energy \rightarrow promoter + RNAP\_transcribing \tag{7.3}$$

$$RNAP\_transcribing \rightarrow RNAP + mRNA \tag{7.4}$$

Translation is modelled in a similar way to transcription as the process is reduced to a single reaction. If the cell has enough free energy translation of an mRNA molecule can occur. Each mRNA molecule will have a length which is once again used, together with the abundance of ribosomes, to generate a reaction time for the translation process. Equation 7.6 shows the generic translation reactions within the model:

$$mRNA + energy \rightarrow mRNA + translating\_mRNA \tag{7.5}$$

$$translating\_mRNA \rightarrow protein \tag{7.6}$$

Transcription regulation

Regulation of transcription is performed by transcription factors binding to the cis-activator, cis-repressor or promoter sites in the regulatory region of a gene. The effect of an activated regulatory region is that the promoter site and RNA polymerase molecule will bind more strongly, increasing the chance of transcription initiation. Conversely, the effect of a repressed regulatory region is that binding is prevented between the promoter site and RNA polymerase, turning off any transcription.

To determine the state of a regulatory region (activated, repressed or neutral) we employ a novel method that we term 'transcription logic'. Transcription logic consists of a Boolean logic table and a corresponding function called the 'expression state'. A column, or Boolean variable, is added to the logic table for each cis-activator and cis-repressor element in the regulatory region. All possible Boolean combinations of these variables are then generated in the table. For each row in the table an 'expression level' is given as shown in Table 7.1 to reflect if transcription is possible and how likely it is to initiate.

State	Description
0 repressed	no possible expression
1 unactivated	basal or leaky expression
2 activated	full possible expression

Table 7.1: Expression states and descriptions.

Using this method, any possible function can be applied to the regulatory region giving the model its complexity and flexibility. For instance, to simulate the expression of the *lac* operon, the regulatory region would consist of a single cis-activator, a single cis-repressor and a single promoter site. The 'transcription logic' function for the *lac* operon is given in Table 7.2.

Activator	Repressor	State
-	-	1
-	+	0
+	-	2
+	+	0

Table 7.2: Example of 'transcription logic' for *lac* regulation, where + is bound and − is unbound.

7.4 Model simulation

Each model is simulated using a modified Gibson-Bruck stochastic algorithm [39], providing an exact stochastic simulation of a trajectory of a molecular system. Therefore, by using a stochastic framework, time is continuous, molecule abundances are discrete values rather than concentrations, and intrinsic noise is introduced. Due to the incorporation of realistic reaction rates for transcription, translation and molecular interactions, accurate timescales for these processes are produced, providing a realistic timescale for model output.

Modifications to the algorithm include 'static reactions' which are non-Markov, fixed time reactions that allow species abundances, or reaction rates to be changed, for example due to environmental changes. Also, Logic-based termination criteria have been introduced for ending each model simulation.

Each model is simulated until one of the following termination criteria is met:

- The model has reached the appropriate 'replication threshold' of free energy:

$$base\ replication\ threshold + genome\ size * additional\ energy\ per\ gene$$

- The model has reached a maximum simulation time threshold [simulated wall time]
- The model does not have enough free energy to produce either an mRNA or a protein, and no protein or mRNA exist, then the model is classed as dead

7.4.1 Model parameters

The model consists of a number of free parameters, which are able to evolve, and fixed parameters. Free parameters include all molecule and DNA element shape parameters (θ, ϕ), with the exception of energy, food and RNAP which are fixed during evolution. Protein and mRNA degradation rates are also free to evolve. The fixed parameters such as transcription and translation rates, food uptake and metabolism rates, and diffusion-limited molecule interaction rates are all derived where possible from *E. coli* experiments. In the simulations we present, all proteins have two domains (with allosteric effects) with a simplified metabolism of a single food molecule that is broken down to form energy. The fixed parameters used in the model are given in Table 7.3.

Parameter	Value(s)	Notes
Food species	1	Represents glucose (1 'food' molecule = approx. 14 glucose)
Initial genome size (number of genes)	1	
Mobile-mobile molecule interaction rate	10^{-4} / second	[83, 27]
DNA-mobile molecule interaction rate	10^{-2} / second	[70]
Regulatory region	*lac* operon regulation	
RNA Polymerase per gene	3	Each cell around 2000 active RNAP [43] and up to 700 operons [104]
Gene length	1080nt	*E. coli* K-12 genome length 4639221 bp with 4289 genes [24]
Transcription rate	50nt / second	[10, 20]
Transcription cost	8 energy	Approx 2000 ATP to transcribe 1080nt [94]
Transcription initiation rate	1 / second	
Activated RNAP-promoter complex off rate	0.1 / second	Gives 90% chance of transcription starting
Unactivated RNAP-promoter complex off rate	1 /second	Gives 50% chance of transcription starting
Protein size	360aa	Each amino acid is 3 nucleotides
Translation rate	15aa / second	[129, 3]
Translation cost	6 energy	Approx 1500 ATP to translate 360aa [94]
Ribosome abundance	4.5 ribosomes / mRNA	18,000 ribosomes per cell and up to 4000 mRNA molecules per cell [20]
Food uptake rate	1.5 / second	Loosely calculated from actual glucose uptake rates
Food metabolism rate	3.5 / second per enzyme	Loosely calculated from glycolytic cycle rates
Energy released from metabolism	2	Glycolysis yields 36 ATP molecule (1 'energy' molecule = 252 ATP)
Initial energy amount (and after replication)	100	
Initial protein amount	10	
Initial food amount	10	
Replication base energy threshold	1000	
Additional energy per gene	100	

Table 7.3: Parameters that are fixed in current model implementation.

7.5 Evolutionary framework

The evolutionary framework used in this work is based on a standard genetic algorithm, in which a population size is defined, and random models are initialised to fill this population. Each model in the initial population is then simulated sequentially. Upon termination of the simulation, the simulated time and energy level are recorded. As the fitness function for the surviving models is inversely proportional to the time taken for replication, models with a quicker replication time are therefore fitter than models with a slower replication time. Model fitness is determined by either:

- If the model reached the replication threshold before the simulation time was exceeded, then the fitness is the simulated time for the model to replicate.
- If the model did not replicate, but still had some free energy, then the fitness is:

$$max\ simulation\ time * (max\ simulation\ time / final\ energy\ level)$$

  Using this fitness function models which were terminated with higher levels of energy will be treated more favourably than those with lower levels.
- If the model died, then its fitness is infinity.

Once the initial population has been created and initially simulated, the evolution process begins. The use of a fixed size population structure provides a source of competition between organisms. Each model in the population (regardless of its previous simulation) replicates to produce an identical model. If the cell survived (replicated or hit the simulated time threshold), then the mobile molecules within the parent cell (proteins, mRNAs and food) excluding RNA polymerase are randomly divided between the two cells using a random normal($\mu = 0.5$, $\sigma = 0.1$) for each molecular species. Dead models receive no molecules. Evolutionary operators are then applied to each model in turn, and each copy of the model is simulated. Once again the simulated time and energy are recorded. The population must then be reduced to its original size using an elitism strategy: models are selected (without replacement) according to their fitness. In this, and subsequent generations, models which did not replicate but did not die are allowed to be selected, however, if not enough surviving models exist, new random models are introduced. This new population is then carried forward to the next generation, where the process starts

again. This evolutionary process is therefore a 'purely elitist' deterministic strategy, such that the models are selected based only on fitness (which due to the simulation algorithm is subject to stochastic fluctuations). Each parameter setting is run three times. Evolution parameters are give in Table 7.4

7.5.1 Evolutionary operators

The evolution framework currently supports three evolutionary operators, *gene duplication*, *gene loss* and *mutation*. These operators are applied to each parameter within each gene with a given probability. The evolutionary results presented in the results section are obtained using low rates of gene duplication, gene loss, and mutation events ($P(d) = P(l) = P(m) = 0.1$).

Gene duplication

Gene duplication has been shown to have had a significant impact on the evolution of genomes [117, 55], and has been used in previous models such as ARN [11, 76] and the mathematical model by Wagner [123]. Therefore, it is important for this process to be included within our model. However, due to real evolutionary timescales and the timescale that can be feasibly simulated computationally, the duplication and loss events are simulated at a much higher rate than has been estimated over the course of millions of years of evolution.

Gene duplication is implemented using the following algorithm:
For each gene in the genome, the gene and its regulatory region are duplicated, with the specified probability $P(d)$. The products of the original and duplicated gene are considered to be different molecular species within the simulation algorithm, but will have identical parameters.

Gene loss

Whilst the genome is able to increase in size using gene duplication, it is also possible to decrease in size by gene loss. Gene loss is also an important process in the evolution of genomes, as it allows the genome to remove useless or non-functional 'junk' genes. This is preferable as 'junk' genes would still be replicated or transcribed, and so waste energy.

Gene loss is implemented using the following algorithm:
For each gene in the genome (whilst there are still at least two genes), the gene, its mRNA

and protein products and its regulatory region are removed from the model, with the specified probability $P(l)$.

Mutation

Mutation (or divergence) is the primary operator for increasing diversity within bacteria, that are reproducing asexually. Any shape including natural and allosteric forms of domains, and cis-regulatory DNA elements, or degradation rate in the model is available to mutate. The mutation operator used is a random normal noise added to the shape, or a log normal random noise to the degradation rate: For each gene in the genome, the binding sites of its regulatory region elements, its mRNA product and its protein product can mutate with a specified probability.

Briefly, the mutation operator on 'shape' is a random deviation of the original position on the surface of the sphere. The random deviation is achieved by generating 'noise' at the pole of the sphere and rotating this 'noise' to the original position to generate the mutation: [1]
Random noise is generated at the pole of the shape sphere

$$\eta = random\ normal(\mu = 0, \sigma = SHAPE\ NOISE\_SDEV)\ [in\ the\ \theta\ direction] \quad (7.7)$$

$$\psi = random[0, 2\pi)\ [in\ the\ \phi\ direction] \quad (7.8)$$

The generated noise is then rotated to the current coordinates (θ, ϕ) using Cartesian algebra:

$$x = \cos(\theta)\cos(\phi)\sin(\eta)\cos(\psi) - \sin(\phi)\sin(\eta)\sin(\psi) + \sin(\theta)\cos(\phi)\cos(\eta) \quad (7.9)$$

$$y = \cos(\theta)\sin(\phi)\sin(\eta)\cos(\psi) + \cos(\phi)\sin(\eta)\sin(\psi) + \sin(\theta)\sin(\phi)\cos(\eta) \quad (7.10)$$

$$z = -\sin(\theta)\sin(\eta)\cos(\psi) + \cos(\theta)\cos(\eta) \quad (7.11)$$

The Cartesian coordinates can then be transformed back into polar coordinates in the standard way. Allosteric shape mutations are applied using the same mutation operator.

As well as the shape mutation, the mRNA and protein product can mutate their degradation

[1] This mutation operator and mathematical formulation was derived by Dr Dov Stekel

rate with log-normal noise:

$$degradation\ rate_{new} = degradation\ rate_{old} * 10^{random\ normal(\mu=0,\sigma=DEG\ RATE\ NOISE\ SDEV)} \tag{7.12}$$

The 'transcriptional logic' functions are currently not subjected to mutation, and so are fixed to *lac* logic.

7.5.2 *In silico* genetics

In silico genetics is the equivalent of performing genetic experiments *in vitro*, except the cells are simulated on a computer. To allow similar experiments to be performed on our model, a custom *in silico* genetics tool was developed, allowing modification of any free or fixed parameter within the cell. This enables 'mutant' cell lines to be created, with changes such as gene or regulatory site knock-out, novel genes, increased or decreased molecule stability, molecular shape etc. Using this technique, it is possible to examine the effects of perturbations in a cell, similar to those techniques used in the laboratory.

Parameter	Value(s)	Notes
Population size	100	
Generations	50	
Maximum simulation time	1 hour	
Mutation rate $P(m)$	0.1, 0.3, 0.5	Varied rates used for sensitivity analysis
Gene duplication rate $P(d)$	0.1, 0.3, 0.5	Varied rates used for sensitivity analysis
Gene loss rate $P(l)$	0.1, 0.3, 0.5	Varied rates used for sensitivity analysis
σ binding affinity value	1, 10, 20, 30, 40, 50	Approximate range determined from model dynamics analysis
α binding affinity value	1	Fixed value determined from model dynamics analysis
Mutation shape random normal s. dev.	0.2	
Initial protein degradation rate	$10^{random\ normal(-2.5,0.5)}$	Average time of 612 seconds
Initial mRNA degradation rate	$10^{random\ normal(-2.5,0.1)}$	Average time of 324 seconds
Mutation protein degradation rate s. dev.	0.2	
Mutation mRNA degradation rate s. dev.	0.05	

Table 7.4: Parameters that are fixed during evolution in current model implementation.

Chapter 8

MODEL PARAMETER AND DYNAMICS ANALYSIS

8.1 Structural parameter analysis

The distribution of random models meeting each of the three termination criteria (replicate, stationary and death) was investigated over a large range of the binding affinity parameters, σ and α. The results of 1000 randomly initialised models for each parameter setting (10^{-3} up to 10^{3}), simulated in a constant and abundant food environment (meaning that the model will always be able to uptake food at the specified rate), are shown in Figure 8.1. Figure 8.1A shows the replicating models for each parameter, where a clear band of 'livable' parameters can be seen. We see in Figure 8.1B that it is highly likely for a random model to be simulated for an hour without replicating, and from Figure 8.1C we have the inverse from A, and we see a band of non-dying parameters and the majority of parameters giving a large percentage of dying models. It is useful to see the parameters ranges in which random models struggle to replicate, as these parameters will provide good starting points for evolution, and will encourage more searching of the solution space. The parameter ranges used for subsequent evolutionary experiments was σ=1 to 50, and α=1.

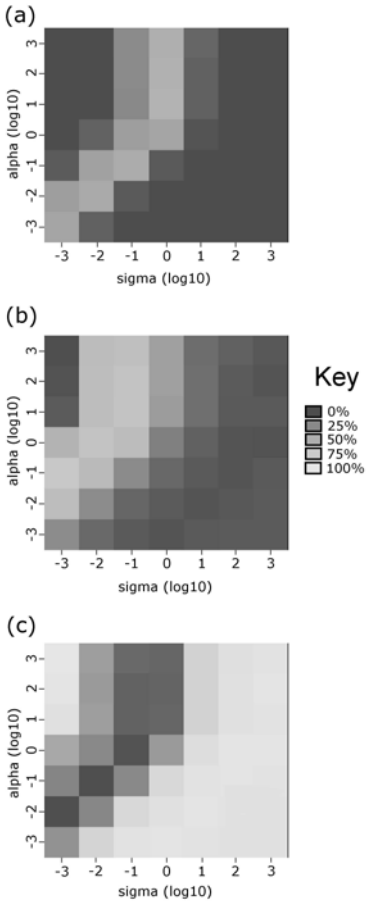

Figure 8.1: Random model simulations over structural parameters. a) shows the proportion of models replicating in each of the environments. A clear light band can be seen passing through from bottom-left to the top middle, indicating environments which are easier to survive in. b) shows the proportion of models that are stationary. A darker band, indicating fewer stationary models following the same pattern as in a), although skewed to the left. c) shows the proportion of models that died in each environment. A light band can again be seen, and follows the same pattern as the dark band in b). Black is 0%, white is 100%.

8.2 Model dynamics

Models consisting of a single gene exhibited four different classes of behaviours, as seen in Figure 8.2. The first behaviour is growth (Figure 8.2A), where the energy level gradually increases up to the replication threshold level; this behaviour, along with a similar behaviour in protein production, may indicate a linear cell volume increase, which is consistent with observed volume growth in *E. coli* [74]. The second behaviour is death (Figure 8.2B), where the energy level hits 0 (or other death criteria level), due to over-expression of the genes and unsustainable usage of energy. These two behaviours are 'primary' behaviours, of which only one is observed over the course of a simulation (the cell either replicates or dies). The second set of behaviours are

'secondary', in that there can be many instances of them observed throughout the simulation. 'Bursting' behaviour can be seen in Figure 8.2C. This behaviour consists of a growth phase, followed immediately by a substantial decrease in energy levels (this can be seen from the figure at around 600 and 1050 seconds into the simulation). Coinciding with the drop in energy is an immediate increase in protein level, indicating that the rapid decrease in energy would be due to rapid transcription and translation events, producing a 'burst' of protein production. The timing of the burst may be due to availability of RNA polymerase, or transcription factor interactions, such as a repressor unbinding or an activator binding. The sudden decrease in energy over a matter of minutes would be expected with the current parameters, as the approximate time to initiate transcription is 1 second, to transcribe the gene is 21.6 seconds and then a further 24 seconds for each fully translated protein to be produced, meaning that in a matter of minutes multiple mRNA molecules can be produced and many more proteins can be produced from them (assuming 4.5 ribosomes per mRNA). 'Stationary' behaviour can be seen in Figure 8.2D. This behaviour consists of the energy level remaining static for a period of time. This indicates a period of transcriptional and translational inactivity, as the figure shows there is enough energy for producing an mRNA transcript, but no transcription takes place, meaning either the gene is repressed, or the limited RNA polymerase molecules are bound to other molecules. A similar behaviour of transcriptional inactivity may be observed in real cells, for instance undergoing a stress, such as heat or acid shock. Stress response often leads to large changes in gene expression, as unimportant genes are switched off and only essential response genes (such as those encoding chaperone or helper proteins) are switched on to conserve energy [94]. Lab based evolution of *E. coli* has shown that mutations reducing the transcription of flagella synthesis genes in the stringent response regulatory network offer a significant fitness advantage [100].

The behaviours and dynamics of the model described above were investigated using random initialisations with the structural parameters $\sigma = 1$ and $\alpha = 1$, which were determined through the previous sensitivity analysis of the structural parameters.

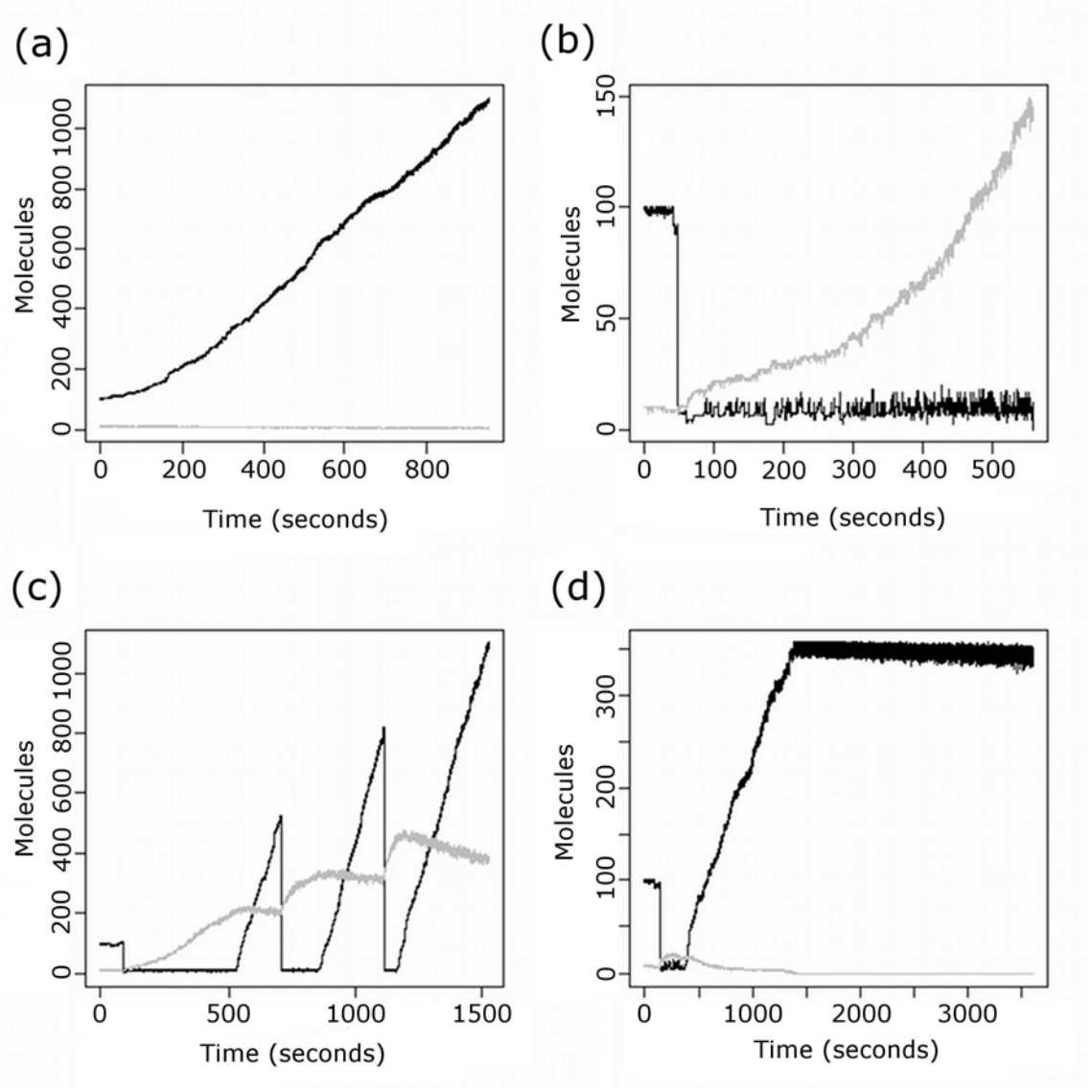

Figure 8.2: Example of single-gene model behaviours. All models have the same structural parameters of $\sigma = 1$ and $\alpha = 1$, and randomly initialised evolvable parameters. A is an example of growth, where the protein level stays constant and the energy level rises steadily reaching the replication threshold of 1100 free energy molecules in around 900 seconds. B is an example of death, where the protein level slowly increases up to around 150 molecules after 500 seconds, whereas the energy level rapidly falls to around 20 molecules and stays close to this amount before it eventually runs out of energy. Energy is very frequently released from metabolism, but is rapidly used in protein production and thus generates the noisy low level of energy. C is an example of 'bursting', where protein level slowly rises to around 200 molecules, before a large rapid increase in energy around 500 seconds, protein level then rises again followed by an even faster drop in energy levels, causing a 'burst' in the protein level. This behaviour is repeated several more times before the energy threshold is reached. D is an example of 'stationary' where after an initial drop, followed by increase in energy, the energy and protein levels appear to reach a steady state around 1500 seconds and remain static. Black line is energy level, grey line is protein level.

8.3 Parameters essential for model replication

To determine important evolvable parameters, 1000 models were randomly initialised and simulated in a constant external food environment, recording evolvable parameters. Each model was simulated 20 times and was classified either as replicating (class 1) or not (class 0), depending on the majority results from the 20 simulations. Both univariate and multivariate methods were used to determine important parameters for replication.

8.3.1 Univariate analysis

Logistic regression [25] with controlled false discovery rate [16] was used to determine significance of each parameter as a predictor for model replication. Table 8.1 shows all parameters, ranked most significant (8 parameters with $q < 0.05$) to least significant. The most significant and accurate predictor for class membership was the protein degradation rate, with an accuracy of over 81%. Other significant parameters included the mRNA degradation rate and various repressor and promoter complex minimum dissociation rates, although classification is only slightly higher than classifying all models as class 0 (non-replicating) which is 55.4%. These results indicate that the most optimal network topology for replication in this simplistic environment would require specific interaction between various molecules and the repressor and promoter sites, and is also very sensitive to the degradation rates of gene products. The significance of sensitivity to protein degradation rate is likely to be due to the protein molecule's role as a metabolic enzyme. As the model can only gain energy by breaking down food molecules and only protein molecules have this potential functionality, the interaction of protein and food molecules is therefore very important. The advantage of a stable protein is that there is an increased probability that the protein will interact with food molecules before it degrades, allowing more metabolic reactions to take place. For metabolism to take place the protein must be bound to the food molecule for a fixed period of time ($\approx$ 0.3sec). Interestingly, the stability of the protein-food complex does not appear to be a significant parameter, due to the rapid rate of metabolism. This indicates that only the rate of food-protein bindings are important, as stochastic effects mean that sufficient numbers of weakly active proteins may be sufficient to allow replication. Stable proteins also need to be replaced less frequently than unstable proteins, requiring less transcription and translation activity and therefore saving energy.

