Artificial Ontogenies:

A Computational Model of the Control and Evolution of Development

by

Nicholas Lewis Geard

BSc (Hons)

A thesis submitted to the

School of Information Technology and Electrical Engineering

The University of Queensland

for the degree of

DOCTOR OF PHILOSOPHY

October, 2006

Statement of originality

I hereby declare that the work presented in this thesis is, to the best of my knowledge, original and my own work, except as acknowledged in the text; and that the material has not been submitted, either in whole or in part, for a degree at this or any other university.

Nicholas Lewis Geard

Abstract

Understanding the behaviour of biological systems is a challenging task. Gene regulation, development and evolution are each a product of nonlinear interactions between many individual agents: genes, cells or organisms. Moreover, these three processes are not isolated, but interact with one another in an important fashion. The development of an organism involves complex patterns of dynamic behaviour at the genetic level. The gene networks that produce this behaviour are subject to mutations that can alter the course of development, resulting in the production of novel morphologies. Evolution occurs when these novel morphologies are favoured by natural selection and survive to pass on their genes to future generations.

Computational models can assist us to understand biological systems by providing a framework within which their behaviour can be explored. Many natural processes, including gene regulation and development, have a computational element to their control. Constructing formal models of these systems enables their behaviour to be simulated, observed and quantified on a scale not otherwise feasible.

This thesis uses a computational simulation methodology to explore the relationship between development and evolution. An important question in evolutionary biology is how to explain the direction of evolution. Conventional explanations of evolutionary history have focused on the role of natural selection in orienting evolution. More recently, it has been argued that the nature of development, and the way it changes in response to mutation, may also be a significant factor.

A network-lineage model of artificial ontogenies is described that incorporates a developmental mapping between the dynamics of a gene network and a cell lineage representation of a phenotype. Three series of simulation studies are reported, exploring: (a) the relationship between the structure of a gene network and its dynamic behaviour; (b) the characteristic distributions of ontogenies and phenotypes generated by the dynamics of gene networks; (c) the effect of these characteristic distributions on the evolution of ontogeny.

The results of these studies indicate that the model networks are capable of generating a diverse range of stable behaviours, and possess a small yet significant sensitivity to perturbation. In the context of developmental control, the intrinsic dynamics of the model networks predispose the production of ontogenies with a modular, quasi-systematic structure. This predisposition is reflected in the structure of variation available for selection in an adaptive search process, resulting in the evolution of ontogenies biased towards simplicity. These results suggest a possible explanation for the levels of ontogenetic complexity observed in biological organisms: that they may be a product of the network architecture of developmental control.

By quantifying complexity, variation and bias, the network-lineage model described in this thesis provides a computational method for investigating the effects of development on the direction of evolution. In doing so, it establishes a viable framework for simulating computational aspects of complex biological systems.

List of publications

The following is a list of publications that were produced during the period of candidature. Publications related to work appearing in this thesis have been highlighted (⋆).

Geard, N., Wiles, J., Hallinan, J. Tonkes, B. & Skellett, B., (2002). A comparison of neutral landscapes – NK, NKp and NKq. In D. B. Fogel, M. A. El-Sharkawi, X. Yao, G. Greenwood, H. Iba, P. Marrow and M. Shackleton (editors), *Proceedings of the 2002 Congress on Evolutionary Computation (CEC2002)*, pp. 205–210, Piscataway, NJ: IEEE Press.

Geard, N. & Wiles, J., (2002). Diversity maintenance on neutral landscapes: an argument for recombination. In D. B. Fogel, M. A. El-Sharkawi, X. Yao, G. Greenwood, H. Iba, P. Marrow and M. Shackleton (editors), *Proceedings of the 2002 Congress on Evolutionary Computation (CEC2002)*, pp. 211–213, Piscataway, NJ: IEEE Press.

Watson, J., Geard, N. & Wiles, J., (2002). Stability and task complexity: a neural network model of genetic assimilation. In R. Standish, M. Bedau and H. Abbass (editors), *Artificial Life VIII: Proceedings of the Eighth International Conference on Artificial Life*, pp. 153–156, Cambridge, MA: The MIT Press.

Geard, N. & Wiles, J., (2003). Structure and dynamics of a gene network model incorporating small RNAs. In R. Sarker, R. Reynolds, H. Abbass, K.-C. Tan, R. McKay, D. Essam and T. Gedeon (editors), *Proceedings of the 2003 Congress on Evolutionary Computation (CEC2003)*, pp. 199–206, Piscataway, NJ: IEEE Press.

Watson, J., Geard, N. & Wiles, J., (2004). Towards more biological mutation operators in gene regulation studies. *BioSystems* 113:239–248.

⋆ Geard, N. & Wiles, J., (2005). A gene network model for developing cell lineages. *Artificial Life* 11(3):249–268.

Copren, K. A. & Geard, N., (2005). An individual based model examining the emergence of cooperative recognition in a social insect. *Sociobiology* 46(2):349–361.

⋆ Geard, N., Willadsen, K. & Wiles, J., (2005). Perturbation analysis: a complex systems pattern. In Abbass, H., Bossamaier, T., and Wiles, J. (editors), *Recent Advances in Artificial Life*, pp. 69–84, Singapore, World Scientific Publishing.

Rudge, T. & Geard, N., (2005). Evolving gene regulatory networks for cellular morphogenesis. In Abbass, H., Bossamaier, T., and Wiles, J. (editors), *Recent Advances in Artificial Life*, pp. 239–252, Singapore, World Scientific Publishing.

Skellett, B., Cairns, B., Geard, N., Tonkes, B. & Wiles, J., (2005). Rugged NK landscapes contain the highest peaks. In H.-G. Beyer et al. (editors) *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2005)*, pp. 579–584, New York, NY, ACM Press.

⋆ Geard, N. & Wiles, J., (2006). Investigating ontogenetic space with developmental cell lineages. In Rocha, L. M., Yaeger, L. S., Bedau, M. A., Floreano, D., Goldstone, R. L., and Vespignani, A. (editors) *Artificial Life X: Proceedings of the Tenth International Conference on the Synthesis and Simulation of Living Systems*, pp. 56-62, Cambridge, MA. The MIT Press/Bradford Books.

Acknowledgments

I thank my advisor, Janet Wiles, for her guidance throughout the research and writing of this thesis: I have learnt much from her unflaggingly enthusiastic approach to research, and also her acumen regarding the more prosaic aspects of academia.

Numerous fellow students deserve thanks for advice, technical support, distraction, and coffee-fuelled discussion of matters both relevant and irrelevant to research. In particular, my thanks go to Kai, James and John in this regard.

Many other people have helped to inspire, develop and clarify the ideas in this thesis. At the outset of my studies, the faculty and students at the Santa Fe Institute Complex Systems Summer School supplied a particularly lively introduction to interdisciplinary research. Over the years I have also benefited from the input of Ricardo Azevedo, John Mattick, Peter Tiňo, Brad Tonkes, and others.

Outside of university, I am grateful to my friends, who have kept me in touch with a world apart from research, and tolerated my thesis-induced absences.

My family have my deep thanks for their continual and unconditional support and encouragement over many years. Their belief in me has sustained me when I may otherwise have faltered.

My final thanks are reserved for Oanh, who, more than anyone else, has provided the love, support and understanding that made this work possible.

The research reported in this thesis was funded primarily by an Australian Postgraduate Award. I gratefully acknowledge two further sources of funding. First, the School of Information Technology and Engineering at The University of Queensland provided a top-up scholarship in my first year, tutoring work and funding for conference travel. Second, the ARC Centre for Complex Systems provided a top-up scholarship in my third year and ongoing research work.

Contents

List of figures

List of tables

Chapter 1

Introduction

Each living organism we see today is a product of two remarkable biological processes: development and evolution. Every multicellular adult organism was once a single egg cell. A complex sequence of genetic, cellular and environmental interactions has resulted in the development of that egg into an adult organism. The ancestors of that organism were simple amoebae. Over millions of years of evolution, the descendants of that amoebae have diversified to produce a rich array of phenotypic forms. An enduring question in evolutionary biology is how organisms have come to have their current forms. The conventional answer provided by the neo-Darwinian paradigm emphasises the role of natural selection in orienting evolution (Fisher, 1930, Mayr, 2001). Out of a multitude of variant forms, some were better adapted to their environments, and these had a better chance of surviving and passing on their genetic material encoding that form to offspring.

In the last two decades, researchers in the field of evolutionary developmental biology have argued that natural selection alone is an incomplete explanation of evolutionary history (Gilbert et al., 1996, Hall, 1999, Raff, 2000, Arthur, 2002a). They suggest that the gradual phenotypic changes wrought by microevolutionary processes such as mutation are an insufficient explanation of the evolution of significant differences between species and body plans. They claim that understanding evolutionary history requires an appreciation of the role played by development. In particular, Arthur (2000, 2002b, 2004a,b) has proposed that development can bias the structure of phenotypic variation that natural selection acts on, and hence plays an equally important role in orienting evolution.

Contemporary evolutionary theory is built upon findings from a diverse range of fields. Genetics, palaeontology, molecular and developmental biology, among others, have all contributed ideas and data (Mayr, 2001, Ridley, 1996). In many of these fields, the use of models has been vital for the formulation, communication and exploration of hypotheses. The form that these models take has changed in response to both the requirements of the questions being addressed and the types of modelling methodologies available. One of the more recent additions to the range of options is computational modelling.

Studies of evolution are plagued by two issues. The first is an absence of data. The evolution of life has been occurring for billions of years, and the only evidence we have for all but the most recent fraction of this history is fossils. Furthermore, we have only one set of data—there are no alternative histories of evolution against which to compare. The second issue is the complexity of evolutionary systems. The history of evolution emerges from dynamic processes at multiple levels of organisation: populations, organisms, cells and genes. The processes that occur at any one of these levels can be complex and difficult to understand; that interactions also occur *between* levels of organisation compounds the difficulty.

Computational models can assist us to address both of these issues. When a computer is used to simulate evolution, all of the data on the history of that simulation can be stored and analysed. The simulation can be run multiple times, under different conditions, to enable the exploration of different parameters, assumptions and hypotheses. As well as allowing repetition, computers are also valuable for dealing with systems containing complex and nonlinear patterns of interaction between their constituent elements. The design and implementation of a computational model requires a researcher to cut through peripheral detail and focus on the core dynamics of a system, assisting in the management of complexity.

Development and its place in evolutionary theory

Development is the transformation of a single egg cell into a multicellular organism. It involves the growth, division, differentiation and organisation of cells to produce ordered patterns and forms (Wolpert, 1998). During embryonic development, the range of potential fates into which a cell can differentiate is progressively restricted to produce complex patterns of terminal identities. Two related problems are to understand how the generation of these patterns is controlled, and how they have evolved (Carroll, 2000, Carroll et al., 2001, Pires-daSilva and Sommer, 2003). A key locus of developmental control is the genetic regulatory system encoded in an organism's genome (Davidson et al., 2003). The interaction between a genetic system, epigenetic properties of cells and tissues, and environmental context determines the features of a developing organism.

Some aspects of development are well understood: the patterns of division and differentiation that produce some organisms (Sulston et al., 1983, Nishida, 1987, Houthoofd et al., 2003); and in a few cases, even the genetic basis for these developmental events (Kaletta et al., 1997, Lin et al., 1998, Yuh et al., 1998, Maduro, 2002, Inoue et al., 2005). In addition, theoretical and empirical advances have been made in understanding the dynamic and structural properties of networks of interacting genes (Barabási and Oltvai, 2004, Albert, 2005), including progress towards mapping the networks controlling the development of specific organisms (Arnone and Davidson, 1997, Davidson, 2001, Davidson et al., 2003, Oliveri and Davidson, 2004). Still, current understanding of developmental control is far from complete (Molin et al., 1999).

Development is important from an evolutionary perspective because it mediates between the genotypic level of description, at which heritable variation occurs, and the phenotypic level of description, at which natural selection acts (Arthur, 2004b). For many decades theories of evolution have focused on genes and how the frequency of their occurrence in a population changes under various sorting processes. The two mechanisms generally considered to play a primary role in sorting variation are: natural selection, a nonrandom sorting process that correlates survival and reproductive success with increased frequency in future generations; and genetic drift, a stochastic sorting process resulting from the finite size of natural populations and the contingencies associated with survival and reproduction (Ridley, 1996). This view—arising from the synthesis between genetics and evolution in the 1940s and driven by the mathematical models of population geneticists—left a strictly secondary role for development (Gilbert et al., 1996).

The calls for a 'new synthesis' to reincorporate development into evolutionary theory result from complaints about the explanatory sufficiency of natural selection in the conventional view of evolution: natural selection succeeds in explaining the conditions under which the frequency of existing types of organism in a population are altered, but is less successful at explaining the appearance of novel types, the occurrence of rapid (in evolutionary terms) transitions between different types and the evolutionary relationships between diverse species (Bonner, 1981, Goodwin et al., 1982, Gilbert et al., 1996, Arthur, 2004a). Related to these issues is the question of what determines the direction of evolution, regardless of whether it involves variation already existing in a population or the appearance of novel phenotypes (Fusco, 2001, Arthur, 2004b). It has been argued that the structure of variation introduced into a population may have a significant effect on the direction of evolution (Yampolsky and Stoltzfus, 2001, Arthur, 2002b). As mediator between genotype and phenotype, development is in a position to bias, constrain or drive evolution through its effect on the structure of variation.

The interpretation of developmental bias as an important evolutionary mechanism has not been universally accepted. One criticism is that quantitative genetics already includes the facility for measuring constraint due to development (Cheverud, 1984). More recently, experiments involving artificial selection on butterfly wing patterns have been used to demonstrate how 'unconstrained' morphological evolution can be in the face of selection (Beldade et al., 2002). However, at least some of these differences of opinion may be due to differences in the definition and interpretation of 'constraints', particularly whether they are relative or absolute with respect to the power of selection to overcome them (Arthur, 2003, Beldade and Brakefield, 2003). Arthur (2004b) argues that a consensus position with regard to developmental bias is slowly being arrived at:

> Perhaps the best way to describe the current state of affairs is that there is general agreement that developmental bias *may* be an important determinant of evolutionary directionality, but that whether it actually *is* so remains in the balance because we lack the relevant evidence to reach a clear conclusion on this issue. (Arthur, 2004b, p. 287)

Computational modelling is a methodology well-suited to exploring the possible sources and effects of developmental bias from a theoretical perspective. The design and implementation of models can clarify, quantify and refine the terms in which hypotheses are phrased. In doing so, it can help establish a conceptual platform from which empirical evidence may be obtained.

Investigating developmental bias

To summarise the relationship between gene regulation, development and evolution: Development consists of a sequence of cellular events (an *ontogeny*) guided by interactions at a genetic, epigenetic and environmental level. The developmental genetic systems governing this process are subject to mutations that can alter ontogenies, resulting in the production of novel phenotypes. Evolution occurs when these novel variants are favoured by natural selection and survive to pass on their genes to future generations.

There are two points at which the direction of this evolutionary cycle can be altered: the introduction of phenotypic variation and the selection of that variation. The effect of natural selection on the direction of evolution has been widely studied—the structure of novel variation, less widely so. To obtain a deeper understanding of how evolution is oriented, we require a framework for investigating how genetic variation is transformed into phenotypic variation by development, and what implication the structure of this variation has for the direction of evolution. This research agenda can be framed as a sequence of logical steps, corresponding to the points of enquiry addressed in this thesis.

- Cell fate is largely a property of gene expression dynamics. How does cell fate potential vary with structural properties of an underlying genetic control system?

- During development, gene expression is coordinated in both a spatial and a temporal fashion. How is this achieved by a genetic network, and how is the space of possible ontogenies shaped by the dynamic properties of a developmental genetic system?

- As discussed, evolution consists of two stages: the generation of variation and the filtering of that variation by natural selection. What effect does the structure of generated variation have on the evolution of ontogeny?

- The definition of a modelling task imposes constraints on the design of a suitable model. What are the requirements of a modelling framework suitable for addressing the above questions? How can a model be designed to satisfy these requirements?

The *theoretical* claim of this thesis is that features of evolved ontogenies can be biased both by the characteristic dynamics of developmental genetic systems, and by the way that mutation to developmental genetic systems produces the novel variation on which selection acts.

The *methodological* claim of this thesis is that computational models are an effective tool for quantifying and exploring the relationship between developmental bias and evolution.

1.1 Overview of the thesis

The overall goals of this thesis are:

- to develop a computational simulation model that integrates mechanisms operating at the level of gene regulation, development and evolution in a way that is efficient, plausible and useful;
- to use this model to investigate the control of development by the dynamics of a genetic system and the evolution of development by modification to this genetic system; and
- to investigate the extent to which the nature of the developmental system can affect the direction of adaptive evolution.

Figure 1.1 illustrates the organisation of this thesis. Biological background for both the model of development and the issue of developmental bias are provided in Chapter 2. Methodological background for the model, and the model itself, are described in Chapter 3. The empirical studies undertaken using the model are reported in Chapters 4, 5 and 6. The implications of the studies for the theoretical issue is discussed in Chapter 7, along with an evaluation of the methodology.

In Chapter 2 development is reviewed from three perspectives. First, the primary mechanisms of development (cell division, differentiation and morphogenesis) are described. Particular attention is paid to the invariant patterns of development observed in invertebrates such as *Caenorhabditis elegans* and *Halocynthia roretzi*. Next, development is reviewed from a control perspective and the roles played by both genetic and environmental factors are described. The central role

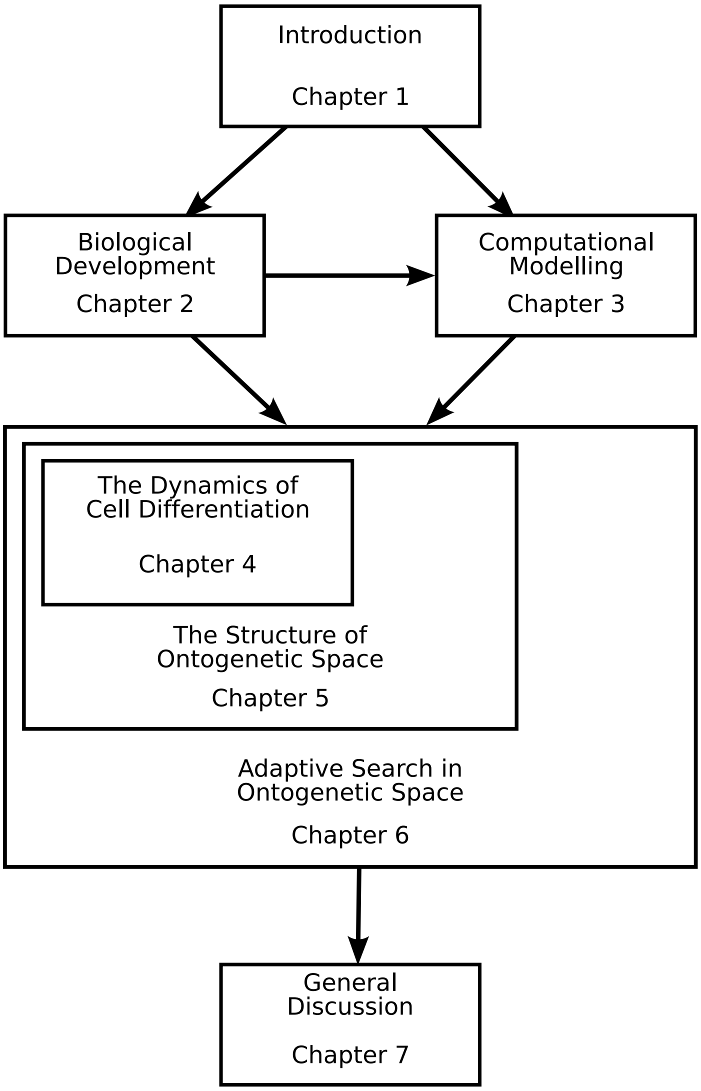


Figure 1.1: The organisation of this thesis. Chapters 2 and 3 review background material and introduce the model used in this thesis. Chapters 4, 5 and 6 report the studies carried out using this model.

played by the genetic regulatory system as a heritable encoding of a developmental programme is emphasised. Finally, development is reviewed from an evolutionary perspective, focusing on the argument that development is an important mechanism in orienting evolution. Chapter 2 concludes by reiterating the questions to be addressed in this thesis and identifying requirements for the design of a suitable computational model.

Chapter 3 takes up the issue of computational modelling in more detail. First, various types of models are reviewed to provide context for why computational modelling is an appropriate methodology for this thesis. Existing models of gene regulation, development and evolution are then reviewed in light of the constraints identified in Chapter 2. The two components of the network-lineage model used in this thesis are then described: The Dynamic Recurrent Gene Network (DRGN) model of genetic control and a cell lineage model of development.

Chapters 4, 5 and 6 report three sets of studies that were carried out to address the thesis questions. These three chapters focus on gene expression, development and evolution respectively, to build up a comprehensive framework for modelling the spaces in which an evolutionary developmental process operates (Figure 1.2). The claims of this thesis about the effect of developmental bias on the orientation of evolution are addressed explicitly in Chapter 6. Prior to that, it is necessary to understand the behaviour of the DRGN model, and the nature of the ontogenies that it generates. Chapters 4 and 5 therefore focus on these two components of the model.

The dynamic recurrent network used to model gene regulation is a complex, high dimensional system capable of a wide variety of dynamic behaviours. The aim of the first set of studies (reported in Chapter 4) was to understand how this repertoire of behaviours—analogous to possible cell fates in a biological system—depends on the structure of the network. Methods for counting and classifying the stable attractors of a network are first reviewed to provide a set of tools for analysing and visualising network behaviour. These tools are then applied to parameterised ensembles of random networks to quantify the relationship between network structure and dynamic behaviour. Further review and analysis reveals that the number of stable attractors displayed by a network of a given size is considerably lower than the theoretical maximum. This discrepancy is investigated and explanations are suggested in terms of the coupling between nodes.

The next set of studies (reported in Chapter 5) extended the scope of investigation to consider the ontogenies that are generated by network dynamics. The aim of these studies was to characterise the space of possible ontogenies, represented here in the form of cell lineages. This task required both a metric for characterising an individual cell lineage and a means of representing and visualising the relationship between cell lineages. Several existing measures of complexity were

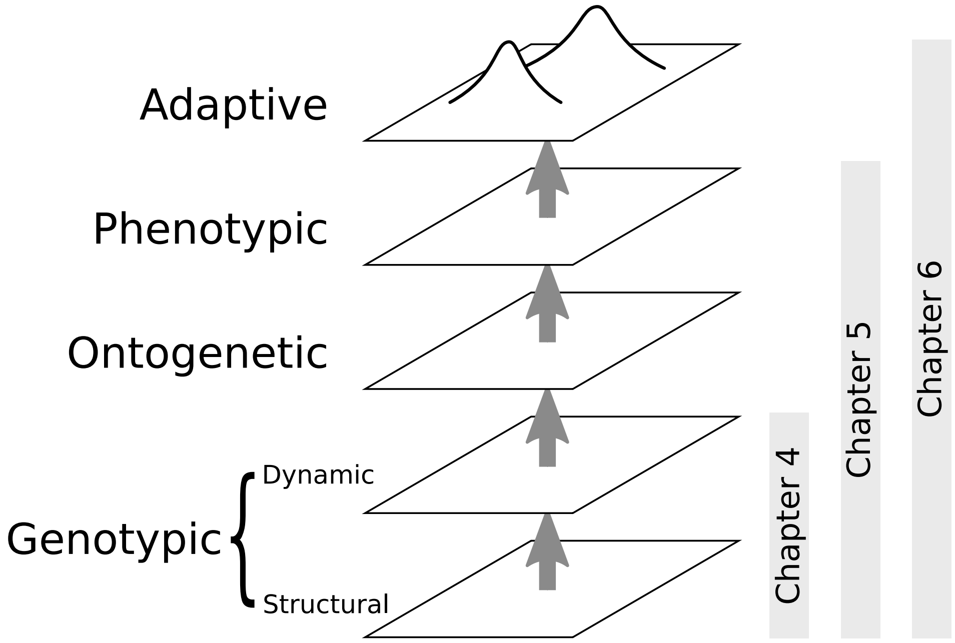


Figure 1.2: The relationship between the different concepts of space used in this thesis and chapters in which they are explored. The structural component of **genotypic space** comprises the set of all possible genomes, represented here as gene regulatory networks. Chapter 4 considers the mapping between the structure of a network and its dynamic behaviours. Ontogeny—the trajectory of developmental events that transforms an egg into an embryo—is derived from a gene network's dynamics. **Ontogenetic space** contains all possible developmental trajectories and characterising this space is the focus of Chapter 5. The end product of an ontogeny is a phenotype, the observable characteristics of an individual. **Phenotypic space** contains all possible phenotypes. Natural selection acts on phenotypes: those that are better adapted to their environment are more likely to survive and pass on their genes to offspring. The level of adaptedness of a phenotype may be quantified through the use of a fitness function; the combined result of a fitness function applied across a range of phenotypes is known as a fitness landscape, or **adaptive space**. The mapping from phenotypic to adaptive space is considered in Chapter 6.

formalised and compared, but found to display certain limitations with respect to intuitive notions of what constituted a 'complex' lineage. A novel complexity metric was therefore designed to address these limitations. An interactive software tool, TreeView, was developed to address the visualisation requirement. This tool provides a qualitative insight into the gradients of complexity that defined ontogenetic space. A complementary quantitative characterisation was obtained through the use of parameterised ensembles of cell lineages. Perturbation analysis was used to evaluate the robustness of ontogeny with respect to both structural and dynamic sources of perturbation. The results of the simulations presented in this chapter suggest that the developmental mapping does bias the types of ontogenies and phenotypes produced by network genotypes.

The final set of studies (reported in Chapter 6) considered the implications of the bias discovered in the previous chapter for evolution. The first issue addressed was the possibility of mutation operators being a second source of bias. A comparison of eight different mutation operators indicated that they *do* differ with respect to the way they structure genetic variation. To eliminate mutation bias as a confounding factor in our investigation of developmental bias, a mutation operator was identified that minimised the amount of bias from this source. Adaptive walks were used to evaluate the capability of the model to generate ontogenies based on targets derived from the observed *C. elegans* and *H. roretzi* cell lineages. The results of these simulations indicate that bias due to the developmental mapping does affect the properties of the ontogenies located by an adaptive process.

Finally, Chapter 7 summarises the results obtained and discusses their implications for our understanding of the sources and effects of bias in evolution. The strengths and limitations of the computational modelling methodology are assessed and avenues for further investigation are briefly outlined.

Chapter 2

Development: A Genetic and Evolutionary Perspective

Development is the transformation of a genotype into a phenotype and is a bridge between understanding genetics and understanding evolution. The development of a multicellular organism from a zygote involves highly complex patterns of dynamic behaviour at the genetic level. Modifications to developmental genetic networks are the basis for the evolution of phenotypic forms.

Development is inherently a computational process: one form of information—the sequence of nucleotides that constitute an organism's genome—is transformed into another—the developed organism itself (Bray, 1995, Flake, 2000, Davidson et al., 2003). Development is also, like most biological phenomena, a very complicated process. In order to build a model of this process that can be simulated on a computer we need to distinguish the core computational features of a developmental system that must be included in a model from the morass of peripheral details.

The aims of this chapter are twofold. Firstly, to review development and gene regulation with an emphasis on identifying the computational aspects of these processes that can be incorporated into a practical model of developmental control. Secondly, to review the issue of developmental bias, framing the questions to be addressed in this thesis and translating them into a set of requirements for a suitable model.

This chapter begins by briefly reviewing the mechanisms of embryonic development (§2.1). We focus specifically on the early embryonic development of invertebrates, which are often characterised by a high degree of lineage invariance. The control aspects of development are then reviewed. While recognising an important role for the environment and epigenetic mechanisms, we focus here on the genetic processes that underly ontogeny (§2.2). Particular attention is paid to how these genes are connected into complex networks, whose dynamics are responsible for the control of cell division and differentiation. Development is then reviewed from an evolutionary perspective (§2.3). The argument that developmental bias can orient evolution is contrasted with the conventional view that natural selection is the primary evolutionary mechanism. The reader already familiar with evolutionary and developmental biology may choose to go directly to §2.4 where the questions motivating this thesis are restated and used to specify requirements for a modelling framework suitable for addressing them.

2.1 Mechanisms of early development

The development of an organism from a fertilised egg cell involves three primary types of change: growth, differentiation and morphogenesis. These processes are not isolated or sequential, but occur in parallel throughout development. This section describes each of these processes in turn. The embryonic development of the nematode *Caenorhabditis elegans* is then described in detail.

Growth

During development, the growth of an organism occurs through an increase in both the number of cells, via cell division, and the size of individual cells. There are two primary modes of cell division that occur in development. Early in development, a series of cleavage divisions occurs, during which the number of cells increases but there is little overall increase in cell mass. Later in development cell proliferation may occur, in which cell division is accompanied by an increase in cell mass (Wolpert, 1998).

When a cell divides, the two newly created cells are not necessarily identical. The contents of the parent cell may be divided asymmetrically between the two daughter cells, resulting in each differing in size or behaviour. Three different modes of division are possible (Stent, 1985):

Proliferative – the two daughter cells have identical behaviour to the parent cell (A → A A);

Stem-cell – one daughter cell has the same behaviour as the parent and the other daughter cell behaves differently (A → A B); and

Diversifying – the two daughter cells behave differently both to each other and to the parent cell (A → B C).

The cell divisions that occur early in embryonic development produce cells which, while physically undistinguished, are already poised to follow their own unique differentiation trajectory.

Differentiation

As cells divide, they undergo changes to their physical and chemical properties that eventually result in them differentiating into specialised cell types. A fertilised egg cell is totipotent: it has the capacity to produce cells that will differentiate into all types found in that organism. As development proceeds, the new cells that are created by division have increasingly restricted potency. At a certain point in its development, a cell's fate is said to be determined; after this point its fate is fixed and it is no longer possible for it to differentiate into any other type. Early evidence for this phenomenon was obtained by transposing cells from a frog blastula at different stages of development (Wolpert, 1998). If the cells were transposed early in development, they differentiated according to their new context. Cells transposed later in development maintained their original fate, despite the new context.

In addition to illustrating the changes in cell potency during differentiation, transposition experiments highlight the role of cellular context in determining cell fate. Inductive interactions between cells play an important part in cell differentiation by providing the signals to bring about specific behaviours. These signals are selective rather than instructive. They do not provide the information necessary to specify the changes to the target cell; their role is as a trigger, to select between the possible differentiation trajectories that the target cell is capable of following.

Besides their separate functional specialisations, differentiated cells have two further important characteristics: discrete identities and irreversibility (Waddington, 1957). While there is some debate about the level of similarity at which cells should be classified into distinct types, any categorisation will be essentially discontinuous. For example, bone cells and blood cells form discrete classes, with no intermediate class of 'half-blood, half-bone' cell. In addition, once a cell has differentiated, it's type remains fixed. Only in rare instances do differentiated cells return to an undifferentiated state, or transform into a cell of a distinctly different type (Wolpert, 1998).

The development of an embryo requires not only that each of its cells be of the appropriate type, but also that the embryo as a whole is organised into a functional form.

Morphogenesis

Like differentiation, the organisation of cells into functional patterns has both a chemical and a mechanical component. Chemical gradients are established that describe the axes along which an organism's body plan will be organised. The embryonic environment can be responsible for these chemical gradients, or they can be established by inductive signalling between cells (Wolpert, 1969). These gradients provide information to cells that enables them to differentiate into the type corresponding to their physical location.

Embryos also undergo a variety of mechanical behaviours that result in changes to the physical arrangement of cells, including cell migration and conformational changes such as contraction, elongation and invagination (Salazar-Ciudad et al., 2003). Physical properties of cells, such as differential adhesion (Hogeweg, 2000a) and cytoskeletal architecture (Ingber, 2005), play an important role in organising cells into appropriate forms.

The processes of growth, differentiation and morphogenesis overlap and inform each other (Salazar-Ciudad and Jernvall, 2004, Wolpert, 1998). Physical changes to the embryo, whether due to division or morphogenesis, bring cells into new chemical contexts that can affect their trajectories of differentiation. Similarly, as cells differentiate, the changes in their functional behaviour can produce morphological effects.

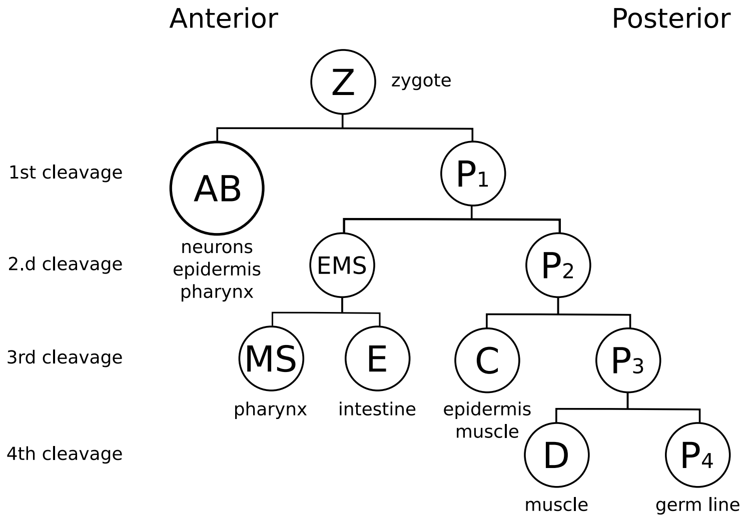


Figure 2.1: The cell lineage of early *C. elegans* embryogenesis. Each precursor cell cleavage results in the production of one somatic cell—which will divide to generate epidermal, intestinal, neural, and muscle cells—and a further precursor cell. The final precursor cell, P_4, gives rise to the germ line (redrawn from Sulston et al., 1983).

2.1.1 Invariant development and cell lineages

The early development of some invertebrate organisms is characterised by invariant patterns of cell division and differentiation. In these organisms, it is possible to identify the function and developmental history of each individual cell. The ontogeny of such an organism can be depicted as a cell lineage diagram (Figure 2.1). A cell lineage diagram captures elements of each of the three developmental processes described above: the pattern of division events, the identification of terminal cells with particular fates, and the allocation of terminal cells to particular positions (García-Bellido, 1985).

One of the first organisms for which a complete map of the cell lineage was obtained was the nematode worm *Caenorhabditis elegans* (Sulston et al., 1983). *C. elegans* occupies an extreme position on the spectrum of lineage determinism. The initial divisions produce a set of six founder cells that go on to generate particular subsets of the final organism. In contrast, most mammals undergo a period of rapid cell proliferation followed by extensive migration. Once close to their final positions, cells differentiate according to combinations of global and local signals. Any of the initially created cells may be destined for a particular fate (Wolpert, 1998).

The invariant development of *C. elegans* was initially thought to be the product of cell-autonomous mechanisms of fate specification. Ablation experiments, in which the neighbours of a particular cell are killed by laser, have been used to observe how each cell develops in the absence of its normal context. Several of the early embryonic cells have been isolated in such a fashion and observed to follow their normal developmental pathway, suggesting a source of control intrinsic to the cell (Sulston et al., 1983). A candidate mechanism for intrinsic control is the asymmetric segregation of regulatory information between daughter cells during division (Kenyon, 1985). Several specific factors that implement such a mechanism have since been discovered, including *pop-1* (Lin et al., 1998) and *lit-1* (Kaletta et al., 1997).

It is now known that several key decisions in the early development of *C. elegans* are controlled by cell-extrinsic mechanisms. In particular, predictable inductive interactions between early cells are responsible for establishing dorsal/ventral polarity and left/right asymmetry (Bowerman, 1998). In such cases, cell lineage may still play an essential role in a cell commitment by ensuring that an appropriately receptive cell is located so as to be exposed to the necessary inductive signals (Stent, 1998).

Sulston et al. (1983) concluded, upon mapping the cell lineage of *C. elegans*, that the complexity of the lineage was one of its most striking features. An open question concerning *C. elegans* development is how to explain this complexity. Although the six founder cells depicted in Figure 2.1 go on to produce predictable sublineages, there is no clear relationship between these initial branches of the cell lineage and terminal cell fate. For example, while the intestine and germ cells are all derived from the E and P_4 cells respectively, epidermal cells are derived from both the AB and C cells. Conversely, the daughters of the AB cell will eventually differentiate into neuronal, epidermal and muscle cells. A possible explanation is that the complexity may be a product of selective forces acting to increase the efficiency of the developmental process. One of the advantages of the polyclonal cell lineage of *C. elegans* is that most cells are created very close to their final location, reducing the need for cell migration (Sulston et al., 1983).

Recent evidence supporting this theory has been obtained by comparing the *C. elegans* cell lineage to that of the related nematode species *Pellioditis marina* and *Halicephalobus* sp. (Houthoofd et al., 2003). The lineage of *P. marina* has many similarities with that of *C. elegans*, including a comparable level of polyclonal specification events. *Halicephalobus* sp., in contrast, has a much simpler lineage consisting of a greater proportion of monoclonal specification events. The disadvantage of having a simpler specification mechanism however, is that more cell migration is required in order to correctly position cells, and the development of *Halicephalobus* sp. is considerably slower—650 minutes until muscle contraction as compared with 430 minutes for *C. elegans*. Houthoofd et al. (2003) examined the phylogenetic relationship between 31 nematode species (Vancoppenolle et al., 1999) and concluded that the simple lineage of *Halicephalobus* sp. was likely to be the primitive character from which the greater complexity of *C. elegans* was derived, under selection for more rapid embryonic development.

Challenging the notion that these cell lineages were as complex as they were perceived, Azevedo et al. (2005) developed an approach to measuring complexity based on the notion of the shortest algorithmic description of a lineage. Using this measure, the *C. elegans* lineage is actually comparatively *simpler* than that of other related nematode species—4.4% less complex than that of *P. marina*. The issue of cell lineage complexity is described further in Chapter 5, where several different measures, including that described by Azevedo et al. (2005) are compared.

In summary, development is a complex process that involves chemical and physical changes occurring both within and among individual cells. The significance of environmental factors and epigenetic mechanisms is evident (Schlichting and Pigliucci, 1998, Wong et al., 2005); however, the reliable transmission from parent to offspring of the developmental trajectory specific to a particular species suggests the importance of a genetic component to developmental control (Davidson et al., 2003). The role of the genetic regulatory system is reviewed in the following section.

2.2 The genetic component of developmental control

During development a cell can exhibit a number of behaviours, including division, apoptosis (programmed cell death), change in shape, movement and changes in state leading to differentiation. The proximate cause for these behaviours is the complement of proteins contained in a particular cell. The set of proteins that are present in a cell is determined by which genes are active in that cell. This section describes how differences in gene expression among cells are responsible for their different behaviours, the network architecture of the gene regulatory system, and the mapping between network dynamics and functional cell behaviours.

2.2.1 Differential gene expression

The primary feature that determines the function of a fully differentiated cell is the proteins it contains (Alberts et al., 1994). Similarly, the most important property characterising a developing cell is its pattern of gene activity (Wolpert, 1998). Externally, the cells of an early embryo may be virtually indistinguishable. Already, however, the particular pattern of gene activity exhibited by a cell determines the role it and its progeny will play in the fully developed organism. When a cell divides, this pattern of active and inactive genes is passed on to its daughter cells via positive regulatory feedback, and physical and chemical modifications to the genome (Wolpert, 1998).

While all cells in a developing embryo contain the same set of genes, the identity of that cell is determined by the subset of those genes that are switched on at any given time. The composition of that subset varies with the spatial and temporal location of the cell, and its expression history. The context of a cell results in its exposure to different chemical signals, either from other cells or the environment, while the history of gene activation events produces a unique regulatory state.

In a eukaryotic cell, the genetic regulatory system is encoded in the genome, which is located in the cell's nucleus. As well as carrying out various functions related to cell maintenance and function, a significant portion of the genetic system is involved in programming embryonic development (Wolpert, 1998, Davidson, 2001).

The process of gene expression begins when an RNA polymerase molecule binds to the start site of a gene, unwinds a section of DNA and uses one of the strands as a template to transcribe messenger RNA (mRNA) molecules. mRNA molecules are transported outside the cell nucleus to the cytoplasm, where they are translated into proteins. Proteins can either be structural, enabling a cell to fulfil its functional role in an organism, or they can re-enter the nucleus to regulate the expression of other genes. These regulatory proteins, known as transcription factors (TFs), interact with the promoter and control regions of a gene to either enhance or inhibit the transcription of that gene. Some TFs are required for any transcription to occur at all. Others play a role as activators, binding to enhancer sites located upstream or downstream of the gene to facilitate transcription. Yet another type acts as a repressor, either by blocking activator TFs, or by preventing the binding of RNA polymerase to a gene start site (Alberts et al., 1994).

When a cell divides, the set of TFs that determine its pattern of gene activation are divided between the daughter cells, so each will generally have a similar pattern of gene expression to its parent. On occasion, the distribution of TFs in the parent cell may be asymmetric (Jan and Jan, 1998). The two daughter cells will therefore inherit different sets of regulatory information and follow two unique developmental trajectories.

Inductive signals originating from other cells or the environment can also alter patterns of gene expression. In general, these signalling molecules bind to receptors found on the cell surface, and the signal is transmitted to the nucleus via a series of chemical events called a signal transduction pathway. The role of these signals is selective rather than instructive. They do not provide new information to a cell about what it should do. Rather the signals select one fate from among the relatively small number of possibilities defined by the current state of the cell (Wolpert, 1998). The fact that signals act as simple triggers, rather than complicated messages, means that signalling pathways can be very generic. A relatively small number of common pathways are used repeatedly, not only during the development of a single organism, but also across different species (Pires-daSilva and Sommer, 2003).

2.2.2 Gene regulatory networks

One of the goals inspired by the molecular revolution in biology was to identify the role of each gene responsible for both the functions of an individual cell, as well as the sequence of events involved in higher order processes like development. The initial models for such control were linear pathways of gene regulation, by analogy with the comparatively better understood metabolic pathways (Greenspan, 2001, Wilkins, 2002). A common experimental technique for identifying gene function was to render a particular gene inactive and observe the phenotypic events. This approach gained several of the first discovered genes their names, which were frequently chosen in light of the morphological response to their inactivation. For example, in the fruit fly *Drosophila melanogaster*, *numb*, *eyeless* and *tinman* mutants lack sensory neurons, eyes and heart respectively (Yohn, 2001). The perspective that there is a single gene for each biological function also resulted in a popular misconception of researchers being able to identify the gene 'for' a particular disease or character trait.

Evidence against the pathway model of gene regulation emerged from the discovery that, while a few genes were specific to a particular phenotypic trait, many others were not. Pleiotropy—the involvement of a single gene in multiple phenotypic functions—appeared widespread. One of the earliest proposals that the structure of regulatory apparatus may be more sophisticated than was then conceived came from Britten and Davidson (1969). They proposed a set of regulatory mechanisms by which multiple changes in gene activity could be initiated by a single event. At the same time, a more abstract model of gene regulation based on networks of Boolean switches was introduced by Kauffman (1969) (The Boolean network model is described further §3.2.1).

The shift from a pathway to a network view of gene regulation provided a new perspective on several aspects of developmental genetics. One was to explain the widespread effects of mutations to some, obviously highly pleiotropic, genes. A second, contrasting aspect was the apparent imperviousness of development to mutations in many other genes (Rutherford, 2000, Hartman IV et al., 2001, Kitano, 2004). That development was a remarkably robust process had been known for some time: Waddington (1942) reported one of the first investigations into what he termed developmental canalisation. He also introduced the concept of an epigenetic landscape, the valleys of which represented the stable phenotypic forms to which developmental trajectories reliably proceeded (Waddington, 1957).

The network architecture of gene regulation results in several mechanisms by which robustness can be attained. The functionality of a system may not be dependent upon a single gene, as implied by the pathway model. In a network, functionality may be duplicated across multiple genes, providing a level of redundancy that ensures robust behaviour even after damage (Krakauer and Plotkin, 2002). The feedback loops inherent in a gene network also appear to play a stabilising role by buffering the intrinsic noisiness of regulatory events involving very small numbers of molecules (Freeman, 2000). Classic examples of how such robustness including the switching mechanism in phage λ (Ptashne, 1992) and chemotaxis in the bacteria *Escherichia coli* (Alon et al., 1999). In each of these cases, system behaviour is stable across a wide range of parameter values.

2.2.3 Cells fates as attractors in dynamic space

A significant example of robustness in biological systems is the stability of individual cell types. It is clear that there is variation of cell fates across local regions of an organism and that these fates are discrete: there are no intermediate cell types between, say, epidermal cells and nerve cells. Yet the cells in these two adjacent regions presumably receive very similar signals from their shared regional environment. It appears that, at some point, a relatively small variation in the input to each set of cells—due to different spatial and temporal cues—has resulted in them following significantly different trajectories of differentiation. Furthermore, once these cells have fully differentiated, they display robustness to perturbation. Very few cells, once fully differentiated, are able to change to another type. What further changes a cell does undergo as a result of external signals are likely to be harmful, such as cancerous growth induced by radiation (Wolpert, 1998). The theory of dynamic systems provides an explanation for cell differentiation in the form of multistability: the possibility of a system existing in different stable states depending on its history. This idea was first suggested in the 1940s, but became more widely known after the development of the random Boolean network model of gene regulation in the 1960s (Thomas, 1998, reviews the history of this idea).

One of the enduring contributions made by Kauffman (1969, 1971) through his work on Boolean gene network models was the popularisation of the idea that attractors in these networks could be interpreted as different cell types. His argument was based on the ensemble properties of randomly connected Boolean networks sharing certain characteristics with the human genomic system (this argument is developed further by Kauffman, 1993). Briefly, cells in a biological organism are classified into types according to the complement of genes expressed within them. Assuming that the expression of a single gene is a binary decision, a system containing N genes can display 2^N different patterns of gene expression. It seems implausible that each of these would correspond to different stable cell types, as many will represent transient states through which a cell passes, or be unattainable in a real system. Kauffman therefore proposed that a subset of these states, those constituting the attractors of the system, correspond to cell types. Supporting this suggestion, he argued, was the fact that a Boolean system of a size equivalent to the human genome would be expected to contain a number of attractors whose order of magnitude was equivalent to the number of cell types currently identified in humans (Kauffman, 1996).

The analogy between basins of attraction in dynamic systems and cell type or, more generally, cell fate is widely used as both an intuitive description of a biological process (Kitano, 2004, Wuensche, 1998) and as a specific feature to be observed or quantified in models of genetic systems (Bagley and Glass, 1996, von Dassow et al., 2000, Albert and Othmer, 2003, Espinosa-Soto et al., 2004). One gap in the argument for basins of attraction has been the lack of empirical evidence at a genetic, rather than a phenomenological, level.

There is empirical evidence that the types of basin dynamic observed in model systems may also exist in real gene networks. Experimental results (reviewed by Huang and Ingber, 2000) demonstrate that particular cell fates can be stimulated by a range of non-specific agents, suggesting a robust process of selection from a limited number of end points. A recent experimental result provides explicit experimental evidence of attractor dynamics in real cellular networks (Huang et al., 2005). In this study, two substances, the solvent dimethylsulfoxide and the hormone all-trans-retinoic acid , were both used to trigger the switch into neutrophils in populations of HL60 cells. Initially, these two biochemically distinct signals target different subsets of genes and hence cause the trajectories to diverge into different regions of the state space. Following this, the attractor hypothesis predicts, and experimental results confirmed, that the states of the two populations converge to the same stable state corresponding to neutrophil differentiation.

2.2.4 Non-genetic aspects of developmental control

The genome is the major, but not sole, repository of the information employed during development. In addition, chemical and physical properties of cells and cell aggregates, epigenetic processes such as methylation and chromatin structure, and the environment all play important roles.

As described above (§2.1), changes to the physical organisation of cells are a critical aspect of development. While many of these physical changes are triggered by underlying genetic events, evidence suggests that the reverse may also be true.

The geometry of cells and tissues can feed back information to affect growth and development (Ingber, 2005).

Even within a single cell, non-genetic factors have an impact on development. Epigenetics concerns changes to gene expression that occur despite the absence of any mutation to the genes, but which are still transferred from parent to child cells during division (Jaenisch and Bird, 2003). The two primary mechanisms by which this occurs are DNA methylation—a chemical change to a genome's nucleotides that can block the transcription of a gene—and chromatin modification—alterations to the physical structure of the genome that affect which genes are accessible for transcription (Li, 2002, Wong et al., 2005). One outcome of these mechanisms is that identical genomes, such as are possessed by monozygotic twins, can nonetheless give rise to different phenotypes (Wong et al., 2005). Some epigenetic differences are a result of stochastic factors; others, however, result from environmental differences.

A cell's local environment affects the inductive signals it is exposed to and may therefore affect the trajectory along which it differentiates. The general environment in which an organism develops may also have an impact on its ontogenetic trajectory. Schlichting and Pigliucci (1998) defined the concept of a reaction norm: the range of phenotypes into which a genotype would develop depending on its environmental context.

2.3 An evolutionary perspective on development

Evolution and development are both processes involving changes to morphology over time. In both cases, these processes result from interactions between genomes and their environment. However, the time scales over which change occurs and the mechanisms of change are very different in each case.

A broad view of development encompasses not only the transformation of a fertilised egg cell into an adult organism (as described above), but also the continuing life-cycle of that organism as it grows, learns and finally dies. All of these changes occur to a single individual, within a single generation and, in general, without any modification of that individual's genome (Gilbert, 2003). Evolution, in contrast, encompasses the adaptation of species or populations to their environment. Over many generations of individuals, the genetic composition of populations alters in response to mutation, recombination, drift, migrations and environmental change (Mayr, 2001).

Despite these differences, development and evolution have important effects on one another. Over evolutionary time scales, and as a result of evolutionary mechanisms, the process of development has changed. Evolution has shaped development: in this direction the relationship is relatively uncontroversial (although still far from fully understood). More controversial is the impact that development has on evolution. The following sections describe the conventional view of evolution that emerged in the mid-20th century, the more recent theory of evolution advocated by evolutionary developmental biologists, and the hypothesised role of developmental bias in orienting evolution.

2.3.1 Early views of development and evolution

One of the earliest conceptual meetings between evolution and development occurred in the mid-19th century, with von Baer's recognition that embryos from different groups of animals share certain common features. Darwin cited this observation in support of his argument for the common descent of various species (as described in Gilbert, 2003).

During the first half of the 20th century, the most dramatic debate was between those disciplines that studied evolution at the level of the organism and those that studied evolution at the level of the gene (Mayr, 1991). The first group ('the naturalists') included palaeontology, ecology and systematics and was concerned primarily with how natural selection produced species that were adapted to their environments. The second group ('the geneticists') focused on Mendel's genetic theory of inheritance and included both experimental and population genetics. The core of the debate was how Mendelian inheritance was compatible with natural selection as a mechanism of evolution.

The integration of genetics and natural selection was accomplished, in the early stages, by population geneticists employing the tools of mathematics (Mayr and Provine, 1980). Work by Fisher, Haldane and Wright established that the continuous changes in phenotypic traits observed by the naturalists could be explained in terms of the discrete changes studied by the geneticists. A decade later, a second wave of publications by Dobzhansky, Simpson and Mayr ushered in the 'evolutionary synthesis'. The outcome of the synthesis was an acceptance that: (a) gradual evolution could be explained in terms of the production of genetic variation by mutation and recombination, and the sorting of this variation by natural selection; and (b) macroevolutionary processes such as adaptation and speciation could be explained in terms compatible with microevolutionary mechanisms (Mayr and Provine, 1980).

The explanatory paradigm forged by the evolutionary synthesis was subjected to several refinements. Most notably, the identification of DNA as the material component of inheritance reinforced the centrality of the gene in evolution. The primacy of natural selection was challenged by the neutral theory of evolution (Kimura, 1983), which argued for the role of genetic drift in evolution. Nonetheless, criticism of the evolutionary synthesis continued, focusing in particular on its perceived reductionism. Researchers from disciplines other than genetics took issue with the fact that, in describing all evolution in terms of genetics, the synthesis excluded more complex mechanisms operating at other levels (Eldredge, 1985).

2.3.2 Evolutionary developmental biology

Developmental biology was one of the disciplines whose relevance to evolution had been marginalised throughout much of the 20th century. While many theorists recognised that the process that transformed heritable genes into selectable traits must be important, development was generally omitted from the conceptual framework. One exception was the theory of genetic assimilation (Waddington, 1953). Waddington observed that artificial selection for an environmentally induced phenotypic modification could lead to the modification becoming ‘assimilated’ such that it was expressed even in the absence of the original environmental trigger. He explained these observations in terms of the canalising effect of development: the funnelling of a range of genetic variation into a relatively uniform phenotype. Furthermore, he suggested that such a process would impart a particular dynamic to evolutionary change.

The discovery that was most significant in raising the profile of development was the discovery of *Hox* genes in the 1980s (Wilkins, 2002). Although developmental genetics had started to uncover some of the genes playing a role in development, there was now a general mechanism underlying the construction of diverse morphological forms. *Hox* genes are highly conserved between animal species, despite their differences in form. The discipline of evolutionary developmental biology arose from the renewed interest in the effect of development on evolution (Raff and Kaufman, 1983, Arthur, 1984). The key foci of evolutionary developmental biology are the developmental mechanisms that underly the formation of organisms, and the evolutionary relationship between the ontogenies of different species. Its goals include understanding:

- how development is governed by gene networks;
- how development has evolved—specifically, how the modification of gene networks can produce novel morphological forms; and
- how phenotypic variation is constrained or biased by development (Raff, 2000).

Further conserved developmental genes were discovered throughout the 1980s and 1990s, leading to the proposal of a common genetic 'tool kit' for building body plans (Carroll et al., 2001). The differences between species appeared to be a result, not of different genes, but of different uses of the same genes. The evolution of different morphologies therefore lies less in modification to structural genes and more in the re-wiring of the regulatory interactions between these genes (Carroll et al., 2001, Davidson, 2001).

The strong correlations observed between genetic changes and phenotypic effects led to a widespread view of development being controlled by a genetic program; a view that has been criticized for oversimplifying the process of development and downplaying the role of non-genetic factors (Nijhout, 1990). As reviewed above (§2.2.4), numerous epigenetic factors play a role in development. Competing with the gene-oriented perspective of developmental control is the structuralist view, which focuses on the 'self organising' properties of developmental processes and argues that these have a primary role in evolution (Goodwin, 1994, Newman and Müller, 2000, Müller and Newman, 2003). One of the inspirations for this view is the apparently sparse and discontinuous distribution of extant morphologies, in which a limited range of body plans recur in different contexts. In the structuralist view, the role of natural selection is not to fit every aspect of a phenotype to some adaptive end, but rather to select between the forms made available by morphogenetic processes.

2.3.3 Developmental constraints

An alternative explanation for the non-appearance of many ‘theoretically’ possible morphologies has been phrased in terms of developmental constraints: “biases on the production of variant phenotypes or limitations on phenotypic variability caused by the structure, character, composition, or dynamics of the developmental system” (Maynard Smith et al., 1985, p. 266). Developmental constraints, it is argued, alter the structure of variation on which natural selection acts, and can therefore affect the probability of evolution proceeding in particular directions, irrespective of an adaptive gradient (Arthur, 2000). This is in contrast to the conventional view of evolution, which assumes that all directions of evolution are equally likely prior to the consideration of an adaptive gradient (Figure 2.2). It is important to note that ‘constraint’ in this context was not intended to be interpreted solely in its negative sense, as forbidding certain evolutionary directions. Developmental constraints may also operate in a positive fashion, by rendering certain evolutionary directions easier to achieve. The term ‘developmental bias’ has been suggested as a more inclusive alternative, incorporating both positive and negative effects on the direction of evolution (Arthur, 2004a).

Arthur (1999, 2002b) proposes the variation in the number of segments in centipedes as a case study for the role of developmental bias. A wide range in the number of segments is observed across several thousand different centipede species, from 15 to 191. However, no species has been observed in which the number of segments is even. Arthur argues that such a distribution is unlikely to be the product solely of natural selection, and that the probability of achieving such a distribution by chance is extremely small. Therefore, he suggests, it seems plausible that some property of the developmental process constrains the number of segments to be odd (Arthur, 1999).

