Evolution of gene networks in sex determination

Thomas Anthony MACCARTHY BANOS

University College London

Submitted for PhD in Modelling Biological Complexity

Abstract

In this work, the evolution of sex determination gene networks is investigated using a modelling approach. Recent evidence indicates that an increase in the complexity of interactions has played an important role in gene network evolution. Sex determination mechanisms offer a good model for studying gene network evolution because, among other reasons, they evolve rapidly. In chapter 2, the potential for evolutionary change of the existing *Drosophila* sex determination gene network is considered. With the aid of a synchronous logical model, theoretical concepts such as a network-specific form of mutation are defined, as well as a notion of functional equivalence between networks. Applying this theoretical framework to the sex determination mechanism, it is found that sex determination networks generally exist within large sets of functionally equivalent networks all of which satisfy the sex determination task. These large sets are in turn composed of subsets which are mutationally related, suggesting a high degree of flexibility is available without compromising the core functionality. The technique for finding functional equivalence between networks suggests a general method for gene network reconstruction, which is explored in chapter 3. Lastly, in chapters 4 and 5, a hierarchical model is presented which integrates population genetics techniques with network dynamics. This model consists of a core population genetics simulation within which parameters such as the sex and fitness of the genotype are calculated from the corresponding network dynamics. The model is used to investigate the early evolution of sex determination networks. Following from a hypothesis proposed by Wilkins (1995), the assumption is made that sex determination networks have evolved in a retrograde manner from bottom to top. Starting from the simplest possible ancestral system, based on a single locus, we explore the way in which more complex systems, involving two or three loci, could have evolved.

Acknowledgements

This work was supported by a Medical Research Council Bioinformatics PhD Studentship awarded by the Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London.

I would like to thank both my supervisors, Andrew Pomiankowski at the department of Biology (UCL) and Rob Seymour at the department of Mathematics (UCL), for their encouragement and support throughout. I would also like to thank Anne Warner, head of CoMPLEX, for her help and good advice.

I want to thank my wife Alejandra for her patience and backing, especially in the final months, as well as my parents who have always done their best to motivate me.

Contents

Chapter 1

Introduction

1.1 Biological networks

1.1.1 Networks and complexity

The study of networks pervades all of science, from neurobiology to statistical physics. In fields related to biology, studies of neural networks, ecological food webs, cell-signalling, gene and biochemical networks have all attracted interest. Beyond biology, a multitude of real-world networks such as electrical power grids, the Internet and World-Wide Web and social networks (for example, the overlapping boards of directors of the largest companies in the United States), have also been analysed. Why is network anatomy so important to characterize? Because structure always affects function. For instance, the topology of social networks affects the spread of information and disease, and the topology of the power grid affects the robustness and stability of power transmission. From this perspective, the current interest in networks is part of a broader movement towards research on complex systems. Networks are inherently difficult to understand, as the following list of possible complications illustrates (from [1]):

1. Structural complexity: the wiring diagram could be an intricate tangle.

2. Network evolution: the wiring diagram could change over time. On the World Wide Web, pages and links are created and lost every minute.

3. Connection diversity: the links between nodes could have different weights, directions and signs. Synapses in the nervous system can be strong or weak, inhibitory or excitatory.

4. Dynamical complexity: the nodes could be nonlinear dynamical systems. In a gene network or a Josephson junction array, the state of each node can vary in time in complicated ways.

5. Node diversity: there could be many different kinds of nodes. The regulatory network that controls cell division in yeast consists of a bewildering number of genes and protein complexes.

6. Meta–complication: the various complications can influence each other. For example, the present layout of a power grid depends on how it has grown over the years – a case where network evolution (2) affects topology (1). When coupled neurons fire together repeatedly, the connection between them is strengthened; this is the basis of memory and learning. Here nodal dynamics (4) affect connection weights (3).

1.1.2 Gene networks

This work will be concerned specifically with transcription-regulatory, or gene networks. The elucidation of gene interactions at the molecular level began in 1960 with the work of two French biologists, Francois Jacob and Jacques Monod, for which they were awarded the Nobel prize [2]. Their research focussed on the response of the gut bacterium *E. coli* to lactose, specifically the method by which the enzyme (β–galactosidase, used to digest lactose) was synthesized in response to a presence of lactose at sufficient concentration in the medium. The mechanism by which the enzyme is induced was found to occur at the level of transcription (the process by which messenger RNA is produced from DNA) of the β–galactosidase gene. It was discovered that adjacent to the gene there is a short sequence of nucleotides in the DNA to which a protein binds. The sequence is called an operator, and the protein that binds to it is called the repressor, since its action is to repress transcription of the β- galactosidase gene. When the repressor is bound to the operator site, no mRNA is produced, and as a consequence, no β–galactosidase enzyme (the protein product of the gene) is produced either. This regulatory mechanism is coupled to the presence of lactose in the environment by the simple fact that lactose binds to the repressor, changing its shape in such a way that it can no longer bind to the operator site. Since the operator site is free, production of β-galactosidase occurs soon afterwards. Because the repressor molecule is itself a product of another *E.coli* gene, the discovery that genes could be "switched on" in this way led immediately to the suggestion that genes might form networks in which genes turn each other on and off.

In the simplest case, two genes might each repress the other. If gene A represses gene B and gene B represses gene A, then such a system might form what is known as a bistable pattern of activity. In the first pattern, gene A would be on and repress gene B; in the second, gene B would be on and repress gene A. Indeed, this simple genetic circuit (a genetic toggle switch) has been constructed using plasmids in *E.coli* [3]. Other simple circuits have also been produced: autoregulatory systems [4], oscillators [5] and logic gates [6], and have established gene network engineering as a discipline in its own right [7].

In spite of a dearth of experimental data, the discoveries of Jacob and Monod led rapidly to attempts at theoretical models of gene networks at both (a) a *macroscopic* level [8], in which the ensemble behaviour of large networks is considered, and (b) a *microscopic* level [9], in which the behaviour of individual networks is considered. To some extent, this division between microscopic vs. macroscopic continues today, though the nascent field of systems biology [10] aims to bridge this gap by using the priciples of systems theory to model and test large (experimentally determined) biological systems. Arguably the most successful of the macroscopic approaches so far is the discovery of patterns in topological features of biological networks such as the scale-free property and "network motifs", which is addressed next. This is followed by an overview of microscopic approaches using modelling.

1.1.3 Topological features

The development of high-throughput data-collection techniques such as microarrays and the yeast two-hybrid method, has unravelled various types of biological networks such as protein-protein, metabolic, signalling and gene networks. An intense field of research has emerged recently in which the topology of such complex networks is analysed using various network measures, such as the degree distribution, which allow comparison and characterisation of different complex networks [11]. Such analysis is made possible by transforming each network into a graph. For example, in a gene network, each node might represent a different gene with the edges representing regulatory interactions. Standard mathematical analysis from graph theory can then be used on such graphs [12]. The most elementary characteristic of a node in a graph is its degree, or connectivity, k, which is the number of connections the node has. If we are dealing with a directed graph (a graph in which each connection is directed, usually shown as an arrow), then we can measure two distinct values, one for incoming links (indegree k_{in}), the other for outgoing links (outdegree k_{out}). Figure 1.1a shows an undirected graph in which the node A has degree $k = 5$, and figure 1.1b shows a directed graph in which the node A has indegree $k_{in} = 4$, and outdegree $k_{out} = 1$.

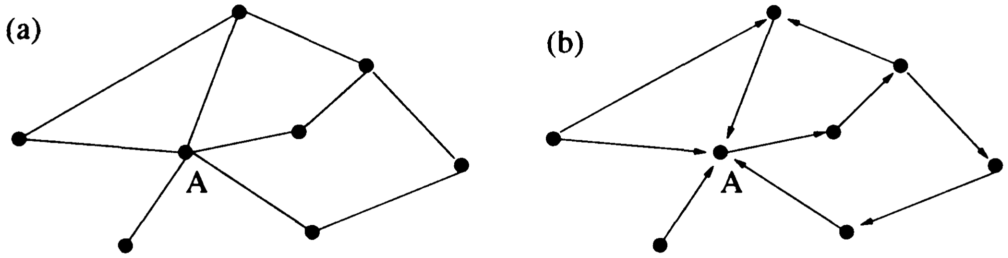


Figure 1.1: (a) In this undirected network, node A has degree $k = 5$. (b) In a directed form of the network, one can further measure indegree $k_{in} = 4$, and $k_{out} = 1$ for node A.

The degree distribution, p_k gives the probability that a selected node has exactly k links, and is obtained by counting the number of nodes n_k with $k = 1, 2, \ldots$ links and dividing by the total number of nodes N (i.e. $p_k = n_k/N$). A particular degree distribution which appears to be common in biological networks is the scale-free distribution, $p_k = k^{-\gamma}$ (γ constant). Figure 1.2 shows the degree distribution for the yeast protein-protein interaction network, which was found to be scale-free [13]. On a log-log plot such as figure 1.2, this relationship should be linear, though for finite networks such as this one, the relationship will often deviate for lower values of k. The protein-protein interaction data for this study was gathered from two data sources [14, 15]. Although this data derives mostly from two-hybrid analyses, a technique known to give many false positives [16], it is assumed in the study that systematic techniques such as this one give a better representation of the underlying degree distribution than would be the case if only evidence from hypothesis-driven experiments were being used.

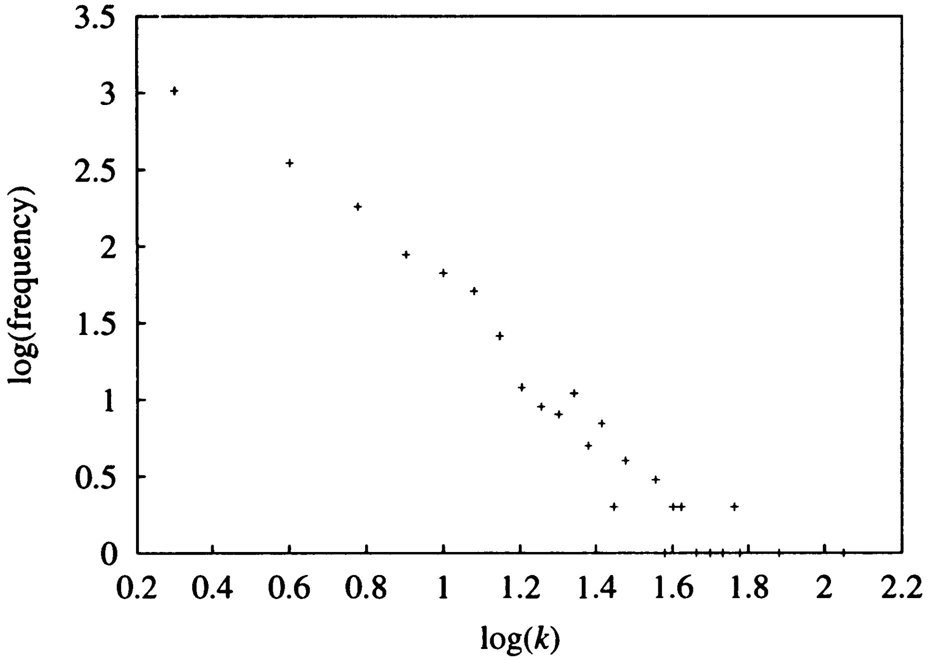


Figure 1.2: Log-log (base 10) plot for the degree distribution of the yeast protein-protein interaction network as used in [13]. Data set available from the URL (`http://www.nd.edu/~networks/database/protein/bo.dat.gz`). Note a scale-free distribution appears to be approximately linear on the log-log plot.

One important characteristic of the scale-free distribution is the existence of a small number of highly-connected "hubs" for which k is large. Scale-free degree distributions have also been observed in metabolic networks [17] and, at least for outdegree (k_{out}), in gene networks [18]. The preferential attachment hypothesis [19] suggests that two requirements are sufficient for scale-free networks to evolve: (1) the network should grow, and (2) new nodes are attached to the existing nodes with probability proportional to the degree k of the existing nodes (i.e. "hubs" are more likely to recruit the new connections, making them more connected still).

A second approach looks at topological features in an attempt to uncover local-scale design principles. "Network motifs" [20] are patterns of interconnections that are found in the gene network of *E. coli* at numbers significantly higher than those in randomized networks. Further studies [21, 22] have extended this analysis to the yeast gene network, as well as a variety of other networks (neural, ecological and electronic). Figure 1.3 shows the 13 possible connected directed subgraphs, or "network motifs", with three nodes and no self-interactions.

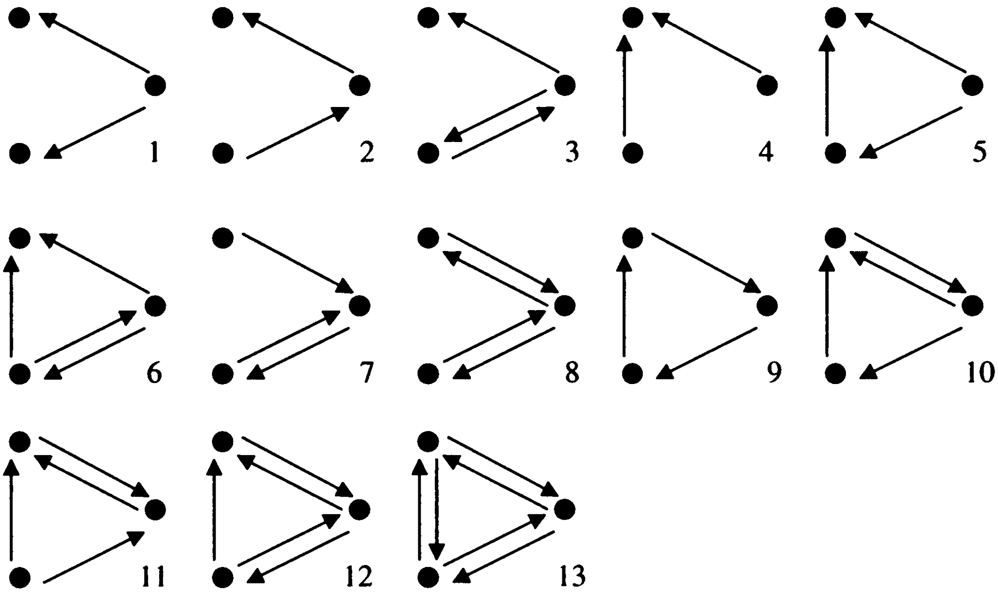


Figure 1.3: All 13 types of three-node connected subgraphs, or "network motifs" (from [21]). Number 5 is the feedforward loop.

In particular, motif number 5 (the "feedforward loop") has been found to occur in far higher numbers in both gene and neural networks (though for the latter there may be a fairly trivial explanation [23]) than in comparable randomized networks [21]. The dynamic properties of the feedforward loop have since been analysed both theoretically [24] and *in vivo* [25], with the finding that this structure acts to delay input activation but not deactivation, a behaviour that is termed "sign-sensitive delay". The three-node motifs (figure 1.3) were subsequently used to define profiles, or "superfamilies", of over(and under)-represented motifs [26], again discovering fundamental similarities in apparently disparate networks. Interestingly, a related bioinformatics study [27] found that, at least for gene networks, over-represented motifs are a consequence of convergent evolution rather than gene duplication (i.e. these structures evolved repeatedly from scratch rather than through repeated duplication of an original template).