Parameter	ID	p-value	q-value	Classification	Sensitivity	Specificity
Protein degradation rate	22	$< 2e^{-16}$	$5.2e^{-15}$	81.7%	0.77803	0.84838
Protein-Repressor complex k_{off}	5	$1.38e^{-10}$	$1.794e^{-9}$	54.7%	0.32287	0.77617
Protein-Promoter complex k_{off}	6	$1.29e^{-8}$	$1.118e^{-7}$	60.4%	0.34978	0.80866
All Promoter complex k_{off}	26	$1.06e^{-7}$	$6.89e^{-7}$	57.6%	0.28027	0.81408
mRNA degradation rate	23	$6.05e^{-7}$	$3.146e^{-6}$	58.1%	0.32287	0.78881
Energy-Promoter complex k_{off}	14	$8.1e^{-5}$	$3.51e^{-4}$	58.3%	0.23318	0.86462
Food-Repressor complex k_{off}	10	0.00273	0.01014	55.3%	0.12556	0.89711
Protein-Protein complex k_{off}	8	0.00346	0.011245	56%	0.12108	0.91336
Protein-RNAP complex k_{off}	3	0.0233	0.06731	56.8%	0.08969	0.95307
Energy-Repressor complex k_{off}	13	0.0375	0.0975	55.1%	0.03587	0.96570
Protein-Food complex k_{off}	1	0.0428	0.097301	55.3%	0.04484	0.96209
All Repressor complex k_{off}	25	0.044908	0.097301	55.5%	0.06278	0.95126
Protein-Energy complex k_{off}	2	0.0975	0.195	55.1%	0.01794	0.98014
Food-Activator complex k_{off}	9	0.38195	0.70934	55.6%	0.00673	0.99819
mRNA-Activator complex k_{off}	17	0.43156	0.7480	55.3%	0	0.99819
mRNA-Promoter complex k_{off}	19	0.4857	0.7893	55.4%	0	1
Protein-Activator complex k_{off}	4	0.5166	0.7901	55.4%	0	1
Energy-Activator complex k_{off}	12	0.56378	0.81435	55.5%	0.00224	1
Protein-mRNA complex k_{off}	7	0.61995	0.84835	55.4%	0	1
mRNA-Repressor complex k_{off}	18	0.6368	0.82784	55.4%	0	1
Food-Promoter complex k_{off}	11	0.739600	0.9156952	55.4%	0	1
All Activator complex k_{off}	24	0.906519	0.9712	55.4%	0	1
RNAP-Repressor complex k_{off}	16	0.9084	0.9712	55.4%	0	1
mRNA-Energy complex k_{off}	21	0.9651	0.9712	55.4%	0	1
RNAP-Activator complex k_{off}	15	0.9655	0.9712	55.4%	0	1
mRNA-Food complex k_{off}	20	0.9712	0.9712	55.4%	0	1

Table 8.1: Univariate analysis of evolvable parameters, ranked by significance. The top eight parameters are significant ($q < 0.05$). For each parameter its ID number, original p-value from a logistic regression, adjusted q-value from controlling the false discovery rate, its classification accuracy, sensitivity and specificity are shown.

8.3.2 Multivariate analysis

Multivariate analysis was performed with GALGO [120], using the diagonal linear discriminant analysis (DLDA) classifier, 200 solutions and a goal fitness of 0.85. All other options were set to default. Models consisting of 2 to 5 parameters were generated, and each model size was repeated 5 times. Multivariate solutions were able to improve classification of more than 5% over univariate solutions. Table 8.2 shows the optimal solutions generated during each GALGO run on each model size and Table 8.3 shows the proportion of parameters selected in the optimal solutions. A model size of 2 generates a solution which includes the two most significant parameters from the univariate analysis, which again indicates a network topology dependent on repressor interaction and protein degradation rate. This solution only improves classification by 2% over the single most significant single parameter, therefore highlighting the major contribution of which this parameter has on model replication. Increasing the model size further only yields slight increases in classification. A 5-parameter classifier achieves only 5% improvement, and a 10-parameter classifier only improves by around 0.5% on this. This result indicates that very few parameters have any significant effect on classification, most of which were found to be significant from the univariate analysis. The parameters which appeared most frequently in the multivariate solutions were again protein and mRNA degradation rate. This indicates the model's major sensitivity to molecule stability through the protein and mRNA degradation rates and minor sensitivity to interactions with the repressor and promoter sites on the DNA through the dissociation rates.

Model size	Optimal solution	Classification	Sensitivity	Specificity
2	5, 22	83.8%	0.80493	0.86462
3 (run 1)	5, 14, 22	84%	0.81839	0.85740
3 (run 2)	5, 6, 22	85.5%	0.82511	0.87906
3 (run 3)	5, 22, 26	84.5%	0.81839	0.86643
3 (run 4)	5, 6, 22	85.5%	0.82511	0.87906
3 (run 5)	5, 22, 23	84.8%	0.82287	0.86823
4 (run 1)	5, 6, 13, 22	86.2%	0.84081	0.87906
4 (run 2)	5, 22, 23, 26	86%	0.84081	0.87545
4 (run 3)	5, 6, 14, 22	86.3%	0.83632	0.88448
4 (run 4)	5, 6, 14, 22	86.3%	0.83632	0.88448
4 (run 5)	3, 5, 6, 22	85.1%	0.82287	0.87365
5 (run 1)	5, 6, 14, 22, 23	86.8%	0.85650	0.87726
5 (run 2)	5, 16, 22, 23, 26	86%	0.84081	0.87545
5 (run 3)	5, 6, 8, 14, 22	86.3%	0.83632	0.88448
5 (run 4)	5, 6, 14, 19, 22	86.3%	0.83632	0.88448
5 (run 5)	5, 14, 22, 25, 26	84.5%	0.82511	0.86101

Table 8.2: Multivariate solutions generated by GALGO. For each model size and run, the best solution, its classification accuracy (using logistic regression), sensitivity and specificity are shown. All five runs with a model size of 2 generated the same optimal solution. All optimal solutions included parameters 5 (Protein-Repressor complex k_{off}) and 22 (Protein degradation rate), which were also the two most significant parameters identified from the univariate analysis. All parameters in the optimal solutions, with the exception of 3 (Protein-RNAP complex k_{off}), 13 (Energy-Repressor complex k_{off}), 16 (RNAP-Repressor complex k_{off}), 19 (mRNA-Promoter complex k_{off}) and 25 (All Repressor complex k_{off}), were identified as significant from the univariate analysis. The only significant univariate parameter not included in the multivariate solutions was 10 (Food-Repressor complex k_{off}).

Parameter	ID	Percentage in model size 2	3	4	5
Protein-Food complex k_{off}	1	0	1.2	0.5	2.4
Protein-Energy complex k_{off}	2	0	1.2	3.8	7.4
Protein-RNAP complex k_{off}	3	0.3	2.4	11.3	4.2
Protein-Activator complex k_{off}	4	0	1.5	3.3	10.1
Protein-Repressor complex k_{off}	5	88.7	81.5	74.4	97.8
Protein-Promoter complex k_{off}	6	7.5	29.4	66.1	65.4
Protein-mRNA complex k_{off}	7	0	2.6	3.9	6.1
Protein-Protein complex k_{off}	8	0	0.1	6	6.3
Food-Activator complex k_{off}	9	0	1.8	4.7	5.2
Food-Repressor complex k_{off}	10	0	0.2	4.4	5
Food-Promoter complex k_{off}	11	0	0.3	4.4	3.8
Energy-Activator complex k_{off}	12	0	0.1	4	7
Energy-Repressor complex k_{off}	13	0	0.8	9.2	19.4
Energy-Promoter complex k_{off}	14	0	13.6	23.3	31.8
RNAP-Activator complex k_{off}	15	0	0.1	4.6	7.4
RNAP-Repressor complex k_{off}	16	0	0.2	3.1	7.1
mRNA-Activator complex k_{off}	17	0.1	0	2.2	5.7
mRNA-Repressor complex k_{off}	18	0	0.4	1.6	6.6
mRNA-Promoter complex k_{off}	19	0	1.6	2.8	11.8
mRNA-Food complex k_{off}	20	0	0	3.9	4.9
mRNA-Energy complex k_{off}	21	0	0.5	1.4	6.1
Protein degradation rate	22	100	100	100	100
mRNA degradation rate	23	0.5	36.7	31.2	23.4
All Activator complex k_{off}	24	0	7.6	1.9	5.4
All Repressor complex k_{off}	25	0	7.1	2.7	9.6
All Promoter complex k_{off}	26	2.9	6.7	17.3	21.9

Table 8.3: Evolvable parameters selected in optimal solutions generated by GALGO. For each parameter its ID number and its percentage in solutions of different model sizes are shown. Parameter proportion in model solutions is averaged over 5 runs.

Chapter 9

EVOLUTIONARY SIMULATIONS

9.1 Realistic replication time is a property of the model

Cell cycle or time to replicate appeared to reach a minimum of around 300 seconds, with typical replication times for the evolved population between 400 and 1000 seconds. The replication time of *E. coli* K-12 depends on growth medium, ranging from 20 minutes up to an hour or more [94]. The replication time of our most efficient evolved cells ranges from around 6 to 15 minutes (the average time for one final generation was 11.5 minutes); therefore it is fair to claim realistic cell replication times as a property of the model, as our cells only model regulatory, metabolic and signalling genes, while processes such as cell growth and DNA replication are not explicitly included in the model. Models consisting of two or more genes evolved similar replication times to those models with only a single gene, indicating that having multiple genes may not always be prohibitive to efficient replication times. Figure 9.1A shows an example simulation of an evolved model, which replicates in 13.4 minutes, and has a protein steady state around 200 molecules.

Replication and replication times were also evolved to be more consistent. 1000 simulations of an evolved model and of its ancestor (the original random model from which the 'evolved' model is descended from) model and were examined. The evolved model replicated in 96.7% of the simulations and the ancestor model achieved only 94.9% replication. 100 replication events were selected from both models for comparison and are shown in Table 9.1.

The maximum speed of replication was similar between the ancestor and evolved models, however, the mean replication time, and standard deviation were reduced in the evolved model. This indicates that the model has evolved a more consistent replication time, but has not

Model	mean replication time	s. dev.	minimum replication time
Ancestor	15.48	5.19	9.27
Evolved	14.14	3.73	9.38

Table 9.1: Mean replication times and standard deviations and minimum replication times in minutes for 100 replication simulations of the ancestor and evolved model.

increased the speed of replication.

9.2 Evolution of stable proteins and unstable mRNA

Investigating other aspects of the evolved model also shows some interesting and life-like trends and principles. The degradation rates of mRNA and protein species within a range of models from different evolutionary environments displayed similar behaviour, selecting for unstable mRNA molecules with typical mean degradation time of under 3 minutes and stable proteins with typical mean degradation time of several hours (see Table 9.2). Molecule stability is defined as the time taken for the molecule to passively degrade, rather than a adopting a particular confirmation from a set of confirmation states.

	protein degradation time			mRNA degradation time		
Generation	mean	s. dev.	range	mean	s. dev.	range
1	10.46	14.21	0.13 - 75.36	5.53	1.44	3.002 - 10.72
50	359.59	384.14	25.5 - 1833.27	2.8	0.505	1.41 - 3.95

Table 9.2: Mean protein and mRNA degradation time and standard deviations for initial and evolved generations in minutes

These are remarkably close to turnover rates in biological cells. The average turnover time for an mRNA molecule in *E. coli* is around 5 minutes [18], and protein degradation rates in *E. coli* and *Saccharomyces cerevisiae*, although wide ranging, are often an order of magnitude higher than those of mRNA [93, 125, 15]. Table 9.2 shows the evolved changes in mean mRNA and protein turnover rates. Although both start at similar levels, mRNA degradation time decreases from 5.53 to 2.8 minutes, whereas protein degradation time increases from 10.46 to around 360 minutes. The increased stability of the protein would allow the same protein to metabolise more food molecules, and so decrease the need for further protein production. However, the large standard deviation of the mean protein degradation rate indicates a very large variation between individual models.

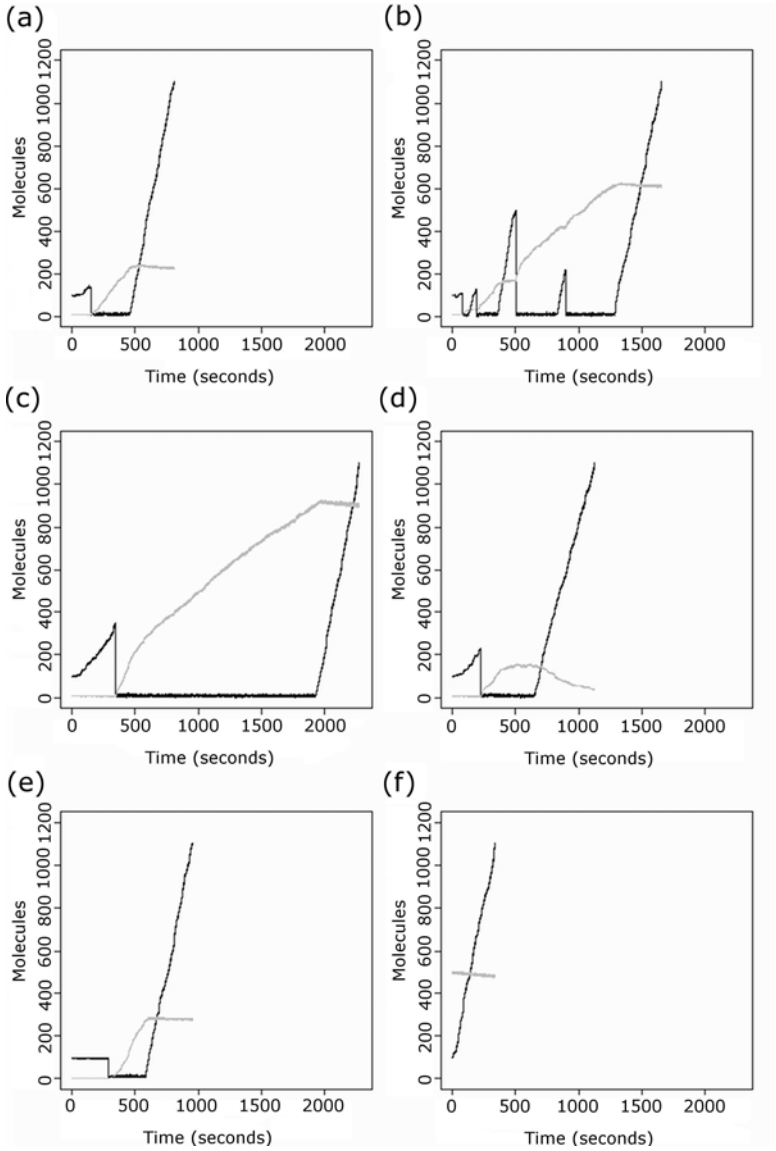

Figure 9.1: Example simulations of evolved single-gene model. A shows the 'wild-type' model, where following the usual 'lag' phase of the energy level dropping to below 20 molecules due to initiation of transcription and subsequent increase in protein, the energy level rapidly increases at around 500 seconds with the protein reaching a steady state of 200 molecules. The energy threshold is reached by 800 seconds. B shows the repressor knock-out mutant, in which 'bursting' has been introduced, as the protein level does not enter steady state due to the increased transcription. After several 'bursts' the energy threshold is reached after 1500 seconds. C shows the increased mRNA stability mutant, in which the 'lag' phase is significantly increased, due to increased translation. 'Wild type' growth behaviour occurs around 2000 seconds, however, with a protein steady state of 850 molecules. D shows the decreased protein stability mutant, in which 'wild-type' growth occurs after 550 seconds, however, a protein steady state is not reached. E shows the 0-protein mutant, in which the 'lag' phase continues for around 500 seconds. After which, transcription and translation occur, causing 'wild-type' growth after 650 seconds with a protein steady state of 300 molecules. F shows the 500-protein mutant, in which no 'lag' phase is evident and 'wild-type' growth occurs immediately, with the energy threshold being reached in around 500 seconds, with a protein steady state of 500 molecules. Black line is energy level, grey line is protein level. All plots are shown on the same scale.

We use *in silico* genetics to investigate the model behaviour to changes in the degradation rates. Decreasing the mRNA degradation rate (mRNA stability increased to 35 minutes from 3.1 minutes) has the effect of increasing steady state protein levels from around 200 molecules in the evolved model to around 850 molecules in the mutated model and increasing the replication time from 13.4 minutes up to 38 minutes (Figure 9.1C). This is because the mRNA molecules exist within the system for longer therefore allowing more transcription. In this example the model is still capable of replication, but with a slower replication time. Increasing the protein degradation rate (protein stability is decreased to 6 minutes from 246.2 minutes) leads to a decreased protein steady state level after the initial transcription activity of around 150 molecules which rapidly decreases, leading to an increased replication time of 18.8 minutes up from 13.4 minutes. Figure 9.1D shows how the initial protein level is reached and then transcription is repressed, as expected. However, the protein level then quickly decreases, rather than staying at a constant level, due to the increased degradation rate. Once again in this example the protein level was high enough to support replication, but with decreased efficiency. The perturbation of degradation rate indicates the adaptive evolution of the mRNA and protein molecules to reduce energy usage from over-expression by limiting availability of mRNA to translate and reducing the need for frequent translation of constantly required enzymes and transcription factors by low protein degradation rates.

The evolution of very stable protein molecules for metabolism is paralleled in real organisms. However, in general, the stability of proteins is highly dependent on their function. Signalling proteins are often very unstable, allowing rapid response to stimulus; proteins which have a negative effect on the cell under stressed conditions may be unstable, or actively degraded to deal with this. Therefore, the environmental conditions and functions required by the cell are likely to strongly influence the evolution of the stability of proteins. It is important to note that even though the protein is very stable, it is being diluted as a result of cell division, and so the cells need to replenish the protein to be able to function.

Rapid mRNA turnover has previously been suggested as a mechanism to enable rapid response to environmental changes [34]. However, we find that mRNA is rapidly turned over in realistic timescales even in an unchanging environment. Thus it would seem that this turnover is an emergent property associated with two-step gene product synthesis that enables greater control over protein production and energy levels. As the level of protein is tightly controlled

over-expression is less likely, therefore less energy would be wasted in translating unnecessary proteins. It is, more than anything else, an adaptation for efficiency. Another possible benefit from rapidly turned over mRNA and the subsequent protein control is that the noise in protein level will be small, producing molecular homeostasis in a constant environment.

9.3 Evolution of basic repressor activity

Models evolved for effective growth in a constant food environment all developed a single-gene repressor regulatory network, where the single gene was repressed by either a product of the gene (mRNA or protein) or by energy or food molecules. This structure was seen in all model lineages. Figure 9.2A shows a typical ancestor model gene regulatory network (a connection between DNA/molecules is shown only if the K_d of the binding is less than 100nM). We can see that the ancestor model already has a simple repressor system, with the protein product negatively self-regulating its own production.

Figure 9.2B shows an evolved model from the ancestor model in Figure 9.2A. Here we can see that the repressor still exists, but the network has grown to include the protein's second binding domain as a TF, with the repressor shape remaining relatively fixed during the course of the evolution. We can also see that the model has evolved to use the promoter site as a 'secondary' repressor, with fairly large changes in the promoter site's shape. The large changes to the promoter site do not affect the RNAP binding (see section 7.3.2). Figure 9.1A shows the cell initially producing protein up to a required threshold and then repressing any more production, saving energy for replication, whereas Figure 9.1B shows a simulation of the knock-out mutant. Without repression 'bursting' behaviour has been introduced to the model dynamics. As shown previously bursting is the result of mass transcriptional and translational activity. Instead of the model repressing protein production when a required level was reached, in the mutant there is no evidence of the repression and proteins are produced in several bursts of transcription and translation activity, which uses large amounts of energy. This causes the model to have several 'false starts' before finally reaching its required energy threshold, therefore doubling the replication time from 13.4 minutes to 27.7 minutes.

Therefore an efficient mechanism within the model, for the single constant, abundant food environment, is a single-gene repressor. This allows the model to metabolise food successfully

and efficiently to ensure that enough energy is saved to reach the replication threshold. This behaviour is fundamentally realistic, as there are more than 2000 known negative regulation interactions in *E. coli* [69], such as the *trp* operon [3, 105], and many of these are auto-regulated. Negative feedback performs several functions: (i) turns off potential transcription of the gene, if not currently required (for example stress-response proteins), and therefore saving energy, (ii) helps to maintain a specific concentration of the protein, or homeostasis, (iii) increases speed of response within a transcription network [103], (iv) minimises mRNA usage [114].

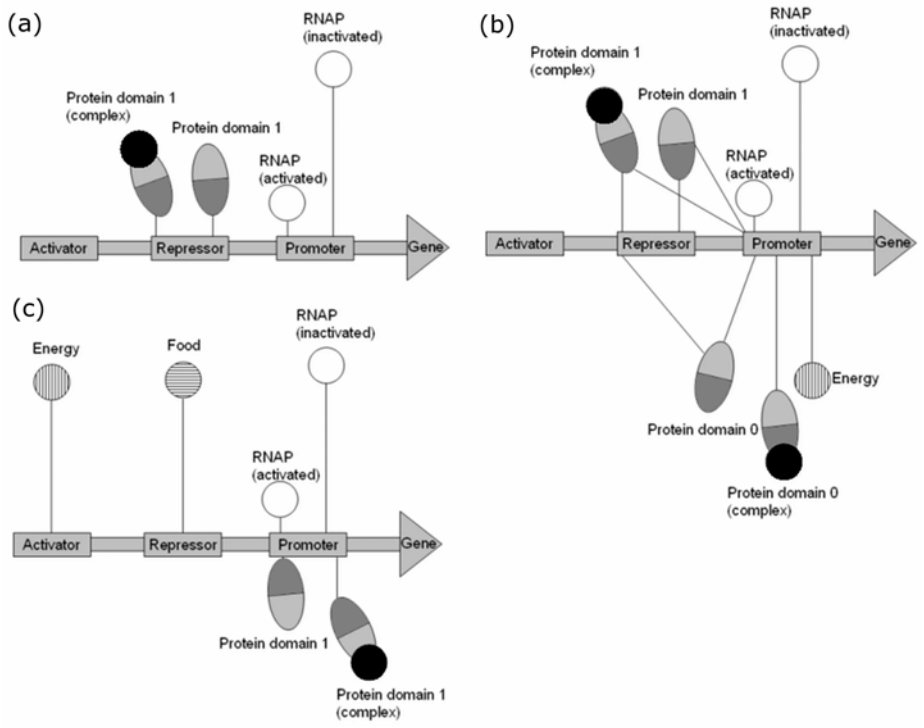

Figure 9.2: Example of ancestor and evolved transcription regulation networks. Two cell lineages were observed, each originating from the initial generation. The final population consisted of 95% of models from the 'major' cell lineage, and the remaining 5% from the 'minor' lineage. Specific bindings of $K_d < 100nM$ are shown. A shows the ancestor network of the major cell lineage from the population. Very strong repressor binding by a single binding domain of the protein is evident. B shows an example evolved network from the major cell lineage. The ancestor repressor connections are still present, although slightly weakened. However, the same protein domain has evolved a specific binding to the promoter site as well. Other evolved specific bindings are the other binding domain of the protein to both the repressor and promoter sites, and energy binding to the promoter site. C shows an evolved network from the minor cell lineage from the same population. Specific binding to the repressor site again exists, although using the food molecule, as does specific binding to the promoter site by a single binding domain of the protein. A specific binding to the activator site using the energy molecule also exists, which was not present in the major cell lineage. Binding strength is approximated by molecules' distance from DNA.

Other network topologies were also evolved. Figure 9.2C shows an evolved network where

both an activator and repressor exist. However, the transcription factors in this example are food and energy molecules, rather than gene products. Energy or food molecules binding directly to the DNA is permitted in this model, although for steric reasons does not appear to happen in life. Real cells have evolved to use energy, food or other types as molecules as signals in regulation by using their binding to transcription factors, causing allosteric changes and effecting the function of the transcription factor. An example would be allolactose in the *lac* regulatory mechanism, in which the lactose metabolite binds to the LacI repressor and prevents it from binding to the DNA, potentially allowing the transcription of the *lacZ* gene. A simple solution would be to restrict the domain shapes which the DNA regulatory elements can take. By limiting the shape space, the food or energy molecules will be unable to bind sufficiently well to the DNA regulatory elements, therefore forcing the model to use a protein as a transcription factor.

The emergence of such a fundamental and life-like network structure indicates the potential power and complexity of the new model as a tool for investigating the evolution of transcription networks.