The hypothesised mechanism by which developmental bias may affect the direction of evolution may be summarised as follows:

- Evolution occurs in two stages: novel phenotypes appear, and then they either do or do not spread throughout a population;

- The conventional view of evolution holds that novel variation appears in a random, unstructured fashion, and that some combination of natural selection and chance is responsible for which phenotypes predominate;

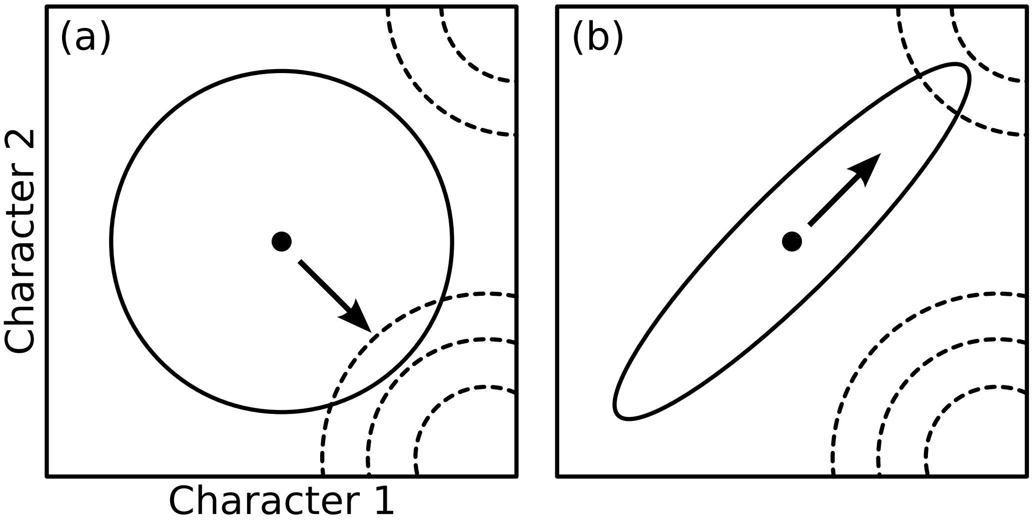


Figure 2.2: The effect of bias on the direction of evolution. The directions (arrows) in which selection takes a population's phenotypic variation when developmental bias is: (a) absent; or (b) present. The solid circle and ellipse indicate the extent of a population's variation; the dashed circles indicate fitness contours (redrawn from Arthur 2004). Note that while Arthur uses the circle and ellipse to represent existing variation in a population, an alternative interpretation is that they represent the *probability* of a mutation introducing variation in a particular direction.

- However, if there are different possible structures to variation, rather than just different quantities, then these structures may affect the direction of evolution, by biasing the variation available for natural selection to act on;

- Therefore there are three mechanisms that may influence the direction of evolution: not only natural selection and chance, but also biases in the structure of variation due to development and mutation.

The relative importance of natural selection and developmental bias as explanatory principles is a topic of debate (Arthur, 2003, Beldade and Brakefield, 2003). Recent contributions suggest that a consensus position seems likely (Arthur, 2004b, Amundson, 2005), with disagreement remaining, not so much about the existence of developmental constraints, but about their ability to affect evolution. A major difficulty with resolving this disagreement is the absence of quantitative data both on phenotypic properties that have resulted from biased evolution, and on the developmental mechanisms that are responsible for bias.

2.4 Directions

Arthur (2004b) has identified two approaches to obtaining evidence for developmental bias as an evolutionary mechanism: indirect and direct. Indirect evidence is that inferred from a fossil record; correlating the ease of producing a particular type of developmental reprogramming with the actual occurrence of that type of change. Direct evidence would require quantifying the structure of *real* adaptive landscapes as well as the structure of variation of a population evolving on those landscapes so that the two could be mapped together. Obtaining the latter type of evidence in laboratory conditions is likely to be a challenging task. An hypothesis of this thesis is that a suitably designed computational model can assist in this task by simulating an evolutionary developmental system in such a way that potential bias can be observed and quantified.

The nature of this task imposes several requirements on the design of such a model:

- The model must be a *developmental* model: it must incorporate both a genotypic and a phenotypic level of description, and a developmental mapping from genotype to phenotype.

- The developmental mapping must be an *implicit* rather than a *explicit* property of the genetic component of the model (Bentley and Kumar, 1999); that is, it must be generated by a dynamic process at the genetic level in a nonlinear fashion, rather than being a direct, linear mapping from gene to developmental or phenotypic feature.

- The model must be an *evolutionary* model: it must enable the necessary conditions for an evolutionary process—variation, differential fitness and heritability (Lewontin, 1970)—to be satisfied.

- The developmental and/or phenotypic components of the model must be *quantifiable*, such that any bias due to development can be measured.

- The model must be computationally *efficient*: evolutionary simulations can require many thousands of iterations to explore complex adaptive spaces. It should be feasible to carry out simulations of sufficient length to enable the observation of interesting behaviours.

- The model must be *plausible* with respect to the biological systems of interest and avoid unrealistic assumptions; the developmental behaviours generated by the model should be grounded in biological data at some level.

The following chapter reviews the role that computational models can play in biology and the existing models that have been proposed for modelling gene regulation, development and evolution. The model used in this thesis, designed to address the requirements above, is then described.

Chapter 3

A Computational Model of Developmental Cell Lineages

The complexity of biological systems is such that their structure and dynamics can be difficult, if not impossible, to understand in an intuitive fashion. Models can assist by abstracting away non-pertinent detail to focus on the critical aspects of a system. This chapter describes the computational modelling methodology used in this thesis. Models have played an important explanatory role throughout the history of science. The forms that models take continue to evolve in response to both the demands of current research questions, and the opportunities offered by the increasing power of the formalisms and tools used to describe and implement models. The argument for computational modelling as a valuable form of inquiry in evolutionary developmental biology is outlined in §3.1.

Computational modelling is a diverse field that is particularly useful for developing insights into the behaviour of complex dynamic systems, including the genetic, developmental and evolutionary systems explored in this thesis. Computational approaches have been applied to the investigation of genetic systems to understand behaviour at the level of individual genes (Ko, 1991, Yuh et al., 1998), small modules controlling a particular function (McAdams and Shapiro, 1995, Barkai and Leibler, 1997, von Dassow et al., 2000, Meir et al., 2002, Oliveri and Davidson, 2004) and entire gene networks (Kauffman, 1971, 1974, Bornholdt, 2001). Computational implementations of developmental systems range from explorations of mathematical models of pattern formation (Gierer and Meinhardt, 1972, Meinhardt, 1982) through to elaborate artificial embryogenies (Stanley and

Miikkulainen, 2003) aimed at both understanding biological development (Fleischer and Barr, 1994, Hogeweg, 2000b, Salazar-Ciudad et al., 2000, Solé et al., 2002, Yoshida et al., 2005) and discovering innovative approaches for the design of other artefacts (Dellaert and Beer, 1996, Bentley, 1999, Kumar and Bentley, 2003, Bongard and Pfeifer, 2003). Similarly, the intersection of evolution and computation has led to multiple research agendas: elucidating the dynamics of natural evolution (Kauffman, 1993, Bullock, 1997, Newman and Engelhardt, 1998) as well as the application of evolutionary principles to the fields of machine learning and optimisation in a wide variety of domains (Holland, 1975, Goldberg, 1989, Harvey and Thompson, 1996, Pollack and Blair, 1998).

Given the diversity of backgrounds and research aims across the computational modelling community, it is unsurprising that there is no unified methodology for modelling evolutionary developmental systems. Existing models of gene regulation, development and evolution that inform the design of the model used in this thesis are reviewed in §3.2. The network-lineage model used in this thesis is then described in detail in §3.3[1]. The model consists of two components: a dynamic network, and a developmental mapping between the dynamics of this network and a cell lineage. Chapter 4 focuses solely on the network component of the model, Chapter 5 focuses on the generation of cell lineages from network dynamics, and Chapter 6 explores the model from an adaptive perspective.

3.1 The case for computational modelling

Modelling a system involves building a formal description of the system on the basis of current knowledge and understanding. Typically, models are constructed to allow a system to be conceptualised and communicated and to assist in determining the course of further research. In its simplest form, modelling is a relation in which a subject takes A as a model of B by identifying properties of both, such that the properties of A are a subset of the properties of B (Edmonds, 1999). A simple example is a road map: a two-dimensional, pen-and-ink diagram containing symbolic representations of buildings, towns or geographical features. A road map may be considered a model of a real landscape by virtue of the fact that some

[1] An early version of this model was described in (Geard and Wiles, 2005). The model described here uses a slightly different input/output mapping, but is otherwise unchanged.

property—the relative positional relationship between features—is possessed by both the map and the landscape. It is important to note that both the map and the landscape may possess additional properties (*e.g.*, an ability to be folded, a breathtaking view) that are *not* shared, but that this does not diminish the usefulness of their shared property. To this end, the modelling relation may be further defined as one in which a subject takes A as a model of B *for a specific purpose*. It is this purpose that will constrain the properties of A and B that must be shared in order for the modelling relation to be a successful one and also inform the choice of which details may be safely omitted. The constraints, or requirements, for the model used in this thesis were outlined in §2.4 of the previous chapter.

Many systems in nature exhibit organisation at multiple levels of description (Solé and Goodwin, 2000). Molecules combine to form amino acids, which are linked together into chains that fold into proteins, which are themselves the building blocks of cells. Cells aggregate together into organs and organisms. Even above the level of the individual, we recognise organisation in the form of family groups, societies and ecosystems. Such systems are not necessarily amenable to reductive analysis. For example, an attempt to explain the emergence of physical form during development in terms of molecular interactions at the level of proteins will struggle to convey an appreciation for a process such as gastrulation, when a massive rearrangement of cells establishes the germ layers and body plan of a developing embryo (Wolpert, 1998). This is not to say that it is not possible to describe morphogenesis in terms of chemical reactions, just that it may not be the most appropriate level of description.

One result of this complexity is that the study of biology has fractured into multiple specialised domains: molecular biology, cell biology, developmental biology, ecology, *etc.* Researchers in each of these fields have focused on understanding phenomena at their level of specialisation. Even as the number of experimental techniques and quantity of available data grow at an increasing rate, there is a recognition that many current issues in biology span multiple research domains (Brenner, 1999, Kitano, 2001, 2002a,b). In this context, the use of models as a tool for communication and collaboration is particularly important.

Models have a long history in the sciences as tools for understanding, communication and prediction (Keller, 2003). The shape that these models take has developed from relatively informal linguistic descriptions of phenomena and diagrams through to formal and mathematical representations of systems and processes. The appearance of new types of data from experimental fields has driven the development of new methods for synthesising, analysing and understanding that data. Given the recent rapid increases in quantity and variety of biological data, computational modelling has become an important supplement to verbal and mathematical models. The following sections review, loosely in chronological order of appearance, some of the broad categories of model that have been employed in biology. Note that some models may be classified into more than one category—models in systems biology, for example, typically employ a variety of modelling techniques.

Model organisms. There is one use of the term 'model' that is quite specific to biology: a model organism is a species that is particularly amenable to study and considered to be representative of some aspects of a wider class of other species. For example, although nematodes such as *C. elegans* contain fewer genes, fewer cells and simpler morphological structures than humans, their gene regulatory systems already display many of the same basic components as are found in the human genome (The *C. elegans* Sequencing Consortium, 1998). Whereas most types of models are artefacts, constructed for a particular purpose, model organisms are natural objects chosen for their property of simplicity, relative to the complexity of the primary object of interest.

Linguistic models. Possibly the simplest type of model, linguistic models are natural language descriptions of a system or phenomenon. Linguistic models have been used ever since people first desired to communicate about the world around them, and are still used by all researchers in the form of a oral or written description of systems under investigation. The limiting factor that determines the usefulness of such models is the size and complexity of the system being described. If a system is particularly large or complicated, natural language descriptions can lack precision or conciseness.

Diagrammatic models. Along with linguistic models, diagrams are an ancient form of representing information on the structure and behaviour of systems. Diagrams are an efficient means of representing interactions between elements of a system, such as the regulatory links between a group of genes, or a temporal sequence of events, such as the transcription/translation process that transforms nucleotide strings into proteins. Again, system size and complexity will limit the usefulness of this type of model.

Formal models. Formal models include those described in mathematical or other symbolic languages that allow a degree of precision not present in linguistic or diagrammatic models. The language of mathematics is largely unambiguous and hence allows models to be communicated more widely with a reduced level of misunderstanding. The process of constructing a formal model of a system can be valuable in itself, by acting as a check upon intuitions and helping to identify inconsistencies in data and understanding. Mathematical and other formal languages are also able to describe models in a concise fashion, which allows larger systems to be modelled than would otherwise be possible. While many linguistic and diagrammatic models are qualitative in nature, mathematical models can be more suited to the expression of quantitative relationships and the generation of quantitative predictions. Such predictions can often be compared with empirical measurements to validate the accuracy of the model.

A further feature of formal models is that they introduce the potential for modelled systems to produce behaviour that is not explicitly contained in the representation of the model. The behaviour of a system can be inferred from a mathematical representation such as a set of equations by solving the equations either analytically or numerically.

Statistical models. One of the features characterising the current era of biological research is the abundance of data. Models are needed that can be used to make sense of the data produced by gene sequencing, microarray analysis and other high-throughput experimental techniques (van Someren et al., 2002). The field of bioinformatics arose around the application of techniques in machine learning and artificial intelligence to the task of finding patterns within this data.

Systems biology models. As mentioned above, researchers are beginning to realise that the open problems in biology frequently require an approach that spans the boundaries of traditional disciplines. The aim of systems biology is to integrate data and knowledge from the broad spectrum of domains between molecular and organismic biology into a unified system-level understanding. The four foci of this approach are the structure, dynamics, control and design of biological systems (Kitano, 2002b). An important feature of systems biology is its emphasis on the necessity of collaboration and sharing of data between research groups. One of the technical requirements that is therefore being addressed is the need for standardised formats for both experimental data and model descriptions to enable effective communication (Crampin et al., 2004, Hucka et al., 2004).

Complex systems models. A complex system can be defined generally as one in which the behaviour of the system emerges from interactions between the components of the system (Holland, 1998). While the rules governing individual components may be very simple, when they combine to interact in large numbers, the behaviour of the system displays a higher level of complexity. One of the earliest biological systems that was recognised as displaying this type of behaviour was the network of interacting genes that regulates physiological behaviour. Kauffman (1969) introduced a random Boolean network model of gene regulation in which each gene was modelled as a simple logical switch, but the system as a whole could display a wide variety of complex behaviours, including periodic and chaotic patterns. Complex systems properties have since been identified in many other natural and artificial systems, including social networks, food webs, weather systems and financial markets.

A significant feature of many complex systems is that their behaviour is generally impossible to predict from a description of their components. The only way to discover what long-term dynamics such a system is going to display is to 'run' it and observe the outcome.

Computational models. Several of the model types described above either rely upon, or can benefit from the use of computers. Formal mathematical models that are too complex to solve analytically can be numerically simulated by a computer in a fraction of the time that would be required by a person. Similarly, bioinformatics models are dependent on computational power to extract patterns from databases too large to be analysed by hand. Systems biology relies upon the data storage and analysis capabilities of computers as well as modern electronic communication systems that enable this data to be rapidly and accurately shared around the world. Computational models are integral to understanding complex systems—

as mentioned above, simulating a complex system is frequently the only way to discover its behaviour.

The application of computers to modelling with which this thesis is primarily concerned is the transformation of a static description of a model into dynamic representation of its behaviour. This type of transformation can be approached in two different ways. If the behaviour of the system is already known, then a computational simulation may be *descriptive*—an animated representation of empirical data collected on a particular phenomena. In this case, the underlying process that generates the behaviour is unimportant.

More commonly, the goal of computational simulation is to understand how a system's behaviour arises, and a *generative* computational simulation may be used. In this case the underlying processes of a system are modelled computationally and the simulation actually generates data on the phenomena of interest. The computational model now plays a more significant role in the process of scientific experimentation. The scientific status of such computational models has been discussed from a variety of perspectives (Miller, 1995, Bullock, 1997, Di Paolo et al., 2000, Peck, 2004). A general conclusion of these enquiries is that caution is required, both in the design of an appropriate model and the sensitive interpretation of results. The associated benefits can be considerable, however: as a type of 'thought experiment', computational simulation can assist researchers to explore their understanding of theoretical terms and clarify their relationship to observed phenomena (Di Paolo et al., 2000). Furthermore, the process of constructing a computational model, which by definition involves an exhaustive and concrete specification of the computational processes contained in a system, can force a novel consideration of previously accepted assumptions.

3.1.1 The suitability of computational models

Given the hypothesis motivating this thesis, that developmental bias may influence the direction of evolution, there are several reasons why computational modelling is a suitable approach:

- Evolutionary developmental biology is an area in which it is difficult to obtain extensive empirical data. Arthur (2004b) proposes several approaches to obtaining data that will allow the effect of developmental bias on evolution to be quantified, but in the absence of more substantial data, statistical approaches will be of limited usefulness;

- The systems involved span multiple levels of description—genetic, cellular, individual and population—and the processes operate across multiple time scales, from molecular to evolutionary. Current systems biology modelling platforms have focused on integrating data at levels of organisation below the individual and are not yet capable of integrating population and evolutionary dynamics;

- The dynamic behaviour of the genetic, developmental and evolutionary systems is inherently nonlinear. Combined with the broad disparity in the temporal and spatial scales of these systems, many analytical approaches become unviable;

- System behaviour at each level of description emerges from interactions at a lower level of description: cell behaviour is a product of gene interactions; development emerges from the dynamics of interacting cells; and the evolution of a population is a product of the differential survival of its individual members. Such emergence is a distinctive characteristic of a complex system, for which computational approaches have previously proven suitable.

3.2 Existing computational models

3.2.1 Gene regulation

There are many different approaches to modelling gene regulation—Smolen et al. (2000a,b), Hasty et al. (2001) and de Jong (2002) provide comprehensive reviews. This section briefly describes Boolean and other logical formalisms, differential equations and neural networks.

Biological processes are highly complicated, and most computational models of gene regulation make two simplifying assumptions. The first is that the control of gene expression resides in the regulation of gene transcription. This assumption is known to be incorrect, as control may also be exercised at a number of other levels, including the post-transcriptional processing and translation of RNA, and the control of RNA and protein degradation (Orphanides and Reinberg, 2002). The second assumption is that genes are expressed and proteins produced at a relatively high, continuous rate, such that any stochastic fluctuations tend to average out. Again, this assumption is known to be an oversimplification—in many systems, the number of molecules is very small, and the stochasticity of molecular events may be important (McAdams and Arkin, 1997).

Boolean networks

One of the earliest approaches to modelling gene regulatory systems was to use networks of logical elements (Kauffman, 1969, 1971, 1993). The Boolean network approach makes three further assumptions to simplify analysis (Somogyi and Sniegoski, 1996). First, the activation of a single gene is represented as a Boolean switch that can be either on or off. In effect, a gene can be either expressed or not expressed and there is no possibility of intermediate levels of activation. This assumption is reasonable when a gene spends most of its time either at a floor value of zero or at some positive saturation level and the time required for a gene to switch is negligible in relation to the time scale of the model. The second assumption is that the regulatory control of a gene is described by a combination of Boolean logic rules (*i.e.*, AND, OR and NOT). The final assumption is that timing is synchronous: all gene states are updated simultaneously at each time step (although Harvey and Bossamaier, 1997, have also explored asynchronous variants).

Kauffman's model of Boolean networks have two primary parameters: network size, N, the number of elements in the network and network connectivity, K, the number of inputs regulating the activity of each element. Each of the N elements is associated with a rule table specifying outputs for each of the 2^K possible input combinations. As each element in the network is updated simultaneously, the system is deterministic and the state at time $t+1$ can be determined on the basis of the state at time t.

The main strengths of the Boolean network model are its analytical tractability and the ease and efficiency with which it can be simulated. One of the immediate advantages of the simplifying assumptions is that the computational requirements of simulating regulatory systems are reduced, allowing the exploration of much larger systems. However, the validity of the assumptions, and the value of the Boolean approach in general, has been questioned by a number of people, particularly in the biological community, where there is a perceived lack of connection between simulation results and empirically testable hypotheses (Endy and Brent, 2001). Some genes are known to have different regulatory effects depending on their level of expression and in some situations the transient period as a gene switches may be significant. While a Boolean representation may be sufficient for a product that tends to be present either in excess, or in insignificant quantities, products whose concentration varies in a more smoothly continuous fashion may require a continuous function to accurately capture their dynamics (Smolen et al., 2000b, Bolouri and Davidson, 2002). A number of researchers have also demonstrated that there is not a direct correlation between the dynamic behaviour of Boolean systems and that of corresponding continuous systems (Glass and Kauffman, 1973, Bagley and Glass, 1996), suggesting a qualitative loss of behavioural information.

Alternative logic networks

Several other logical network formalisms have been proposed as alternatives to Boolean networks. One of the more widely used variations is generalised logic, a formalism for modelling genetic regulatory systems that has been developed over the past three decades (Thomas and Kaufman, 2001). While its origins lie in similar areas to the Boolean models described above, it is distinguished by several features: it is inherently asynchronous, it allows variables to take multiple logical values and it allows for a more sophisticated definition of logical interactions, involving multiple thresholds and parameters. Generalised logic is also motivated by a different set of questions. While Kauffman's networks were developed to investigate the theoretical properties of an entire class of networks, generalised logic tends to focus on models of actual systems. It provides a set of tools with which to characterise and analyse networks derived either from known interactions or from measured patterns of gene expression in terms of their dynamic steady states.

Although the initial version of the generalized logic formalism described the state of a gene in a Boolean fashion (Thomas, 1973), later variants introduced the possibility of state variables assuming more than two levels (Thomas, 1991). The argument for multivariate logic is that when a particular element acts in more than one context, it cannot necessarily be assumed that the thresholds required for each action to occur is going to be equal. For example, product X may have an effect on gene Y when it reaches concentration c_1 and also have a further effect on gene Z at concentration c_2.

The generalised logic formalism has been applied to the analysis of a number of real genetic systems, including phage-λ (Thieffry and Thomas, 1995), dorsoventral patterning in *Drosophila* (Sánchez et al., 1997) and flower morphogenesis in *Arabidopsis thaliana* (Mendoza et al., 1999).

Differential equations

There is a long history of using systems of differential equations to model the reaction kinetics of regulatory systems (Tyson and Othmer, 1978, Chen et al., 1999, Weaver et al., 1999). Continuous differential equations have several advantages over logical approaches. In principle, their more detailed representation of regulatory interactions provides a more accurate representation of the physical system under investigation. Additionally, there is a large body of dynamic systems theory that can be used to analyse such models.

Two major disadvantages of differential equations are the large number of kinetic parameters for which accurate values have not been measured, and the intractable nonlinearity of many systems. When analytic solutions are not possible, two approaches can be followed. In some cases, qualitative properties can be established, such as existence of steady states, limit cycles and critical points (Tyson et al., 2001), even in the absence of a complete characterisation of system dynamics. Alternatively, computers can be used to solve sets of equations numerically. In numerical simulation, the exact solution of an equation is approximated by calculating values for each of the state variables at a series of discretized time steps.

A significant problem for the numerical approach is the lack of measurement of the various kinetic parameters in a system. The number of systems for which detailed parameter values are known is very small, and the size of most systems makes it unfeasible to obtain *in vitro* or *in vivo* measurements of many parameter values. Some researchers have dealt with this problem by using automated search to locate parameter combinations that allow the qualitative behaviour of a system to be reproduced (von Dassow et al., 2000, Meir et al., 2002, Goutsias and Kim, 2004). A second approach is to apply machine learning techniques to the task of inferring parameter values from gene expression data (van Someren et al., 2002).

Neural networks

Artificial neural networks are mathematical models of information processing originally inspired by networks of neurons in the brain (Hertz et al., 1991). A neural network typically consists of a collection of nodes, some of which may be designated as input or output nodes, connected by weighted links. Each node contains a transfer function that transforms a set of weighted input signals into an output signal. These networks can be trained to match particular patterns of activation via a variety of learning processes. While early neural networks had a feed-forward structure in which the emphasis was on learning a mapping between a set of input features and a particular output, later *recurrent* networks used layers of fully connected nodes (Elman, 1990). The addition of recurrent connections enables a network to maintain an internal state such that a system's behaviour is a product of both the inputs received in the current time step as well as a history of past activation.

A relatively straightforward analogy may be drawn between an information processing system in which the constituent elements are neurons and the links are synaptic interactions, and a system in which the elements are genes and the links are regulatory interactions. Consequently, neural network like approaches have been used to model several biological systems. Mjolsness et al. (1991) developed a phenomenological model of segmentation in the *Drosophila* blastoderm that used a neural network model to describe the internal dynamics of a cell as well as a generative grammar that described higher-level developmental processes such as cell division and differentiation. This model has also been applied to other aspects of pattern formation and neurogenesis in *Drosophila* (Reinitz et al., 1995, Reinitz and Sharp, 1995, Marnellos and Mjolsness, 1998). In these models, network parameters were trained such that the dynamics matched observed experimental behaviour.

Vohradský (2001a,b) used a similar approach to model the lysis/lysogeney decision in phage λ. Here, the network structure was determined *a priori* from known interactions and the interaction weights were learned from experimental data. Several variations on the basic network were investigated, including connected networks and multi-compartment models, in which protein and RNA products are represented by separate network layers (Vohradský, 2001b).

Mathematically, a neural network can be described as a system of differential equations. Conceptually, there is little to distinguish the parameter search used by von Dassow et al. (2000) from the parameter training used by Marnellos and Mjolsness (1998). One distinction is that differential equation models tend to use activation functions based on a model of a specific molecular process, whereas neural networks tend to use a generalised activation function, usually some form of nonlinear monotonic function, such as a logistic sigmoid or tanh.

3.2.2 Development

As with gene regulation, there are many different ways of modelling the process of development (Stanley and Miikkulainen, 2003, provides an extensive review). Unlike gene regulation however, there are relatively few standardised approaches. This section therefore considers how various models have addressed two specific issues: the control of development and representation of developing entities.

Developmental control

Broadly speaking, two approaches to modelling the control of a developmental process have been considered, termed *grammatical* and *cell chemistry* approaches by Stanley and Miikkulainen (2003). Grammatical approaches describe development using a set of production rules, which are applied iteratively to transform an initial state into a final phenotype. Each production rule consists of a non-terminal symbol on the left, which is replaced by some combination of terminal and non-terminal symbols on the right. The use of grammars to model biological systems was first introduced by Lindenmayer (1968), who developed L-systems as a means of describing the complex fractal patterns observed in nature. L-systems remain in wide use today, particularly for describing the architecture of plants (Prusinkiewicz, 2004). Various extensions have been proposed to extend the descriptive capabilities of L-systems, including parameterised rules, environmental interactions and stochasticity.

Grammatical approaches have also been used for the evolutionary design of neural networks (Kitano, 1990, Gruau, 1995), and robot morphologies and controllers (Hornby and Pollack, 2002). The motivation for using grammatical (or 'generative') encodings in these contexts was to increase evolvability through the use of a representation that was scalable and inherently modular. The issue of representation is reviewed further below (§3.2.3).

The cell chemistry approach to developmental control is a bottom-up approach to simulating a growth process. Rather than explicitly specifying phenotypic change using rewriting rules, developmental events are derived from the dynamics of an underlying cell chemistry system. One of the earliest such models was the reaction-diffusion system described by Turing (1952), which was capable of producing a variety of natural-looking spatial patterns. The interacting components of Turing's system were abstract chemicals known as morphogens; more recent cell chemistry approaches have used metabolic or gene regulatory networks as developmental controllers. Many of the network modelling formalisms reviewed above (§3.2.1) were designed to address issues in biological development.

In a cell chemistry model of development, the dynamics of the underlying network are used to control cellular events such as division and differentiation. In many cases, specific components of the network (*i.e.*, genes or metabolites) are assigned a particular function, such as cell division or cell death. When this component becomes active, its assigned event takes place (Fleischer and Barr, 1994, Fleischer, 1996, Eggenberger, 1997). In other models, cell division and cell death depend on a cell's volume: a cell whose volume exceeds some upper bound will divide and form two cells, while falling below a lower bound initiates apoptosis (Hogeweg, 2000a,b). Cell differentiation is typically defined in terms of patterns of gene activity, with each stable pattern indicating a distinct type (Kaneko and Yomo, 1994, Furusawa and Kaneko, 1998, Hogeweg, 2000b, Solé et al., 2003, Keränen, 2004). Cell behaviour is usually modelled as a product of both internal dynamics and external signals. The external signals may originate from other cells, either via direct contact or diffusion; or a pre-specified field, such as a morphogen gradient.

Phenotypic representation

The second issue arising in the design of a developmental model is how a developing entity is represented. Existing models may be classified according to their treatment of the spatial field in which development occurs:

- In pattern formation models, space is modelled as a continuous field over which chemicals diffuse and react (Turing, 1952, Meinhardt, 1982). In such models, there are no discrete cells, and often no notion of organism growth (*i.e.*, the size of the field is fixed).

- Most developmental models take individual cells as the fundamental building blocks. In some models, all cells are of an identical shape and size, and the arrangement of cells is fixed to a regular square (Eggenberger, 1997, Keränen, 2004) or hexagonal (*e.g.*, Marnellos and Mjolsness, 1998) grid. In other models, cells may be located freely within space, and even adopt irregular shapes and sizes (Hogeweg, 2000b, Kumar and Bentley, 2003). The dimensionality of the space in which development occurs ranges from one (Salazar-Ciudad et al., 2000) to three (Eggenberger, 1997, 2003, Kumar and Bentley, 2003). In general, most models use two dimensions, which allows for a reasonably diverse range of morphological forms, without the computational expense involved in scaling up to three dimensions (Fleischer, 1996). Discrete cell models may exist in a continuous substrate, through which signals can diffuse between cells or across which morphogen gradients may be defined (Rudge and Geard, 2005).

- A further class of models utilises building blocks at a level above that of an individual cell. Many of the generative grammar models of developmental control described above fall into this category: each developmental unit represents a morphological module. A single module may be a branch segment or leaf in a plant model, or a body or limb segment in an animal model. Cell chemistry models may also adopt this approach. Bongard and Pfeifer (2003) uses macrocellular units as components of a robot morphology: each unit is a complex morphological component involving sensors, actuators and neural control elements.

- One final possibility is that models may not include an explicit spatial component. The grammatical approaches to modelling neural networks (Kitano, 1990, Gruau, 1995) are one example: the important feature of the network phenotypes is the relationship between individual neurons, rather than their location in space. Another form of phenotypic representation that is an organisational rather than a spatial description is a cell lineage (described in a biological context in §2.1.1). Yoshida et al. (2005) use a cell lineage representation of a developing phenotype in order to detect the occurrence of recursive patterns of cell differentiation. A cell lineage representation of development may be depicted and represented as a one dimensional array of cells whose length increases over time. Alternatively, individual division events may be labelled with their orientation (*e.g.*, left-right, dorsal-ventral or anterior-posterior), making it possible to interpret a cell lineage in multiple dimensions. The cell lineage model used in this thesis is described further in §3.3.2.

3.2.3 Evolution

Computational approaches to modelling evolution may be categorised into two classes on the basis of their motivation. First, there are evolutionary models whose purpose is to explore and explain some aspect of biological evolution. Second, there is a large class of 'biologically inspired' approaches to adaptive search and optimisation. In reality, these two classes overlap at a technical level, and, so long as the fact is kept in mind that evolution frequently does *not* act as an optimisation process, both purposes have much to gain from each other.

One form that computational models of evolution may take is simply an implementation of a mathematical quantitative genetic model, where the role of the computer is simply to perform large numbers of iterated calculations. In the last few decades however, evolutionary computing is more likely to consist of a bottom-up computational implementation of an adaptive process, rather than a top-down analytical approach. Evolutionary algorithms are a class of adaptive search algorithms based on natural evolution. Numerous varieties have been proposed, including genetic algorithms (Holland, 1975, Goldberg, 1989), evolutionary programming (Koza, 1992) and evolutionary strategies (described in Bäck et al., 1997).

In essence, an evolutionary algorithm consists of a population of individuals (which may be as small as one member) representing either a set of candidate solutions to a particular problem or agents located in a particular environment. Each solution is assigned a fitness value, representing its proximity to the target solution or level of adaptedness to its environment. The fittest individuals are selected to reproduce, either asexually or via recombination, and the newly created offspring, possibly modified by some form of mutation, constitute the next generation of solutions. Theoretically, and in practice, the average fitness of the population will increase over successive iterations of this selection/mutation cycle.

Genotypic representation

During evolution, new individuals are created via the modification of existing individuals in a population. The range of individuals that are mutationally accessible from a given individual will depend on what types of genotypic change are possible. In turn, the range of possible changes will depend on how a genotype is represented. The original genetic algorithm proposed by Holland (1975) used a bit-string representation, in which each locus represented a single binary allele. Mutations to such a representation involved randomly 'flipping' the bits at some loci to create a new individual. Possible alternatives to bit-string representations include: continuous value representations, consisting of a sequence of real valued numbers that were mutated by adding small amounts of random noise (Goldberg, 1989); tree representations, in which a solution (usually an algorithm) is encoded as a binary tree of operators and values. Considerable effort has been invested in determining which representation and associated mutation operators are most effective for the solution of particular search problems.

One way of interpreting the evolutionary implications of a particular genotypic representation is in terms of its effect on the adaptive landscape (as described in §2.3). The choice of representation (and the nature of the problem) will affect how the correlation between the fitness of a given individual and that of its adaptive neighbours. If neighbouring fitness values are closely correlated, the resulting landscape will be smooth, and potentially easy to search. As correlation decreases, the landscape becomes increasingly rugged, and the number of local optima in which search can become trapped will increase. One landscape characteristic of considerable interest in the last decade is *neutrality*: the presence of plateaus or ridges of mutationally adjacent individuals of equal (or very nearly equal) fitness. Many natural and artificial systems display the hallmarks of neutrality (Kimura, 1983, Shipman et al., 2000). The dynamics of populations evolving on neutral landscapes are of particular interest because they provide potential explanations for periods of evolutionary stasis (Bornholdt and Sneppen, 1998), escape from local optima (van Nimwegen and Crutchfield, 2000), robustness (Wilke et al., 2001) and open-ended evolution (Wilke, 2001).

Another implication of different genotypic representations is that the mutation operators associated with a particular representation may bias the distribution of variation they produce (Bullock, 1999, 2001). As in studies of natural evolution (§2.3.2), the potential effects of variational structure on artificial evolution are relatively unexplored. The issue of mutation bias is addressed further in Chapter 6.

Evolving development

In the discrete and real valued representations described above, an individual solution is generally encoded directly into the genotype. This situation clearly differs from biology where the object of selection, the phenotype, is derived from the genotype via a complex dynamic process. The developmental models described in §3.2.2 above embody a similar level of indirection. The application of developmental mappings to real world design problems is currently a topic of much interest. By exploiting properties of developmental mappings such as modularity, redundancy and canalisation, it is anticipated that the scalability and robustness of evolutionary systems can be increased (Roggen and Federici, 2004). The choice of genotype representation remains important when evolving a developmental system. While most cell chemistry models share a common interaction network structure, a wide variety of different schemes for encoding this network have been proposed. The simplest approaches involve using weight matrices to specify the strength of interaction between nodes (Wagner, 1996, Siegal and Bergman, 2002). Other models have used elaborate 'artificial chemistries' to determine affinities between genes and regulatory factors. In these models, the strength of binding may be derived from sequence matching (Eggenberger, 1997, Reil, 1999), fractal patterns (Bentley, 2003) or production rules (Suen and Jacob, 2003).

Azevedo et al. (2005) took an unusual approach to modelling the evolution of development. They were interested in studying how the complexity of an ontogeny (represented as a cell lineage) could be reduced during evolution with stabilising selection on the phenotype (represented by the terminal cell fates of the lineage). This model did not use an explicit representation of the genotype that produced an ontogeny, and mutation operators were therefore defined directly in terms of modifications to the cell lineage. The methodology and results of Azevedo et al. (2005) are of particular relevance to the studies of this thesis and are described further in Chapters 5 and 6.

3.3 The network-lineage model

This section describes the network-lineage model that will be used throughout this thesis. The network-lineage model consists of two components: a network component that generates the gene expression dynamics controlling development and a cell lineage component that defines how these dynamics are interpreted to define an ontogeny. The network component is based on a standard recurrent neural network architecture (Elman, 1990, Hertz et al., 1991). The cell lineage component is a novel contribution. As reviewed above, two previous studies have explicitly modelled aspects of development using a cell lineage (Azevedo et al., 2005, Yoshida et al., 2005); however, the model presented here differs from each of these by virtue of the mapping from dynamics to development. The evolutionary modelling techniques used in this thesis are employed for the purpose of investigating aspects of the network-lineage model, rather than as focus in themselves. Description of these techniques is therefore deferred to the relevant sections of Chapter 6.

3.3.1 The DRGN component

In the DRGN model, a genetic system is defined as a network of interacting nodes (Figure 3.1). The network is structured in three layers, consisting of N_I input nodes, N_R regulatory nodes and N_O output nodes respectively. The input nodes are used to provide information to the DRGN on its current regulatory context (see §3.3.2 below) and their activation is determined by extracellular events rather than the DRGN dynamics. The regulatory nodes represent genes that play a regulatory role only. That is, they have no direct effect on functional behaviour, but mediate between the input nodes and output nodes. The output nodes represent a subset of genes that specifies the functional behaviour of the network (see §3.3.2 below). These nodes have no regulatory outputs, that is, their level of expression has no direct influence on the future dynamics of the network. The activation state of each node is a continuous variable in the range $[0, 1]$, where 0.0 represents a completely inactive gene and 1.0 a fully expressed gene.

Information flows through the DRGN from the input nodes, through the regulatory nodes to the output nodes. The state of the output nodes at a given time is a product, not only of the DRGN's current inputs, but also its dynamic history, as stored by the recurrent interactions between the regulatory nodes. The inter-

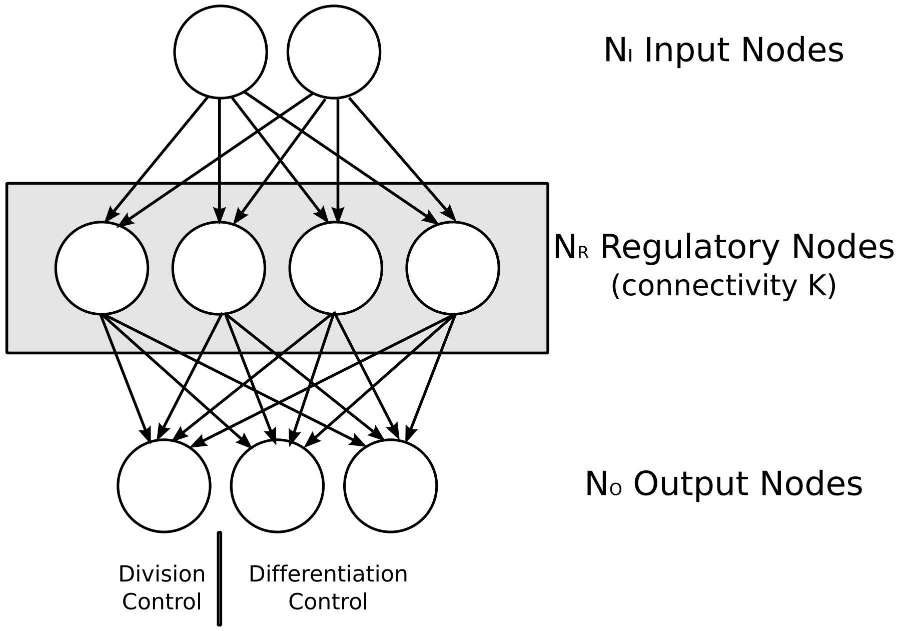


Figure 3.1: The structure of a DRGN network, showing input, regulatory and output nodes. In this model, the input layer and the regulatory layer are fully connected, the regulatory layer is randomly connected such that each node has inputs from K regulatory nodes (including possible self connections), and the regulatory and output layers are fully connected. The first output node controls the division of a cell; all other output nodes represent possible cell fates.

actions between nodes in the three layers can be summarised as follows. All input nodes are connected to all regulatory nodes, all regulatory nodes are connected to all output nodes, and all regulatory nodes are optionally connected to all regulatory nodes (including self connections). The level of regulatory connectivity of the network (K) determines the number of inputs each regulatory node receives from other regulatory nodes.

The interactions between two DRGN layers can be represented by a weight matrix, in which the entry at row i, column j specifies the influence that gene j has on gene i. These entries may be positive or negative, depending on whether the product produced by gene j is an activator or a repressor in the regulatory context of gene i. A zero entry indicates that there is no interaction between the two genes. The inclusion of self-connections (i.e. from node i to node i) allows for the possibility of genes influencing their own regulation. When a random network is created, each of the non-zero entries in its weight matrix are typically initialised to a value drawn from the Gaussian distribution $G(0, W)$, where W defines the interaction strength (or weight scale) of the network. The effect of W on network dynamics is explored in Chapter 4. Collectively, the three parameters N, K and W are referred to as the genotypic parameters, as they define a class of network genotypes.

The state of the network was updated synchronously in discrete time steps, with the activation of regulatory node i at time $t+1$, $a_i(t+1)$, given by

$$a_i(t+1) = \sigma(\sum_{j=1}^{N_I} w_{ij}a_j(t) + \sum_{j=1}^{N_R} w_{ij}a_j(t) - \theta_i) \tag{3.1}$$

where N_I ad N_R are the number of nodes in the input and regulatory layers, w_{ij} is the level of the interaction from node j to node i, θ_i is the activation threshold of node i, and $\sigma(.)$ is the sigmoid function, given by

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{3.2}$$

The activation of output node i at time $t+1$, $a_i(t+1)$, was given by

$$a_i(t+1) = \sigma(\sum_{j=1}^{N_R} w_{ij}a_j(t) - \theta_i) \tag{3.3}$$

where the definitions of all symbols follows that of Equation 3.1.

The sharp distinction between output and regulatory roles for genes reflects the traditional definition of a gene as the region of DNA encoding a single protein (Alberts et al., 1994). Proteins are typically classed as having either a functional or regulatory role and genes have traditionally been classified according to the type of protein they encode. It is now recognised that the regulation of gene expression is a significantly more complicated process than initially thought (Orphanides and Reinberg, 2002). Genes may code for multiple products via alternative splicing, and regulation may occur at stages other than transcription, such as RNA editing and translation control. Furthermore, genes may not be the only—or even primary—source of regulatory signals. Considerable evidence is beginning to amass that suggests RNA-based signals encoded in intronic and intergenic regions of the genome may play an important role (Mattick, 2001, 2004, Mattick and Gagen, 2001). In this respect therefore, the DRGN model represents a first approximation to the complexity of the regulatory process, with the potential for future refinement.

The dynamic properties of the DRGN are explored in Chapter 4. The following section describes how the DRGN component of the model was used to generate an ontogeny.

3.3.2 The cell lineage component

A cell lineage is a record of an entire ontogenetic trajectory. Considering the lineage as a tree: the root node represents the fertilised egg cell; the non-terminal nodes represent the transient states that cells pass through whilst differentiating; the terminal nodes represent the final differentiated cells that exist at the end of the developmental process. Therefore, it is the terminal nodes of the cell lineage that constitute an organism's phenotype, while its ontogeny encompasses the complete set of nodes (both terminal and non-terminal) that existed at some point during the developmental process. The DRGN model, as described above, is a general purpose computing device. In a developmental system, the computation performed is the transformation of a temporal sequence of contextual inputs into an ordered pattern of cell division and differentiation events.

For all of the simulations reported in Chapters 5 and 6, two input nodes were used to specify the relative position of a cell with respect to its sibling. After division, the activation of these nodes was set to $(0, 1)$ in the left daughter and $(1, 0)$ in the right daughter. This minimal external input reflects the combined effects of the different contextual signals received by the two cells resulting from their respective positions in the embryo. A clear example of this type of signal is the *pop-1* gene in *C. elegans*, which is differentially expressed in the two daughters produced following an anterior–posterior cell division (Lin et al., 1998). At the level of abstraction of the DRGN model, these inputs were not assigned a specific biological role. Rather than explicitly requiring that cell fate be specified by any particular mechanism (such as asymmetric division or inductive signals), the DRGN inputs simply indicate that there is some difference in regulatory context between two daughter cells.

The nodes in the output layer of the DRGN were used for the control of a cell's division and differentiation decisions. If the activation of the first output node was above a certain division threshold θ_d, that cell would divide, otherwise it would

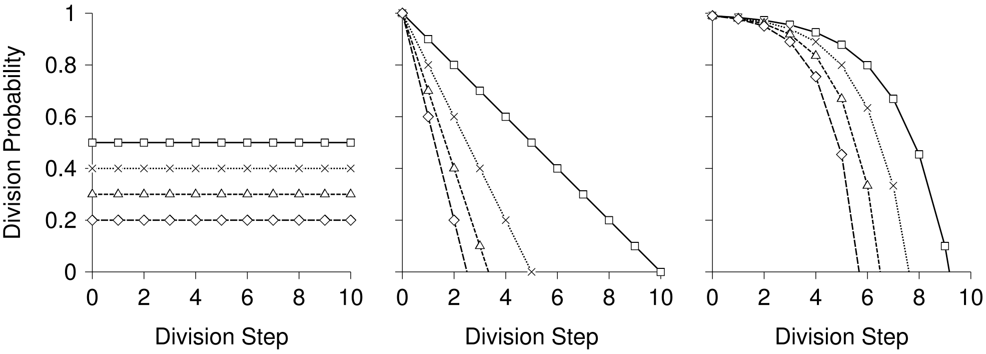


Figure 3.2: The effect of different division threshold scaling methods: constant (left), linear (centre) and exponential (right). $\lambda = 0.5$ (squares), 0.6 (crosses), 0.7 (triangles) and 0.8 (diamonds).

differentiate. As development proceeded, the likelihood of a cell continuing to divide decreases. To simulate this, the division threshold was scaled dynamically. Three different parameterised scaling regimes were implemented:

$$\begin{aligned} \text{constant}: \quad & \theta_d = 1 - \lambda \\ \text{linear}: \quad & \theta_d = 1 - \lambda d \\ \text{exponential}: \quad & \theta_d = 1 - 0.01e^{\lambda d} \end{aligned} \tag{3.4}$$

where d was the depth of the current cell and λ was a parameter controlling the rate at which the division threshold was scaled (Figure 3.2).

Depending on the scaling regime used and the value of λ, it was possible for a DRGN to continue dividing indefinitely. To ensure that simulations completed in a reasonable time, an upper limit was imposed on the number of levels of division that could occur (*i.e.*, the maximum depth of the cell lineage tree). Any cells that had not differentiated by this time were labelled as undifferentiated. Unless otherwise specified, the exponential scaling regime was used for the studies reported in this thesis.

Once a cell had stopped dividing, the remaining $N_O - 1$ output nodes were used to determine its differentiation type. A simple 'one-hot', or exclusive, encoding scheme was used, in which each output node corresponded to a single cell type. A cell was assigned the type corresponding to the output node with the highest activation values.

A DRGN was used to generate a cell lineage as follows:

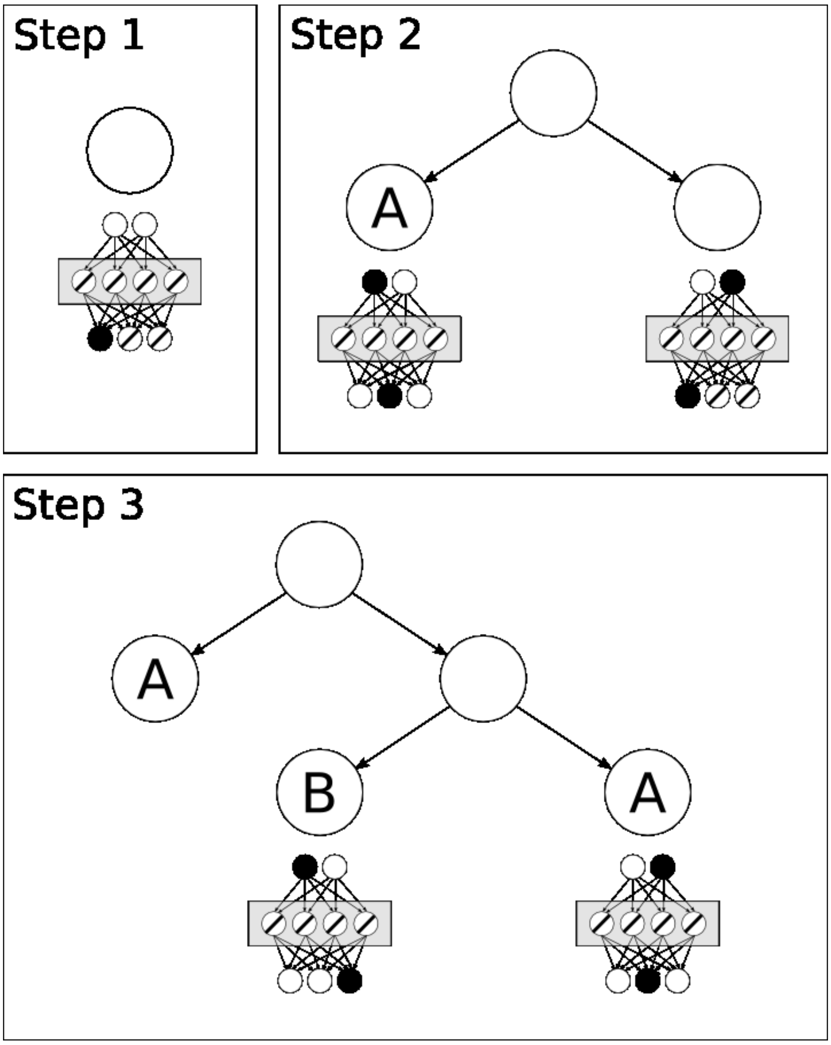


Figure 3.3: Generation of a cell lineage from network dynamics. Although node activations are actually continuous, they are represented here as binary switches for simplicity—nodes are coloured black (on), white (off) or barred (unimportant). **Step 1:** The developing system is initialised with a single cell and both input nodes are switched off. The network is updated once. The *Division Control* output node (left) switches on indicating that cell division occurs. **Step 2:** The initial cell has divided and the regulatory network has been copied into each of the two daughter cells. The left input node has been switched on in the left daughter and the right input node has been switched on in the right daughter. The network is updated a second time. The *Division Control* output indicates that the right daughter will divide, but that the left daughter will not. The activation of the first *Differentiation Control* output node (centre) is greater than that of the second (right), therefore the left daughter cell adopts fate A. **Step 3:** The undifferentiated cell divides again. This time, both daughters differentiate and adopt fates B and A respectively. The 'phenotype' associated with this lineage consists of two cells of type A and one cell of type B.

1. A single DRGN, representing a fertilised egg cell, was initialised by setting the activation of all of its nodes to 0.0 (Figure 3.3, Step 1).
2. The activation of each of the nodes in the regulatory and output layers was updated.
3. If the activation of the division output node was less than θ_d, division occurred (Figure 3.3, Step 2):
   (a) Two copies of the DRGN were created with identical weights and node activations.
   (b) The activation of the two input nodes was set to $(0, 1)$ in the left daughter and $(1, 0)$ in the right daughter.
4. Otherwise, if the activation of the division output node was greater than θ_d, differentiation occurred and the current cell was assigned the type corresponding to its most active differentiation node (Figure 3.3, Step 3).
5. Cells that had differentiated underwent no further change. Steps 2 to 4 were repeated for each of the remaining cells.
6. Development ceased when all cells had been differentiated, or some predefined limit on division depth had been reached. Any remaining undifferentiated cells at this stage were labelled as such.

The properties of both cell lineages and phenotypes generated by network dynamics are explored in Chapter 5.

3.3.3 Addressing the requirements

The network-lineage model described above addresses the requirements identified in §2.4 in the following ways:

- The *developmental* requirement is satisfied by the described model by virtue of the dynamic mapping from genotype to phenotype. The parameters of a gene network (*i.e.*, number of nodes, pattern of interactions and weight matrices) constitute the genotypic level of description. The number and type of terminal cells in a cell lineage constitute the phenotypic level of description. The generating procedure described above constitutes the developmental mapping.

- The requirement that development be *implicitly* specified is satisfied by the network architecture of the developmental controller and the dynamic nature of the mapping. The nodes within the regulatory layer of the gene network do not correspond to specific phenotypic characteristics. Similar to biological gene networks, functions may be controlled by the action of multiple genes (polygeny), while any given gene may contributed to the control of multiple functions (pleiotropy).

- The described model enables the conditions for an *evolutionary* process to be satisfied: both genotypic and phenotypic variability are possible; several approaches to assigning fitness values to phenotypes are described in §6.2.1; and the pattern of network interactions forms, as represented by a weight matrix, is a heritable unit. Evolutionary aspects of the network-lineage model are addressed further in Chapter 6.

- A cell lineage representation of ontogeny maintains a complete history of a developmental process, rather than just the final phenotype. This representation constitutes an organisational, rather than a spatial, description of development. As such it it is straightforward to *quantify*. Several ways in which a cell lineage can be quantified are described and compared in Chapter 5.

- By focusing on the organisational aspects of development, the cell lineage model could be implemented without computationally costly physical and mechanical processes. This choice of representation resulted in an *efficient* computational implementation.

- Cell lineage data is available for variety of organisms (§2.1.1). Defining the evolutionary tasks used in Chapter 6 on real data ensures that task complexity falls within plausible bounds (not too simple, but not unrealistically complex) and establishes a valid link between the simulation experiments and real biology.

Chapter 4

The Dynamics of Cell Differentiation

The dynamics of gene networks play a key role in the control of development. During development, changes to a cell's pattern of gene expression lead it to differentiate into one of several possible fates. As reviewed in Chapter 2, cell types share several characteristics with the attractors of a dynamic system: they are discrete and generally stable, but can be transformed under certain conditions. The dynamic properties of a system, including its attractors, depend on structure. Therefore the repertoire of fates into which a cell can differentiate will be influenced by the structure of its genetic control network. In order to explore how a DRGN's dynamics influence a developmental process, it is first necessary to understand the range and characteristics of these dynamics and their relationship to system structure.

Many dynamic systems exhibit a wide variety of complex behaviours. This complexity can make it difficult to determine which features of the system are responsible for behaviour. The approaches that have been identified for investigating real gene regulatory networks may be broadly classified into analytic, inductive and ensemble techniques (Kauffman, 2004). Analytic techniques involve the derivation of detailed models from knowledge of the underlying mechanisms. Inductive techniques take the reverse approach and use automated processes to derive models from observed data. The third approach proposes that individual systems may be considered as exemplars of broader *classes* of systems, defined by their size, topology and other features. Therefore, by studying the statistical properties of random members of these classes (an ensemble) it becomes possible to draw inferences about the typical behaviour of a particular type of system.

One limitation of the analytic approaches to characterising network dynamics is the constraints they impose on the size and topology of the systems under investigation. These constraints can limit the general applicability of results to the behaviour of networks with arbitrary structure (Pasemann, 1995, 2002). A second limitation of existing analyses is the restriction to fixed point attractors (rather than cyclic attractors) for reasons of tractability (Beer, 1995, Tiňo et al., 2001, Mochizuki, 2005). Furthermore, as the methodology of computational simulation involves drawing inferences from observations of real (simulated) systems, we need to have a good understanding of how the DRGN model will behave *in practice* as well as in theory.

The aims of the set of studies reported in this chapter were to explore the range of dynamic behaviours that a DRGN can generate and to characterise how the diversity and stability of these behaviours change with structural properties. The focus of these studies was the use of ensembles to identify how dynamic behaviour depends on structural features (§4.3). Prior to this, a preliminary study developed the methodology used to analyse and visualise the attractors in a dynamic landscape (§4.2). A final study compared the results obtained from random ensembles to behaviour of hand-crafted systems, in order to understand how patterns of connectivity can affect the number of attractors found in a system.

Insights from these studies, focusing on a single cell, inform the investigation of multicellular development (to be considered in Chapter 5). Understanding the dynamics of gene expression at a local level is an important first step in exploring how these dynamics guide the process of development.

4.1 Basins of attraction

Attractors in a nonlinear dynamic system may be defined as a closed set A with the following properties (Strogatz, 1994):

1. A is invariant: any trajectory that starts in A remains in A for all time.

2. A attracts an open set of initial conditions: there is an open set U containing A such that any trajectory that starts in U will approach A as $t \to \infty$. The largest U is described as the basin of attraction of A.

3. A is minimal: there is no proper subset of A that satisfies conditions 1 and 2.

The two types of attractors most commonly of interest in biological systems are stable fixed points, and stable limit cycles. A fixed point of a system, x^*, may be considered stable if, after some perturbation to the system, the system returns to that fixed point (*i.e.*, if all trajectories near to x^* approach x^* as $t \to \infty$). Stability may be similarly defined for limit cycles: if all nearby trajectories approach a limit cycle, it is stable or attracting.