Many of the "network motifs" studies have used interaction databases which rely heavily on hypothesis-driven experimental results, as reported in the literature. Such results may in turn be biased by the type of research undertaken. Therefore, as the number of reported interactions approaches the full number (version 4.0 of RegulonDB, an *E. coli* database, contains an estimated ~20-25% of all interactions [28]), there will be a need for such studies to be updated. One such update [29] for the transcription regulatory network of *E. coli* has indeed found qualitatively similar results to the original study [20]. For example, the updated study found that the "feedforward loop" is still highly represented in the updated network.

Biological networks are abstract representations of complex biological systems, necessarily capturing essential characteristics only. Furthermore, since real data have only recently become available, the field is still in an early phase of development. Network descriptions such as the scale-free exponent and the motif "superfamily" profile are likely to become useful tools for assessing network function, particularly in view of the abundance of data produced by novel high-throughput techniques. These first attempts at analysing biological networks have (reasonably enough) used existing tools borrowed from graph theory and physics, which often use crude representations, such as directed graphs. Obviously, one way in which progress will be made in the field, is in the form of greater detail [30]: representing the strength and sign (repressing or activating?) of the interactions, or the functional form of the nodes (do they represent AND and OR type logic operations?), and so on. These details are not necessarily easy to discover experimentally and, at least in the short-term, theorists will often need to work with incomplete data. The principal shortcoming of topological feature analysis is a static view of the network which does not take into account dynamical behaviour. Significant progress will surely come from integrating these two. Already, one bioinformatics-oriented approach has been made which integrates degree-distribution analysis with dynamical data [31], though modelling-oriented approaches are likely to play an important role in the future, as the techniques improve.

1.1.4 Modelling approaches

The development of mathematical and computational techniques to model and analyse gene networks is an integral part of understanding the complexity of gene networks [32]. Spurred on by a vast increase in computational power and an improved mathematical understanding of the mechanisms involved [33], gene network models have advanced significantly since the early attempts in the 1970s. Moving beyond techniques which deal merely with network structure (such as the degree distribution), a wide range of mathematical formalisms have been used to describe dynamical behaviour in gene networks [34]. At the simpler extreme are logic models [8, 35] in which the expression level of a gene is a simple binary ON/OFF value. At the other extreme, are models which represent more fully the system dynamics, such as those based on nonlinear Ordinary Differential Equations (ODEs) [3, 7], delay differential equations [36], and stochastic algorithms [37]. Specifically relevant to the field of developmental biology, network models are now capable of reproducing the development of complex structures such as somites [38], insect abdominal segments [39], and bristle patterns [40].

One topic which has recently become a popular target for investigation using models is the robustness of gene networks, understood here as the resistance of a biological system to noise and/or mutations [41]. Robustness is in turn related to the concept of canalization introduced in the 1950s by Waddington [42]. One approach to this problem [39, 40] defines a system of ODEs integrated into a lattice model (each cell of the lattice represented by an equivalent set of equations). The robustness of the system is tested by varying the model parameters at random and observing the effect. Interestingly, it was found that the system was largely robust to even large variations in the individual parameter values. Another investigation [43] questions the assumption made by Waddington that stabilizing selection is needed for canalization (robustness) to evolve. The authors developed a modelling framework in which the strength of selection towards an optimal expression pattern could be modulated, simulating different degrees of stabilizing selection. Stability in gene expression was also selected for (it was argued that this is a necessity in development). Since canalization was observed independently of the degree of stabilizing selection, the authors conclude that selection for stable (i.e. reaching steady state) network dynamics alone is a sufficient condition for producing robust networks without the need to invoke evolutionary assumptions such as stabilizing selection. Moving beyond theory, the availability of fitness data for yeast [44] has made it possible to analyse the role of duplicate genes in robustness [45], with the finding that duplicated genes do indeed compensate for null mutations more than singleton genes.

Model organisms such as baker's yeast (*Saccharomyces cerevisiae*), the fruit fly *Drosophila melanogaster*, and the nematode worm *Caenorhabditis elegans*, have been subjected to intense experimental, molecular and genetic studies. As a consequence, some gene networks in these organisms together with the network properties (e.g., gene interactions and activation thresholds) are well defined. To gain an evolutionary perspective though, comparisons are needed with related species, and this is becoming increasingly feasible as high-throughput techniques generate genomic, proteomic and functional data (e.g. RNAi , microarrays) from an increasingly diverse range of organisms. As far as sex determination mechanisms are concerned, three organisms have been studied more intensely than any others: *C. elegans* (figure 1.4a), *D. melanogaster* (figure 1.4b), and the mammalian system. From an evolutionary perspective, the most convenient of these for study is currently *D. melanogaster*, given recent advances in understanding of related insect sex determination networks: in the Mediterranean fruitfly (*Ceratitis capitata*) [46], domestic house fly (*Musca domestica*) [47], the phorid fly *Megaselia scalaris* [48], honeybee (*Apis mellifera*) [49], and silkworm (*Bombyx mori*) [50].

1.1.5 Models of gene network evolution

Relatively few attempts have been made to model the evolution of developmental gene networks in a biologically realistic context which includes both dynamical behaviour and evolution. One reason for this, mentioned above, is the lack of experimental (comparative) data, but a second important reason is a dearth of theoretical techniques combining dynamical models with evolutionary models (though each is individually highly developed). Those studies which have been done, have tended to focus on either (a) general aspects of network evolution, or (b) the evolution of patterning networks.

Within the first category, one important contribution [51] considered the effect of gene duplications on network dynamics, confirming the intuitively obvious notion that duplications involving either (a) very few genes or, (b) nearly all genes, have the smallest effect on dynamics. A less obvious result of the study showed that the largest changes in temporal expression pattern are likely to occur when ~40% (rather than 50%) of the genes in the network are duplicated. Extensions of this particular model have since been used to address several important questions in evolutionary biology such as canalization (discussed above)[43], genetic assimilation [52] (when an acquired trait loses it dependency on the environmental trigger to become an inherited trait [53]), and the prevalence of "evolutionary capacitors" (genes which suppress phenotypic variation under normal conditions but release the variation when functionally compromised, such as the *Drosophila* gene *Hsp90*) [54].

A second important area which has attracted interest is the evolution of patterning gene networks. Two recent theoretical studies [55, 56], evolve networks together with lattice model simulations to show that as few as three or four nodes are sufficient to generate complex spatial expression patterns. Another study [57] has also shown the tendency for small "emergent" (reaction-diffusion) type networks to be replaced by hierarchical (cascade) networks, as in modern *Drosophila*.

As may be obvious from this survey, there is remarkably little consensus on modelling methodology. Biological network modelling is similar to many other branches of science where modelling is used, in that although more progress can be made with simpler models, these simpler methods will always be dismissed by critics who believe that more sophisticated methods (for example, using stochastic formulations) are indispensable [41]. One major problem in network modelling is a lack of useful data. In particular, parameters such as rate constants, and even interactions, must often be guessed by random sampling even when most of the network interactions are known. For example, in [39], a hypothetical interaction is introduced to make the network behaviour more realistic. Many studies are still obliged to deal with simulated network structures [58], although the discovery of topological features specific to gene networks, is allowing simulated structures to become ever more realistic [18]. The focus of experimentalists on model organisms makes lack of data an even greater problem for researchers interested in the evolution of gene networks, since comparative data are particularly scarce. The experimental emphasis on model organisms has caused a parallel emphasis on increasingly detailed models of networks in model organisms, with no evolutionary perspective. Hope for evolutionary biologists may come from technological advances in high throughput methods, which should be capable of producing the desired data at low cost in the near future.

Another important point that should be mentioned is that modelling has a long history in certain areas of biology, and is often highly developed mathematically - population genetics, for example [59]. However, many of the new tools being used in biological networks come from different areas such as graph theory and engineering. There is an urgent need to reconcile concepts such as pleiotropy which have long been used in genetics, with newer ones such as robustness [60]. For example, dominance can be considered as equivalent to robustness, though at the level of a single gene, since dominant phenotypes are more robust to genetic perturbations than recessive phenotypes [61]. This integration will be impossible without common definitions, which must eventually be quantitative to allow meaningful comparisons of model and experiment. Thus far, such statistics have adopted only rather crude forms: for example, to measure robustness, the degree to which null mutations cause an increase in gene expression variation (which in turn is correlated with decreased fitness) [54], and for modularity, the frequency of interactions in the network [62] (and even this first attempt at a definition may be of limited usefulness [63]).

1.2 Evolution of sex

The question of why sexual reproduction evolved is a fundamental question in biology. Behind this broad question, a number of more specific questions can be asked. These in turn fall into two principal categories [64]: Firstly, why did sex evolve at all, and why is it so pervasive? Most current theories suggest that sex offers an evolutionary advantage through genetic recombination. Recombination seems to be good at removing harmful mutations and allowing new combinations of genes to come together, providing more opportunities for improved fitness and offering the flexibility to adapt to new environments. Secondly, how did these complex sexual systems evolve? Once sexual reproduction is in place, we can ask what mechanisms determine the differential development of the sexes (sex determination). Here, we will be concerned mostly with sex determination, and in particular the evolution of the genetic mechanisms involved.

In all organisms that produce two different sexes, sexual development is the result of the modification of a basic developmental program in order that one of the sexes can develop [65]. Initially, the male and female embryos are similar with sexual differences developing at later stages. The determination of the somatic sexual phenotype (i.e. the development of the individual as either male or female) is usually quite different, and consequently considered separately, from the determination of the germline (whether the future germ cells become sperm or egg). Another important aspect of sex determination is the mechanism by which the embryo compensates for the differences in chromosomal composition between males and females, or dosage compensation.

1.2.1 Variety and flexibility of sex determination mechanisms

The single requirement of a sex determining system is that it should cause some members of a species to develop as one sex, the rest as the other. Unsurprisingly, there is more than one solution to this problem. Sex determination systems can be either chromosome-based (genetic sex determination) or environmentally controlled (environmental sex determination), or even both [66]. In genetic systems, sex chromosomes can contain a single dominant regulator (for example, the mammalian Y chromosome contains *Sry*, a dominant masculinizer), or several dosage dependent regulators (such as in the systems of *D. melanogaster* and *C. elegans* discussed below).

In many genetic systems, the sex that produces gametes with different sex chromosomes is termed *heterogametic* (the other sex is *homogametic*). If males are heterogametic, as in mammals, then the sex chromosomes are called X and Y (XY males are *heterogametic*, and XX females are *homogametic*). If the females are *heterogametic* though (as they are in birds and butterflies), the chromosomes are called Z and W. A female mouse is therefore XX, but a female butterfly is ZW. In some species, such as *C. elegans*, males have only one sex chromosome (denoted XO), and females (actually hermaphrodites) have two (XX). Genetic sex determination can also occur without sex chromosomes. In the order Hymenoptera (wasps, bees and ants), the copy number of the entire genome is used to determine sex: unfertilised (haploid) eggs develop as males, whereas fertilised (diploid) eggs develop as females. This mechanism allows females to control the sex ratio among their progeny by allocating stored sperm as needed.

An enormous variety of environmental sex determining systems also exists. For example, certain turtles (*Trachemys scripta*) have a temperature dependent system in which eggs developing at cool temperatures become male, and eggs developing at warm temperatures become female. Other reptiles, such as alligators, have the opposite developmental program, with cool eggs developing as female, warm eggs developing as male.

Social environment cues can also determine sex. The fish species *Pseudanthias sqamipinnis* are sequential hermaphrodites and develop initially as females, possibly becoming males afterwards, depending on the social context. These fish form harems in which one male oversees numerous females. If the male dies, the dominant female of the harem will undergo a sex change from female to male and replace it.

Separate mechanisms can operate in different tissues of the same organism. For example, in marsupial mammals, germline sex is determined by the presence of a Y chromosome, but the choice of a female pouch versus a male scrotum depends on X chromosome dosage (with XX leading to a pouch). Therefore, mutant XXY kangaroos develop both testes and a pouch [67].

1.2.2 Sex determination as a model for network evolution

Although we have an intuitive notion of biological complexity, in terms of morphological complexity, or the variety of cell types, the term itself is hard to define. Traditionally, it was thought that biological complexity was largely reflected by the number of genes [68]. More recently though, the various genome projects have shown that this is not necessarily the case [69], which suggests a better definition is required. At the same time, an increase in network complexity appears to be correlated with the evolution of higher organisms, whether considered as an increase in the complexity of protein interactions [70, 71], or of gene regulation [72, 73]. It has therefore been suggested [12] that complexity measures based upon gene network connectivity would correlate better with biological complexity, understood as morphological or behavioural complexity, or the variety of cell types. A growing body of evidence supports the notion that plasticity in gene regulation (for example, changes in the *cis*-regulatory systems of genes) more often underlies the evolution of morphological diversity, than do changes in gene number or protein function [74]. A deeper theoretical understanding of how gene networks evolve will therefore become increasingly important to evolutionary biologists. As gene networks from various organisms are determined in the laboratory and data become more plentiful, we can begin to ask how networks change as species diverge.

Sex determination mechanisms represent a good model for the study of gene network evolution for several reasons:
a) They evolve relatively rapidly [75].
b) Certain sex determination networks (in *C. elegans* and *D. melanogaster*) are among the best understood of any gene network [76, 77].
c) When modelling sex determination networks, spatial aspects can, and indeed have, been ignored [78]. This considerably reduces model complexity.
d) The genes that constitute the network often perform secondary functions, which may or may not be sex-specific: for example, in *D. melanogaster*, *transformer* plays a important role (via *fruitless* [79]) in male courtship [80], *Sex-lethal* regulates sex-specific dosage compensation [81], and *scute/sisb* is vital for neurogenesis [82]. At a first approximation, we can say that these secondary functions (and their contributions to overall fitness) do not interact significantly. This is in contrast to other developmental networks, (in pattern formation, for example) in which genes interact in nontrivial ways at many different levels [83].

Although sex determination systems are known to be diverse, certain common features between species do exist [76]. Firstly, sex determination is triggered by a primary signal, often as a result of differential expression of genes on sex chromosomes (though in other cases the signal can be an environmental cue). Secondly, one highly conserved gene (*doublesex*) does appear to exist, usually at a downstream position in the pathway. Homologs of the *D. melanogaster doublesex* gene have been found to be involved in sex determination across a wide variety of species including *C. elegans* [84], humans [84] and birds [85]. The discovery of this highly conserved gene suggests that sex determination is not as plastic as previously thought, since certain constraints exist.