9.4 Protein is regulated to a realistic low copy number

The protein copy numbers observed within evolved models are typically between 50 and 400 molecules, and in the majority of simulations a stable level was reached within this range. The protein copy number per *E. coli* cell of enzymes within the glycolytic pathway range from only 100 copies up to several thousand copies of other enzymes, each varying depending on cell-cycle [115], although the numbers for many enzymes are unknown. Although our simulations appear to be at least an order of magnitude different, enzyme copy number is likely to be a function of substrate copy number, and so we would observe different levels of protein under different conditions. Each food molecule in our model is equivalent to 14 glucose molecules (see Table 7.3), and therefore once we have taken the scaling of food molecules within the model into consideration the levels of our simulated cell's enzymes are similar to the values observed in biological cells. For example, the enzyme phosphoglycerate kinase has a copy number of around 3000 molecules in the growth phase [115] and assuming each molecule can metabolise only a single 1,3-bisphosphoglycerate molecule at a time, our model would require around 200 proteins to metabolise the equivalent 'food' molecules. This enzyme copy number is well within the

observed simulated copy number of many evolved cells, however, it must be noted that as our model approximates the glycolytic pathway into a single reaction and so is a much simplified, inexact pathway.

Using the *in silico* genetics tool we investigated the impact on the cell's ability to replicate by changing the starting protein level to simulate biased cell replication, which leaves the cell with an extreme amount of protein (very little or large amounts). It is important for biological cells to cope with extremes of protein level as the replication process may create these situations. Figure 9.1 shows an example of each extreme case; no protein (Fig. 9.1E) and 500 protein (Fig. 9.1F). The behaviour for no protein is similar to the 'wild-type' cell, with the exception of a longer lag period at the beginning of the simulation as there are no proteins to metabolise the food, nor any transcription taking place. Once the cell has started to transcribe the gene and proteins are produced, the growth of the cell is very similar to wild-type growth. In the opposite case, where there is a large number of proteins at the beginning of the simulation, we see a different dynamic. Due to the large number of free proteins in the cell, the food molecules entering the cell are immediately consumed producing large amounts of energy. Also due to the high level of protein molecules, the gene is immediately repressed preventing any transcription and so the protein level remains constant.

9.5 Affects of σ and mutation rate on population dynamics

Results from the experiments indicate that the σ binding affinity parameter has substantial impact on the potential for evolving models. In Figure 9.3A we can see a low mutation environment with a very small σ.

The population is very quickly (within 5 generations) dominated by models which are replicating, and by the end of the short evolution of 50 generations we can see that the final population had around 95% replicating models, with no models being in the stationary phase after 1 hour. Comparing to the initial population in which less than 40% of models were replicating, around 20% were in the stationary phase, leaving the remaining 40% of models dying. However, we see a very different population dynamic if we have a larger σ value. Figure 9.3B shows a low mutation environment, but with a large σ binding affinity value of 50. Here 95% of the initial population die, with no models replicating at all. In the early stages of the evolution we see a

slight increase in the number of models which do not die, but do not replicate either, and we start to see models which are persistently capable of replicating, but do not quickly take over the population as was seen in Figure 9.3A. Around halfway through the evolution by generation 25, we start to see a monotonic increase in replicating models, and by generation 40 up to 75% of the population consists of replicating models. However, the amount of stationary models each generation remains fairly constant between 10 and 20%, as do replicating and dying models after generation 40.

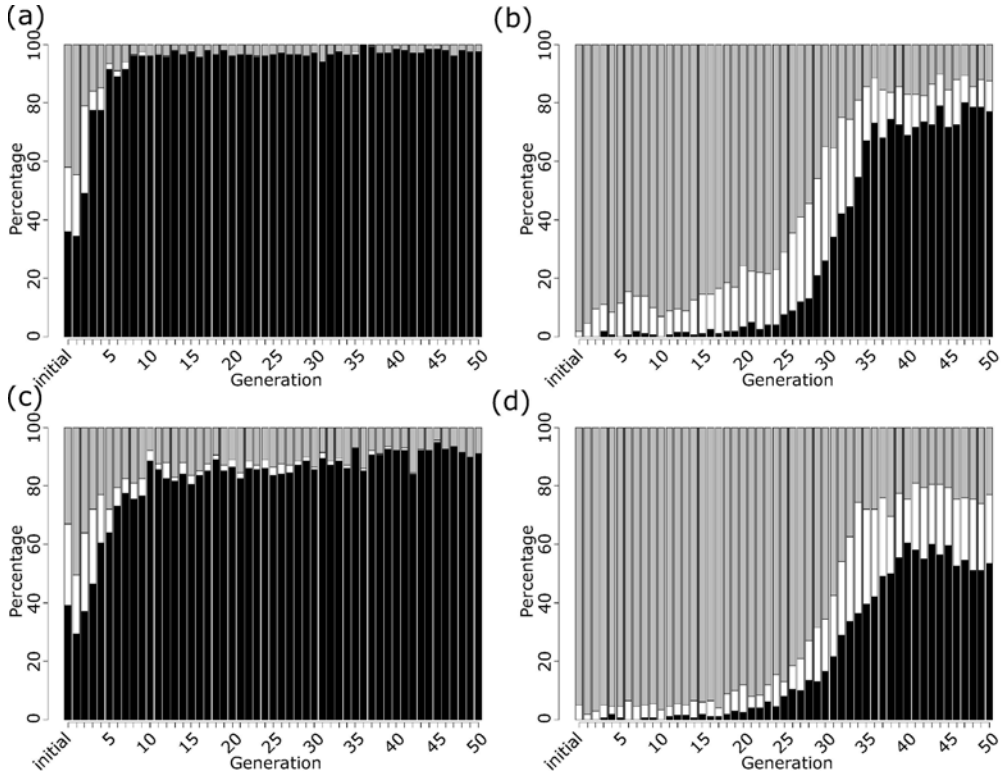

Figure 9.3: Population status each generation. Black replication, white stationary, grey death. A shows a low mutation environment and σ=1, in which the population very quickly becomes dominated by replicating models. A small proportion of models in each generation die, likely due to stochasticity. B shows a low mutation environment and σ=50, where the population initially consists mainly of dying models, but after around 20 generations replicating models begin to establish themselves within the population. The population then rapidly becomes dominated by the replicating models, reaching an equilibrium around generation 35. C shows a high mutation environment and σ=1, where the population again rapidly becomes dominated by replicating models. However, the proportion of dying models each generation is higher than in a low mutation environment, indicating more detrimental mutations taking place. D shows a high mutation environment and σ=30, which again shows a substantial number of generations which are dominated by dying models. Replicating models again begin to establish themselves within the population around generation 20, and rapidly dominate the population. Proportion of models replicating when the population has reached equilibrium is smaller than in other regimes.

In a high mutation environment we see a similar population dynamic (Figure 9.3C). The small value of σ once again quickly reaches a large number of models which replicate, albeit slightly slower than in the low mutation environment. Also the replicating models only consist of around 90% of the population, lower than in the low mutation environment. In a high mutation environment with a larger σ we can also see a similar behaviour as in a low mutation environment. Figure 9.3D shows $\sigma = 30$, where the same initial lag period before the models capable of replicating begin to fill the population as occurs in the low mutation environment. Once again the maximum number of replicating models is lower than in the low mutation environment. In some cases of larger σ values, no replicating models are able to establish themselves within the population, as shown in Figure 9.4.

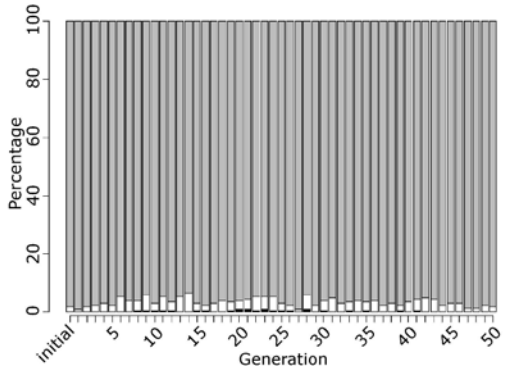

Figure 9.4: Population status each generation. Black replication, white stationary, grey death. High mutation environment and σ=50. Due to the large σ value, it is very difficult to generate replicating models, and so population consists mainly of dying models.

This change in behaviour can be explained in two ways. Firstly, the size of the σ binding affinity parameter determines the shape space complexity. In a low value shape space, there are effectively fewer shapes that each molecule can take, and so it is more likely to have an initial population with a number of models which have similar enough shapes to provide the required interactions and dynamics. With a larger value, the shape space increases in size, therefore molecule shapes have to be more accurate to achieve the same interactions and dynamics required. As we can see from the Figures low σ environments start with a larger number of replicating models up to 40% of the population, whereas in a larger σ environment it takes many generations of searching the shape space for a random model to survive and replicate well enough to start propagating through the population. Secondly, the mutation rate affects the rate of evolution within the population. In the low mutation environment, the population

quickly reaches equilibrium where around 95% of the population are replicating models. The low mutation rate also means that once a model has achieved the required interactions and dynamics to replicate, it is unlikely to mutate away from this state, and so we see only 5% of the population dies each generation. This death rate would also be due to stochastic affects, as there is a small probability that even the most highly optimised model could die. In a high mutation environment, we can see that the mutation rate of 50% is having a detrimental affect on the evolution. Whilst the initial population has a similar distribution of models as the low mutation environment, it takes several generations longer to reach equilibrium, and once it has reached this there is a larger proportion of dying models, as it is more likely that a model will mutate and lose required interactions.

9.6 Genome size in the population

Genome size was also recorded during evolution. In a low mutation environment, the average genome size is no larger than 1.3 genes per model, and larger models are quickly selected out, and a low equilibrium genome size is achieved within the population. Whereas, in the high mutation environment the average genome size quickly reaches around 1.7 genes per model, and again stays around this size. A small genome size is expected due to the simple environmental challenges requiring little complex regulation and also due to the selection pressure implicitly imposed from the replication criteria. Each gene requires an extra 10% energy of the replication threshold, which means that 'junk genes' will be detrimental to reaching its replication threshold. Results of evolutionary runs where the genome size was initially larger than 1 gene also show an average final gene size of slightly larger than 1, indicating that the extra genes are not likely to be required for efficient and fast replication.

9.7 Cell lineages in the population

Examining the final population of each evolution shows that the majority of all models in the population come from a single common ancestor. In the low mutation, low σ environment the final population usually consisted of two 'lineages', in one case the population was split approximately equally, however in another case 95% of models had the same initial ancestor. In all cases in this environment, all the lineages could be traced back to the initial population.

As the evolutionary environment may generate new random models if the population cannot be filled by surviving and replicating models, model ancestors may only be several generations old. Therefore, the 'generational age' of the lineage may imply the evolutionary fitness landscape ruggedness, and ease of randomly generating models initially capable of surviving the simulation. In a high σ environment we see a different pattern. Instead of multiple lineages competing for space in the population, we see a dominance of a single model. In two cases 100% of the final population consisted of models derived from a single model (emerging in generation 1 and 3), and the third case consisted of two lineages with offspring from one model from generation 19 contributing to 98% of the population, the other 2% were from a model from the initial population. In a high mutation environment we see different population distributions. In a low σ environment there is a mixture between complete dominance of a single model or partial dominance of one model against either one or two other models. However, in all cases, all the models trace back to the initial population. In the high σ environment a different population distribution is dominant. In each case the final population consists of over 50 lineages, each occupying only 1-10% of the population, but tracing back up to 20 generations.

Chapter 10

MODEL DISCUSSION

10.1 Limitations of model

The maximum genome size of the models is currently limited to six genes. This is limited due to computational requirements of the simulation algorithm (Gibson-Bruck) when simulating large genomes. Use of an alternative simulation algorithm such as the Gillespie algorithm [40] may reduce the computational requirements of larger genomes. A further limitation imposed by the computational complexity of the simulation paradigm is the length of time that can feasibly be simulated. With current evolutionary parameters (100 models per generation and 1 hour simulated time per model) it takes approximately one day to simulate 10 generations using a single CPU. Generating a simulation or evolutionary environment which utilises multi-CPU or GRID technology would be a relatively straight-forward modification that could dramatically increase the performance of the model and its evolution simulation.

Due to the modelling approach and necessary efficient simulation of the model, polymerisation is limited to generating complexes of up to three molecules. However, as previously noted, physical structures of the cell, for instance long polymers, are not modelled, and so this limitation does not negatively affect the model.

A further limitation is the simulation of the environment and evolution. A generational approach, using a genetic algorithm, does not accurately model a growing bacterial culture, in which cells would be dividing at different times. This could lead to some cells dividing several times, whilst others divide only once. A more realistic simulation and evolutionary environment such as spatial structure, may be included in future model formulations.

10.2 Future directions

10.2.1 Environments with increasing complexity

The results presented previously indicate that even in the simplest of evolutionary environments, we observe non-trivial and realistic behaviours and mechanisms, such as the evolution of rapidly turned over mRNA and repressor activity. However, the evolved network structures are relatively simple, which is an indication of the simplicity of a chemostat-like environment (an environment that free-living bacteria such as *E. coli* would not normally encounter, nor be adapted to). We predict that in increasingly complex environments, that would be more representative of evolutionary conditions in nature, the model would produce even more complex network structures and solutions. Due to the flexibility of the model, it will be straightforward to create more complex environments, each presenting different problems to be solved.

An example complex environment to investigate would consist of multiple food sources. *E. coli* is able to grow on a number of sugars; in the presence of multiple sugars it is able to selectively metabolise the most energy efficient food first, using regulatory mechanisms such as the *lac* operon. As the model already implements the *lac* operon 'transcription logic', it would be fair to assume that a similar switching mechanism requiring both activation and repression activity may evolve within an environment with two or more different food sources.

Another environment may consist of a food source that is varying in a predictable way, analogous to a day/night cycle. Organisms have evolved mechanisms for responding to these cycles, known as circadian rhythms, by developing 'circadian or biological clocks'. Prokaryotic circadian clocks, found within cyanobacteria such as *Synechococcus*, consist of only three genes *kaiA, kaiB* and *kaiC*, that are able to exhibit rhythmic behaviour [33]. The proposed regulation of the 'circadian clock' is a feedback loop by all three proteins, with unknown interactions between them, and both activation and repression of the genes [58]. Eukaryotes including mammals and plants have evolved more complex clocks that include multiple oscillating loops that are thought to provide robustness to noise and seasonal effects [53, 32]. This proposed feedback loop could be represented within our model and could produce a circadian clock, which is tuned to the availability of food.

Many organisms live in an environment in which food, or resources are limited and their availability to the organism may fluctuate. The organism therefore requires mechanisms to

optimally use these limited resources, for example the starvation response in *E. coli* governed by RpoS (σ^s, σ^{38}) [98]. It is predicted that our model would behave in a similar, albeit simpler, way to that of *E. coli* cells when faced with starvation. Upon detection of carbon starvation, RpoS up regulates the transcription of hundreds of genes which help to protect the cell against stresses, whilst down regulating hundreds of other genes. The cell enters 'stationary-phase' in which the cell has an increased chance for survival. Although RpoS is in fact a *sigma factor* which binds to RNAP helping it to recognise specific promoter sequences, our model could still simulate a similar mechanism. For instance, if the 'starvation' TF could bind to the enzyme's repressor site as an unbound monomer strongly, but when in a complex (with food) it is unable to bind, then the same response would be observed; when food is available the repressor site is unbound, allowing production of the enzyme, whereas if no food is detected by the 'starvation' TF, then enzyme production is prevented. Whilst the current model does not support σ factors, they can easily be incorporated by removing the specific RNAP molecule, and allow any protein with appropriate shape to function as an RNA polymerase.

Some exploratory simulations were performed in environments with randomly available food sources, and indeed a 'switching' mechanism was observed in a dual-function transcription factor/enzyme protein in several cases. However, many more simulations with similar food environments and analysis are required before any conclusions can be drawn from these results, and as such is left as future work.

The current formulation of the model, with its constant environment and generational population structure, is in some ways analogous to a chemostat. Further developments could include an explicit spatial structure, which could potentially lead to coexistence of different "species" [71]. In the current formulation, replication time is a fair measure of fitness, since in a chemostat the fastest growing bacteria will dominate the population; in a spatially explicit environment, this is not necessarily the case, and an alternative approach to fitness may be necessary.

10.2.2 Multi-generation fitness function

The current evolutionary framework does not take offspring into consideration, and as such 'selfish' solutions, which produce a very quick replication time for the parent model, but potentially unfavourable initial starting conditions for offspring, may be selected for. One alternative fitness function would assign fitness not only based on the result of the parent simulation, but also its

offspring and their simulation. This additional constraint of offspring survival is biologically relevant as cell line persistence is essential in evolution, and also may facilitate further exploration of the fitness and parameter landscapes within the model. A prototype multi-generational evolutionary framework was developed to investigate the effects of offspring fate on evolution. If the parent model achieved replication then instead of a fitness based on speed of replication being assigned, the offspring models were also simulated. This continues until either all models have died, or a fixed time has passed. The offspring receive protein and food molecules from the parent cell as in the original framework, which can lead to the situation where one or both offspring receive no protein or food and are therefore potentially disadvantaged. The fitness of the model is then determined by not only speed of replication (and therefore number of offspring), but also the number of surviving offspring. Preliminary experiments with this evolutionary framework indicated that different, and potentially less 'selfish' solutions were selected. However, very limited simulations were performed using this experimental framework, due to the performance issues discussed earlier, which are dramatically increased with the larger simulation time and population size. Further work into a multi-generation evolutionary framework, such as the one proposed, may yield interesting insights into potential evolutionary trade-offs between 'selfish' solutions and lineage propagation.

10.2.3 Molecule shape dimensionality and evolution

Our scalable 2D continuous shape space is a substantial simplification of the high dimensional protein shapes in real cells. Other models such as the model proposed by van Noort *et al* [122] and extended by Cordero and Hogeweg [29], and used in the 'coarse-grained model', use an even greater simplification, with a 1D discretised shape space, and yet are still able to produce complex and realistic networks. This indicates that a high dimensional shape space is not essential for the evolution of complex networks; however, an adequately large shape space is required. Future work could investigate the impact of shape space dimensionality on the evolution of complex networks.

10.2.4 Recombination

Recombination is essential for higher order eukaryotes, and is also thought to be a major source of genetic variation in primeval genomes [106]. Whilst modern day bacteria such as *E. coli*

and *Campylobacter jejuni* do not use sexual recombination, they do have other mechanisms for DNA exchange such as DNA-uptake, horizontal gene transfer via plasmids and phages (HGT) and internal genome recombination [94]. In the current model formulation, genes can only be transferred vertically (VGT), that is are passed from parent cell to daughter cell only. Future model formulations may include processes of DNA exchange between organisms such as HGT.

10.2.5 Potential network analysis techniques

The analysis of both the final networks, and the dynamics of the evolution of these networks are likely to become increasingly difficult as the complexity of the environment and hence networks increases. The analysis of biological networks currently suffers from a number of issues, such as obtaining networks from data and determining functionality of particular sections of the network, due to the size of the whole-genome networks and noise in the data collected [92]. The concept of 'network motifs' has been introduced to analyse the building blocks of complex networks as a way to elucidate function from the networks, and has been applied to several genomes including artificial networks [90, 12, 85]. Such an approach may be required to analyse the structure and function of the evolved networks from the model. Applying such a technique across several generations and large simulated evolutionary timescales may help to identify how and why specific network structures are evolved, which is currently not possible with laboratory experiments.

Analysing the evolutionary dynamics using techniques such as evolutionary activity statistics [14, 23] may provide valuable information and details about the evolutionary process, and may also highlight specific and important components of the system. Once these components have been identified, this information can be used along with traditional network analysis, such as 'network motifs', to help identify and separate functional modules within the networks.

10.3 Summary and conclusion

In this study a new model of evolving transcription control networks in prokaryotic cells was introduced. The model incorporates several novel mechanisms, realistic and evolvable parameters and a scalable level of complexity. The models are simulated using a stochastic framework, from which the dynamics of the model were investigated over a range of parameters. Several

key realistic network structures and model behaviours were observed and important parameters determining whether a model would replicate are presented and discussed.

Evolutionary runs of the models were performed using a standard genetic algorithm incorporating realistic evolutionary operators, in an idealised constant food environment. The initial and evolved networks are presented, as well as overall population dynamics. The results of these evolutions show that over the short evolutionary time-frame used, the models optimise their initial network configurations to produce a more robust and faster replication time, and a few novel network interactions were introduced. Several realistic behaviours emerged during the simulated evolution. A realistic cell replication time was evident, and the most efficiently replicating models consisted of a single gene, which controlled its own expression through a repressor mechanism, indicating a necessity to remove non-essential genes. This network structure (or motif) is prevalent in many instances in all organisms, and typically one of its purposes is maintaining protein levels. Realistic mRNA and protein degradation rates evolved, which also follow similar general principles found in *E. coli* and *S. cerevisiae* of typically up to several orders of magnitude difference between stabilities of mRNA to proteins.

The robustness of the evolved models was investigated using *in silico* genetics to produce mutant models consisting of various knock-outs and perturbations to molecule and complex stability. Many models were shown to be resilient against fairly large perturbations, however, the dynamics of the models after certain mutations such as regulatory site knock-outs were substantially changed, as would be expected from real cells.

The importance of energy and stochastic processes within gene regulatory networks is again highlighted in this model. Additionally, the importance of mRNA and protein stability was indicated in metabolism and gene regulation. The incorporation of protein-protein interactions and polymerisation had limited influence on the evolved networks, however, in increasingly complex environments these processes may become more essential as discussed.

The exploratory results presented in this study indicate the model allows reasonably realistic modelling and evolution of transcription control networks in an abstracted prokaryotic cell allowing complex behaviours to in simple environments, and provides the functionality to easily simulate more complex and biologically interesting environments.

Part IV

EXTENDED COARSE-GRAINED MODEL

This part contains an extended version of the coarse-grained model. Chapter 11 discusses limitations of the coarse-grained model and introduces the extended model, Chapter 12 presents a detailed analysis of deterministic and stochastic formulations of the model and its simulation and the consequences for realistic gene regulatory network modelling and Chapter 13 presents a detailed analysis of evolution in more complex environments and discusses the biological implications of the results obtained. The preceding chapters are formed from two manuscripts, "Stochasticity vs determinism: Consequences for realistic gene regulatory network modelling and evolution" and "*De novo* evolution of complex, global and hierarchical gene regulatory mechanisms", which are currently in submission to different journals. Chapter 14 discusses the limitations and future directions of the extended model.

Chapter 11

EXTENDED COARSE-GRAINED MODEL INTRODUCTION

11.1 Overview of changes to coarse-grained model

Whilst the original coarse-grained model was able to replicate many biological properties and was shown to be a suitable model for evolutionary experimentation, a number of important extensions and refinements were identified, from results obtained using the original 'coarse-grained' model and the 'fine-grained' model, to further accurately model biological systems. Firstly, the 'on-demand' protein production mechanism, whereby for each protein species binding to the DNA, the exact number of proteins is assumed to be produced, is inadequate to fully capture protein dynamics. Fluctuations in protein levels (due to stochastic expression and turn over) will mean that an inexact number of proteins are likely to be present in the cell (either too many, or too few), and so the assumption that the exact number of proteins is synthesised is incorrect. Results from the 'fine-grained' model indicated the importance of protein copy number in cellular function. As a result, the extended model incorporates a different representation of protein and protein levels. During simulation each protein species has a discrete abundance of molecules, which is determined by protein synthesis, turn over and binding status. Additionally, the cost of producing protein is increased with a cost for mRNA production as well as per protein, again reflecting transcription and translation within cells more accurately. Therefore, protein level is now a limiting factor, and must be adequately controlled by the model, as in biological cells. Further, proteins were assumed to only exist for a single time-step, meaning required

proteins would need to be constantly synthesised. Whilst some house-keeping proteins may be constantly synthesised and rapidly turned over, many proteins have a much slower turn over rate, and subsequently are synthesised far less frequently. The rate of protein synthesis is often a reflection of the protein's function. The results obtained from the 'fine-grained' model further underline the importance of protein stability in gene regulation. The original model is therefore unable to capture the protein turnover dynamics, and so is modified such that proteins in the extended model have evolvable stabilities.

A further limitation of the original model was the lack of interaction with the environment. 'Food' molecules were the only external source to the model, and as such interaction with the environment consisted of food availability, and lack thereof. Whilst carbon starvation is a severe stress which cells must be able to respond to, cells must also respond to many more environmental factors. As such, biological cells must be able to cope with a multitude of different stresses simultaneously, due to often highly variable, harsh and complex environments. As a result, a new environmental factor has been introduced to the extended model, 'environmental stress'. The new 'stress' factor introduces harmful molecules, which if not responded to adequately will cause the model to die. The harmful molecules can be thought of as, for example, denatured proteins after a heat-shock event, metal ions or other toxic molecules. The response required by the model is the removal of the stress molecules, for instance, by refolding of denatured proteins or excretion of toxic molecules, before a lethal level of the stress has accumulated within the model. However, due to the complexity of the response required, an additional energy cost is needed to perform the required processes, which can be thought of as synthesis of chaperone proteins, or transportation of the toxic molecules to the cell membrane and excreted by membrane proteins. The interaction between the model and the environment is dramatically increased and more complex, resulting in more realistic evolutionary environments. The stress concept is implemented within the model by the introduction of two additional 'specialised' gene types: stress receptor genes, *rcp*, which detect the presence of specific stress molecules, and stress response genes, *rsp*, which perform the required actions to remove the specific stress molecules.