A distinguishing feature of both stable fixed points and stable limit cycles is that the behaviour of trajectories within their respective basins of attraction are predictable. Given any initial condition in the basin of attraction of a fixed point or limit cycle, we know that it will approach a stable attractor as $t \to \infty$. There is a further type of attractor, a chaotic attractor, that may exist in nonlinear dynamic systems for which this property does not hold. One of the defining conditions of a chaotic attractor is a sensitive dependence on initial conditions. That is, the trajectories of two initial conditions, x_0 and x'_0, whose initial separation δ_0 is very small, will diverge at an exponential rate. Therefore, at some point t in the future, it will no longer be possible to predict the behaviour of the trajectory x'_t based on x_t.

4.2 Study 1: Tools for exploring network dynamics

A characteristic feature of high-dimensional nonlinear systems is that their dynamics can be difficult to analyse and visualise. The first aim of *Study 1: Tools for exploring network dynamics* was to review and compare several techniques for visualising the dynamics of high-dimensional systems. An additional aim was to determine an effective empirical means of counting the attractors displayed by a DRGN and classifying them in terms of type and stability. The DRGN model was used as defined in §3.3.1 with one modification: N_I and N_O, the number of input and output nodes, were set to zero. Thus, in the absence of any functional context for the network, this study focused on the intrinsic dynamic properties of the regulatory component.

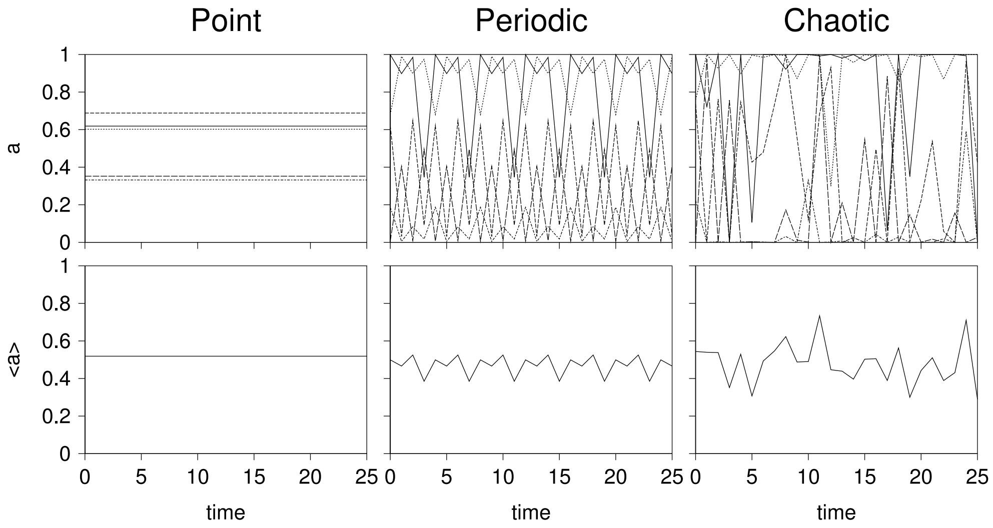


Figure 4.1: Examples of trajectories in fixed point, cyclic and chaotic attractors. Each of the upper figures shows the individual activation trajectories of each of five network nodes; the lower figures show the corresponding average activation trajectory.

4.2.1 Bifurcation diagrams

The trajectory of an N-dimensional system can be thought of as a path through N-dimensional space. For any $N > 3$, this trajectory will be difficult to visualise. A straightforward option is to plot the activation over time of each element of the system independently (Figure 4.1, upper plots). Each of these plots represents one possible trajectory of a dynamic system and already the amount of information generated is difficult to visualise in a meaningful fashion. This information can be condensed by plotting the average activation of all the nodes at each time point (Figure 4.1, lower plots):

$$\langle \mathbf{a} \rangle = \frac{1}{N}(a_0 + \ldots + a_{N-1}). \tag{4.1}$$

A common technique for exploring the behaviour of a dynamic system is to investigate the behaviour of a family of parameterised functions. For example a family of linear systems may be described by the function $f_m(x) = mx$ where m is varied. It is then possible to observe how the dynamics of the system change as the function is changed. The same technique may be applied to investigate the behaviour of dynamic networks by the inclusion of a parameter W that scales the

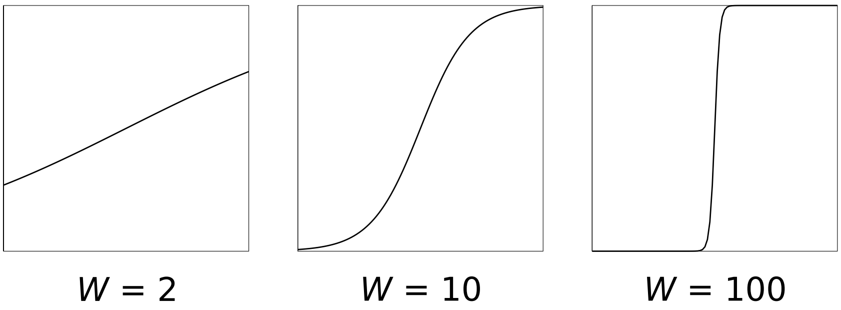


Figure 4.2: The effect of W on the slope of the sigmoid function. As W is increased, the sigmoid function passes from the linear range, through the nonlinear range and approximates a Boolean step function when W is very large.

net input i_{net} into the node activation function σ

$$f_W(x) = \sigma(W * i_{net}). \tag{4.2}$$

The scaling parameter W affects the slope of the sigmoid function. When W is very small, $f_W(x)$ is linear. As W increases, $f_W(x)$ passes through the nonlinear range, eventually saturating and approximating a Boolean function when W is very large (Figure 4.2).

To obtain an insight into the dynamic behaviour of a network with a particular pattern of interactions, the trajectories originating from a single initial condition were recorded as interactions were scaled from very weak to very strong. For the examples in this section, a fully connected DRGN with 20 regulatory nodes was created with weights and biases drawn from a Normal distribution with mean 0 and variance 1.

1. The state of the DRGN was initialised to $\mathbf{I} = (I_0, \ldots, I_{N-1})$ where I_n was a uniform random value in the range $[0, 1]$.

2. The scaling factor W was initialised to a small value (0.01).

3. The DRGN was iterated 1,000 steps to ensure that the trajectory was located on an attractor.

4. The system was iterated a further 500 steps and the average activation (Equation 4.1) at each step was recorded.

5. The network was then reset to the initial state **I**, and W was incremented by 0.01.

6. Steps 3 to 5 were repeated until $W = 20.0$.

A qualitative picture of the dynamics of the parameterised network was obtained by plotting the average activation of each of the states visited in the attractor orbit (Figure 4.3).

Several general statements can be made about DRGN dynamics on the basis of an orbit diagram. It is possible to clearly distinguish three different types of long-term dynamic behaviour (*i.e.*, after discarding the first 1,000 time steps to eliminate transient fluctuations):

1. if the trajectory is located on a point attractor, all 500 values in the set are identical, and a single point appears on the plot (*e.g.*, when $W < 1.0$);

2. if the trajectory is located on a periodic attractor, each of the states visited appear as discrete points (*e.g.*, the period 2 cycle that appears when $1.0 < W < 1.5$);

3. if the trajectory is located on a chaotic attractor, a smear of points is produced as the system visits a series of unique points within a given neighbourhood (*e.g.*, as occurs when $2.3 < W < 5.0$).

In general, the location of a basin of attraction in dynamic space moves in a gradual fashion as W is varied. At some values of W bifurcations occur and the nature of the attractor changes (*e.g.*, around $W = 1.0$ and $W = 1.5$ the attractor bifurcates into a two-cycle and a four-cycle respectively). Between some adjacent values of W, the system dynamics change in a discontinuous fashion, suggesting that the trajectory may be jumping between different basins of attraction (*e.g.*, around $W = 5.2$ and $W = 6.0$).

While bifurcation diagrams provide a qualitative picture of how the type of attractor changes as W is varied, they do not give any indication of the stability of that attractor. One possible measure of dynamic stability, the Lyapunov exponent, is investigated in the following section.

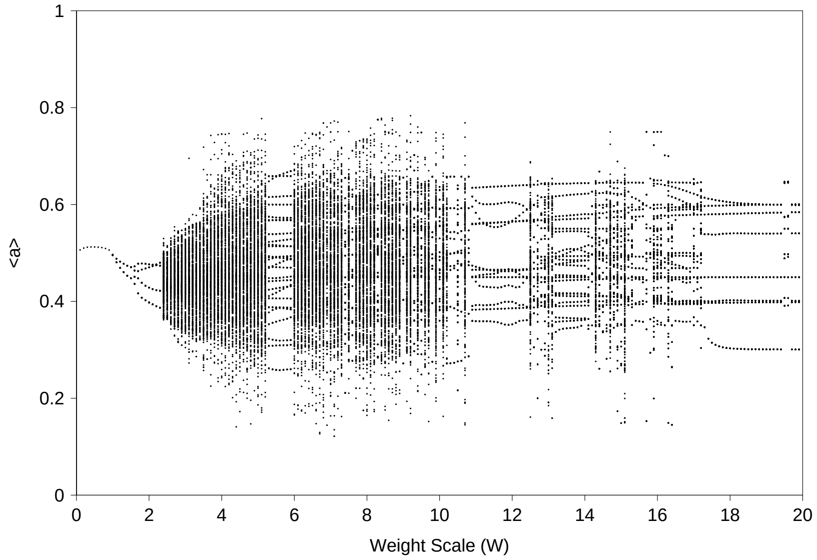


Figure 4.3: Orbit diagram for a fully connected network with 20 nodes. Each point represents the average activation $\langle \mathbf{a} \rangle$ for a single trajectory state. Each vertical slice represents all states visited on the trajectory originating from a single initial condition. Qualitative features that are visible include fixed point, cyclic and chaotic attractors, bifurcations and discontinuities. Full details of the generation and interpretation of the diagram are provided in the text.

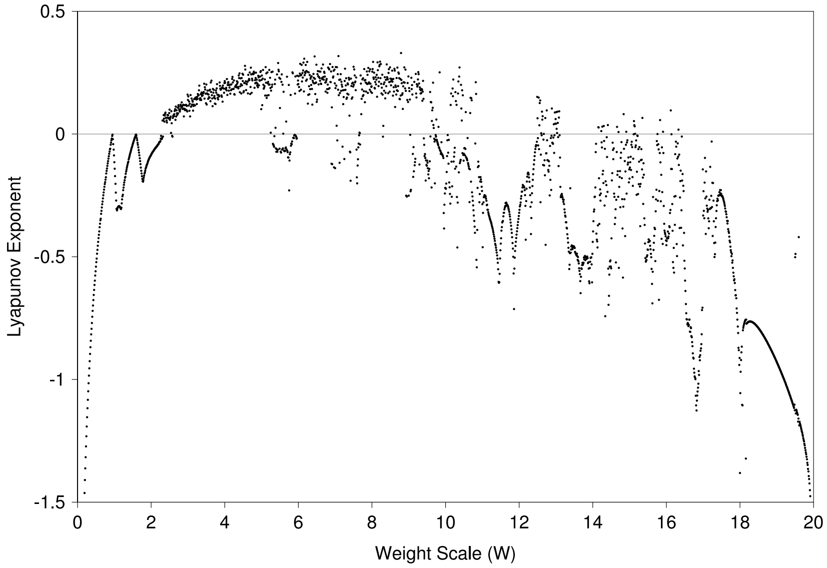


Figure 4.4: An example Lyapunov diagram for the network used to generate Figure 4.3. Each trajectory is associated with a single Lyapunov value: positive values indicate diverging trajectories (chaotic attractors) and negative values indicate converging trajectories (fixed point and periodic attractors). Full details of the generation and interpretation of this diagram are provided in the text.

4.2.2 Lyapunov exponents

As described in §4.1 above, in a chaotic system, two trajectories that have separation δ_0 at time 0 will diverge over time. If δ_t is the separation at time t, and $|\delta_t| \simeq |\delta_0|e^{\lambda t}$ then λ is known as the Lyapunov exponent, and it measures the exponential rate at which the two trajectories will diverge (Strogatz, 1994). Alternatively, in a stable system, λ will be negative, indicating the rate at which two nearby trajectories converge to an attractor.

In reality, for an N-dimensional system, there are actually N Lyapunov exponents. The spectrum of Lyapunov exponents of an N-dimensional dynamic system can be conceptualised by imagining the time evolution of an infinitesimally small, N-dimensional sphere. Over time the sphere will become an ellipsoid and if we let $\delta_k(t), k = 1, \ldots, N$ denote the kth principal axis of the ellipsoid, then $|\delta_k(t)| \simeq |\delta_k(0)e^{\lambda_k t}|$ where λ_k are the Lyapunov exponents, each describing the expansion or contraction of the ellipsoid in the n dimensions. Over time, the diameter of the ellipsoid will be dominated by the most positive λ_k, therefore λ is the *largest* Lyapunov exponent (Wolf et al., 1985). From this point on in this thesis, any mention of the Lyapunov exponent will refer to the largest Lyapunov exponent:

$$\lambda = lim(t \to \infty)log(\frac{\delta_t}{\delta_0}) \tag{4.3}$$

The magnitude of the Lyapunov exponents provides a quantitative picture of a system's dynamics in information theoretic terms, measuring the rate at which systems create or destroy information (Wolf et al., 1985). In a practical sense, the magnitude of a positive exponent corresponds to the time scale on which a system's dynamics become unpredictable. The magnitude of a negative exponent corresponds to the rate at which a system approaches an attractor.

While it is possible to calculate the spectrum of Lyapunov exponents directly from a set of differential equations, the difficulty of doing so increases with the size of the system under consideration (Wolf et al., 1985). Fortunately, knowledge of the largest Lyapunov exponent is sufficient to identify the qualitative behaviour of a system, and several methods exist for estimating the value of this exponent from time series data (Wolf et al., 1985, Bryant et al., 1990, Sprott, 2003). This technique has previously been applied to the analysis of high-dimensional neural networks (Dechert and Gencay, 1992, Albers et al., 1998, Albers, 2004). In particular, Albers (2004) carried out a substantial investigation of the transition from order to chaos in a parameterised class of high-dimensional dynamic systems. The procedure used to estimate the value of the largest Lyapunov exponent in this study was based on that described by Sprott (2003) (pp.116–117):

1. The state of the network $\mathbf{I}$ was initialised to $(I_0, \ldots, I_{N-1})$.

2. The network was iterated for 1,000 steps to ensure that the trajectory was located on an attractor, rather than a transient.

3. A duplicate network was created and its activation state $(\mathbf{s}')$ was perturbed such that its separation from the state of the unperturbed network $(\mathbf{s})$ was δ_0.

4. Both networks were iterated for a single step.

5. The new distance, δ_1, between the states of the original and perturbed networks was measured as:

$$\delta_1 = [(\mathbf{s}'_0 - \mathbf{s}_0)^2, ..., (\mathbf{s}'_{N-1} - \mathbf{s}_{N-1})^2]^{1/2}.$$

6. The log of the ratio of the two distances, r_t, was calculated and recorded as:

$$r_t = ln(\frac{|\delta_1|}{|\delta_0|})$$

7. The state of the perturbed network was modified such that the direction of its trajectory was unchanged, but its distance from the trajectory of the original network was restored to δ_0 (see Figure 4.5) by adjusting the activation of each node as follows:

$$\mathbf{s}'_n(t) = \mathbf{s}_n(t-1) + \frac{\delta_0}{\delta_1}[\mathbf{s}'_n(t-1) - \mathbf{s}_n(t-1)]$$

8. Steps 4 to 7 were repeated for $t = 500$ iterations and the value of the largest Lyapunov was estimated by taking the average of the log ratios:

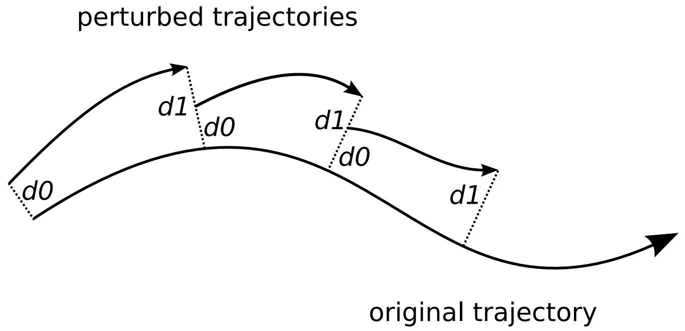


Figure 4.5: A graphical representation of the procedure used to estimate the largest Lyapunov exponent. The continuous line represents the original (unperturbed trajectory). $d0$ is the original separation between the unperturbed and perturbed trajectories. $d1$ is the separation between the two trajectories after a single iteration. After each iteration, the perturbed trajectory is adjusted so that its separation is $d0$ in the direction of $d1$. (Figure redrawn from Sprott, 2003, p. 117)

$$\lambda = \frac{1}{t}\sum_{i=1}^{t} r_t$$

The number of time steps required to assure sufficient accuracy in steps 2 and 8 was found to vary with the size of the system. Preliminary trials were used to determine the number of time steps required to ensure that the network was located on an attractor (in step 2) and that the value of the Lyapunov exponent had converged (in step 8). As with the orbit diagram, this procedure was repeated for 100 values of W in the range $[0.1, 20.0]$ and the values of λ obtained were plotted (Figure 4.4).

Figure 4.4 provides a complementary view of network dynamics to that provided by the orbit diagram shown in Figure 4.3. Whereas Figure 4.3 showed the locations of attractors, Figure 4.4 shows their stability. The nature of the attractors can now be verified:

1. When W is below 1.0 and the system is stable, the Lyapunov exponent is negative;

2. Around $W = 1.0$, the attractor bifurcates and the Lyapunov exponent approaches 0.0.

3. When the system is chaotic, the Lyapunov exponent is positive. Positive Lyapunov exponents are much noisier due to the non-repeating nature of the trajectory; the exact value of the exponent varies throughout the trajectory.

As W becomes very large, the activation function saturates, the system begins to approximate a Boolean network (Kauffman, 1993) and the Lyapunov exponent drops below 0.0. At this stage, the network is very robust to small perturbations to activation.

Even when there is no qualitative change in the type of an attractor, its stability changes as W is varied and the location of the attractor changes. The sudden jumps in the location of the attractor that were observed in the bifurcation diagram are matched by sudden changes in the value of the Lyapunov exponent. One limitation of the method described above is that, unless the DRGN contains only a single attractor, only a portion dynamic space is being measured (that of the basin containing the initial state **I**). The following section addresses this shortcoming by sampling more widely from the set of initial conditions.

4.2.3 Multiple initial conditions

Figures 4.3 and 4.4 illustrate the behaviour of a system when it is initialised to a single starting condition. For systems with more than one basin of attraction, this means that only a subset of the possible dynamics are captured. A more comprehensive picture of system dynamics can be obtained by using multiple initial conditions.

Using multiple initial conditions raises the issue of how to sample the state space of the system. An initial condition consists of a vector of continuous values, therefore there is an infinite number of them, and no possibility of exhaustively testing the entire space of the system (as may be possible for a reasonably sized discrete system). Two different methods for sampling were investigated: random, in which the activation of each node was chosen from a uniform distribution in the range $[0.0, 1.0]$; and systematic, in which the initial states were a subset of the 2^N corners of a N-dimensional hyper-cube embedded in the state space of the system.

A system with 10 nodes has 2^{10} (1024) hypercube corners—a feasible number of initial conditions to test. A system with 20 nodes has 2^{20} (over one million) corners—a less feasible number to test exhaustively. For large systems, a subset of these can be chosen in a systematic fashion to ensure even coverage with a reasonable number of samples.

Figure 4.6 shows the largest Lyapunov exponents for the basins found in the sample system after testing 100 random initial conditions. This figure supports several intuitions about system behaviour described above. At low values of W, all initial conditions lead to trajectories with very similar Lyapunov values, suggesting a single basin of attraction. At higher values of W, different trajectories lead to markedly different Lyapunov values. In the region $7.6 < W < 9.2$ at least three different negative values are visible, as well as at least one positive value. This range suggests that some of the discontinuities in the Lyapunov exponent seen in Figure 4.4 may be due to the initial condition 'jumping' between different basins of attraction. In contrast, in the region $6.0 < W < 6.4$, *all* of the initial conditions led to stable trajectories, while at slightly lower or higher values of W they led to unstable trajectories. It is implausible that all 100 of the random initial conditions are, by chance, located on the boundary between an unstable and a stable attractor, and a more significant alteration in the structure of dynamic space appears likely.

When a system contains only a single basin of attraction (*e.g.*, when W is very small) there will be little difference in behaviour of trajectories originating at different points: they may have slightly different transients, but the long term behaviour will be identical. When a system contains multiple basins of attraction, depending on the nature of the bifurcation that created them, they may have different levels of stability (*i.e.*, one steep and one shallow basin). This phenomenon may explain the discontinuities observed in Figures 4.3 and 4.4: The Lyapunov exponent is a measure local to a given basin of attraction. If, due to the scaling of the weights, the boundaries of the basin of attraction shift in such a way as to place the initial condition in a different basin of attraction, an apparent discontinuity will result. It is also possible that multiple basins of attraction in a single system may be of different qualitative types (*e.g.*, one chaotic and one periodic) which could explain the periodic windows in the otherwise chaotic regions.

The use of Lyapunov exponents as a method for characterising attractors does have limitations. One issue highlighted by the use of multiple initial conditions is the sensitivity of the procedure used to estimate Lyapunov exponents to the parameters of the method. Different initial conditions may produce slightly different

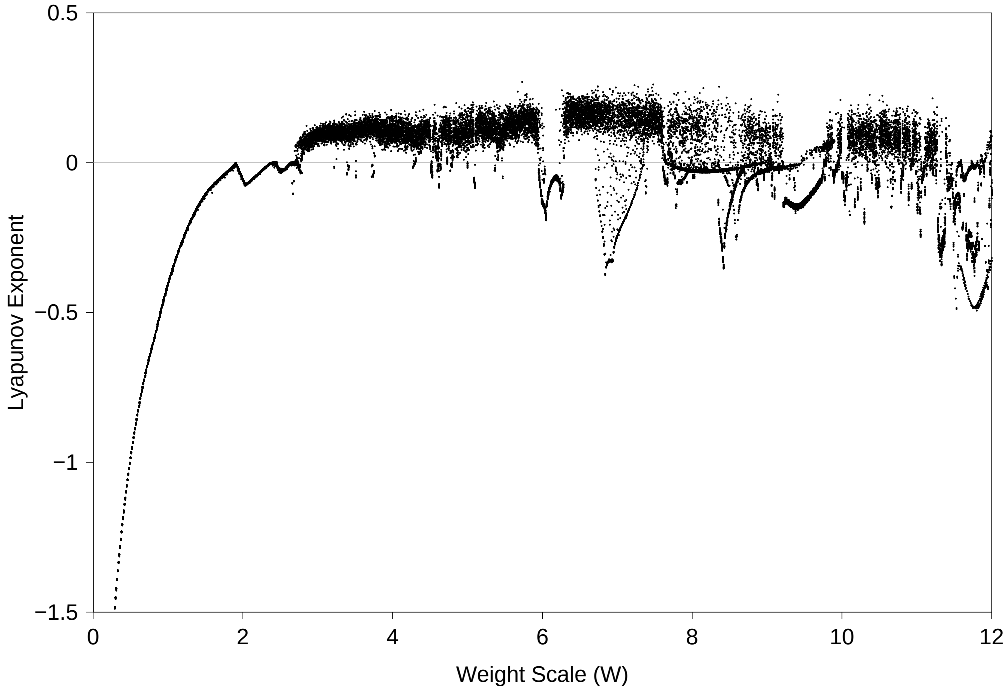


Figure 4.6: The Lyapunov exponents of each of the attractors found from 100 different random initial conditions.

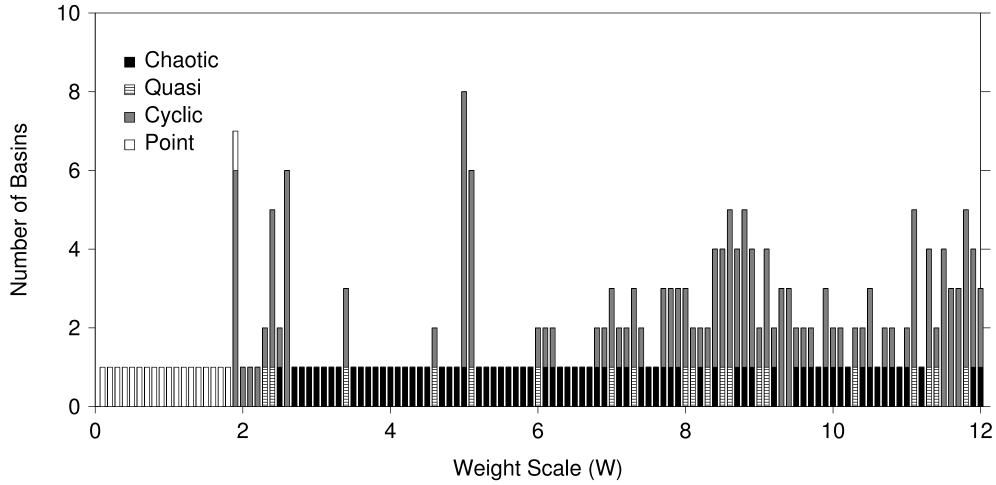


Figure 4.7: The number and type of attractors found for the network illustrated in Figure 4.6. Quasi attractors refer to trajectories for which the Lyapunov was negative (indicating a stable orbit), but no repeating states were observed over the duration of the calculation. See text for further discussion.

Lyapunov exponent values (visible as vertical ‘smears’ in Figure 4.6). In addition, the direction and magnitude of the initial perturbation in step 3 of the Lyapunov estimation procedure will influence the exact value of the exponent. Therefore the value of a Lyapunov exponent is not a unique identifier of an attractor: a more accurate value could be achieved by averaging across several initial conditions and perturbations. Similarly, it is not possible to estimate the exact value of the Lyapunov exponent of a chaotic attractor using this procedure. Due to the fact that the structure (and stability) of a chaotic basin may vary depending on the exact location of a trajectory, there is significant deviation in the estimated values.

4.2.4 Counting and classifying basins

To efficiently characterise a particular system, it is useful to be able to count and classify the number of unique attractors in an automated fashion. This section introduces the methodology used to do so, and provides an estimate of the level of coverage achieved with a given density of initial condition sampling.

When distinguishing between basins, the value of the Lyapunov exponent alone is not necessarily a unique identifier, as multiple basins may be equally stable in the case of a symmetric bifurcation of the state space. Furthermore, as observed above, some basins have an anisotropic structure that results in local variation in the value of the Lyapunov exponent. The one clear discrimination that the Lyapunov exponent can be used to make is between chaotic and stable attractors.

The sign of the Lyapunov exponent was therefore used as the first step in classification, to identify the existence of at least one chaotic attractor. Next, the set of states in the attractor was used as the basis for distinguishing between non-chaotic attractors. Point attractors could be identified by the equality of successive states ($f_t(x) = f_{t+1}(x)$). Periodic attractors of length k could be identified by the equality of states k time steps apart ($f_t(x) = f_{t+k}(x)$). A final possibility was that the Lyapunov exponent was negative (indicating stability), but no repeated states were observed during the 500 iterations over which it was calculated. Possible explanations for this include very long transients or cyclic attractors, quasi-periodic attractors (which are stable but non-repeating) or estimation errors in the Lyapunov calculation.

Figure 4.7 summarises the number and type of attractors observed in the DRGN shown in Figure 4.6. For $W < 1.9$, all initial conditions converge to a single fixed point attractor, whose stability gradually decreases until a bifurcation occurs at $W = 1.9$. For $1.9 < W < 2.3$, all initial conditions still converge to a single attractor; however, the nature of the attractor has changed; it is now cyclic. At approximately $W = 2.5$, a further bifurcation occurs and all initial conditions converge to a chaotic attractor. For $W > 7.0$, the system contains multiple attractors, both cyclic and chaotic.

4.2.5 Summary of preliminary observations

The preliminary investigations reported above indicate that the dynamic behaviour of a DRGN tends to vary with the scale of the weights, W, in a predictable way. Systems with very small weights contain a single fixed point attractor. As W increases, the number of attractors increases and cyclic and chaotic attractors are frequently observed. As W becomes large, chaotic attractors become less frequent, and cyclic attractors predominate. As well as containing multiple attractors, systems frequently contain different types of attractors—for example, both stable and chaotic attractors—depending on the initial state. While significant regularities were observed between the behaviour of different systems of the same size, there was also considerable variation. *Study 2: Characteristics of dynamic space* (reported in §4.3) used the methodology developed in this study to quantify the probability of observing certain types of attractor and the number of attractors observed in a single system. A further open question concerns the manner in which system behaviour changes as W is varied. Two different types of change were observed: gradual and sudden. The structural features responsible for these phenomena were explored in *Study 3: The formation of attractors* (reported in §4.4).

4.3 Study 2: Characteristics of dynamic space

The aim of *Study 2: Characteristics of dynamic space* was to use ensembles of parameterised DRGNs to quantify the relationship between the size, connectivity and weight scale of a DRGN and the number and type of basins it contains.

4.3.1 DRGN ensembles

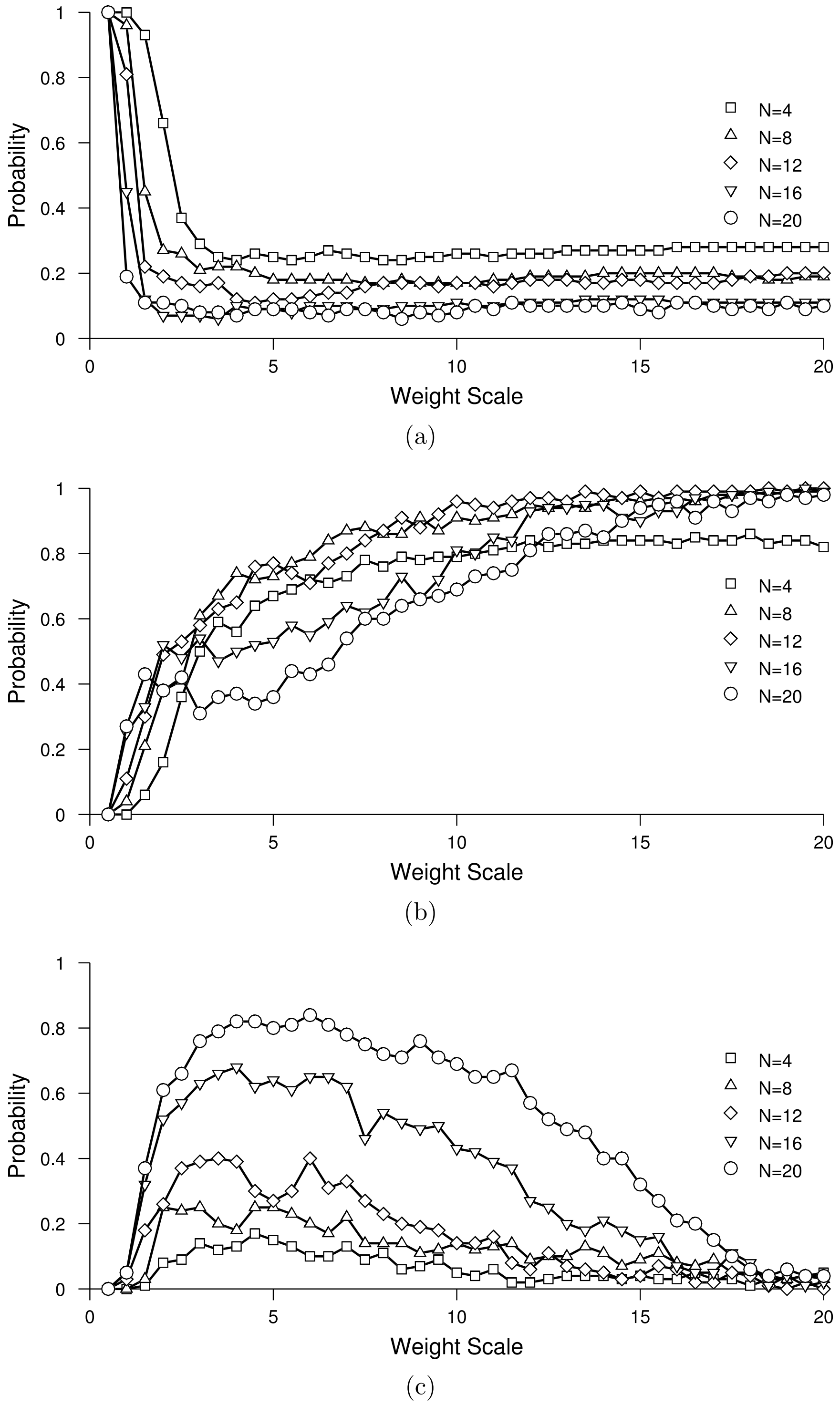
The structure of a DRGN's regulatory component, as defined in §3.3.1, is characterised by the number of nodes (N) and the level of connectivity between those nodes (K). Two series of ensembles were used in this study to investigate the effect of varying each of these parameters. For the first series, five ensembles were generated with $N = \{4, 8, 12, 16, 20\}$. All of the networks in this series of ensembles were fully connected (*i.e.*, $K = N$). For the second series of ensembles $N = 20$ and $K = \{2, 4, 8, 12, 16, 20\}$[1]. Each ensemble consisted of 100 randomly generated base networks. For each of the base networks, 40 values of W were tested in the range $[0.5, 20]$. Therefore, 4,000 networks in total were generated per ensemble. 256 different initial conditions were tested for each network (representing a uniformly distributed subset of hypercube corners). The system was initially iterated 2,000 iterations to ensure that the trajectory was located on an attractor, and the Lyapunov and basin type were calculated over the subsequent 500 time steps. For each network, the number and type of unique attractors was recorded. In order to simplify analysis, trajectories with negative Lyapunov values but no repeated states (as discussed in §4.2.4 above) were omitted.

4.3.2 Ensemble results

The first statistic calculated was the probability of finding *at least* one attractor of a particular type (fixed point, cyclic or chaotic) in a network.

Figure 4.8 shows how the probability of a DRGN containing a fixed point, cyclic or chaotic attractor varies with N and W. The following trends were observed. For all network sizes, the probability of finding a fixed point attractor was 1.0 for very small weight scales ($W = 0.5$) but dropped rapidly to between 0.08 ($N = 16, 20$) and 0.24 ($N = 4$) by $W = 3.0$ and remained at this level as W increased further. The probability of finding a cyclic attractor was 0.0 for $W = 0.5$ and increased gradually as W increased, approaching 0.8 for $N = 4$ and 1.0 for $N = 8, 12, 16, 20$. The probability of finding a chaotic attractor was 0.0 for $W = 0.5$ and increased rapidly as W increased to 5.0 before decreasing gradually to approach 0.0 as W increased to 20.0. As N increased, the peak probability of finding a

[1] The parameters for final ensemble in each of these series are identical ($N = 20$, $K = 20$). This ensemble was therefore only simulated once.



(a)

(b)

(c)

Figure 4.8: Probability of finding a (a) fixed point, (b) cyclic or (c) chaotic attractor for each of the N-series ensembles. Each data point corresponds to a probability calculated over the 100 members of an ensemble for the scaling factor W. The probabilities for given values of N and W can total more than one due to a system containing, for example, both a point and cyclic attractor.

chaotic attractor increased from 0.18 (for $N = 4$, $W = 4.5$) to 0.81 (for $N = 20$, $W = 6.0$).

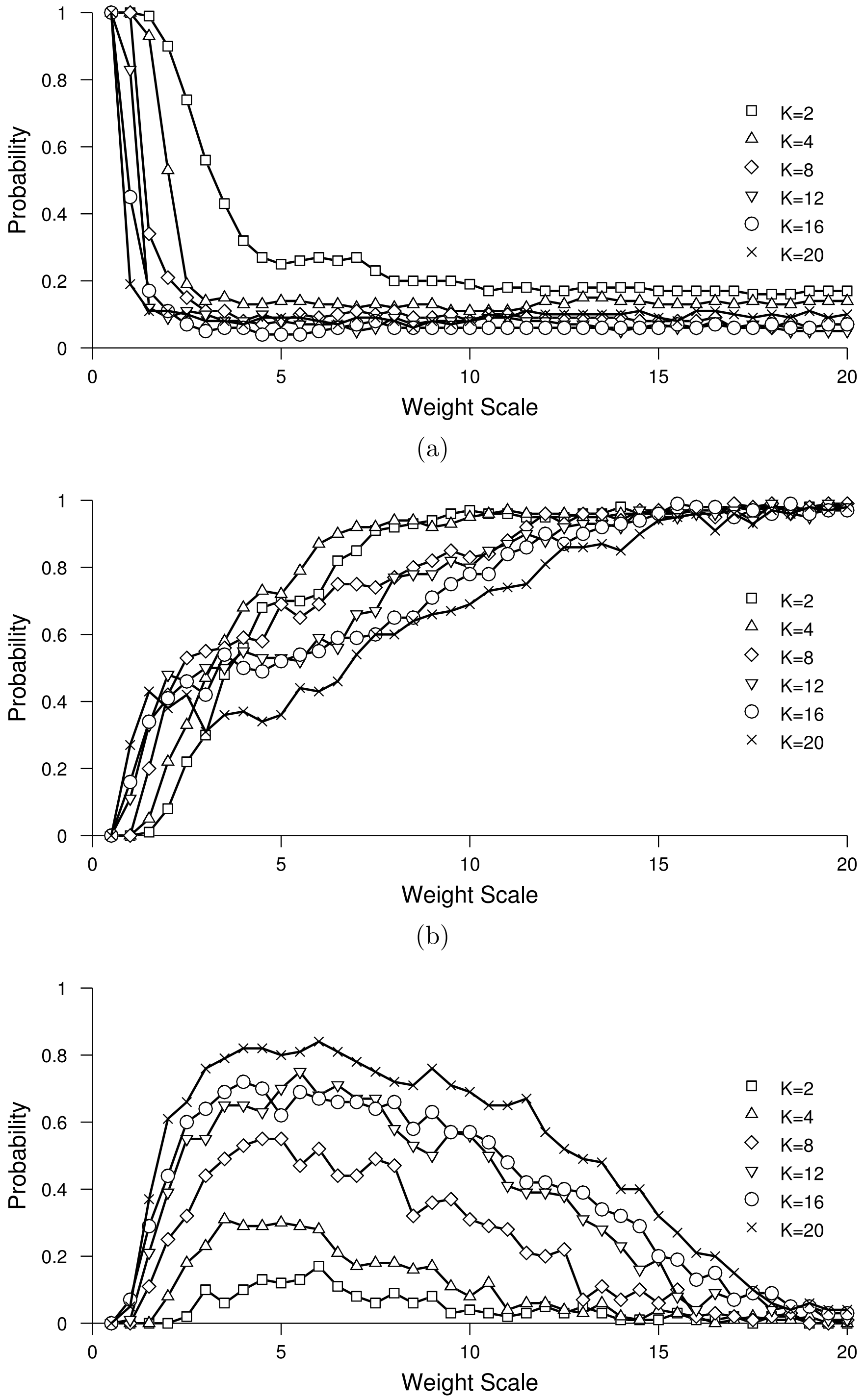
Figure 4.9 shows how the probability of a DRGN containing a fixed point, cyclic or chaotic attractor varies with K and W. In general, the trends in probabilities observed as W was varied mirrored those of the N-series described above (Figure 4.8). Across the range of weight scales, the probability of finding a fixed point attractor decreased slightly as connectivity increased, from 0.19 (for $K = 2$,$W = 20.0$) to 0.03 (for $K = 12$, $W = 20.0$). The probability of finding a cyclic attractor approached 1.0 as W increased to 20.0 for all values of K. The peak probability of finding a chaotic attractor increased as connectivity increased, from 0.17 (for $K = 2$,$W = 6.0$) to 0.82 (for $K = 20$, $W = 6.0$).

The second statistic calculated on the basis of the ensembles was the mean number of stable basins (*i.e.*, fixed point or cyclic) found in a particular network. Figure 4.10 shows the mean number of stable basins for the N-series and the K-series of ensembles. The general trend for both series was that the number of stable basins increased as W increased. While larger networks generally contained more basins than smaller networks, above $N = 12$ the difference was minimal. Similarly, more sparsely connected networks generally contained more basins, but the difference was minimal until network connectivity was *very* sparse ($K = 2$).

4.3.3 Discussion

The number of stable attractors in a DRGN does increase with network size. However, if all genes are fully connected, the rate at which these new attractors are created decreases as further genes are added (Figure 4.8(a)). For large networks, the number of stable attractors increases as the network becomes more sparsely connected (Figure 4.9(b)). One factor that may limit the appearance of new attractors as gene number increases is the increasing probability of some region of the dynamic space being occupied by one or more chaotic attractors (Figure 4.8(c)). Correspondingly, as connectivity is reduced, the probability of chaotic attractors decreases (Figure 4.9(c)).

The results of this study provide quantitative support for the intuitions described earlier concerning the nature of DRGN behaviour changes as the strength of the interactions between nodes (W) is scaled. Networks with weak interactions contain a single point attractor. As the strength of interactions increases, the

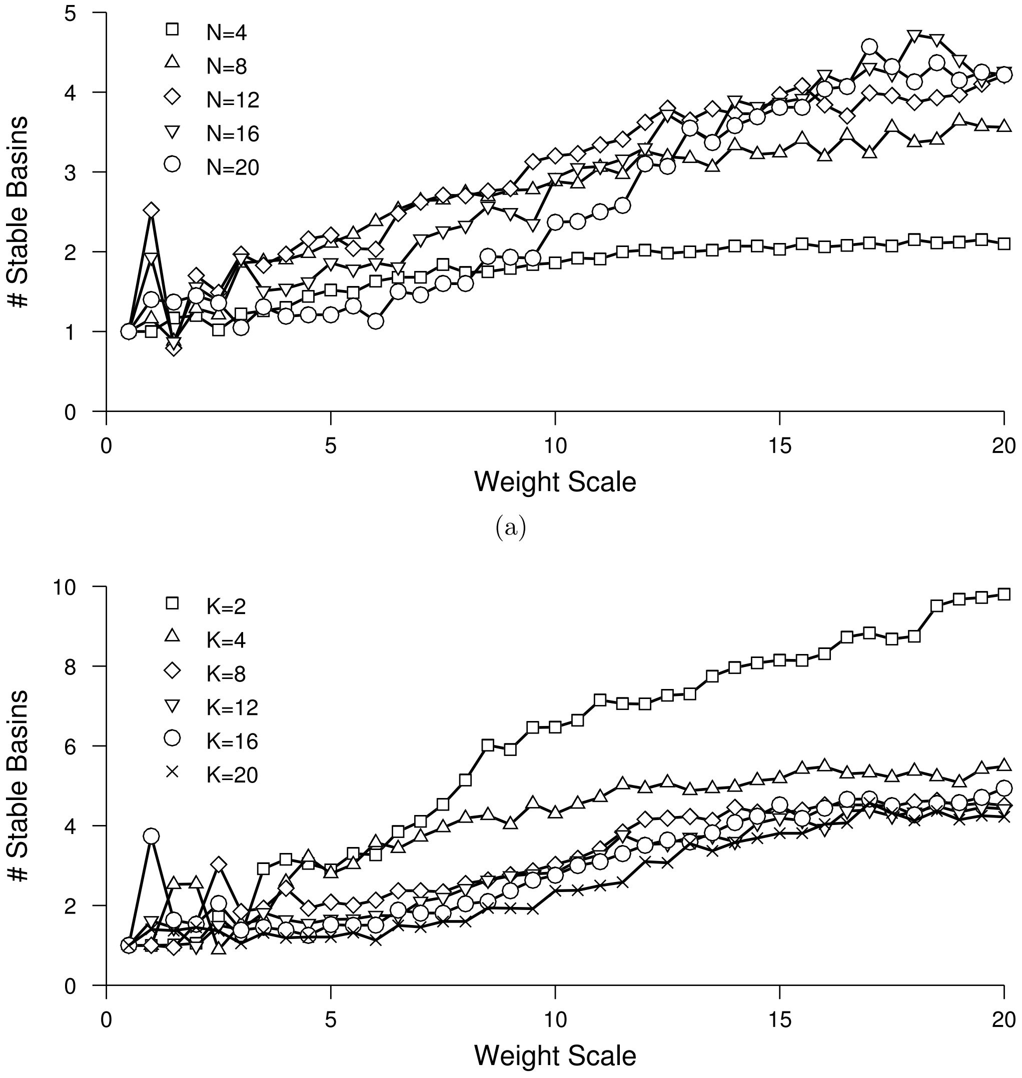


(a)

(b)

(c)

Figure 4.9: Probability of finding a (a) fixed point, (b) cyclic or (c) chaotic attractor for each of the K-series ensembles. Each data point corresponds to a probability calculated over the 100 members of an ensemble for the scaling factor W. The probabilities for given values of K and W can total more than one due to a system containing, for example, both a point and cyclic attractor.



(a)

(b)

Figure 4.10: The average number of stable attractors in a DRGN for the (a) N and (b) K series of ensembles.

probability of a network containing cyclic and chaotic attractors also increases, while the likelihood of fixed point attractors decreases. Further increases in the strength of interactions produces an overall increase in the number of stable attractors and a decrease in the probability of chaotic attractors in a network. These new attractors are overwhelmingly likely to be cyclic, although there is a small but reliable probability of fixed point attractors occurring.

4.4 Study 3: The formation of attractors

Study 2: Characteristics of dynamic space provided an indication of how the attractor dynamics of a DRGN depend on its structural parameters: size, connectivity and weight scale. However, it was observed in our preliminary investigations that the dynamics of individual networks in the same parameter class can vary widely. The aim of *Study 3: The formation of attractors* was to understand how features of network structure interact to produce the patterns of dynamic behaviour observed in the previous studies. In the first section we consider how attractor basins are created as the strength of interactions is scaled. Following that, we demonstrate the existence of upper limits on the occurrence of several types of attractor. Finally, we explore the structural features that interact to reduce the number of attractors that are actually observed in the random networks.

4.4.1 How are attractors formed?

The most straightforward way to understand how the scaling factor W affects a network's dynamics is in terms of its effect on the sigmoid updating function. To simplify the exploration, we began by considering a minimal network consisting of a single node, with a single self-connection. Despite the small size of the network, a range of dynamic behaviours are possible.

When the net input to a node is positive and the value of W is sufficiently small, the graph of $f(x)$ intersects $y = x$ at exactly one point, $x = k$ (Figure 4.11 (a)). In this situation the activation of the node, regardless of its initial state, will eventually converge to $f(k)$. As the value of W is increased, the steepness of $f(x)$ increases until a bifurcation occurs and the graph of $f(x)$ intersects $y = x$ at three points, representing two stable fixed points (derivative less than one, attracting) and one unstable fixed point (derivative greater than one, repelling) (Figure 4.11 (b)). Unstable fixed points are of limited interest in biological systems due to their inherent noisiness; it is unlikely that a system would remain at an unstable fixed point for long before being perturbed into the basin of one of the stable fixed points. This analysis will therefore focus on the stable fixed points of a system. There are two possible stable states for this single node system. If the input is less than k_b, then the system will eventually settle to the fixed point located at k_a; if the input is greater than k_b, then the system will settle to the fixed point located

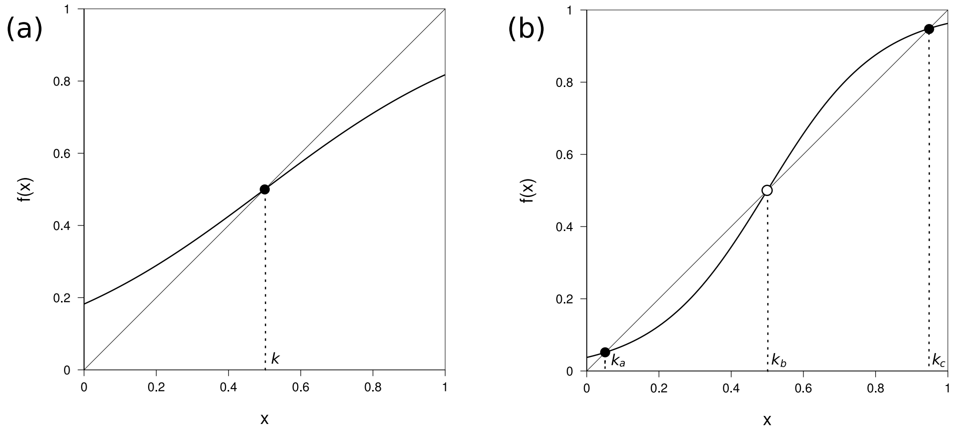


Figure 4.11: The sigmoid activation function for a single node with a positive self-connection showing the location of (a) a single stable fixed point (solid circle) ($W = 3.0$); and (b) two stable fixed points (solid circles) plus an unstable fixed point (hollow circle) ($W = 6.5$).

at k_c (Figure 4.11 (b)).

When the net input to a node is negative, the system can either contain a single attracting fixed point at $x = k$ or a stable cyclic attractor of period two plus an unstable fixed point, depending on the steepness of the sigmoid (Figure 4.12). In the general case, where the activation function is not bounded, it is also possible for such a system to contain only a single repelling fixed point. In the specific case of the DRGN model, the use of the sigmoid function restricts node activation to a fixed range, producing a limit cycle.

As the number of nodes increases, so does the dynamic repertoire of the system. Pasemann (2002) and Thomas (1999) have each described minimal requirements for chaotic dynamics using formalisms related to the DRGN model. The smallest such network consists of two nodes with one positive and one negative feedback loop. While it is arguable whether chaotic dynamics are desirable in a biological system, any system able to generate chaotic behaviour will also, under different parameter settings, be able to generate a wide range of other dynamic behaviours, including cyclic attractors of various periods and fixed point attractors (Pasemann, 2002).

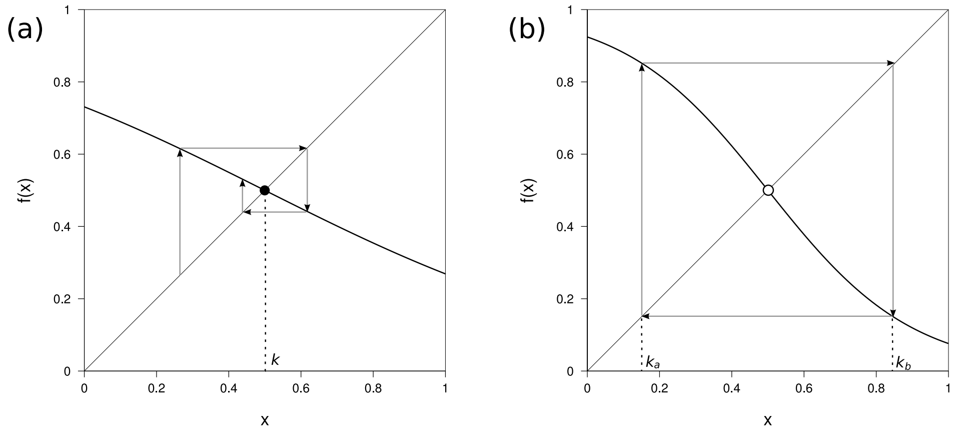


Figure 4.12: The sigmoid activation function for a single node with negative self-connection, showing the location of (a) a single stable fixed point (solid circle) ($W = -2.0$) and (b) a stable limit cycle plus an unstable fixed point (hollow circle) ($W = -5.0$).

4.4.2 How many attractors can a system contain?

The next question addressed concerned the upper limit on the number of stable attractors that a system of a given size may contain. In the general case, this is an open research question (Bagley and Glass, 1996). We therefore focused on a subset of networks that, while restricted, was nonetheless sufficient to demonstrate the point that the number of attractors observed in the ensemble studies reported in §4.3 are far from the maximum possible number.

The class of networks considered consisted of N nodes in which each node was connected to itself but there were no connections between nodes. If each of these nodes is governed by a sigmoid function with a slope in the range $(0, 1)$ then the system has only a single fixed point. On the other hand, if each node is governed by a sigmoid function with slope > 1 at the point of inflection, then the maximum number of fixed point attractors in any one system is 2^N. In addition, $3^N - 2^N$ unstable fixed points are also created (Thomas and Kaufman, 2001).

To demonstrate that this limit is achievable with the DRGN system, we constructed a network with four regulatory nodes in which each node had a self-connection with a unique weight (1.0, 2.0, 4.0 and 8.0) and there were no interconnections between nodes (*i.e.*, the weight matrix was constructed such that all entries on the diagonal were unique and all entries off the diagonal were zero). The

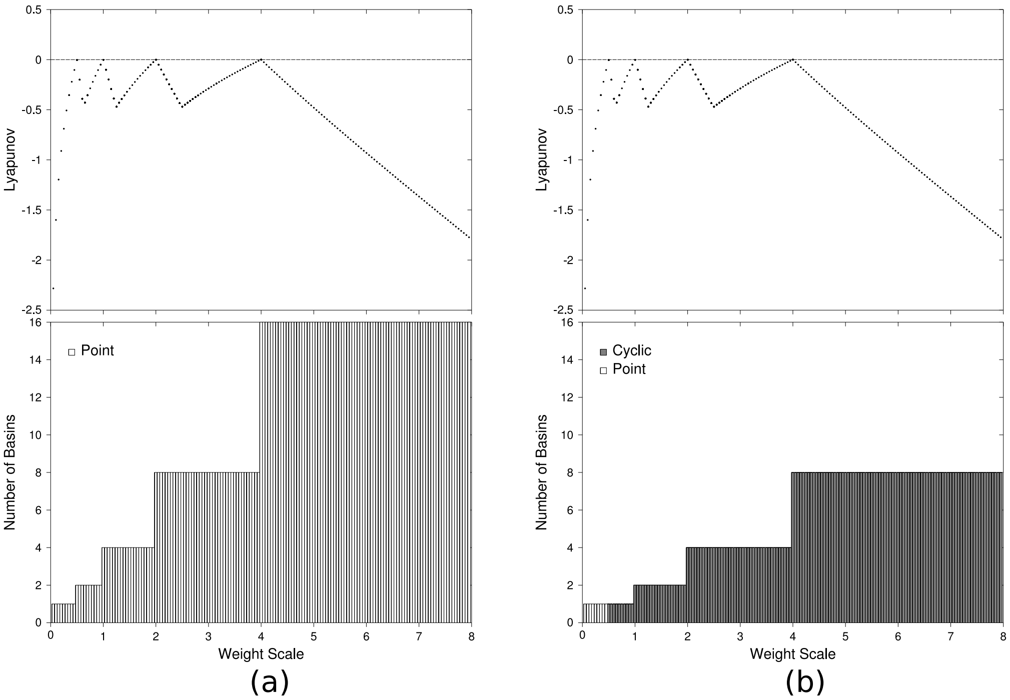


Figure 4.13: Lyapunov exponents and attractor number in an uncoupled system with (a) positive and (b) negative self-connections. In both cases, new attractors are created whenever the slope of the sigmoid governing one of the nodes increases past ±1.0. Further details of the networks are provided in the text.

threshold for each node was therefore equal to half the value of the self-connection. The procedure described in §4.3 was used to count and classify the attractors. 160 values of W in the range $[0.05, 8.0]$ and 16 different initial conditions were tested for each value of W.

As the strengths of the self interactions are scaled the number of basins increases by powers of 2 until the maximum of 16 is reached (Figure 4.13 (a)). Each bifurcation of the basin occurs when the increment to the scaling factor results in the slope of one of the self weights growing above one at the point of inflection. The stability of each of these basins (as indicated by their Lyapunov exponents) is equal, as is their size.

If each of the N nodes in a system is governed by a sigmoid function with slope in the range $(0, -0.5)$ at the point of inflection, the system will converge to a single fixed point attractor. If each of the slopes is less than -0.5, then each node will oscillate with period two. The maximum number of stable periodic attractors is therefore equal to the number of possible combinations of these cycles(*i.e.*, $2^{(N-1)}$).

To demonstrate this limit, the previous investigation was repeated with a second series of four node networks in which each of the self-weights were negative (-1.0, -2.0, -4.0 and -8.0). Again, there were no connections between nodes. All other parameters were as above. For $W < 0.5$, the system contained a single point attractor (Figure 4.13 (b)). As W increased past 0.5, the slope of the sigmoid governing the node with self-connection equal to -8.0 crossed -1, and the attractor changed to a period two cycle. As the strength of the remaining self interactions increased, so did the number of period two attractors, reaching the expected maximum of 8 once the slope of all four sigmoid functions had dropped below -1.