The discovery of the conserved role of *doublesex*-like genes added support to a hypothesis [86] concerning the evolution of sex determination networks. The core idea of the hypothesis is that sex determination networks evolve in reverse order from the final step in the pathway up to the first. Evidence across insect species also supports the hypothesis. Studies analysing the genes of the *D. melanogaster* sex determination network (figures 1.4b and 1.5) in other insect species suggest that indeed the genes may be less conserved as we move up the pathway. As mentioned previously, the gene *doublesex* appears to be highly conserved, and has proven to be so in other insects which have been studied. The next gene, *transformer*, is conserved in a sex determination role in the Mediterranean fruit fly *Ceratitis capitata* (separated from *D. melanogaster* approx. 100 mya) [46]. At the same time, the next gene, *Sex-lethal*, has been found to perform a sex determining role within the genus *Drosophila* [87], but not in more distantly related species such as *Ceratitis capitata* [88] and *Musca domestica* [89]. The sex determination pathway is not the only network which contains both conserved and variable genes. A recent study of the gene network underlying wing development in ants [90] showed that network changes leading to a wingless phenotype could occur in a number of different ways (alterations in expression patterns), whereas the expression pattern of genes such as *Ultrabithorax* (*Ubx*) and *distal-less* (*dll*) remain largely conserved.

As mentioned above, relatively few sex determination networks are known in real depth. The best understood are probably *D. melanogaster*, and *C. elegans* [91]. However, interest in sex determination evolution has led to studies in other insects, such as *C. capitata*. What is known of the structures of these three networks is shown in figure 1.4. There are two common features: a) the shared homolog (*dsx*, *mab-3*) at the end of all three pathways, and b) *D. melanogaster* and *C. capitata* both share *transformer* (*tra*). Apart from this, all three networks are different. In *C. elegans* and *D. melanogaster*, the primary signal is defined by the X:A ratio (albeit by different genes in each), whereas in *C. capitata* it is defined by a male determining factor (as it is in mammals). *Musca domestica*, which is more distantly related to *D. melanogaster*, is likely to have a system based on a dominant feminizing allele

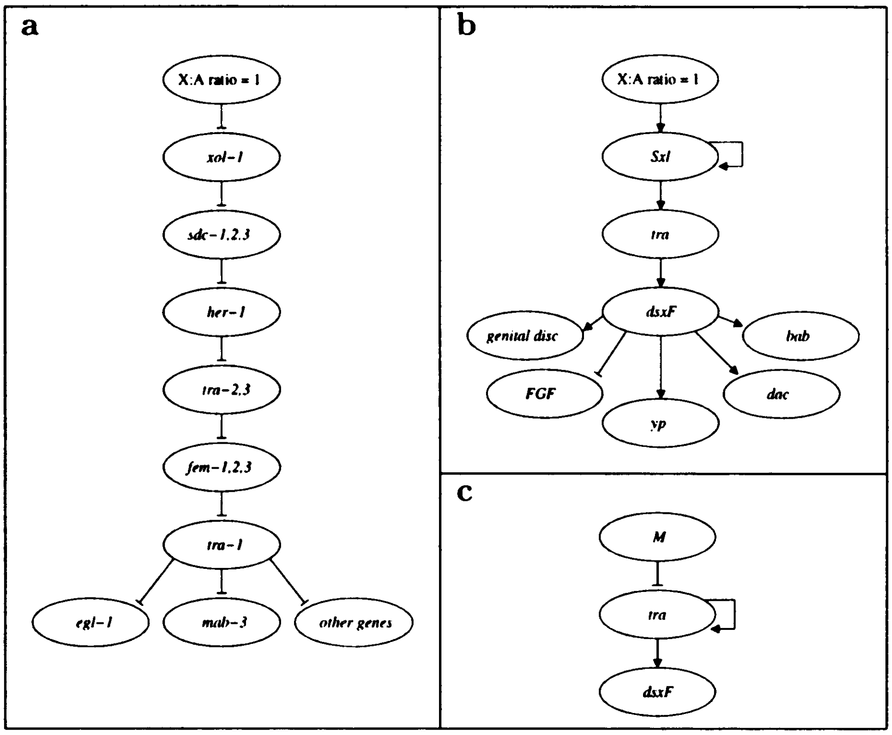


Figure 1.4: Sex determination networks of a) *C. elegans*, b) *D. melanogaster*, and c) *C. capitata*. This diagram is a conventional gene network description in which each node represents one or more genes that participate at that point in the network. Arrows indicate positive interactions, flat tips indicate negative interactions. The gene *doublesex* (*dsx*) is denoted *dsxF*, since it is the female form which is activated by *tra* in each case.

F [92].

1.2.3 Sex determination in *D.melanogaster*

The main features of sex determination in *D. melanogaster* are now described - for general reviews see [91, 76]. The pathway is shown in figure 1.5. The key early signal leading to sexual differentiation (whether the fly becomes male or female) is the ratio X:A, of X chromosomes to autosomes (X and Y chromosomes are sex chromosomes, other chromosomes are known as autosomes). The directive for establishing the sexual phenotype is carried out by the differential expression of the key gene *Sex-lethal* (*Sxl*), together with several

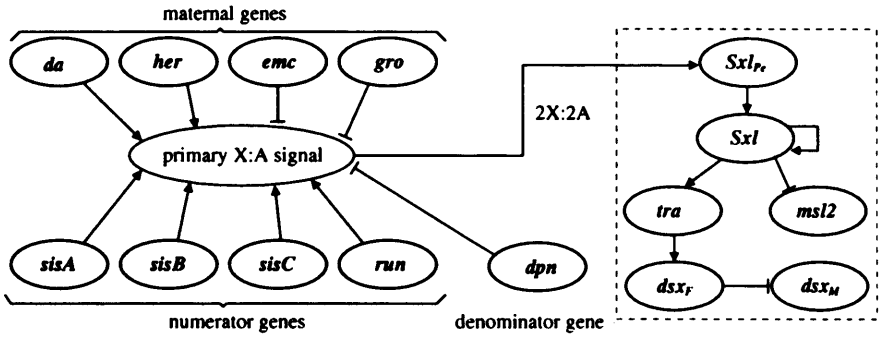


Figure 1.5: The sex determination gene network in *D. melanogaster*. The X:A signal components are integrated and activate the network on the right in females (2X:2A). In males (1X:2A) the network on the right hand side is not activated. The dotted box delimits a simplified version of the "known" network which will be used in chapter 2.

downstream sex-specific genes. If *Sxl* is switched "off", then the pathway produces the male pathway of determination, whereas the "on" position shunts the system into female mode of sex determination. The default mode of the pathway culminates in the production of the male-determining transcription factor (a generic name for gene regulatory proteins) DSXM, whereas the non-default pathway culminates in the production of the female-determining transcription factor DSXF (by convention, gene names and their abbreviations are denoted in italics, but the protein product is denoted in capitalised non-italics) Although *Sxl* is activated by differential transcription regulation (qualitatively equivalent to that described for *E. coli* β-galactosidase at the beginning of this chapter), subsequent links in the pathway occur via a mechanism termed alternative splicing.

DNA segments that code for proteins, contain intervening sequences called *introns*. Splicing occurs after transcription of the DNA into the primary RNA transcript (pre-mRNAs), and involves removing the *introns* to bring together the coding regions, or *exons*, to form a mature mRNA which codes for a protein. Alternative pathways of splicing can produce different mRNAs, and subsequently different proteins from the same primary transcript. Whether one splicing variant is chosen over another will often depend on the presence (or absence) of specific RNA-binding proteins. Often, alternatively spliced proteins (such as *Sxl*) can themselves be RNA-binding and regulate their own production (forming autoregulatory loops).

Looking now at this process in more detail: In XX females, the X:A ratio of 2 X chromosomes to 3 pairs of autosomes leads to activation of *Sxl*, whereas in XY males, the corresponding ratio is 1:3 and *Sxl* is not activated. The early signal is determined by a key group of "numerator" genes (so called due to their differential role in increasing early signal strength) on the X chromosome, as shown in figure 1.5. Four numerator genes have so far been identified: three *sisterless (sisA, sisB, sisC)* and *runt (run)*. All encode transcription factors which positively regulate *Sxl*. Other genes are essential to the sex determination process, but are expressed equally in males and females, and therefore do not have discriminative power. These other genes are: a) the single autosomal "denominator" (which tends to reduce the signal) gene so far found: *deadpan (dpn)*, and b) four maternal genes, of which : *daughterless* (*da*) and *hermaphrodite* (*her*) are positive regulators ("activators") of *Sxl*, and *extramacrochaetae* (*emc*) and *groucho* (*gro*) are negative regulators ("repressors") of *Sxl*. These transcription factors (numerators, denominator and maternal) have only one sex determination role: in a narrow time window in the early *Drosophila* embryo - roughly from 2 to 3 hours after fertilization - they determine if the *Sxl* regulatory switch gets flipped on.

The outcome of the early signal is activation of Sxl_{Pe}, the "establishment promoter" (promoters are the regions of DNA that signal initiation of transcription) in females only, leading to production of *Sxl* mRNA transcripts, which in turn leads to production of early SXL protein. At this point, Sxl_{Pe} is replaced by the "maintenance promoter" Sxl_{Pm}, active in both sexes. From this stage onwards, the SXL protein itself is necessary for alternative splicing leading to more functional SXL protein (i.e. it is autoregulatory). Male pre-mRNA transcripts contain an exon with what is known as a "stop codon", which is spliced out in females. A stop codon is the signal for the translation machinery (which later converts the RNA to protein) to terminate translation at that point, which in the case of male Sxl_{Pm} will create a truncated and non-functional protein. Since the early Sxl_{Pe} transcripts naturally lose the stop-codon containing exon during RNA processing, the early burst of SXL is possible in the absence of any SXL protein, which permits the initiation of this autoregulatory process in females. These steps are summarised in figure 1.6.

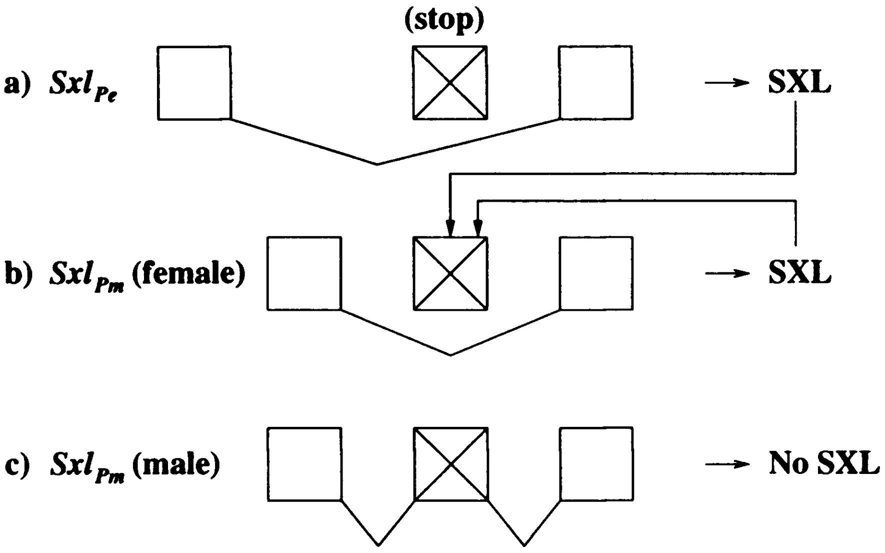


Figure 1.6: The initiation and maintenance of the *Sxl* switch. The boxes represent exons, with the crossed box representing the stop-codon-containing exon. (a) Early promoter transcripts (activated in females only) naturally lose the stop-codon-containing exon. (b) Maintenance promoter transcripts have the stop-codon-containing exon spliced out in females due to maintained presence of SXL protein. (c) In males, the absence of initial SXL protein leads to inclusion of the stop-codon-containing exon and no production of SXL as a consequence.

Not only does SXL have to autoregulate, but it must be capable of activating the shunt pathway that will lead to female-specific gene expression. This activation is accomplished again through RNA-binding. The main target of SXL protein is the gene *transformer* (*tra*), which is spliced (again in females only) to produce an mRNA-encoding active TRA protein. In turn, TRA protein is an RNA-binding protein that produces female-specific splicing of the *doublesex* (*dsx*) pre-mRNA. The mRNA produced by this splicing pattern encodes a DSXF protein, a global female-determining transcription factor. In the absence of active SXL protein, the splicing pattern of *tra* primary transcript produces an mRNA which contains a stop-codon, and produces nonfunctional protein. In the absence of functional TRA protein, the default splicing of the *dsx* pre-mRNA leads to the production of DSXM, a global male-determining transcription factor.

It should be noted that there are many other genes involved in the sex determination process, which, though crucial, are expressed equally in both sexes, and are therefore not discriminatory. For example, the product of the *transformer 2* (*tra2*) gene is required for TRA to correctly splice *doublesex* in females [93].

Until recently, the only known targets of *dsx* were two genes encoding the terminal differentiation proteins: *Yolk protein-1* and *Yolk protein-2* (the *yp* genes), with DSXF activating and DSXM repressing their expression. However, several new targets of *dsx* have recently been discovered [94]. These genes (including *yp*) are shown downstream of *dsxF* in figure 1.4. Perhaps most importantly, *dsx* acts to promote sex-specific growth of the genital disc, leading to development of appropriate genitalia in each sex [95]. Of particular interest to evolutionary developmental biologists (since it appears to have evolved in the *D. melanogaster* species) is the interaction of *dsxF* with *bric-a-brac* (*bab*) [96]. The net effect of the interaction of DSXF with *bab* is to prevent male-specific pigmentation in two abdominal segments (A5 and A6), a distinguishing characteristic of *D. melanogaster* males.

The above steps describe sex determination of the soma (i.e. not of the germ cells). Additionally though, *Sxl* mediates two other downstream pathways: a) dosage compensation and b) germline development. Control of dosage compensation is important because XX females have two copies of each X-chromosomal gene whereas XY males have only one, and these dosage differences could lead to potentially fatal imbalances. In *D. melanogaster* dosage compensation is achieved by hyperactivation of the single X chromosome in males. All five *male-specific lethal* genes (*msl1, msl2, msl3, mle, mof*) are required for dosage compensation, though only *msl2* is expressed exclusively in males. The *Sxl* gene downregulates *msl2* in females, thereby halting hypertranscription.

The third pathway controlled by *Sxl*, germline sex determination (the genetic mechanism which decides whether egg or sperm is produced), is less well understood. Although *Sxl* is required for oogenesis (egg production) [97], it does not function as the key gene of sex determination and its regulation is different from that of the soma. If *Sxl* were the key to sexual development in the germline, then female (XX) germ cells lacking *Sxl* function, in analogy to the effect of the gene in the soma, should form sperm, and XY cells with mutations of *Sxl* such that functional *Sxl* is always produced ("constitutive" mutations of *Sxl*), should form eggs. However, XX cells lacking *Sxl* function do not form sperm, and XY germ cells with constitutive *Sxl* mutations produce fertile sperm in male hosts [98]. Additionally, *Sxl* is activated much later in germ cells than in somatic cells, and does not require the "numerator" genes for it to be activated. In summary, it is unlikely that *Sxl* plays a key "switch" role in the germline as it does in the soma. The key gene for sex determination in germ cells is still not known, and indeed may not even exist [76], since it is possible that several different signals are involved in controlling distinct aspects of sexual development in germ cells. For the purpose of this study, we will be concerned only with the somatic sex determination network.