11.2 Model components

Briefly, the model consists of a GRN which contains regulatory genes, which act solely as transcription factors (TFs) and a number of types of specialised genes, each performing a different cellular process. Energy signalling, consisting of genes, *nrgX*, responding to the levels of energy within the cell. These are analogous to unbound CRP in *E. coli* reflecting a high-energy state, and bound CRP-cAMP complexes reflecting a low-energy state. Metabolism, consisting of genes, *fodX*, reflecting the status of enzymes actively involved in sugar metabolism. Stress detection, consisting of genes, *rcpX*, responding to the presence of 'stress' molecules within the cell. Stress molecules can be thought of, for example, as toxic or heavy metal ions or denatured proteins due to heat or acid shock. Biosynthesis, consisting of genes, *synX*, reflecting the activation of biosynthesis pathways responsible, for example, amino acid synthesis. Stress response, consisting of genes, *rspX*, reflecting the removal of stress molecules by, for example excretion from the cell or protein refolding by a chaperone. These specialised genes can also act as TFs. Additionally, the specialised genes are classed as either 'input' (*nrg, fod, rcp*) or 'output' (*syn, rsp*) genes. The model also abstractly models ATP, nucleotides and amino acids as a single 'energy' molecule, represented as an integer value within the model. Energy, as in biological organisms, is essential for fuelling cellular processes; if the energy level falls to, or below, 0, then the cell dies. Biomass is modelled as an integer value representing how much biomass has been produced. Biomass, or yield, is an indicator of growth and as such is the primary measure of fitness of the cells. Stress molecules are modelled as integer values representing the quantity of each molecular stress species within the cell.

Genes

Each gene, g_i (with the exception of input genes), has an associated regulatory region consisting of a set, J, of binding sites, bs. Each binding site bs_j can be either activating, $r_{ij} = 1$, or inhibitory, $r_{ij} = -1$ and has an occupancy value, o_{ij}, which is 1 if bound and 0 otherwise. Activation, a_i of a gene is dependent on activating and inhibitory binding site occupation according to the formula:

$$a_i = \sum_{j \in J} r_{ij} o_{ij}$$

where i is gene, j is binding site, r_{ij} is binding site type (1 if activating; -1 if inhibitory) and o_{ij} is binding site occupancy (1 if occupied; 0 if unoccupied). Input gene activation is dependent on other cellular or environmental states. Each gene, g_i encodes a protein p_i product which has a number of parameters: shape s_i, an integer value from a 1D circular shape space (of size s^{max});mean protein production value, p^{prod}, an integer value determining the number of proteins produced per gene expression event; protein stability, p^{stab}, an integer value determining the stability of the protein before it passively degrades. Each binding site, bs_j also has a shape, s_j, parameter, drawn from the same shape space as protein shape.

11.2.1 Protein-DNA binding and affinity

Each gene and binding site has a shape, and the complementarity between the two shapes determines the binding affinity between a protein and a binding site. The 1D circular shape space was introduced by Cordero and Hogeweg [29], and provides an abstracted representation of protein binding domains and binding site structure. The binding affinity, b_{ij} between two shapes s_i and s_j is given by the following equation:

$$b_{ij} = \begin{cases} \frac{1}{d_{ij}+1} & if\ d_{ij} \leq d^{max} \\ 0 & \text{otherwise} \end{cases} \tag{11.1}$$

$$d_{ij} = \|s_i - s_j\| \tag{11.2}$$

where d^{max} is the largest integer Euclidean distance which two shapes can bind.

11.2.2 Input gene activation

The input genes each have specific criteria governing their activation state:

- Energy signalling is modelled as the activation of one of the energy signalling genes, *nrgX*, when the cellular energy level is above a given threshold, T_X^{energy} for the gene, X.
- Metabolic activity is modelled as the activation of one of the metabolic genes, *fodX*, when the cell has access to a given sugar or food molecule, and has generated energy. Each food or sugar within the environment has a unique *fodX* gene within the model. Further, each food species has an energetic value, allowing high- and low-energy sugars to be modelled.

- Stress detection is modelled as the activation of one of the stress receptor genes, *rcpX*, when any quantity of the specific stress molecules are present in the cell. Each stress, X, within the the environment has a unique *rcpX* gene and a threshold, T_X^{stress}, within the model.

11.2.3 Transcription, translation and basal expression

Transcription and translation are modelled as bursts, in which several mRNA or protein molecules are synthesised simultaneously, reflecting the experimental work of Cai, Freidman and Xie [21], with an energetic cost associated with the transcription (C^{mRNA}) and translation ($C^{protein}$) events. Any gene, g_i, whose activation level, a_i, is ≥ 1 (more positive bindings) will be expressed. Additionally, any gene with an activation of 0 (equal positive and inhibitory binding) can randomly express with a given probability, K^{basal}, representing random RNA polymerase binding events. Variable rates of basal transcription reflect abundance of RNA polymerase within the cell. The mean number of proteins (and subsequent energetic cost) generated is dependent on the protein production value.

11.2.4 Output gene expression

When output genes (*syn* and *rsp*) are expressed, additional processes take place:

- Biosynthesis (*syn* genes) - each *synX* gene has an associated biomass production value (P_X^{bio}) and an energetic cost (C_X^{bio}) for producing the biomass. The biomass produced is added to the existing cells produced biomass value.
- Stress response (*rsp* genes) - each *rspX* gene has an associated stress molecule removal value (P_X^{stress}) and an energetic cost (C_X^{stress}) for the removal. Each *rspX* activation removes a number of stress molecules of a unique stress species from the cell immediately.

11.2.5 Protein degradation

Protein degradation is a passive process, which is determined by the stability value of a protein, p_i^{stab}. This value represents the number of time-steps the protein remains stable and functional. Once a protein has degraded, it is removed from the cell.

11.3 Model simulation

The models are simulated using discrete stochastic Boolean networks. Simulation of each cell consists of a fixed number of time-steps, consisting of a sequential ordering of sub-steps, and and equal starting energy level:

1. Determine ordering of protein and binding site interactions: Each species of protein within the cell is selected to interact with binding sites in a specific order. The ordering of the binding sites is also specified each time-step. Each protein species will attempt to interact with unoccupied binding sites, until either no free protein molecules are available, or all binding sites have been selected. In the deterministic formulation, the ordering of both protein and binding sites is fixed, such that input genes and proteins are selected first, followed by regulatory genes and proteins and finally output genes and proteins are selected last. In the stochastic formulation, the ordering of protein and binding sites is randomised each time-step.

2. Determine activation status of input genes: *nrg* genes are activated based on specific energy levels, *fod* genes are activated based on environmental food availability and *rcp* genes are activated based on the presence of stress molecules.

3. Transcribe and translate input genes: Each input gene which was activated in step 2 is expressed, producing protein. In the deterministic formulation the number of proteins generated is fixed (to the genes protein production value, g_i^{prod}), and no basal expression can occur ($K^{basal} = 0$ in this case). In the stochastic formulation the number of proteins generated is a random normal, with $\mu = g_i^{prod}$ and $\sigma = 0.5$, rounded to the nearest non-negative integer.

4. Determine protein-DNA interactions: Using the protein and binding site order from step 1, protein-DNA interactions are determined. The binding affinity, b_{ij}, is used to determine if binding occurs between protein i and binding site j. In the deterministic formulation a protein will bind if $b_{ij} > 0$. In the stochastic formulation a protein will bind if $b_{ij} > R[0, 1)$ (where $R[0, 1)$ is a random value between 0 (inclusive) and 1 (exclusive)).

5. Determine activation status of regulatory and output genes: Activation, a_i, is calculated based on binding site occupancy, o_{ij}.

6. Transcribe and translate regulatory and output genes: Activated genes ($a_i > 0$) and basally expressed genes ($a_i = 0$ and $K^{basal} > R[0,1)$) are transcribed and translated (including output processes). In the deterministic formulation the number of proteins generated is fixed (to the genes protein production value, g_i^{prod}), and no basal expression can occur ($K^{basal} = 0$ in this case). In the stochastic formulation the number of proteins generated is a random normal, with $\mu = g_i^{prod}$ and $\sigma = 0.5$, rounded to the nearest non-negative integer.
7. All proteins unbind from the DNA ($o_{ij} = 0$) and all genes are inactivated ($a_i = 0$).
8. Determine protein degradation: In the deterministic formulation proteins are degraded based on the fixed number of time-steps each is stable. In the stochastic formulation proteins are degraded with the probability of $1/p_i^{stab}$.
9. Check simulation termination criteria: 1) required number of time-steps completed 'cell replicates', 2) energy level falls to, or below, 0 'cell death', 3) stress levels exceed specified threshold 'cell death'.

11.4 Evolutionary framework

The evolutionary environment is a genetic algorithm [54], which consists of a fixed-size population of cells. The initial generation consists of randomly generated cells, and is non-adaptively evolved for 10 generations to generate a more biologically realistic network (for evolutionary operators see below). A single offspring from each non-adaptively evolved initial network is generated, creating a population with twice the initial number of cells. Each cell within the population is then simulated independently, but with identical environmental conditions. After simulation, each model is assigned a fitness value, f_i:

$$f_i = \begin{cases} \text{biomass generated + time-steps} & \text{if cell replicates} \\ \text{time-steps survived} & \text{otherwise} \end{cases} \tag{11.3}$$

Each successive generation consists of the fittest 50% of the population (elitist selection). The selected models then each replicate once, creating a new generation of models. During the replication process both the parent and daughter cell can mutate (see below). The evolutionary processes is then repeated with the new population. Whilst no direct competition between cells is present (such as competition for food), space is a limiting factor within the evolution-

ary environment and as such introduces an indirect competition between the cells, generating evolutionary pressure.

11.4.1 Evolutionary operators

A number of evolutionary operators are defined at the individual gene level and genome level:

- Gene duplication: The entire gene (including protein parameters) and its regulatory region is duplicated and added to the genome with probability 1×10^{-3}. If either an input or output gene is duplicated then the gene and its associated parameters and regulatory region are duplicated, however, the duplicate gene does not function as an input/output gene.
- Gene loss: The entire gene and its regulatory region is removed from the genome with probability 1×10^{-3}. Input and output genes cannot be lost. This ensures a 'minimal' genome will always exist consisting of the initially defined number of input and output genes.
- Protein shape mutation: The protein shape is mutated by a random normal, with $\mu = 0$ and $\sigma = \log_{10} s^{max}$, with probability 5×10^{-3}.
- Protein production mutation: The protein production is mutated by a random normal, with $\mu = 0$ and $\sigma = 2$, with probability 5×10^{-3}.
- Protein stability mutation: The protein stability is mutated by a random normal, with $\mu = 0$ and $\sigma = 2$, with probability 5×10^{-3}.
- Binding site duplication: A random binding site from the genome is duplicated with probability 8×10^{-3}.
- Binding site loss: A binding site is lost with probability 8×10^{-3}.
- Binding site shape mutation: The shape of the binding site is mutated by a random normal, with $\mu = 0$ and $\sigma = \log_{10} s^{max}$ with probability 8×10^{-4}.
- Binding site regulation flip: The regulation type (activating or inhibitory) is flipped with probability 8×10^{-4}.
- Horizontal gene transfer: A portion of the donor genome is duplicated into the host genome with probability 5×10^{-5}.
- Basal rate mutation: The basal rate exponent ($log_{10} K^{basal}$) is mutated by a random normal, with $\mu = 0$ and $\sigma = 0.4$, with probability 5×10^{-3}.

Mutation parameter values (with the exception of binding site regulation flip, horizontal gene transfer and basal rate mutation) are taken from Cordero and Hogeweg[29].

11.5 Fixed model parameters

The extended model has a large number of parameters, which can be fixed or variable. A number of fixed parameters are common to all experiments, and are given in Table 11.1.

Parameter	Value
s^{max}	128
d^{max}	3
C^{mRNA}	3
$C^{protein}$	2
Initial genome size	32
Max initial r size	3
Max initial g^{prod}	8
max initial g^{stab}	3
Simulation starting energy	1000
Simulation time-steps	2000
Population size	100

Table 11.1: Fixed parameters of extended model

Chapter 12

STOCHASTICITY VERSUS DETERMINISM

12.1 Investigation aims

This investigation sets out to answer three questions: i) What are the paradigms under which artificial regulatory networks can be simulated and evolved to the biologically realistic fitness goal of growth and biomass production? We are specifically interested in comparing the classical deterministic Boolean network with a stochastic Boolean paradigm, with and without basal gene expression, and exploring the impact of realistic evolutionary simulation on network architecture under these paradigms. ii) Are these paradigms, together with the simple biological goal of biomass production, sufficient to produce a range of living solutions, or is a more complex environment necessary? iii) What are the interactions between a biologically realistic fitness goal and the evolutionary dynamics? In particular, does such a goal automatically entail a trade-off between maximizing growth and mutational robustness?

12.1.1 Experimental and environmental conditions

Internal and external environments of the *in silico* organisms were generated and simulated. These environments offer a range of simple problems to be solved: efficient growth with or without a single central carbon metabolism energy signalling system. A single, constant food supply is present within the environment at all times. Initial populations of organisms were randomly initialised and evolved both adaptively (use of a fitness function to determine selection of

cells each generation) and non-adaptively (cells are selected at random each generation). Separate populations for stochastic models with basal expression, stochastic models without basal expression and deterministic organisms were maintained during the evolutionary process. All organisms were recorded during evolution, allowing cell lineages and the evolution of individual genes within the population to be traced back to the origin population. The specific model and evolution parameters for this investigation are given in Table 12.1.

Parameter	Value	Note
K^{basal}	1×10^{-1} to 1×10^{-6}	Range of initial basal rates
F_1^{energy}	15	
P_1^{bio}, P_2^{bio}	50	
C_1^{bio}, C_2^{bio}	150	
T_1^{energy}	500	
Generations	1000	

Table 12.1: Model and evolution parameters for 'Stochasticity vs determinism' investigation

12.2 Results

12.2.1 Stochastic basal gene expression dynamics implicitly shrink genomes

Populations under all experimental conditions with a fixed rate of basal gene expression (1×10^{-2}) evolved minimal networks, that were much smaller both in terms of number of genes (nodes) and number of regulatory interactions (edges) than the initial and adaptively evolved stochastically simulated model populations and also all deterministically simulated, or stochastic networks without basal expression, model populations (Figure 12.1). This indicates large scale adaptation in the stochastic networks, caused by the basal gene expression.

A common solution was observed in many of the evolved populations with basal expression: strong activation of one or both biosynthesis pathways by the energy signal (Figure 12.2A). The network allows the organism to grow when the energy signal, *Nrg1*, is on, which positively regulates the biosynthesis pathways, *Syn1* and/or *Syn2*; when the energy signal is off, the biosynthesis pathways are unactivated and growth ceases, thus preventing expression when the cell has low energy.

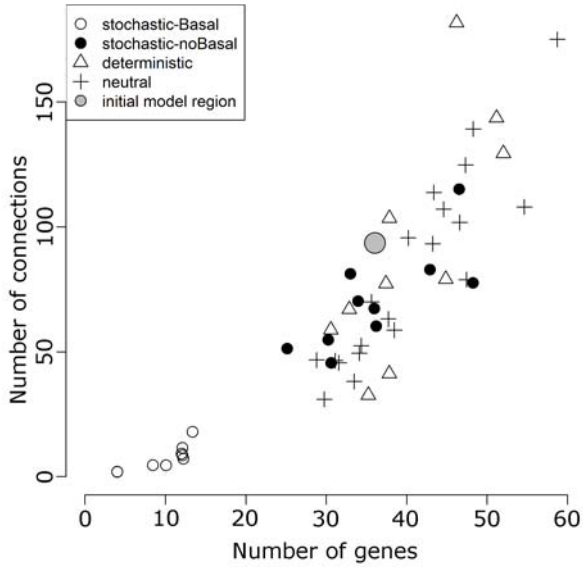

Figure 12.1: Genome size and number of regulatory interactions in 10 replicates of deterministic and stochastic (with and without basal expression) network evolved populations. The stochastic network populations with basal expression are closely clustered with small genome sizes and a low number of regulatory interactions, with a strong trend away from the non-adaptively evolved networks (two-sample t-test of genome: $p = 1.233 \times 10^{-7}$; interactions: $p = 2.2246 \times 10^{-4}$). Deterministically evolved populations are not clustered and do not show a consistent trend when compared to the non-adaptively evolved networks (genome: $p = 0.6823$; interactions: $p = 0.4307$). Stochastic networks without basal expression also do not show a trend (genome: $p = 0.1466$; interactions $p = 0.2862$).

12.2.2 Deterministic Boolean dynamics and stochastic networks without basal expression produce 'bloated' genomes

The topologies of one class of evolved deterministic Boolean networks and stochastic networks without basal expression solutions were very different to those with evolved with basal expression. The networks were much larger in terms of both genome size and regulatory interactions, yet were able to generate fitness values comparable with the networks with basal expression (Table 12.2). Figure 12.2B(i) and (ii) show examples of final deterministic network topologies. The highlighted area of the networks (dotted rectangle) contain the same solution observed in the networks with basal expression: activation of the biomass pathway by the energy signal. However, a large number of genes and regulatory interactions are also present, causing a 'bloating' of the networks. Dynamics analysis reveals that a large proportion of the genes in the deterministic networks were superfluous and could never be activated. The only active genes are those in the subset of genes present in the basal expression network solutions: energy signal and biomass

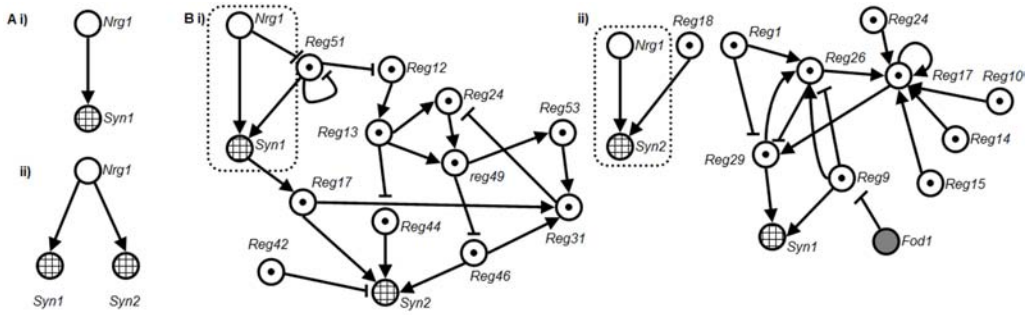

Figure 12.2: Minimal stochastic networks and 'deterministic bloat'. Nodes represent genes; $\rightarrow$ are positive interaction; $\dashv$ are negative interaction. All interactions (both in stochastic and deterministic networks) in these example networks have a binding affinity of 1. Networks A i) and ii) are examples of stochastic networks with basal expression after 1000 generations of simulated evolution. The simple solution of utilising the energy signal (*Nrg1*) to positively regulate one, or both, of the biosynthesis pathways (*Syn1* and *Syn2*) is evident in both networks. Networks B i) and ii) are examples of 'deterministic bloat' (network B ii) consists of two disconnected subgraphs). Here, the deterministic solutions are much larger, both in terms of genome size and regulatory interactions. However, the highlighted areas (dotted rectangle) show the only active network components allowing the cell to grow, which is the same as the solution evolved under stochastic dynamics. The remaining portion of the networks are never activated and so superfluous to the regulation of the biosynthesis pathways, but due to the deterministic dynamics they are *fitness neutral* and do not infer a penalty to the fitness of the cell. Therefore, there is no evolutionary pressure to remove them, and so leads to the characteristic 'deterministic bloat'.

pathways. Networks resulting from stochastic dynamics without basal dynamics displayed a similar 'bloating'. Thus, despite the large difference in network size and connectivity, the core solution of the deterministic networks is the same as in the basal expression networks.

The evolution of genome size and their fitness in example stochastic models with basal expression and deterministically evolved model populations is shown in Figure 12.3.

In the first 1000 generations, the networks evolved stochastically with basal expression show the rapid loss of excess genes and regulatory connections, while the deterministic solutions grow and bloat. Yet, both cells are able to obtain equivalent fitness values. If the simulation environment is "switched" from stochastic to deterministic dynamics after 1000 generations, the network very quickly starts to become 'bloated', without loss of fitness (the slight increase in fitness observed can be attributed to the stochastic nature of gene expression, where a gene may not express even under identical conditions). Conversely, when the dynamics are switched from deterministic to stochastic, the networks are initially unable to survive, since basal gene expression in the bloated networks uses energy; slimmer networks are rapidly selected for within the population, and efficient solutions are evolved, due to the additional pressure of basal expression.

	Fitness		Genome size		Network edges	
Population	Stoch	Det	Stoch	Det	Stoch	Det
1	6750	9950	12	46	9	188
2	7050	9950	15	38	17	79
3	9700∗	9950	4∗	31	2∗	61
4	7050	9950	11	35	6	36
5	9700∗	9950	4∗	52	2∗	123
6	7350	9950	10	46	3	86
7	9600∗	9950	4∗	33	2∗	62
8	6800	9950	12	52	10	153
9	6500	9950	13	38	8	39
10	7400	9950	8	37	4	97

Table 12.2: Fitness values, genome sizes and number of regulatory interactions for best individual in evolved populations. Ten populations of stochastic (with fixed level of basal expression) and deterministic models were independently evolved. Three of the stochastic populations evolved the fully optimised 'high-yield' network (denoted by $*$), and several others had a partially optimised 'high-yield' network. The fitness of the 'high-yield' solutions are very close to the deterministic populations, even though the genome sizes and regulatory structure are very different.

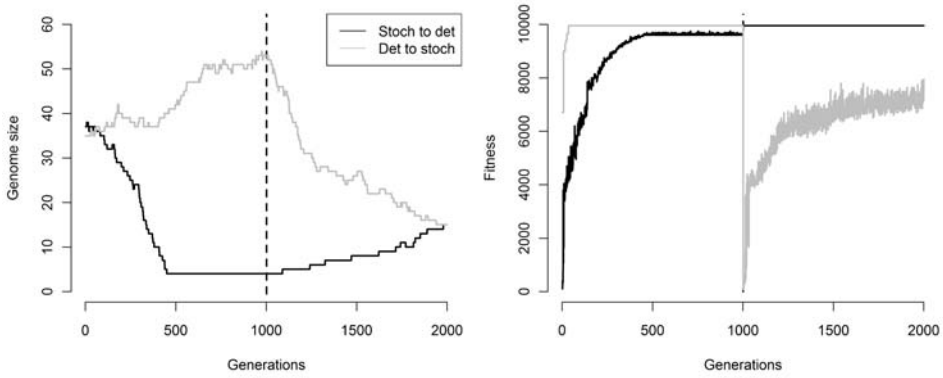

Figure 12.3: Evolved stochastic (with basal expression) and deterministic models have different network sizes, but similar fitness. Populations of stochastic and deterministic cells (with an energy signal) were evolved for 1000 generations. After the evolution period, a population derived from the fittest individuals from each population is evolved for a further 1000 generations using the alternate simulation dynamics. The left panel shows the genome size evolution and the right panel shows the fitness value of the model during evolution. The genome size during the first 1000 generations of evolution shows a rapid decrease in genes in the stochastic model, to a minimum size of 4 genes, but a large increase in the deterministic model, reaching over 50 genes, nearly twice the initial size. The fitness values for both models very quickly reaches an optimal value around 9000, despite the large difference in genome size. However, when the simulation dynamics are changed after 1000 generations, the model which was initially evolved stochastically starts to accumulate 'bloat', ending with a genome size threefold larger than starting, yet the fitness is unaffected. 'Deterministic bloat' therefore, will occur even from an optimised, minimal solution. Conversely, the model initially evolved deterministically, is unable to survive when changed to stochastic dynamics. Very quickly the genome size decreases, whilst the fitness dramatically increases. The model is therefore able to recover by removing detrimental genes and interactions.

12.2.3 Novel solutions only found with stochastic dynamics

When an energy signalling mechanism is not available (*nrg1* is knocked-out), populations evolved with deterministic dynamics were unable to evolve a solution to allow survival and growth within 1,000 generations. Stochastically simulated populations (with or without basal expression) were able to evolve solutions within several hundred generations. Solutions can also be observed within some stochastic (with or without basal expression) energy signalling populations that cannot be observed in the deterministic populations with energy signalling, indicating a limitation to the deterministic paradigm, rather than the lack of energy signalling.