In an uncoupled system of N nodes in which N_+ nodes have positive self-connections and N_- nodes have negative self-connections, the maximum number of stable attractors will be $2^{N_+} \times 2^{(N_- - 1)}$, with the maximum length of any cyclic attractors remaining at 2. To obtain attractors with higher periods, or chaotic attractors, some coupling between nodes is required, as described above. However, while increasing the range of possible dynamic behaviours, coupling can also act to reduce the diversity of observed attractors. The following section considers this situation.

4.4.3 What factors reduce the number of attractors?

As nodes in a system become coupled to one another, they begin to affect each other's behaviour. One effect of this interaction is that a wider range of possible behaviours become possible (Pasemann, 2002). The activation of each node is now a function, not only of its own previous activation state, but also that of the other nodes from which it receives input. To investigate the effect of introducing coupling between nodes, we created four ensembles of DRGNs, one for each combination of positive and negative self-connection and positive and negative coupling. Two weight scaling parameters were used: one controlled the strength of the self-connection and threshold and the other the strength of the interconnections between nodes. 160 values of each of the two parameters in the range $[0.05, 8.0]$ were used, resulting in 25,600 combinations for each of the four ensembles. As before, 16 initial conditions were tested for each network.

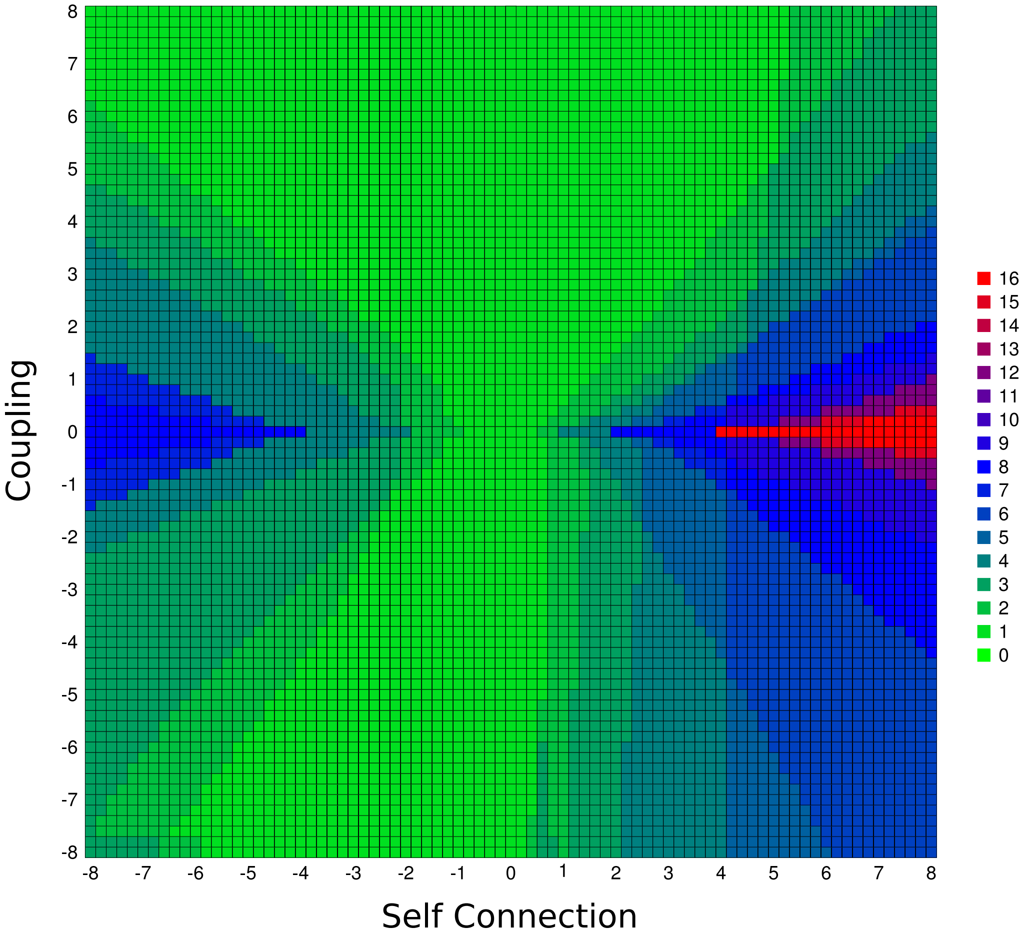


Figure 4.14: Number of basins in a coupled system as the strength of self connection and coupling are varied. Coupling between nodes limits the number of stable attractors that exist in a system.

Figure 4.14 summarises the results of these ensembles. To assist in orienting the heatmap, consider that the horizontal strip through the centre of the diagram (coupling = 0.0) corresponds to the two ensembles shown in Figure 4.13. Moving towards the top or bottom of the heatmap corresponds to increasing the amount of positive or negative coupling between nodes. Clearly, as the strength of coupling increases, the number of stable attractors in a system decreases rapidly. Taking similar slices through the diagram at other locations reveals the gradual disintegration of the basin structure as coupling is increased (Figure 4.15). Not only does the rate at which new attractors are created by bifurcation decrease, but stability of individual attractors diverges. Rather than all attractors being roughly equal in stability, there is now considerable variation.

In addition, the way in which new attractors appear in the system has changed. In the uncoupled system, all attractors had a Lyapunov value very close to zero at the point where a new attractor was created by bifurcation (Figure 4.13). While this does occur, it is more frequently the case that the existing attractors remain more or less stable while a new, barely stable attractor is introduced (indicated by the Lyapunov curves starting near zero in Figure 4.15).

The changes introduced by coupling in these ensembles can be understood by considering the shape of the sigmoid function. In particular, as the relationship between a node's connections and the value of its threshold changes, the point of inflection of its sigmoid function is shifted from $x = 0.5$ (Figure 4.16). If the point of inflection is outside a particular range (depending on weight scale, up to $[0, 1]$), then the sigmoid will only ever cross y = x once, meaning that no bifurcation will ever occur (Figure 4.16). The location of the point of inflection is modified by the threshold on the node. In the network in Section 4.2, the bias was set to half the value of the self weight to ensure maximum sensitivity. As the threshold varies above or below this value, the probability of a system containing the maximum number of basins decreases.

In the uncoupled systems simulated in §4.4.2, all bifurcations followed the same pattern: as the slope of the sigmoid function increased past 1, a single steady state split into two steady states (Figure 4.17). In contrast, when the nodes are coupled, the point of inflection of the sigmoid function could occur at a location other than $x = 0.5$ and a different type of bifurcation could occur (Figure 4.18). As before, the bifurcation results in the creation of one unstable and one additional stable fixed point. The primary qualitative difference is that whereas before the two new fixed points were created very close to original fixed point, now the second fixed point is introduced at the opposite end of the dynamic landscape.

The bifurcation in Figure 4.18 also suggests explanations for the discontinuities in the orbit and Lyapunov diagrams reported in §4.2. Before bifurcation, an input of $x = 0.1$ would be in the basin of attraction of the steady state located at k_a. After bifurcation, the same input would be located in the basin of attraction of the

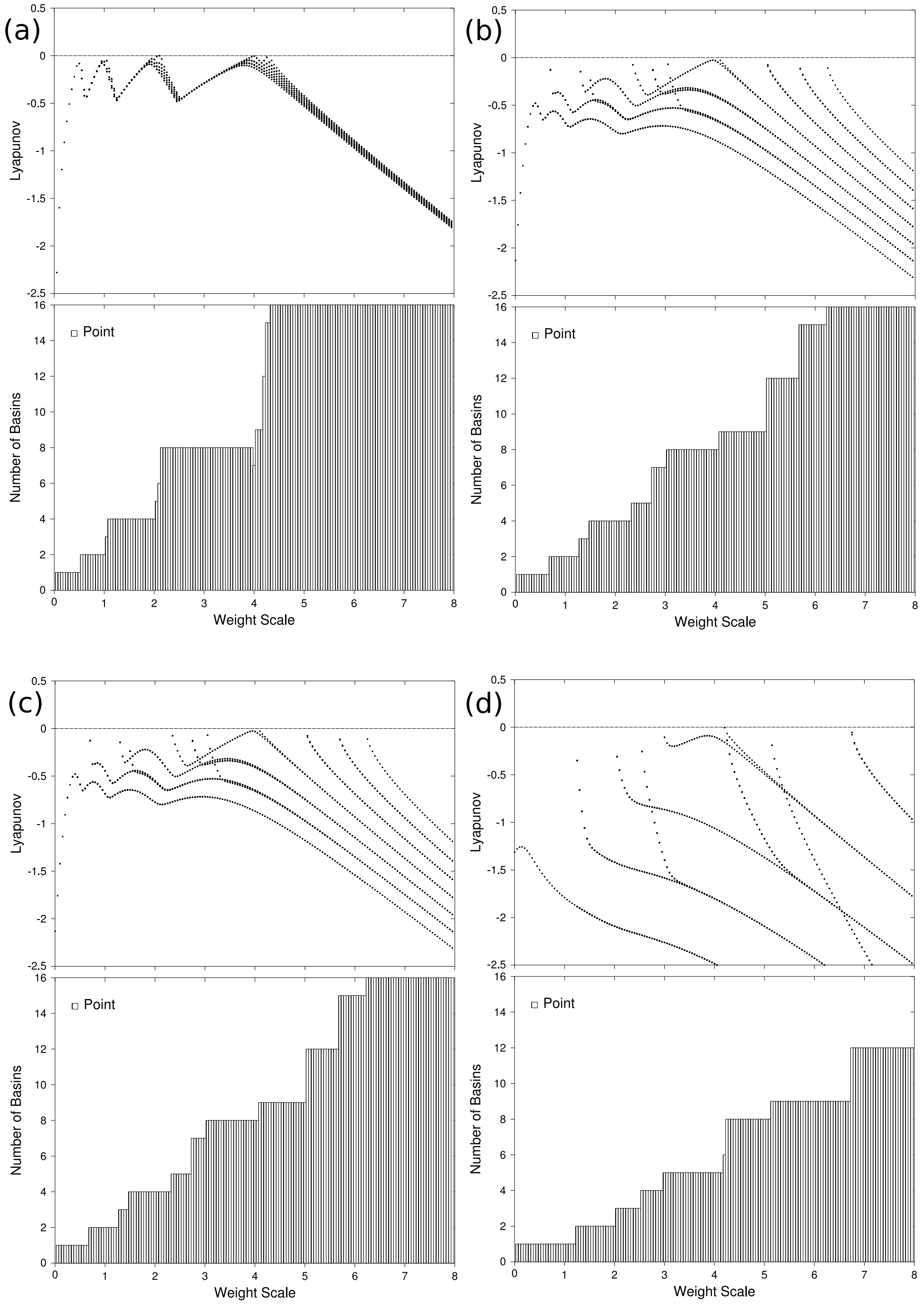


Figure 4.15: The effect of increasing coupling on basin structure. Each of the four nodes has a positive self-connection and coupling between nodes of strength (a) 0.02; (b) 0.08; (c) 0.32; or (d) 0.128. The self-connection was scaled with W but coupling remained fixed.

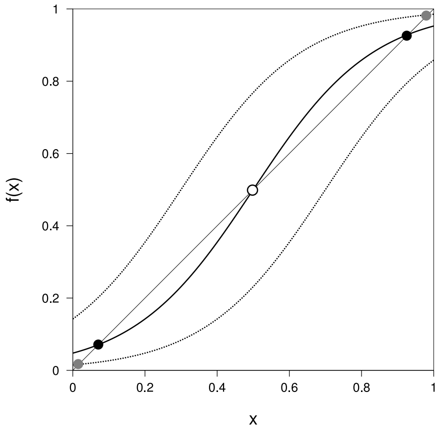


Figure 4.16: The effect of modifying threshold. The unmodified function ($W = 5; \theta = 0.5$) has bifurcated to produce two stable fixed points and one unstable fixed point. As θ is either increased/decreased by 0.2, the function shifts to the left/right and the bifurcation fails to occur, despite the slope being > 1.

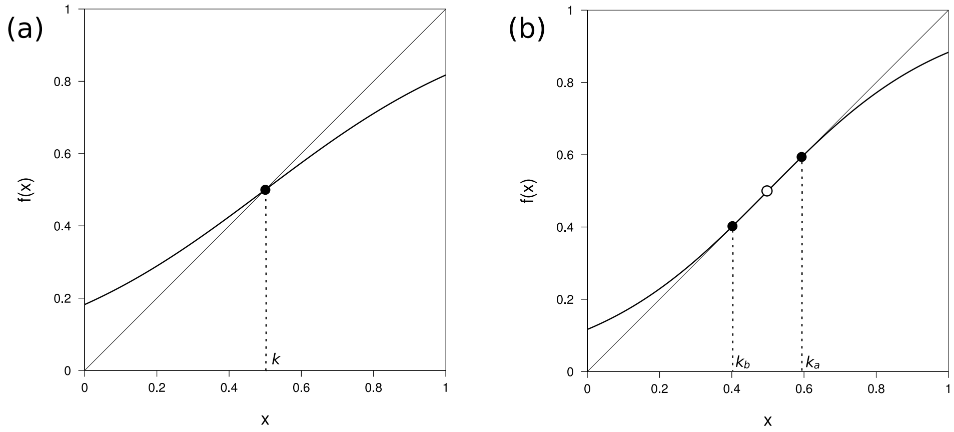


Figure 4.17: One type of bifurcation that can occur as W is increased from (a) 3.0; to (b) 4.0 ($\theta = 0.5$). The original fixed point splits into two new fixed points in a gradual fashion.

steady state located at k_b. As indicated by Figure 4.15, the newly created basins are likely to have different Lyapunov values.

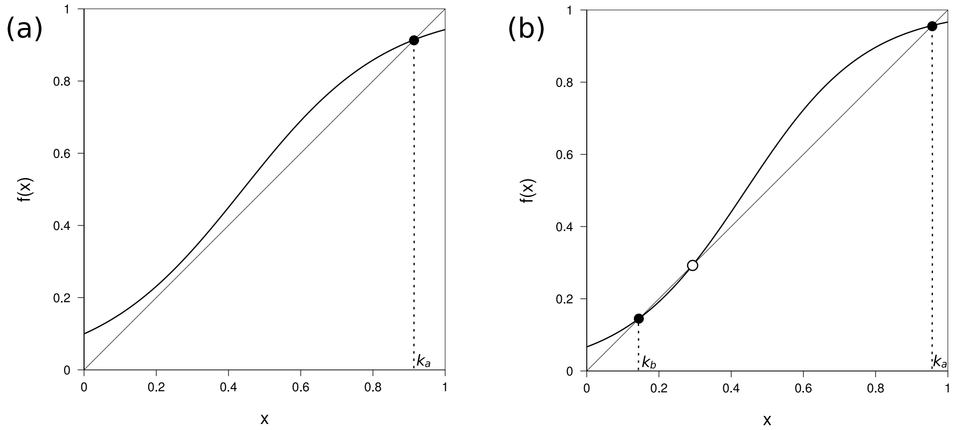


Figure 4.18: A second type of bifurcation that can occur as W is increased from (a) 5.0; to (b) 6.0 ($\theta = 0.44$). The new steady state appears at a considerable distance from the original fixed point.

4.5 Discussion

This chapter has addressed the question of how the dynamic behaviour of a DRGN depends on its structural properties: the number of regulatory nodes, and the degree and strength of their connectivity. The dynamic features of interest, the number and stability of attractors, vary widely among different network topologies. While network dynamics are generally robust to small structural changes, they also exhibit sensitive regions where small structural changes can radically alter the resulting dynamics. The studies reported in this chapter developed empirical methods to characterise dynamic behaviour, applied these methods to quantify how dynamics vary with structural parameters and explored some of the reasons for these variations.

Study 1: Tools for exploring network dynamics (§4.2) assembled a set of tools that could provide both a qualitative and a quantitative account of a dynamic trajectory. Orbit diagrams indicate the type of attractor—fixed point, cyclic or chaotic—that a particular initial condition is located in and show how the nature of this attractor changes as the strength of network interactions (W) is scaled. Both gradual and discontinuous changes in dynamics were observed. Gradual changes included variations in both the location of an attractor and its type; for example, a fixed point attractor bifurcating to form a cyclic attractor of period 2. Discontinuous changes were also observed, where the location or type of an attractor altered rapidly. Lyapunov diagrams added to this picture a quantitative measure of the stability of an attractor, confirming our classification of stable and chaotic attractors. The Lyapunov diagrams also indicated that, as an attractor underwent a bifurcation, its stability decreased temporarily, approaching zero at the point of bifurcation. Furthermore, abrupt changes to the location or type of an attractor also resulted in discontinuities in its stability.

By exploring multiple initial conditions, it was possible to construct a more complete picture of the dynamic space of a network, quantifying both the number and type of attractors observed. Several issues emerged when using this approach to automatically count and classify attractors. The first was the issue of sampling: unlike a discrete dynamic system (such as a random Boolean network) there is no way of exhaustively testing every initial condition of the system. A preliminary investigation comparing two different methods of sampling, one stochastic and one systematic, revealed no indication of bias due to the sampling method used. The second issue concerned the accuracy of the attractor classification procedure when the Lyapunov value of an attractor was very close to zero. For a small proportion of trajectories, typically those occurring around the point at which a system bifurcated, the Lyapunov value was negative (indicating a stable attractor) but no repeated states were observed. Three explanations for this type of behaviour can be identified: quasiperiodic orbits, intermittency, and measurement error. A quasiperiodic orbit occurs under certain conditions; a trajectory is not chaotic (*i.e.*, two nearby points do not diverge over time), but also not repeating. Such a trajectory may be visualised as a line that wraps around a torus, continuing indefinitely but never returning on itself (Strogatz, 1994). Intermittency occurs when a chaotic trajectory comprises periods of reasonably constant values that are interspersed with erratic bursts; the Lyapunov value may therefore be characterising only a local region of an attractor. Finally, measurement error can occur due to the choice of initial condition and perturbation direction: it is possible that a chaotic attractor may display a negative Lyapunov value for some parameter settings.

The application of these methods to ensembles of parameterised networks in *Study 2: Characteristics of dynamic space* (§4.3) indicated several generalisations about the relationship between network dynamics and network structure. Irrespective of the size (N) or connectivity (K), all ensembles showed a similar trend with respect to (W) (Figures 4.8 and 4.9). When W was very small, the network contained a single fixed point attractor. As W increased, the probability of observing a fixed point attractor decreased rapidly, and the probability of observing a cyclic or chaotic attractor increased. As W increased further, the probability of observing a chaotic attractor began to decrease until, for large values of W, almost all attractors were cyclic. The most noticeable effects of varying N and K were on the probability of observing a chaotic attractor, which increased with both size and connectivity (Figures 4.8(c) and 4.9(c)). The number of stable attractors in a system increased slowly with the size of the system; however, above $N = 8$ there was little difference. For larger systems, reducing connectivity produced an increase in the number of stable attractors observed, although values of $K = 2$ were necessary before there was a significant increase.

The relationship between the size of an attractor and the number of observed basins agrees with results obtained by Kauffman (1969) for Boolean systems when W is very large ($\simeq 20.0$). This can be explained by the fact that the sigmoid updating function saturates and approximates a Boolean function in this situation. At lower values of W the number of stable basins scales more slowly with N. This finding agrees with the observation of Bagley and Glass (1996) that the number of stable attractors differs between discrete and continuous systems, primarily due to the appearance of quasiperiodicity and deterministic chaos in continuous systems. The number of observed basins is higher than that predicted by Mochizuki (2005) (who observed that attractor number saturated at a small value as N increased) due to our inclusion of cyclic attractors. Omitting cyclic attractors and considering only fixed point attractors resulted in comparable observations.

Study 3: The formation of attractors sought to explain why increasing the size of a system did not produce an accompanying expansion of its repertoire of dynamic behaviours. Hand-crafted networks were designed that demonstrated upper limits on the number of certain types of attractors. One of the relationships that proved important in maintaining a large number of attractors in a system was a balance between the input connections into a node and its activation threshold. This balance determined the location of the point of inflection of a node's sigmoid updating function. When the point of inflection shifted outside of a certain range, which depended on the weight scale, the number of dynamic attractors in a system was reduced. The location of the point of inflection was also found to have implications for how new attractors are created.

In summary, this chapter has focused on the long-term dynamic behaviour of a single cell—the repertoire of attractor states that can be reached from various initial states. These different attractor states have been equated to the different types that a cell can differentiate into. As size and connectivity increase, greater precision in specifying patterns of interaction is required if the number of attractors is to scale accordingly. What more commonly occurs is that there is a diminishing increase of the number of distinct dynamic behaviours as size and connectivity increase. In order to produce the appropriate number of cells of each type in an organism, development must trigger the correct differentiation trajectory in each cell. In terms of our model, this task may be equated to selecting the initial state of each cell such that its intrinsic dynamics achieve the appropriate long-term behaviour. The following chapter addresses the question of how the initial states of each cell in a developing system can be configured to produce desired patterns of cell fates.

Chapter 5

The Structure and Composition of Ontogenetic Space

Ontogenetic space comprises all possible developmental trajectories. Phenotypic evolution occurs via modifications to ontogeny, therefore understanding the structure and composition of this space can provide insight into the space in which adaptation occurs. During development, the differentiation of cells into specific types is coordinated in a spatial and a temporal fashion to produce organised forms. These patterns of differentiation are guided by gene expression dynamics. Different gene networks will produce different dynamics and, as a result, different ontogenetic trajectories will be generated.

The studies in Chapter 4 considered the long term dynamics of DRGNs and demonstrated that a single DRGN could display multiple dynamic behaviours depending on the basin of attraction in which its initial conditions are located. Multicellular organisms generally contain cells of many different types and during development each cell follows a unique trajectory that leads it to differentiate appropriately. In the context of the DRGN model, during development the dynamic state of each cell has become positioned in the appropriate basin of attraction.

In the previous chapter an assumption was made that the initial dynamic state of a DRGN could be located in a particular basin of attraction. What was not considered was how this configuration of initial conditions occurred. This chapter focuses on the control processes that set up the initial state for each cell of an organism (Figure 5.1).

Dynamic recurrent networks, as a class of computational devices, are known

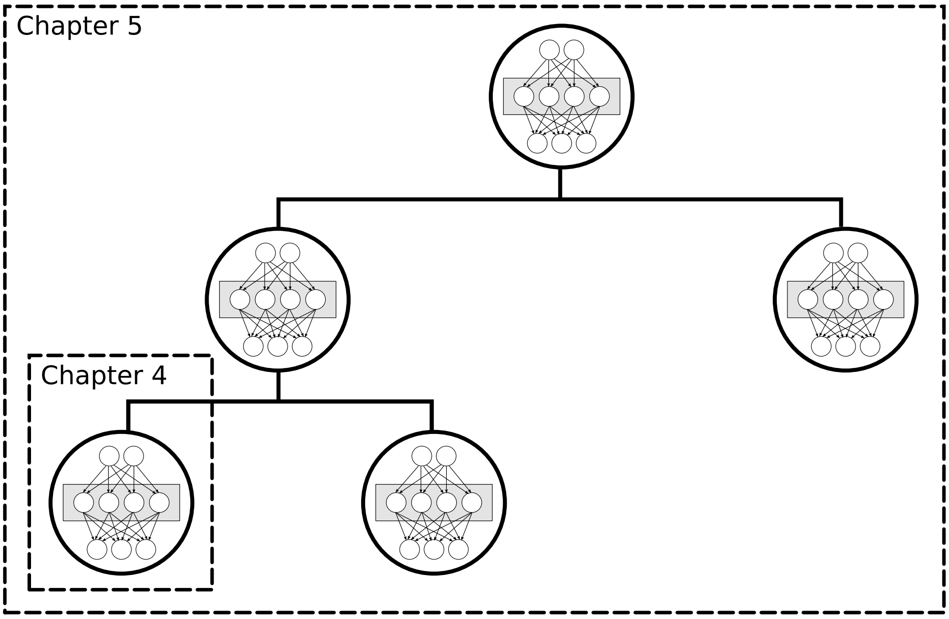


Figure 5.1: The relationship between the DRGN model explored in Chapter 4 and the combined network-lineage model explored in Chapter 5. While the previous chapter considered the long term dynamics characteristic of cell differentiation, this chapter focuses on the transient dynamics that establish developmental patterns.

to be capable of a highly flexible range of behaviours. It is likely that their characteristic dynamic properties, such as cyclic and chaotic behaviour, will affect the distribution of lineages they generate. The overall goal of this chapter is to investigate the space of ontogenies generated by DRGNs. The specific aims of this chapter were to develop tools to characterise and explore ontogenetic space, to quantify how the composition of ontogenetic space varies with DRGN size and connectivity and to evaluate the robustness of DRGN-generated ontogenies to structural and dynamic perturbation.

There were two requirements for characterising ontogenetic space: one or more metrics for classifying and comparing lineages; and a means of visualising and exploring the parameter space. The first two studies reported in this chapter (§5.1 and §5.2) addressed these requirements. The third study (5.3) used these metrics and tool to characterise the distribution of ontogenies and phenotypes generated by DRGNs. Comparisons were carried out both among DRGNs with different structural properties and between DRGNs and stochastic processes. A final study (§5.4) investigated the robustness of phenotypes and ontogenies to perturbation.

To address the questions about the interaction between evolution and development raised in Chapter 3, we required DRGNs capable of generating cell lineages comparable to those observed in biological organisms. The studies in this chapter provide a guide to the organisation of ontogenetic space, and an indication of which combinations of genotypic and developmental parameters are likely to produce interesting behaviour.

5.1 Study 4: Characteristics of cell lineage complexity

As with the dynamic attractors studied in the previous chapter, classifying lineages required a metric that could be used to distinguish and differentiate in a quantitative fashion. At the scale of a single cell, the feature of interest was the long term dynamic behaviour of the genetic system. At a multicellular scale, the feature of interest is the pattern of division and differentiation events that produce organisation in a developing embryo—represented here as a cell lineage. Early developmental biologists tended to compare the development of different organisms in a qualitative fashion (Gilbert, 2003). When quantitative measurements are possible, they typically apply to individual characters, such as limb length (Young and Hallgrímsson, 2005), or specific properties, such as speed of development (Vancoppenolle et al., 1999), rather than entire ontogenies. In one sense, development seems too rich and diverse a process to be encapsulated in a single metric. One interesting property that has potential to be quantified is complexity (Bonner, 1988). A question of particular interest to evolutionary biologists is whether complexity has increased or decreased over evolutionary history, or whether there is a measurable trend at all (Valentine et al., 1994, McShea, 1996, 2005). A challenge of this research agenda is how to quantify complexity.

Complexity is an amorphous subject: while easy to recognise in an intuitive fashion, it has proven difficult to crystallise these intuitions into a formal definition (Adami, 2002). One standard view from complex systems is that dynamic processes may be divided into a spectrum of different categories; at one end of the spectrum lie stable or periodic processes and at the other end lie random processes. Both stable and random processes are considered to be simple, while complex processes fall somewhere in between the transition from ordered, periodic processes to disordered, random processes (Solé and Goodwin, 2000).

One class of complexity measures that has been proposed focuses on the nucleotide sequences that constitute an organism's genome. Because sequences are amenable to formal characterisation, concepts from mathematics and physics were readily applicable (see, *e.g.*, Badii and Politi, 1997). Measures of sequence complexity suffer from two related problems: they tend to focus on characterising the difficulty of predicting the next symbol in a sequence, rather than the meaning of the sequence; furthermore, the mapping from nucleotide sequences to protein structure and organismic function is highly nonlinear, and sequence complexity may not necessarily equate to ideas about complexity at a functional level (Adami, 2002).

A second class of complexity measures focuses on the structural form of an organism (reviewed by McShea, 1996). These measures focus on the number of different types of parts or interactions in an organism and the degree of hierarchical structure. Definitions of structural complexity may be further divided into those that measure the complexity of an object (morphological complexity) and those that measure the complexity of a process (developmental complexity).

The results presented in §4.3 indicated that, in certain parameter combinations, DRGNs produce a very limited range of dynamic behaviour that is unlikely to produce interesting (non-trivial) developmental patterns. In order to identify which regions of parameter space are associated with the most interesting regions of ontogenetic space for the studies reported in this chapter, it was necessary that the measure of complexity could be applied in an automated fashion. Given the lack of consensus on how to measure complexity, a definition was needed that suited our requirements for classifying lineages. Therefore, a pragmatic approach to assessing measures of complexity was adopted: each metric was compared to *a priori* notions of what constituted a non-trivial, or 'interesting', lineage from the perspective of developmental control. The first step toward defining a measure of complexity was to identify the intuitions that it needs to capture:

- lineages containing more cells are likely to require more complex control than those with fewer cells; and

- heterogeneous lineages are likely to require more complex control than homogeneous lineages, where heterogeneity may be measured both at the cellular level (number of cell types) and the multicellular level (number of different arrangements of cells).

In the following section several notions of structural complexity are reviewed and formalised in such a way that they can be applied to cell lineages. Four metrics were based on concepts from the literature. A fifth metric, weighted complexity, was designed to address several shortcomings identified with the application of existing measures to cell lineages. For conciseness, all metrics, including weighted complexity are presented together. These metrics are then applied to a set of sample lineages to assess how they correspond to our intuition of what constitutes a non-trivial control task.

5.1.1 The complexity metrics

Morphological metrics

The metrics discussed below focus broadly on morphological aspects of an organism; that is, in terms of how a cell fate distribution is described.

Number of cells: One of the simplest proposed indicators of the complexity of an organism is its size. Bonner (1988) argues that as as organisms grow larger they must by necessity become more complex as the internal requirements for supporting their larger size become more specialised. The advantage of this metric is its simplicity. However, interpreting this metric requires caution: applied strictly it would imply that larger organisms are always more complex than smaller organisms, which is clearly not always the case (despite being larger, a blue whale is not necessarily more complex than a dolphin). The number of cells of a cell lineage was defined as its number of terminal nodes.

Number of cell types: Bonner (1988) also proposed that an increase in the complexity of an organism would be reflected by an increase in the number of specialised cell types it contained. As with the number of cells, this metric has considerable intuitive appeal. A potential problem with applying this metric to real organisms is the difficulty in classifying different types of cells and the potential for bias in favour of more well-studied organisms (McShea, 1996). The number of cell types of a cell lineage was defined as the number of different types of terminal nodes. Given the definition of the DRGN-lineage model, an upper limit was imposed on this metric by the structure of the underlying control network. Therefore, while there was some scope for variation due to a DRGN not employing all possible cell types during development, its range was limited by the number of output nodes in the DRGN.

Number of hierarchical levels: Between two organisms containing an equal number and type of components, there may be a difference in how those components are arranged. Hierarchical, in this context, refers to the number of levels of nestedness of a morphology (McShea, 2001). For example, in the sequence "organelle, cell, organ, organism", each component contains and partially constrains the behaviour of the earlier components. Given that a cell lineage captures an ontogenetic, rather than a physical, relationship between cells, it is not possible to define a formal measure of hierarchy in this context. However, if we accept a relationship between the ontogeny of a cell and its morphological context, the algorithmic measures of complexity described below may be taken as a proxy for levels of hierarchical organisation.

Developmental metrics

An alternative approach to defining complexity metrics in terms of describing an object is to describe a process. The following metrics focus broadly on the development of an organism; in terms of how a cell fate distribution is generated.

Algorithmic complexity (deterministic): One approach to measuring the complexity of a system is by considering the length of its shortest algorithmic description—an approach formalised as Kolmogorov complexity (Badii and Politi, 1997). A measure of cell lineage complexity based on Kolmogorov complexity was introduced by Braun et al. (2003) and further refined by Azevedo et al. (2005). This measure is calculated by transforming a cell lineage into a series of unique production rules of the form $X \rightarrow \{Y, Z\}$, indicating that a cell of type X divides to form cells of type Y and Z. X is necessarily an undifferentiated (non-terminal) cell, while Y and Z may be differentiated (terminal) or undifferentiated. The plane of cell divisions is lost ($X \rightarrow \{Y, Z\}$ is equivalent to $X \rightarrow \{Z, Y\}$), but otherwise these rules provide a complete description of a lineage. This initial set of rules

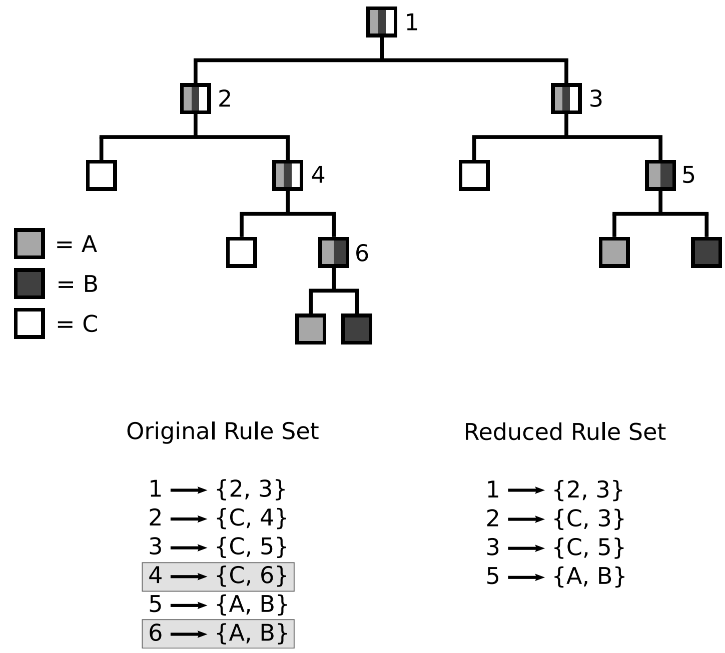


Figure 5.2: An example application of the deterministic algorithmic complexity metric. First, a cell lineage is transformed into a set of production rules by creating a rule for each division. Redundant rules (highlighted) from this set are then removed. Algorithmic complexity is measured as the proportion of unique cell divisions (Alg. Cx. = $4 \div 6 \simeq 0.67$).

is then reduced by removing equivalent rules until a minimal set is arrived at. Deterministic algorithmic complexity is then defined as the size of this minimal set as a proportion of the total number of divisions (Figure 5.2).

Algorithmic complexity (non-deterministic): One of the implications of the algorithmic complexity measure used by Azevedo et al. (2005) is that each of the rules corresponds to an intermediate cell state that will always produce an identical sublineage in a deterministic fashion. An alternative view is that an intermediate cell state could define the subset of terminal cell fates possible in that sublineage, but not necessarily the exact structure of that sublineage. To investigate the effect of this definition, we defined a second algorithmic complexity metric in which the production rules were non-deterministic. Rules now took the form of $\{A, B, C\} \rightarrow \{\{A, B\}, C\}$, where $\{A, B, C\}$ and $\{A, B\}$ are undifferentiated cells that will eventually give rise to differentiated cells of types A, B and C, or A and B, respectively, and C is either a differentiated cell that may or may not continue

Table 5.1: Complexity values for sample lineages

Lineage (Figure 5.3)	Number of cells	Number of cell types	Algorithmic complexity (det.)	Algorithmic complexity (non-det.)	Weighted complexity
A	**32**	1	0.1613	0.0645	2.064
B	**32**	**4**	0.1935	0.1290	4.128
C	18	**4**	0.7059	0.6471	**11.65**
D	4	**4**	**1.0**	**1.0**	4.0
E	2	2	**1.0**	**1.0**	2.0

to divide in a proliferative fashion (*i.e.*, to give rise to more cells of type C). As with deterministic algorithmic complexity, redundant rules are removed from this set, and non-deterministic algorithmic complexity is defined in terms of the size of this minimal set as a proportion of the total number of divisions.

Weighted complexity: As discussed further below, each complexity metrics displayed certain limitations when applied to cell lineages. A final complexity metric was designed to address these limitations, that combined both morphological and developmental aspects of complexity. Weighted complexity was defined as the product of the number of cells in a lineage and the non-deterministic algorithmic complexity of a lineage.

5.1.2 A comparison of complexity metrics

The previous section described five formalised complexity metrics: two morphological, two developmental and one that combines both morphological and developmental aspects. To provide a comparison of how each of these metrics worked in practice, they were applied to a set of five sample lineages (Figure 5.3). The sample lineages were chosen to exhibit a range of lineage types: a large, homogeneous lineage (A); a large, heterogeneous lineage with regular structure (B); a medium-sized heterogeneous lineage with irregular structure (C); and two smaller lineages (D and E). The results of applying each of the metrics to this sample set are summarised in Table 5.1. For each metric, the most complex lineage(s) are indicated in bold.

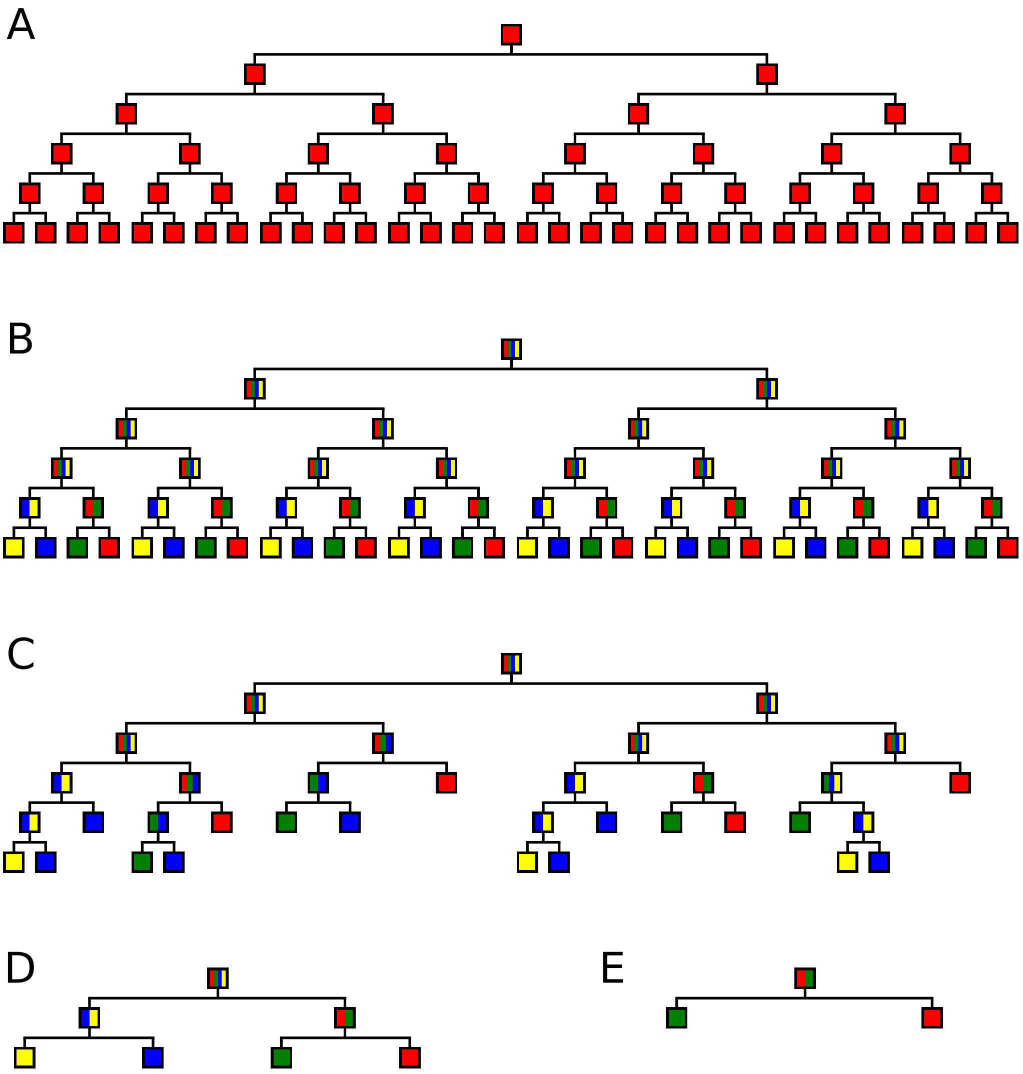


Figure 5.3: The five sample lineages used to compare complexity metrics: a large, homogeneous lineage (A); a large, heterogeneous lineage with regular structure (B), a medium-sized heterogeneous lineage with irregular structure (C) and two smaller lineages (D and E).

Number of cells: Using number of cells as a measure of complexity clearly satisfies our first intuition about complexity: that lineages containing more cells are more complex. However, one significant disadvantage is apparent: counter to our second intuition, the large, homogeneous lineage (A) is ranked as being of equal complexity to the large, heterogeneous lineage (B) and of greater complexity than all of the smaller, heterogeneous lineages (C, D and E).

Table 5.2: Deterministic and non-deterministic rule sets for sample lineages

	Deterministic			Non-deterministic		
Lineage	Rule Set			Rule Set		
A	1	$\rightarrow$	$\{2,2\}$	$\{R\}$	$\rightarrow$	$\{R\},\{R\}$
	2	$\rightarrow$	$\{3,3\}$			
	3	$\rightarrow$	$\{4,4\}$			
	4	$\rightarrow$	$\{5,5\}$			
	5	$\rightarrow$	$\{R,R\}$			
Y	1	$\rightarrow$	$\{2,2\}$	$\{R,G,B,Y\}$	$\rightarrow$	$\{R,G,B,Y\},\{R,G,B,Y\}$
	2	$\rightarrow$	$\{3,3\}$	$\{R,G,B,Y\}$	$\rightarrow$	$\{R,G\},\{B,Y\}$
	3	$\rightarrow$	$\{4,4\}$	$\{R,G\}$	$\rightarrow$	$\{R\},\{G\}$
	4	$\rightarrow$	$\{5,6\}$	$\{B,Y\}$	$\rightarrow$	$\{B\},\{Y\}$
	5	$\rightarrow$	$\{Y,B\}$			
	6	$\rightarrow$	$\{G,R\}$			
C	1	$\rightarrow$	$\{2,3\}$	$\{R,G,B,Y\}$	$\rightarrow$	$\{R,G,B,Y\},\{R,G,B,Y\}$
	2	$\rightarrow$	$\{4,5\}$	$\{R,G,B,Y\}$	$\rightarrow$	$\{R,G,B,Y\},\{R,G,B\}$
	3	$\rightarrow$	$\{6,7\}$	$\{R,G,B,Y\}$	$\rightarrow$	$\{R,G,B\},\{B,Y\}$
	4	$\rightarrow$	$\{8,5\}$	$\{R,G,B,Y\}$	$\rightarrow$	$\{R,G,\},\{B,Y\}$
	5	$\rightarrow$	$\{9,R\}$	$\{R,G,B,Y\}$	$\rightarrow$	$\{R\},\{G,B,Y\}$
	6	$\rightarrow$	$\{8,10\}$	$\{R,G,B\}$	$\rightarrow$	$\{R\},\{G,B\}$
	7	$\rightarrow$	$\{11,R\}$	$\{B,Y\}$	$\rightarrow$	$\{B,Y\},\{B\}$
	8	$\rightarrow$	$\{12,B\}$	$\{R,G\}$	$\rightarrow$	$\{R\},\{G\}$
	9	$\rightarrow$	$\{G,B\}$	$\{G,B,Y\}$	$\rightarrow$	$\{G\},\{B,Y\}$
	10	$\rightarrow$	$\{G,R\}$	$\{G,B\}$	$\rightarrow$	$\{G\},\{B\}$
	11	$\rightarrow$	$\{G,12\}$	$\{B,Y\}$	$\rightarrow$	$\{B\},\{Y\}$
	12	$\rightarrow$	$\{Y,B\}$			
D	1	$\rightarrow$	$\{2,3\}$	$\{R,G,B,Y\}$	$\rightarrow$	$\{R,G\},\{B,Y\}$
	2	$\rightarrow$	$\{Y,B\}$	$\{R,G\}$	$\rightarrow$	$\{R\},\{G\}$
	3	$\rightarrow$	$\{G,R\}$	$\{B,Y\}$	$\rightarrow$	$\{B\},\{Y\}$
E	1	$\rightarrow$	$\{G,R\}$	$\{G,R\}$	$\rightarrow$	$\{G\},\{R\}$

Number of cell types: Using number of cell types as a measure of complexity mitigates the problem with large homogeneous lineage (A); however, it suffers from two disadvantages. At the scale of the lineages considered here, it is a very coarse measure: lineages B, C, and E are all classified as being equally complex despite the differences in the size and regularity of the lineage. Furthermore, as mentioned above, in the DRGN-lineage model, cell type number is constrained at the point of model definition: a lineage can never contain more than N_O cell types.

Algorithmic complexity (deterministic): The first of the developmental complexity metrics had several advantages over the morphological metrics. The large, heterogeneous lineage (B) was ranked as having greater complexity than the large, homogeneous lineage (A). The smaller but less regular lineage (C) was ranked as having greater complexity than both lineages A and B. However, problems emerge with the remaining small lineages (D and E). Lineage D contains only a single cell division event, which is by definition unique, and therefore obtains a maximal complexity value of 1.0, as does lineage E, with three unique division events. A further limitation is less obvious but may be detected by considering the rule set that describes lineage A. Intuition suggests that this lineage is a product of a single rule (X → X, X) applied a fixed number of times. However, the procedure described generates unique rules for each level of non-terminal cells, resulting in a larger rule set than anticipated (Table 5.2).

Algorithmic complexity (non-deterministic): The ordering of the complexity values according to the non-deterministic algorithmic complexity measure is identical to that of the deterministic complexity measure, although the actual values differ. Lower complexity values assigned to lineages A (60% lower), B (33% lower) and C (9% lower) reflect a loss of information about the structure of the lineage. Equal complexity values are assigned to lineages D and E indicating that the first problem identified with deterministic algorithmic complexity persists: lineages D and E are assigned disproportionately high complexity values. Considering the rule sets indicates that the problem of repeated proliferative divisions producing new rules at each level has been addressed by the introduction of non-deterministic rules (Table 5.2). One possible disadvantage of this metric is that there is no longer a unique mapping between a rule set and a lineage. The rule set describing lineage A can be used to describe homogeneous lineages containing any number of levels of cell division.

Weighted complexity : The final complexity metric addresses several of the limitations of the previous metrics. By incorporating lineage size into the measure, the bias of the algorithmic complexity measures towards small lineages has been balanced. Specifically, the small lineages (D and E) are no longer assigned disproportionately high complexity values as with the other algorithmic definitions. Conversely, the large, homogeneous and regular lineages (A and B) no longer have excessively large complexity values due to their size alone.

5.1.3 Discussion

Any measure of complexity will have strengths and limitations and be, to some extent, specific to a particular task or observer. In the absence of a single accepted definition of complexity, the decision to use any one metric over another was guided by pragmatic requirements. Given the focus of this research on the control of developmental processes, the most suitable complexity metric was deemed to be one which reflected the number and diversity of control decisions required to produce a given lineage.

This study illustrated the strengths and limitations of several different measures of complexity from the perspective of control decisions. The results suggest that a size-weighted algorithmic complexity measure, based on a modified version of the lineage complexity metric introduced by Azevedo et al. (2005), accords with our intuitions about which lineages should be considered more or less complex.

Study 5: Visualisation of ontogenetic space used the metrics described in this section to explore how lineages, as quantified by complexity, varied over parameterised regions of space.

5.2 Study 5: Visualisation of ontogenetic space

While the complexity metrics described above are helpful in providing a quantitative measure of a lineage, they lack descriptive power. To understand how ontogenetic space is structured complementary techniques were required to provide a visual representation of how lineages are located with respect to one another in ontogenetic space.

The studies reported in Chapter 4 revealed that the space of possible long-term dynamic behaviours of DRGNs can be very large. Cell lineages are a product not of long-term dynamics, but of the transient dynamics experienced by a system prior to reaching a stable attractor (Bolouri and Davidson, 2003). The number of possible transients in a system is significantly greater than the number of stable states, resulting in a space of possible ontogenies that is very large.

Methods for the effective visualisation of tree structures (of which cell lineages are an example) are of interest in several research domains. In particular, researchers in phylogenetics, the field of evolutionary biology concerned with resolving the relationships between organisms, have developed a number of computational tools to assist with the visualisation of large trees (*e.g.*, Trooskens et al., 2005, Sanderson, 2006). In addition Braun et al. (2003) describe ALES, a software tool for interactively visualising cell lineages either imported from data files or generated according to stochastic rules. In general, the tools that exist were developed to address requirements that differed from those arising in this study. A major focus of many tree visualisation tools is how to usefully convey the information contained in very large sets of hierarchically structured date. In comparison, in the initial stages of development, we were concerned not so much with displaying lineages that were individually very large as we were with allowing large numbers of smaller lineages to be rapidly compared and related to one another.

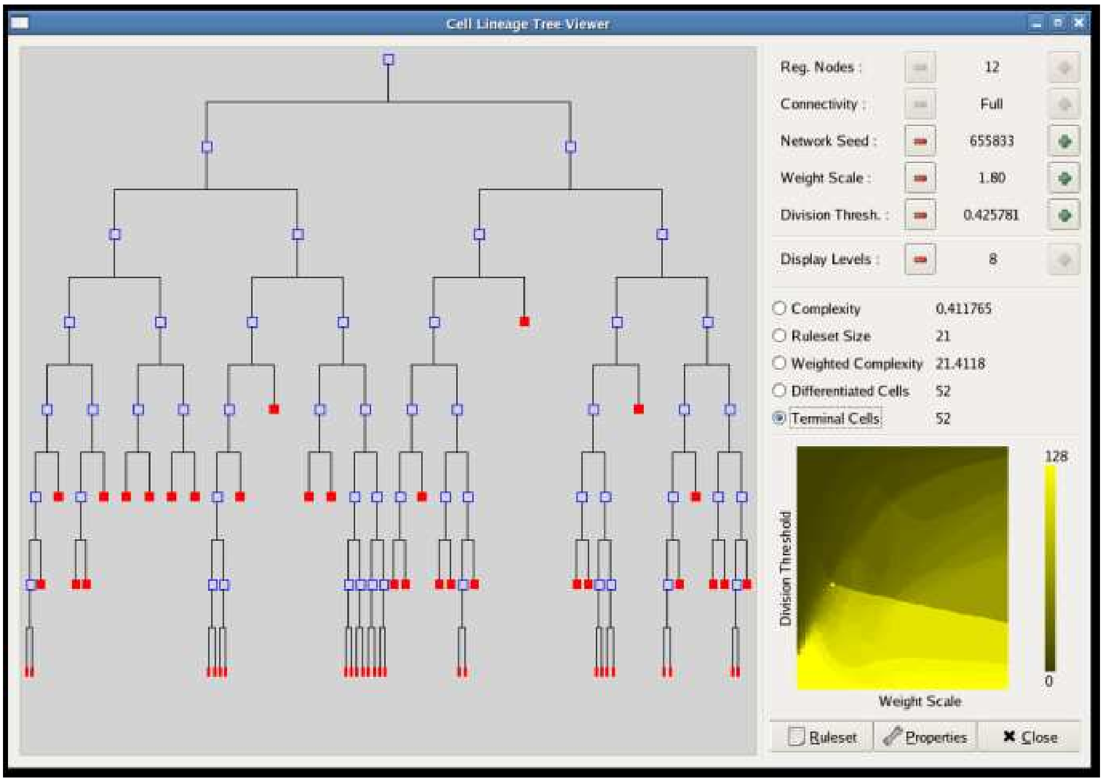
The design of our visualisation tool was also subject to the following requirements:

- there must be some notion of 'relatedness' between lineages that defines neighbourhoods in ontogenetic space;
- it must be possible to visualise how the complexity metrics vary across these neighbourhoods; and
- it must be possible to visualise a large number of lineages.

5.2.1 *TreeView*: Lineage visualisation tool

An interactive visualisation tool, *TreeView*, was designed that addressed each of these requirements[1]. The main issue in defining relatedness between lineages was how to reduce the dimensionality of ontogenetic space such that it could be visualised effectively. The approach adopted—parameterizing two of the variables in the system—is described in the following section. The second and third requirements were addressed by using an interactive heatmap as an interface to the individual lineage trees. The design of a suitable interface raised several practical issues; a full description of the interface is provided in Appendix A.

[1] *TreeView* and preliminary results pertaining to Studies 5, 6 and 7 were reported in (Geard and Wiles, 2006).



Reg. Nodes : 12

Connectivity : Full

Network Seed : 655833

Weight Scale : 1.80

Division Thresh. : 0.425781

Display Levels : 8

○ Complexity 0.411765

○ Ruleset Size 21

○ Weighted Complexity 21.4118

○ Differentiated Cells 52

● Terminal Cells 52

Figure 5.4: A screenshot from *TreeView*: a tool for exploring ontogenetic space. The heatmap in the bottom right represents the complexity gradients over a parameterised slice of ontogenetic space and can be coloured according to each of the complexity measures described in §5.1.1. The main panel on the left shows the current cell lineage. The controls in the top right allow the size, connectivity and random seed of the DRGN to be altered.

Parameterizing ontogenetic space

Both the DRGN and the cell lineage components of the model are amenable to parameterisation. As described in Chapter 4, DRGNs can be parameterised by size (N), connectivity (K) and weight scale (W). When considering the individual behaviour of a single network (rather than the average behaviour of an ensemble of networks), varying either N or K causes a discontinuous transformation. That is, changes to the pattern of interaction in a single network produces new behaviour that is effectively uncorrelated with its prior behaviour. In contrast, W may be varied continuously in an incremental fashion, resulting in smoother transitions between behaviours.

The developmental component of the model can also be parameterised. The most obvious axis along which to parameterise is the value of the division threshold scaling parameter λ. As described in Chapter 3, a cell will divide if the activation of its division node is above a certain value, θ_d. This value increases over developmental time according to $\theta_d(n) = \theta_d(0)e^{(\lambda n)}$, where n is the number of cell division events separating the current cell from the initial zygote.

One property of parameterizing λ is that the sequence of lineages obtained is monotonic; that is, once a particular transition from lineage l_a to l_b has been observed upon increasing from λ_a to λ_b, no further increases to λ_b will result in l_a being observed again. Therefore if two lineages l_a and l_b are equivalent, all values of λ between λ_a and λ_b will also produce equivalent lineages. As a result of this property, it is possible to efficiently explore the space of possible lineages in the direction of the λ axis in a recursive fashion as follows:

```
RecursiveExplore($\lambda_a, \lambda_b$, $depth$):
    generate lineages $l_a$ and $l_b$
    declare float $\lambda_c = \frac{\lambda_a+\lambda_b}{2}$
    generate lineage $l_c$
    if ($l_a \neq l_c$) and ($depth > 0$):
        RecursiveExplore($\lambda_a, \lambda_c$, $depth - 1$)
    else if ($l_c \neq l_b$) and ($depth > 0$):
        RecursiveExplore($\lambda_c, \lambda_b$, $depth - 1$)
    else return
```

Initially, the RecursiveExplore procedure was called with λ_a equal to 0.0 and λ_b equal to 1.0. As the procedure was called recursively, this range was continually subdivided, with regions of equivalent lineages being ignored and regions of varying lineages being explored in greater detail. The depth parameter imposed a limit on the level of recursion. Increasing depth resulted in a map with greater resolution along the λ axis, at the expense of increased processing time.