1.2.4 Evolution of sex determination in *D.melanogaster*

In [92], a detailed hypothesis is put forward concerning the evolution of the Drosophila sex determination pathway. Using the available molecular data together with standard population genetics models, the authors postulate a step-by-step reconstruction of the pathway in which sexual selection plays a fundamental role. Starting with a simple ancestral system in which the discriminatory signal resides at the *dsx* locus, the following hypothetical steps are proposed: (1) the discriminatory signal passes to the *tra* locus by means of a stop codon mutation causing *tra* mRNA transcripts to be truncated prematurely in males (as occurs in modern Drosophila). (2) Recruitment of *Sxl* as a splicing regulator of *tra* (removing the stop codon from *tra* mRNA transcripts) provokes three further changes in *Sxl*, as follows: (3) *Sxl* autoregulation, (4) appearance of a null allele of *Sxl* containing a stop codon, (5) recruitment of *sis* (representing the "numerator" signal) as an activator of the early promoter Sxl_{Pe}. (6) Lastly, the appearance of an X-linked null allele of *sis*, changes the discriminatory signal to the *sis* locus. (7) The null *sis* allele leads to degeneration of the male X chromosome where it resides, leaving us with the modern XX/XY Drosophila system (female homogamety and male heterogamety). Each hypothetical transition leads to increased fidelity of the sex determining signal.

That evolution has favoured increased signal fidelity is supported by a recent theoretical study [78], which deals specifically with activation of *Sxl*. The authors have shown how through mechanisms such as dimerization (in the primary signal elements) and autoregulation of *Sxl*, the initial 2:1 female to male signal ratio is amplified to ~80:1. It is reasonable to suppose that such a mechanism has evolved through positive selection for robustness in the signal.

1.2.5 Sex determination in *C. elegans*

The sex determination mechanism of the nematode worm *Caenorhabditis elegans* (reviewed in [91]) is perhaps the only other sex determination system understood at a comparable level of detail to that of *D. melanogaster*. As in *D. melanogaster*, the primary signal in *C. elegans* sex determination is determined by the ratio of X chromosomes to autosomal chromosomes (X:A). Worms with two X chromosomes develop as hermaphrodites (phenotypic "females" generating a limited amount of sperm which can be used for self-fertilization), whereas XO worms develop as males. The key difference, as can be seen in figure 1.4, between the *C. elegans* sex determination network and that of *D. melanogaster* is the reliance of the *C. elegans* system on negative genetic switches, each one reversing the action of the previous one, where *D. melanogaster* has a positive cascade.

In *C. elegans*, the X:A signal itself is determined by at least four *numerator* signals which act, in XX hermaphrodites, to repress (recall that the *D. melanogaster* numerator genes are activators) the gene XO-lethal 1 (*xol-1*) at the top of the hierarchy. At least two such numerator genes have been found:

Signal Element on X (*sex-1*) and *Feminizing gene On X* (*fox-1*), though two other, as yet unidentified, signals are known to act in concert with these two to repress *xol-1* [99]. More specifically, SEX-1 protein is known to repress *xol-1* at the transcription level, whereas FOX-1 (an RNA-binding protein) represses *xol-1* post-transcriptionally. In XO males, where this repression does not occur, *xol-1* is active.

In worms, dosage compensation occurs by hypo-activation (a reduction in overall activity) of the two X chromosomes in XX hermaphrodites, whereas no dosage compensation occurs in XO males. The gene *xol-1* represses the three *Sex determination and Dosage Compensation defect* genes (*sdc-1*,*sdc-2* and *sdc-3*) at the next step in the hierarchy, as shown in figure 1.4. Since *xol-1* is inactive in XX hermaphrodites, this repression does not occur, causing (a) *her-1*, the next gene in the hierarchy, to be repressed via transcription regulation, and (b) activation of dosage compensation in which SDC-2 plays a crucial role. In this way, the *sdc* genes play a dual role (in both dosage compensation and the sex determination hierarchy) comparable to that played by *Sex-lethal* in flies.

We find that, as we move down the pathway, the remaining genes act essentially as switch genes, i.e. their function is principally to transmit the sex-determining signal. Inactivation of *her-1* in hermaphrodites causes activation of two *transformer* genes *tra-2* and *tra-3* by default, whereas in males these genes are inactive (note that the *C. elegans transformer* genes are unrelated to the *D. melanogaster transformer* gene). In turn, activation of *tra-2* and *tra-3* causes repression of the *FEMinization of XX and XO animals* genes *fem-1*, *fem-2* and *fem-3*. Finally, activation of the key "switch" gene *tra-1* will lead to phenotypic hermaphrodites, whereas inactive *tra-1* lead to phenotypic male worms. The gene *tra-1* also represses several other genes including *Male ABnormal 3* (*mab-3*), the previously mentioned homolog of *doublesex*.

Note that many of the genes described in this pathway only exist to control the activity below them. It has been shown that if the appropriate signal can be generated by other means, the sex determination system can change markedly and the upstream regulators can become dispensable to the process. In one experiment [100], two mutations in the (autosomal) *tra-1* gene were used to generate a stable and fertile strain of worms in which *heterogametic* (XY) males were *homozygous* for the mutant allele, and *homogametic* (XX) females were *heterozygous*. Here, the *tra-1* locus had become the discriminatory sex-determining locus. Similar experiments have shown how temperature-sensitive mutations can transform a previously genetically determined system into an environmentally-determined system [101]. These experiments demonstrate the plasticity of sex-determining pathways.

1.2.6 Sex determination in the mammalian system

Although it has been the object of intense research, sex determination in the mammalian system is less well understood than for either *C. elegans* or *D. melanogaster*. Mammalian sexual development can be divided into several steps:
(a) Formation of the undifferentiated gonad,
(b) Gonad commitment to testis or ovary (sex determination),
(c) Differentiation into testis or ovary,
(d) Hormone production which leads to full somatic (rest of the body) sexual differentiation.
In the absence of a functional testis, the rest of the body develops as female [66]. More specifically, studies have shown that male development is caused by making functional testicular Sertoli cells, rather than ovarian follicle cells. In turn, Sertoli cells secrete Mullerian Inhibiting Substance (MIS) to promote differentiation of Leydig cells. Leydig cells produce testosterone which antagonises female differentiation, whereas MIS promotes male differentiation.

Although the sex determination pathway is not well understood, several crucial genes have been identified. The Y-linked testis-determining factor *Sry*, is an essential trigger of male gonad differentiation [102], and its activity is required for expression of the related gene *Sry-box containing gene 9* (*Sox9*). Another important gene, *Dax1*, antagonizes male differentiation in females [103]. *Dmrt1*, a mammalian homolog of the *Drosophila* gene *doublesex*, has been found to play a role in differentiation of different cell types in the testis [104].

1.3 Thesis overview

This thesis is concerned with modelling the evolution of gene networks, and in particular networks involved in sex determination. Specifically, we attempt to model the evolution of these gene networks in a biologically realistic context which includes both dynamical behaviour and evolution. The thesis is organised around two models: a simpler synchronous logic model (chapters 2 and 3), and a more complex hierarchical model (chapters 4 and 5). In chapter 2, the logic model is used to consider the potential for evolutionary change of the existing Drosophila sex determination gene network. With the aid of this model, theoretical concepts are introduced, such as a network-specific form of mutation, as well as a notion of functional equivalence between networks. These concepts are then applied to the sex determination mechanism and compared to a population of random networks (which constitute a suitable null hypothesis in this case). It is found that sex determination networks generally exist within large sets of functionally equivalent networks all of which satisfy the sex determination task. These large sets are in turn composed of subsets which are mutationally related, suggesting that the networks can change significantly without compromising the core functionality, namely the sex determination task. The technique for finding functional equivalence between networks suggests a general method for gene network reconstruction, which is explored in chapter 3. The technique is used to suggest ways in which experiments involving large-scale perturbations can be designed to obtain reasonably accurate reconstructions. It is found that a relatively small number of perturbations significantly improve inference accuracy, particularly for low-order inputs, as long as the perturbations themselves alter the expression level of approximately half the genes in the network. Lastly, in chapters 4 and 5, a hierarchical model is presented which integrates population genetics techniques with network dynamics (using Ordinary Differential Equations). The model consists of a core population genetics simulation within which parameters such as the sex and fitness of the genotype are calculated from the corresponding network dynamics. The model is used to investigate the early evolution of sex determination networks. Following from a hypothesis proposed by A.S. Wilkins [86], the assumption is made that sex determination networks have evolved in a retrograde manner from bottom to top. Starting from the simplest possible ancestral system, based on a single locus, we explore the way in which more complex systems could have evolved. Specifically, transitions from single locus to two locus sex determination systems are considered in chapter 4, and transitions from two to three loci in chapter 5. Changes in heterogamety are also considered for both ancestral conditions.

Chapter 2

The evolutionary potential of the Drosophila sex determination gene network

2.1 Background

The fact that certain aspects of the sex determination mechanism seem to be highly conserved whereas others have greater plasticity (discussed in chapter 1), leads us to conclude that an understanding of the evolutionary constraints and flexibility imposed by gene networks will be vital to explaining the evolutionary diversification of sex determination. Gene network models, reviewed in [34], have been used to investigate many aspects relevant to network evolution. These include gene distribution in the genome [51], robustness [39], the preservation of topological motifs [20] and differentiation [57].

In this chapter, we investigate mutability in sex determination networks. Network behaviour is characterised using a synchronous logic model, which in turn requires a network architecture to be defined. In the network architecture, each gene is represented as a node and each interaction as a directed link. Taking advantage of the fact that the sex determination system has been extensively studied in *D. melanogaster*, we have both a known network architecture as well as the pattern of gene expression through time. We proceed by analysing various network characteristics, in each case comparing the *Drosophila* network to a population of randomly generated "sex determination" networks in order to find special properties of the *Drosophila* network. It is quite possible that the general characteristics found for the population of random sex determination networks may apply more broadly, to other classes of gene network.

2.2 Methods

We use a synchronous logical network model. Similar models have been used extensively in neural network modelling [105], as well as for gene networks [9, 106] - often focussing on the global properties of large-scale genetic regulatory systems [8, 107, 108]. Here we are concerned with studying the potential for mutational variation in network architecture. We use this theoretical approach to propose possible constraints on the evolution of the sex determination gene network.

2.2.1 A discrete gene network model

For a system of n nodes, the state of each node s_i $(i = 1, .., n)$ is represented by the binary values 0(OFF) and 1(ON). Note that genes may sometimes be represented by more than one node if, for example, they represent different products of the same gene as produced by alternative splicing. Additionally, each node is assigned a default ON/OFF state $\theta_i \in \{0, 1\}$. The node interactions are described by an $(n \times n)$ matrix C, composed of elements $C_{ij} \in \{-1, 0, +1\}$, representing the positive(+1), zero(0) or negative(-1) influence of node j on gene i. State transitions are calculated as follows:

$$s_i(t+1) = \sigma(u_i(t)) \tag{2.1}$$

$$\text{where} \quad u_i(t) = \sum_j C_{ij} s_j(t), \quad \sigma(x) = \begin{cases} 1 & \text{if} \quad x > -\theta_i \\ 0 & \text{otherwise} \end{cases}$$

The state of the ith node at the next timestep, $s_i(t+1)$, is therefore determined by the balance of positive versus negative inputs which are ON at the previous timestep t. If the balance is positive, then $u_i(t) > 0$ and the next state will be 1(ON). Similarly, if the balance is negative, then $u_i(t) < 0$ and the next state will be 0(OFF). If $u_i(t) = 0$ (indicating either that there are no active input connections, or that they balance out), then the default value θ_i determines the next state. This default value needs to be given *a priori*, though how it is defined in practice will depend on the problem. For example, one might take the expression level before a certain developmental event of interest begins. Alternatively, one might consider the expression level in the absence of a mutation or chemical treatment. In the *D. melanogaster* sex determination gene network, the male state for each gene will be used, which is reasonable since the default sex in this species is male. By stating that the default sex in this species is male we mean that, in the absence of the primary signal, the state of the downstream nodes in the pathway will remain unchanged, the result of which is a male phenotype.

2.2.2 Network Mutations and their *neighbourhoods*

We define the following distance metric for networks represented by interaction matrices C and C':

$$d(C, C') = \sum_i \sum_j |C_{ij} - C'_{ij}| \tag{2.2}$$

We define any network P to be a *neighbour* of C if $d(C, P) = 1$. This is simply the case where P is identical to C except for a single difference: an edge deletion (from $-1 \rightarrow 0$, or $+1 \rightarrow 0$), or edge insertion (from $0 \rightarrow -1$, or $0 \rightarrow +1$). Any given network can have between n^2 and $2n^2$ such *neighbours.* Why this is so becomes clear if we consider the $n \times n = n^2$ entries of the matrix C: if the network is fully connected, then only deletions are possible (n^2 possible changes from $\pm 1 \rightarrow 0$), whereas if there are no connections, both positive or negative edges can be inserted ($2n^2$ possible changes from $0 \rightarrow -1$, or $0 \rightarrow +1$) for each entry in the matrix C. We define a *neighbourhood* to be

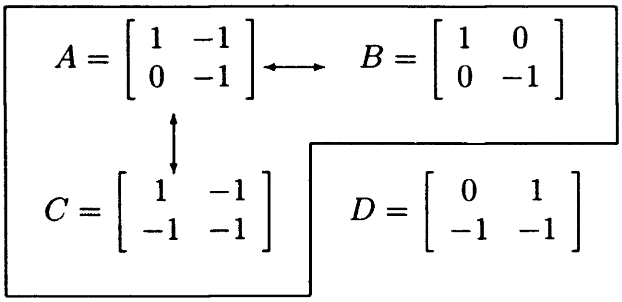

$A = \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix}$
$B = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
$C = \begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix}$
$D = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}$

Figure 2.1: Four interaction matrices in a 2-gene system. The arrows show *neighbour* relationships. The three networks within the boundary (A, B and C) represent a particular *neighbourhood*, even though B and C are not themselves *neighbours*. D is not in this *neighbourhood*.

a set of networks in which each element has at least one *neighbour* in the set (Fig. 2.1).

The concept of *neighbour* is analogous to that of genetic mutations, though specific to gene networks. As networks mutate, multiple *neighbours* may exist within a species at any given time. Selection will then act upon these. For example, if a particular *neighbour* has an effect on sexual dimorphism, then sexual selection may favour this and eventually lead to its fixation. Our main assumption is that an evolutionary change in network architecture is more likely to involve a succession of single interaction changes as opposed to multiple simultaneous changes.

By saying that a network may evolve into any *neighbour*, we are proposing a general model of network evolution in which all changes of a single interaction are considered equally probable. In real biological systems, certain interactions may be more easily evolved than others. For example, an input from a transcription factor may evolve with higher probability [109] than an input, say, from a structural gene. However, these probabilities are hard to specify *a priori* and are difficult to determine experimentally. It is also important to point out that although network mutations are represented as a single events in our model, they may involve more than one genetic mutation.

2.2.3 Determining solution networks

Assuming we are given the state dynamics $s(t)$ and the default vector θ, the problem is to find the necessary model parameters which will reproduce these dynamics. Specifically, a system initialised at $s(0)$ should reproduce the given dynamics $s(t)$ for $t > 0$. Note that multiple $s(t)$ expression patterns may be defined: for *D. melanogaster* sex determination, both female ($s^f(t)$) and male ($s^m(t)$) dynamics will be given. Our problem is to find one or more interaction matrices that will reproduce the given dynamics $s(t)$. The set of such matrices constitutes the solution set $\mathcal{G}$. The idea of a solution set combined with that of *neighbourhood* is similar to that of a "neutral network", a concept first introduced for a model of RNA folding [110].