Switching simulation dynamics during evolution, from stochastic to deterministic, affects not only the network structure, but also its ability to grow (Figure 12.4), generating unfit solutions. Models which have evolved using deterministic dynamics are unable to grow. Yet, when these models are simulated under stochastic dynamics, they are initially unable to survive (due to the 'deterministic bloat'), but within 100 generations more efficient slimmer networks are selected for. Stochastic processes are therefore essential for specific realistic solutions to function.

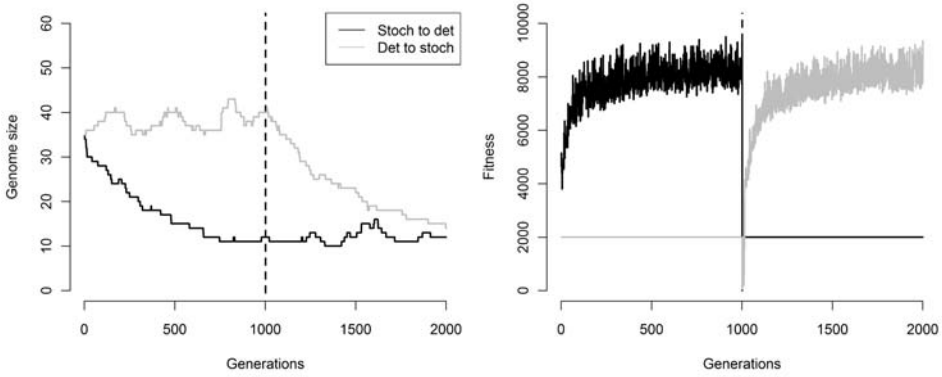

Figure 12.4: Deterministic dynamics are unable to discover solutions in specific environments. The experimental conditions from Figure 12.3 are replicated for populations without energy signalling. Genome size evolution is shown in the left panel and the corresponding fitness values in the right panel. Populations initially evolved with stochastic dynamics are able to discover efficient solutions within 100 generations. The lineage shows a rapid decrease in genome size and rapid increase in model fitness. Within 1000 generations, the lineage reaches a minimal genome size of 12 genes and a fitness of around 8000. After 1000 generations, the dynamics are switched from stochastic to deterministic or vice versa. Once using deterministic dynamics the model survives simulation, but does not biosynthesize. This fitness remains stable until the end of the evolution, but the genome size fluctuates. Whilst in the population initially evolved with deterministic dynamics the genome size remains large, and is unable to produce a network capable of survival and growth. When the dynamics are switched to stochastic, the genome size starts to steadily decrease, whilst growth rapidly increases. Within 1000 generations, an efficient network is obtained.

12.2.4 Multiple attractor states in basal expression rate evolution under identical conditions

Energy-signalling populations which were able to evolve their rate of basal expression typically reached one of two attractor values, largely determined on initial starting rate. Figure 12.5A indicates the first attractor state, with a high level of basal expression ($\approx 10^{-2}$), generates much smaller genomes, yet less fit individuals than the second attractor state, with very little basal expression ($< 10^{-5}$). Populations that evolved to high basal rates have a fitness below 8,000, whereas, the populations evolved to very low basal rates have a fitness of nearly 10,000.

A number of viable solutions can therefore exist within the same environmental conditions. Populations evolved without energy-signalling did not show the same bimodal distribution of basal rates. Instead, a larger range of basal rates were evolved. As a result a large number of networks had a basal expression rate which led to a slight bias against 'bloating', and as such slightly smaller genomes by around 10% on average (≈ 28 compared with ≈ 30). The basal expression mechanisms are therefore also influenced by the environmental conditions.

Due to the large range of final basal expression rates in the populations without energy signalling, the final values may not equate to specific attractors, but are the result of 'glassy' behaviour. The spurious attractors, analogous to 'glassy' local minima energy states, capture the evolving parameters in local maxima fitness states, but do not correlate to basal rates representing global maxima, or high model fitness. Further analysis of the parameter fitness landscape is therefore important in determining the true attractor states (if any).

12.2.5 Different solutions can evolve under the same environmental conditions with stochastic dynamics

A number of network topologies were observed under the different experimental conditions and simulation dynamics. All the evolved deterministic networks share the same underlying solution. Yet, the actual networks vary hugely in their size and structure. However, the evolved stochastic networks can display a variety of solution types within the same experimental conditions. In populations with an energy signal and fixed basal expression rates (ranging from 1×10^{-1} to 1×10^{-6}), three classes of solutions were observed. The first is the simple solution previously described, involving positive regulation of the biosynthesis pathways (Figure 12.2A).

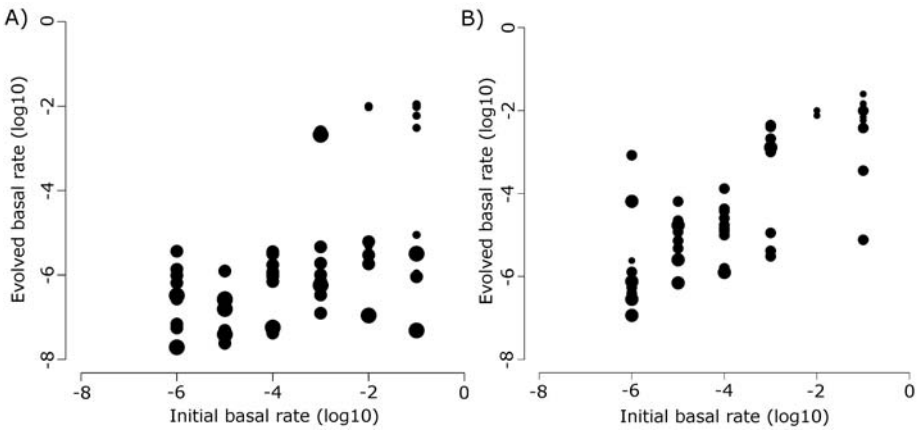

Figure 12.5: Attractor states in evolution of basal expression rates. Populations of models start with basal transcription rates ranging from 1×10^{-1} to 1×10^{-6} either with an energy signalling gene (A) or without (B). Each initial rate was replicated 10 times, and the fittest final individual is represented. The size of data point indicates genome size (< 20, $20 - 39$,> 40). Two attractor basal rates exist in the energy signalling populations: a high basal (1×10^{-2}), but lower yield and smaller genome solution, and a low basal ($< 1 \times 10^{-5}$), higher yield and larger genome solution. The no-energy signalling populations do not show a bimodal distribution of evolved basal rates, instead a larger range of rates is observed (from 10^{-1} to 10^{-7}).

This solution generates high levels of biomass, and so is a 'high-yield' solution. This solution was observed in both networks with and without basal expression. The second class of solution is very different, relying on basal transcription for activation, rather than the input energy or food signals. Example topologies are shown in Figure 12.6 and Figure 12.7 show further examples which also demonstrate the presence of non-adaptive gene duplication events.

Here, the networks have a larger number of genes and larger number of regulatory interactions. The biosynthesis pathways are more heavily regulated, weakly activated by a number of TFs. The number of TFs interacting with the biosynthesis pathways is dependent on the rate of basal expression; higher levels of basal expression have fewer activating TFs than lower levels. This solution appears to act as a 'noise filtering' mechanism, generating a steady signal from noisy levels of gene expression. This solution is only observed in networks with basal expression.

The third class of solution is observed only in networks with very high basal expression. In this case, the network relies on large-scale repression by either an input gene, or other TF. If an energy signal is present, then this is often utilised as in the 'high-yield' solution indicating some hybrid solutions are viable. Again this solution is only observed in the networks with basal expression. Interestingly, the three types of solutions have different biomass yields, with the simple high-yield solution able to generate biomass growth levels approximately 33% higher

on average than the filtering and repression solutions (see Table 12.3).

Multiple solutions were also observed in populations without an energy signal, which are all related to the solutions observed in energy signalling populations. A 'high-yield' solution, using the food signal (*fod1*), rather than the energy signal is evolved in no basal expression networks.

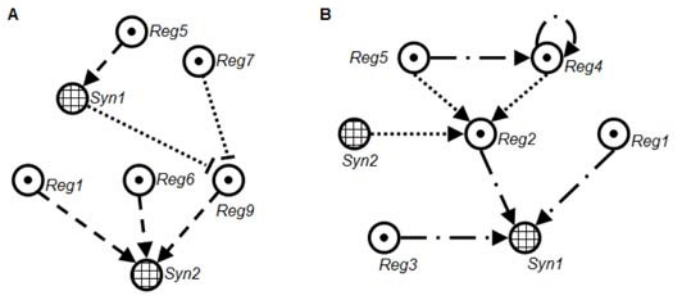

Figure 12.6: Heterogeneity of 'filtering' solutions. Nodes represent genes; $\rightarrow$ are positive interaction; $\dashv$ are negative interaction. Weighting of line represents binding affinity: — = 1, - · - = 0.5, - - - = 0.33, $\cdots$ = 0.25. Two examples of the 'filtering' solution observed in the stochastic populations showing no utilisation of the energy signalling mechanism and a larger number of interactions regulating the biosynthesis pathways. The large number of weak, positive regulatory interactions on the biosynthesis pathways act as a filtering mechanism of the noisy input signals from the basal expression levels of the TFs. A number of variations of the 'filtering' solution, utilising more or fewer TFs were observed in the no-signalling populations. This second type of solution is not present in the deterministic populations, due to the lack of basal gene expression.

	Fixed K^{basal}		
Population	1×10^{-1}	1×10^{-2}	none
1	$7750^{\dagger\circ}$	7500^{*}	$9700^{\dagger}$
2	$7300^{\dagger\circ}$	7350^{*}	$9700^{\dagger}$
3	$8200^{\dagger\circ}$	$9700^{\dagger}$	$9700^{\dagger}$
4	$8950^{\dagger\circ}$	7800^{*}	$9550^{\dagger}$
5	5650°	$9700^{\dagger}$	$9550^{\dagger}$
6	7050°	7400^{*}	$9700^{\dagger}$
7	$8500^{\dagger\circ}$	$9750^{\dagger}$	$9700^{\dagger}$
8	7200°	7550^{*}	$9750^{\dagger}$
9	7000°	7500^{*}	$9750^{\dagger}$
10	$8550^{\dagger\circ}$	7650^{*}	$9700^{\dagger}$

Table 12.3: Fitness and network solution in stochastic models with different fixed basal expression rates. † is high-yield, $*$ is noise-filtering, and $\circ$ is repressing network solution (multiple symbols indicate a hybrid solution). The increased biomass yield is evident in high-yield solution, with up to 33% larger yields than the other network solutions. The level of basal expression is a dominant factor in which solutions are selected.

This solution utilises the stochastic nature of protein production to infrequently produce activator proteins from the always-on signal. Consequently, this solution, whilst not utilising basal expression, is not possible within the deterministic system, due to the lack of stochasticity

within protein production. A 'noise-filtering' solution is observed in networks with moderate basal expression, but with high basal expression, the 'repression' solution is observed.

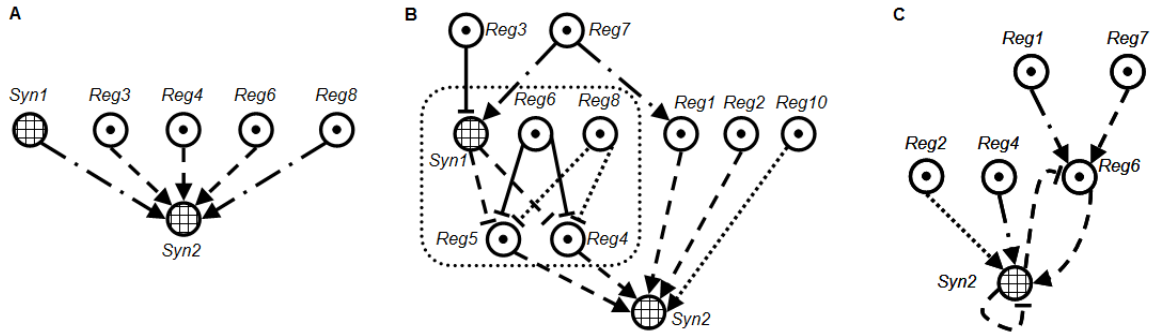

Figure 12.7: Further heterogeneity of 'filtering' solutions evolved in the stochastic populations, with differing genome sizes and large variations in number and type of interactions. The solutions can also contain 'bloat', as can be seen in the highlighted region on network B. This 'bloat' or redundancy in the network is due to two recent gene duplication events. Both artificial and biological networks are constantly evolving, only providing a snapshot at the current time, and so some potentially *fitness negative* parts may not have been removed. However, this should not be confused with 'deterministic bloat' which is permitted by *fitness neutral* interactions.

When the basal rate is evolved in energy signalling populations, only two solutions are observed (as indicated by the two attractor states in figure 12.5A). The 'high-yield' solution is only observed in the no basal expression attractor, whilst the basal expression attractor corresponds to a 'noise-filtering' solution. Whereas, in populations without energy signalling, despite the lack of dominant attractor states, two solutions are also observed; 'high-yield' in low/no basal expression networks, and 'noise-filtering' in networks with higher basal expression.

12.2.6 Evolutionary trade-offs: Yield vs efficiency vs robustness

The existence of two distinct classes of networks with different biomass growth rates can be explored using a combination of *in silico* genetics and evolution. Sixteen mutants were created from both the high-yield network (Figure 12.2A(i)) and the noise filtering network (Figure 12.6A), including binding affinity mutants, single and multi-gene knock-out and knock-in mutants (see Tables 12.4 and 12.5). Clonal populations of the 'wild-type' (WT) and mutant models were generated and simulated. Biomass generation (yield) and survival rate (efficiency) were averaged over each population (Figure 12.8). The high-yield solution is able to generate large yields, but relies on a very constrained network topology, namely the presence of an activating interaction from the energy signal to the biosynthesis pathways. A single mutation removing

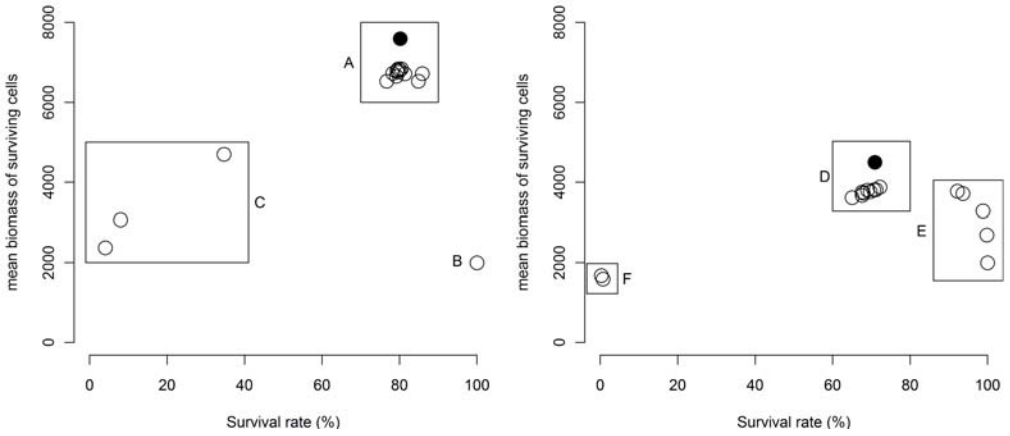

Figure 12.8: Survival rate (efficiency) and growth (yield) of wild-type and mutants of 'high-yield' (left) and 'filtering' (right) solutions. • are wild-type cells; ∘ are knock-out/knock-in mutations. In the high-yield solution, the high biomass growth levels in the wild-type cell is evident. The knock-in mutations of genes and regulatory interactions activating the biosynthesis pathways cause little difference to either survival (around 5% decrease in cell survival) or biosynthesis (around 10% decrease) (cluster A). This indicates that the regulation strategy is robust against duplication and gain of regulatory interactions. However, the single knock-out mutant (B) produces a dramatically different result; the survival rate increases to around 100%, but biosynthesis drops to the basal level of biomass production. Cluster C shows knock-in mutants of higher production and stability genes, similar to those found in the filtering solution. The network becomes both inefficient and low yield, indicating that a combination of the two solutions is not viable. The filtering wild-type solution shows a lower level of biosynthesis production, but similar survival rate compared to the efficient solutions. The solution is again robust against most knock-in mutations (cluster D). The filtering solution is also more robust against knock-out mutations, as evident by the larger number of knock-out mutants that have biomass productions above the base level (cluster E). The network is able to survive losing several genes, whilst still maintaining a higher than basal level of biosynthesis production. Cluster F shows the two knock-in mutants, with an activating interaction from the energy signal to each biosynthesis pathway, which have very low survival rates ($< 5\%$) and low biosynthesis production. The network is also unable to survive using a combination of both solutions. The two solutions are, therefore, local maxima in fitness space.

this interaction produces a network with a lower growth rate than the filtering networks. The yield of the knock-out mutant is equivalent to a model which has no regulation, and is the base level of growth under stochastic dynamics (around 2000 with $K^{basal} = 1 \times 10^{-2}$). The filtering networks, however, are able to sustain a number of gene or interaction knock-outs, and still generate a substantial amount of biomass. Evolution therefore seems to have a number of solutions available to explore: a greedy, 'all or nothing', high-yield solution, but that is also highly susceptible to mutations; or a lower yield solution, that is more robust to mutation. However, mutational robustness is likely to be a selective pressure only in high-mutation environments, and is therefore unlikely to be a adaptive property of the solutions. Each evolved population

exhibited only one of these solutions, possibly due to historical contingency, as each of the two classes of solutions represents a local maximum in the fitness space. The level of basal expression is a dominant factor in what hybrid solutions are therefore viable. In direct competition simulations, in which the initial population is seeded with equal numbers of models from each of the two classes of solutions, the high-yield solution always out-competes the filtering solution.

Interestingly, in both solutions, the survival rate (efficiency) of the wild-type cells is only around 70-80%, yet several other solutions are easily obtainable (one or two mutations away) which provide a much higher chance of survival. In the 'high-yield' solution, a mutant with reduced binding affinity increases the efficiency from 80.2% (WT) to 85.9%, but reduces the average yield from 7593 (WT) to 6719. Similarly, in the 'filtering' solution, the WT efficiency is 70.9%, however, a single regulator knock-out mutant has a much higher efficiency of 93.6%, but with an average yield reduced from 4496 to 3869. Whilst yield is the primary selection pressure, efficiency may be a small selection pressure due to the limited population size. However, the model is unlikely to benefit with a much higher efficiency, and it only needs to ensure survival a large proportion of the time. Therefore, it appears that the evolved solutions support an evolutionary trade-off between efficiency and yield. Mutants combining the two classes of solutions are not viable (clusters C and F in Figure 12.8). Complete efficiency and yield results from each mutant can be found in Tables 12.4 and 12.5.

12.2.7 Effects of different mutation rates

A number of mutation rates were used to evolve populations. All stochastic populations (with $K^{basal} = 1 \times 10^{-2}$) in an environment with a mutation rate 10 times the default, evolved an optimised 'high-yield' network. In populations with mutations rates 100 times the default, a weaker activation by the energy signal is evolved, but is still characteristic of the 'high-yield' solution. The deviation from the optimised high-yield solution may solely be due to the increased mutation rate causing mutations to the energy signal interactions. The 'filtering' network was not observed in the replicates in either increased mutation environment. Therefore, increasing the mutation rate of the evolutionary environment, somewhat surprisingly, does not appear to favour the selection of the more robust filtering solution.

Mutant	Efficiency (%)	Yield
WT	80.2	7592.58
Nrg1 KO	100	1993.3
Nrg1-Syn1 affinity 0.5	81.4	6714.1
Nrg1-Syn1 affinity 0.333	85.9	6718.75
Nrg1-Syn1 affinity 0.25	84.9	6528.05
Syn1 single-KI	79.7	6836.7
Syn1 double-KI	80.1	6805.4
Syn1 triple-KI	79.4	6790.25
Syn2 single-KI	80.5	6838.55
Syn2 double-KI	79.4	6798.95
Syn2 triple-KI	79.7	6753.55
Syn1,2 single-KI	78.3	6727.8
Syn1,2 double-KI	79.2	6654
Syn1,2 triple-KI	76.7	6524.75
Syn1 single filtering-KI	34.7	4697.8
Syn1 double filtering-KI	8	3062.05
Syn1 triple filtering-KI	4.1	2369.7

Table 12.4: Wild-type and mutants derived from 'high-yield' network. Mutants consist of either a gene knock-out (KO), gene knock-in(KI), or binding affinity stronger/weaker. Default KI mutations consist of a regulatory gene with parameters: stability = 1; production = 1; affinity to existing gene = 0.5. 'Filtering' KI mutations consist of an average regulatory gene from the noise filtering solution with: stability = 2; production = 2; affinity to existing gene = 0.333.

Mutant	Efficiency (%)	Yield
WT	70.9	4496.33
Reg1 KO	93.6	3713.3
Reg1,6 KO	98.8	3283.45
Reg1,6,9 KO	99.8	2688.05
Syn1 KO	92.2	3773
Reg1,6,9,Syn1 KO	100	1991.15
Syn1 single-KI	71.2	3809.2
Syn1 double-KI	69	3794.75
Syn1 triple-KI	69.7	3757.65
Syn2 single-KI	72.2	3876.15
Syn2 double-KI	70.6	3807.7
Syn2 triple-KI	67.9	3721.2
Syn1,2 single-KI	67.6	3752.1
Syn1,2 double-KI	65	3610.55
Syn1,2 triple-KI	67.6	3668.2
Nrg1 to Syn1 KI	0.3	1672.15
Nrg1 to Syn2 KI	0.7	1587.65

Table 12.5: Wild-type and mutants derived from 'noise filtering' network. Mutants consist of either a gene knock-out (KO), gene knock-in(KI). Default KI mutations consist of a regulatory gene with parameters: stability = 1; production = 1; affinity to existing gene = 0.5. In the case where the biosynthesis gene, *Syn1*, is knocked-out, the protein shape is modified to remove any interaction, but the gene is maintained to ensure the same base level of growth

12.3 Discussions

This study sets out to answer three questions about the evolution of transcription regulatory networks: i) How do the different simulation paradigms of classical deterministic and stochastic Boolean networks, with and without basal expression, and realistic biological goals impact on the structure of *in silico* networks? ii) Are these simulation paradigms, combined with the realistic, but simple goal of biomass production, sufficient to generate a variety of solutions? iii) What are the interactions between a realistic fitness goal and the evolutionary process; in particular, will a trade-off between growth and mutational robustness emerge?

The combination of stochastic Boolean network dynamics, with realistic cellular processes such as basal gene expression, generate networks which are inherently minimal and efficient. In contrast, the networks observed in models evolved using the classical deterministic Boolean network dynamics or stochastic networks without basal expression displayed 'bloating', with an accumulation of superfluous genes and regulatory interactions, which do not have a role in the actual network solution for generating biomass. The explanation behind the different network sizes and topologies in 'high-yield' solutions can be found in the network dynamics; using deterministic Boolean network dynamics, genes can only be expressed due to direct activation from a TF, yet any unrepressed gene can express due to basal transcription with stochastic dynamics. The lack of basal transcription means that large parts of the genome, whilst potentially highly connected, will never be expressed and therefore never use energy. These regulatory interactions and genes are therefore *fitness neutral*. However, when basal transcription is introduced, all genes and regulatory interactions will no longer be fitness neutral and there is implicit pressure on the regulation of energy levels within the cell. The evolutionary process therefore has a number of potential solutions to solve this additional energy pressure: 1) removal of unrequired regulatory interactions, 2) removal of unrequired genes and 3) additional repressive regulation of genes. Indeed, it is evident in the stochastic networks that mutations removing interactions and genes have been fixed, as most regulatory interactions and genes can have a negative affect on fitness. Whilst classical deterministic Boolean dynamics can evolve efficient and realistic solutions, the network architectures are not realistic due to the 'bloat'. The stochastically evolved minimal networks are therefore a closer parallel to actual biological networks and so should be a consideration in future gene regulatory network modelling. Therefore we would conclude that

in trying to simulate evolution of GRNs towards biological goals, it might be inappropriate to use a deterministic Boolean network paradigm, but reasonable to use a stochastic variant.

However, it is important to also recognise that there may be alternative approaches to resolve the 'bloating' problem. For example, an ODE-based simulation, with a sigmoidal transcription response function, would allow basal transcription under an alternative deterministic paradigm, and thus potentially lead to a selection against non-functional elements. However, such ODE dynamics are themselves an abstraction of stochastic, leaky basal gene expression and so the approach we describe is both sufficient and more true to biology. Another approach to resolve bloat would be to incorporate a fitness penalty associated with genome size.