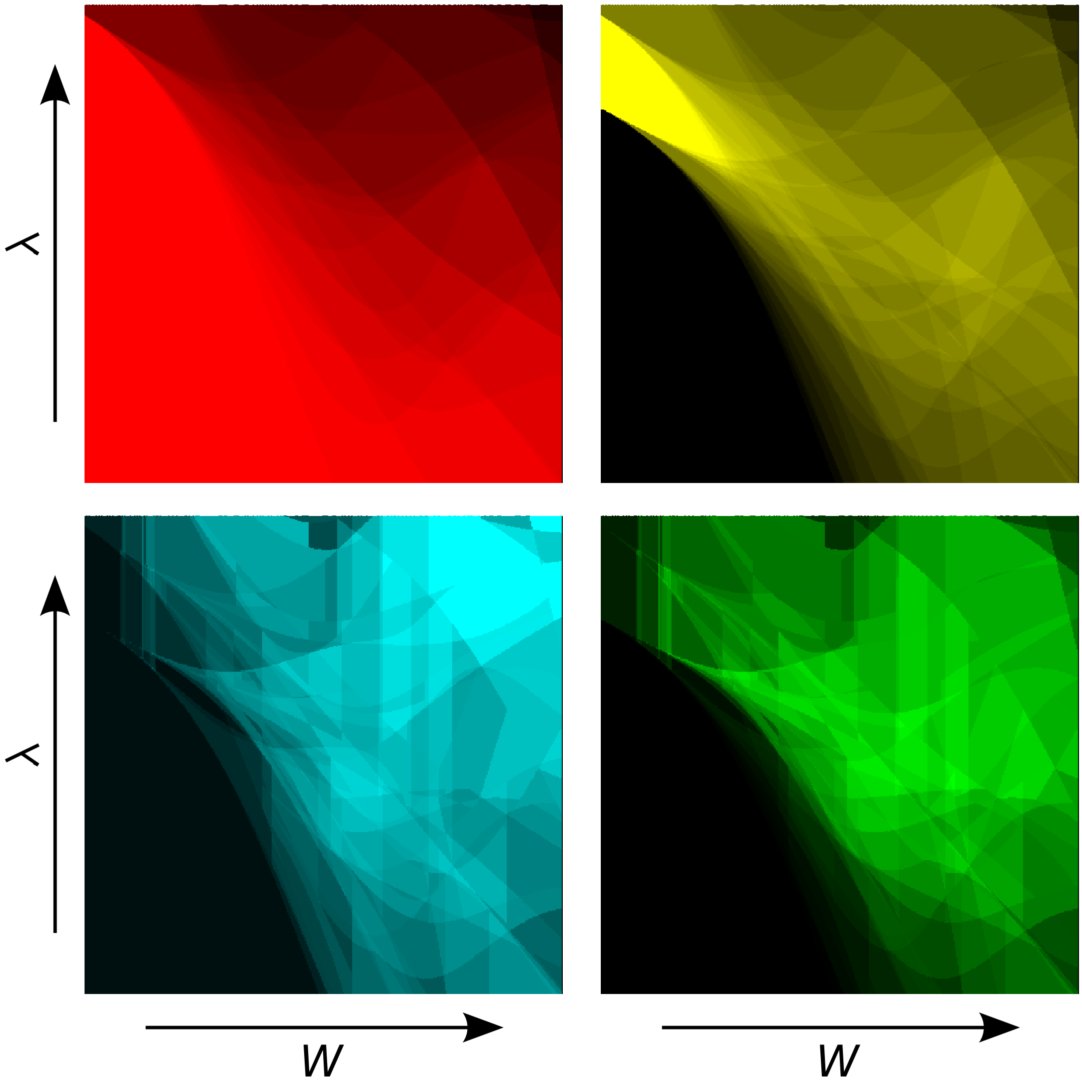


Figure 5.5: Four Complexity heatmaps for the slice of ontogenetic space around the network ($N = 8, K = 8, \lambda = [0, 1.0], W = [0.01, 2.0]$). The four complexity metrics are: number of terminal cells (top left), number of differentiated cells (top right), non-deterministic complexity (bottom left) and weighted complexity (bottom right).

5.2.2 Insights into ontogenetic space

The most immediate observation was that different patterns of network interaction (*i.e.*, generated from different random seeds) displayed considerable variability in the complexity heatmaps that were produced. Whereas some DRGNs mapped to regions of ontogenetic space filled with a diverse range of lineages (such as that shown in Figure 5.5), others mapped to much more homogeneous regions containing only a limited number of different lineages. This observation corroborates the finding reported in §4.3 that networks may differ widely in the class of dynamics they produce despite sharing the same basic parameters.

Some aspects of the complexity maps *did* recur over multiple random networks. For example, if W was low, high values of λ were required for any differentiation to occur. As W increased, the probability of an initial cell never dividing increased, particularly when λ was high. Otherwise, the shape of the transitions from low to high complexity was strongly dependent on the properties of the individual network.

Comparing the complexity maps produced by different metrics on the same set of lineages revealed areas of both similarity and difference. While the location of the major complexity transitions were consistent across metrics, the size and orientation of these transitions varied considerably. The top left map in Figure 5.5 indicates that the total number of terminal cells in a lineage decreases as W and λ increase. In contrast, the top right map indicates that many of the lineages with a large number of terminal cells never cease dividing and hence contain no differentiated cells. The bottom left map (non-deterministic algorithmic complexity) highlights once again the disproportionately high complexity values this metric assigns to very small lineages, indicated by the bright patch that occurs for high values of both W and λ. The weighted complexity metric (bottom right) rectified this anomaly and suggested a link between high complexity and lineage diversity: the most dense concentrations of different lineages and the most complex lineages both co-occur in a region of parameter space between uncontrolled cell proliferation and absolute cell quiescence.

Within a single complexity map generated by a parameterised network, transitions between complexity values tended to be relatively smooth, with occasional large jumps. That is, as W and λ were varied, neighbouring lineages tended to share similar levels of complexity. Occasionally, larger jumps in complexity were observed (*e.g.*, the dominant transition running diagonally from top-left to bottom-right in the maps shown in Figure 5.5). In these cases, increasing the resolution of the map (*i.e.*, decreasing the size of increments for W and λ) frequently (but not always) resulted in the appearance of intermediate lineages.

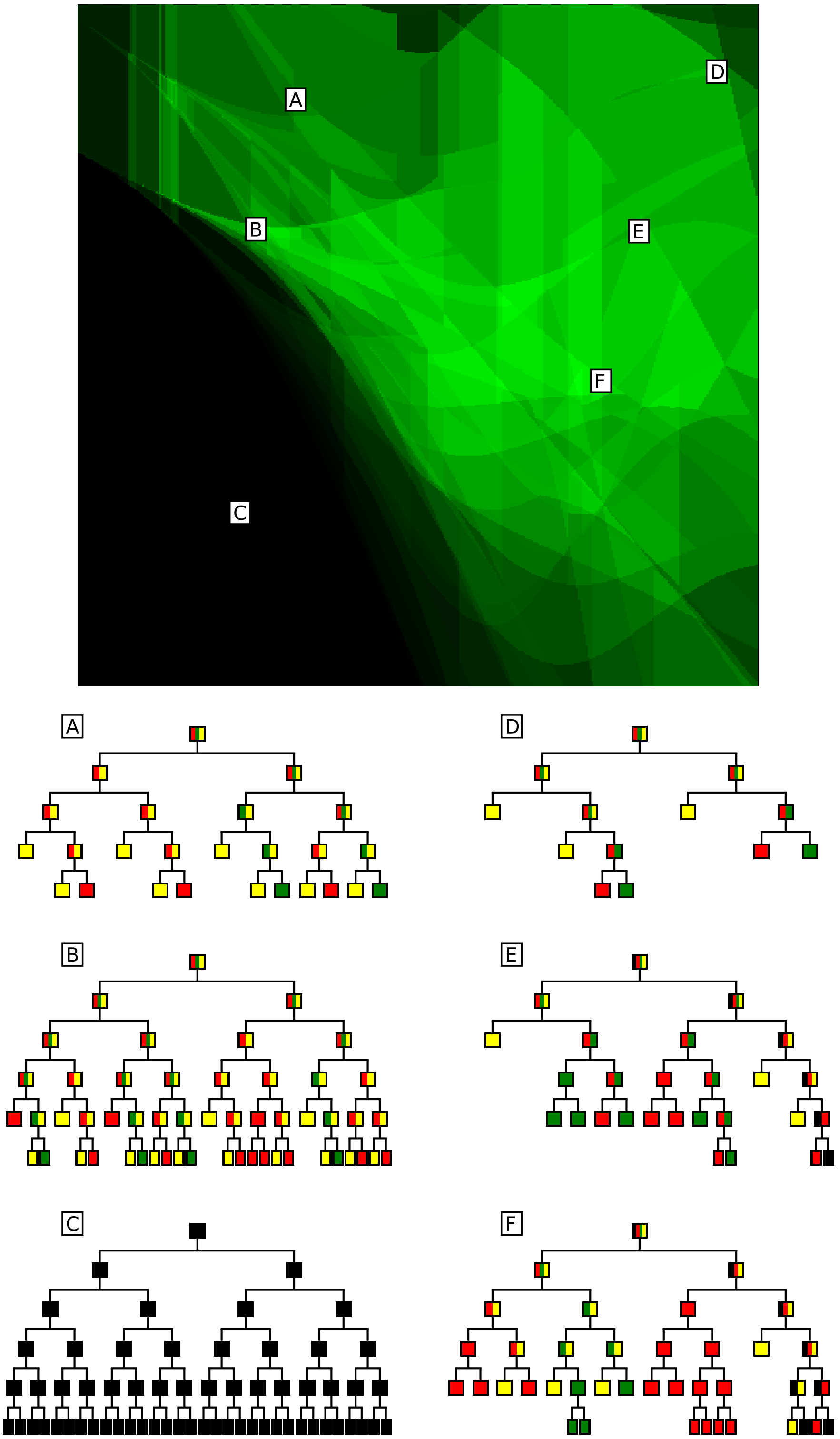


Figure 5.6: Six cell lineages from the region of ontogenetic space shown in Figure 5.5. The heatmap is coloured according to weighted complexity. Further details of the lineages are provided in Table 5.3.

Table 5.3: Details of cell lineages shown in Figure 5.6

Lineage	Weight Scale (W)	Division Scale (λ)	Weighted Complexity
A	1.0	0.92	8.67
B	0.5	0.85	11.42
C	0.4	0.40	0.0
D	3.25	0.97	4.67
E	2.9	0.69	12.0
F	2.3	0.55	15.11

5.2.3 Discussion

The interactive visualisation tool *TreeView* provided an effective means of developing insights into ontogenetic space. In particular, it was discovered that high complexity lineages tend to be clustered into particular regions of parameter space. It has previously been noted that complex or interesting behaviours tend to be boundary phenomena, occurring in the transition from one type of simple behaviour to another (Langton, 1990). The two types of simple behaviour that bracket interesting behaviours in this situation are unchecked proliferation—a cell lineage that continually divides without ever differentiating—and quiescence—a cell lineage in which the initial cell differentiates immediately without ever dividing. In between these two extremes lie a wide variety of more complex structures.

A significant advantage of the methodology was the ability to recursively subdivide the parameter range, which automatically increased the resolution of the maps in regions with a high diversity of lineages. Furthermore, exploration of different network parameters and seeds could be carried out in an efficient fashion by starting with a low resolution analysis and selectively increasing the resolution for interesting network seeds and parameter combinations. The interactive nature of *TreeView* was similarly valuable for rapidly obtaining insight into the effect of modifying various parameters. A further use of *TreeView* that was briefly investigated was to observe the effects of changes to the DRGN-lineage model definition; for example, replacing the exponential scaling of λ with logarithmic or linear scaling (as described in §3.3.2).

The high degree of variability between the ontogenetic spaces generated by different random DRGNs with identical parameters raised the question of what predictions could be made about the properties of a lineage generated by a given

DRGN. Effectively, what is the probability distribution of complexities at a given location in an ontogenetic map? The following study addressed this question by measuring the complexity distributions of cell lineages generated by parameterised DRGN ensembles.

5.3 Study 6: Characteristic properties of network-lineages

The results of *Study 5: Visualisation of ontogenetic space* indicate the range of lineages that DRGNs are capable of producing. In *Study 6: Characteristic properties of network-lineages* we sought to quantify how the distribution of cell lineages, as measured by weighted complexity[2], depends on the properties of the DRGN controller.

Chapter 4 demonstrated how the dynamic behaviour of a DRGN depends on its structural properties. The patterns of division and differentiation events that occur in development are generated by the dynamic behaviour of the underlying control system. Developmental trajectories are a product of DRGN dynamics, therefore these characteristic dynamic behaviours are likely to affect the generated ontogenies.

The primary aim of this study was to measure the distribution of cell lineage complexity and phenotypes generated by DRGNs in order to understand the relationship between the structural properties of DRGNs (size, connectivity and weight scale) and the ontogenies they generate. A further aim was to investigate the extent to which these distributions were attributable to the DRGN control structure, rather than being an artifact of the mechanism of lineage generation. Finally, this study was intended to guide to the choice of suitable parameters for use in future evolutionary simulations.

5.3.1 Ensemble comparisons

One of the strengths of the complexity heatmaps used in the previous study was the high parameter value resolution that could be obtained. The associated limitation

[2] For the remainder of the studies in this and the following chapter, weighted complexity is the sole metric used to measure cell lineage complexity; all references to 'complexity' alone therefore refer to 'weighted complexity'.

Table 5.4: Parameters for ensemble simulations

Parameter	Values
Size (N)	1, 2, 4, 8, 16, 32
Connectivity (K)	1, 2, 4, 8, 16, 32
Weight scale (W)	1.0, 2.0, 4.0, 8.0
Division (λ)	0.6, 0.7, 0.85

was that only a relatively small number of random networks could be compared in this fashion. This study reversed these conditions: much larger ensembles of networks were evaluated, with the compromise that the resulting distributions were obtained for only a selected set of parameter combinations. The values that were explored for each parameter (N, K, W and λ) are shown in Table 5.4. An ensemble of 10,000 DRGNs was created for each combination of these parameters, with the exception that the maximum connectivity for a given network was equal to its size. Therefore a total of 252 ensembles were tested (21 size/connectivity combinations × 4 weight scales × 3 division thresholds). For all ensembles the number of possible cell types was 2.

Ontogenies – complexity distribution

The first statistic recorded for each DRGN in an ensemble was the complexity of the lineage it generated. Given the large number of distributions that resulted (one for each of the 252 ensembles), a selection of representative samples are shown in Figures 5.7, 5.8 and 5.9. Each series of three plots was selected to illustrate the trend in the shape of the complexity distribution as one of the parameters (N, K and W respectively) was varied while the remaining parameters were held fixed.

A notable feature of these distributions is that they are (for the most part) bimodal, suggesting a natural division of generated lineages into two distinct categories. The first category consists of trivial lineages that undergo only one or two divisions and is responsible for the significant complexity ‘spike’ at the lower end of the distribution. The second category, the non-trivial lineages, produces a roughly symmetric distribution around a peak at a higher complexity value (between 7 and 12 in the distributions shown). The effect that each parameter has may therefore be described in terms of how it effects the proportion of lineages in each category and the properties of each of the sub-distributions.

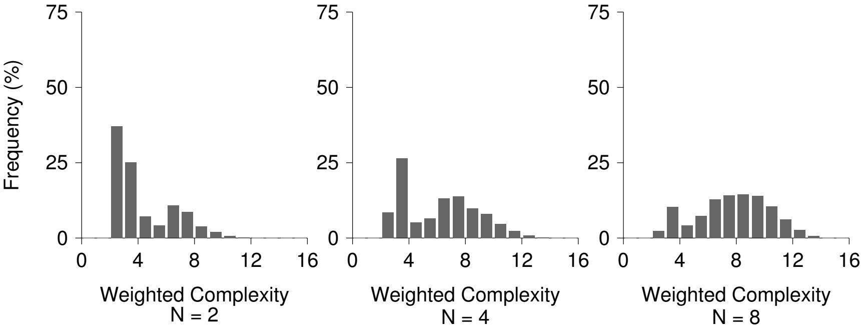


Figure 5.7: Distributions of weighted complexity for varying network sizes ($N = 2, 4, 8; K = \text{full}; W = 2.0; n_f = 2; \lambda = 0.85; 10,000$ samples).

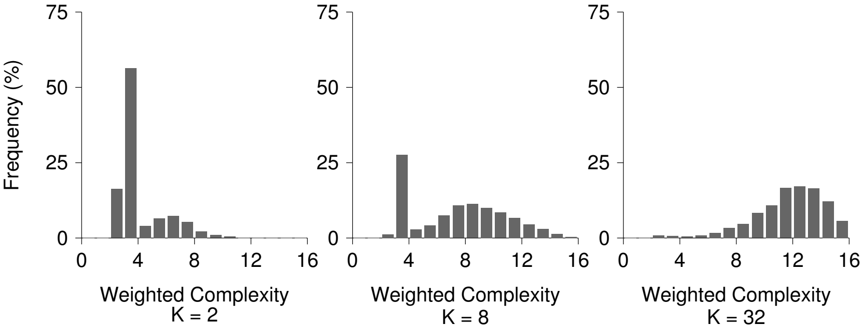


Figure 5.8: Distributions of weighted complexity for varying network connectivities ($N = 32; K = 2, 8, 32; W = 1.0; n_f = 2; \lambda = 0.6; 10,000$ samples).

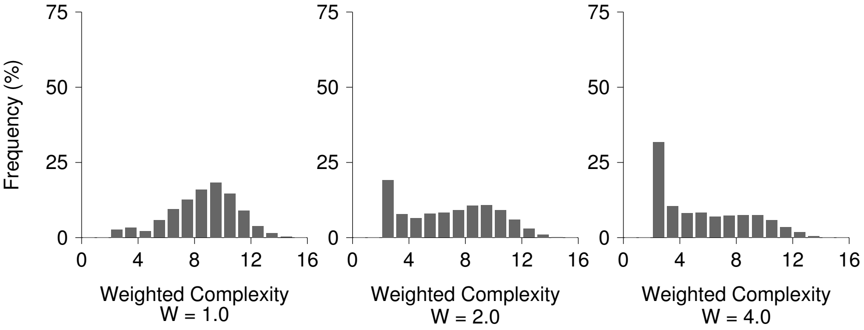


Figure 5.9: Distributions of weighted complexity for varying weight distributions ($N = 32; K = 32; W = 1.0, 2.0, 4.0; n_f = 2; \lambda = 0.85; 10,000$ samples).

As network size (N) increases, the proportion of non-trivial lineages increases, as does the average complexity of non-trivial lineages (Figure 5.7). A similar trend is observed as network connectivity (K) increases (Figure 5.8). In contrast, as network weight scale (W) increases, the proportion of non-trivial lineages and the average complexity of those lineages both decrease; the upper limit of the distribution remains constant however (Figure 5.9).

Phenotypes – cell fate distributions

The second statistic recorded was the number of terminal cells of each type that were generated by each lineage. Each lineage could therefore be mapped to a two-dimensional phenotypic space in which the x and y axes indicated the number of cells of the first and second type respectively (Figure 5.10).

Figure 5.11 shows a selection of heat maps plotted over these phenotypic spaces and illustrates how the patterns of cell fate distributions change with the size and connectivity of the network. As λ varied, the overall size of the cell lineages decreased (as cells tended to differentiate earlier), and so the distributions moved closer to the origin; however, their shape was otherwise similar. Two important features are apparent in these plots. As with the complexity distributions above, the distribution of phenotypes changes as the size and connectivity of their networks are varied. Furthermore, the distribution of phenotypes for any given parameter combination is not uniform: there is a clear structure to the plots. It was expected that no systematic bias in favour of one cell type over another would be evident and this was confirmed by the symmetrical distribution of phenotypes around the diagonal. Lines running perpendicular to the diagonal indicate lineages with an equal number of terminal cells, but different distribution of cell types. Horizontal and vertical lines indicate lineages in which the number of cells of one type is fixed, but that of the other type varies. The strong curved lines ('wings') visible in the plots from the larger networks reflect relationships between cell type numbers introduced by the presence of repeated sublineages within cell lineages. The dense concentration of points in the centre of the plots indicates a strong bias towards lineages containing approximately equal proportions of each of the two cell types.

In summary, the structure of the distribution indicates that not all combinations of cell types are equally likely to be generated by a random network. The existence of isolated points not conforming to these structures suggests some flex-

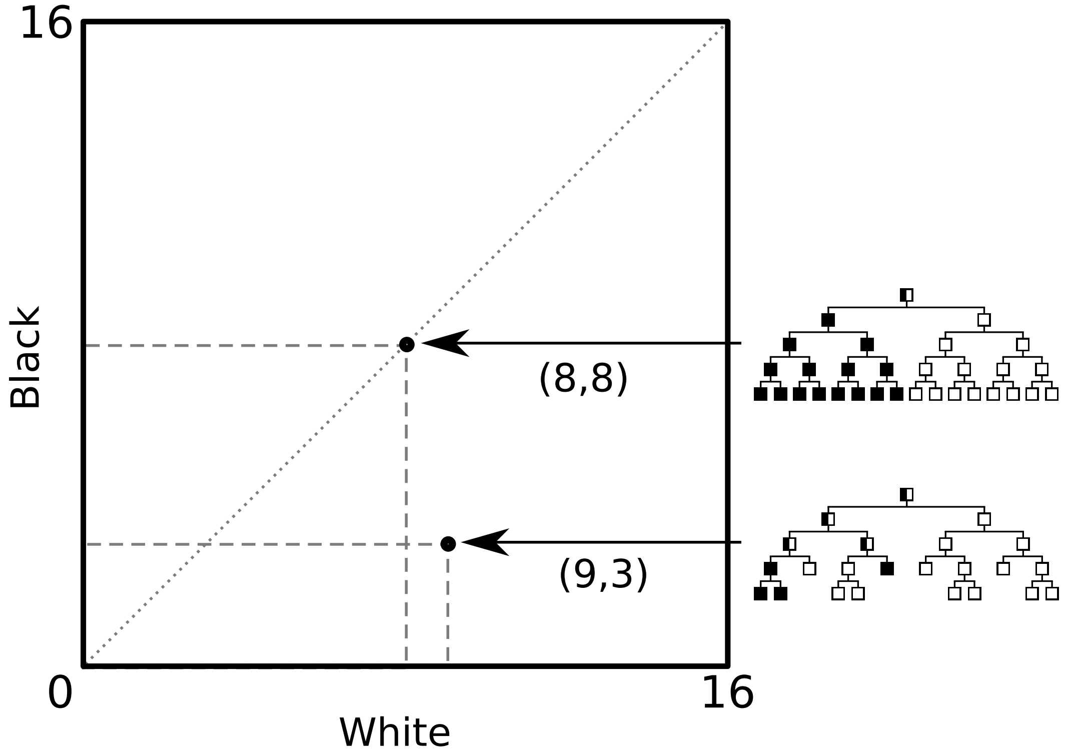


Figure 5.10: Mapping cell lineages into phenotypic space. Each cell lineage can be mapped to a single point in the two-dimensional space according to the distribution of its terminal cells. Here the two cell fates are represented as black and white, and the counts of each are plotted on the x and y axes respectively. Any lineage on the dotted diagonal line (such as the top lineage) will have equal numbers of each cell type.

ibility: cell fate distribution is not subject to strong constraints that cannot be broken, but rather to biases that make certain distributions less likely to appear.

5.3.2 Comparison with stochastic processes

Figures 5.7 through 5.11 indicate that cell lineages generated by DRGNs are distributed in a non-uniform fashion with respect to both their complexity and the composition of their terminal cell types. What is not clear from these results is the extent to which these distributions can be attributed specifically to the fact that development is controlled by the dynamics of a DRGN, rather than being an artifact of aspects of the lineage component of the model, such as the division threshold scaling regime.

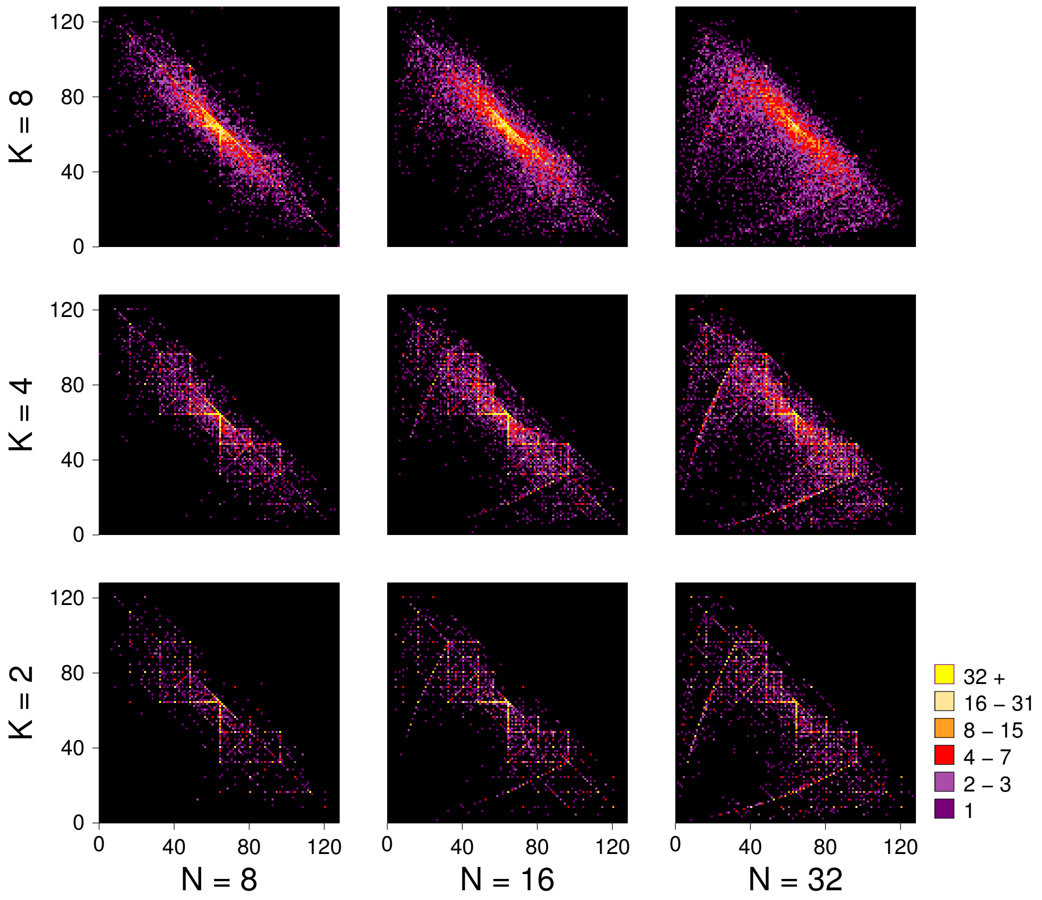


Figure 5.11: Phenotype distributions for varying network sizes and connectivities ($N = 8, 16, 32; K = 2, 4, 8; W = 1.0; n_f = 2; \lambda = 0.6$; 10,000 samples).

In order to investigate the possibility that the observed distributions were a result of the developmental component of the model rather than the DRGN dynamics, two further ensembles were generated: one in which the DRGN controller was replaced with a stochastic process while the manner in which cell division produced lineage structure remains unchanged; and a second in which both the DRGN controller and the cell division decision were replaced by a stochastic process.

Replacing the controller

For the first set of comparison lineages, the decision by each cell whether to divide or not was made randomly rather than on the basis of DRGN output. The probability of a cell dividing was equal to the division threshold for a cell at the same

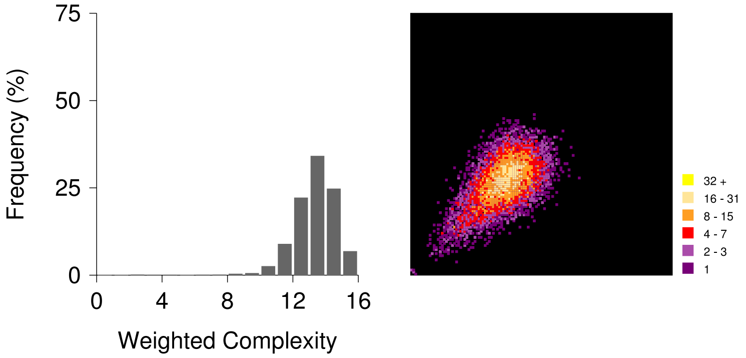


Figure 5.12: Ensemble of lineages generated by a stochastic controller ($\lambda = 0.6; 10,000$ samples).

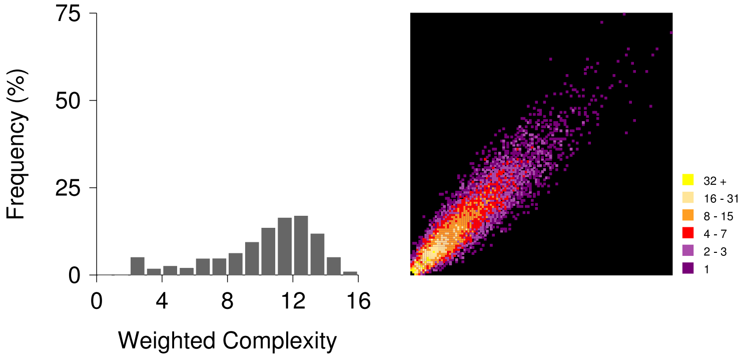


Figure 5.13: Ensemble of lineages generated by a stochastic (Markovian) process, with the size distribution drawn from a previously generated ensemble ($N = 32; K = 8; W = 1.0$) from Figures 5.8 and 5.11 above. (10,000 samples)

level in a DRGN-generated lineage. For example, where a DRGN development process may have had a division threshold sequence of 1.0, 0.8, 0.5, ... at each successive cell division, the lineages developed for this ensemble had a probability of division of 1.0, 0.8, 0.5, ... at each successive division. Once all cells had ceased dividing, cell fates were assigned randomly to the terminal cells. Thus development occurred in an identical fashion to lineages generated from DRGNs, except that each of the division control decisions was now made in a stochastic fashion.

The *complexity distribution* of the first stochastic ensemble was similar in shape to certain of the DRGN ensembles, with the notable feature that all of the lineages appear to fall into the non-trivial category (Figure 5.12). In contrast, the *distribution of phenotypes* was markedly different from any of the DRGN ensembles. Instead of the structured distribution of phenotypes shown in Figure 5.11, the phenotypes generated by the stochastic process generate a dense, relatively uniform cloud of points (Figure 5.12).

Replacing both the controller and the developmental process

The second set of comparison lineages was generated according to a Markovian process as described by Braun et al. (2003). In this model, each cell has an equal chance of dividing, regardless of level. At each step, one of the terminal cells is randomly chosen (with equal probability) to divide. This process is repeated until the lineage contains a given number of terminal cells. Because this process requires the number of terminal cells to be specified prior to generation of the lineage, we used a distribution of lineage sizes drawn from one of the previously generated DRGN ensembles ($N = 32; K = 8; W = 1.0$). Again, cell fates were randomly assigned to each of the terminal cells. Therefore, the second set of comparison lineages differed from the DRGN controlled lineages both due to the nature of the controller (stochastic rather than deterministic) and the developmental process (no longer based on a λ-scaled division threshold).

Comparing the second stochastic ensemble with the equivalent (in terms of size distribution) DRGN ensemble, two differences are apparent (Figure 5.13). Again, the ontogenetic complexity distribution indicates that a lower proportion of trivial lineages are being created, and that the average complexity of the non-trivial lineages is approximately one-third greater. The phenotype distribution indicates a more striking difference: there is a strong bias in the stochastic process towards lineages containing an equal number of each of the two cell types. Given that each terminal cell has an equal probability of being of either type, this bias is not necessarily surprising in its own right. However, at a structural level, the DRGN also imposes no bias on the production of one cell type over another. Therefore the deviation from this equality observed in Figure 5.11 is a property of the DRGN dynamics.

5.3.3 Discussion

The results of *Study 6: Characteristic properties of network-lineages* provide two additional insights into the nature of the ontogenetic space produced by DRGNs. The first concerns the distribution of lineages as measured by complexity. Rather than containing a continuous level of variation between simpler and more complex lineages, ontogenetic space is divided into two distinct classes, described here as trivial and non-trivial lineages. The trivial lineages consist of those lineages that fail to divide or divide only once, and hence remain at the single or two-cell stage; and those lineages that fail to differentiate, and hence proliferate indefinitely. These lineages are represented by a single low-complexity spike in Figures 5.7 to 5.9. The non-trivial lineages consist of all remaining lineages that divide at least twice and in which a significant proportion of the cells differentiate. These lineages are represented by a symmetric distribution around a higher-complexity peak. As the genotypic parameters (size, connectivity and weight scale) are varied, the proportion of ontogenetic space that falls into each of these classes changes.

The second insight concerns the distribution of phenotypes produced by the network-lineage model. Using cell fate composition as a simple proxy for phenotype, the structure of phenotypic space produced by DRGNs shows a clearly non-uniform structure (Figure 5.11). In comparison with the stochastically generated ensembles, the DRGN controlled ontogenies display a lower complexity distribution and a more diverse range of phenotypes.

5.4 Study 7: The robustness of development

The robustness of developmental systems in biology—their ability to produce viable phenotypes in a wide range of environmental conditions—is one of their most remarkable features (Gibson, 2002, Kitano, 2004). Developmental systems may display robustness to two different sources of perturbation (Wagner and Altenberg, 1996). The genotype may remain static while the course of development is subject to noise from the environment, potentially leading to the appearance of novel phenotypes; a phenomenon termed phenotypic plasticity (Debat and David, 2001). Robustness to such plasticity has been termed environmental canalisation (Waddington, 1942, 1959). Alternatively, the genotype itself may be perturbed—for example, by mutation—leading to the appearance of novel phenotypes. Robustness to such mutations has been termed genetic canalisation. It is the variability produced via such genotypic change that can be selected for and inherited.

The aim of *Study 7: The robustness of development* was to measure the robustness and variability of lineages to structural and dynamic perturbation. The following section describes how structural and dynamic perturbations were applied to the network-lineage model and how robustness and variability were measured. The results of ensemble studies are then presented and discussed.

5.4.1 Perturbation analysis

This section describes the methodology that was used to measure the robustness of the DRGN-lineage model to genetic and environmental sources of perturbation. We used 21 ensembles of DRGNs parameterised by size ($N = 1, 2, 4, 8, 16, 32$) and connectivity ($K = 1, 2, 4, 8, 16, 32$). The remaining parameters, weight scale, threshold scaling and number of cell types, were fixed for all ensembles ($W = 2.0; \lambda = 0.8; N_O = 2$). Each ensemble consisted of 200 randomly generated DRGNs. For each DRGN, the unperturbed lineage was generated and stored. Four sets of perturbed lineages were then generated for each DRGN. The four sets varied the source of perturbation (genetic or environmental) and the rate of perturbation (absolute or relative). Each set consisted of 100 perturbed lineages. The robustness for the four sets was calculated by comparing each of the perturbed lineage to the unperturbed lineages. Further implementation details relating to the source of perturbation, the rate of perturbation and the lineage comparison algorithm are described below.

Source of perturbation

A developmental system may experience perturbation from two different sources: genetic and environmental (Wagner and Altenberg, 1996). In the context of the DRGN-lineage model, these can be interpreted as structural and dynamic perturbation respectively. Genetic perturbation, or mutation, is a heritable change to an organism's genome, here represented as the pattern of interactions between nodes in the DRGN. Environmental perturbation by contrast is a non-heritable and transient disturbance affecting development, here represented by the dynamic behaviour of the DRGN. Structural perturbations were implemented by modifying the connection strengths between nodes. Each perturbed DRGN was generated by adding Gaussian noise with distribution $G(0, 0.1)$ to 20% of randomly chosen connections. This implementation of mutation corresponds to the random walk assumption used in population genetics models (Zeng and Cockerham, 1993). Dynamic perturbations were implemented by adding probabilistic noise to a subset of node activations. After each cell division, each node activation had a 10% chance of being modified by the addition of Gaussian noise with distribution $G(0, 0.05)$.

Rates of perturbation

In biology, mutation rates are known to vary widely both among different species and among different regions of a single genome (Kumar and Subramanian, 2002). Similarly, the level of stochasticity in the regulatory events involved in gene expression is not known precisely (McAdams and Arkin, 1997). As a first approximation, we considered the two approaches to determining rates of perturbation.

As the number of parts and interactions in a system varies, there are two possible ways of measuring the amount of perturbation applied to the system. The amount of variation may be absolute, in the sense that a fixed number of perturbations are applied regardless of the structure of the system. Alternatively, the probability of any part or interaction being perturbed may be held fixed, in which case the number of perturbations will be relative to the size or connectivity of the system. In this study, we explored the effect of both absolute and relative rates of perturbation.

Measuring robustness and variability

In order to measure the effect of a perturbation, a metric was required for quantifying the difference between the unperturbed and perturbed cell lineages. Measuring the distance between tree structures is a common task in phylogenetics for which there are a number of widely used methods (Felsenstein, 2004). In general, these methods rely on the terminal nodes of the trees (*i.e.*, the extant species in a phylogenetic tree) being fixed with the variation between trees being in the branching relationship that links the terminal nodes. When considering perturbations to cell lineages however, the set of terminal nodes cannot be assumed to remain constant. Perturbations may result in the terminal cells of a lineage increasing or decreasing in number and also changing from one type to another.

For organisms with invariant patterns of development, the physical location of a cell is often closely tied to its position in the lineage (Sulston et al., 1983, Nishida, 1987, Houthoofd et al., 2003). We therefore decided to base our comparison between lineages on the similarity between the order and composition of their terminal cells[3]. It was important for this measure that, not only could sequences be of dissimilar lengths, but common sub-sequences could be recognised despite being shifted in location. The degree of similarity between two phenotypes was based on the Levenshtein distance (Sankoff and Kruskal, 1983) between the unperturbed fate sequence U and the perturbed fate sequence P. Levenshtein distance is defined in terms of the minimum number of transformations required to change U into P, where possible transformations are the insertion, deletion and substitution of cell fates. The dynamic programming algorithm used to calculate the distance between two sequences U and P was as follows:

```
LevenshteinDistance($U$, $P$):
    declare int $d$[length($U$)+1, length($P$)+1]
    declare int $i, j$
    for $i$ from 0 to length($U$)
        $d[i, 0] = i \times spacePenalty$
    for $j$ from 0 to length($P$)
        $d[j, 0] = j \times spacePenalty$
    for $i$ from 1 to length($U$)
        for $j$ from 1 to length($P$)
            if ($U[i] = P[i]$)
                $currentValue = matchValue$
            else
                $currentValue = mismatchValue$
            $d[i, j] =$ minimum(
                                $d[i-1, j] + 1$,
                                $d[i, j-1] + 1$,
                                $d[i-1, j-1] + currentValue$,
                              )
    return $d$[length($U$), length($P$)]
```

[3] Several alternative approaches to comparing cell lineages are described and investigated in the following chapter.

$matchValue$ and $mismatchValue$ were the scores assigned for a correct or incorrect match at a particular position. *spacePenalty* was the score assigned for an insertion or deletion. The values used for $matchValue, mismatchValue$ and *spacePenalty* were $1, -1$ and -2 respectively. The similarity between two sequences was then defined as:

$$similarity(U, P) = \frac{\textsc{LevenshteinDistance}(U, P)}{|U|} \tag{5.1}$$

where $|U|$ was the length of the unperturbed sequence U. A similarity of 1.0 indicated a perfect match between the perturbed and unperturbed sequence—the phenotype was robust to the perturbation. Any value less than 1.0 indicated an imperfect match—the perturbation resulted in a modified phenotype.

For each ensemble we calculated two statistics: the percentage of perturbations that left the order and composition of terminal cell fates completely unchanged (a similarity of 1.0) and the percentage of perturbations that produced only minor changes to the terminal cell fates (a similarity greater than 0.9). The latter statistic reflects the possibility that small changes to an organism's phenotype may have a negligible effect from the point of view of natural selection (Ohta, 2002).

5.4.2 Ensemble results and discussion

A number of trends in robustness were observed across ensembles as DRGN size and connectivity were varied. When DRGN structure was perturbed in a relative fashion, robustness decreased as either size or connectivity increased (Figure 5.14 (a) and (b)). In contrast, when structure was perturbed in an absolute fashion, robustness increased as size and connectivity increased (Figure 5.14 (c) and (d)). The reversal in the direction of this trend can be explained as follows. As either size or connectivity increased, the total number of interactions also increased, therefore applying a fixed number of perturbations decreased the *relative* rate of perturbation.

When DRGN dynamics were perturbed in a relative fashion, robustness decreased slightly with an increase in size, but more strongly with an increase in connectivity (Figure 5.15 (a) and (b)). The strong response to increased connectivity can be explained by the rate at which a perturbation spreads throughout the network. In a sparsely connected network, perturbing the activation of a single node has an immediate effect only on the few nodes to which the perturbed node is directly connected. In a more densely connected network, the number of nodes directly connected to the perturbed node is correspondingly higher. Hence the effect that the perturbation has on network dynamics is proportionally greater.

When DRGN dynamics were perturbed in an absolute fashion, robustness increased as size increased, but decreased as connectivity decreased (Figure 5.15 (c) and (d)). The increase in complexity as size increased occurs for similar reasons that structural robustness increased in the same scenario. As size increased, the number of genes also increased, and therefore the relative rate of dynamic perturbation decreased. In contrast, increasing the connectivity of a network without changing the size has no effect on the number of genes, therefore there is no difference between absolute and relative rates of perturbation. However, as before, increasing connectivity means that the number of nodes directly connected to a perturbed node is higher. Therefore the effect of a dynamic perturbation is greater and robustness decreases.

These results demonstrate that the developmental cell lineages generated by DRGN dynamics can be robust to perturbations to both network structure and network dynamics. The level of robustness changes in a predictable way as DRGN size and connectivity are varied. Figures 5.14 and 5.15 show the relationship between network parameters and robustness at a general level. One thing these figures fail to show is how robustness may vary across more local regions of ontogenetic space, such as are reachable by a network with fixed size and connectivity.

Within each ensemble, the robustness of an individual DRGN lineage can be calculated on the basis of the 100 perturbed variants that are generated from that particular lineage. These individual values could then be compared with other properties of the lineage such as complexity. Figure 5.16(a) shows one such relationship, indicating that more complex lineages are, on average, less robust than simpler lineages. The number of variant lineages produced by perturbations to more complex lineages is correspondingly greater. It is also possible to compare the robustness to the two genetic and environmental sources of perturbation. Figure 5.16(c) indicates that structural and dynamic robustness are strongly correlated.

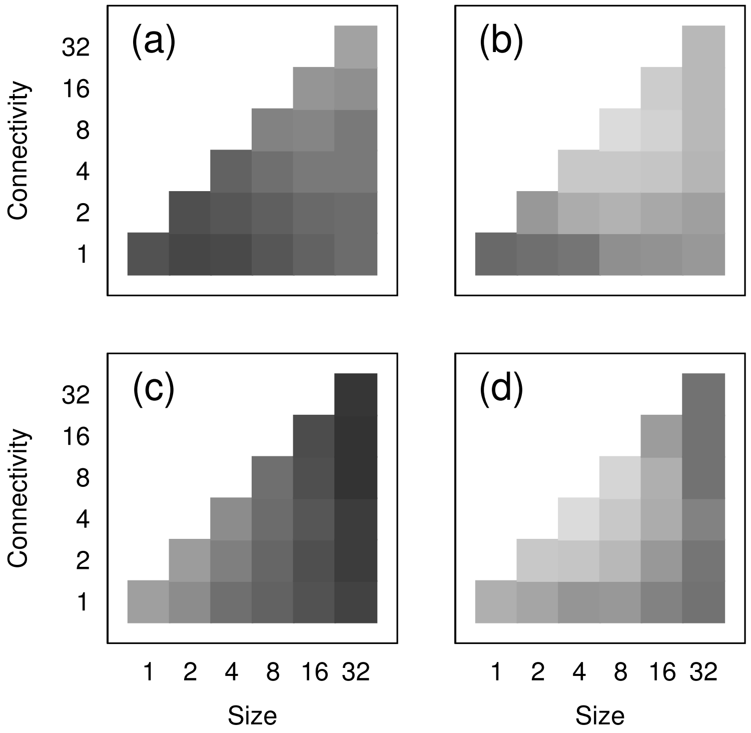


Figure 5.14: Robustness of development to structural perturbation. Perturbations were applied to interactions between nodes in either a relative ((a) and (b)) or an absolute ((c) and (d)) fashion. Robustness was measured on the basis of either 0.9 similarity ((a) and (c)) or perfect similarity ((b) and (d)). Each square represents the level of robustness over an ensemble of 20,000 perturbed lineages with darker squares indicating greater robustness (black = 100%; white = 0%).

5.5 Discussion

The focus of this chapter was the cell lineage component of the DRGN-lineage model. The results of the studies reported in this chapter demonstrate that DRGNs are able to generate a wide variety of complex lineages in a robust fashion.

Study 4: Characteristics of cell lineage complexity (§5.1) formalised several existing notions of morphological and developmental complexity to produce metrics that could be applied to cell lineages. A comparison of these metrics revealed limitations in how well they corresponded with our intuitive notions of which cell lineages constituted a complex control task. A novel metric, weighted complexity, was introduced that addressed the shortcomings of the existing metrics by combining measurements of size and organisational heterogeneity.

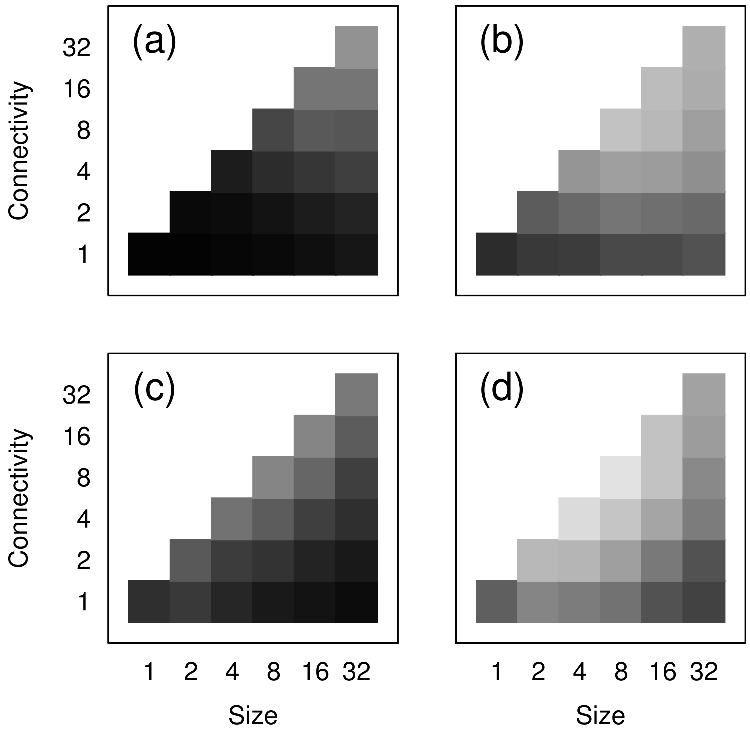


Figure 5.15: Robustness of development to dynamic perturbation. Perturbations were applied to nodes in either a relative ((a) and (b)) or an absolute ((c) and (d)) fashion. Robustness was measured on the basis of either 0.9 similarity ((a) and (c)) or perfect similarity ((b) and (d)). Each square represents the level of robustness over an ensemble of 20,000 perturbed lineages with darker squares indicating greater robustness (black = 100%; white = 0%).

Study 5: Visualisation of ontogenetic space (§5.2) addressed the need for a software tool to visualise ontogenetic space. *TreeView* employed an interactive interface to enable large numbers of cell lineages to be rapidly explored in an intuitive fashion. This study also constituted a first evaluation of the DRGN model's ability to generate interesting developmental patterns, revealing that even very small networks (four regulatory nodes) are capable of generating a range of interesting developmental patterns.

The complexity heatmaps used for navigation in *TreeView* also provide insights into the complexity gradients that characterise ontogenetic space. Complexity, regardless of which metric is used, is distributed in a non-uniform manner throughout ontogenetic space. While intermediate values of both W and λ consistently produced the greatest density of high-complexity lineages, the exact locations of these

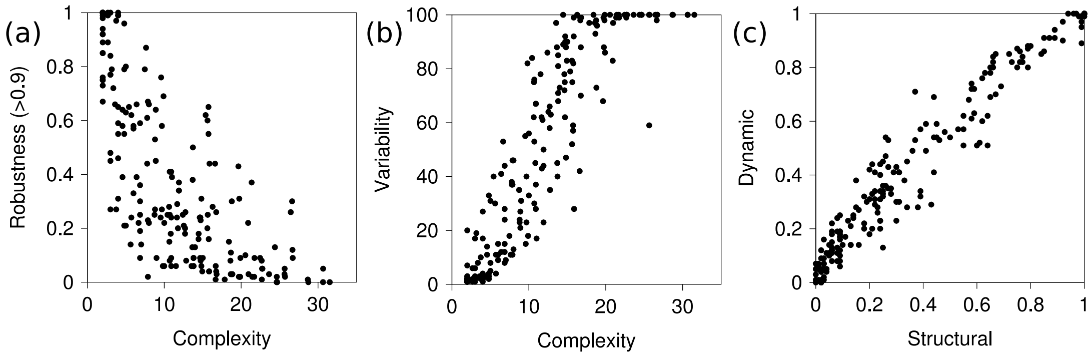


Figure 5.16: Three perspectives of robustness of an individual ontogeny: (a) shows the relationship between complexity and structural robustness; (b) shows the relationship between complexity and variability; and (c) shows the relationship between structural and dynamic robustness.

regions varied among networks. One consistent feature was a phase transition between trivial lineages (zero or very low complexity, occurring when both W and λ are low) and nontrivial lineages (here defined as complexity greater than 4.0). Figure 5.17 illustrates how complexity tends to vary across this transition. In general, the most complex lineages are located in a boundary region whose location is defined by both W and λ. The lineages to the left of this region fail to differentiate, while the lineages to the right of this region fail to divide. The size of this boundary region was observed to vary substantially between different network seeds. In some cases, the region was virtually non-existent, and small increases in W and/or λ transformed a proliferating lineage into a quiescent lineage without an intervening complex regime.

As with the dynamic behaviours of DRGNs observed in the previous chapter, there was a high level of variability between the lineages produced by networks of identical size and connectivity. To quantify how the distribution of lineage complexity in an ontogenetic space varied as a function of the size and connectivity of the generating network, *Study 6: Characteristic properties of network-lineages* (§5.3) used ensemble studies similar to those used to quantify network dynamics in Chapter 4. The distinction between trivial and nontrivial lineages was clearly reflected in the bimodal complexity distributions observed in many of the ensembles (*e.g.*, Figure 5.7). The major effect of the genotypic parameters, N, K and W, was to change the proportion of lineages that fall into each of these two classes. The

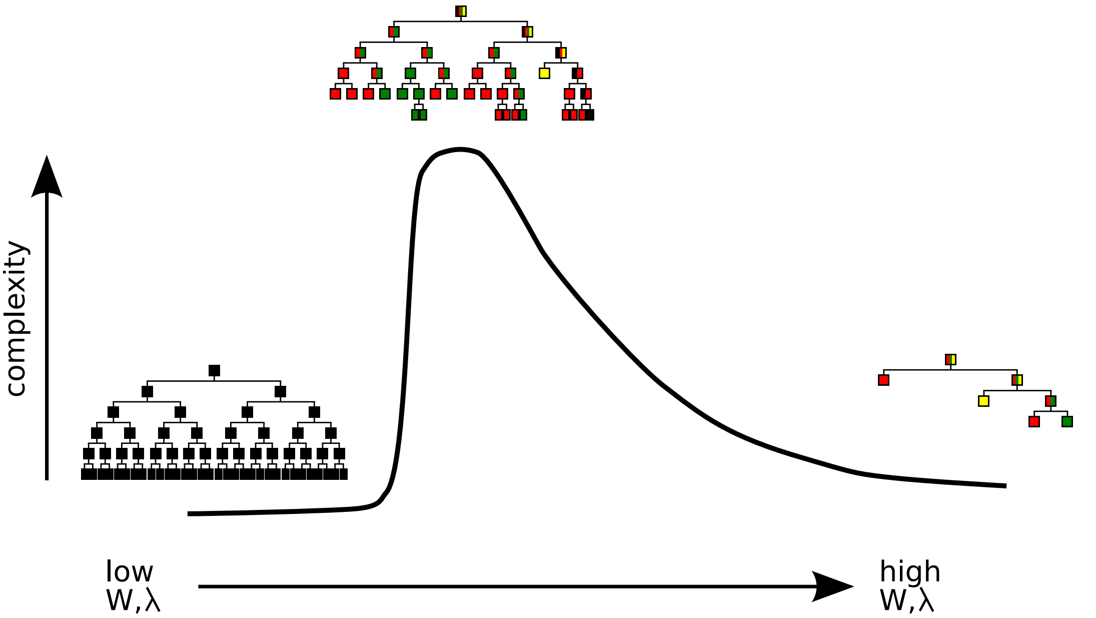


Figure 5.17: A cartoon view of the complexity phase transition in ontogenetic space. At low values of W and λ, cells proliferate indefinitely and never differentiate. At high values of W and λ, cells divide at most one or two times, if at all. The most complex lineages are found between these two extremes, although the exact location varies among networks.

proportion of nontrivial lineages increased as N and K increased, but decreased as W increased. The mean complexity of nontrivial lineages followed a similar trend.

Considering a simple quantification of phenotypes—the ratio between two cell types—over these ensembles revealed a highly non-uniform distribution in phenotypic space. Certain combinations of cell numbers are exponentially more likely to be produced than others. Furthermore, these distributions are not clustered in clouds, but display a high level of structure (Figure 5.11). This structure suggests that, depending on genotypic parameters, networks are strongly constrained in the patterns of ontogeny and cell fate distribution that they can generate. In comparison with stochastically generated lineages, the most notable feature of lineages generated by DRGNs was this quasi-systematic structure.

Study 7: The robustness of development (§5.4) focused on the local structure of ontogenetic space as indicated by the response of the DRGN-lineage system to structural and dynamic perturbation. The primary observations were that robustness to structural perturbations tended to decrease as the relative proportion of perturbed connections increased, and robustness to dynamic perturbations tended to decrease as the probability that those perturbations would be propagated through the network increased.

Considering individual lineages drawn from an ensemble, robustness tended to decrease as complexity increased. This relationship has a bearing on an observation by Houthoofd et al. (2003) that the lineages of two related nematode species, *Halicephalobus* sp. and *C. elegans* differ with respect to the concentration of monoclonal (simpler) and polyclonal (more complex) cell fate distributions. Houthoofd et al. (2003) suggest that evolutionary pressure for increased speed of development could contribute to the transition from monoclonal to polyclonal lineages. Furthermore, they hypothesise that a disadvantage of polyclonal specification may be the greater complexity of the specification mechanism required, and the corresponding increased likelihood of developmental errors. The results of this study support their hypothesis that the networks controlling the development of more complex cell lineages are less robust to errors that occur both during development and in the control mechanism itself.

Comparing individual lineages revealed that those lineages that were robust to dynamic, or environmental, perturbations also tended to be robust to structural, or genetic, perturbations. This correlation supports the contention by Meiklejohn and Hartl (2002) that a single mode of canalisation is responsible for robustness to both genetic and environmental sources of perturbation. These results also corroborate the findings of Wagner and Altenberg (1996) and Siegal and Bergman (2002), who observed that increasing genetic canalisation was a side-effect of selection for robustness to environmental noise.

In summary, the studies reported in this chapter demonstrate the following:

- Lineages can be classified and compared using a variety of complexity metrics; we have defined a measure weighted algorithmic complexity that conforms to our *a priori* notions of what constitutes a complex pattern of behaviour.

- Interactive visualisation tools are a useful means of exploring and developing insight into the structure of complicated parameter spaces.

- The lineages generated by DRGNs cover a broad range of complexities and patterns, though in a highly non-uniform fashion. Depending on the values of genotypic and ontogenetic parameters, certain lineage structures and phenotypic patterns are more likely to be generated than others.

- The DRGN-lineages can be robust to both structural and dynamic sources of perturbation. These two types of robustness tend to be correlated, such that networks that are robust to dynamic perturbation will also be more robust to structural perturbation and vice versa. Furthermore, robustness to structural perturbation (mutation) tends to decrease as lineages become more complex. This relationship has important implications for the adaptive exploration of ontogenetic systems that are explored further in the following chapter.

The claim of this thesis is that development introduces biases into the structure of evolutionary variation. The results presented in this chapter indicate that the space of possible ontogenies, from which novel variation will be drawn, is structured by the dynamic properties of the underlying genetic control systems. The following chapter expands the scope of investigation to consider the implications of this structure in an adaptive context.

Chapter 6

Adaptive Search in Ontogenetic Space

Novel phenotypic forms arise from changes to existing ontogenies. In order for these changes to be preserved in future generations, they must be reflected in changes to the heritable component of developmental control—that is, the gene regulatory network. Arthur (2000) has suggested the term *developmental reprogramming* for changes to ontogeny caused by mutations to developmental genes. Chapter 5 focused on the global properties of the ontogenetic space in which this reprogramming occurs. In this chapter, we return to the questions motivating this thesis. Do properties of the developmental mapping make some types of reprogramming more likely to occur than others? What is the effect of this bias on the direction of evolution? Addressing these questions requires us to focus on the local properties of ontogenetic space: the distribution of ontogenies that are mutationally accessible from any given point.

The studies reported in the previous chapter indicated that both an ontogenetic property (complexity) and a phenotypic property (cell fate distribution) are distributed in a non-uniform fashion with respect to the space of possible DRGNs. The primary aim of this chapter is to explore the implications of this non-uniformity for the types of ontogeny found in an adaptive context. We therefore required a model of the adaptive process. Researchers into evolutionary models of computation typically focus on three aspects of an adaptive process: mutation operators, fitness functions and selection methods (Mitchell, 1996).

In both biological and artificial evolution, mutation operators are, along with recombination, a primary source of innovation. In an evolutionary algorithm, the choice of mutation operator will affect the diversity of novel solutions that are generated and therefore have an impact on the efficacy of the search process. The adaptive task used in this chapter involved searching for DRGNs able to generate lineages based on the cell lineages of real organisms. Choosing an appropriate mutation operator was therefore of pragmatic importance.