We now define a *parsimony* measure for any network C, as the number of non-zero entries in C:

$$k(C) = \sum_i \sum_j |C_{ij}| \tag{2.3}$$

Since $\mathcal{G}$ may be very large in practice, it is convenient to use this parsimony measure to define a representative subset. The set $\mathcal{M} \subset \mathcal{G}$ for which k is a minimum will be referred to as the *Most Parsimonious Solution Subset* of $\mathcal{G}$. The minimum value of k will be denoted as $k_{\mathcal{M}}$.

The problem of finding any $C \in \mathcal{G}$, may be broken up into n sub-problems, since the expression pattern for each node i may be solved independently from the others (for a more detailed explanation, see chapter 3, section 3.2.2). This reduces the search space from $O(3^{n^2})$ down to $O(n3^n)$. Since we deal below with a small system of just 6 representative nodes, we used a straightforward enumerative approach to determine the entire set $\mathcal{G}$. For each node, all 3^n possible row vectors are evaluated.

Once the solutions have been found for each node, *neighbourhoods* within the set of row vectors can be easily determined using the distance metric. The number of *neighbourhoods* in $\mathcal{G}$ is simply the product of the number of *neighbourhoods* for each node. Finding $\mathcal{G}$ and determining the *neighbourhoods* are both combinatorial problems which scale exponentially with n. Finding only $\mathcal{M}$ will usually require less computation, but generally speaking this approach is not practical for more than a small number of genes.

2.2.4 The state dynamics of the sex determination gene network

In order to be able to combine the two models: state dynamics and network evolution,.it is our strategy to work with as small a number of genes as possible. We adapt the scheme suggested in [76], and attempt to represent the *Drosophila* network in terms of elementary processes which are common to sex determination in many species. In order to simplify the network to one node per process, we have chosen to condense the upstream components which determine the initial signal ("numerators", "denominator" and maternal factors) down to a single primary signal node denoted as Sxl_{P_e}. The network we will use is shown in the dotted box on the right hand side of figure 1.5 and will be referred to as the "known" network. It should be noted that, although the resulting network appears simple, it is in fact fairly complex, given it includes alternative promoters, alternative splicing, stop codons, and a complex primary signal, all of which has taken many years of effort to elucidate [76]. The processes used and their associated description are as follows:

1. Primary signal: Sxl_{P_e}, the *Sxl* gene 'establishment promoter', activated by the X:A signal.

2. Key gene: *Sxl*, the key sex determination gene, active expression of which is achieved by alternative splicing in XX embryos only.

3. Subordinate control gene: *tra*, spliced by SXL to a productive form in females only.

4. Dosage compensation: *msl2*, a target of SXL required for dosage compensation in males.

5. Female switch: dsx_F, female-specific form of *doublesex*, a target of *tra*, which activates female-specific and represses male-specific genes.

6. Male switch: dsx_M, male-specific form of *doublesex*, produced by default, which activates male-specific genes.

Now that we have a defined network, we need initial states $(s^f(0), s^m(0))$ and a default vector θ to produce the state dynamics. Since the initial states of these genes are known, and we stated previously that we would use $\theta = s^m(0)$, this task is straightforward, and the results are shown in figure 2.2. Since these are the state dynamics produced by the "known" network, we will refer to them as the "known" dynamics.

Female (s^f)

$$\begin{array}{l} Sxl_{Pe} \\ Sxl \\ tra \\ msl2 \\ dsx_F \\ dsx_M \end{array} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \circlearrowleft$$

Male (s^m)

$$\begin{array}{l} Sxl_{Pe} \\ Sxl \\ tra \\ msl2 \\ dsx_F \\ dsx_M \end{array} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \circlearrowleft$$

Figure 2.2: Female and male patterns of the "known" dynamics. The U-shaped arrow to the right of each final state indicates a steady state.

2.2.5 A measure of local dynamic diversity

Given a particular network and its associated pair of expression patterns (one female, one male), it is useful to know the potential it has to evolve changes in the expression pattern, while retaining its fundamental function (sex determination). We will now define a general statistic which measures the diversity of viable expression patterns available from the local *neighbourhood* of a sex determination network.

We first need a general definition of viability, applicable to any sex determination network. An ordered (female/male) pair of expression patterns (s^f, s^m) will be defined as *viable* if it fulfils the requirements of sex determination described as follows. Since our gene network model is deterministic, the temporal pattern is known completely as soon as we encounter a repeated state $s(t_r)$ that has ocurred previously at a time t_p (i.e. $s(t_r) = s(t_p), t_p < t_r$). If the repeated state occurred at the previous timestep $(t_p = t_r - 1)$, then the system is in a steady state. For the two expression patterns s^f and s^m, we denote this time t_p as t_p^f and t_p^m respectively. To define a pattern as *viable*, we focus on the final state of two sex-defining genes, which we will call **F** for female and **M** for male (in *D. melanogaster*, these would be dsx_F and dsx_M respectively). The criterion for the female expression pattern (s^f) in a *viable* pattern pair is:

$s^f_{\mathbf{F}}(t_i) = 1$ and $s^f_{\mathbf{M}}(t_i) = 0$ for $t_i \geq t^f_p$

In other words, the female-defining node **F** should be ON and the male-defining node should be OFF from time t_p onwards. Conversely, the criterion for the male expression pattern (s^m) is:

$s^m_{\mathbf{F}}(t_j) = 0$ and $s^m_{\mathbf{M}}(t_j) = 1$ for $t_j \geq t^m_p$.

Together, these two statements simply state that, in a *viable* pattern pair the two genes **F** and **M** should go to a steady state (even if other genes do not) and should be correctly expressed in both sexes.

We now need a second definition. Consider a network C which produces a pattern pair $(s^f$,$s^m)$ when initialised with the state vectors $s^f(0)$ and $s^m(0)$ respectively. The default vector θ is equal to $s^m(0)$ or $s^f(0)$ depending on the organism (for *D. melanogaster*, $\theta = s^m(0)$). Each *neighbour* C' of C also produces a pattern pair (s'^f, s'^m) under the same conditions ($s^m(0)$,$s^f(0)$ and θ). Two pattern pairs $(s^f$,$s^m)$ and (s'^f, s'^m) will be considered *distinct* if there are any differences in female expression (e.g. $s^f \neq s'^f$), male expression (e.g. $s^m \neq s'^m$) or both. Furthermore, some of these pattern pairs may be *viable*, others may not.

We now define the *local dynamic diversity* measure $l(C)$ of the network

C as equal to the number of *viable* and *distinct* pattern pairs which can be generated from C and its *neighbours*. $l(C)$ will be an integer value ranging from 1 (if C and all its neighbours share a single *viable* pattern pair), to $2n^2+1$ (if C and its n^2 neighbours all have *distinct* and *viable* pattern pairs). For example, the *Drosophila* network C_d was calculated to have $l(C_d) = 19$, which means that the 67 relevant networks (C_d and its 66 *neighbours*) are capable of generating 19 *distinct* and *viable* pattern pairs.

2.2.6 Comparing with random sex determination networks

We have presented several concepts which we will use to study the *Drosophila* sex determination network. However, we need some kind of null model to indicate which of the results describe sex determination networks in general, and which indicate something particular about the *Drosophila* network. We therefore generated a population of random networks with six nodes (there are six nodes in our simplified *Drosophila* network). There are a total of 3^{36} such networks and these were sampled uniformly using a random number generator for large integers (part of the GNU MP library). Two random and distinct initial condition vectors $s^f(0)$ and $s^m(0)$ were also generated each time, and the network was tested in turn with both defaults $\theta = s^f(0)$ and $\theta = s^m(0)$. If the female/male pattern pair (s^f,s^m) was *viable* (as defined in the previous section) using either default θ, then it was accepted as a valid sex determination network. A population of 100,000 such randomly generated sex determination networks (together with their initial conditions, default and resulting pattern pair) constitute our null model. Each attempt to find a *viable* pattern pair has a success rate of approximately 0.0036 (1 in 276), indicating that over 2.75×10^7 attempts were needed in order to obtain the sample of 100,000.

2.3 Results

2.3.1 The set $\mathcal{G}$ of dynamically equivalent networks

The enumerative search method described above in section 2.2.3 was implemented in the C++ programming language, and used to find $\mathcal{G}$ and $\mathcal{M}$ for both the random network population and the *Drosophila* "known" dynamics. The result of counting the solutions found for each node in the *Drosophila* network is shown in Table 2.1. Since the solutions for each node are independent of each other, any combination is possible. Using this knowledge, we can easily calculate the size of $\mathcal{G}$, $|\mathcal{G}| = 2,295,645,300$ (2.29×10^9), as the product of the number of solutions for each node.

node	solutions
Sxl_{Pe}	215
Sxl	13
tra	26
msl2	26
dsx_F	45
dsx_M	27

Table 2.1: Number of solutions found for each node.

Although this may seem like a high number, to put it into context, we need to compare it with the results for the population of random networks. The distribution for $\log(|\mathcal{G}|)$ (using $\log(|\mathcal{G}|)$ is appropriate here since $|\mathcal{G}|$ is calculated as a product of the solutions for each node) is shown in figure 2.3a, and we can immediately see that 2.29×10^9 (indicated by an arrow on the graph) is typical for $|\mathcal{G}|$.

2.3.2 *Neighbourhoods* within the set $\mathcal{G}$

We now look at the number of *neighbourhoods* in $\mathcal{G}$, and find that in the case of the *Drosophila* sex determination network, $\mathcal{G}$ is divided into 135 *neighbourhoods*. What is important from the evolutionary point of view though, is the *neighbourhood* $\mathcal{N} \subset \mathcal{G}$ in which the particular network actually exists, since this defines how the network can evolve. The value $|\mathcal{N}|$ can give us

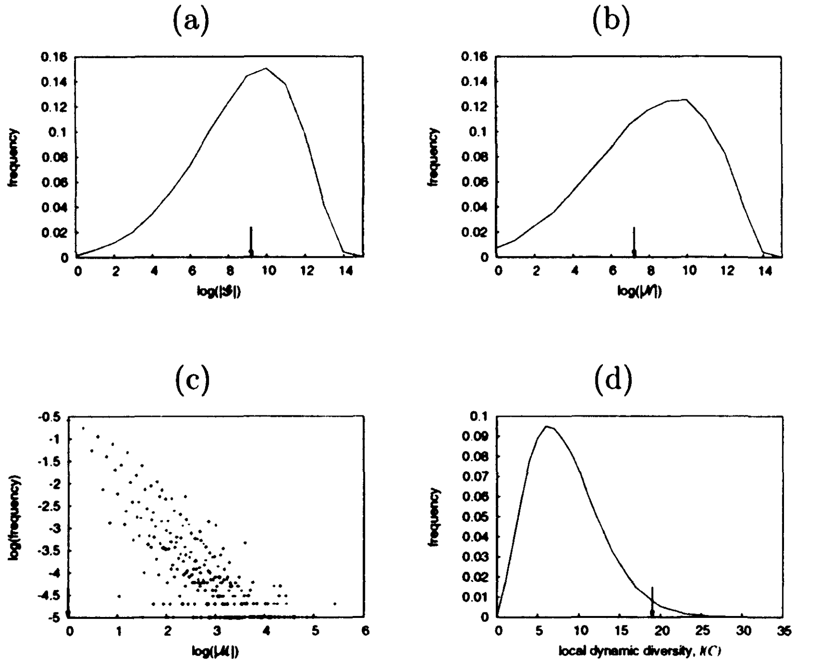


Figure 2.3: Frequency distribution of (a) log($|\mathcal{G}|$) and (b) log($|\mathcal{N}|$), *neighbourhood* size, in the population of random networks. (c) Distribution of log(frequency) vs. log($|\mathcal{M}|$ in the population of random networks. (d) Frequency distribution of $l(C)$, or *local dynamic diversity*. The arrows indicate the relevant position on the x-axis for the "known" network. All logarithms are base 10.

an idea as to how much a member network can evolve while maintaining phenotypic equivalence. The *neighbourhood* in which the *Drosophila* network exists ($\mathcal{N}_d$), contains a total of 24,187,500 (2.4 x 10^7) networks. Because $\mathcal{N}_d$ is a *neighbourhood*, this means that starting from any network in $\mathcal{N}_d$, one can reach any other network in $\mathcal{N}_d$ by going through successive *neighbours* and *without ever leaving* $\mathcal{N}_d$. The *Drosophila* network is just one member of this large *neighbourhood* $\mathcal{N}_d$, and in figure 2.4 we show a very different member of $\mathcal{N}_d$ as an example of the potential range within this set. Although it may seem surprising that a small system of just 6 genes should exhibit such a high degree of flexibility, a comparison with the population of random networks (figure 2.3b) again shows this value to be typical.

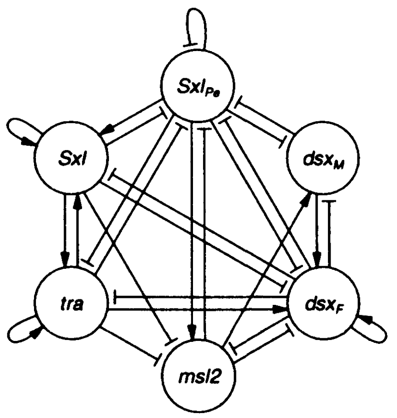


Figure 2.4: Example of a highly connected network in the *neighbourhood* $\mathcal{N}_d$. With a total of 26 connections out of a possible 36, this network sits at the opposite extreme in parsimony to the "known" network, with just 6 connections.

2.3.3 $\mathcal{M}$ represents a single network

We also found the *Most Parsimonious Solution Subset*, $\mathcal{M}$. $\mathcal{M}$ was found to contain a single network ($|\mathcal{M}| = 1, k_{\mathcal{M}} = 6$), which means no other network with 6 (or fewer) interactions can reproduce the dynamics $s_f(t)$ and $s_m(t)$ shown in Fig. 2.2. This single network matches the "known" network (Fig. 1.5, dotted box).

In order to find how common this feature is, we again compare with the population of random networks. For each random network, we determined $\mathcal{G}$ and $\mathcal{M}$ from the expression pattern pair. The distribution of $|\mathcal{M}|$ is shown in figure 2.3c (a log scale is appropriate on the frequency axis to improve visualisation of this distribution). Since 26.4% of these *most parsimonious* sets contain a single network, we can again see that this feature is typical. However, the probability that a random network should itself be the single most parsimonious network (as with the *Drosophila* network) is very small: 47 in 100,000 (0.047%). The fact that the *Drosophila* network is also the *most parsimonious* is therefore a particular characteristic of this network.