Despite the large difference in network architectures, the deterministic and stochastic networks, in many cases, share the same underlying solution. The 'high-yield' solution, consisting of strong positive activation of one or both biosynthesis pathways by the energy signal, is able to generate very high levels of biomass under both paradigms. This regulation strategy is also often observed in biological organisms; for example the CRP-cAMP global regulator in *E. coli* [94] positively regulates many biosynthesis pathways as shown in the EcoCyc database [69], or the CcpA global regulator in *Bacillus subtilis* [118]. The 'noise-filtering' solution, which provides efficient growth without an energy signalling mechanism, shares several mechanisms with organisms that exist in constant food environments, such as the endosymbiont *B. aphidicola*, which utilises basal gene expression in many pathways [109].

The evolution of the basal expression rate revealed two attractors - one with and one without basal expression - each generating solutions using different aspects of the network and processes. However, the extent and mechanism by which the low/no basal expression attractor can be achieved by evolution is debatable. On the one hand, the rate of basal expression is mediated by a number properties, for example, the availability of RNA polymerase, or promoter sequence modification to produce a stronger binding to increase expression, or a weaker binding to decrease expression. While evolution can act to reduce that availability, the reduction is constrained by the need for RNA polymerase for expression of other genes, and polymerase cannot be excluded altogether. On the other hand, cells have evolved many mechanisms to repress gene expression, including global repressors in prokaryotes [130] and histone modifications in eukaryotes [5]. Nonetheless, the evolution of different basal transcription rates in our model in different environments suggests that tuning RNA polymerase levels and/or global regulation

may be an approach for an organism to evolve appropriate and efficient regulatory networks.

The ability to evolve GRNs to biologically realistic goals is extremely important in understanding the relationship between topology and function. For example, Shen-Orr *et al.* [107] claimed that Feed-Forward Loop (FFL) motifs are over-represented, using a modified Erdos-Renyi random graph model, and so proposed adaptive functions such as noise filtering for these motifs. However, the model proposed by Shen-Orr *et al.* has been shown to be a poor description of genome architecture [13].

Subsequent work demonstrated that in non-adaptively growing networks, but with realistic growth mechanisms, FFLs would be highly represented by chance [76, 29]. Here, with a realistic biological goal, but the absence of a noisy environment, such motifs are not observed. This might indicate that the functional adaptation suggested by Mangan and Alon [85] might be correct after all. However, due to the small networks evolved, network motif analysis was not possible, and so remains an open question.

Some heterogeneity of solutions has been observed in a number of 'digital organisms' such as Avida [96, 80] and the work by François and Hakim [36]. However, no clear conclusions can be drawn as to what are the conditions, either experimentally or environmentally, to facilitate the rise of this range of solutions. Nor can the huge biodiversity we see be replicated to a sufficient extent *in silico*. The observed classes of solutions are very different to each other, each utilising different input signals. Whilst individual populations exhibited only one class of solution, the range of solutions between populations was very high. Noise filtering solutions displayed a wide range of regulation mechanisms, yet were able to generate biomass to equivalent levels as each other. Hybrid cross-class solutions were also not viable. The introduction of stochasticity to traditional deterministic Boolean networks increases the solution space available to the evolutionary process, provided by additional 'parameters' available to the evolutionary process. In particular, stochastic binding interactions and the introduction of basal gene expression allow novel solutions to be evolved which are unreachable with purely deterministic dynamics. We have therefore shown that diversity can emerge even in very simple environmental conditions, provided a stochastic simulation paradigm is used. However, much more work is required into inferring the specific processes or conditions required to generate diversity.

Investigation into the evolved networks indicated that a small trade-off between biomass production and efficiency had occurred throughout the evolutionary process, due to the presence

of models capable of increased efficiency, but reduced yield. Therefore, the use of a biological fitness function strongly affected the evolutionary process. The filtering class of solutions was more robust to mutations, capable of surviving several gene knock-outs, yet, in direct competition simulations, was consistently out-competed by the high-yield solution, even in environments with very high mutation rates. However, the chemostat-like environment used to evolve the populations promotes a 'survival of the fastest' regime, in which the organisms that can replicate fastest (or specifically within the model, grow fastest) will ultimately prevail. Biological organisms do not have such a strong 'survival of the fastest' pressure, nor such a rigid competition environment; many other pressures exist, which will ultimately affect the competitiveness and survival chances of a species. Introduction of a more realistic environment may promote mutational robustness as a stronger evolutionary pressure and may also allow the coexistence of multiple network solutions.

This study investigated a number of important questions in both gene regulatory modelling and evolutionary biology. Stochastic dynamics, along with a realistic fitness goal and basal gene expression, have been shown to be able to produce biologically realistic gene regulatory network mechanisms and architectures, whilst deterministic dynamics or networks lacking basal gene expression are unable to generate realistic architectures, but can produce appropriate mechanisms. The introduction of energetic cost for biological processes, and basal gene expression, facilitates the generation of realistic networks. Further, the basal expression mechanisms are influenced by environmental conditions. A variety of solutions was observed, in stochastically evolved populations, from identical environmental and experimental conditions. Whilst each independent population displayed only a single solution, there is a large difference between the independent solutions. Further, evolutionary trade-offs between growth rate and efficiency were observed, indicating the strong effect of realistic goal-based evolution.

Chapter 13

EFFECTS OF ENVIRONMENTAL COMPLEXITY ON EVOLUTION

13.1 Investigation aims

Richard Lenski and colleagues' work with the Avida model [96] has shown the adaptive evolution of complex features within 'digital organisms' arises in an incremental fashion [80]. These results show that increasingly complex functions were built from simpler functions, and in many cases functionality, or lack of, differed by a single mutation between parent and offspring. In several cases deleterious mutations provided stepping-stones for further complex functionality to evolve.

However, the model used is abstract and not reflective of natural biological systems. The genomic structure, which consists of a sequential, circular list of CPU-like instructions, is an abstraction of a real genome, with limited interaction between 'genes'. Functionality is based on logic functions, such as NOT, NAND and XOR, whereas biological systems process information through molecular mechanisms, such as gene regulation. This begs the question as to whether the results of such an abstract model be applied directly to biological systems.

We aim to address the fundamental question of 'adaptive vs non-adaptive' evolution of gene regulatory network topology by using a biologically realistic computational approach analogous to that used in Avida. We set out to answer a number of questions: 1) Is it possible to identify whether complex features, in particular global regulators, evolve as a result of adaptive or non-adaptive processes? 2) Is the evolved topology and complexity of a network dependent on the complexity of the environment? 3) How does complexity arise within an evolving network? To

do this we investigate the evolution of each of the network properties defined by Lynch [81], and their adaptive or non-adaptive origins, to various degrees, but focus primarily on the concepts of complexity and modularity.

13.1.1 Experimental and environmental conditions

Two types of environment were simulated: 1) a stress-free, base environment, and 2) a 'stress' environment, which introduces a number of stresses to the base environment. The base environment consists of nine food sources, each providing 5, 10, 15, 20 (two sources of each) or 25 molecules of 'energy'. In the 'stress' environment, each food source is randomly varying and available for approximately 12% of the simulation. Four biosynthesis pathways, combinations of high and low yield, high and low cost, are present in the genome. Two energy signalling genes, one detecting high energy concentrations (500 molecules) and the other detecting low energy concentrations (333 molecules) are also present. The 'stress' environment consists of two further 'stresses' (representing denatured proteins within the cell) and two corresponding stress response pathways. Every 25 simulation time-steps 25 'stress molecules' enter the cell. Each activation of a stress response pathway removes 25 molecules of the associated stress. If the number of a specific type of stress molecules reaches the given lethal 'stress limit' (100), then the cell dies. The specific model and evolution parameters for this investigation are given in Table 13.1.

13.1.2 Functional complexity

Measurement of the complexity of biological systems is inherently a very difficult task. Gene regulatory networks can be viewed as consisting of a number of integrated components or systems. In this model we define three functional systems, each corresponding to a specific biological function:

1. The 'energy regulation' component consists of all output genes/pathways which are repressed by at least one input or output gene. These interactions conserve energy by down regulating over-expression.
2. The 'stress response' component consists of stress response pathways and the input and output genes that activate them. A number of stress response sub-systems can be present in the network, and is dependent on the number of stress receptor/stress response pathways

in the network.

3. The 'growth' components consists of biosynthesis pathways and the input and output genes that activate them. A number of growth sub-systems can be present in the network, and is dependent on the number of biosynthesis pathways in the network.

Therefore in this model, a qualitative representation of complexity is the emergence and interaction of these functional systems.

13.1.3 *In silico* 'global regulators'

Global regulators have previously been defined by Gottesman [42] and further defined by Martínez-Antonio and Collado-Vides [86], as transcription factors (TF) that: i) regulate several metabolic pathways, or responses to environmental stimuli, ii) regulate large numbers of genes and operons, iii) will form regulation cascades, providing a hierarchy of regulation, iv) are likely to co-regulate with other TFs or global regulators and v) regulate operons which are transcribed by different σ factors.

In this model, the primary criteria for global regulation classification is based on properties i) and ii). Thus as the model consists of two classes of outputs (biosynthesis and stress response), a global regulator is defined as regulating both biosynthesis and stress response pathways.

To test significance of percentage of genome regulated, percentage of network edges regulated and percentage of positive edges regulated by the regulators, a χ^2 test could not be used as the genome and regulated edges sizes of the evolved models were so small that the expecteds were much less than 5 for each network. Therefore, the non-parametric Wilcoxon rank-sum test was used to test the proportions (computed using R).

13.2 Results

13.2.1 Complexity of evolved network is strongly influenced by environmental complexity

We evolved a number of randomly initialised model populations under 'stress-less' and 'stress' environmental conditions, reflecting increasing complexity. The network architectures were dramatically different between the two types of populations (Figure 13.1). Networks evolved under

Parameter	Value	Note
K^{basal}	1×10^{-2}	Derived from Section 12.2.4
P_1^{bio}, P_3^{bio}	50	
P_2^{bio}, P_4^{bio}	10	
C_1^{bio}, C_2^{bio}	75	
C_3^{bio}, C_4^{bio}	5	
$P_1^{stress}, P_2^{stress}$	25	
$C_1^{stress}, C_2^{stress}$	100	
$T_1^{stress}, T_2^{stress}$	100	
T_1^{energy}	500	
T_2^{energy}	333	
Generations (stress-less)	1000	
Generations (stressed)	10000	

Table 13.1: Model and evolution parameters for 'Complexity' investigation.

stress conditions had a large and dominant energy regulation system (red box), which mainly consisted of one or several co-regulating global regulators (Figure 13.1a). Two stress response systems (yellow boxes) were observed in all final models. The double activation by the associated stress receptor signal is a mechanism to over-ride the global energy regulation system, and thus indicates a co-evolving relationship between the two systems. The presence or absence of growth systems (blue boxes) within the network is highly dependent on whether the network can replicate. Networks a(i) and a(ii) each contain at least one growth system. Networks evolved under stress-less conditions have a much smaller energy regulation system, although still usually consisting of a single global regulator (Figure 13.1b). However, the global regulators often perform both activation and repression of different output pathways. The growth systems are the more dominant systems (examples are shown in networks b(i-ii)). The systems are largely inter-connected, with many input pathways activating several biosynthesis pathways.

13.2.2 The evolution of global regulators is adaptive

We examined the most highly connected regulators within each population, evolved under stress, stress-less or non-adaptive conditions. Table 13.2 shows the percentage of genome regulated and percentage of all network interactions regulated by the largest regulator in each of the 10 replicate populations (including mean values over replicates) under each experimental condition (networks with multiple different regulators were excluded). Under stress conditions 62.9% of all genes were regulated by the largest regulator, and a similar proportion (57.1%) of genes were regulated under stress-less conditions. In non-adaptively evolved populations, the percentage of

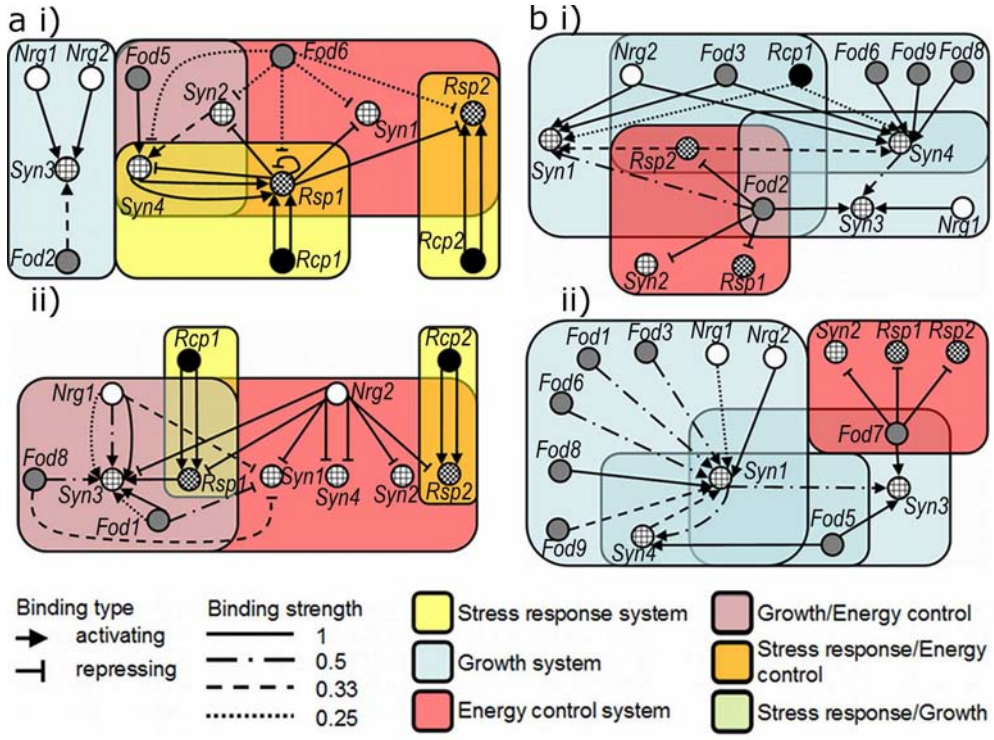

Figure 13.1: Different network topologies evolve in different environments. Two example replicates are shown from each environment. Populations evolved in environments with stresses (starvation and heat-shock-like stresses) have networks with a wide variety of all possible sub-systems a(i-ii). A similar functional structure is observed, consisting of a large energy regulation system (pink), two stress response systems (orange) and growth systems (blue). Global regulators are evident in each example (*Rsp1* and *Fod6* in (i); *Nrg2* in (ii)), each performing only repression. Populations evolved in a stress-less environment do not show as much variety of sub-systems b(i-ii). Energy regulation is on a smaller scale than in stressed populations, but growth systems are more heavily utilised, usually consisting of at least three growth systems. Global regulators are also present in all examples, but have a different structure to those found in the stressed populations. The global regulators perform the energy regulation, but are also incorporated into the growth systems, meaning the global regulators are dual-function.

genes regulated by the largest regulator was significantly smaller, with an average of 11.2% of genes regulated (stress: $p = 1.717 \times 10^{-4}$; stress-less: $p = 1.817 \times 10^{-4}$). The total proportion of network interactions regulated by the largest regulator was also significantly higher in the stress (22.8%) and stress-less (19.6%) populations, than the 1% in non-adaptive populations (stress: $p = 1.083 \times 10^{-5}$; stress-less: $p = 1.083 \times 10^{-5}$). The number of positive and negative interactions by the largest regulator in the non-adaptive populations was statistically equivalent ($p = 0.23$), whereas stress populations had a significant bias towards negative interactions (100%; $p = 6.386 \times 10^{-5}$), and the stress-less populations had a less significant bias towards negative interactions (66.2%; $p = 2.305 \times 10^{-3}$). None of the largest regulators within the non-adaptively

evolved populations were classed as global regulators (as defined in Section 13.1.3), but in stress populations 100% and in stress-less populations 90% of the largest regulators were classed as global regulators. Global regulation was therefore strongly selected for under both stress and stress-less environmental conditions, and adaptive regulator structure is significantly different from non-adaptive regulator structure. The ancestor, randomly generated network data are given Table 13.3. There is a slight bias towards negative regulation in the ancestor stress and stress-less population than in the non-adaptive populations, however, the proportion of genome regulated is smaller and very few (up to 10%) of the largest regulators were global regulators. This further strengthens the adaptive evolution and selection of global regulation mechanisms.

		Number of genes		Number of interactions			
		Total	Largest	Total	Largest regulator		Is global
	Replicate	network	regulator	network	Total	Activating	regulator?
Stress	1	6	4 (66.7%)	15	4 (26.7%)	0 (0%)	yes
	2	6	5 (83.3%)	32	5 (15.6%)	0 (0%)	yes
	3	6	3 (50%)	15	3 (20%)	0 (0%)	yes
	4	6	5 (83.3%)	21	5 (23.8%)	0 (0%)	yes
	5	7	4 (57.1%)	14	4 (28.6%)	0 (0%)	yes
	6	7	2 (28.6%)	17	3 (17.7%)	0 (0%)	yes
	7	6	5 (83.3%)	12	5 (41.7%)	0 (0%)	yes
	8	8	4 (50%)	18	4 (22.2)%)	0 (0%)	yes
	9	7	3 (42.9%)	20	3 (15%)	0 (0%)	yes
	10	6	2 (83.3%)	22	5 (22.7%)	0 (0%)	yes
	mean		$62.9\% \pm 20.1$		$23.4\% \pm 7.8$	$0\% \pm 0$	100%
Stress-less	1	10	6 (60%)	48	6 (12.5%)	2 (33.3%)	yes
	2	12	3 (25%)	37	4 (10.8%)	0 (0%)	yes
	3	6	5 (83.3%)	10	5 (50%)	2 (40%)	yes
	4	10	4 (40%)	24	4 (16.7%)	2 (50%)	yes
	5	6	6 (100%)	23	7 (30.4%)	4 (57.1%)	yes
	6	6	3 (50%)	23	3 (13.0%)	1 (33.3%)	yes
	7	28	6 (21.4%)	68	6 (8.8%)	2 (33.3%)	no
	8	6	5 (83.3%)	52	8 (15.4%)	2 (25%)	yes
	9	7	6 (85.7%)	30	6 (20%)	2 (33.3%)	yes
	0	23	5 (21.7%)	33	6 (18.2%)	2 (33.3%)	yes
	mean		$57.1\% \pm 29.7$		$19.6\% \pm 12.3$	$33.8\% \pm 15.2$	90%
Non-adaptive	1	297	34 (11.5%)	5793	36 (0.6%)	18 (50%)	no
	2	287	29 (10.1%)	4611	31 (0.7%)	18 (58.1%)	no
	3	196	32 (16.3%)	1681	33 (2.0%)	20 (60.6%)	no
	4	293	31 (10.6%)	5234	32 (0.6%)	15 (46.9%)	no
	5	186	20 (10.8%)	2080	22 (1.1%)	13 (59.1%)	no
	6	226	27 (12.0%)	2983	27 (0.9%)	15 (55.6%)	no
	7	189	23 (12.2%)	1963	23 (1.2%)	9 (39.1%)	no
	8	225	25 (11.1%)	3105	25 (0.8%)	11 (44%)	no
	9	280	25 (8.9%)	4603	29 (0.6%)	17 (58.6%)	no
	10	179	16 (8.9%)	1449	16 (1.1%)	10 (62.5%)	no
	mean		$11.2\% \pm 2.1$		$1.0\% \pm 0.4$	$53.4\% \pm 8.0$	0%

Table 13.2: Network data for each replicate evolved population; 'stress', 'stress-less' and 'non-adaptive'. The number of genes regulated and number of interactions by the largest single regulator in the network is shown, along with its global regulator status.

		Number of genes		Number of interactions			
		Total	Largest	Total	Largest regulator		Is global
	Replicate	network	regulator	network	Total	Activating	regulator?
Stress	1	38	4 (10.5%)	130	4 (3.1%)	3 (75.0%)	no
	2	40	5 (12.5%)	123	5 (4.1%)	4 (80.0%)	no
	3	37	7 (18.9%)	158	7 (4.4%)	1 (14.3%)	no
	4	38	6 (15.8%)	143	6 (4.2%)	2 (33.3%)	no
	5	38	7 (18.4%)	130	7 (5.4%)	1 (14.3%)	no
	6	38	6 (15.8%)	130	6 (4.6%)	1 (16.7%)	no
	7	38	6 (15.8%)	153	6 (3.9%)	0 (0.0%)	no
	8	39	4 (10.3%)	123	4 (3.3%)	2 (50.0%)	no
	9	37	6 (16.2%)	126	6 (4.8%)	4 (66.7%)	no
	10	37	6 (16.2%)	125	6 (4.8%)	3 (50.0%)	no
	mean		$15.0\% \pm 3.0$		$4.3\% \pm 0.7$	$40.0\% \pm 28.4$	0%
Stress-less	1	40	6 (15.0%)	164	7 (4.3%)	4 (57.1%)	no
	2	38	6 (15.8%)	108	6 (5.6%)	2 (33.3%)	no
	3	38	5 (13.2%)	142	5 (3.5%)	2 (40.0%)	no
	4	37	6 (16.2%)	119	6 (5.0%)	2 (33.3%)	no
	5	38	5 (13.2%)	147	5 (3.4%)	1 (20.0%)	yes
	6	39	8 (20.5%)	156	9 (5.8%)	5 (55.6%)	no
	7	38	5 (13.2%)	140	5 (3.6%)	1 (20.0%)	no
	8	39	6 (15.4%)	187	6 (3.2%)	1 (16.7%)	no
	9	39	8 (20.5%)	176	10 (5.7%)	8 (80.0%)	no
	10	39	5 (12.8%)	143	5 (3.5%)	1 (20.0%)	no
	mean		$15.6\% \pm 2.9$		$4.4\% \pm 1.1$	$37.6\% \pm 20.8$	10%
Non-adaptive	1	39	7 (17.9%)	142	8 (5.6%)	4 (50.0%)	no
	2	38	9 (23.7%)	123	12 (9.8%)	3 (25.0%)	yes
	3	38	6 (15.8%)	139	7 (5.0%)	3 (42.9%)	no
	4	40	8 (20.0%)	202	9 (4.5%)	4 (44.4%)	no
	5	39	8 (20.5%)	165	9 (5.5%)	6 (66.7%)	no
	6	37	7 (18.9%)	145	7 (4.8%)	4 (57.1%)	no
	7	39	7 (17.9%)	181	8 (4.4%)	6 (75.0%)	yes
	8	39	6 (15.4%)	128	6 (4.7%)	2 (33.3%)	no
	9	38	6 (15.8%)	133	6 (4.5%)	4 (66.7%)	yes
	10	38	10 (26.3%)	178	11 (6.2%)	6 (54.5%)	yes
	mean		$19.2\% \pm 3.6$		$5.5\% \pm 1.6$	$51.6\% \pm 15.7$	40%

Table 13.3: Network data for each replicate ancestor population; 'stress', 'stress-less' and 'non-adaptive'. The number of genes regulated and number of interactions by the largest single regulator in the network is shown, along with its global regulator status. The ancestor network is the original, randomly generated network from which the best model in the final generation is evolved from.

13.2.3 Complexity of a network arises in stages

We examined the entire evolutionary history of a number of models from each type of population. Figure 13.2 shows selected 'snapshots' of the best performing model in one of the stressed populations over its 10,000 generation lineage. The initial network (randomly generated) consists of a small energy regulation system regulating a single biosynthesis pathway, and an interconnected growth system (Figure 13.2a). Additionally, a stress response system is present. This network is non-viable (able to survive no more than 50 time-steps) and with the minimal system architecture is not complex.

The network evolution progresses in two broad phases. In the first phase (Figure 13.2a-d), the networks are unable to replicate, and selection is for longevity. In the second phase (Figure 13.2e-h) the networks are able to replicate, and selection is for rate of growth. After 100 generations the network has substantially changed from its initial state (Figure 13.2b). The complexity of the energy regulation system is increased, with three biosynthesis pathways regulated. A second stress response system has evolved, whilst the complexity and efficiency of the original response system has increased, with the addition of activation by its associated receptor. The growth system has been lost, with increasing efficiency of energy regulation. This together with the stress response systems allow the network to survive around 100 time-steps. The network after 250 generations has increased the efficiency of the second stress response system, also evolving activation by its receptor (Figure 13.2c). The increased efficiency of the stress response systems now prevents the cell from dying due to the lethal stress levels. The efficiency of the energy regulation systems has further increased, with additional input pathways regulating the biosynthesis pathways. This network is able to survive between 150 and 250 time-steps. The network after 500 generations again has an increasingly complex energy regulation system, with three co-regulating global regulators providing an efficient, but redundant, energy saving mechanism (Figure 13.2d). A growth mechanism has also reappeared, but the energetic cost is still unsustained by the network. These adaptations increase the survival to between 300 and 400 time-steps.