The definition of mutation operators is important for an another reason: in addition to developmental mappings, they are a further possible source of bias on the structure of variation. The existence and influence of mutation biases has been explored in the context of both biological (Yampolsky and Stoltzfus, 2001) and artificial (Bullock, 1999, 2001) evolution. The distinction between mutation bias and developmental bias may be conceived as follows. The study reported in §5.3 indicated that a random sample of networks in genotypic space does not map to a uniformly distributed set of cell fate distributions in phenotypic space. Developmental bias affects the distribution of phenotypes such that some cell fate distributions are more likely to be generated than others. Even if mutations were to occur in the uniformly random fashion assumed by early theorists (Fisher, 1930), the structure of phenotypic variation would be non-uniform (Figure 6.1 (b)). Alternatively, if there was no developmental bias, and a uniform sample of networks did generate a similarly uniform sample of phenotypes, bias in the discovery of novel networks due to properties of the mutation operator could also produce a non-uniform distribution of phenotypic variation (Figure 6.1 (c)).

A fitness function is a mapping that assigns a numerical value to a candidate solution rating how 'good' it is. The choice of fitness function defines the adaptive gradient of the search space. Choosing an inappropriate fitness function can make a phenotypic landscape very difficult to search. As a trivial example, a fitness function that assigns maximal fitness to the target solution and zero to all other solutions will be very difficult to search. An adaptive process searching this space receives no information on the location of the target solution until it has actually found it.

Together, a mutation operator and a fitness function define the structure of an adaptive space: the fitness function provides the 'height' of each possible solution; and the mutation operator describes the neighbourhood relationship between each of these solutions. The final component of an evolutionary model concerns the

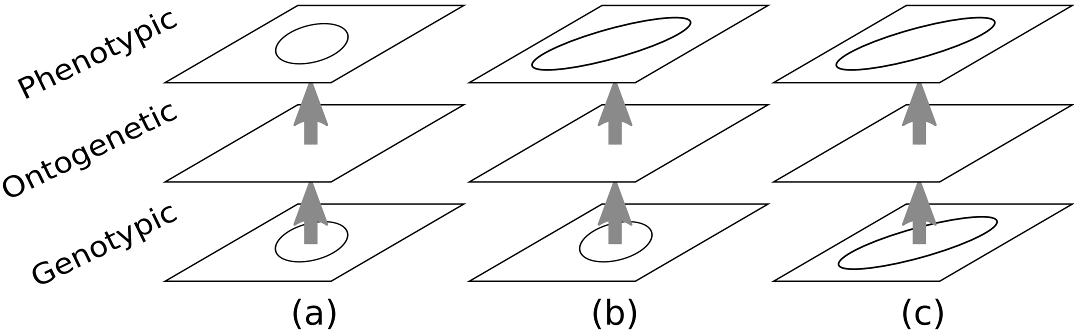


Figure 6.1: The structure of genotypic and phenotypic variation when there is (a) no bias: variation at both levels is uniformly distributed; (b) developmental bias only: uniform genotypic variation is transformed into non-uniform phenotypic variation; and (c) mutation bias: non-uniform variation at the genotypic level is transformed into non-uniform phenotypic variation. It is also possible that both mutation bias *and* developmental bias may occur concurrently.

dynamics of the adaptive process: the selection procedure that chooses which solutions are used to create the next generation. In population genetics, selection is modelled as a coefficient determining the degree to which one allele is favoured over another variant at a particular locus. The strength of this selection pressure will, in combination with the likelihood of mutation generating that variant, determine the equilibrium frequency of that mutant. Evolutionary computation uses a variety of different mechanisms for modelling the selection of individuals. In the simplest variant, the likelihood of each individual contributing to the next generation is proportional to its fitness (Mitchell, 1996). More complicated variants include the use of tournaments between subsets of the population and emphasising the relative rank, rather than the absolute value, of individual fitnesses. In both analytical and empirical approaches to investigating the relationship between mutation and selection, the size of the population is an important factor (van Nimwegen et al., 1997, Hartl, 2000)

Throughout this chapter, adaptive walks have been used as a model for evolutionary processes (Kauffman and Levin, 1987). Adaptive walks are computationally efficient and capable of providing information on the structure of landscapes[1].

[1] 'Adaptive walks' are used as a general term to describe the type of search process used in this chapter. The approach used is very similar to several other algorithms, including the (1+1) EA (Bäck et al., 1997) and the hill-climbing algorithm, which itself exists in a wide range of forms.

Population based approaches, while possibly a more faithful representation of biological evolution, have two disadvantages in the context of this thesis. They can require a significantly greater amount of computation, reducing the range and number of simulations that can be run. Furthermore, they can confound two issues: (a) the relationship between the phenotype distribution that constitutes the adaptive space and the mutation operators that give that space its structure; and (b) the relationship between the mutation operators and selective mechanism. While imposing some limitations on generality (discussed in §7.3), the use of adaptive walks rather than populations provides a foundation of understanding on which more complex models can be built.

In summary, the studies reported in this chapter had three aims. The first was to determine how different mutation operators bias the adaptive exploration of ontogenetic space. In *Study 8: Evaluation of mutation bias* (§6.1) eight variant mutation operators were defined and random walks were used to compare their effect on genetic and ontogenetic properties. The second aim was to investigate the extent to which an adaptive process can locate DRGNs able to generate specific phenotypic cell fate distributions. *Study 9: The evolution of ontogenies* (§6.2) consisted of three series of adaptive walks. Series A and B compared the effects of different mutation operators and fitness functions on the evolvability of the system. Series C compared the performance of the adaptive walks on a succession of increasingly complex phenotypic targets based on the cell lineages of *C. elegans* and *H. roretzi.* The final aim was to analyse the results of these studies for evidence that the direction of evolution was biased by the properties of the developmental process (§6.2.5).

6.1 Study 8: Evaluation of mutation bias

Mutation operators are the generators of change in an evolutionary adaptive process. They define the set of possible steps that may be taken at any given point and so structure the way in which an adaptive walk can explore genotypic space. As described above, the way in which mutation operators are implemented has the potential to bias the direction of evolution (Bullock, 2001).

Two scenarios in which mutation bias may play a role are genetic drift and neutral evolution. Genetic drift refers to changes in gene frequency that occur as a result of chance rather than selection. In any real population, selection occurs in a stochastic fashion leading to fluctuations in the genetic composition of the population over time. The neutral theory of evolution proposes that, rather than ‘climbing hills’ in fitness landscapes, evolving populations actually spend much of their time diffusing through networks of mutationally adjacent genotypes that each produce phenotypes of equal (or nearly equal) selective fitness (Kimura, 1983). Both genetic drift and the presence of neutrality will result in the properties of a mutation operator having a significant effect on the rate and extent of exploration of an adaptive space.

The aim of *Study 8: Evaluation of mutation bias* was to characterise and compare the bias resulting from several different mutation operators. In particular, the effect of mutation operators on the distribution of genotypes (in terms of weight distribution) and ontogenies (in terms of complexity) was measured. In addition, *Study 9: The evolution of ontogenies* (§6.2) used adaptive walks to locate DRGNs capable of generating specific phenotypes. In order to perform these simulations in an effective fashion it was necessary to understand how well various mutation operators were able to explore phenotypic space. We therefore also compared the rate at which novel phenotypes were located by each of the mutation operators.

6.1.1 The mutation operators

Biological mutations affect the nucleotide sequences that make up the genome. They can be classified as either point mutations, in which a single nucleotide is substituted, or sequence mutations, in which a contiguous stretch of nucleotides is duplicated, deleted, transposed or inverted (Alberts et al., 1994). Depending on where in the genome mutations occur, they can cause a variety of changes to the structure of a gene network: new genes may be created, existing genes may be destroyed, and the patterns of interaction between genes may be altered (Watson et al., 2004). How these different effects are implemented in a simulation will depend on the type of genotypic representation used. When the genotype is modelled as a sequence, as in the various ‘artificial genome’ models (Reil, 1999, Geard and Wiles, 2003, Watson et al., 2004, Quayle and Bullock, 2005), it is possible to map biological mutation operators directly onto the simulation model. When the genotype is modelled as a network, mutation operators must be implemented in terms of their effect on the structure of that network.

The type of network mutation operators depends on aspects of the model definition, in particular whether the number of genes in a network is fixed or variable. If the number of genes is fixed, then mutation operators act solely on the pattern of interaction between a given set of genes. If the number of genes can vary, mutation operators may also result in the new genes being created, or existing genes being destroyed. All of the simulations reported in this chapter used networks containing a fixed number of genes.

For this study, eight different mutation operators were defined, consisting of two base operators that could optionally be combined with either, or both, of two modifiers. The two base mutation operators were additive noise and weight replacement (abbreviated as 'Noise' and 'Replace' respectively in tables and figures)[2]. *Additive noise* involved an incremental modification to a subset of network weights: each weight in the network was modified, with probability μ_p, by the addition of a value drawn from the Gaussian distribution $G(0, \mu_s)$. *Weight replacement* involved the replacement of one weight with a randomly chosen value that did not depend on the previous value of the weight: a single network interaction was chosen at random and its weight replaced by a value drawn from the distribution $G(0, W)$ (*i.e.*, the same distribution from which the initial network weights were drawn).

The two different post-mutation modifiers were investigated: normalisation and threshold adjustment. *Normalisation* involved adjusting all weights in a mutated network such that it was located on the surface of the same hypersphere in weight space as the original network (*i.e.*, with radius W). This modifier was motivated by a technique used with neural networks to prevent their weights from growing infinitely large (Haykin, 1999). *Threshold adjustment* involved adjusting the threshold of each node such that the ratio between its inputs and its threshold remained at a constant value (usually 0.5). This modifier was motivated by the analysis presented in §4.4.3, which suggested that networks tended to display a more diverse range of dynamic behaviours when their activation functions were located so as to be sensitive to the expected range of inputs.

Each of these modifiers could be applied separately or in combination. When applied in combination, network weights were first normalised and the threshold

[2] These correspond, respectively, to the random walk and house of cards models described by Zeng and Cockerham (1993).

Table 6.1: Robustness Under Different Mutation Conditions

Base Operator	θ Adjusted	Normalised	Robustness (%)
Additive Noise	No	No	32.53
	No	Yes	32.57
	Yes	No	42.25
	Yes	Yes	41.67
Weight Replacement	No	No	16.73
	No	Yes	16.42
	Yes	No	21.76
	Yes	Yes	21.08

was then adjusted. Therefore, eight different mutation operators were possible: two base operators, each with four combinations of modifiers.

6.1.2 Initial comparison: Perturbation analysis

Perturbation analysis was used to provide an initial indication of how disruptive each of the mutation operators was. The methodology used was identical to that used to measure robustness to structural perturbation in §5.4 except that the eight mutation operators described above were used to create the perturbed ensembles. To recap, for each mutation operator, an ensemble of 20,000 individuals was created by applying 100 independent mutations to 200 randomly chosen networks ($N = 8; K = 8; W = 2.0$). For the four additive noise ensembles, $\mu_p = 0.2$ and $\mu_s = 0.1$ (*i.e.*, each mutant was created by modifying 20% of its weights by the addition of Gaussian noise with distribution $G(0, 0.1)$). Cell lineages were generated from each network and compared using the sequence comparison metric described in §5.4. As before, robustness was measured as the percentage of ensemble members for which the mutant lineage and the original lineage had a similarity greater than 0.9.

The robustness of development to perturbation by the eight mutation operators is reported in Table 6.1. Initially it appears that additive noise is a less disruptive base mutation operator (development was more robust to its effects). However, direct comparisons between the two base mutation operators are precluded since each is parameterised differently. By either increasing the proportion of weights modified by noise (μ_p) or the size of the noise (μ_s), it is possible to decrease the robustness of development to the additive noise mutation operators.

What *can* be compared is the effect of the post-mutation modifiers. Considering the two base mutation operators independently, each combination of post-mutation modifiers had comparable effects. Adjusting the threshold after applying the mutation increased robustness by approximately one third in either case; for both base operators, normalisation had a negligible effect.

The following section extends the scope of investigation to focus on the effect of a sequence of mutations, modelled as a random walk.

6.1.3 Random walks

One way of investigating whether a particular mutation operator exhibits a bias is to consider the effect of repeated application of the relevant operators and modifiers in the absence of any selection pressure (Bullock, 2001). In order to investigate the effect of mutation biases in the absence of any directional selection we considered random walks on a flat fitness landscape (one in which each phenotype was associated with an identical fitness value). For each of the eight mutation operators, we used an ensemble of 50 networks ($N = 8, K = 8, W = 2.0$); each network was used as the starting point for a random walk of 20,000 steps; at each step, a new network was created from the old network by application of one of the mutation operators described above. Selection was not stochastic: the newly created network *always* replaced the original network. The distribution of network weights, ontogenetic complexity and the final phenotype were recorded at each step.

Bias at the genotypic level

The level of mutation bias at the genotypic level was analysed by considering the effect of each of the different mutation operators on the distribution of each of the network weights over each of the 50 members of the ensemble.

The additive noise operator, applied without any post-mutation modifiers, is clearly biased with respect to its effect on weight distribution (Figure 6.2 (a)). After the 20,000 steps, the standard deviation of network weights grew from 1.98 to 6.65. A similar change occurred when the threshold adjustment modifier was used (Figure 6.2 (c)). While the standard deviation of weights did not change significantly when the normalisation modifier was used, the shape of the distribution changed: the proportion of weights with values very close to zero increased by around one half and the minimum and maximum values of the distribution increased from $[-7.4, 6.7]$ to $[-16.2, 15.2]$ (Figure 6.2 (b)). When both modifiers were applied the weight distribution remained unchanged, suggesting any bias that exists at the genotypic level supports the initial distribution (Figure 6.2 (d)).

In contrast, all four of the weight replacement mutation operators displayed very little change in weight distribution (Figure 6.3). With respect to genotypic variation, weight replacement results in less bias than additive noise.

Bias at the ontogenetic level

Mutation bias at the ontogenetic level was analysed by comparing the distribution of ontogenetic complexity over the 50 members of the ensemble before and after undergoing random walks.

The first point to note is that, unlike the weight distributions considered above, which were equivalent for all initial ensembles, there is a difference between the complexity distributions *before* the random walk occurs depending on whether or not the networks have had their thresholds adjusted. In simulations *without* the threshold adjustment modifier, the additive noise mutation operator resulted in an increase in the proportion of lineages with zero complexity (*i.e.*, those in which the first cell failed to divide), regardless of whether or not the normalisation modifier was used (Figure 6.4 (a) and (b)). In the simulation *with* the threshold adjustment modifier only, the additive noise mutation operator again resulted in significant shift towards less complex lineages (Figure 6.4 (c)). The use of both post-mutation modifiers resulted in the original complexity distribution being preserved over the duration of the random walks, indicating an absence of bias due to mutation.

As was the case with the weight distribution, the weight replacement mutation operator did not significantly affect the complexity distribution over the course of the random walks regardless of which combination of post-mutation modifiers was used.

It appears that one way in which mutation can bias evolution is by changing the shape of the genotypic weight distribution. The gradual addition of random noise, in the absence of selection, alters the shape of a weight distribution, with a corresponding reduction in the complexity of generated lineages.

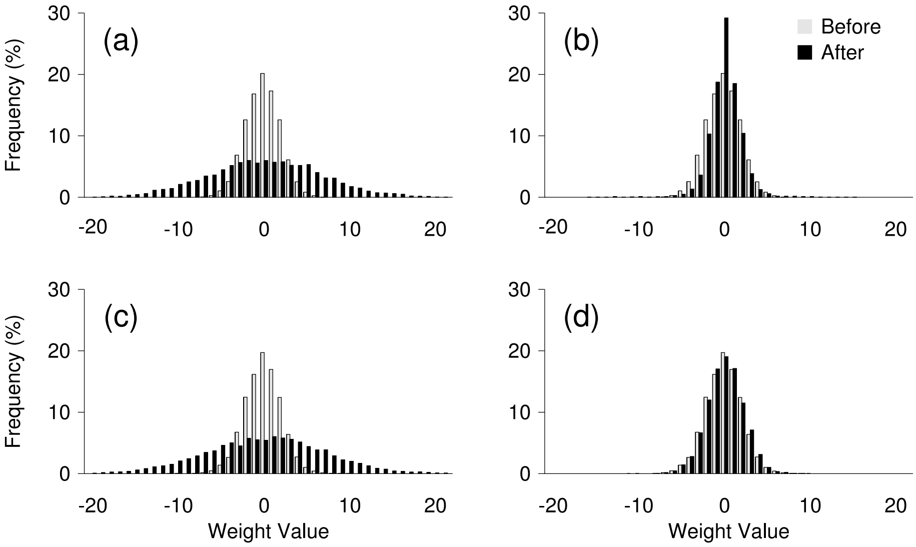


Figure 6.2: Weight distribution over the 50 networks before (light grey) and after (black) undergoing a random walk using noise mutation operators with (a) no modifiers, (b) normalisation, (c) threshold adjustment and (d) both normalisation and threshold adjustment.

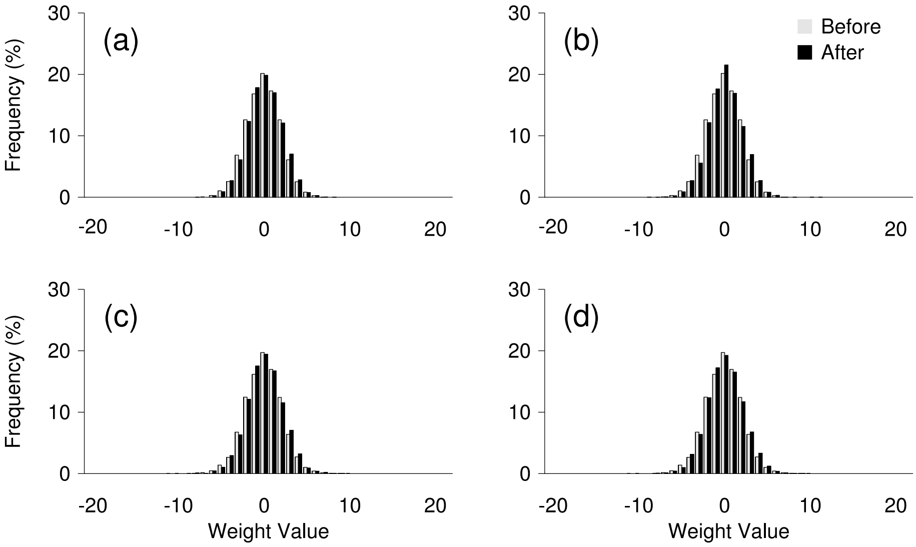


Figure 6.3: Weight distribution over the 50 networks before (light grey) and after (black) undergoing a random walk using weight replacement mutation operators with (a) no modifiers, (b) normalisation, (c) threshold adjustment and (d) both normalisation and threshold adjustment.

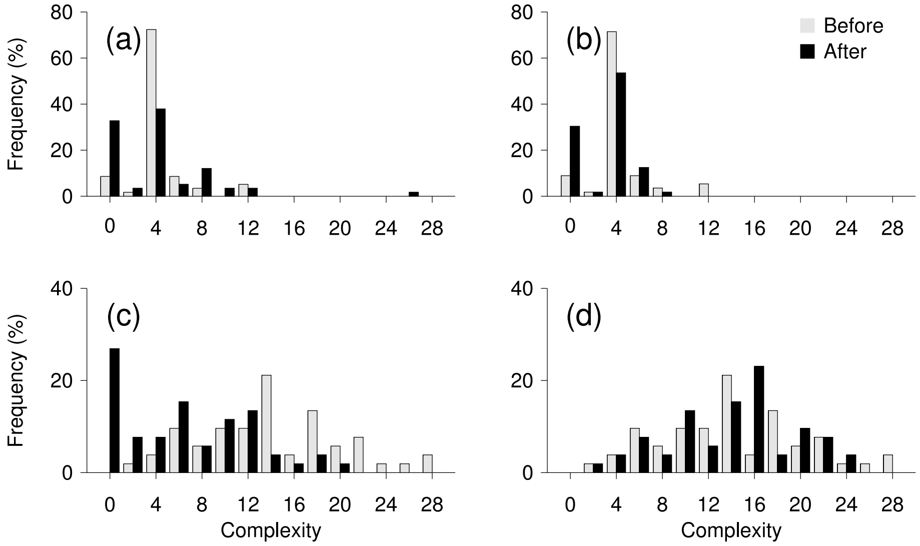


Figure 6.4: Complexity distribution over the lineages generated by the 50 networks before (light grey) and after (black) undergoing a random walk using the noise mutation operator with (a) no modifier, (b) normalisation, (c) threshold adjustment and (d) both normalisation and threshold adjustment.

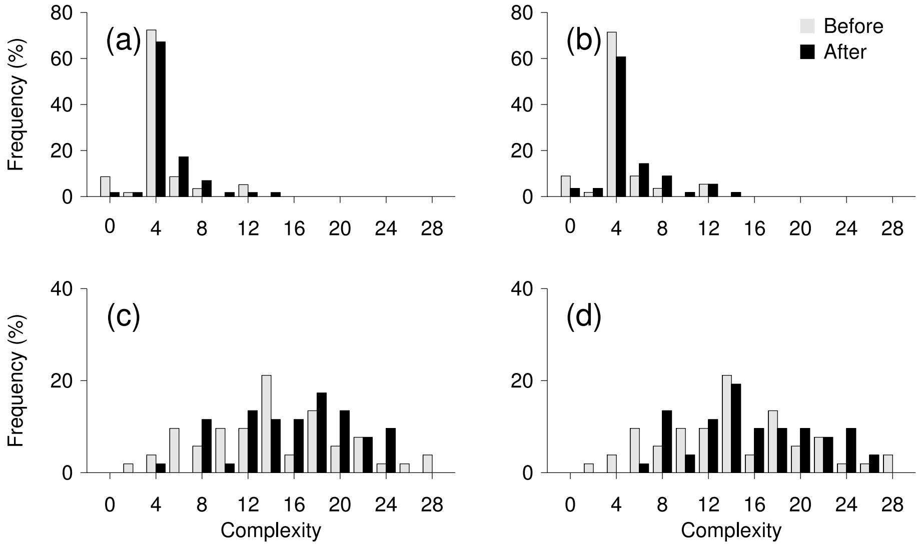


Figure 6.5: Complexity distribution over the lineages generated by the 50 networks before (light grey) and after (black) undergoing a random walk using the replacement mutation operator with (a) no modifier, (b) normalisation, (c) threshold adjustment and (d) both normalisation and threshold adjustment.

Rate of phenotypic discovery

Figures 6.2–6.5 provide snapshots of how mutation bias affects the distribution of genotypic and ontogenetic properties at the beginning and end of random walks. However, they do not contain any information about the regions of space that were explored during the random walks. A mutation operator that is effective for adaptive search must generate a sufficient amount of phenotypic variation for selection to act on. In order to measure the range of phenotypic variation generated, we kept track of how many unique phenotypes were discovered during each random walk.

The number of unique phenotypes discovered during a random walk was greatest for the weight replacement operator with either threshold adjustment or both modifiers (Figure 6.6). When both modifiers were used, an average of 15,487 unique phenotypes were found over 20,000 steps—almost four out of every five mutations resulted in the generation of a novel cell fate distribution. Additive noise with both modifiers also resulted in a high number of unique phenotypes being discovered (10,663). All of the other mutation conditions resulted in the discovery of substantially fewer unique phenotypes. Comparing the average number of unique phenotypes discovered with the average complexity of the lineages visited on a random walk reveals a strong relationship between level of variation and complexity (Figures 6.6 and 6.7 respectively).

To further explore the relationship between complexity and rate of phenotypic discovery, we examined several individual random walks in greater detail. Using the additive noise operator with no modifiers, random walks tended to spend an initial period of time in moderate to high complexity regions of space before entering a low complexity region from which they escaped only sporadically, if at all. Figure 6.8(a) shows a representative example: after approximately 14,000 steps complexity remains low and very few new phenotypes are discovered. The relationship between complexity and number of unique phenotypes discovered is apparent. To highlight this relationship, we calculated moving averages of both complexity and the percentage of unique phenotypes. Figure 6.8(b) indicates that high complexity regions of space are associated with periods of rapid discovery while low complexity regions are associated with periods of little discovery of phenotypic novelty. Furthermore, once additive noise mutation has moved networks into less complex regions of space, they tend to become trapped there.

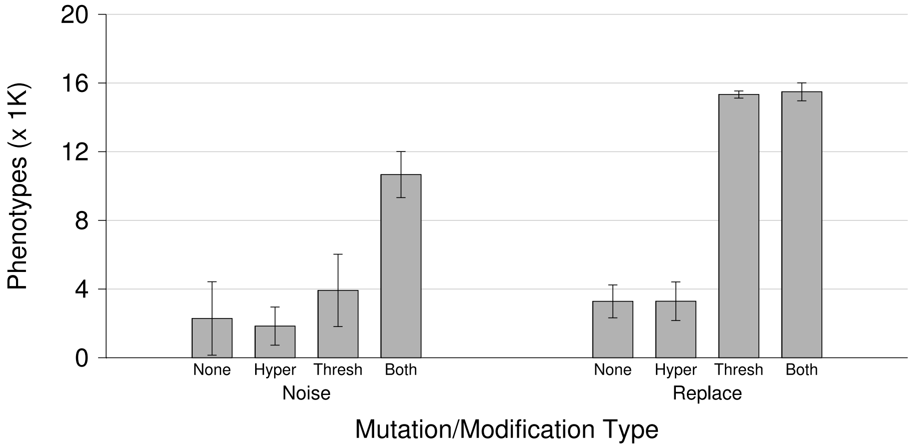


Figure 6.6: Number of unique phenotypes discovered by random walks using each of the different additive noise and weight replacement mutation operators.

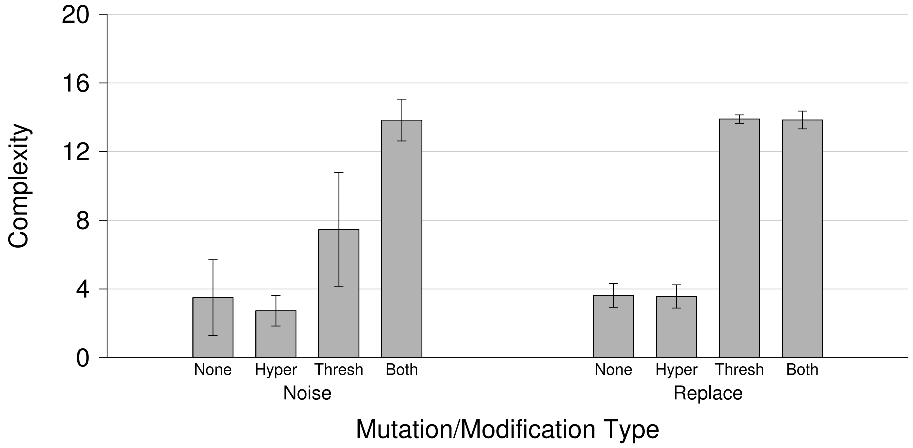
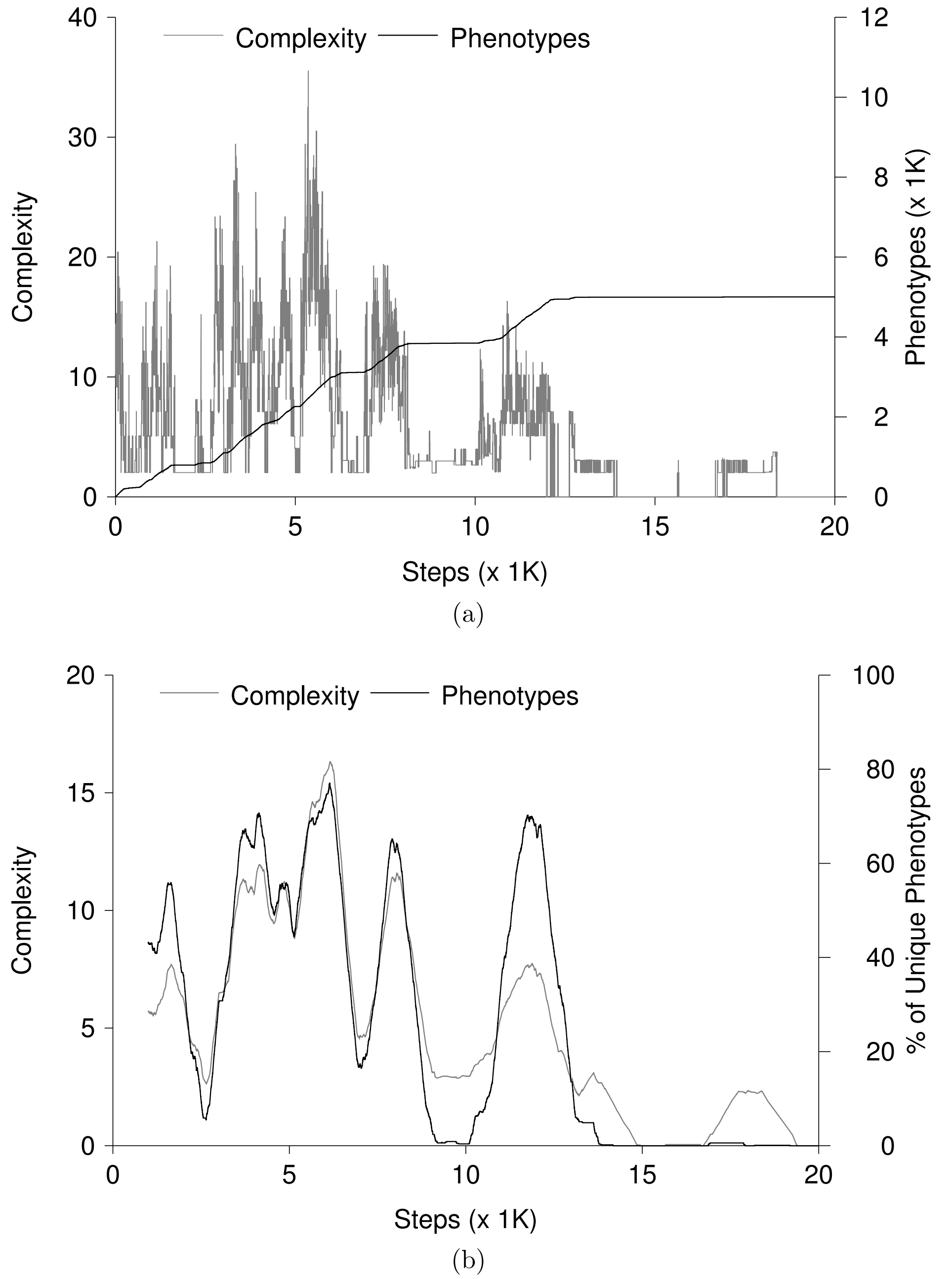


Figure 6.7: Average weighted complexity discovered over random walks using each of the different additive noise and weight replacement mutation operators.

When additive noise was used with both modifiers, random walks still visited both high and low complexity regions of space; however, they were less likely to become trapped in the low complexity regions (*e.g.*, the period of low complexity around steps $14 - 15,000$ in Figure 6.9). The relationship between complexity and frequency of phenotypic novelty remained.

Compared to those using additive noise mutation, random walks that use the weight replacement mutation operator were more uniform with respect to complexity and rate of phenotypic discovery. Figure 6.10 shows a random walk using weight replacement with both modifiers. Other weight replacement operators pro-



(a)

(b)

Figure 6.8: Additive noise—no modifiers: Complexity and phenotypic discovery; (a) shows the complexity of the current lineage and the number of unique phenotypes discovered so far; (b) shows the complexity and percentage of unique phenotypes discovered averaged over 1000 steps.

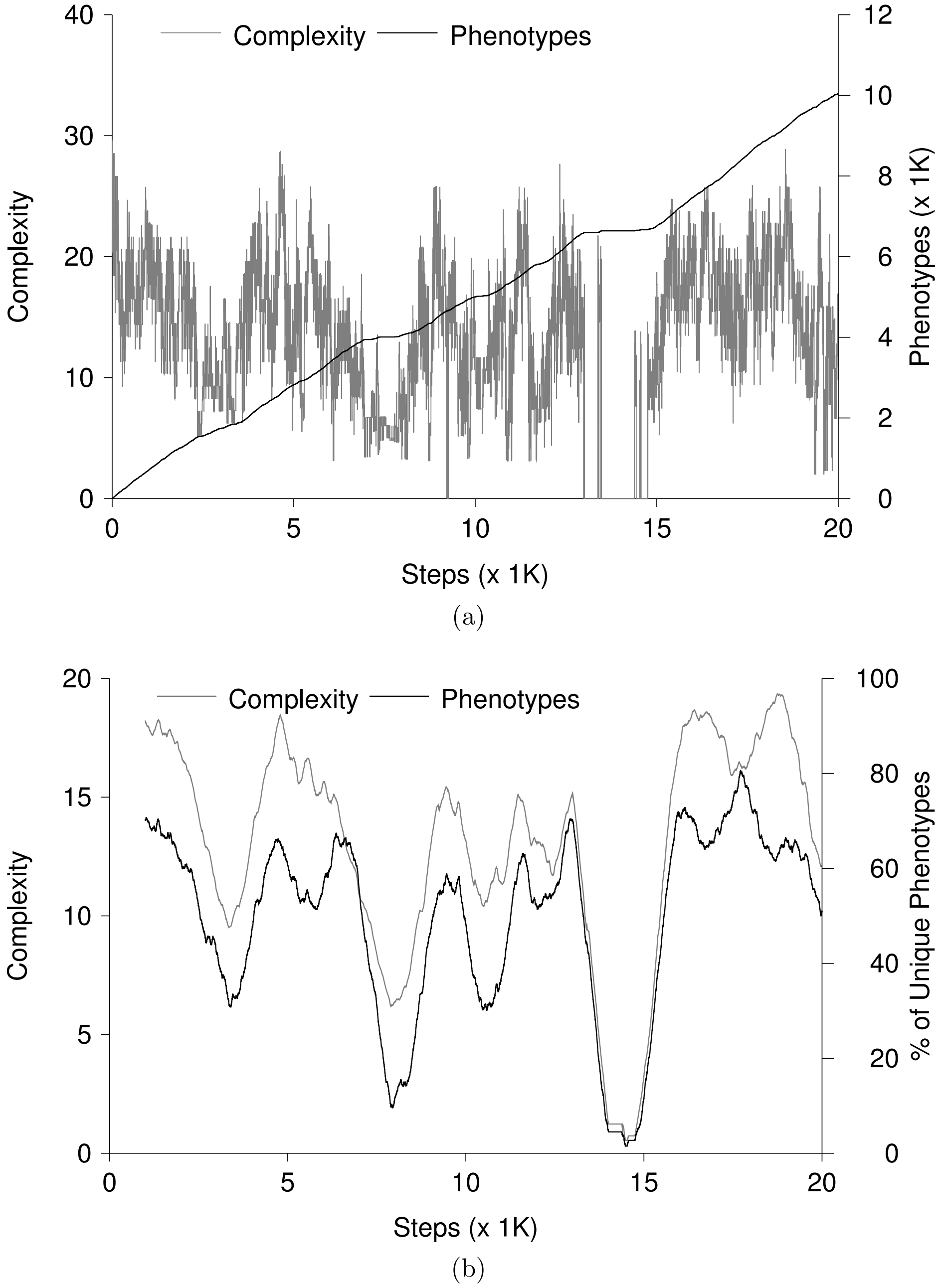


Figure 6.9: Additive noise—both modifiers: Complexity and phenotypic discovery; Compared to Figure 6.8, phenotypic discovery is more consistent and periods of low complexity are more transitory.

duced similar behaviours except that, without threshold adjustment, walks spent more time revisiting a small number of low complexity lineages. Thus, average complexity was reduced and the rate of phenotypic discovery, while remaining constant over the course of the walk, was lower.

6.1.4 Discussion

The preliminary investigation of each of the mutation operators using perturbation analysis raised the issue that it is difficult to compare the effects of different mutation operators directly in a quantitative fashion due to the fact that they are parameterised differently. However, the additive noise and weight replacement mutation operators clearly differ with respect to their effect on weight distribution (Figures 6.2 and 6.3), which biases the regions of ontogenetic space that are explored (Figure 6.5).

Of the two post-mutation modifiers, adjusting the threshold had the most significant effect on the number and complexity of ontogenies that could be reached by the mutation operators (Figures 6.6 and 6.7). Normalising network weights had a less noticeable effect, except when combined with threshold adjustment and the additive noise operator, in which case it resulted in a significant increase in both the number and complexity of ontogenies found.

Without threshold adjustment, the random walks were initialised in less complex, and hence less diverse, regions of space. It is likely that normalisation has minimal effect on the weight replacement operator because of the fact that replacing a weight with another drawn from the same distribution does not significantly change the weight distribution, and hence the radius of the hypersphere on which the network is located.

The results of this study suggest that the networks that produce complex lineages are be clustered in one or more regions of genotypic space, surrounded by large regions containing networks that produce only trivial lineages. Networks that produce complex lineages are surrounded by networks that produce a diverse range of similar, but not identical, lineages. In contrast, networks that produce trivial lineages occupy far less diverse regions of space. The random walks that spend the most time in these diverse regions of space, and hence discover novel phenotypes more rapidly, are those using either the weight replacement base operator with threshold adjustment or both modifiers, or the additive noise base operators with

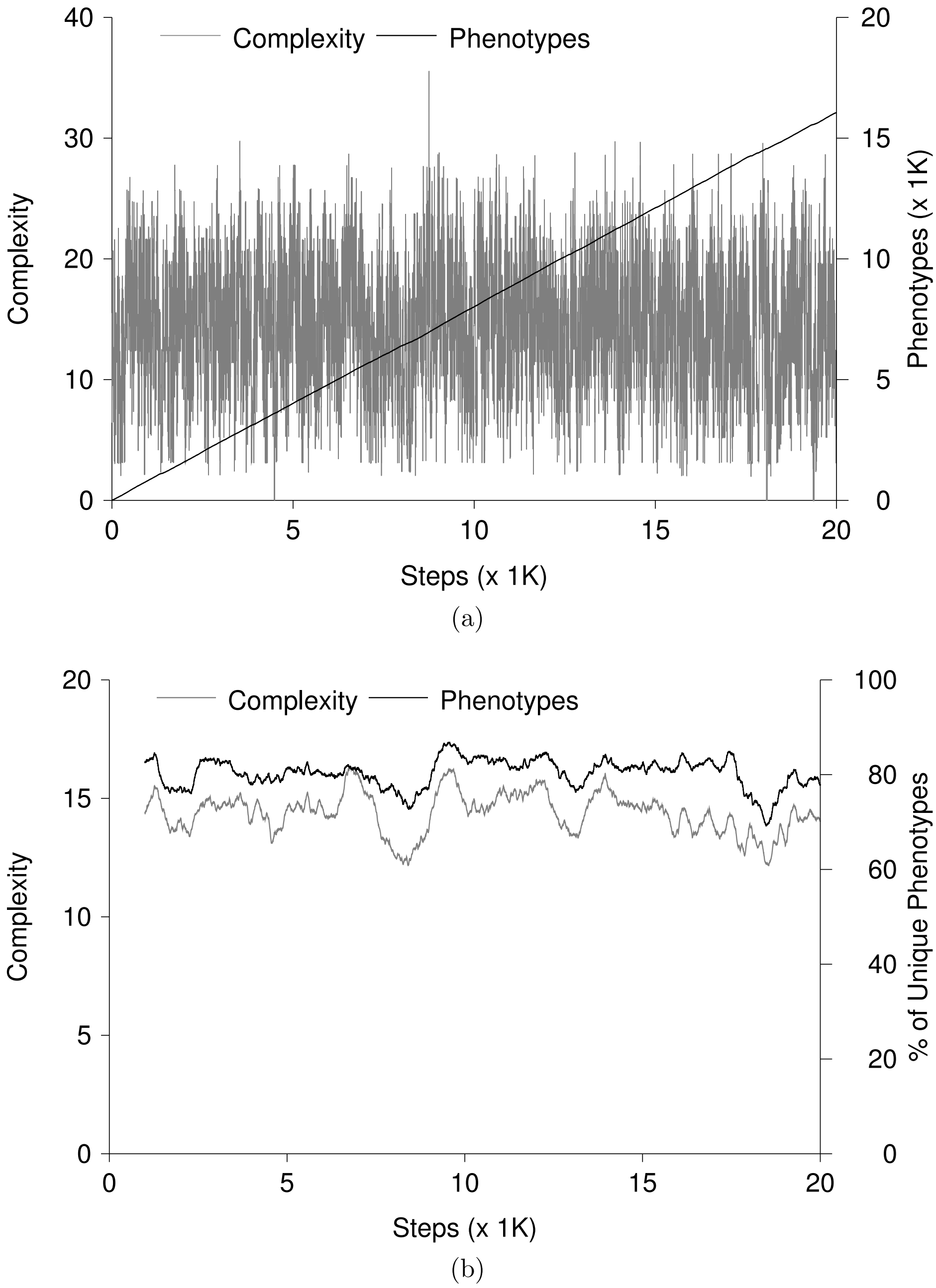


Figure 6.10: Weight replacement—both modifiers: Complexity and phenotypic discovery; Compared to Figures 6.8 and 6.9, phenotypic discovery is more consistent. The unsmoothed complexity is 'noisier' suggesting greater phenotypic change between subsequent steps.

both modifiers. We therefore anticipate that these three mutation operators will be the most effective for adaptive search.

6.2 Study 9: The evolution of ontogenies

The adaptive landscape used in *Study 8: Evaluation of mutation bias* was flat: from the point of view of selection, all individuals were equally fit. In most situations, evolution does not occur on such a level playing field. Certain individuals, by virtue of some heritable trait, will stand a better chance of surviving to pass on their genes to offspring than others. Which phenotypic traits will increase an organism's chance of reproduction will depend on the nature of the ecological niche it inhabits. It is therefore possible to imagine an adaptive gradient mapped to phenotypic space. The idea of an adaptive phenotypic space was introduced by Simpson (1944), who described a two-dimensional landscape representing the possible combinations of two phenotypic characters in which elevation corresponded to fitness. The highest point in the landscape represents the phenotype that is most adapted to the current environment. Because environments are dynamic, the location of this optimum point will move over time. Simpson's adaptive phenotypic landscape is a descendent of the fitness landscape described by Wright (1932) but differs in two respects. First, the axes of Wright's fitness landscape represent gene frequencies rather than phenotypic characters. Second, the structure of fitness landscapes is typically more complex due to epistatic interactions between genes.

There is an important relationship between genotypic and phenotypic landscapes. The adaptive phenotypic landscape specifies the direction of evolution favoured by selection. However, any movement from phenotype A to phenotype B in phenotypic space is contingent upon genotype B being mutationally accessible from genotype A in genotypic space (Figure 6.11). The mapping from a genotype to a phenotype is defined by the developmental process; therefore, properties of the developmental process will affect adaptation. In order to determine if development is biasing adaptation, we require a better understanding of the mapping between genotypic and phenotypic space.

Study 9: The evolution of ontogenies had two aims: The first aim was to investigate the ability of adaptive walks to locate networks capable of generating specific phenotypic targets, and to evaluate the relative merits of (a) the mutation

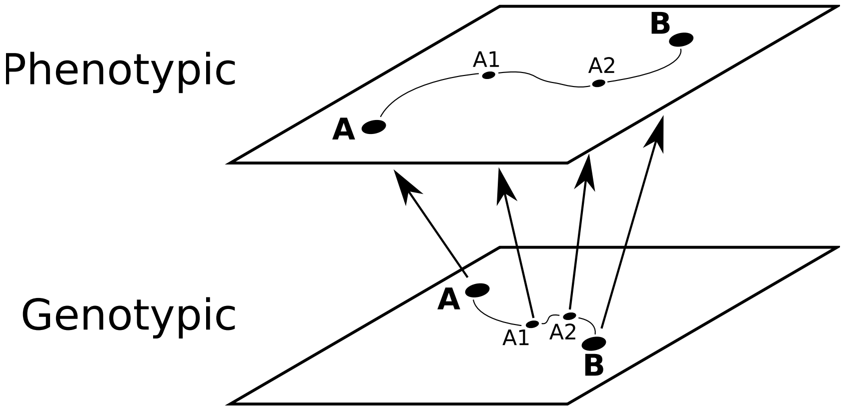


Figure 6.11: Phenotypic adaptation depends on mutational accessibility. In order for phenotype adaptation to proceed from phenotype **A** to phenotype **B**, there must be a mutationally accessible path of genotypes between genotypes **A** and **B**. The mapping from genotypic to phenotypic space will be affected by the nature of development.

operators introduced in the previous study and (b) different ways of defining adaptive gradients. The second aim was to investigate the types of ontogenies generated by those networks that successfully produced the desired phenotypic target and to analyse the successful adaptive walks in order to gain insight into what effect the developmental mapping has on the adaptive landscape.

The specific adaptive tasks used in this study are derived from the lineages of the organisms *C. elegans* and *H. roretzi.* The use of targets derived from real lineages is important because we know that they have been evolved once, and are of a realistic complexity level. The range of lineage complexities DRGNs are capable of producing was demonstrated in Chapter 5. Therefore, it was possible to match the complexity of the task against the known range of complexities of the controller to provide an indication of how challenging a particular task is for a particular network.

We make a simplifying assumption in this study that evolution is occurring in a fixed environment and the target phenotype is the most highly adapted to that environment. Fitness is then calculated in terms of minimising the distance between the current phenotype and the target phenotypes. In a real environment, ecological niches are highly dynamic, changing as environments change or according to fluctuations in co-evolutionary relationships. However, so long as the rate of evolution is more rapid than the rate of environmental change, the assumption of a static fitness landscape is not implausible. The next section defines four different measures of phenotypic distance based on varying levels of constraint. Three series of simulations are then described, focusing in turn on evaluating mutation operators, comparing the different phenotypic fitness measures and exploring the limitations on the size and complexity of phenotypes that may be found by search.

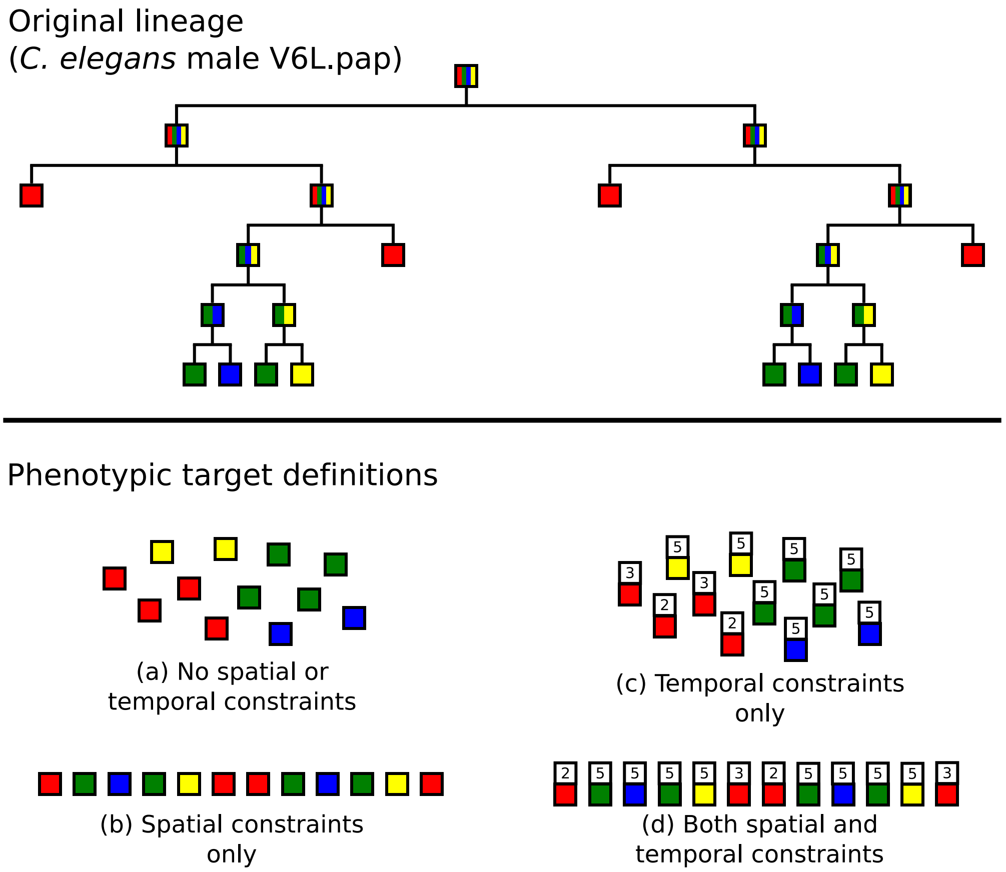
6.2.1 Phenotypic fitness metrics

Cell lineages are an organisational, rather than a morphological, description of a phenotype and can be quantified and compared in an automated fashion. For this study we defined metrics on the basis of the phenotypic component of a cell lineage—that is, the terminal cells.

A phenotypic target can be defined as the intersection of three types of constraint on the cells it contains: the identity of each cell, their spatial location and the time of their appearance. The first and most basic constraint is on the cell fate distribution: the requirement that a certain number of cells of each specific type are present at the end of development. The second constraint requires that each terminal cell be correctly positioned in relation to the other cells in the phenotype. The final constraint requires each cell to be produced at the correct point in time during development.

Four different base fitness metrics were used. The identity constraint was considered fundamental and always used, on top of which temporal and spatial constraints could be applied either separately or together. When there were no spatial or temporal constraints, the only requirement was for a lineage to contain the correct number and type of terminal cells. When there were spatial constraints only, each of the terminal cells had to be correctly ordered, regardless of its depth in the lineage. When there were temporal constraints only, the terminal cells had to appear at the correct depth, regardless of order. When there were both spatial and temporal constraints, each terminal cell had to appear at the correct depth and in the correct order. The practical implication of each of these constraints and their intersection is illustrated in Figure 6.12.

No temporal or spatial constraints. When there were no temporal or spatial constraints, a phenotype was considered as an unordered set of cell fates and the



(a) No spatial or temporal constraints

(c) Temporal constraints only

(b) Spatial constraints only

(d) Both spatial and temporal constraints

Figure 6.12: The four phenotypic distance metrics as applied to the *C. elegans* male V6L.pap lineage (Braun et al., 2003): (a) **No spatial or temporal constraints:** the target phenotype was described as an unordered set of cell fates; (b) **Spatial constraints only:** the target phenotype was described as an ordered set (or vector) of cell fates; (c) **Temporal constraints only:** the target phenotype was again described as an unordered set of cell fates, but each cell fate was now also tagged with the depth of the lineage at which it appeared; (d) **Both spatial and temporal constraints:** the target phenotype was described as a vector of depth-tagged cell fates. See text for equations.

fitness $f(C,T)$ of the current cell fate set C with respect to the target cell fate set T was defined as:

$$f(C,T) = \frac{|(C \cap T)| - |(C \ominus T)|}{|T|} \tag{6.1}$$

where $|T|$ is the size of set T, $C \cap T$ is the intersection of sets C and T and $C \ominus T$ is the symmetric difference of sets C and T.

Temporal constraints only. When temporal constraints were used, each cell fate was tagged with its depth in the lineage and equation 6.1 was used to calculate fitness.

Spatial constraints only. When spatial constraints were used, a phenotype was considered as an ordered sequences of cell fates and the fitness $f(C, T)$ of the current cell fate sequence C with respect to the target cell fate sequences T was defined as:

$$f(C, T) = \frac{\textsc{LevenshteinDistance}(C, T)}{|T|} \tag{6.2}$$

where $\textsc{LevenshteinDistance}(C, T)$ was the Levenshtein distance between sequences C and T (as defined in §5.4.1) and $|T|$ was the length of sequence T.

Both temporal and spatial constraints. When both temporal and spatial constraints were used, each cell fate was tagged with its depth in the lineage and equation 6.2 was used to calculate fitness.

After calculating a base fitness value according to one of the metrics described above, that value was normalised such that the maximum fitness (a perfect match between the current and target fates) was 1.0 and the bulk of the remaining fitness values were in the range $[0.0, 1.0)$. Due to the open-ended nature of the possible solutions (*i.e.*, within the constraints of the simulation, a lineage could be arbitrarily large), it was not possible to fix an absolute lower bound on the values that fitness could take. In practice, it was observed that adaptive walks rapidly found solutions with fitness > 0.0 (within the first few successful steps) and so the majority of the walk occurred in the fitness range $[0.0, 1.0]$.

6.2.2 Series A: Varying mutation operators

The first series of adaptive walk simulations compared the performance of each of the mutation operators described in the previous study on a search task. The target phenotype was the *C. elegans* male V6L.pap lineage with spatial, but not temporal constraints (*i.e.*, as illustrated in Figure 6.12 (a)). Based on the results of *Study 8: Evaluation of mutation bias*, it was expected that the mutation conditions that

Table 6.2: Performance of walks using different mutation operators

Mutation Operator	θ Adjusted	Normalised	Perfect Runs (of 100)	Fitness Avg. (Std. Dev.)
Additive Noise	No	No	21	0.798 (0.165)
	No	Yes	18	0.810 (0.129)
	Yes	No	37	0.902 (0.091)
	Yes	Yes	26	0.696 (0.333)
Weight Replacement	No	No	61	0.940 (0.084)
	No	Yes	64	0.943 (0.079)
	Yes	No	48	0.905 (0.112)
	Yes	Yes	30	0.747 (0.305)

achieved high rates of phenotypic discovery—additive noise with both modifiers, and weight replacement with threshold adjustment or both modifiers—would be the most effective for search, given that they generated a greater level of diversity on which selection could act.

To evaluate the performance of each mutation operator on adaptive search, ensembles of 100 random networks ($N = 8, K = 8, W = 2.0$) were created and eight adaptive walks performed for each individual using one of each of the eight mutation operators described above (§6.1.1). Each adaptive walk consisted of 20,000 steps. At each step, a new network was created using the current mutation operator. The newly created network was accepted if its fitness was greater than or equal to that of the current network (*i.e.*, neutral mutations were always accepted).

The results from Series A clearly demonstrate that, on average, random walks using the weight replacement mutation operator outperform those using the additive noise mutation operator (Table 6.2). In addition to achieving higher average fitnesses, a larger number of the weight replacement runs resulted in networks that generated perfect phenotypes.

One possible explanation for these results is that the greater rate of phenotypic discovery achieved by the weight replacement operator simply enabled a greater number of perfect solutions to be obtained in the time allowed (20,000 steps). However, examination of the unsuccessful additive noise runs revealed more 'non-starters'—that is, walks that, within the first 1,000 steps, became trapped in regions of low fitness from which mutation was unable to locate any fitter individuals.

As anticipated on the basis of the previous study, constraining the adaptive walk to the surface of a hypersphere (via normalisation) had a negligible effect on performance. Contrary to expectations, adjusting the threshold, while beneficial for the walks using the noise operator, resulted in a decrease in performance of walks using the replacement operator. Similarly, for either base mutation operator, using both post-mutation modifiers together—which resulted in the greatest increases in rate of phenotypic discovery on random walks—actually decreased performance on adaptive walks. One possible explanation is that the systemic modifier involved in normalising weights, while increasing the probability of generating a previously unseen phenotype, was actually disruptive in an adaptive context. One strategy that may be used by the adaptive process is to preserve a particular subset of network weights, while experimenting with a disjoint subset. Normalisation will modify all weights, causing the loss of both the well adapted as well as the experimental subsets.

One of the aims of this series of simulations was to ascertain which mutation operator was to be used for the subsequent simulation series. While the normalisation modifier did result in a small improvement in search performance, this came at the cost of an increase in the computational cost of running simulations. The weight replacement mutation operator without either of the post-mutation modifiers, which performed only slightly worse but was computationally more efficient, was therefore used for all simulations in Series B and Series C.

6.2.3 Series B: Varying phenotypic distance metrics

The second series of adaptive walks compared the effect of using the four different phenotypic distance metrics described in §6.2.1 as fitness measures. Two questions were addressed by these simulations. How does the level of phenotypic constraint affect the difficulty of the search task? What types of ontogenies does adaptation find to satisfy the different phenotypic definitions?

To address these questions, an ensemble of 500 networks ($N = 8, K = 8, W = 2.0$) were created and four adaptive walks using one each of the four phenotypic fitness metrics were performed for each individual. Each adaptive walk consisted of 20,000 steps. The weight replacement mutation operator was used to create all mutants. As before, a newly created lineage replaced the current lineage if its fitness was either equal to or greater than that of the current lineage.