2.3.4 The *Drosophila* network has high *local dynamic diversity*

Measuring the *local dynamic diversity*, for the *Drosophila* network C_d, we mentioned previously that it has a value $l(C_d) = 19$. This means that 19 *distinct* and *viable* male/female pattern pairs can be generated from C_d and its *neighbours*. The distribution of $l(C)$ for networks C in the population of random networks is shown in figure 2.3d. The *Drosophila* network, with $l(C_d) = 19$ lies 2.4 standard deviations from the mean ($\mu = 8.456, \sigma = 4.393$), or in the top 2.54%, which indicates a second particular characteristic of this network. A logical explanation for this might be that it is related to parsimony k, since more parsimonious networks also have more *neighbours* as potential candidates (a network with no connections has $2n^2$ *neighbours* whereas a fully connected network has only n^2 *neighbours*). This would lead to a negative correlation between *parsimony* (k) and *local dynamic diversity* (l). However, the correlation between these two measurements is weak ($r = -0.133$), which appears to discard this explanation.

2.4 Discussion

We have put forward a simple theoretical framework useful for the study of gene network evolution in the short term. Two key theoretical concepts were proposed which relate to network mutations: how they change through mutation (the *neighbour*, defined as the insertion or deletion of an interaction), and how networks are mutationally related (the *neighbourhood* set). The synchronous logic model then permits us to define equivalence between two networks - defined as reproducing a given dynamics provided we start with the same initial state(s). Given a particular dynamics, an exhaustive algorithm can be used to reconstruct the entire set of equivalent networks, $\mathcal{G}$, though this algorithm is only appropriate for small systems.

It should be noted that our definition of network equivalence is conservative in that only networks reproducing the exact same expression pattern are considered equivalent. The concept of a *viable* network (as used for *lo-*

cal dynamic diversity) introduces a more liberal definition of equivalence (or neutrality [110]) which allows variation in the expression pattern while fulfilling minimal functional requirements. Because the algorithm for finding the set $\mathcal{G}$ depends upon the first definition, finding the equivalent set for *viable* networks (a superset of $\mathcal{G}$, and most likely much larger than $\mathcal{G}$) is computationally intractable. The question remains though, as to which of the two definitions of neutrality is more relevant to evolution. We cannot know whether neighbours which change the expression pattern while remaining "viable" (second definition) will not have a detrimental effect, through their connections to genes outside the sex determination system (for example, if a gene such as *tra* were oscillating). This is almost certainly not the case with neighbours in which the expression pattern remains exactly the same (first definition). The two definitions therefore represent respectively lower and upper bounds for neutrality, with evolution most likely adopting a path between these.two extremes.

We applied our theoretical framework to the *D. melanogaster* sex determination mechanism. Our strategy was to work with as small a number of genes as possible, and represent the network as a system of six nodes (derived from four genes). An expanded version of the model might have taken into consideration non-functional nodes such as "MALE TRA", the non-functional form of *tra* derived from the stop-codon containing transcript in males, or indeed genes outside the sex determination system.

Since the system of *D. melanogaster* has been extensively studied, there exists a known biological network for the chosen genes. We found $\mathcal{G}$ to be relatively large, indicating that many networks, including the "known" network, can perform sex determination by producing the same time course. Furthermore, these networks in $\mathcal{G}$ exist in large *neighbourhoods* within which any two networks in the set are mutually accessible through single network mutations. By comparing against a population of random sex determination networks, we have also been able to show that these two characteristics are general to sex determination networks, and may also therefore apply more broadly to other classes of gene network – a topic for future work. However, these general results do tell us that a great deal of plasticity is available through network mutations whilst maintaining the core function of sex determination. For example, *Cctra*, a homolog of *tra* in the medfly (*Ceratitis capitata*), has been shown both to autoregulate and act as a sex determination switch [46]. Because one *neighbour* of the "known" network in $\mathcal{G}$ adds a positive autoregulatory interaction to the gene *tra*, such a change indicates that, in an ancestral version of the network, *D. melanogaster tra* might have had this characteristic. An autoregulatory *tra* is an example of a "spurious" interaction (in that it does not qualitatively affect the behaviour) which might actually exist in *D. melanogaster*, but is unlikely to be discovered experimentally precisely because it is "spurious".

Within the set $\mathcal{G}$, the "known" network occupies a special place in that it is the network with fewest interactions (*most parsimonious*). This result is shown to be particular to the *Drosophila* network and is not necessarily to be expected, since: a) there could well be more than one most parsimonious network (as occurs in 73.6% of the random population), or b) the "known" network might contain additional interactions either for redundancy, or which participate in some other process unrelated to sex determination.

Lastly, we show that *local dynamic diversity* is high in the *Drosophila* network relative to the population of random networks, which tells us that the network has access to many pattern variations within the distance of a single network mutation. This characteristic pre-adapts the network into a good position for evolutionary adaptation of its expression pattern, while preserving its sex determination functionality. If these special characteristics of the *Drosophila* network (parsimony, *local dynamic diversity*) were found to be common to other "real" sex determination networks, then it would be worth investigating whether "real" sex determination systems are different in this respect to other classes of gene network (for example, the highly conserved embryonic pattern system). If this is the case, then this would suggest that network characteristics are major contributors to the rapid diversification observed in sex determination mechanisms.

In the case of gene networks, a dual constraint is imposed: by the required temporal dynamics on the one hand and the existing network architecture on the other. Evolution overcomes this dual constraint by taking advantage of the combinatorial nature of networks, which lends itself to creating flexibility. Although this analysis has focused specifically on the sex determination gene network of *D. melanogaster*, we hope the general method of analysis used may serve to elucidate such constraints and flexibility in other systems.

2.5 Appendix A: Definitions

Term	Description
neighbour	two networks P and C are *neighbours* if P is identical to C except for a single difference: an edge deletion (from $-1 \to 0$, or $+1 \to 0$), or edge insertion (from $0 \to -1$, or $0 \to +1$)
neighbourhood	a set of networks in which each element has at least one *neighbour* in the set
parsimony	number of connections in the network
viable	a male/female pair of expression patterns are *viable* if they fulfil the requirements of sex determination
distinct	two pattern pairs (s^f,s^m) and (s'^f, s'^m) will be considered *distinct* if there are any differences in either female expression (e.g. $s^f \neq s'^f$) or male expression (e.g. $s^m \neq s'^m$)
local dynamic diversity	the number of *viable* and *distinct* pattern pairs which can be generated from a network C and its *neighbours*

Chapter 3

Using large-scale perturbations in gene network reconstruction

3.1 Background

Recent technological advances have led to an explosive growth in high-throughput genomic and proteomic data such as DNA microarrays. The rapid growth in available data has led in turn to a need for novel quantitive methods for analysis. As a consequence of this need, the reconstruction of gene network architectures from DNA microarray expression data has become a major goal in the field of systems biology. An increased understanding of the network architectures and their respective dynamics will enable novel approaches to disease treatments by allowing us, for example, to identify drug targets *in silico* which manipulate the functional outputs of these networks. This process is expected to lead to novel classes of drug based on a network approach to cellular dynamics.

Frequently, the expression data themselves are derived from so-called experimental "perturbations" of the gene expression levels in an organism. Such experiments usually involve a treatment either at the microscopic level (e.g. over-expression of a transcription factor), or at a more macroscopic level (e.g. stress conditions, temperature shifts, and chemical treatments, which are expected to affect the gene expression level of many genes). Following the immediate changes in gene expression as a consequence of the perturbation, network connectedness will likely cause the expression level of other genes to be changed through time. In this study we use a modelling approach to simulate expression patterns for artificial networks, and consider two types of expression pattern. The first type is the simulated time-course which reflects the "normal" (unperturbed) dynamic behaviour, given a particular set of initial conditions. The second type ("perturbation") is the simulated time-course obtained as a consequence of changing those same initial conditions at random. The degree to which the initial conditions are thus changed is mediated via a tunable "perturbation intensity" parameter q (see Methods).

Although these global perturbations are frequently carried out in order to reveal causality between genes, it is not always clear how experiments should be designed so as to reveal as much causality as possible, while both minimising costly experimentation and remaining computationally tractable.

A range of computational and mathematical techniques have been adopted in the effort to find a successful gene network reconstruction technique. Reconstruction methods often have to negotiate a tradeoff between intensive (often intractable) computations, and having to perform a large number of costly experiments. Certain progress can be achieved by making simplifications, such as imposing a limit on the number of inputs to each gene, or making steady state assumptions about the system [111, 112]. Some techniques described in the literature offer efficient algorithms, but require a large number of experiments, perhaps as many as there are genes [113, 114, 115]. On the other hand, theoretical work on Boolean models has shown [116] that perhaps as few as $O(log(n))$ experiments (input/output pairs) might be required for n genes, but that to infer these relationships requires the use of computationally costly enumeration methods.

In this chapter, we propose to explore the issue of how perturbation microarray experiments might be designed, and to suggest how such experiments might be optimised so as to maximize inference capability (sex determination is not addressed in this chapter). Logic gene network models, in which gene states are represented as binary ON/OFF values, are only able to represent gene networks at a simple qualitative level, where gene expression (as measured in microarray data, for example) is clearly quantitative. However, the simplicity of logic models has enabled some progress to be made towards understanding complex systems such as gene networks, which would otherwise have been more difficult. Logic models have been used to investigate a variety of topics related to gene networks, including robustness [117], perturbation dynamics [118] and evolutionary potential [119]. This class of model forms the basis of the inference method used in this study. The inference method [119] is similar to others in which networks with a minimal number of connections are reconstructed through enumeration [120, 121]. A recent study in yeast has found that, for a sample of genes, 93% of these had between 1 and 4 known inputs [18]. In spite of the fact that the study only deals with known inputs, and also that the number of inputs per gene is likely to be larger in higher eukaryotes, it is still reasonable to say that, in general, most genes will have few inputs. Additionally, given the significant speed advantage of integer computation over floating point computation, an enumerative reconstruction method is considered to be adequate for this investigation.

In this work, we proceed by generating artificial gene networks with biologically realistic in/out degree characteristics. A gene network reconstruction algorithm is then used to study the effect on inference quality, of adding (simulated) perturbed expression patterns. The reconstruction algorithm uses an enumeration technique to evaluate up to a maximum of 4 inputs of both positive and negative sign (see Methods). Enumeration is computationally feasible on an ordinary desktop computer for medium-sized networks ($n \sim 100$), and still tractable for large networks ($n \sim 1000$), though this would require some parallelisation. The effect on inference quality is considered for two experimental parameters: a) the number of perturbations required, P, and b) the perturbation intensity, q, mentioned above.

3.2 Methods

3.2.1 Discrete dynamical model

For a system of N genes (the number of genes was n in previous chapter), the state of each gene s_i ($i = 1, .., N$) is represented by the binary values 0(OFF) and 1(ON). Additionally, each gene is assigned a default ON/OFF state $\theta_i \in \{0, 1\}$. The gene interactions are described by an $(N \times N)$ matrix C, composed of elements $C_{ij} \in \{-1, 0, +1\}$, representing the positive(+1), zero(0) or negative(-1) influence of gene j on gene i. State transitions are calculated as follows:

$$s_i(t+1) = \sigma(u_i(t)) \tag{3.1}$$

$$\text{where} \quad u_i(t) = \sum_j C_{ij} s_j(t), \quad \sigma(x) = \begin{cases} 1 & \text{if} \quad x > -\theta_i \\ 0 & \text{otherwise} \end{cases}$$

The state of the ith gene at the next timestep, $s_i(t+1)$, is therefore determined by the balance of positive versus negative inputs which are ON at the previous timestep t. If the balance is positive, then $u_i(t) > 0$ and the next state will be 1(ON). Similarly, if the balance is negative, then $u_i(t) < 0$ and the next state will be 0(OFF). If $u_i(t) = 0$ (indicating either that there are no active input connections, or that they balance out), then the default value θ_i determines the next state. This default value needs to be given *a priori*, and for the purpose of this study will be random.

3.2.2 Network inference method

Assuming we are given the gene expression pattern (time-course) $s(t)$ and the default vector θ, the inference problem is to find the necessary model parameters (the interaction matrix C) which will reproduce $s(t)$. Specifically, a system initialised at $s(0)$ should reproduce the time-course $s(t)$ for $t > 0$. Note that more than one time-course such as $s(t)$ may be defined. The time-courses will be denoted as $s^r(t)$ for $r = 0, .., P$, and will correspond to the "normal" (unperturbed) time-course $s^0(t)$, together with P perturbation time-courses $s^r(t)$. Our problem is to find at least one interaction matrix C that will reproduce all given dynamics $s^r(t)$. The problem of finding an appropriate matrix C may be broken up into N sub-problems, since each gene i may be solved independently from the others. More precisely, the inputs to gene i (i.e. C_i, the ith row of C), can be found independently of those for the other genes. Thus the search space is reduced from $O(3^{N^2})$ (if every entry $\{0, \pm 1\}$ in the $N \times N = N^2$ matrix is evaluated), down to $O(N3^N)$ (when only the 3^N entries for each of the N rows are evaluated).

Each input z^i to gene i is represented as an ordered pair (j, g), $j \in \{1, .., N\}$, $g \in \{\pm 1\}$, indicating an input from gene j of sign g. A solution $y(i)$ for gene i is a set of K inputs $\{z_1^i, z_2^i, ..., z_K^i\}$ (with $y(i) = \phi$ if $K = 0$). For K inputs there are $\binom{N}{K}2^K$ solutions to evaluate. Starting with $K = 0$ (no inputs), we progress up to a maximum of $K = 4$, exhaustively evaluating all possible solutions for each K. Making a parsimony assumption, if solutions are found for some $K_s < 4$, the method no longer continues the evaluation for $K > K_s$. Note that the method does not stop as soon as a solution is found, but evaluates all possible solutions for K_s. The failure rate (percentage of genes for which no solution was found for $K \leq 4$) never exceeded 3% of the genes in any single network for which reconstruction was attempted.

3.2.3 Global Perturbations and the perturbation intensity measure

The control time series $s^0(t)$ is generated by setting $s^0(0) = \theta$. The other time series $s^r(t)$, $r > 0$ are obtained from initial conditions which are perturbations of θ, and correspond to standard experiments such as stress conditions, or chemical treatments. Since, experimental perturbations can usually be modulated in intensity (for example, a temperature shift), this was represented using modulated artificial perturbations. Perturbed initial states $s^r(0)$ were generated by randomly changing each state $s^0(0)$ with probability q.

3.2.4 Measuring inference accuracy

Assuming one or more solutions $y_1(i), y_2(i), \ldots$ are found for gene i, these are consolidated into a solution set, $Y_i = \bigcup_l \{y_l(i)\}$. Note that some information about the solutions has been lost using this approach. For example, a solution set $Y_i^{(2)}$ obtained from a single two-input ($K = 2$) solution: $Y_i^{(2)} = \{y(i)\} = \{z_1^i, z_2^i\}$, may be equal to another solution set $Y_i^{(1)}$ resulting from two single-input ($K = 1$) solutions: $Y_i^{(1)} = \{y_1(i), y_2(i)\}$ with $y_1(i) = \{z_1^i\}$ and $y_2(i) = \{z_2^i\}$.

The consolidation process is convenient in that the solution set is easily compared with the known network structures using standard accuracy measures such as *sensitivity* and *specificity*, which are in turn defined in terms of:

1. *true positives* (TP): members of the solution set Y_i which are also true inputs (true inputs are known because the networks will be generated artificially).
2. *true negatives* (TN): members not in Y_i, which are not true inputs. Note that by *not* belonging to Y_i, they are implicitly predicted *not* to exist.
3. *false positives* (FP): members of Y_i, which are *not* true inputs (i.e. incorrect predictions).
4. *false negatives* (FN): members not in Y_i, which are really true inputs.