The second phase begins at generation 1166 with the emergence of the first replicating network. This network shows the first appearance of what becomes the primary global regulator, *Rsp1*, in the energy regulation system (Figure 13.2e). A co-regulating global regulator, *Fod2*,

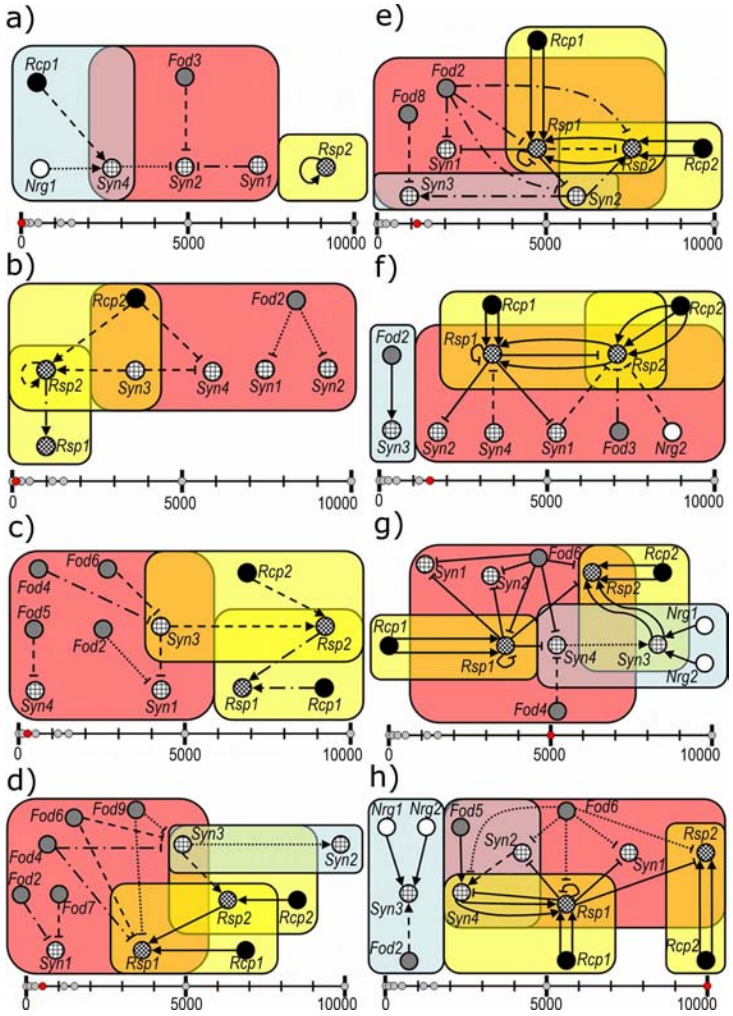

Figure 13.2: Incremental evolution of a functionally complex gene regulatory network. The network initially starts with each type sub-system, yet is unable to survive more than a few tens of time-steps (a). After 100 generations the network structure has changed dramatically, losing the growth system, but gaining an efficient stress response system (b). By 250 generations a second stress response system has evolved, and also the energy regulation system has continued to grow (c). By 500 generations a number of small global regulators have evolved, further increasing the energy regulation system (d). The network also has a number of redundant regulators each performing identical roles. A growth system has also reemerged which is interacting with stress and energy regulation systems. The first replication event after 1166 generations shows the network has a very efficient set of stress response systems, and also the emergence of just two co-regulating global regulators performing the majority of the energy regulation (e). The different systems have become increasingly interconnected. After 1500 generations only a single global regulator now performs the key role in global regulation (f). An independent growth system has also emerged, which is now viable due to efficient energy regulation and stress systems. Network functionality remains similar after 5,000 generations, with the global regulator increasing the number of pathways regulated, and recruitment of another gene as a transient global regulator (g). The growth system has increased in efficiency, now utilising the energy signals. After 10,000 generations network structure and function is again similar (h). The energy regulation system is still controlled by the same global regulator, and a secondary weaker connected regulator. The main growth system is now independent of other systems, and a second has evolved within the energy regulation and stress systems. The network functionality is evolved in stages, with certain systems as prerequisites for the sustainability of others.

is also present in this system. The complexity of the stress response systems has increased, with each receptor binding to multiple binding sites of each response gene. This additional complexity has evolved in response to the incorporation of both stress response pathways into the energy regulation system, indicating an adaptive response to the other systems. By 1,500 generations, the network whilst maintaining similar stress response systems, has lost the co-regulating global regulator *Fod2* (Figure 13.2f). The growth system, *Syn3*, has been modified to be more efficient, using a food signal. Moreover, the network is now able to sustain this system and biomass production (now the measure of fitness) has increased fitness from around 3,500 to over 5,500 (Figure 13.3). The network after 5,000 generations has evolved a new co-regulator, *Fod6*, in the energy regulation system (Figure 13.2g). However, this global regulator is redundant, showing the transient state of the network and influences of non-adaptive processes. The functional systems are increasingly inter-connected, producing an increasingly complex network. The growth system remains and is increasingly more efficient, with regulation from the energy signalling pathways (*Nrg1* and *Nrg2*), resulting in an increase of biomass production and fitness to around 19,500. At the end of the evolution, 10,000 generations, the network structure is similar to the previous network. The energy regulating system still consists of two global regulators, but with weakened interactions from the redundant *Fod6* regulator (Figure 13.2h). This indicates that these interactions are non-essential and are in the process of being lost through genetic divergence. The main growth system has become decoupled from the energy regulation system. Protein production and stability rates have also been modified, leading to an increase in biomass production and fitness of around 22,000 (see Table 13.4).

	Production		Stability		
Generation	Rsp1	Genome mean	Rsp1	Genome mean	Fitness
initial	3	3 ± 2.4	1	1.6 ± 0.8	43
100	3	2.3 ± 2.3	1	1.7 ± 1.1	101
250	3	1.8 ± 2.5	7	2.2 ± 1.8	154
500	3	0.8 ± 1.5	9	2.2 ± 2.0	353
1166	7	0.8 ± 1.9	20	2.3 ± 4.0	3570
1500	7	1.0 ± 2.1	34	3.2 ± 7.5	5810
5000	4	0.5 ± 1.1	73	7.8 ± 16.6	19800
10000	3	0.7 ± 1.1	96	8.4 ± 21.5	22980

Table 13.4: Protein production and stability parameters of global regulator, *Rsp1*, and genome mean

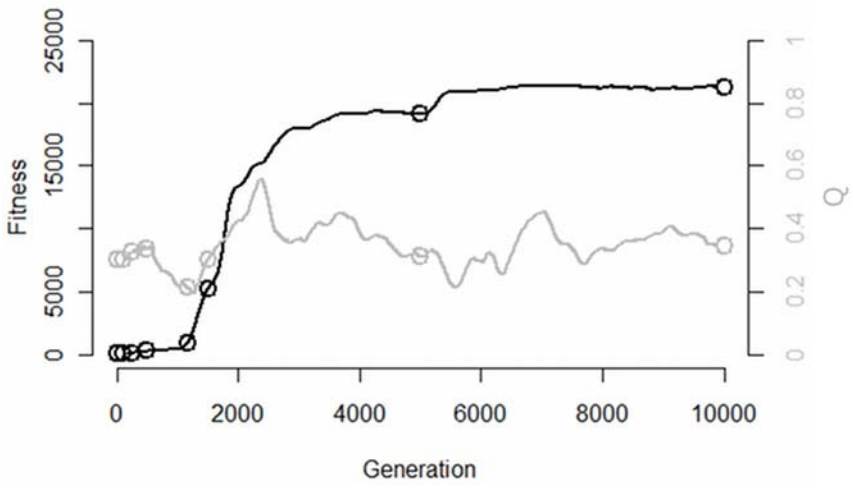

Figure 13.3: Fitness (black) and modularity 'Q' (gray) during evolution. Data are de-noised using an exponential central-moving average algorithm ($\alpha = 0.01$, points=175), and $\circ$ are the measured points in Figure 13.2. The fitness value slowly increases during the first 1000 generations. After 1,000 generations the fitness rapidly increases as the network is able to replicate. The fitness begins to plateau around 3,000 generations, but a final increase in fitness occurs around generation 5,000. The modularity shows no such clear trend. The network modularity varies between 0 and 0.5, but after 10,000 generations has a smaller modularity than the initial network (around 0.35 from 0.45). This indicates that modularity is not correlated to fitness.

13.2.4 Complexity consists of modularity and functional information

Applying the structural 'modularity' measure [95] to the evolving network reveals no trend in modularity, whilst fitness increases monotonically (Figure 13.3). However, examining the network 'modules' of the model lineage reveals very similar structures to the functional systems that are extracted (Figure 13.4 compare with Figure 13.2). Despite the single-module membership limitation of each node in the network, several networks share almost identical groupings. The initial network, (a), generates three modules which correspond (apart from the overlapping of *Syn4*) to the functional systems. In network (d), the modularity measure is able to extract the two separate stress response systems, but does not extract the single growth system. However, as the network evolves it is unable to extract as many systems. For example, in the final network (h), the modularity measure detects a growth and a stress response system, however, is unable to extract the second growth and stress response systems due to the high inter-connectivity between the functional modules. The structural method is able to extract functional sub-systems in the early stages of network evolution, but as the apparent network complexity, and model fitness, increases it is unable to extract all the functional systems due to inter-connectivity.

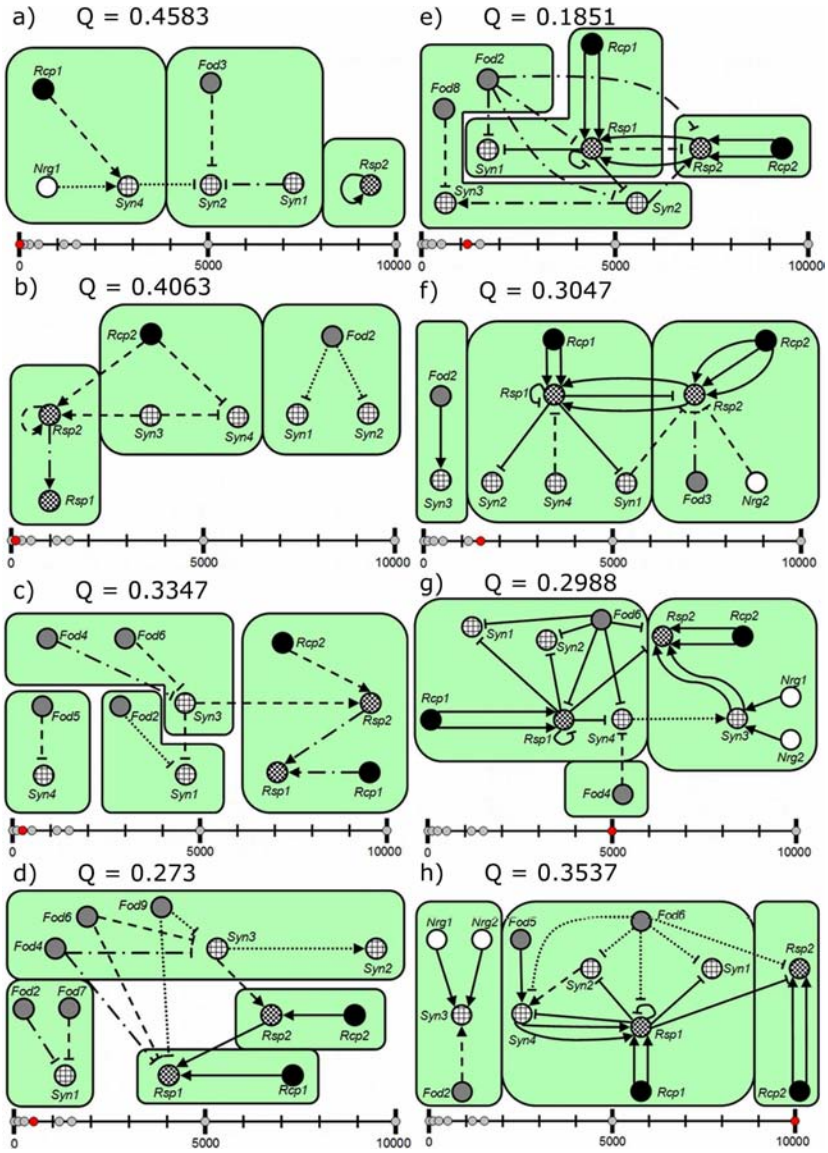

Figure 13.4: Modularity during network evolution. Modularity values vary during evolution, with no trend evident. During the early generations, up to 500 (a-d), the extracted modules in many cases are similar to the functional sub-systems, indicating that functionality evolves in clearly defined modules. After the initial rapid stage of evolution, > 1000 generations (e-h), the functional systems interact with each other causing the clear modules to be lost.

13.2.5 Highly adapted models consist of essential and non-essential components

The network contains several very fragile sub-systems (Figure 13.5 and Table 13.5). Removal of either stress response system (*str1KO/str2KO*) or energy regulation system (*nrgKO*) are completely lethal. Major reductions (90% or greater) in the production or stability rates of the global regulator, *Rsp1*, are also detrimental to survival rate. However, the network is also robust to other mutations and is able to withstand, to varying degrees, entire removal of some systems. Removal of the decoupled growth system (*grwKO*) severely reduces the growth rate of the model, but increases its survival to almost 100%. Small reductions in production or stability rates ($< 50\%$) are mostly non-lethal.

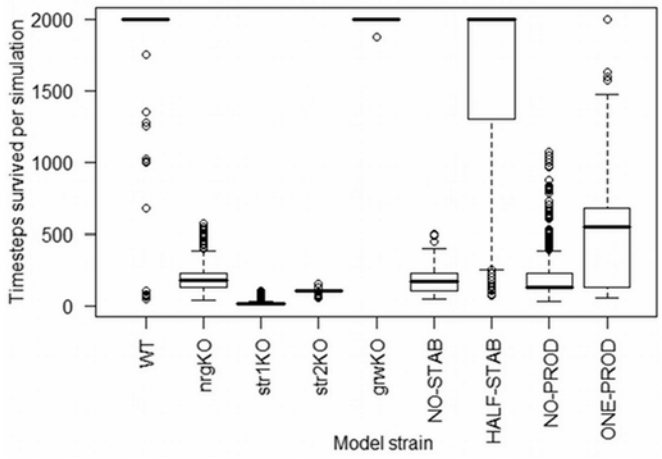

Figure 13.5: Robustness and fragility to mutations in network components. Wild-type model is evolved network from Figure 13.2, and each strain is simulated 1,000 times. The wild-type (WT) consistently reaches the termination criteria of 2,000 simulation steps, indicating a robust and efficient network. Removing the global regulators (nrgKO) governing the energy regulation system reduces survival rate to 0, but is able to survive around 150 time-steps. Removal of the first stress response system (str1KO) also reduces the survival rate to 0, and can survive only tens of time-steps. A similar result is observed removing the second stress response system (str2KO), but survives around 100 time-steps. The mutants *str1KO* and *str2KO* cause the network to die at different points in simulation, due to the additional global regulator activity of elements of the *str1* system. Removal of the independent growth system (grwKO) has a positive effect on survival rate, reaching nearly 100%. Therefore, certain sub-systems are pre-requisites for survival, whilst others can be lost with little effect on survival rate. Perturbing the global regulator, *Rsp1*, also dramatically effects survival rate. Halving the protein stability (HALF-STAB) causes the network to die at any point, but mostly replicates. Reducing the protein production rate also has a large impact on survival, indicating the highly tuned state of the network.

Strain	Survival (%)	Biomass
Wild-type (WT)	97.5	18628.3
Energy regulation (nrgKO)	0	-
Stress response 1 (str1KO)	0	-
Stress response 2 (str2KO)	0	-
Growth (grwKO)	99.9	1039.6
Rsp1 no stability	0	-
Rsp1 half stability	61.9	20441.9
Rsp1 no production	0	-
Rsp1 one production	1.4	17397.9

Table 13.5: Wild-type and mutant strain survival and biomass values. Wild-type model is evolved network from Figure 13.2, and each strain is simulated 1000 times. Removal of the energy regulation or stress response 1 or 2 systems is a lethal mutation, whereas removal of the growth system increases survival rate slightly, but reduces biomass yield around 18-fold. Halving the stability of the global regulator, *Rsp1*, reduces survival rate by 1/3, but has little effect on biomass. Reducing the protein production rate dramatically reduces the survival rate to almost 0%, but again little effect on biomass. No protein stability or no production mutations are completely lethal

13.3 Discussions

In summary, we find that gene regulatory network structure and function is strongly influenced by environmental conditions. The evolution of functional complexity occurs in stages, in which essential energy regulation and stress response systems are required before growth systems can be sustained. Also, the network is more robust to mutations to the non-essential growth systems, than the energy regulation and stress response systems. Evidence of redundancy is observed during multiple points during evolution, indicating that duplication of systems is used to provide exploratory material for further functional evolution. These genes are also transient, and can eventually be lost through mutation. Also observed was the *de novo* evolution of global regulation mechanisms, which are strongly selected for under specific conditions.

Evolution in biology is inherently difficult to observe in action, due to the enormous timescales required. However, computational evolution allows more realistic timescales on which we can observe evolution. The evolution of increasingly complex gene regulatory mechanisms has also been observed in other *in silico* bacterial models, for instance, in evolving chemotaxis dynamics, simple mechanisms were observed in environments of constant stimuli, whereas under fluctuating stimuli environmental conditions more complex mechanisms were observed [41]. This further implies a strong connection between environmental and network complexity. The incremental functional evolution observed during our experiments is also an exciting result. Randomly gen-

erated networks are non-viable due to the energetic cost of over-expression and/or lethal stress levels. Therefore, the solution is to remove the energetic requirements, which can be achieved in a number of ways: 1) remove non-essential/non-functional genes, 2) reduce the expression rate or 3) regulate the expression of genes. It is evident that all three actions are utilised, as genome size very quickly reaches a small size, and many gene expression rates are also reduced. Global regulation of gene expression was a selected mechanism and the evolution of similar global regulatory structures was observed in many populations. The global regulation mechanism is a very energy efficient solution, requiring expression of only a single gene to regulate many. The relative ease, in living systems, of adaptive evolution of a binding site via point mutations to a specific transcription factor, in a reasonable evolutionary timescale [17], would further strengthen the selection of such a regulatory mechanism in the model. This energy efficiency, and ease of evolving new regulatory interactions, along with the similar structure observed in many populations, may be strong evidence for the adaptive selection of global regulation mechanisms observed in many biological networks, in contrast to the non-adaptive mechanisms proposed [22, 81, 82]. Once energy regulation is resolved, the models adapt to counter lethal stress levels, which are only encountered once energy regulation is in place. When both energy regulation and lethal stress levels are resolved, the next adaptation is for speed of growth. Growth systems were observed at multiple points during the network evolution, however, it is only once the 'core' systems are in place that the growth systems become fixed. Thus, a reasonable hypothesis is that, early in evolution the 'core' survival systems of energy regulation and detoxification might have evolved prior to efficient growth and replication systems. However, as the evolutionary fitness function used within the model is separated into distinct regimes, which first assigns fitness based on survival time (the efficiency of energy regulation) followed by fitness based on growth (the amount of biomass production), the observation of the evolution of functional complexity in steps may merely be an artefact of this fitness function. Whilst this function allows the evolutionary framework to evolve surviving models, an alternative function, using only a single regime of fitness based on growth could be used if the evolutionary process is seeded with models already capable of surviving simulation in the environment. This could be achieved by selecting models evolved in simpler environments and then introducing the evolved models to the complex environments. However, this alternative approach may introduce bias into the evolutionary process, as hand-made, or hand-picked models may be required, which may influence

how the model can evolve.

We do not understand biological systems, or complexity measures, enough to assign a complexity value to a particular system or network, and so assessment of complexity remains qualitative. Biological networks are often thought of consisting of modular, independent units. Indeed, other *in silico* experiments have found modularity to increase with network complexity [64, 52]. However, the observed network structures, whilst displaying some clearly modular functional systems, were not independent with many cases of inter-connected systems. Examining biological networks in more detail we see a similar inter-connected functional structure. For instance the global regulator *CRP* regulates the central carbon metabolism of *E. coli*. Yet, it also regulates many other metabolic and stress response pathways, creating a centrally connected hub structure, rather than independent functional modules [69]. Although it is convenient to attempt to separate a biological network into smaller independent sub-graphs, such as the network motif approach, it is also possible to 'lose the bigger picture'. Such an approach may yield some dynamical or functional behaviour from a network, but without taking all other interactions and connections into account, not all behaviours will be identified. As such, we suggest that a true biological complexity measure should not only take structural information, such as modularity, into account, but necessarily requires functional information, such as the functional systems approach taken in this investigation.

Chapter 14

MODEL DISCUSSIONS

14.1 Limitations of model

There are a number of limitations in the current model formulation. Whilst the lack of polymerisation and complex formation did not appear to hinder the evolution of complex gene regulation networks, simulations using the fine-grained model indicated that complex formation and conformational changes can play a large role in gene regulation mechanisms. Therefore, implementation of complex formation may facilitate the evolution of novel gene regulation mechanisms within the model.

Basal gene expression rates were found to be extremely important in gene regulation mechanisms. However, the current implementation of basal gene expression, a probabilistic process using a fixed value, allowed the possible evolution of somewhat infeasible gene regulatory networks. Incorporation of RNA polymerase as a distinct and discrete protein species, the production of which could be evolved using the same mechanism as currently implemented for proteins, would be a more accurate implementation for gene expression. Therefore, the number of free RNA polymerase within the network would determine transcription of both activated genes and levels of basal expression, introducing competition of the limited resource required for gene expression.

With the evolution of increasingly complex networks and environmental conditions, sigma σ factors may begin to play an important role in transcription initiation. Implementation could consist of multiple σ factors and a single RNA polymerase species, or multiple RNA polymerase species, each representing the polymerase with a specific σ factor.

14.2 Future directions

14.2.1 Spatial environments and organism interactions

As identified with the fine-grained model, a spatial environment, in which the models can interact with each other and the environment, could be implemented, using for example cellular automata, to produce a more realistic environmental framework. Incorporation of such an environment would also allow implementation of more bacterial processes, for example, chemotaxis and cell-cell communication, such as quorum sensing. Further, the models could output products to the environment, which could either be used as food by other models, or be toxic. The environmental complexity would thus be much greater than in the current model formulation, and the results presented indicate that this would generate increasingly complex gene regulatory networks.

14.2.2 Plasmid and bacteriophage ecology

Whilst horizontal gene transfer is already modelled, a more sophisticated mechanism could be implemented along with the spatial environment. A number of important questions surround the existence of plasmids, such as why plasmid-based genes are not incorporated into the host genome, or why certain genes have been transfered to an extra-chromosomal existence and energetic and fitness costs or benefits for maintenance of plasmids. Whilst plasmid costs and benefits have previously been explored both experimentally and theoretically, such questions could be investigated using the existing model and a spatial environment, potentially yielding new insights. Associated with plasmids, bacteriophage dynamics could also be implemented and investigated.

Part V

SUMMARY

This part summarises the models, results of evolutionary simulations and future directions.

Chapter 15

SUMMARY

The aim of this thesis was to develop novel computational models for investigating the evolution of prokaryotic gene regulation networks. A number of objectives were defined: i) review existing computational models for their suitability to realistic gene regulatory network modelling and evolution ii) development of realistic biologically-informed computational models allowing identification of 'essential' biological processes to accurately model gene regulatory network evolution, iii) use the computational models to investigate the effects of stochasticity on gene regulatory network evolution and diversity, in particular how basal expression is used and iv) use the computational models to investigate the evolution of complexity in gene regulatory networks.

Part I briefly reviews the biological processes used in gene regulatory network and also discusses many of the currently used computational models.

Part II introduced the first novel computational model, the 'coarse-grained' model which added functional dynamics to an existing evolutionary model. Evolutionary simulations revealed the importance of modelling 'energy' within a computational model, due to which realistic regulatory mechanisms were evolved. Further, incorporation of basal gene expression, along with energetic constraints, led to the observations of 'stochastic shrinkage' and 'deterministic bloat' of genomes. Basal gene expression was also utilised in networks which did not have a central energy signalling system, and consequently deterministic systems could not grow under these conditions. Non-adaptively evolved networks were also structurally different to the networks evolved to a fitness function, indicating the strong adaptive pressure due to a fitness function.

Part III introduced the second novel computational model, the 'fine-grained' model which was developed to be as biologically realistic as possible. Again a number of realistic network

properties were evolved, including energy regulation and unstable mRNA molecules and stable protein molecules, highlighting the importance of molecule stability in gene regulation.