As anticipated, as phenotypic definition became more constrained, the diffi-

Table 6.3: Performance of walks using different phenotypic distance metrics

Temporal Constraint	Spatial Constraint	Perfect Runs (of 500)	Unique Lineages
No	No	499	496
No	Yes	288	103
Yes	No	201	113
Yes	Yes	27	1

culty of the search process increased (Table 6.3). With no spatial or temporal constraints, only one of 500 walks failed to find a perfect solution. In contrast, with both spatial and temporal constraints, only 27 of 500 walks were able to find lineages that produced the target phenotype. When the phenotypic definition incorporated either spatial or temporal constraints, around half of the runs found lineages that produced the target phenotype. Spatial constraints were moderately easier to satisfy than temporal constraints (288 compared to 201 perfect solutions).

The phenotypic definition had a significant effect on the variety of lineages that were found. Of the 499 solutions found with no spatial and temporal constraints, 496 of the lineages generating these phenotypes were unique. In contrast, the intersection of spatial and temporal constraints restricted the space of possible solutions to a single lineage, that of the original data set. One explanation for the lower rate of success under this phenotypic definition appears to be the structure of the adaptive landscape. Using the least constrained phenotypic definition means that a greater number of lineages map to the target phenotype, and hence a larger proportion of genotypic space maps, via ontogeny, to a perfect fitness value. When the most constrained phenotypic definition is used, only a single lineage maps to the target phenotype, and hence a much smaller proportion of genotypic space maps to a perfect fitness value.

6.2.4 Series C: Varying phenotypic targets

The third and final series of adaptive walks compared the performance of adaptive walks on five target lineages derived from real data sets. The first target lineage was the *C. elegans* male V6L.pap used in Series A and Series B (shown in Figure 6.12). Three further target lineages from *C. elegans* were also used: the sublineage of the C founder cell, which produces the muscle and epidermis cells

Table 6.4: Target Lineage Details

Lineage	Number of Cells	Number of Cell Types	Maximum Depth	Weighted Complexity
C. elegans maleV6Lpap	12	4	5	6.55
C. elegans C	48	4	6	11.23
C. elegans MSp	46	5	7	22.49
C. elegans MSa	48	5	7	26.55
H. roretzi (half)	55	7	6	31.57

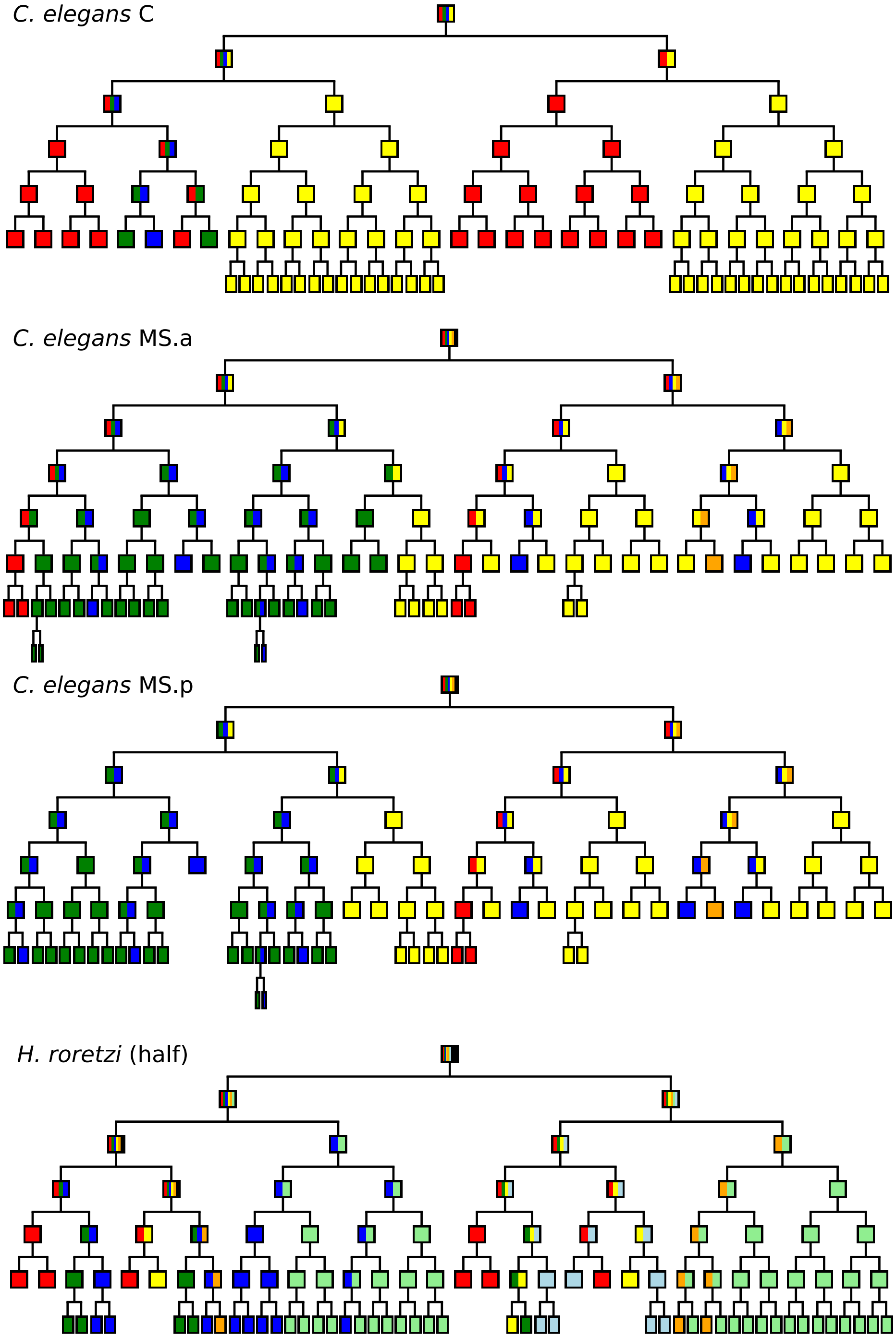
Table 6.5: Performance of walks using targets of varying complexity

Target Lineage	Best Fitness	Remaining Errors	Avg. Fitness (Std. Dev.)	Perfect Runs (of 50)
C. elegans maleV6Lpap	1.0	-	0.938 (0.071)	24
C. elegans C	1.0	-	0.950 (0.038)	6
C. elegans MSp	0.956	3	0.852 (0.068)	-
C. elegans MSa	0.958	3	0.834 (0.076)	-
H. roretzi (half)	0.982	1	0.745 (0.074)	-

in the posterior region of the worm's body; and two sublineages, MSa and MSp, of the MS founder cell, which primarily produces the pharynx (a digestive organ), but also some muscle cells and the somatic gonad precursors (Sulston et al., 1983). The final target lineage was taken from the ascidian *H. roretzi* (Nishida, 1987). The properties of the five lineages are summarised in Table 6.4 and the cell lineages themselves illustrated in Figure 6.13.

An ensemble of 50 random networks ($N = 16, K = 16, W = 2.0$) was generated and adaptive walks were performed using each of the five different targets for each individual. The weight replacement mutation operator was used to create all mutants. The second phenotypic definition (spatial constraints only) was used to evaluate the fitness of each network's phenotype. The additional targets are of significantly greater complexity than that used in the previous simulations. Preliminary trials indicated that 8 node networks performed poorly on the larger lineages, therefore networks of 16 nodes were used. Larger networks contain more weights and the genotypic space in which the adaptive walk must search is consequently larger. Therefore the maximum length of the walks was increased by a factor of three to 60,000 steps.

The results of Series C demonstrate that adaptive search becomes more difficult



C. elegans C

C. elegans MS.a

C. elegans MS.p

H. roretzi (half)

Figure 6.13: Four additional lineages used as adaptive walk targets: Three sub-lineages from *C. elegans*—C, MSa and MSp—together with half of the *H. roretzi* lineage (the second half of the lineage is identical to the first). The initial target lineage, *C. elegans* male V6Lpap, is illustrated in Figure 6.12.

as the complexity of the target lineage increases. While almost half of the walks were able to locate the simplest lineage (*C. elegans* maleV6Lpap), the best performing walk on the most complex lineage (*H. roretzi*) contained a single incorrect cell after 60,000 steps.

In order to demonstrate that the MSp, MSa and *H. roretzi* tasks were in fact achievable, the best performing networks on each of these targets were re-run with no limitations on the maximum length of the walk. At least one walk was able to locate each of the target lineages; however the search times required were on the order of 300,000 steps.

6.2.5 Analysis of adaptive search results

This section analyses two aspects of the adaptive walk simulations. First, a single adaptive walk is examined in detail. The lineages found by the adaptive walks of Series B and C are then compared to stochastic lineages generated by a Markovian process (Braun et al., 2003) (described in §5.3.2).

Analysis of an adaptive walk

The progress of an adaptive walk towards a target may be measured in several ways (Figure 6.14). The walk shown achieved a fitness of 0.98 (1 cell remaining incorrect) on the *H. roretzi* target in the initial Series C simulations. An additional 250,000 steps were required to correct the remaining error, therefore only the initial 60,000 steps are illustrated here.

Fitness followed a hyperbolic trajectory over the duration of the walk. This trajectory is commonly observed in observations of evolution in both computational (Kauffman, 1993) and *in vitro* (Lenski and Travisano, 1994) conditions and has also been modelled analytically (Orr, 1998).

Complexity tended to increase over the course of the adaptive walk, achieving the complexity of the target lineage after approximately 7,000 steps and thereafter fluctuating about that value. One anomaly is the initial spike in complexity within the first 100 steps (more clearly visible in Figure 6.15). Comparing the fitness and complexity plots, it is evident that there is a degree of neutrality in the mapping from ontogenetic space (measured by complexity) to the fitness landscape. Clearly it is possible for multiple lineages to share an equal fitness value. Furthermore, it

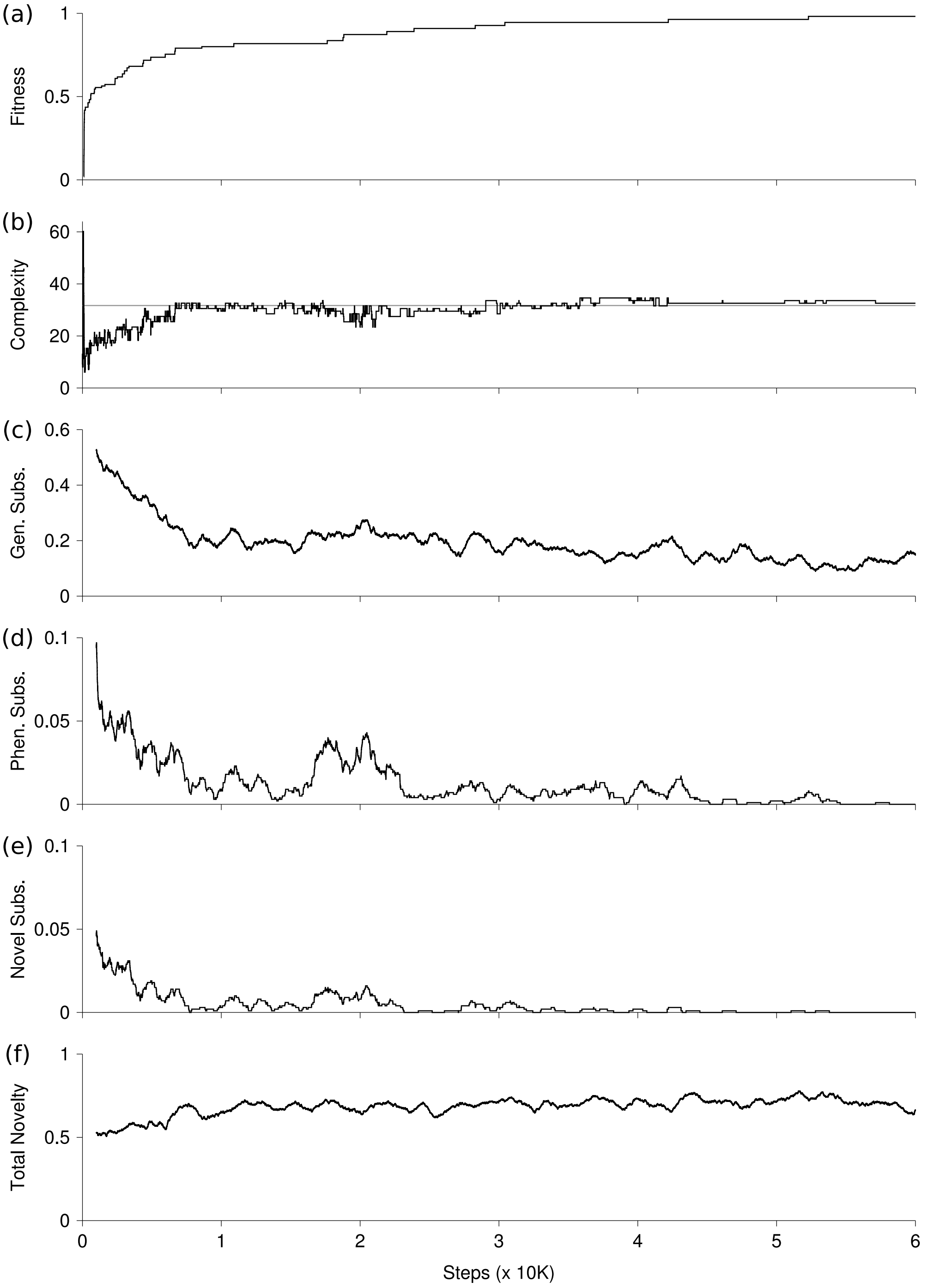


Figure 6.14: Analysis of a single adaptive walk using the *H. roretzi* target lineage. From top to bottom, the plots show: (a) fitness; (b) complexity; (c) genotypic substitution rate; (d) phenotypic substitution rate; (e) accepted phenotype novelty rate; (f) generated phenotype novelty rate. See text for further details.

is possible for an adaptive walk to move between these equivalent lineages via the weight replacement mutation operator.

Genotypic substitution rate measures the acceptance of newly created networks. Initially, around 60% of mutations are accepted (*i.e.*, are either beneficial or neutral). This probability decreases at a constant rate until around step 7,000. After this point, approximately 20% of mutations are accepted with a moderate decrease over the remainder of the walk. Should this statistic ever reach 0, it is possible that no further adaptation could occur as the network weights would be so finely tuned that any mutation would be detrimental. In practice, this phenomenon was never observed in any of the simulations reported here: there was sufficient neutrality in the gene network to lineage mapping to ensure that some change was possible. When walks were continued past the point where a perfect solution had been found, a high genotypic substitution rate continued to be observed.

Phenotypic substitution rate measures the acceptance of networks that generated a different phenotype to the previous network. Initially, around 10% of accepted networks generate different phenotypes. This probability decreases to almost zero after approximately 10,000 generations and thereafter fluctuates. Towards the end of the adaptive walk, the probability of phenotypic substitution falls to zero. The discrepancy between the probability of genotypic and phenotypic substitution can be explained by the degree of neutrality in the mapping from genotypic to phenotypic space: while a relatively large number of mutations are accepted throughout the adaptive walk, the proportion of these that result in phenotypic change decreases. The ontogenetic substitution rate (the acceptance of a network generating a different lineage) was also measured and found to be identical to the phenotypic substitution rate. Therefore, while the results of Series B (§6.2.3) indicated that it is possible for more than one lineage to produce the same phenotype, all of the ontogenetic changes in this adaptive walk were accompanied by phenotypic change.

Accepted phenotype novelty rate measures the acceptance of networks that generated a previously unseen phenotype. Again, a rapid initial decrease was followed by a gradual decrease to zero as the adaptive walk proceeded. Given the many-to-one mapping from genotypic to phenotypic space, it is possible that a previously seen phenotype could be rediscovered from an entirely different position in genotypic space. This rediscovery could therefore be advantageous if the new genotype responsible is located in a more promising region of genotypic space—one in which the mutationally accessible ontogenies result in more fit phenotypes.

Generated phenotype novelty rate: measures the generation of novel phenotypes by a newly created network, irrespective of whether its fitness is better than, equal to or worse than the current best. Phenotypic discovery remained high (above 50%) over the entire duration of the adaptive walk. This constant rate of discovery suggests that, while more accurate lineages do become harder to find, it is not due to the potential diversity of the system being exhausted. Novel phenotypes continue to be generated; however, the vast majority of these are less fit than the current best phenotype.

The statistics plotted in Figure 6.14 show how the adaptive walk proceeds towards the target in terms of rate of discovery, but provide little information about how the structure of the current best lineage changes. In Figure 6.15, the first 10,000 steps of the adaptive walk have been enlarged and eight positions have been highlighted on the graph (labelled (a) through (h)). The lineages produced by the current best network at each of these positions are shown in Figure 6.16.

(a) At the beginning of the walk there is clearly no correlation between the current and target phenotypes. The lineage generated by the random initial network is large and relatively homogeneous in terms of structure and cell fate distribution.

(b) After 60 steps (19 successful mutations) the network has adapted to generate a more diverse range of cell fates, although it still bears little resemblance to the target phenotype. There is little apparent modular structure in the lineage at this point, resulting in a significant increase in complexity (60.39, compared with 9.04 at generation 0).

(c) Step 112 (37 successful mutations) produces the first network to generate a phenotype containing 55 terminal cells (the number in the target phenotype). The reduction in the size of the lineage is accompanied by a corresponding reduction in its complexity.

(d) Step 147 (45 successful mutations): produces the simplest lineage accepted by selection. After this point, the size of the lineage remains relative constant

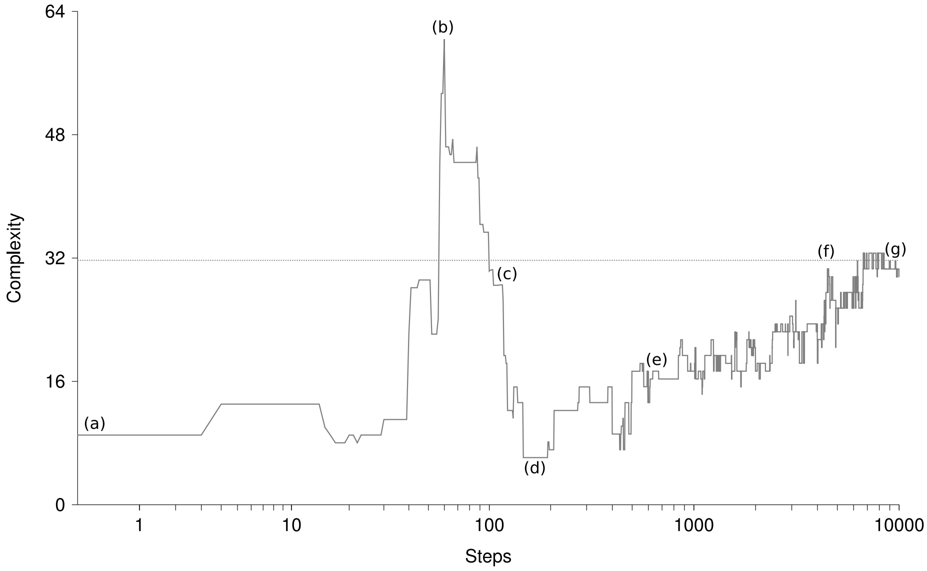


Figure 6.15: An enlarged view of complexity for the first 10,000 steps of the adaptive walk shown in Figure 6.14 with a log scaled x-axis. The horizontal line indicates the complexity of the *H. roretzi* lineage used to define the target. The seven lineages from the points labelled (a) through (g) are described further in the text.

and selection acts to establish the dominant cell fate (light green). Complexity continues to reduce as the number of cell fates appearing in the lineage decreases.

(e) Step 592 (80 successful mutations) produces the first network to generate a lineage with recognisable translational symmetry between its left and right halves. This step represents the last major change to the structure of the lineage.

(f) Step 5964 (284 successful mutations): The first appearance of a lineage containing all 7 cell fates occurring in the target phenotype.

(g) Step 9977 (345 successful mutations): The complexity of the current lineage has now approached that of the *H.roretzi* lineage. For the remaining steps of the simulation selection acted to fine tune the identities of the terminal cell fates with only minor changes to lineage structure.

(h) Step 336,460 (1,217 successful mutations) eliminated the final incorrect cell

Table 6.6: Statistics for the lineages shown in Figure 6.16

Lineage	Fitness	Number of Cells	Number of Cell Types	Maximum Depth	Weighted Complexity
(a)	-3.05	236	2	8	9.04
(b)	-1.29	156	5	8	60.39
(c)	0.18	55	4	8	28.52
(d)	0.41	54	3	8	6.11
(e)	0.49	53	4	8	17.33
(f)	0.75	52	7	8	29.56
(g)	0.80	54	7	8	29.54
(h)	1.0	55	7	8	34.63

and achieved a 100% accurate phenotype. While the lineage is quite different in appearance from that of the original data set, the spatial distribution of terminal cells is equivalent, as is the complexity (Table 6.6).

Comparison of stochastic and evolved lineages

In order to ascertain whether evolution had been biased by the network-lineage developmental process, the evolved *C. elegans* male V6L.pap lineages from Series B were compared to ensembles of stochastic lineages that produced equivalent phenotypes. Two of the four phenotypic distance metrics were considered: *No spatial or temporal constraints* (Figure 6.12(a)) and *Spatial constraints only* (Figure 6.12(b)).

No temporal or spatial constraints. The ensemble of evolved lineages contained the final lineages found by the 499 perfect runs from the first set of runs in Series B (Table 6.3, first row). The stochastic lineages were generated using the Markovian procedure described in §5.3.2. To recap, a lineage was generated by successively choosing a terminal cells to divide at random, until a lineage with 12 terminal cells was obtained. Once this ontogeny had been generated, the 12 phenotypic cell fates were randomly assigned to the terminal nodes.

Spatial constraints only. The ensemble of evolved lineages contained the final lineages found by the 288 perfect runs from the second set of runs in Series B (Table 6.3, second row). The stochastic lineages were again generated using the Markovian procedure described in §5.3.2. After each stochastic ontogeny had been

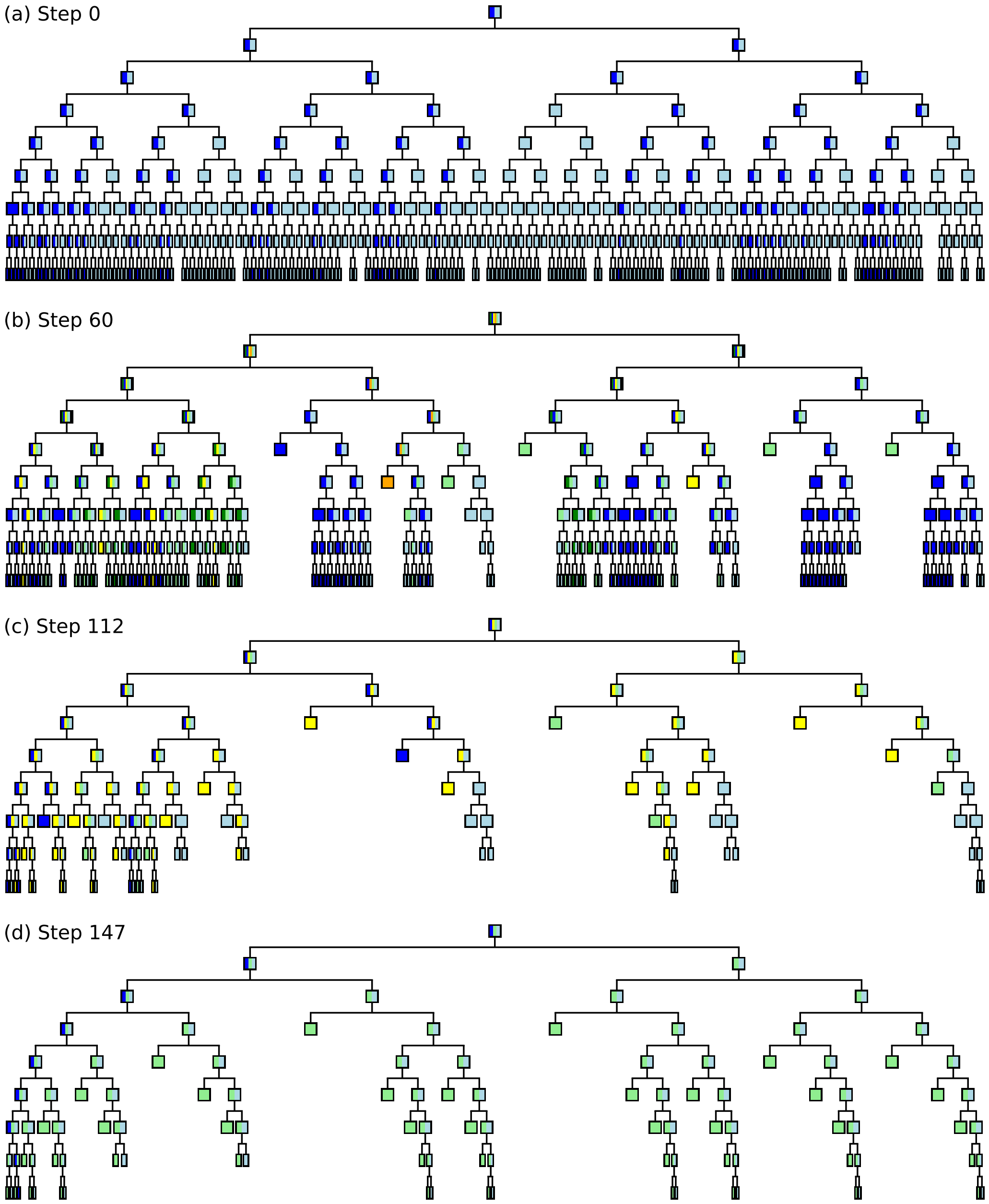


Figure 6.16: The lineages generated by the best solution at each of the points labelled (a)–(d) in Figure 6.15.

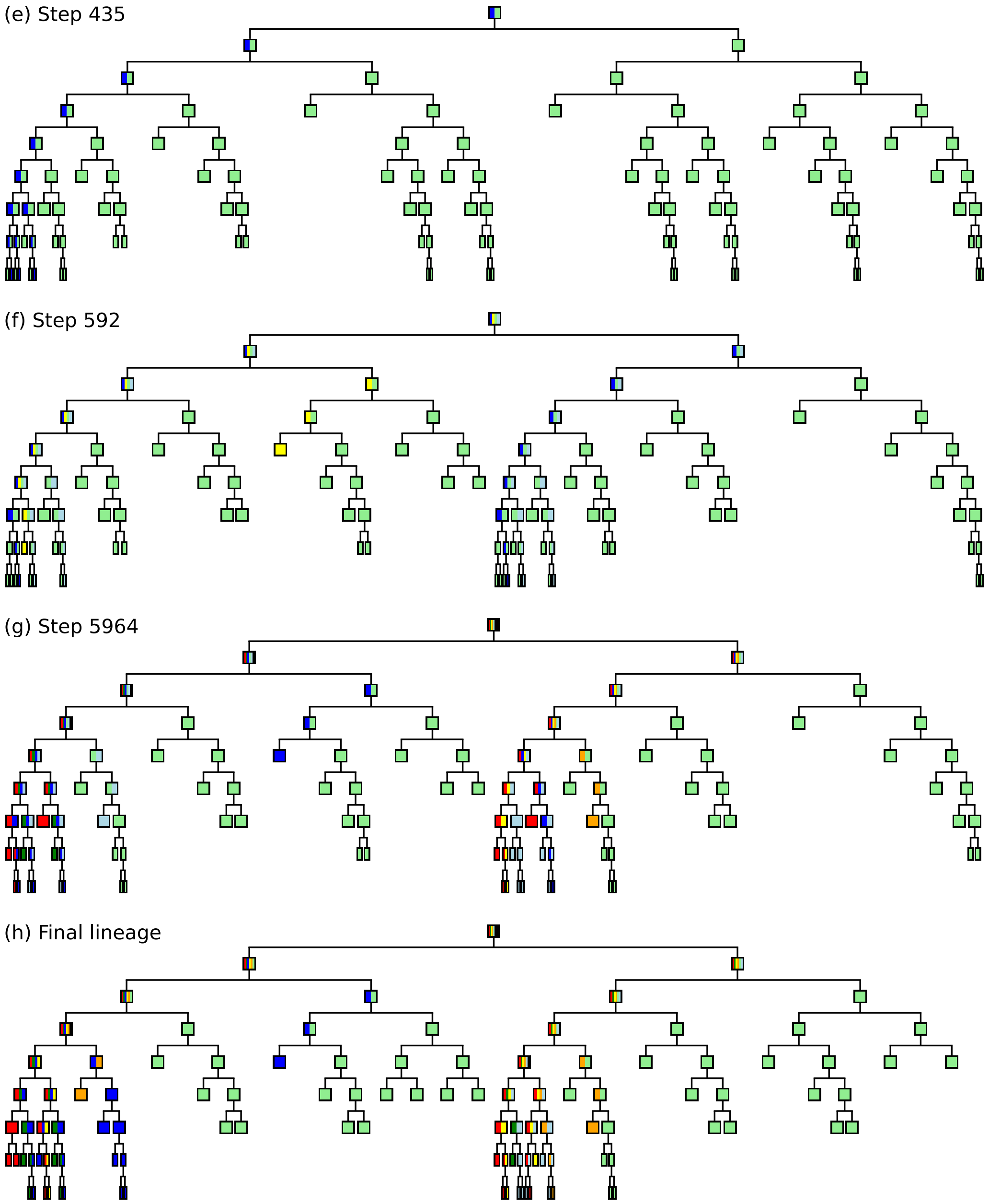


Figure 6.16: (continued) The lineages generated by the best solution at each of the points labelled (e)–(g) in Figure 6.15; and (h) the first lineage discovered that produced the target phenotype with 100% accuracy.

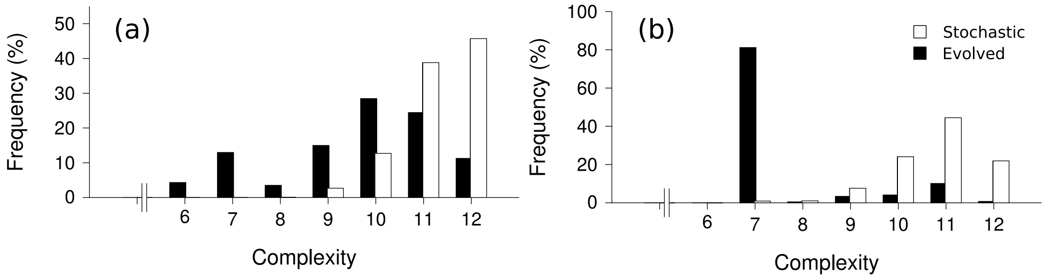


Figure 6.17: Complexity distributions of stochastic and evolved ensembles of lineages with (a) no spatial constraint on the target phenotype; and (b) spatial constraints on the target phenotype.

generated, the 12 phenotypic cell fates were assigned to the terminal nodes in the same order that they appear in the target lineage.

The complexity of each lineage in the four ensembles (two evolved and two stochastic) was then calculated and the complexity distributions for each pair of sets were compared (Figure 6.17). In both the presence and absence of spatial constraints on phenotypic cell fates, the evolved lineages are consistently less complex than equivalent stochastic lineages. With no spatial constraints (Figure 6.17(a)), the complexity distribution peak for the stochastic ensemble occurs at a complexity 12 (the maximum possible complexity given the size of the target lineage), while that for the evolved ensemble occurs at 10. When spatial constraints are present (Figure 6.17(b)), the complexity distribution peak for of the stochastic ensemble occurs at 11, while that for the evolved ensemble occurs at 7 (the complexity of the original target lineage).

A possible explanation for the difference between the distributions with and without spatial constraints is that spatial constraints impose a restriction on the set of possible lineages that can be generated by either an evolutionary or an adaptive process. In particular, the twelve cell phenotypic fate distribution consists of a single sequence of length six repeated twice (Figure 6.12). As a result, the complexity distributions for both the evolved and stochastic lineages are shifted to the left in the second pair of ensembles (when spatial constraints are present). The more significant contrast however, is between the evolved and stochastic ensembles using the same phenotypic distance metric. In both cases, the peaks for the evolved lineages are lower than those for the stochastic lineages.

6.3 Discussion

The studies reported in this chapter applied an adaptive process to the task of searching ontogenetic space for phenotypic targets and demonstrated the control capabilities of DRGNs. Chapter 5 indicated that DRGNs were able to generate a diverse range of ontogenies. The results in this chapter indicated that DRGNs are also capable of generating the *specific* ontogenies, comparable to those observed in real biological organisms. This section discusses, in turn: mutation bias, factors affecting search performance, neutrality and the distribution of mutation sizes.

Mutation operators can bias the structure of variation

The results of *Study 8: Evaluation of mutation bias* (§6.1) indicate that the choice of mutation operator can bias the structure of variation produced. Two different types of mutation operator—additive noise and weight replacement—were investigated. In the absence of selection, repeated applications of the additive noise mutation operator were found to alter the distribution of genotypic weights, essentially increasing the value of W. As observed in §5.3, when W increases, the distribution of ontogenetic complexities shifts downwards (Figure 5.9) and the structure of phenotypic variation changes accordingly. As expected, the complexity of the ontogenies visited by a random walk using random noise also decreased (Figure 6.4). Two post-mutation modifiers—hypersphere normalisation and threshold adjustment—were introduced that could alleviate the effects of the mutation bias resulting from the additive noise mutation operator. In contrast, the weight replacement mutation operator preserved the distribution of genotypic weights.

As discussed by Bullock (1999, 2001), the existence of different levels of mutation bias between operators does not necessarily provide a sufficient basis for choosing any one operator over another. Knowledge of the intrinsic biases of different operators does allow their effects on adaptation to be anticipated. In the context of this thesis, knowledge of mutation biases allows us to distinguish between the effects of mutation bias (which appears at the genotypic level) and developmental bias (which appears at the phenotypic level).

An additional finding of this study was that different regions of genotypic space map to different levels of phenotypic diversity. In general, the regions of ontogenetic space containing more complex lineages also produce the most diverse range of phenotypes. Thus, the weight replacement mutation operators, by preserving genotypic weight distribution, keep random walks in more complex regions of ontogenetic space and produce the greatest phenotypic diversity.

Factors affecting adaptive search performance

Study 9: The evolution of ontogenies (§6.2) used three series of adaptive walks to investigate different factors that could affect search performance: mutation operator, phenotype definition (fitness function), and task complexity.

Mutation operator: The weight replacement mutation operators outperformed the additive noise mutation operators. The post-mutation modifiers, despite increasing the rate of phenotypic discovery as described above, did not produce consistently better performance on adaptive walks. The most likely explanation for this observation is that the modifiers were disruptive and prevented well adapted subsections of the network from being preserved.

Phenotype definition: Increasing the level of phenotypic constraint increased the difficulty of the search task by restricting the number of candidate lineages that satisfied the phenotypic definition. Almost 500 different lineages (out of 500 walks) were found to satisfy the least constrained definition, which required only that the correct number of each cell type be present in the set of terminal cells. In comparison, the intersection of spatial and temporal constraints required every cell to be in the correct position in the lineage with respect to both its depth and its order. For the data set used (the *C. elegans* maleV6Lpap lineage), there was only a single cell lineage that satisfied these requirements. The difficulty of the search task was therefore greatly increased, as an adaptive landscape containing many global optima was replaced by one containing a single global optimum.

Task complexity: The complexity of the target lineage was found to be a reasonable indicator of search performance, with mean fitness decreasing linearly as complexity increased, although there was considerable variation in the fitnesses achieved for each target.

Preliminary simulations for this series of walks indicated that there was some benefit to be gained from using a larger network ($N_R = 16$), particularly for the more complex targets. However, this benefit was not observed across all target lineages. The proportion of perfect runs on the *C. elegans* male V6L.pap target in Series C ($N_R = 16$, 48%) was less than obtained in Series A ($N_R = 8$, 61%) or Series B ($N_R = 8$, 57.6%), despite the adaptive walks having a threefold increase in maximum number of steps (60,000 compared to 20,000; all other parameters were equivalent). It appears that increasing the size of the control structure—and the upper limit on performance—comes at the cost of increasing the difficulty of searching genotypic space. While not explicitly explored here, this trade-off has been observed previously (Geard and Wiles, 2005) and emerges implicitly from the results of the second study.

Neutrality in the ontogenetic mapping

Two types of neutrality were observed to affect the adaptive exploration of genotypic space. The first is in the mapping from genotype and ontogeny. There are many different combinations of network weights that produce identical cell lineage trees. This neutrality accounts for the robustness of networks to structural perturbations reported in §5.4 as well as the high rate of genotypic substitution observed in the adaptive walks of this chapter (Figure 6.14(c)).

The second type of neutrality is in the mapping from phenotype to fitness. Considering for a moment just the spatially constrained phenotype definition: a mutation which swaps the identities of two incorrect terminal cells in such a way that they are still incorrect will produce a novel phenotype without any change in fitness. The adaptive walks revealed that phenotypes were frequently substituted while on a plateau of neutral fitness (Figure 6.14(a) and (d)). For example, during one long period of stasis (approximately steps 1200–1800) there was considerable neutral substitution until, around step 1800 the neutral plateau was escaped and a burst of novel phenotypic substitution ensued, resulting in further fitness increases. Two interpretations of this dynamic are possible. First, the neutrality may have been beneficial, as it allowed search to continue where it would otherwise have become trapped at a local optima. Second, the neutrality may have been a hindrance, introducing a long period of drift where a more rapid transition to a more fit phenotype could otherwise have been achieved. Distinguishing between these

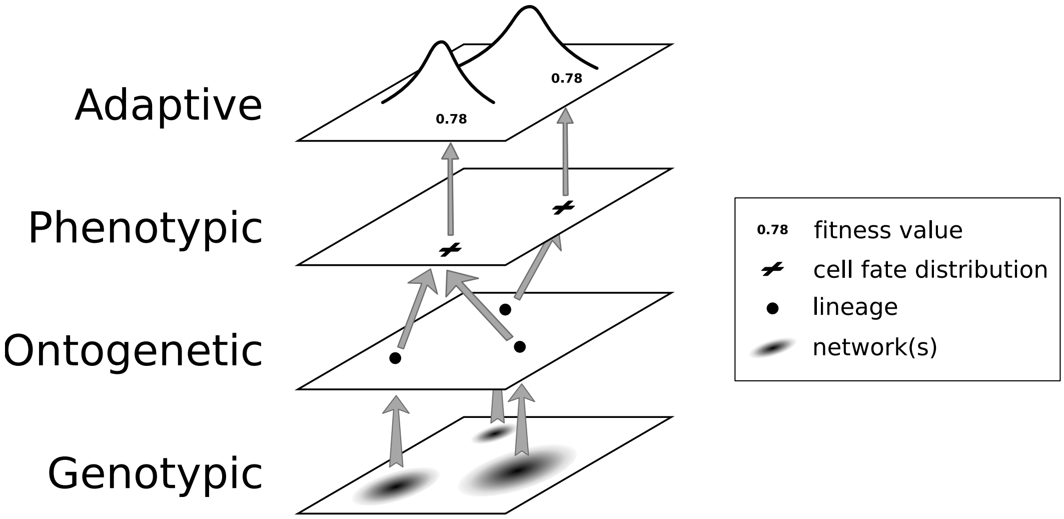


Figure 6.18: Summary of different types of neutrality affecting adaptive search. Many different genotypes map to a single ontogeny. More than one ontogeny may map to a given phenotype; however, these equivalent ontogenies do not appear to be accessible via mutation in ontogenetic space. Finally, multiple phenotypes have equivalent fitness values.

two possibilities is difficult, as it implies a comparison with a search landscape that lacks neutrality, but is otherwise identical.

A third type of neutrality—in the mapping from ontogeny to phenotype—is known to be possible, depending on the phenotype definition. In Series B, under all but the strictest set of phenotypic constraints, multiple lineages were located that mapped to the target phenotype. In practice, none of the adaptive walks were observed to exploit this form of neutrality. One possible explanation for this is that these neutral lineages are located at some distance from one another with respect to genotypic space, such that they are not mutationally accessible to one another. Figure 6.18 summarises the different types of neutrality that were observed or inferred from the adaptive walks.

Phenotypic improvements occur across range of scales

Analysis of the accepted mutations over the adaptive walk shown in Figure 6.14 revealed that mutations can cause phenotypic improvement across a wide range of scales. At the lower end of the spectrum were those mutations that modified the identity of a single terminal cell, and those that added or removed a single

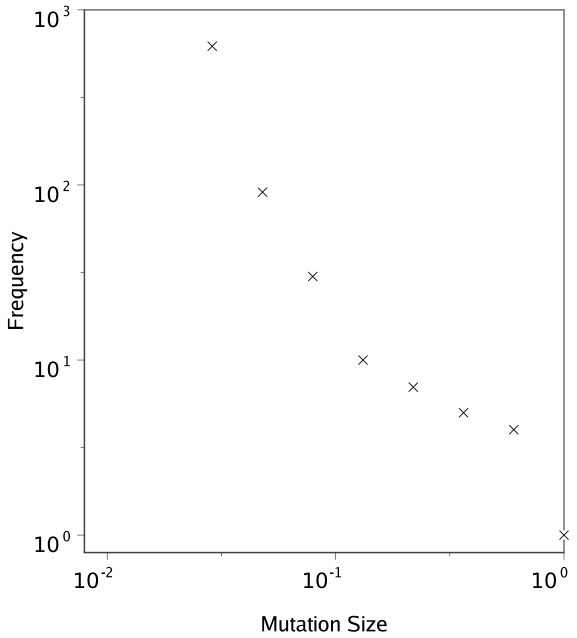


Figure 6.19: The distribution of phenotypic improvements indicates that beneficial mutations can occur across a range of scales. Although the majority of accepted mutations resulted in small phenotypic changes, some mutations of larger effect were also accepted. The size of a mutation was measured as the distance between the initial and mutant lineages for each of the accepted mutations in an adaptive walk. Mutations are sorted into exponentially scaled bins. The fit of the distribution to a power-law with exponent 1.56 was $R^2 = 0.92$.

cell. At the upper end of the spectrum were those mutations that introduced or removed a new cell type, and those that added or removed an entire branch of the cell lineage. The size of a phenotypic improvement was estimated by applying the fitness function using the pre-mutation lineage as the current solution and the post-mutation lineage as the target solution. The sizes of such changes follow a power law distribution (Figure 6.19).

The scale of evolutionary change is a subject of ongoing debate in evolutionary biology (Leroi, 2000). The essence of the debate concerns how to explain the evolution of species as inferred from the fossil record: is the selection of individual mutations a sufficient mechanism, or are higher-level evolutionary forces necessary? One context in which this debate arose was the argument by Fisher (1930) that mutations of large effect would be far less likely to be beneficial, and hence only mutations of small effect were likely to be significant. Kimura (1983) challenged this claim, pointing out that if very rare large beneficial mutations *did* occur, they would be more likely to be fixed, and hence the distribution of mutation sizes would be skewed. More recently, Orr (1998) extended Kimura's model to consider the distribution of mutations fixed on an adaptive walk towards an optimum. He predicted a negative exponential distribution, in which many mutations of small effect were fixed, but so too were a small number of larger mutations. The distribution observed in Figure 6.19 supports the claim that mutations causing both large and small phenotypic changes will occur in an adaptive walk.

Figure 6.19 also highlights one of the benefits of the gene network approach to modelling ontogeny. If the cell lineage representation had been modified directly by the adaptive process, we would have needed to specify the sizes and types of mutations that were possible (*e.g.*, swapping sublineages, adding and deleting terminals, etc.). As it is, we did not need to impose a preconceived step size on the adaptive process—it emerged naturally as a consequence of the dynamic mapping.

Implications for explaining cell lineage complexity

Azevedo et al. (2005) suggest that (a) the apparently complex cell lineages of organisms such as *C. elegans* are actually simpler than they appear; (b) this simplicity may be a product of selection for faster development or for more efficient genetic encoding; and (c) these cell lineages are almost as simple as they could be given the requirements of precisely positioning cells in a developing embryo.

The adaptive walks reported in this chapter suggest another possible explanation for the simplicity of observed cell lineages: it is a side-effect of the dynamics of the gene networks that control their development. With selection for a spatial distribution of cell fates, but not for either increasing or decreasing complexity, adaptation consistently located cell lineages of an equivalent level of complexity as those observed in nature. As observed by Azevedo et al. (2005), this level is consistently lower than that of random cell lineages with the same cell fate distribution (Figure 6.17(b)). Furthermore, we showed that if selection for the correct spatial distribution of cell fates was removed and the only requirement on networks was to generate lineages containing the correct complement of cells, the evolved lineages were more complex—but still simpler than random lineages generated under the same conditions (Figure 6.17(a)).

Chapter 7

General Discussion

This thesis has introduced a model of ontogeny and used it to explore the hypothesis that biases in the structure of phenotypic variation can affect the orientation of an adaptive process. In particular this work has focused on the role that the intrinsic dynamics of a gene network controller play in shaping the structure of variation. This chapter reviews the results presented in this thesis and summarises their implications for understanding the role of developmental bias. The computational modelling methodology used to obtain these results is evaluated and avenues for future research are described.

7.1 Summary and review

The dynamics of the gene regulatory system play a central role in the development of an organism. Alterations to a cell's identity and behaviour during development follow from changes at the level of gene expression. The network architecture of the regulatory system suggests explanations for several observed properties of development, including its robustness and the discreteness of differentiated cell types, which can be likened to the attractive states in network dynamics. In addition, phenotypic evolution occurs via modification to the gene regulatory systems that govern development. The question motivating this thesis was whether bias due to the dynamic nature of developmental control affects the direction of evolution?

Addressing questions in evolutionary developmental biology is challenging due to both the complexity of the systems involved and the limited availability of suitable empirical data. Computational modelling—the methodology used in this thesis—is a useful means of exploring the theoretical issues arising from the interaction between evolution and development. A novel network-lineage model of development was designed (§3.3), based on a cell lineage representation of ontogeny. The primary advantages of this model were: it provided a concise representation of both a final phenotype and its developmental history; it could be measured and compared in a quantitative fashion; it was evolvable; and it was computationally efficient to simulate. The capabilities and limitations of the network-lineage model are discussed further below (§7.3).

The first three studies (reported in Chapter 4) considered the gene network component of this model. Tools from dynamical systems were applied to characterise the space of dynamic behaviours of the DRGN model and how these behaviours depended on macro-level network parameters: size, connectivity and weight scale. Results indicated that the probability of a network containing: (a) point attractors was high when weights were very small, but low otherwise; (b) cyclic attractors increased with increasing weight scale; and (c) chaotic attractors was greatest for large, highly connected networks with intermediate weight scales. Furthermore, the number of stable attractors in a network increased only very slowly with the size of the network, but increased more rapidly as the connectivity of large networks was reduced. The relationship between dynamics and micro-level structural features was explored. The primary finding of these studies was that the dynamic behaviour of networks combines a high level of structural and dynamic stability with a small but significant potential for sensitivity to perturbation.

The scope of investigation was then expanded to consider the ontogenies that are generated by network dynamics (Chapter 5). A novel complexity metric for classifying cell lineages, weighted algorithmic complexity, was introduced. Weighted algorithmic complexity refines existing measures of cell lineage complexity (Braun et al., 2003, Azevedo et al., 2005), matching intuitive conceptions of complexity over a wider class of lineage structures. *TreeView*, a novel software tool for visualising ontogenetic space, was also introduced. The qualitative picture produced by *TreeView*, combined with quantitative results obtained using network ensembles, provided several insights into the structure of ontogenetic space:

- Complex ontogenies are distributed in a nonuniform fashion: the most complex ontogenies tend to be densely clustered in a region around the phase transition between proliferating and quiescent lineages;

- Ontogenies can be separated into two distinct classes of control—trivial, containing proliferating and quiescent lineages, and non-trivial—on the basis of their complexity. These classes form two separate peaks in the distribution of lineage complexities, the width and height of which vary depending on genotypic and ontogenetic parameters;
- A combination of the weight scale and division threshold parameters (W and λ) defines the behaviour of lineage complexity: when both parameters are low, the result is proliferating lineages; when both parameters are high, the result is quiescent lineages; complex lineages emerge at an intermediate point between these two extremes, the exact location of which depends on the structure of the network.

Having demonstrated that the network-lineage model was capable of generating plausible ontogenies, and that the distribution of phenotypes produced by these ontogenies *did* display a characteristic structure, we then used adaptive search to investigate the possible effects of this bias (Chapter 6). *Study 8: Evaluation of mutation bias* used random walks to demonstrate the potential for a second source of bias: the choice of mutation operator. The results of this study indicate that while mutation bias and developmental bias are related, in that the biased structure of genotypic variation introduced by mutation will be further transformed by developmental bias on the production of phenotypic variation, they can each be varied and measured independently. In addition, it was found that those mutation operators that most biased the structure of genotypic variation tended to drive random walks towards lower complexity regions of space.

Study 9: The evolution of ontogenies used three series of adaptive walks in which the mutation operator, the fitness function and the phenotypic target were varied. It was in these simulations that the network-lineage model could be grounded in biological data, through the use of real cell lineages (up to 55 cells) as phenotypic targets. The first series of walks determined that weight replacement was the most effective mutation operator for the type of adaptive task under investigation. The second series of walks demonstrated a relationship between increasing levels of constraint in the definition of the phenotypic target and search difficulty. The final series demonstrated a relationship between increasing the complexity of the phenotypic target and search difficulty. The results of these simulations demonstrated the capacity of the network model to generate cell lineages matching those observed in organisms such as *C. elegans* and *H. roretzi.* In addition, the type of cell lineage structures that were found shed light onto a possible effect of developmental bias: Evolved lineages were substantially less complex than random lineages that generate the same distribution of cell fates. Azevedo et al. (2005) had previously observed that the complexity of lineages observed in nature was substantially less than that of random lineages, given the constraints on the spatial distribution of terminal cells. The results of these studies suggest that a possible explanation for lineage simplicity may be bias due to the intrinsic dynamics of the developmental gene regulatory system.

7.2 The origins and implications of bias

Four questions were identified in Chapter 1 as necessary steps toward an understanding of the role of developmental bias. The final question concerned the methodology and will be discussed in the following section 7.3. The first three questions concerned the dynamics of a developmental control network, the nature of ontogenies controlled by this network and the evolution of development via modification to this network.

How does cell fate potential vary with structural properties of an underlying genetic control system?

The repertoire of dynamic behaviours of the network model used in this thesis varied with each of the number of the nodes in the network and their connectivity, both in terms of the number of inputs each received and also the strength of those inputs. When the strength of interactions was large, the system approximated a Boolean network, and the dynamic behaviours observed were comparable with previous results in this area (Kauffman, 1969, 1993, Bagley and Glass, 1996), both with respect to the number of stable attractors and the absence of deterministic chaos. Lowering the strength of interactions produced a simultaneous decrease in both the number of stable attractors and their level of stability (as estimated by their Lyapunov exponent), and an increase in the probability of observing deterministic chaos. In this region, the networks displayed robustness and flexibility. In general, their behaviour was stable to small dynamic and structural perturbations. Less frequently, such perturbations would either shift the location of a trajectory in dynamic space, or alter the structure of the dynamic space itself, producing a change in long term term behaviour. The combination of robustness and flexibility has been recognised as an important characteristic of biological systems (Csete and Doyle, 2002, Wuensche, 2002): they must be stable enough to buffer temporary environmental fluctuations but also able to adapt dynamically should circumstances demand.

One example of a biological system balancing robustness and flexibility is the Heat-shock protein 90 (Hsp90), which has been termed a "capacitor of phenotypic variation" (Rutherford and Lindquist, 1998, Quietsch et al., 2002). Under normal conditions, Hsp90 is abundant in *Drosophila melanogaster* and interacts with a wide variety of different signalling pathways. When Hsp90 is perturbed by mutation under experimental conditions, a wide range of morphological disturbances are observed. A similar phenomena is observed in the plant *Arabidopsis thaliana*. Quietsch et al. (2002) hypothesise that Hsp90 may act as a global buffer on developmental stability, allowing the accumulation of neutral mutations that, under extreme conditions, may be released, producing a burst of novel variation that may contain phenotypes better suited to the changed environment.

A second example of such a balance was described in Chapter 2. During development, cells undergo a process of differentiation into one of many possible types. Early in this process, they are highly responsive to contextual signals and will adopt the fate of their neighbours if transplanted; later in development, they are robust to signals and will maintain their original fate irrespective of their context (Wolpert, 1998). The underlying mechanisms involved in Hsp90 buffering and the maintenance of cell identity are considerably different. However, the variation of dynamic behaviours observed when interaction strength was scaled (Chapter 4) suggests a common way of conceptualising the global change in network dynamics that may occur under certain situations.

How is the space of possible ontogenies shaped by the dynamic properties of the genetic control system?

The class of ontogenies generated from network dynamics is characterised by a quasi-systematic structure: both the structure of a lineage, generated by the pattern of cell divisions, and the distribution of cell fates across its terminal nodes, generated by the pattern of differentiation, display a certain level of regularity.

Specific regularities that were observed include:

Translational symmetry: a cell divided to produce two daughter cells with identical potentials, that is, two cells giving rise to identical sublineages. When all cells divided to produce two daughter cells with the same potential as the parent cell, this led to proliferation. However, it was also possible that the first cell division (for example) would produce two daughter cells, with the same potential as each other, that would go on to produce non-homogeneous sublineages, in which case the entire cell lineage would display translational symmetry without being homogeneous.

Recursive production: a cell divided to produce one daughter cell with the same potential as its parent and one with different potential. It was commonly observed that either the left or right cell in a lineage would continually divide, while the other differentiated, producing a pattern analogous to the stem-cell mode of cell division.

Modularity: Identical sublineages could appear at multiple locations in a cell lineage, suggesting that the cells producing these sublineages share a common potential. A further implication of this phenomenon is that a particular cell fate potential can be achieved via multiple developmental trajectories, since each cell in a lineage has received a unique sequence of inputs.

Such regularities have also been recognised in biological lineages (Sulston et al., 1983, Kenyon, 1985). The complexity metrics described by Braun et al. (2003) and Azevedo et al. (2005) are, in fact, measures of such regularity.

In this thesis, we chose to define the structure of ontogenetic space in terms of the underlying genotypic and ontogenetic parameters. As described above, the complexity of cell lineages observed in a particular region of space was to a considerable extent a property of the strength of interactions in the generating network (W) and the parameter controlling the rate at which the cell division threshold was scaled (λ). As these two parameters are scaled, the generated cell lineages change from homogeneous cell proliferation, through a complex region, to virtual quiescence (Figure 5.17). The idea of complexity being a threshold phenomena—of life existing ‘on the edge of chaos’—is widespread in the domain of complex systems (originally proposed by Langton, 1990). The observed transition shows some similarities with this phenomena, in that complex behaviour exists in a region between two quite different, but essentially uninteresting, forms of simpler behaviour.

The major implication of the quasi-systematic structure of cell lineages generated from network dynamics is that not all cell lineages are equally likely to appear. A random sampling of genotypic space does not produce a uniform distribution of either ontogenies (as measured by complexity) or phenotypes (Figures 5.7–5.11). Comparison with samples of stochastically generated lineages indicates that this non-uniformity is a feature specifically of ontogenies generated from network dynamics, rather than a general feature of the cell lineage representation (Figures 5.12 and 5.13).

What effect does the structure of generated variation have on the evolution of ontogenies?

The structure of variation has the potential to influence the direction of evolution because it constitutes the raw material on which natural selection acts. If more mutant phenotypes display, for example, trait X than have trait Y, but there is otherwise no selective advantage to either trait, then purely by chance, more individuals are likely to appear with trait X, despite the fact that it has not been actively selected for. The non-uniform phenotypic distribution observed in the network-lineage model suggests that there may be a base level of bias in the space from which variation was drawn. The adaptive walks reported in Chapter 6 confirmed that this bias could affect the outcome of an evolutionary process. Despite there being no explicit selection for simpler rather than more complex lineages, adaptive walks consistently found ways of generating ordered cell fate distributions that were simpler than those generated by random lineages. The intrinsic bias of dynamic networks towards simpler, more regular lineages results in the evolution of ontogenetic simplicity.

One of the dominant features of the adaptive space defined by the network-lineage model is the multiple levels at which neutrality occurs (Figure 6.18). Neutrality has been observed in a variety of natural and artificial search spaces (Kimura, 1983, Huynen et al., 1996, Barnett, 1998, Newman and Engelhardt, 1998, Geard et al., 2002). Neutral search spaces are characterised by plateaus or networks of genotypes that map to equal (or nearly equal) fitness values. In some instances, this neutrality can enable evolving individuals or populations to escape from local optima in the landscape, by moving across these neutral regions until a ‘portal’ to a higher fitness network is located (van Nimwegen and Crutchfield, 2000, Shipman et al., 2000, Barnett, 2001). In other situations however, neutrality may have either no effect on search performance (Smith et al., 2001) or be detrimental (Bullock, 2002).