Two standard accuracy measures are defined as:
sensitivity = TP / (TP+FN), and
specificity = TN / (TN+FP).
Here, accuracy was measured using *sensitivity* only. The relatively large number of true negatives, makes *specificity* an uninformative statistic (see Discussion). We observe many cases where we correctly infer that there are no inputs ($Y_i = \phi$), which gives TP=FN=0. Although technically undefined, this special case will be assigned *sensitivity* 1.

Accuracy statistics were gathered from inferences performed on a large number of medium-sized random networks ($20 \leq N \leq 70$). Inferences on R random networks (each with N genes), will produce approximately RN *sensitivity* measurements (slightly fewer due to the nonzero failure rate).

Statistics reported, such as average *sensitivity* and its standard deviation, apply to all *RN sensitivity* measurements. Note that, since *sensitivity* has a maximum value of 1, the standard deviation must be zero if mean *sensitivity* is 1.

3.2.5 Artificial gene network generation

It appears to be the case in gene networks that indegree follows an exponential distribution, whereas outdegree appears to follow a scale-free distribution. More specifically, for the yeast network, the probability distribution for indegree k follows $p_k \sim C_{in}e^{-\beta k}$ with $\beta \sim 0.45$, whereas the distribution for outdegree follows $p_k \sim C_{out}k^{-\tau}$, with $\tau \sim 1$ (C_{in},C_{out} constants) [18]. As discussed above, because these results were inferred only for known interactions in yeast, these characteristics may not apply in higher eukaryotes, or indeed, even in yeast (due to unknown interactions).

Here, artificial gene networks [58] were created using the algorithm for generating directed graphs with arbitrary in/out degree distributions described in [122]. The exponential probability distribution for indegree k is given by:
$p_k = (1 - e^{-\beta})e^{-\beta k}$,
where $\beta = 0.45$ is a constant. Similarly, the power law distribution (including an exponential cutoff term which is both biologically realistic and necessary analytically when $\tau < 2$ [122]) for outdegree k is described by:
$p_k = Ck^{-\tau}e^{-\gamma k}$,
where C,γ, and $\tau = 1$ are constants. Since the algorithm begins by generating in/out-degree pairs for each node, we require equal means for both indegree ($< k_{in} >$) and outdegree ($< k_{out} >$). Following [122], we obtain expressions for the mean in/out degree:

$$< k_{in} >= \frac{e^{-\beta}}{1 - e^{-\beta}} \quad , \qquad < k_{out} >= \frac{-e^{-\gamma}}{(1 - e^{-\gamma})\ln(1 - e^{-\gamma})}$$

Since β is given, we obtain a value $< k_{in} >= 1.76$, and fit the free parameter $\gamma = 0.436$ to obtain $< k_{out} >=< k_{in} >$. Since the resulting networks are unweighted, non-zero weights ($C_{ij} \in \{-1, +1\}$) are assigned at random with probability 0.5, as in [58]. It should be noted that autoregulatory interactions can be (and indeed were) generated, and that these present no particular problem for the inference method. An example of a network which was used in the analysis is shown in figure 3.1.

3.3 Results

3.3.1 Additional perturbations improve accuracy

A discrete dynamical model was used to generate time series data from random networks (see Methods). To measure the effect of adding perturbations on inference ability, inference *sensitivity* (defined as true positives/true positives + false negatives, see Methods) was measured against P, the number of additional perturbations. Figure 3.2 shows the results for predicted solutions with one and two inputs, as well as overall sensitivity. The top graph in figure 3.2 shows that overall sensitivity is clearly enhanced by including more perturbation experiments, with lower order solutions (one and two inputs) reaching higher levels of sensitivity. Although for low P, we see *sensitivity* going down in some cases, this would appear to be a consequence of the limited number of simulations performed. The bottom graph shows the corresponding inverse relationship for the standard deviation of the sensitivity (lower for higher P).

It should be noted that the algorithm tends to underestimate the number of inputs a gene may have. This is to be expected in genes for which dynamics cannot be informative: for example, consider a gene i which has one or more negative inputs, as well as having default value OFF. Since the discrete dynamics for this gene will be the same as if it had no inputs at all (i.e. zero gene expression for $t > 0$), the presence of the inputs is impossible to infer. This underestimation effect is clear in table 3.1, which compares the distribution of inferred solution set sizes ($|Y_i|$, see Methods) with the actual solution sizes (i.e. the indegree distribution), and shows that the method is only able to produce roughly half the number of one and two input solution

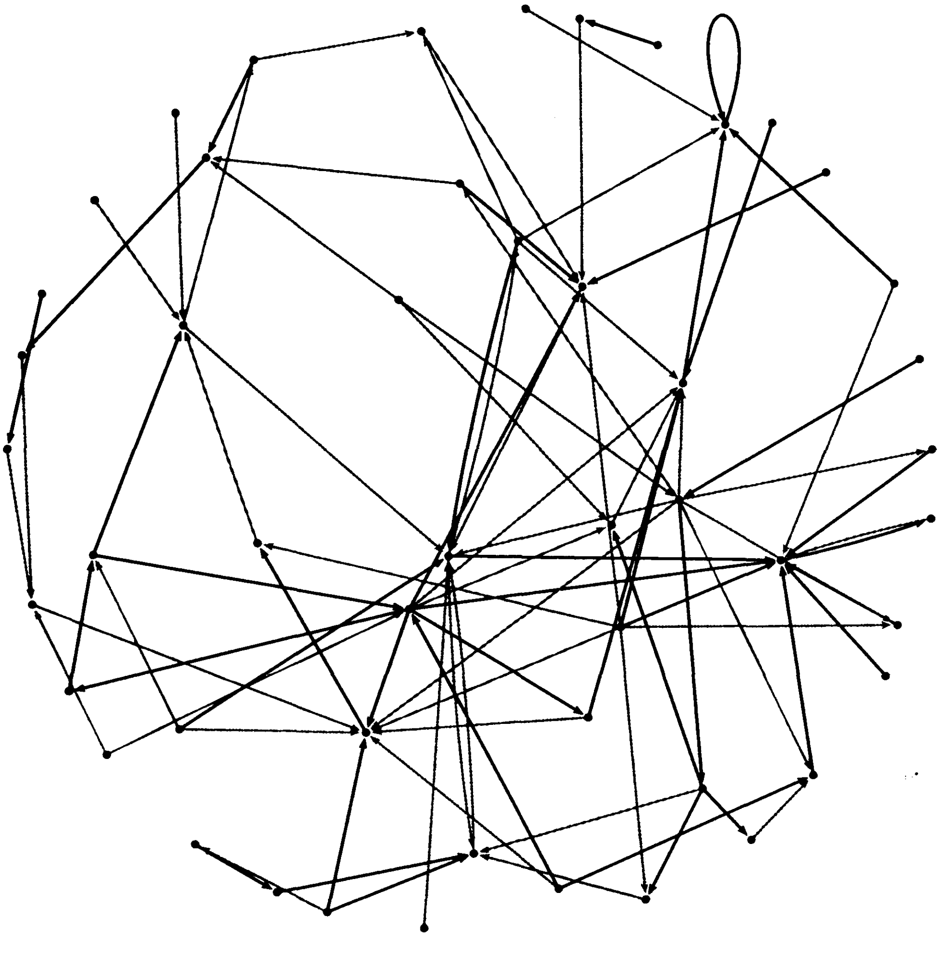


Figure 3.1: Example of an artificial gene network with $N = 50$. Positive interactions are shown in black, negative interactions in grey. Note the autoregulatory interaction on the upper right hand side. This diagram was generated using Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek).

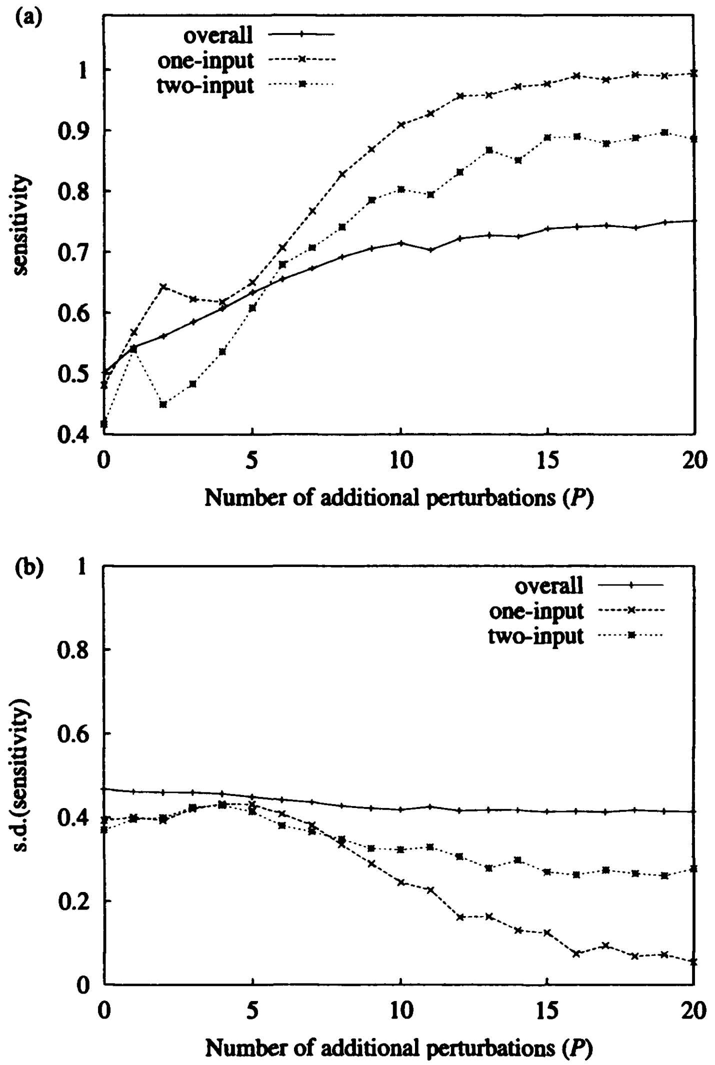


Figure 3.2: Sensitivity vs. P. (a) Sensitivity vs. number of additional perturbations used. (b) The corresponding standard deviation is shown here separately for clarity. The curves represent results for overall (i.e. all solutions) sensitivity, and specific sensitivity for (predicted) one and two-input solutions. Sensitivity is generally lower for higher order of inputs. Accuracy increases significantly with the number of additional perturbations used. The results shown are average values for 250 random networks at each data point. The remaining parameters are fixed: network size N =50, perturbation intensity q =0.5.

$\lvert Y_i \rvert$	0	1	2	3	$\geq$4
inferred	0.57	0.12	0.07	0.05	0.19
actual	0.37	0.24	0.15	0.10	0.14

Table 3.1: Solution set sizes. Distribution for the inferred solution set sizes, compared to the distribution of indegree in the actual network for the simulations. These statistics were produced from 250 random networks run using the following parameter values: $N = 50$, $P = 12$, and $q = 0.5$. The table illustrates how the algorithm overestimates the number of solutions with zero inputs.

sets that actually exist.

The increase in sensitivity with P can be explained at least partially, in the following way. Since the time series are discrete, many of the genes may have identical behaviour over time despite having different inputs (i.e. $s_i(t) = s_j(t)$ for two different genes i and j). If we define a "concatenated" time series vector $S_i = \{(s_i^0(t), s_i^1(t), ..., s_i^P(t)) : t \geq 0\}$ for gene i, and then map each gene i onto S_i, we obtain a many-to-one mapping. As we increase the number of perturbations, we might expect the number of distinct time series also to increase. We define a simple measure to quantify this mapping, $M = n'/N$ where n' is the number of distinct vectors S_i, and N is the number of genes. The maximum value of $M = 1$ indicates that the mapping of genes to time series is one-to-one, whereas lower values indicate degenerate mappings. The manner in which M increases with the number of perturbations is shown in figure 3.3 , and shows how the increase in M reflects the corresponding increase in sensitivity (figure 3.2).

3.3.2 Network size and minimum perturbation intensity

The experiments described above were repeated to consider variations in two other parameters: the network size N, and the perturbation intensity parameter q (roughly, the proportion of genes whose initial expression level is changed in each perturbation experiment - see Methods).

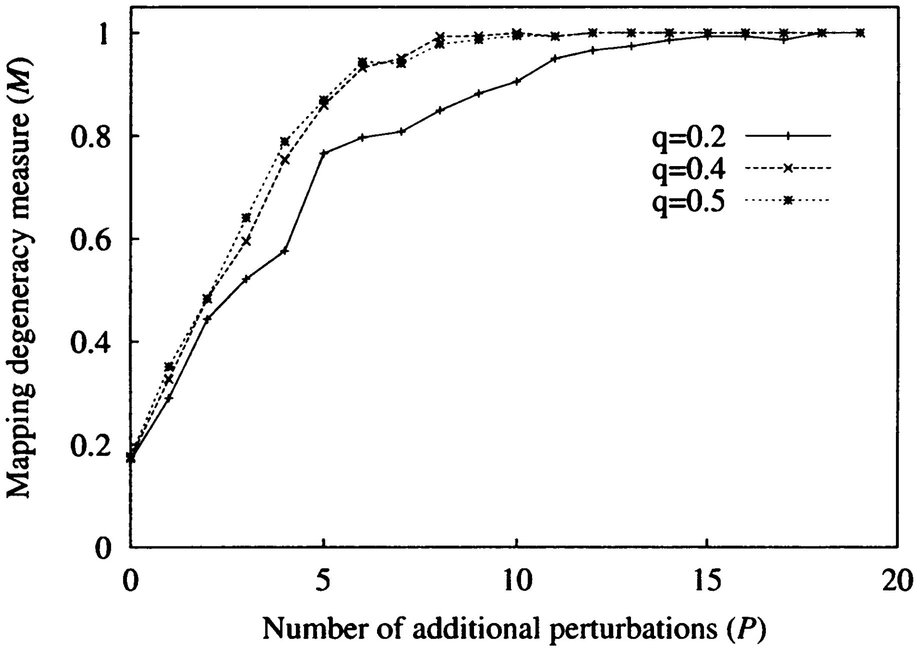


Figure 3.3: M vs. P. M (the number of distinct "concatenated" vectors S_i divided by N, the number of genes) increases in value, as the number of perturbations (P) is increased. The graph shows curves for three values of perturbation intensity q.

To consider the first case, the minimum number of perturbations P^* required to reach a given high accuracy criterion was measured for different values of the network size N. The high accuracy criterion was defined as average sensitivity, A=0.95 for one-input solution sets (A is found using a default value q=0.5 and averaging for all the $\sim$ 250N sensitivity measurements obtained from 250 random networks). To find P^*, we first find the number of perturbations P^+, such that $A(P^+) \geq 0.95$, and $A(P^+ - 1) < 0.95$. If $A(P^+) \geq 0.95$, (i.e. $A(P^*) = 0.95$ lies between $A(P^+)$ and $A(P^+ - 1)$), then P^* is estimated by simple linear interpolation.