Part IV introduced the 'extended coarse-grained' model, which built on the 'coarse-grained' model. This model included several important modifications and additions which were identified from evolutionary simulations using the 'coarse-grained' and 'fine-grained' models. As a result, a more biologically-informed model, which is also computationally very efficient, was produced.

The 'extended coarse-grained' model was used in an investigation into the effects of stochasticity on network evolution and diversity, and was presented in Part IV. These simulations show that in systems that are able to evolve rates of basal expression, two attractors, one with and one without basal expression, are observed. Simulation paradigms without basal expression generate bloated networks with non-functional elements. Further, a range of functional solutions were observed under identical conditions only in stochastic networks, however diversity of solutions within a single population was not evolved. Moreover, there are trade-offs between efficiency and yield, indicating an inherent intertwining of fitness and evolutionary dynamics.

The evolutionary simulations using the 'extended coarse-grained' model presented in Part IV have also shed much light on the property of complexity and its evolution. These results show, to the authors knowledge, the first observation and analysis of the *de novo* evolution of adaptive global gene regulation mechanisms using *in silico* models. Using functional definitions, complex networks with a hierarchical structure were identified, with 'core' global energy regulation mechanisms providing a base on which further functional systems could be sustained by the network. Further, such complexity arises in stages; first, the energy regulation system evolved along with rapid loss of redundant genes, environmental responses are then evolved before finally evolving systems to increase biomass production (fitness). Environmental complexity was directly related to the complexity of the resultant networks, generating very different functional topologies. The network topologies and gene regulation mechanisms of the *in silico* models share many striking parallels with bacterial gene regulatory networks. As a result, the strong adaptive evolution observed shaping the *in silico* networks provides strong evidence that biological networks evolve due to adaptive selection, rather than by non-adaptive evolutionary mechanisms. Experimental work by a number of research groups further supports the 'adaptive view' of network evolution.

The models and evolutionary simulations presented in this thesis suggest that *in silico* modelling provides an ideal framework for studying evolution. The use of biologically-informed

models evolved to the single goal of any organism, survival and replication, is capable of answering fundamental questions in the evolution of living systems, such as the evolution of complexity and heterogeneity. Indeed, the 'extended coarse-grained' model is a powerful and easy to use tool for studying the evolution of prokaryotic gene regulatory networks, from which several important insights into the role of stochastic processes in gene regulation and expression have been observed. In particular, the role of stochastic basal gene expression, an often overlooked process in modelling, has been shown to have a great impact on the dynamics and structure of a network and can also be selected for and adaptively tuned during evolution. The adaptive evolution of complex, global and hierarchical regulation mechanisms in the model is also a result of importance, as they are key mechanisms in many biological gene regulatory networks, the evolutionary origins of which are the subject of much debate.

15.1 Future work

A number of extensions to both the 'fine-grained' and 'extended coarse-grained' models were presented and discussed in Chapter 10 and Chapter 14 respectively. Whilst the 'fine-grained' model is a very suitable model of gene regulatory networks, it is not well suited to evolutionary work, due to its computational complexity. As such use of this model is recommended for modelling specific systems, along with performing *in silico* genetics analysis. The 'extended coarse-grained' model is highly suited, and recommended, for use in evolutionary modelling: it is computationally efficient, biologically-informed and capable of replicating biological phenomena. A worthwhile extension to this model would be a spatial environment, allowing interactions between models.

Refinement of a quantitative biological complexity measure would also be a valuable addition to assist in the analysis of gene regulatory networks. The work on functional complexity and modularity in Part IV indicates that a biologically relevant network complexity measure would consist of both approaches. Such a complexity measure would simplify observing the evolution of complexity, as the method used in this work was a time-consuming process in which each network was analysed by hand and did not produce a metric to compare two networks.

List of References

[1] Adami, C., Ofria, C. & Collier, T. C. (2000) Evolution of biological complexity. *Proc Natl Acad Sci USA*, **97(9)**:4463–4468.

[2] Albert, R. (2005) Scale-free networks in cell biology. *J Cell Sci*, **118(21)**:4947–4957.

[3] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002) *Molecular Biology of the Cell*. Garland Science, 270 Madison Avenue, New York, NY 10016, U.S.A., 4th edition.

[4] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol*, **215(3)**:403–410.

[5] An, W. (2007) Histone acetylation and methylation: Combinatorial players for transcriptional reglation. *Subcell Biochem*, **41**:351–369.

[6] Andrianantoandro, E., Basu, S., Karig, D. K. & Weiss, R. (2006) Synthetic biology: New engineering rules for an emerging discipline. *Mol Syst Biol*, **2(2006.0028)**.

[7] Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. (2006) Construction of *Escherichia coli* k-12 in-frame, single-gene knockout mutants: the keio collection. *Mol Syst Biol*, **2(2006.0008)**.

[8] Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. & Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, **14(3)**:283–291.

[9] Babu, M. M., Teichmann, S. A. & Aravind, L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol*, **358(2)**:614–633.

[10] Bai, L., Santangelo, T. J. & Wang, M. D. (2006) Single-molecule analysis of RNA polymerase transcription. *Annu Rev Biophys Biomol Struct*, **35**:343–360.

[11] Banzhaf, W. (2003) On the dynamics of an artificial regulatory network. In *Proceedings of the 7th European Conference on Artificial Life*.

[12] Banzhaf, W. & Kuo, P. D. (2004) Network motifs in natural and artificial transcriptional regulatory networks. *Journal of Biological Physics and Chemistry*, **4**:85–92.

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

[14] Bedau, M. A., Synder, E. & Packard, N. H. (1998) A classification of long-term evolutionary dynamics. In *Proceedings of the sixth international conference on Artificial Life*. MIT Press, Cambridge, MA, U.S.A.

[15] Belle, A., Tanay, A., Shamir, R. & O'Shea, E. K. (2006) Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci USA*, **103(35)**:13004–13009.

[16] Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Statist Soc B*, **57(1)**:289–300.

[17] Berg, J., Willmann, S. & Lässig, M. (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol*, **4(42)**.

[18] Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S. & Cohen, S. N. (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci USA*, **99(15)**:9697–9702.

[19] Blount, Z. D., Borland, C. Z. & Lenski, R. E. (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc Natl Acad Sci USA*, **105(23)**:7899–7906.

[20] Bremer, H. & Dennis, P. P. (1996) *Escherichia coli and Salmonella: Cellular and molecular biology*, volume 2, chapter Modulation of cell parameters by growth rate, pages 1553–1569. ASM Press, Washington, D.C., 2nd edition.

[21] Cai, L., Friedman, N. & Xie, X. S. (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature*, **440(7082)**:358–362.

[22] Cases, I. & de Lorenzo, V. (2005) Promoters in the environment: Transcriptional regulation in its natural context. *Nat Rev Microbiol*, **3(2)**:105–118.

[23] Channon, A. (2001) Passing the ALife test: Activity statistics classify evolution in Geb as unbounded. In *Advances in Artificial Life: Proceedings of the Sixth European Conference on Artificial Life (ECAL2001)*. Springer Verlag, Heidelburg, Germany.

[24] Chaudhuri, R. R., Loman, N. J., Snyder, L. A., Bailey, C. M., Stekel, D. J. & Pallen, M. J. (2008) xBASE2: A comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res*, **36(Database issue)**:D543–D546.

[25] Christensen, R. (1997) *Log-linear models and logistic regression*. Springer texts in statistics. Springer Verlag, Heidelburg, Germany, 2nd edition.

[26] Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311(5765)**:1283–1287.

[27] Clarkson, J., Shu, J.-C., Harris, D. A., Campbell, I. D. & Yudkin, M. D. (2004) Fluorescence and kinetic analysis of the SpoIIAB phosphorylation reaction, a key regulator of sporulation in *Bacillus subtilis*. *Biochemistry*, **43(11)**:3120–3128.

[28] Cooper, T. F., Rozen, D. E. & Lenski, R. E. (2003) Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc Natl Acad Sci USA*, **100(3)**:1072–1077.

[29] Cordero, O. X. & Hogeweg, P. (2006) Feed-forward loop circuits as a side effect of genome evolution. *Mol Biol Evol*, **23(10)**:1931–1936.

[30] Crombach, A. & Hogeweg, P. (2008) Evolution of evolvability in gene regulatory networks. *PLoS Comput Biol*, **4(7)**:e1000112.

[31] Dekel, E., Mangan, S. & Alon, U. (2005) Environmental selection of the feed-forward loop circuit in gene-regulation networks. *Phys Biol*, **2(2)**:81–88.

[32] Dodd, A. N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., Hibberd, J. M., Millar, A. J. & Webb, A. A. R. (2005) Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science*, **309(5734)**:630–633.

[33] Dvornyk, V., Vinofradova, O. & Nevo, E. (2003) Origin and evolution of circadian clock genes in prokaryotes. *Proc Natl Acad Sci USA*, **100(5)**:2495–2500.

[34] El-Samad, H., Kurata, H., Doyle, J. C., Gross, C. A. & Khammash, M. (2005) Surviving heat shock: Control strategies for robustness and performance. *Proc Natl Acad Sci USA*, **102(8)**:2736–2741.

[35] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002) Stochastic gene expression in a single cell. *Science*, **297(5584)**:1183–1186.

[36] François, P. & Hakim, V. (2004) Design of genetic networks with specified functions by evolution *in silico*. *Proc Natl Acad Sci USA*, **101(2)**:580–585.

[37] Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A. M., Treviño-Quintanilla, L. & Collado-Vides, J. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res*, **36(Database issue)**:D120–D124.

[38] Gans, J., Wolinsky, M. & Dunbar, J. (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, **309(5739)**:1387–1390.

[39] Gibson, M. A. & Bruck, J. (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. *J Phys Chem A*, **104**:1876–1889.

[40] Gillespie, D. T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys*, **22**:403–434.

[41] Goldstein, R. A. & Soyer, O. S. (2008) Evolution of taxis responses in virtual bacteria: Non-adaptive dynamics. *PLoS Comput Biol*, **4(5)**:e1000084.

[42] Gottesman, S. (1984) Bacterial regulation: Global regulatory networks. *Ann Rev Genet*, **18**:415–441.

[43] Grainger, D. C., Hurd, D., Harrison, M., Holdstock, J. & Busby, S. J. W. (2005) Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc Natl Acad Sci USA*, **102(49)**:17693–17698.

[44] Gregory, R., Saunders, J. R. & Saunders, V. A. (2006) The Paton individual-based model legacy. *Biosystems*, **85(1)**:46–54.

[45] Gregory, R., Saunders, J. R. & Saunders, V. A. (2008) Rule-based modelling of conjugative plasmid transfer and incompatibility. *Biosystems*, **91(1)**:201–215.

[46] Gregory, R., Saunders, V. A. & Saunders, J. R. (2008) Rule-based computing system for microbial interactions and communications: Evolution in virtual bacterial populations. *Biosystems*, **91(1)**:216–230.

[47] Guet, C. C., Elowitz, M. B., Hsing, W. & Leibler, S. (2002) Combinatorial synthesis of genetic networks. *Science*, **296(5572)**:1466–1470.

[48] Guido, N. J., Wang, X., Adalsteinsson, D., McMillen, D., Hasty, J., Cantor, C. R., Elston, T. C. & Collins, J. J. (2006) A bottom-up approach to gene regulation. *Nature*, **439(7078)**:856–860.

[49] Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol*, **210(9)**:1518–1525.

[50] Hardiman, T., Lemuth, K., Keller, M. A., Reuss, M. & Siemann-Herzberg, M. (2007) Topology of the global regulatory network of carbon limitation in *Escherichia coli*. *J Biotechnol*, **132(4)**:359–374.

[51] Herring, C. D., Raghunathon, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., Albert, T. J., Blattner, F. R., van den Boom, D., Cantor, C. R. & Palsson, B. O. (2006) Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat Genet*, **38(12)**:1406–1412.

[52] Hintze, A. & Adami, C. (2008) Evolution of complex modular biological networks. *PLoS Comput Biol*, **4(2)**:e23.

[53] Hirota, T. & Fukada, Y. (2004) Resetting mechanism of central and peripheral circadian clocks in mammals. *Zoolog Sci*, **21(4)**:359–368.

[54] Holland, J. H. (1992) *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence.* MIT Press, Cambridge, MA.

[55] Hughes, A. L. (2005) Gene duplication and the origin of novel proteins. *Proc Natl Acad Sci USA*, **102(25)**:8791–8792.

[56] Ingram, P. J., Stumpf, M. P. H. & Stark, J. (2006) Network motifs: Structure does not determine function. *BMC Genomics*, **7(108)**.

[57] Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M. & Serrano, L. (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, **452(7189)**:840–845.

[58] Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C. R., Tanabe, A., Golden, S. S., Johnson, C. H. & Kondo, T. (1998) Expression of a gene cluster *kaiABC* as a circadian feedback process in cyanobacteria. *Science*, **281(5382)**:1519–1523.

[59] Jacob, F. & Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, **3**:318–356.

[60] Jenkins, D. J. & Stekel, D. J. (2008) Effects of signalling on the evolution of gene regulatory networks. In S. Bullock, J. Noble, R. Watson & M. A. Bedau, editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 289–296. MITPress, Cambridge, MA.

[61] Jenkins, D. J. & Stekel, D. J. (2009) A new model for investigating the evolution of transcription control networks. *Artif Life*, **15(3)**.

[62] Kærn, M., Elston, T. C., Blake, W. J. & Collins, J. J. (2005) Stochasticity in gene expression: From theories to phenotypes. *Nat Rev Genet*, **6(6)**:451–464.

[63] Kaplan, S., Bren, A., Dekel, E. & Alon, U. (2008) The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Mol Syst Biol*, **4(203)**.

[64] Kashtan, N. & Alon, U. (2005) Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci USA*, **102(39)**:13773–13778.

[65] Kashtan, N., Noor, E. & Alon, U. (2007) Varying environments can speed up evolution. *Proc Natl Acad Sci USA*, **104(34)**:13711–13716.

[66] Kauffman, S. A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*, **22(3)**:437–467.

[67] Kauffman, S. A. (1993) *Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, Oxford, U.K.

[68] Kepler, T. B. & Elston, T. C. (2001) Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys J*, **81(6)**:3116–3136.

[69] Keseler, I. M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A. G. & Karp, P. D. (2009) Ecocyc: A comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res*, **37(Database issue)**:D464–D470.

[70] Koch, S. J. & Wang, M. D. (2003) Dynamic force spectroscopy of protein-DNA interactions by unzipping DNA. *Phys Rev Lett*, **91(2)**:028103.

[71] Kreft, J.-U. (2004) Biofilms promote altruism. *Microbiology*, **150(8)**:2751–2760.

[72] Kreft, J.-U., Booth, G. & Wimpenny, J. W. (1998) BacSim, a simulator for individual-based modelling of bacterial colony growth. *Microbiology*, **144(12)**:3275–3287.

[73] Kreimer, A., Borenstein, E., Gophna, U. & Ruppin, E. (2008) The evolution of modularity in bacterial metabolic networks. *Proc Natl Acad Sci USA*, **105(19)**:6976–6981.

[74] Kubitschek, H. E. (1990) Cell volume increase in *Escherichia coli* after shifts to richer media. *J Bacteriol*, **172(1)**:94–101.

[75] Kunin, V., Goldovsky, L., Darzentas, N. & Ouzounis, C. A. (2005) The net of life: Reconstructing the microbial phylogenetic network. *Genome Res*, **15(7)**:954–959.

[76] Kuo, P. D., Banzhaf, W. & Leier, A. (2006) Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems*, **85(3)**:177–200.

[77] Kurata, H., El-Samad, H., Iwasaki, R., Ohtake, H., Doyle, J. C., Grigorova, I., Gross, C. A. & Khammash, M. (2006) Module-based analysis of robustness tradeoffs in the heat shock response system. *PLoS Comput Biol*, **2(7)**:e59.

[78] Lawrence, J. G. & Hendrickson, H. (2003) Lateral gene transfer: When will adolescence end? *Mol Microbiol*, **50(3)**:739–749.

[79] Leier, A., Kuo, P. D. & Banzhaf, W. (2007) Analysis of preferential network motif generation in an artificial regulatory network model created by duplication and divergence. *Advances in Complex Systems*, **10**:155–172.

[80] Lenski, R. E., Ofria, C., Pennock, R. T. & Adami, C. (2003) The evolutionary origin of complex features. *Nature*, **423(6936)**:139–144.

[81] Lynch, M. (2007) The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet*, **8(10)**:803–813.

[82] Lynch, M. (2007) The frailty of adaptive hypothesis for the origins of organismal complexity. *Proc Natl Acad Sci USA*, **104(Suppl 1)**:S8597–S8604.

[83] Magnin, T., Lord, M. & Yudkin, M. D. (1997) Contribution of partner switching and SpoIIAA cycling to regulation of σ^F activity in sporulating *Bacillus subtilis*. *J Bacteriol*, **179(12)**:3922–3927.

[84] Maharjan, R., annd L. Notley-McRobb, S. S. & Ferenci, T. (2006) Clonal adaptive radiation in a constant environment. *Science*, **313(5786)**:514–517.

[85] Mangan, S. & Alon, U. (2003) Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci USA*, **100(21)**:11980–11985.

[86] Martínez-Antonio, A. & Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol*, **6(5)**:482–489.

[87] Mazurie, A., Bottani, S. & Vergassola, M. (2005) An evolutionary and functional assessment of regulatory network motifs. *Genome Biol*, **6(4)**:R35.

[88] McAdams, H. H. & Arkin, A. (1997) Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA*, **94(3)**:814–819.

[89] Meshi, O., Shlomi, T. & Ruppin, E. (2007) Evolutionary conservation and over-representation of functionally enriched network patterns in the yeast regulatory network. *BMC Syst Biol*, **1**:1.

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

[91] Moran, N. A. & Mira, A. (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol*, **2(12)**:research0054.

[92] Myers, C. L., Robson, D., Wible, A., Hibbs, M. A., Chiriac, C., Theesfeld, C. L., Dolinski, K. & Troyanskaya, O. G. (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol*, **6(13)**:R114.

[93] Nath, K. & Koch, A. L. (1970) Protein degradation in *Escherichia coli*: 1. measurement of rapidly and slowly decaying components. *J Biol Chem*, **245(11)**:2889–2900.

[94] Neidhardt, F. C., Ingraham, J. L. & Schaechter, M. (1990) *Physiology of the bacterial cell: A molecular approach.* Sinauer Associates Inc., Sunderland, MA.

[95] Newman, M. E. J. (2006) Modularity and community structure in networks. *Proc Natl Acad Sci USA*, **103(23)**:8577–8582.

[96] Ofria, C. & Wilke, C. O. (2004) Avida: A software platform for research in computational evolutionary biology. *Artif Life*, **10(2)**:191–229.

[97] Palsson, B. O. (2006) *Systems biology: Properties of reconstructed networks.* Cambridge University Press, Cambridge, U.K.

[98] Peterson, C. N., Mandel, M. J. & Silhavy, T. J. (2005) *Escherichia coli* starvation diets: Essential nutrients weigh in distinctly. *J Bacteriol*, **187(22)**:7549–7553.

[99] Pfeiffer, T., Soyer, O. S. & Bonhoeffer, S. (2005) The evolution of connectivity in metabolic networks. *PLoS Biol*, **3(7)**:e228.

[100] Philippe, N., Crozat, E., Lenski, R. E. & Schneider, D. (2007) Evolution of global regulatory networks during a long-term experiment with *Escherichia coli*. *Bioessays*, **29(9)**:846–860.

[101] Quayle, A. P. & Bullock, S. (2006) Modelling the evolution of genetic regulatory networks. *J Theor Biol*, **238(4)**:737–753.

[102] Reil, T. (1999) Dynamics of gene expression in an artificial genome - implications for biological and artificial ontogeny. In D. Floreano, J. D. Nicoud & F. Mondada, editors, *Proceedings of the 5th European Conference on Advances in Artificial Life*, pages 457–466. Springer Verlag, Heidelburg, Germany.

[103] Rosenfeld, N., Elowitz, M. B. & Alon, U. (2002) Negative autoregulation speeds the response times of transcription networks. *J Mol Biol*, **323(5)**:785–793.

[104] Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. & Collado-Vides, J. (2000) Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc Natl Acad Sci USA*, **97(12)**:6652–6657.

[105] Santillán, M. & Mackey, M. C. (2001) Dynamic regulation of the tryptophan operon: A modeling study and comparison with experimental data. *Proc Natl Acad Sci*, **98(4)**:1364–1369.

[106] Santos, M., Zintzaras, E. & Szathmáry, E. (2004) Recombination in primeval genomes: A step forward but still a long leap from maintaining a sizeable genome. *J Mol Evol*, **59(4)**:507–519.

[107] Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet*, **31(1)**:64–68.

[108] Shetty, R. P., Endy, D. & Knight, Jr, T. F. (2008) Engineering biobrick vectors from biobrick parts. *J Biol Eng*, **2**:5.

[109] Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. aps. *Nature*, **407(6800)**:81–86.

[110] Siegal, M. L. & Bergman, A. (2002) Waddington's canalization revisited: Developmental stability and evolution. *Proc Natl Acad Sci USA*, **99(16)**:10528–10532.

[111] Simpson, M. L., Cox, C. D. & Sayler, G. S. (2004) Frequency domain chemical langevin analysis of stochasticity in gene transcriptional regulation. *J Theor Biol*, **229(3)**:383–394.

[112] Singh, A. H., Wolf, D. M., Wang, P. & Arkin, A. (2008) Modularity of stress response evolution. *Proc Natl Acad Sci USA*, **105(21)**:7500–7505.

[113] Soyer, O. S. & Bonhoeffer, S. (2006) Evolution of complexity in signalling pathways. *Proc Natl Acad Sci USA*, **103(44)**:16337–16347.

[114] Stekel, D. J. & Jenkins, D. J. (2008) Strong negative self regulation of prokaryotic transcription factors increases the intrinsic noise of protein expression. *BMC Syst Biol*, **2(6)**.

[115] Sundararaj, S., Guo, A., Habibi-Nazhad, B., Rouani, M., Stothard, P., Ellison, M. & Wishart, D. S. (2004) The CyberCell Database (ccdb): A comprehensive, self-updating, relational database to coordinate and facilitate *in silico* modeling of *Escherichia coli*. *Nucleic Acids Res*, **32(Database issue)**:D293–D295.

[116] Swain, P. S., Elowitz, M. B. & Siggia, E. D. (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA*, **99(20)**:12795–12800.

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

[118] Tobisch, S., Zühlke, D., Bernhardt, J., .Stülke, J. & Hecke, M. (1999) Role of ccpa in regulation of the central pathways of carbon catabolism in *Bacillus subtilis*. *J Bacteriol*, **181(22)**:6996–7004.

[119] Tononi, G., Sporns, O. & Edelman, G. M. (1999) Measures of degeneracy and redundancy in biological networks. *Proc Natl Acad Sci USA*, **96(6)**:3257–3262.

[120] Trevino, V. & Falciani, F. (2006) GALGO: An R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, **22(9)**:1154–1156.

[121] van Nimwegen, E., Crutchfield, J. P. & Huynen, M. (1999) Neutral evolution of mutational robustness. *Proc Natl Acad Sci USA*, **96(17)**:9716–9720.

[122] van Noort, V., Snel, B. & Huynen, M. A. (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep*, **5(3)**:280–284.

[123] Wagner, A. (1994) Evolution of gene networks by gene duplications: A mathematical model and its implication on genome organization. *Proc Natl Acad Sci USA*, **91(10)**:4387–4391.

[124] Wagner, A. (2005) Energy constraints on the evolution of gene expression. *Mol Biol Evol*, **22(6)**:1365–1374.

[125] Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D. & Brown, P. O. (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci USA*, **99(9)**:5860–5865.

[126] Wilcox, J. L., Dunbar, H. E., Wolfinger, R. D. & Moran, N. A. (2003) Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Mol Microbiol*, **48(6)**:1491–1500.

[127] Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E. & Adami, C. (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, **412(6844)**:331–333.

[128] Woese, C. R. (2002) On the evolution of cells. *Proc Natl Acad Sci USA*, **99(13)**:8742–8747.

[129] Yildirim, N. & Mackey, M. C. (2003) Feedback regulation in the lactose operon: A mathematical modeling study and comparison with experimental data. *Biophys J*, **84(5)**:2841–2851.

[130] Zheng, D., Constantinidou, C., Hobman, J. L. & Minchin, S. D. (2004) Identification of the crp regulon using *in vitro* and *in vivo* transcriptional profiling. *Nucleic Acids Res*, **32(19)**:5874–5893.