In their simulation models of the artificial evolution of the *C. elegans* cell lineage towards less complex configurations, Azevedo et al. (2005) observed that, beyond a certain point, evolution was unable to locate any simpler lineages, even though they were known to exist. They attribute this decrease in evolvability to the mutational inaccessibility of the simplest lineages. An additional reason why these simpler lineages may become increasingly more difficult for evolution to locate is the robustness of simple lineages. The flip-side of structural robustness is phenotypic variability: every mutation (structural perturbation) that results in a new gene network producing the same phenotype as the old gene network (*i.e.*, a neutral mutation) acts to decrease the amount of phenotypic variability available for selection to act on. As evolution moves into less complex regions of ontogenetic space, the proportion of these neutral mutations increases, reducing the proportion of mutations that will result in simpler lineages.

7.3 An evaluation of the methodology

This section discusses the strengths and limitations of the methodology that emerged from the studies reported in this thesis, both of computational modelling in general and the specific model employed here.

Strengths of the model

The computational modelling methodology used in this thesis proved to be well suited to the task of generating insights into complex relationships between gene regulation, development and evolution. One important aspect of the model was the ability to quantify many of the concepts in the domain, such as robustness, complexity and bias, that are typically discussed in more qualitative terms. Their description in concrete and measurable terms provides a platform for future development.

The network model was able to display a wide variety of dynamic behaviours consistent with previous observations for this class of recurrent neural networks. The results reported in this thesis demonstrate that, given a relatively simple mapping between dynamics and developmental events, these networks are capable of generating a diverse range of ontogenetic patterns. Furthermore, the network model proved to be a highly evolvable representation of developmental control that was able to generate biologically realistic ontogenies. Using adaptive walks, networks were identified that could generate spatial cell fate distributions up to 55 cells, containing 7 different cell types.

The lineage model was an intuitively accessible model of biological development. In comparison with more common morphological models of development (reviewed in Chapter 3), the organisational representation of cell lineages had several advantages:

- The entire history of development was contained in the representation, and hence accessible for comparison and measurement, rather than just the final phenotype;
- It was possible to quantify, not only in terms of phenotypic features, such as cell type, position and number, but also in terms of developmental features, such as complexity;
- Defining distance metrics for comparing different cell lineages was straightforward;
- Cell lineages are widely used as representations of development in biology, therefore it was possible to directly compare between the results of computational experiments and real data sets; and
- Cell lineages can be simulated without implementing physical or mechanical aspects of development, allowing for more efficient computation.

Limitations on the capabilities of the model

Lineages with more than 55 terminal cells were attempted, but the adaptive process tended to stall at an early stage. There are two possible explanations for the inability of the system to scale to larger phenotypes. One explanation is that the networks used are not capable of the level of control required to generate larger lineages. Theoretical results suggest that recurrent neural networks (with appropriate architectures) can be computationally equivalent to Turing machines; therefore, their formal computational power should not be the limiting factor (Siegelmann and Sontag, 1991). However, the studies reported in Chapter 4 indicated that as the size of a DRGN increases so does the sensitivity of its dynamics to the precise pattern of interactions between its regulatory nodes. That is, the set of network structures that will produce a desired behaviour becomes smaller. The second (related) explanation is that the adaptive process used to evolve networks may be responsible: capable networks may exist somewhere in genotypic space, but they are not being found. There are reasons to expect that both of these factors play a role. Preliminary trials carried out for the Series C simulations indicated that, for the larger phenotypic targets, increasing the number of regulatory nodes did improve performance. This improvement suggests that larger networks may well be capable of more complex control tasks. However, comparing the results of Series B and Series C, the larger networks found the smaller targets less frequently, suggesting some cost to increasing the size of the network. One explanation for this cost is that, as the number of nodes and interactions in a network increases, so does the number of free variables in the system, and hence the size and dimensionality of the space that an adaptive process has to explore. It appears that there is a trade-off between the control capability of the network and the complexity of the adaptive space.

Considering the adaptive tasks from a biological perspective, certain limitations of the methodology are exposed. Firstly, the definition of an adaptive task in terms of a single, fixed, optimum phenotype is an extremely simplistic view of evolution. The precise structure of evolutionary landscapes is still a topic of debate (see, *e.g.*, Gavrilets, 2002, Skipper, 2004). However, they are likely to be dynamic, with the mapping from phenotype to fitness shifting over time in response to changing environmental conditions. The succession of evolutionary changes from single celled bacteria to *C. elegans* was not *directed* towards that end. Rather, over evolutionary history, the size and complexity of organisms gradually increased via a series of modifications—some larger, some smaller—all of which built incrementally on a succession of viable, functioning platforms. In contrast, our adaptive process had no concept of any intermediate targets, and so correspondingly less information to guide the search process. Secondly, and more importantly, the complexity of the gene regulatory network controlling development has not remained constant throughout evolutionary history (Bonner, 1988, Zuckerkandl, 2001). In our simulations, adaptation occurred via modifications to the interaction strengths of a fixed network topology. However, the topology of biological gene networks is far from being fixed, especially over macroevolutionary timescales. Evidence suggests that gene duplication has been an important mechanism in the growth and evolution of gene networks (Wagner, 1994, Teichmann and Babu, 2004). Therefore, early in the adaptive process, the space of possible solutions is relatively small. At some point, after adaptation has produced organisms well suited to the current ecological niches, a growth in the size and complexity of the genotype may introduce new dimensions into the adaptive space that can be used to increase the complexity of the phenotype. The key issue is that the control requirements for highly complex phenotypes would be met through the adaptation of genetic networks already capable of generating phenotypes of a lower degree of complexity.

A dynamic approach to the complexity of solution representations has been explored in several contexts. Stanley (2004) developed NEAT (the NeuroEvolution of Augmenting Topologies) and demonstrated that a process of *complexification*—building up a solution's complexity throughout evolution—was able to evolve more sophisticated behaviours than static topologies. Watson (2002) considered an alternative mechanism by which hierarchical levels of complexity could be introduced into a population, namely via symbiosis, and demonstrated that the use of such a mechanism enabled populations to evolve solutions to a particular class of hierarchically decomposable problems. Thus, we do not feel that the limitations on adaptation observed in these studies are a reflection of any inherent constraints on the ability of the control system. Rather, they suggest that the adaptive processes used could benefit from the incorporation of more sophisticated models of evolution.

Limitations on the generality of the results

The most obvious caveat that accompanies any results obtained using a simulation model is that their value depends on the validity of the underlying model. To reduce the chance of results being artifacts of model design, the network-lineage model was designed to be as simple as was feasible, given the nature of the research questions. In particular, three of the abstractions chosen may limit the generality of the results obtained:

- The emphasis of the network-lineage model is on genetic, as opposed to epigenetic (or environmental), factors in the control of development. As reviewed in Chapter 2, there are strong arguments for the central importance of the genetic regulatory system in the control and evolution development (Carroll et al., 2001, Davidson, 2001). There is also a corresponding case for the importance of non-genetic factors (Nijhout, 1990, Müller and Newman, 2003, Minelli, 2003, West-Eberhard, 2003). Most researchers accept a role for both components, with some disagreement as to their relative importance. At this stage, the mechanisms of epigenetic control and epigenetic inheritance are not as well understood as their genetic counterparts, and there is little research into how epigenetic components of development should be modelled.

- The network-lineage model omits any description of morphological aspects of development. As described above, this abstraction brings many benefits for implementing and analysing the model. However, it sacrifices much of the richness of development that emerges from properties of the physical substrate in which the entire process is embedded. Newman and Müller (2000) have argued that it is physical processes in morphogenesis that play the primary role in generating morphological novelty, with genetic change playing a consolidating, rather than an innovating, role. Models that incorporate more realistic physical components have demonstrated that the mechanical aspects of morphogenesis are capable of generating complex phenotypic structures even under conditions of minimal genetic control (Rudge and Geard, 2005, Rudge and Haseloff, 2005). One perspective on the model presented in this thesis is that it illustrates possible limitations of purely genetic control. In order to improve evolvability on more complex phenotypic targets it may be necessary to incorporate a richer physical model.

- The network-lineage model was explored using adaptive walks as a model of evolution. Again, this had the advantage of computational efficiency, and was sufficient for investigating the effects of the structure of variation. However, it also imposed a limitation on the extent to which the interaction between bias and selection could be investigated. The simulations in this thesis have considered how developmental bias affects the type of ontogenies found to achieve a particular phenotypic target. While there were many different ontogenies capable of performing this task, the adaptive walks tended to find some of them—those favoured by developmental bias—more frequently than others. In this situation, there was no selective difference between any of the candidate ontogenies, bias merely guided adaptation towards one out of several equally fit solutions. An open question is the extent to which development could bias the direction of evolution *against* the direction of an adaptive gradient (Amundson, 1994). That is, could bias produce adaptation to a less fit solution than would be possible in the absence of that bias? Alternatively, can natural selection 'break' the constraints imposed by developmental bias? (Beldade et al., 2002) Such questions raise the issue of the relative strength of bias and selection in a natural or artificial context—an issue which will require population-based models of evolution to resolve.

7.4 Further work

Incorporating additional levels of biological detail, such as a physical model of morphogenesis or a population model of evolution, could broaden the generality of the results described here and provide a deeper understanding of the role that development plays in orienting evolution. Extending the network-lineage model to address the limitations noted above suggests three possible directions for future work in this area:

- The genotypic component of the model (the DRGN) could be extended to incorporate post-transcriptional and epigenetic mechanisms, such regulatory control by noncoding RNA (Mattick, 2004) and chromatin remodelling (Li, 2002). Doing so would enable a more complete model of the role that non-genetic mechanisms play in controlling development.

- The phenotypic component of the model (the cell lineage) could be embedded in a simulation environment incorporating a richer physical and mechanical dimension. Models capable of simulating complex morphological processes have been implemented (*e.g.*, Cickovski et al., 2005); however, these lack an explicit representation of the organisational aspects of ontogeny such as are represented in a cell lineage. Integrating both an organisational *and* a morphological view in a common framework could provide a deeper insight into the relationship between developmental and structural complexity than is possible with either model alone.

- The adaptive walk methodology used in Chapter 6 could be replaced with a population-based evolutionary model. Doing so would enable a more comprehensive investigation of the balance between the relative strength of selection and bias.

7.5 Conclusions

This thesis has used computational models of gene regulation, development and evolution to address a fundamental question in evolutionary developmental biology: how bias due to development can affect the direction of evolution.

Networks of interconnected elements, such as the gene regulatory systems modelled in this thesis, are capable of a wide range of dynamic behaviours. In the context of network-lineage model, these dynamic behaviours predisposed the generation of cell lineages with a simpler, more regular structure than stochastic lineages producing identical cell fate distributions. In an adaptive context, this predisposition affected the structure of selectable phenotypic variation that was generated, with the outcome that ontogenies found by adaptive search were biased towards simplicity. The results obtained in this thesis suggest a possible explanation for the level of complexity observed in the cell lineages of biological organisms: that it may be a product of bias resulting from the network architecture of developmental control.

By quantifying complexity, variation and bias, the network-lineage model described in this thesis provides a computational method for investigating the effects of development on the direction of evolution. In doing so, it establishes a viable framework for simulating computational aspects of complex biological systems.

References

Adami, C. (2002). What is complexity? *BioEssays*, 24:1085–1094.

Albers, D. J. (2004). *A Qualitative Numerical Study of High Dimensional Dynamical Systems.* PhD thesis, University of Wisconsin-Madison.

Albers, D. J., Sprott, J. C., and Dechert., W. D. (1998). Routes to chaos in neural networks with random weights. *International Journal of Bifurcation and Chaos*, 8(7):1463–1478.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957.

Albert, R. and Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila. Journal of Theoretical Biology*, 223(1):1–18.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1994). *Molecular Biology of the Cell.* Garland Publishing Inc., New York, NY, 3rd edition.

Alon, U., Surette, M. G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, 397:168–171.

Altman, R. B., Dunker, A. K., Hunter, L., and Klein, T. E., editors (1998). *Pacific Symposium on Biocomputing '98*, Singapore. World Scientific.

Altman, R. B., Lauderdale, K., Dunker, A. K., Hunter, L., and Klein, T. E., editors (1999). *Pacific Symposium on Biocomputing '99*, Singapore. World Scientific.

Amundson, R. (1994). Two concepts of constraint: Adaptationism and the challenge from developmental biology. *Philosophy of Science*, 61(4):556–578.

Amundson, R. (2005). *The Changing Role of the Embryo in Evolutionary Thought.* Cambridge University Press, Cambridge, NY.

Arnone, M. I. and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124:1851–1864.

Arthur, W. (1984). *Mechanisms of Morphological Evolution.* Wiley, New York.

Arthur, W. (1999). The pattern of variation in centipede segment number as an example of developmental constraint in evolution. *Journal of Theoretical Biology*, 200:183–191.

Arthur, W. (2000). The concept of developmental re-programming and the quest for an inclusive theory of evolutionary mechanism. *Evolution & Development*, 2:49–57.

Arthur, W. (2002a). The emerging conceptual framework of evolutionary developmental biology. *Nature*, 415:757–764.

Arthur, W. (2002b). The interaction between developmental bias and natural selection: from centipede segments to a general hypothesis. *Nature Heredity*, 89:239–246.

Arthur, W. (2003). Developmental constraint and natural selection. *Evolution & Development*, 5(2):117–118.

Arthur, W. (2004a). *Biased Embryos and Evolution.* Cambridge University Press, Cambridge.

Arthur, W. (2004b). The effect of development on the direction of evolution: toward a twenty-first centure consensus. *Evolution & Development*, 6:282–288.

Azevedo, R. B. R., Lohaus, R., Braun, V., Gumbel, M., Umamaheshwar, M., Agapow, P. M., Houthoofd, W., Platzer, U., Borgonie, G., Meinzer, H. P., and Leroi, A. M. (2005). The simplicity of metazoan cell lineages. *Nature*, 433:152–156.

Bäck, T., Fogel, D., and Michalewicz, Z. (1997). *Handbook of Evolutionary Computation.* Oxford University Press, Oxford.

Badii, R. and Politi, A. (1997). *Complexity – Hierarchical Structures and Scaling in Physics.* Cambridge University Press, Cambridge.

Bagley, R. J. and Glass, L. (1996). Counting and classifying attractors in high dimensional dynamical systems. *Journal of Theoretical Biology*, 183:269–284.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–114.

Barkai, N. and Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, 387:913–917.

Barnett, L. (1998). Ruggedness and neutrality – the NKp family of fitness landscapes. In Adami, C., Belew, R. K., Kitano, H., and Taylor, C. E., editors, *Artificial Life VI: Proceedings of the Sixth International Conference on Artificial Life*, pages 18–27, Cambridge, MA. The MIT Press/Bradford Books.

Barnett, L. (2001). Netcrawling – optimal evolutionary search with neutral networks. In *Proceedings of the 2001 Congress on Evolutionary Computation (CEC2002)*, pages 27–31, Piscataway, NJ. IEEE Press.

Bedau, M., McCaskill, J., Packard, N., and Rasmussen, S., editors (2000). *Artificial Life VII: Proceedings of the Seventh International Conference on Artificial Life*, Cambridge, MA. The MIT Press/Bradford Books.

Beer, R. D. (1995). On the dynamics of small continuous-time recurrent networks. *Adaptive Behaviour*, 3:471–511.

Beldade, P. and Brakefield, P. M. (2003). The difficulty of agreeing about constraints. *Evolution & Development*, 5(2):119–120.

Beldade, P., Koops, K., and Brakefield, P. M. (2002). Developmental constraints versus flexibility in morphological evolution. *Nature*, 416:844–847.

Bentley, P. J., editor (1999). *Creative Evolutionary Design.* Morgan Kaufman Publishers Inc., San Francisco, CA.

Bentley, P. J. (2003). Evolving fractal proteins. In Tyrrell, A. M., Haddow, P. C., and Torresen, J., editors, *Evolvable Systems: From Biology to Hardware, 5th International Conference (ICES 2003)*, volume 2606 of *Lecture Notes in Computer Science*, pages 81–92, Berlin. Springer.

Bentley, P. J. and Kumar, S. (1999). Three ways to grow designs: A comparison of embryogenies for an evolutionary design problem. In Banzhaf, W., Daida, J., Eiben, A. E., Garzon, M. H., Honavar, V., Jakiela, M., and Smith, R. E., editors, *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 25–43, Orlando, FL. Morgan Kaufmann.

Bolouri, H. and Davidson, E. H. (2002). Modeling transcriptional regulatory networks. *BioEssays*, 24:1118–1129.

Bolouri, H. and Davidson, E. H. (2003). Transcriptional regulatory cascades in development: Initial rates, not steady state, determine network kinetics. *Proceedings of the National Academy of Science, USA*, 100(16):9371–9376.

Bongard, J. and Pfeifer, R. (2003). Evolving complete agents using artificial ontogeny. In Hara, F. and Pfeifer, R., editors, *Morpho-functional Machines: The New Species (Designing Embodied Intelligence)*, pages 237–258. Springer-Verlag, Berlin.

Bonner, J. T., editor (1981). *Evolution and Development.* Springer-Verlag, Berlin.

Bonner, J. T. (1988). *The Evolution of Complexity.* Princeton University Press, Princeton, New Jersey.

Bornholdt, S. (2001). Modeling genetic networks and their evolution: A complex dynamical systems perspective. *Biological Chemistry*, 382:1289–1299.

Bornholdt, S. and Sneppen, K. (1998). Neutral mutations and punctuated equilibrium in evolving genetic networks. *Physical Review Letters*, 81(1):236–239.

Bowerman, B. (1998). Maternal control of pattern formation in early *Caenorhabditis elegans* embryos. *Current Topics in Developmental Biology*, 39:73–117.

Braun, V., Azevedo, R. B. R., Gumbel, M., Agapow, P. M., Leroi, A. M., and Meinzer, H. P. (2003). ALES: cell lineage analysis and mapping of developmental events. *Bioinformatics*, 19:851–858.

Bray, D. (1995). Protein molecules as computational elements in living cells. *Nature*, 376:307–312.

Brenner, S. (1999). Theoretical biology in the third millennium. *Philosophical Transactions of the Royal Society of London, Series B*, 354:1963–1965.

Britten, R. J. and Davidson, E. H. (1969). Gene regulation for higher cells: A theory. *Science*, 165:349–357.

Bryant, P., Brown, R., and Arbarbanel, H. D. I. (1990). Lyapunov exponents from observed time series. *Physical Review Letters*, 65(13):1523–1526.

Bullock, S. (1997). *Evolutionary Simulation Models: On their character, and application to problems concerning the evolution of natural signalling systems.* PhD thesis, University of Sussex.

Bullock, S. (1999). Are artificial mutation biases unnatural? In Floreano et al. (1999), pages 64–73.

Bullock, S. (2001). Smooth operator? understanding and visualising mutation bias. In Kelemen and Sosik (2001), pages 64–73.

Bullock, S. (2002). Will selection for mutational robustness significantly retard evolutionary innovation on neutral networks. In Standish, R., Bedau, M. A., and Abbass, H. A., editors, *Artificial Life VIII: Proceedings of the Eigth International Confereonce on the Synthesis and Simulation of Living Systems*, pages 192–201, Cambridge, MA. MITB.

Carroll, S. B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, 101:577–580.

Carroll, S. B., Grenier, J. K., and Weatherbee, S. D. (2001). *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. Blackwell Science, Oxford.

Chen, T., He, H. L., and Church, G. M. (1999). Modeling gene expression with differential equations. In Altman et al. (1999).

Cheverud, J. M. (1984). Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology*, 110:155–171.

Cickovski, T. M., Huang, C., Chaturvedi, R., Glimm, T., Hentschel, H. G. E., Alber, M. S., Glazier, J. A., Newman, S. A., and Izaguirre, J. A. (2005). A framework for three-dimensional simulation of morphogenesis. *IEEE/ACM Transactions on Computational Biology and Bioninformatics*, 2(4):273–288.

Crampin, E. J., Halstead, M., Hunter, P., Nielsen, P., Noble, D., Smith, N., and Tawhai, M. (2004). Computational physiology and the physiome project. *Experimental Physiology: Translation and Integration*, 89(1):1–26.

Csete, M. E. and Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, 295:1664–1669.

Davidson, E. H. (2001). *Genomic Regulatory Systems*. Academic Press, San Diego, CA.

Davidson, E. H., McClay, D. R., and Hood, L. (2003). Regulatory gene networks and the properties of the developmental process. *Proceedings of the National Academy of Science, USA*, 100(4):1475–1480.

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103.

Debat, V. and David, P. (2001). Mapping phenotypes: canalization, plasticity and developmental stability. *Trends in Ecology and Evolution*, 16(10):555–561.

Dechert, W. D. and Gencay, R. (1992). Lyapunov exponents as a nonparametric diagnostic for stability analysis. *Journal of Applied Econometrics*, 7:S41–S60.

Dellaert, F. and Beer, R. D. (1996). A developmental model for the evolution of complete autonomous agents. In Maes, P., Mataric, M. J., Meyer, J.-A., Pollack, J., and Wilson, S. W., editors, *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 393–401, Cambridge, MA. The MIT Press/Bradford Books.

Di Paolo, E. A., Noble, J., and Bullock, S. (2000). Simulation models as opaque thought experiments. In Bedau et al. (2000), pages 497–506.

Edmonds, B. (1999). *Syntactic Measures of Complexity.* PhD thesis, University of Manchester.

Eggenberger, P. (1997). Evolving morpohologies of simulated 3D organisms based on diffferential gene expression. In Husbands and Harvey (1997), pages 205–213.

Eggenberger, P. (2003). Genome-physics interaction as a new concept to reduce the number of genetic parameters in artificial evolution. In Sarker, R., Reynolds, R., Abbass, H., Tan, K.-C., McKay, R., Essam, D., and Gedeon, T., editors, *Proceedings of the IEEE 2003 Congress on Evolutionary Computation*, pages 191–198, Piscataway, NJ. IEEE Press.

Eldredge, N. (1985). *Unfinished Synthesis.* Oxford University Press, New York, NY.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Endy, D. and Brent, R. (2001). Modelling cellular behaviour. *Nature*, 409:391–395.

Espinosa-Soto, C., Padilla-Longoria, P., and Alvarez-Buylla, E. R. (2004). A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*, 16(11):2923–2939.

Felsenstein, J. (2004). *Inferring Phylogenies.* Sinauer Associates, Sunderland, MA.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

Flake, G. W. (2000). *The Computational Beauty of Nature.* The MIT Press, Cambridge, MA.

Fleischer, K. (1996). Investigations with a multicellular developmental model. In Langton, C. G. and Shimohara, T., editors, *Artificial Life V: Proceedings of the Fifth International Conference on the Synthesis and Simulation of Living Systems*, pages 229–236, Cambridge, MA. The MIT Press/Bradford Books.

Fleischer, K. and Barr, A. H. (1994). A simulation testbed for the study of multicellular development: The multiple mechanisms of morphogenesis. In Langton, C., editor, *Artificial Life III*, pages 389–416, Redwood City, CA. Addison-Wesley.

Floreano, D., Nicoud, J.-D., and Mondada, F., editors (1999). *Advances in Artificial Life: 5th European Conference, ECAL '99*, volume 1674 of *Lecture Notes in Artificial Intelligence*, Berlin. Springer-Verlag.

Freeman, M. (2000). Feedback control of intercellular signalling in development. *Nature*, 408:313–319.

Furusawa, C. and Kaneko, K. (1998). Emergence of rules in cell society: differentiation, hierarchy and stability. *Bulletin of Mathematical Biology*, 60:659–687.

Fusco, G. (2001). How many processes are responsible for phenotypic evolution. *Evolution & Development*, 3(4):279–286.

García-Bellido, A. (1985). Cell lineages and genes. *Philosophical Transactions of the Royal Society of London, Series B*, 312:101–128.

Gavrilets, S. (2002). Evolution and speciation in hyperspace: the roles of neutrality, selection, mutation and random drift. In Crutchfield, J. P. and Schuster, P., editors, *Evolutionary Dynamics - Exploring the Interplay of Selection, Accident, Neutrality, and Function*, Santa Fe Institute Studies on the Science of Complexity, chapter 7. Oxford University Press, New York, NY.

Geard, N., Hallinan, J., Tonkes, B., Skellett, B., and Wiles, J. (2002). A comparison of neutral landscapes - nk, nkp and nkq. In Fogel, D. B., El-Sharkawi, M. A., Yao, X., Greenwood, G., Iba, H., Marrow, P., and Shackleton, M., editors, *Proceedings of the 2002 Congress on Evolutionary Computation (CEC2002)*, pages 205–210, Piscataway, NJ. IEEE Press.

Geard, N. and Wiles, J. (2003). Structure and dyanmics of a gene network model with RNA regulation. In Sarker, R., Reynolds, R., Abbass, H., Tan, K.-C., McKay, R., Essam, D., and Gedeon, T., editors, *Proceedings of the 2003 Congress on Evolutionary Computation (CEC2003)*, pages 199–206, Piscataway, NJ. IEEE Press.

Geard, N. and Wiles, J. (2005). A gene network model for developing cell lineages. *Artificial Life*, 11(3):249–268.

Geard, N. and Wiles, J. (2006). Investigating ontogenetic space with developmental cell lineages. In Rocha, L. M., Yaeger, L. S., Bedau, M. A., Floreano, D., Goldstone, R. L., and Vespignani, A., editors, *Artificial Life X: Proceedings of the Tenth International Confereonce on the Synthesis and Simulation of Living Systems*, pages 56–62, Cambridge, MA. The MIT Press/Bradford Books.

Gibson, G. (2002). Developmental evolution: getting robust about robustness. *Current Biology*, 12:R347–R349.

Gierer, A. and Meinhardt, H. (1972). A theory of biological pattern formation. *Kybernetik*, 12:30–39.

Gilbert, S. F. (2003). *Developmental Biology*. Sinauer Associates, Sunderland, MA, 7th edition.

Gilbert, S. F., Opitz, J. M., and Raff, R. A. (1996). Resynthesizing evolutionary and developmental biology. *Developmental Biology*, 173:357–372.

Glass, L. and Kauffman, S. A. (1973). The logical analysis of continuous non-linear biochemical control networks. *Journal of Theoretical Biology*, 39:103–129.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, MA.

Goodwin, B. C. (1994). *How the Leopard Changed Its Spots: The Evolution of Complexity.* Princeton University Press, Princeton, NJ.

Goodwin, B. C., Holder, N., and Wylie, C. C., editors (1982). *Development and Evolution.* Cambrudge University Press, Cambridge.

Goutsias, J. and Kim, S. (2004). A nonlinear discrete dynamical model for transcriptional regulation: construction and properties. *Biophysical Journal*, 86(4):1922–1945.

Greenspan, R. J. (2001). The flexible genome. *Nature Reviews Genetics*, 2:383–387.

Gruau, F. (1995). Automatic definition of modular neural networks. *Adaptive Behaviour*, 3(2):151–183.

Hall, B. K. (1999). *Evolutionary Developmental Biology.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 2nd edition.

Hartl, D. L. (2000). *A Primer of Population Genetics.* Sinauer, Sunderland, MA, 3rd edition.

Hartman IV, J. L., Garvik, B., and Hartwell, L. (2001). Principles for the buffering of genetic variation. *Science*, 291:1001–1004.

Harvey, I. and Bossamaier, T. (1997). Time out of joint: attractors in asynchronous random boolean networks. In Husbands and Harvey (1997).

Harvey, I. and Thompson, A. (1996). Through the labyrinth evolution finds a way: A silicon ridge. In Higuchi, T., Iwata, M., and Wexlin, L., editors, *First International Conference on Evolvable Systems*, pages 406–422, Berlin. Springer-Verlag.

Hasty, J., McMillen, D., Isaacs, F., and Collins, J. J. (2001). Computational studies of gene regulatory networks: *in numero* molecular biology. *Nature Reviews Genetics*, 2:268–279.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation.* Prentice Hall, Upper Saddle River, NJ, 2nd edition.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* Addison Wesley, Redwood City, CA.

Hogeweg, P. (2000a). Evolving mechanisms of morphogenesis: on the interplay between differential cell adhesion and cell differentiation. *Journal of Theoretical Biology*, 203:317–333.

Hogeweg, P. (2000b). Shapes in the shadow: Evolutionary dynamics of morphogenesis. *Artificial Life*, 6:85–101.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.* University of Michigan Press, Ann Arbor, MI.

Holland, J. H. (1998). *Emergence: From Chaos To Order.* Addison-Wesley, Redwood City, CA.

Hornby, G. S. and Pollack, J. B. (2002). Creating high-level components with a generative representation for body-brain evolution. *Artificial Life*, 8(3):223–246.

Houthoofd, W., Jacobsen, K., Mertens, C., Vangestel, S., Coomans, A., and Borgonie, G. (2003). Embryonic cell lineage of the marine nematode *Pellioditis marina*. *Developmental Biology*, 258:57–69.

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. E. (2005). Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical Review Letters*, 94:128701.

Huang, S. and Ingber, D. E. (2000). Shape-dependent control of cell growth, differentiation, and apoptosis: Switching between attractors in cell regulatory networks. *Experimental Cell Research*, 261:91–103.

Hucka, M., Finney, A., Bornstein, B. J., Keating, S. M., Shapiro, B. E., Matthews, J., Kovitza, B. L., Schilstra, M. J., Funahashi, A., Doyle, J. C., and Kitano, H. (2004). Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. *Systems Biology*, 1(1):41–53.

Husbands, P. and Harvey, I., editors (1997). *Fourth European Conference on Artificial Life*, Cambridge, MA. The MIT Press/Bradford Books.

Huynen, M. A., Stadler, P. F., and Fontana, W. (1996). Smoothness within ruggedness: the role of neutrality in adaptation. *Proceedings of the National Academy of Science, USA*, 93:397–401.

Ingber, D. E. (2005). Mechanical control of tissue growth: Function follows form. *Proceedings of the National Academy of Science, USA*, 102(33):11571–11572.

Inoue, T., Wang, M., Ririe, T. O., Fernandes, J. S., and Sternberg, P. W. (2005). Transcriptional network underlying *Caenorhabditis elegans* vulval development. *Proceedings of the National Academy of Science, USA*, 102(14):4972–4977.

Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33:245–254.

Jan, Y. N. and Jan, L. Y. (1998). Asymmetric cell division. *Nature*, 392:775–778.

Kaletta, T., Schnabel, H., and Schnabel, R. (1997). Binary specification of the embryonic lineage in *Caenorhabditis elegans*. *Nature*, 390:294–298.

Kaneko, K. and Yomo, T. (1994). Cell division, differentiation, and dynamic clustering. *Physica D*, 75:89–102.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22:437–467.

Kauffman, S. A. (1971). Gene regulation networks: a theory for their global structure and behaviours. *Current Topics in Developmental Biology*, 6:145–182.

Kauffman, S. A. (1974). The large-scale structure and dynamics of gene control circuits: An ensemble approach. *Journal of Theoretical Biology*, 44:167–190.

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

Kauffman, S. A. (1996). *At Home in the Universe: The Search for Laws of Self-organization and Complexity*. Penguin, London.

Kauffman, S. A. (2004). A proposal for using the ensemble approach to understand genetic regulatory networks. *Journal of Theoretical Biology*, 231(1):581–590.

Kauffman, S. A. and Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128:11–45.

Kelemen, J. and Sosik, P., editors (2001). *Advances in Artificial Life: 6th European Conference, ECAL 2001*, volume 2159 of *Lecture Notes in Computer Science*, Berlin. Springer-Verlag.

Keller, E. F. (2003). *Making Sense of Life : Explaining Biological Development with Models, Metaphors, and Machines*. Harvard University Press, Cambridge, MA.

Kenyon, C. (1985). Cell lineage and the control of *Caenorhabditis elegans* development. *Philosophical Transactions of the Royal Society of London, Series B*, 312:21–38.

Keränen, S. V. E. (2004). Simulation study on effects of signaling network structure on the developmental increase in complexity. *Journal of Theoretical Biology*, 231:3–21.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Kitano, H. (1990). Designing neural networks using genetic algorithms with graph generation system. *Complex Systems*, 4(4):461–476.

Kitano, H. (2001). *Foundations of Systems Biology.* The MIT Press/Bradford Books, Cambridge, MA.

Kitano, H. (2002a). Computational systems biology. *Nature*, 420:206–210.

Kitano, H. (2002b). Systems biology: a brief overview. *Science*, 295:1662–1664.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5:826–837.

Ko, M. (1991). A stochastic model for gene induction. *Journal of Theoretical Biology*, 153:181–194.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, Cambridge, MA.

Krakauer, D. C. and Plotkin, J. B. (2002). Redundancy, antiredundancy, and the robustness of genomes. *Proceedings of the National Academy of Science, USA*, 99(3):1405–1409.

Kumar, S. and Bentley, P. J. (2003). Biologically inspired evolutionary design. In Tyrrell, A. M., Haddow, P. C., and Torresen, J., editors, *Proceedings of the International Conference on Evolvable Systems 2003*, pages 57–68, Berlin. Springer.

Kumar, S. and Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Science, USA*, 99(2):803–808.

Langton, C. (1990). Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D*, 42:12–37.

Lenski, R. E. and Travisano, M. (1994). Dynamics of adaptation and diversification: a 10,000 generation experiment with bacterial populations. *Proceedings of the National Academy of Science, USA*, 91:6808–6814.

Leroi, A. M. (2000). The scale independence of evolution. *Evolution & Development*, 2(2):67–77.

Lewontin, R. C. (1970). The units of selection. *Annual Review of Ecology and Systematics*, 1:1–18.

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics*, 3:662–673.

Lin, R., Hill, R. J., and Priess, J. R. (1998). POP-1 and anterior-posterior fate decisions in *C. elegans* embryos. *Cell*, 92:229–239.

Lindenmayer, A. (1968). Mathematical models for cellular interaction in development: parts I and II. *Journal of Theoretical Biology*, 18:280–315.

Maduro, M. F. (2002). Making worm guts: The gene regulatory network of the *Caenorhabditis elegans* endoderm. *Developmental Biology*, 246:68–85.

Marnellos, G. and Mjolsness, E. (1998). A gene network approach to modeling early neurogenesis in *Drosophila*. In Altman et al. (1998), pages 30–41.

Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. *EMBO reports*, 2(11):986–991.

Mattick, J. S. (2004). RNA regulation: a new genetics? *Nature Reviews Genetics*, 5:316–323.

Mattick, J. S. and Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Molecular Biology and Evolution*, 18(9):1611–1630.

Maynard Smith, J., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., Lande, R., Raup, D., and Wolpert, L. (1985). Developmental constraints and evolution: A perspective from the Mountain Lake Conference on Development and Evolution. *Quarterly Review of Biology*, 60(3):265–287.

Mayr, E. (1991). *One Long Argument: Charles Darwin and the Genesis of Modern Evolutionary Thought*. Harvard University Press, Cambridge, MA.

Mayr, E. (2001). *What Evolution Is*. Basic Books, New York, NY.

Mayr, E. and Provine, W. B., editors (1980). *The Evolutinoary Synthesis: Perspectives on the Unification of Biology*. Harvard University Press, Cambridge, MA.

McAdams, H. H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Science, USA*, 94:814–819.

McAdams, H. H. and Shapiro, L. (1995). Circuit simulation of genetic networks. *Science*, 269:650–656.

McShea, D. W. (1996). Metazoan complexity and evolution: is there a trend? *Evolution*, 50(2):477–492.

McShea, D. W. (2001). The minor transitions in hierarchical evolution and the question of a directional bias. *Journal of Evolutionary Biology*, 14:502–518.

McShea, D. W. (2005). The evolution of complexity without natural selection, a possible large-scale trend of the fourth kind. *Paleobiology*, 31(2, Supplement):146–156.

Meiklejohn, C. D. and Hartl, D. L. (2002). A single mode of canalization. *Trends in Ecology and Evolution*, 16(10):468–473.

Meinhardt, H. (1982). *Models of Biological Pattern Formation.* Academic Press, London.

Meir, E., von Dassow, G., and Munro, E. (2002). Robustness, flexibility, and the role of lateral inhibition in the neurogenic network. *Current Biology*, 12:778–786.

Mendoza, L., Thieffry, D., and Alvarez-Buylla, E. R. (1999). Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. *Bioinformatics*, 15(7):593–606.

Miller, G. F. (1995). Artificial life as theoretical biology: How to do real science with computer simulation. Cognitive Science Research Paper 378, School of Cognitive and Computing Sciences, University of Sussex.

Minelli, A. (2003). *The Development of Animal Form: Ontogeny, Morphology and Evolution.* Cambridge University Press, Cambridge.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms.* The MIT Press/Bradford Books, Cambridge, MA.

Mjolsness, E., Sharp, D. H., and Reinitz, J. (1991). A connectionist model of development. *Journal of Theoretical Biology*, 152:429–453.

Mochizuki, A. (2005). An analytical study of the number of steady states in gene regulatory networks. *Journal of Theoretical Biology*, 236:291–310.

Molin, L., Schnabel, H., Kaletta, T., Feichtinger, R., Hope, I. A., and Schnabel, R. (1999). Complexity of developmental control: Analysis of embryonic cell lineage specification in *Caenorhabditis elegans* using *pes-1* as an early marker. *Genetics*, 151:131–141.

Müller, G. B. and Newman, S. A. (2003). *Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology.* MIT Press, Cambridge, MA.

Newman, M. E. J. and Engelhardt, R. (1998). Effects of neutral selection on the evolution of molecular species. *Proceedings of the Royal Society of London, Series B*, 256:1333–1338.

Newman, S. A. and Müller, G. B. (2000). Epigenetic mechanisms of character origination. *Journal of Experimental Zoology (Molecular and Developmental Evolution)*, 288:304–317.

Nijhout, H. F. (1990). Metaphors and the role of genes in development. *BioEssays*, 12(9):441–446.

Nishida, H. (1987). Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme. III. Up to the tissue restricted stage. *Developmental Biology*, 121:526–541.

Ohta, T. (2002). Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy of Science, USA*, 99(25):16134–16137.

Oliveri, P. and Davidson, E. H. (2004). Gene regulatory network controlling embryonic specication in the sea urchin. *Current Opinion in Genetics & Development*, 14:351–360.

Orphanides, G. and Reinberg, D. (2002). A unified theory of gene expression. *Cell*, 108:439–451.

Orr, H. A. (1998). The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution*, 52(4):935–949.

Pasemann, F. (1995). Characterization of periodic attractors in neural ring networks. *Neural Networks*, 8(3):421–429.

Pasemann, F. (2002). Complex dynamics and the structure of small neural networks. *Network: Computation in Neural Systems*, 13:195–216.

Peck, S. L. (2004). Simulation as experiment: a philosophical reassessment for biological modeling. *Trends in Ecology and Evolution*, 19(10):530–534.

Pires-daSilva, A. and Sommer, R. J. (2003). The evolution of signalling pathways in animal development. *Nature Reviews Genetics*, 4:39–49.

Pollack, J. B. and Blair, A. D. (1998). Co-evolution in the successful learning of a backgammon strategy. *Machine Learning*, 32(3):225–240.

Prusinkiewicz, P. (2004). Modeling plant growth and development. *Current Opinion in Plant Biology*, 7:79–83.

Ptashne, M. (1992). *A Genetic Switch: Phage Lambda and Higher Organisms.* Blackwell Science, Oxford.

Quayle, A. P. and Bullock, S. (2005). Modelling the evolution of genetic regulatory networks. *Journal of Theoretical Biology*, 238(4):737–753.

Quietsch, C., Sangster, T. A., and Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, 417:618–624.

Raff, R. A. (2000). Evo-devo: the evolution of a discipline. *Nature Reviews Genetics*, 1:74–79.

Raff, R. A. and Kaufman, T. C. (1983). *Embryos, Genes and Evolution: The Developmental-Genetic Basis of Evolutionary Change.* Macmillan, London.

Reil, T. (1999). Dynamics of gene expression in an artificial genome - implications for biological and artificial ontogeny. In Floreano et al. (1999), pages 457–466.

Reinitz, J., Mjolsness, E., and Sharp, D. H. (1995). Model for cooperative control of positional information in *Drosophila* by *bicoid* and *hunchback*. *Journal of Experimental Zoology (Molecular and Developmental Evolution)*, 271:47–56.

Reinitz, J. and Sharp, D. H. (1995). Mechanisms of *eve* strip formation. *Mechanisms of Development*, 49:133–158.

Ridley, M. (1996). *Evolution.* Blackwell Science, Oxford, 2nd edition.

Roggen, D. and Federici, D. (2004). Multi-cellular development: is there scalability and robustness to gain? In Yao, X., Burke, E., Lozano, J., and al., editors, *Proceedings of Parallel Problem Solving from Nature 8 (PPSN 2004)*, volume 3242 of *Lecture Notes in Computer Science*, pages 391–400, Berlin. Springer-Verlag.

Rudge, T. and Geard, N. (2005). Evolving gene regulatory networks for cellular morphogenesis. In Abbass, H. A., Bossamaier, T., and Wiles, J., editors, *Recent Advances in Artificial Life*, volume 3 of *Advances in Natural Computation*, pages 239–252, Singapore. World Scientific Publishing.

Rudge, T. and Haseloff, J. (2005). A computational model of cellular morphogenesis in plants. In Capcarrere, M. S., Freitas, A. A., Bentley, P. J., Johnson, C. G., and Timmis, J., editors, *Advances in Artificial Life: 8th European Conference (ECAL 2005)*, volume 3630 of *Lecture Notes in Artificial Intelligence*, pages 78–87, Berlin. Springer-Verlag.

Rutherford, S. L. (2000). From genotype to phenotype: buffering mechanisms and the storage of genetic information. *Bioessays*, 22:1095–1105.

Rutherford, S. L. and Lindquist, S. (1998). Hsp90 as a capacitor of morphological evolution. *Nature*, 396:336–342.

Salazar-Ciudad, I., Garcia-Fernández, J., and Solé, R. V. (2000). Gene networks capable of pattern formation: From induction to reaction-diffusion. *Journal of Theoretical Biology*, 205:587–603.

Salazar-Ciudad, I. and Jernvall, J. (2004). How different types of pattern formation mechanisms affect the evolution of form and development. *Evolution & Development*, 6(1):6–16.

Salazar-Ciudad, I., Jernvall, J., and Newman, S. A. (2003). Mechanisms of pattern formation in development and evolution. *Development*, 130:2027–2037.

Sánchez, L., van Helden, J., and Thieffry, D. (1997). Establishment of the dorso-ventral pattern during embryonic development of *Drosophila melanogaster*: A logical analysis. *Journal of Theoretical Biology*, 189:377–389.

Sanderson, M. J. (2006). *Paloverde*: An OpenGL 3D phylogeny browser. *Bioinformatics*, 22(8):1004–1006.

Sankoff, D. and Kruskal, J., editors (1983). *Time warps, string edits and macromolecules.* Addison-Wesley, Reading, MA.

Schlichting, C. D. and Pigliucci, M. (1998). *Phenotypic Evolution: A Reaction Norm Perspective.* Sinauer Associates, Sunderland, MA.

Shipman, R., Shackleton, R., Ebner, M., and Watson, R. (2000). Neutral search spaces for artificial evolution: a lesson from life. In Bedau et al. (2000), pages 162–169.

Siegal, M. L. and Bergman, A. (2002). Waddington's canalization revisited: Developmental stability and evolution. *Proceedings of the National Academy of Science, USA*, 99(16):10528–10532.

Siegelmann, H. T. and Sontag, E. D. (1991). Turing computability with neural nets. *Applied Mathematical Letters*, 4:77–80.

Simpson, G. G. (1944). *Tempo and Mode in Evolution.* Columbia University Press, New York, NY.

Skipper, R. A. (2004). The heuristic role of Sewall Wrights 1932 adaptive landscape diagram. *Philosophy of Science*, 71:1176–1188.

Smith, T. M. C., Husbands, P., and O'Shea, M. (2001). Evolvability with complex genotype-phenotype mapping. In Kelemen and Sosik (2001), pages 272–281.

Smolen, P., Baxter, D. A., and Byrne, J. H. (2000a). Mathematical modeling of gene networks. *Neuron*, 26:567–580.

Smolen, P., Baxter, D. A., and Byrne, J. H. (2000b). Modeling transcriptional control in gene networks - methods, recent results, and future directions. *Bulletin of Mathematical Biology*, 62:247–292.

Solé, R., Fernández, P., and Kauffman, S. (2003). Adaptive walks in a gene network model of morphogenesis: insights into the cambrian explosion. *International Journal of Developmental Biology*, 47:693–701.

Solé, R. V. and Goodwin, B. (2000). *Signs of Life: How Complexity Pervades Biology*. Basic Books, New York, NY.

Solé, R. V., Salazar-Ciudad, I., and Garcia-Fernández, J. (2002). Common pattern formation, modularity and phase transitions in a gene network model of morphogenesis. *Physica A*, 305:640–654.

Somogyi, R. and Sniegoski, C. A. (1996). Modelling the complexity of of genetic networks: Understanding multigenic and pleiotropic regulation. *Complexity*, 1:45–63.

Sprott, J. C. (2003). *Chaos and Time-Series Analysis*. Oxford University Press, Oxford.

Stanley, K. O. (2004). *Efficient Evolution of Neural Networks Through Complexification*. PhD thesis, The University of Texas at Austin.

Stanley, K. O. and Miikkulainen, R. (2003). A taxonomy for artificial embryogeny. *Artificial Life*, 9(2):93–130.

Stent, G. (1985). The role of cell lineage in development. *Philosophical Transactions of the Royal Society of London, Series B*, 312:3–19.

Stent, G. (1998). Developmental cell lineage. *International Journal of Developmental Biology*, 42:237–241.

Strogatz, S. (1994). *Nonlinear Dynamics and Chaos*. Addison-Wesley, Reading, MA.

Suen, G. and Jacob, C. (2003). A symbolic and graphical gene regulation model of the *lac* operon. In *Fifth International Mathematica Symposium*, pages 73–80, London, England. Imperial College Press.

Sulston, J. E., Schierenberg, E., White, J. G., and Thompson, J. N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental Biology*, 100:64–119.

Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nature Genetics*, 36:493–496.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, 282:2012–2018.

Thieffry, D. and Thomas, R. (1995). Dynamical behaviour of biological regulatory networks - II: Immunity control in bacteriophage lambda. *Bulletin of Mathematical Biology*, 57:277–297.

Thomas, R. (1973). Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42:563–585.

Thomas, R. (1991). Regulatory networks seen as asynchronous automata: A logical description. *Journal of Theoretical Biology*, 153:1–23.

Thomas, R. (1998). Laws for the dynamics of regulatory networks. *Int. J. Dev. Biol.*, 42:479–485.

Thomas, R. (1999). Deterministic chaos seen in terms of feedback circuits: analysis, synthesis, 'labyrinth chaos'. *International Journal of Bifurcation and Chaos*, 9:1889–1905.

Thomas, R. and Kaufman, M. (2001). Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behaviour. *Chaos*, 11(1):170–179.

Tiňo, P., Horne, B. G., and Giles, C. L. (2001). Attractive periodic sets in discrete time recurrent networks. *Neural Computatation*, 13:1379–1414.

Trooskens, G., Beule, D. D., Decouttere, F., and Criekinge, W. V. (2005). Phylogenetic trees: visualizing, customizing and detecting incongruence. *Bioinformatics*, 21(19):3801–3802.

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London, Series B*, 237:37–72.

Tyson, J. J., Chen, K., and Novak, B. (2001). Network dynamics and cell physiology. *Nature Reviews Molecular Cell Biology*, 2(12):908–916.

Tyson, J. J. and Othmer, H. G. (1978). The dynamics of feedback control circuits in biochemical pathways. *Progress in Theoretical Biology*, 5:1–62.

Valentine, J. W., Collins, A. G., and Meyer, C. P. (1994). Morphological complexity increase in metazoans. *Paleobiology*, 20:131–142.

van Nimwegen, E. and Crutchfield, J. P. (2000). Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths? *Bulletin of Mathematical Biology*, 62(5):799–848.

van Nimwegen, E., Crutchfield, J. P., and Mitchell, M. (1997). Finite populations induce metastability in evolutionary search. *Physics Letters A*, 229:144–150.

van Someren, E. P., Wessels, L. F. A., Backer, E., and Reinders, M. J. T. (2002). Genetic network modeling. *Pharmacogenomics*, 3(4):507–525.

Vancoppenolle, B., Borgonie, G., and Coomans, A. (1999). Generation times of some free-living nematodes cultured at three temperatures. *Nematology*, 1:15–18.

Vohradský, J. (2001a). Neural model of the genetic network. *The Journal of Biological Chemistry*, 276(39):36168–36173.

Vohradský, J. (2001b). Neural network model of gene expression. *The FASEB Journal*, 15:846–854.

von Dassow, G., Meir, E., Munro, E. M., and Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature*, 406:188–192.

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, 150:563–565.

Waddington, C. H. (1953). Genetic assimilation of an acquired character. *Evolution*, 7(2):118–126.

Waddington, C. H. (1957). *The Strategy of the Genes.* George Allen & Unwin Ltd., London.

Waddington, C. H. (1959). Canalization of development and genetic assimilation of acquired characters. *Nature*, 183:1654–1655.

Wagner, A. (1994). Evolution of gene networks by gene duplication: a mathematical model and its implications on genome organization. *Proceedings of the National Academy of Science, USA*, 91:4387–4391.

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

Wagner, G. P. and Altenberg, L. (1996). Complex adaptation and the evolution of evolvability. *Evolution*, 50(3):967–976.

Watson, J., Geard, N., and Wiles, J. (2004). Towards more biological mutation operators in gene regulation studies. *BioSystems*, 113:239–248.

Watson, R. A. (2002). *Compositional Evolution: Interdisciplinary Investigations in Evolvability, Modularity and Symbiosis.* PhD thesis, Brandeis University.

Weaver, D. C., Workman, C. T., and Stormo, G. D. (1999). Modeling regulatory networks with weight matrices. In Altman et al. (1999), pages 112–123.

West-Eberhard, M. J. (2003). *Developmental Plasticity and Evolution*. Oxford University Press, Oxford.

Wilke, C. O. (2001). Adaptive evolution on neutral networks. *Bulletin of Mathematical Biology*, 63:715–730.

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

Wilkins, A. S. (2002). *The Evolution of Developmental Pathways*. Sinauer, Sunderland, MA.

Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985). Determining Lyapunov exponents from a time series. *Physica D*, 16:285–317.

Wolpert, L. (1969). Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology*, 25:1–47.

Wolpert, L. (1998). *The Principles of Development*. Oxford University Press, Oxford.

Wong, A. H. C., Gottesman, I. I., and Petronis, A. (2005). Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human Molecular Genetics*, 14(Spec No 1):R11–R18.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the 6th International Congress on Genetics*, 1:356–366.

Wuensche, A. (1998). Genomic regulation modeled as a network with basins of attraction. In Altman et al. (1998), pages 89–102.

Wuensche, A. (2002). Basins of attraction in network dynamics: a conceptual framework for biomolecular networks. Working paper 02-02-004, Santa Fe Institute.

Yampolsky, L. Y. and Stoltzfus, A. (2001). Bias in the introduction of variation as an orienting factor in evolution. *Evolution & Development*, 3(2):73–83.

Yohn, C. B. (2001). FlyNome: A Database of *Drosophila* nomenclature. Accessed at `http://www.flynome.com` on 9 June, 2006.

Yoshida, H., Furusawa, C., and Kaneko, K. (2005). Selection of initial conditions for recursive production of multicellular organisms. *Journal of Theoretical Biology*, 233:501–514.

Young, N. M. and Hallgrímsson, B. (2005). Serial homology and the evolution of mammalian limb covariation structure. *Evolution*, 59(12):2691–2701.

Yuh, C.-H., Bolouri, H., and Davidson, E. H. (1998). Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene. *Science*, 279:1896–1902.

Zeng, Z.-B. and Cockerham, C. C. (1993). Mutation models and quantitative genetic variation. *Genetics*, 133:729–736.

Zuckerkandl, E. (2001). Intrinsically driven changes in gene interaction complexity. I. Growth of regulatory complexes and increase in number of genes. *Journal of Molecular Evolution*, 53:539–554.

Appendix A

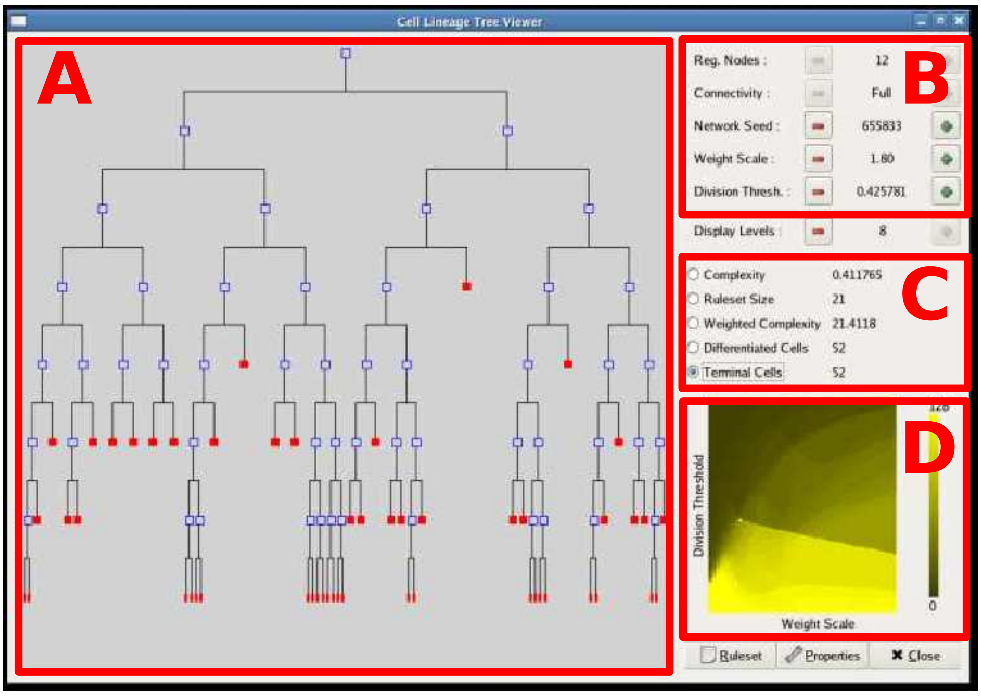
TreeView Technical Details

TreeView is a tool that enables large numbers of cell lineage diagrams to be visualised and explored in a rapid and intuitive fashion. The *TreeView* interface consists of four main components (Figure A.1):

A – Lineage View The largest panel displays the currently selected lineage, as described in Chapter 3. Developmental time runs from top to bottom (root to leaves). Terminal nodes are coloured according to the cell fate—black nodes indicate a cell line that has not yet stopped dividing and hence is currently undifferentiated. Non-terminal nodes may be either blank (as in Figure A.1) or coloured according to the set of potential fates into which they can differentiate (as in the figures in Chapters 5 and 6). The lineage may be scaled to fit within the panel, or shown at full size, in which case scroll bars can be used to navigate the diagram.

B – Parameters The parameters panel lists the parameters of the current lineage, including: the size, connectivity and generating seed of the generating DRGN; and the values of λ and W for the current lineage. The plus/minus buttons allow the size, connectivity or seed of the network to be altered; and the current heatmap to be navigated.

C – Complexity Metrics The complexity measures panel displays the complexity of the current lineage according to the metrics described in Chapter 5. Selecting any of these metrics changes the global heatmap to represent the gradient associated with that metric.



Reg. Nodes : 12

Connectivity : Full

Network Seed : 655833

Weight Scale : 1.80

Division Thresh. : 0.425781

Display Levels : 8

○ Complexity 0.411765

○ Ruleset Size 21

○ Weighted Complexity 21.4118

○ Differentiated Cells 52

● Terminal Cells 52

Figure A.1: Components of the TreeView interface: A – Lineage View. B – Parameters. C – Complexity Metrics. D– Heatmap View.

D – Heatmap View The heatmap view provides a summary of a parameterised slice through ontogenetic space, as described in Chapter 5. Each heatmap displays the possible lineages generated by a single network as weight scale (W) and division threshold (λ) are varied. The colour of each point represents the complexity value of each lineage according to the currently selected metric in the Complexity Metric panel. Selecting any point displays the corresponding cell lineage in the Lineage View panel.