The resulting values for P^* are shown in figure 3.4. Since the relationship is expected to be logarithmic [116], the plot shows $log(N)$ against P^* (logarithms used are base 10). A least squares best fit gives $P^* \simeq$ 1.75 $log(N)$ + 7.02, which, for N = 1000, gives $P^* \simeq$ 12.26. The relative straightness of the line shown in figure 3.4 indicates that, at least for one-input solution sets, P^* scales reliably with N, though we should remember that only a limited range for N is shown, and that, for higher N, the deviations may be larger. In order to obtain a measure of variance for P^*, we would need to calculate P^*-equivalent values for many individual networks separately, then consolidate these values to obtain the relevant statistics. However, because it was only feasible to consider medium-sized networks ($20 \leq N \leq 70$), and for any such network we often find only a small number of one-input solution sets, such statistics were found to be unreliable.

The second case (varying perturbation intensity) suggests an optimal range for q. Figure 3.5a shows the inference sensitivity over a range of values for q, and figure 3.5b shows the corresponding standard deviation. Again, inference sensitivity for one-input solutions is higher than for two-input solutions, which in turn is higher than overall sensitivity. For one-input solutions, the results show a clear peak for sensitivity in the range $0.5 < q < 0.6$. Together with a corresponding minimisation of the standard deviation in this interval (though it still remains fairly high in absolute terms), these results suggest that perturbation intensity should be in this range to optimise inference accuracy.

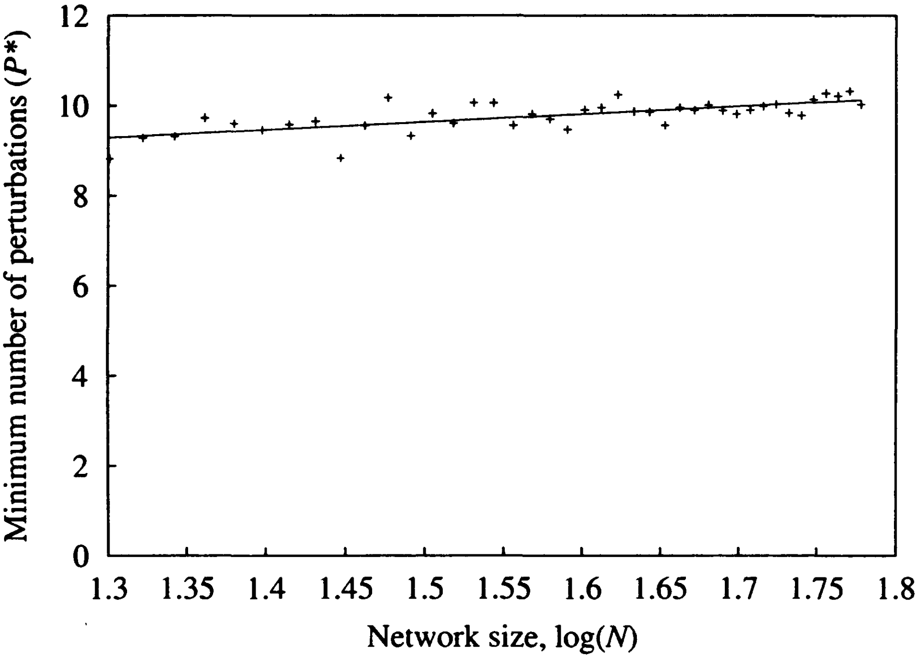


Figure 3.4: Perturbations required for high accuracy. The minimum number of perturbations (P^*) required to reach the high accuracy criterion (average sensitivity=0.95) for different values of the network size N. Each point represents the average value for 250 random networks inferred. This is equivalent to finding the value of P for which sensitivity=0.95 on the one-input curve of figure 3.2 (a) for different values of N (figure 3.2 (a) shows N=50). A linear fit is also shown.

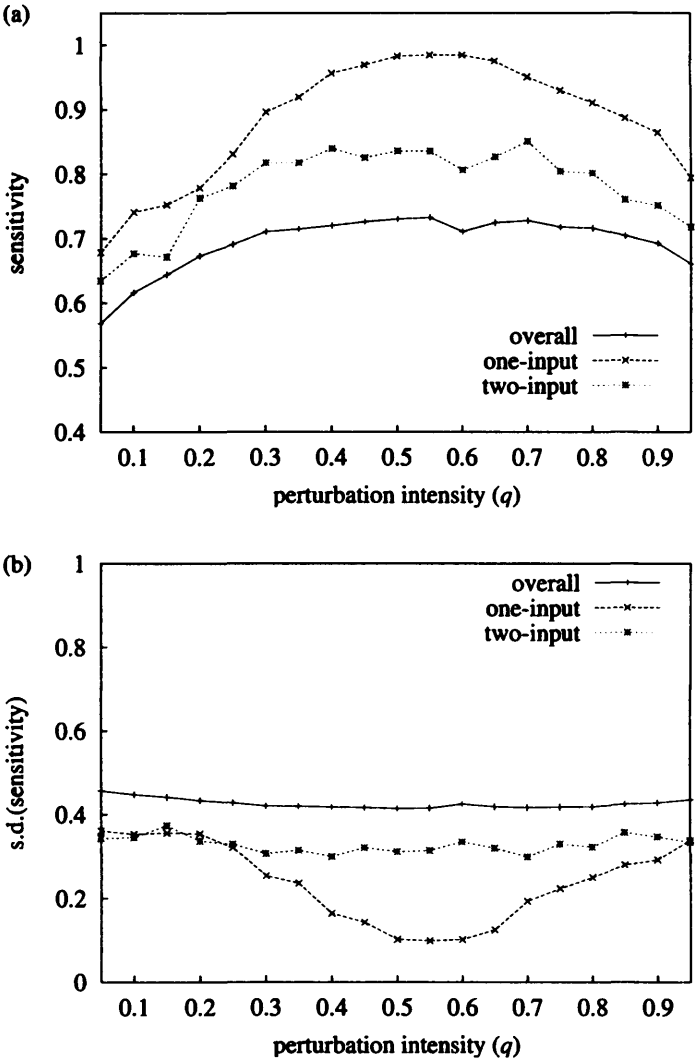


Figure 3.5: Sensitivity vs. q. (a) Average inference sensitivity vs. perturbation intensity q. (b) The variance (one standard deviation) is shown here separately for clarity. The results show sensitivity for (predicted) one and two-input solutions being generally higher than the overall case. The results shown are average values for 250 random networks inferred. The remaining parameters are fixed: network size N =50 and P =12.

3.4 Discussion

A recent analysis of the yeast genetic network has shown that 93% of genes are regulated by between 1 and 4 genes [18]. This suggests that enumerative network reconstruction methods can be useful within computationally feasible limits. Experiments involving large-scale perturbations (such as temperature shifts, chemical stress) are a standard way of obtaining time-series of gene expression data [123, 124]. A key result of [18] is that indegree appears to follow an exponential distribution, whereas outdegree follows a scale-free distribution, which has enabled the generation of realistic artificial gene networks used here. A logic model [119] was used to simulate the perturbed expression data. Subsequently, experimental parameters were considered in relation to inference accuracy, namely: a) number of perturbations required, P, and b) perturbation intensity, q.

Using *sensitivity* as the sole accuracy measure has potential drawbacks. Conventionally, both *sensitivity* and *specificity* are used to quantify the tradeoff which often exists between the two. To see why this is so, consider a prediction method which simply increases the number of positive predictions in such a way that every gene is predicted to receive inputs from all genes. Inevitably, the number of *false negatives* will become zero, due to the fact that there are no *negative* predictions at all. Because *sensitivity* does not take *false positives* into account, it will become 1 in this case, with the low accuracy being reflected in the *specificity* measure. However, as mentioned above, in this case, the large number of *true negatives* make *specificity* an uninformative statistic. More importantly though, the fact that we predict high *sensitivity* for low-order inputs (i.e. solution sets with a small number of positive predictions) indicates that the reconstruction algorithm is not boosting *sensitivity* in this spurious fashion.

The inference method itself is most useful for low order inputs, with inference accuracy maximized for predicted single input genes. More accurate methods have been proposed, though these generally require a much larger number of experiments [113, 32]. Methods such as the one proposed here, which infer relationships from expression data may well be more successful when used in conjunction with other methods such as promoter analysis [125, 126], or when used to drive experimental procedure [127]. Here, the results show that only a relatively small number of perturbations are necessary in order to achieve a substantial inference accuracy, even for large N. These relatively modest experimental requirements would presumably imply lower experimental costs. The results also suggest that the perturbations should be calibrated (by changing stress intensity, for example), so as to alter the expression levels of approximately half the genes in each experiment. Note that in this study we have represented the alteration of gene expression levels as an extreme (ON $\leftrightarrow$ OFF) change. How relevant these extreme changes are to the real biological situation remains to be elucidated. Generating perturbations which alter the expression level of half the genes at random may be difficult to achieve in practice, though experiments can be designed to come as close to this goal as possible. Even in the absence of optimal perturbations, we hope the simulation approach described here will still serve as a useful tool for planning experiments.

3.5 Appendix A: Definitions

Term	Description
solution set	Assuming one or more solutions $y_1(i), y_2(i), ...$ are found for gene i, these are consolidated into a solution set, $Y_i = \bigcup_l \{y_l(i)\}$
true positives(TP)	members of the solution set Y_i (predicted inputs) which are also true inputs
true negatives (TN)	members not in Y_i, which are not true inputs
false positives (FP)	members of Y_i, which are *not* true inputs (i.e. incorrect predictions)
false negatives (FN)	members not in Y_i, which are true inputs
sensitivity	TP / (TP+FN)
specificity	TN / (TN+FP)

Chapter 4

The early evolution of gene networks in sex determination

4.1 Background

Few studies have attempted to model the evolution of gene networks at the scale of individual genes and their mutational variants. One important reason for this is that population genetics and developmental biology have historically been separate disciplines, leading to a dearth of theoretical techniques which combine the two [128]. Sex determination mechanisms represent a good model for the study of gene network evolution since they evolve relatively rapidly [75], and are among the best understood gene networks at the molecular level in two model organisms, *C. elegans* [77] and *D. melanogaster* [76]. In addition, when modelling sex determination networks, spatial aspects can be ignored [119, 78], considerably reducing model complexity.

Discussions concerning the evolution of sex determination systems have, until recently, been limited by lack of knowledge of the molecular and genetic mechanisms involved. However, as this knowledge has accumulated, more attention has been drawn to the issue. One early study of insect sex determination suggested that the observed heterogeneity might reflect diversity only in upstream pathway genes [129]. A later analysis of the *C. elegans* sex determination pathway by Wilkins [86], postulated the hypothesis that the pathway evolved in reverse order from the final step in the pathway up to the first. It was suggested that an initial deviation in sex ratio might recruit a new determiner which favoured the minority sex as a consequence of frequency dependent selection. Under normal circumstances, a new (dominant) determiner might be expected to plateau when the minority sex reaches about 50% of the population, but if the minority sex were to overshoot in size to become a majority sex, a cycle would be initiated in which successive determiner genes are recruited, each one reversing the action of the previous one. In order for the overshoot to occur, it was suggested that either the new dominant allele is (a) tightly linked to another allele under positive selection, or (b) itself under direct selection for a trait independent of its effect on sexual development. It is conceptually difficult to see why this overshoot would occur consistently though, since it requires that two conditions (selection on sex ratio, and an additional selective advantage) be fulfilled simultaneously, and that, furthermore, they be repeated each time a new determiner is recruited. However, the core idea of reverse-order pathway evolution has since found significant support support from molecular studies [47].

More recently, in [92], the evolution of the *Drosophila* sex determination pathway was considered. In contrast to *C. elegans* (which contains a succession of negative genetic switches), the *Drosophila* system contains positive genetic switches and complexities such as alternative splicing patterns, autoregulation and stop codons, which are less easily explained. A detailed hypothesis was presented using available molecular data and standard population genetics models, with sexual selection proposed as the main driving force. Although this study was necessarily more complicated than that proposed by Wilkins for *C. elegans*, it does retain the common principle that the *Drosophila* pathway also evolved in reverse order (last to first) in stepwise fashion from a much simpler ancestral system. A consequence of Wilkins hypothesis is that there must exist a simplest-possible ancestral sex determination system, in which the primary signal and the "switch" gene are the same (single locus sex determination). We use a modelling approach to ask how the system can evolve from this simplest-possible system to a more complex two-locus system.

We develop a general approach to modelling network evolution, and consider the conditions which permit a single locus system to recruit a gene from a previously unrelated locus. In particular, we address the following questions: (1) Does recruitment of the new gene occur due to an evolutionary change in the recruited gene, or in the existing discriminatory gene? (2) Is recruitment due to the effect on downstream regulation, or is it due to the direct fitness effects of the genes involved? (3) How does the ancestral heterogamety (whether it is male or female) affect the recruitment process? We not only consider whether genes were recruited or not, but look into the necessary conditions for recruitment, specifically how the fitness contributions of the genes involved may be favourable to one sex over the other [130].

The chapter is structured as follows: We begin by describing the model components at both the network and population levels and how the two levels interact. We then describe an application of the model to explain early evolution of sex determination pathways in general, assuming the hypothesis of [86] mentioned above. Finally, we look at how our results may explain some key features of known sex determination systems.

4.2 Methods

4.2.1 Gene expression

In a network, gene expression is a result of the interaction of genes. To model this, we adopt a network-orientated definition of allele. Let an allele i have three properties:

(1) A vector $I_{ij} \in \{-1, 0, +1\}$ represents the inputs allele i can receive from other alleles j (e.g. *cis*-regulatory elements, RNA splicing sites).

(2) A scalar value $R_i \in \{0, +1\}$ represents the existence of a *regulatory* domain in allele i which can influence the expression of other alleles (e.g. the allele codes for a transcription factor domain or RNA recognition motif).

(3) A binary reduction parameter, $T_i \in \{0, +1\}$ reduces the output of allele i if $T_i = 1$.

The network dynamics for a genotype with l loci (i.e. $2l$ alleles) are described by a system of $2l$ equations, used to calculate an output S_i (between 0 and 1) for each allele i:

$$\frac{dS_i}{dt} = \sigma(u_i) - S_i \tag{4.1}$$

$$u_i = \left[\sum_{j=1}^{2l} I_{ij} R_j S_j\right] - k_T T_i \tag{4.2}$$

where $\sigma(x) = 1/(1 + e^{-ax})$ is a sigmoid function with steepness a. k_T is a global positive constant which modulates the reducing effect of T_i (when $T_i = 1$) as shown in figure 4.1. The presence of the *regulatory* domain R_j represents the potential for allele j to influence any other allele. Whether or not it actually does so (and how) depends on I_{ij}. The baseline expression level of an allele is $S_i(0) = \sigma(-k_T T_i)$. If there are no gene interactions, then $S_i = \sigma(0) = 1/2$ when $T_i = 0$, and is lower $S_i = \sigma(-k_T)$ when $T_i = 1$. Taking $S_i(0) = \sigma(-k_T T_i)$ to be thè initial values at time $t = 0$, the steady state expression levels $\hat{S}_i$ for each allele, can be estimated in a standard way for ODEs (see appendix A).

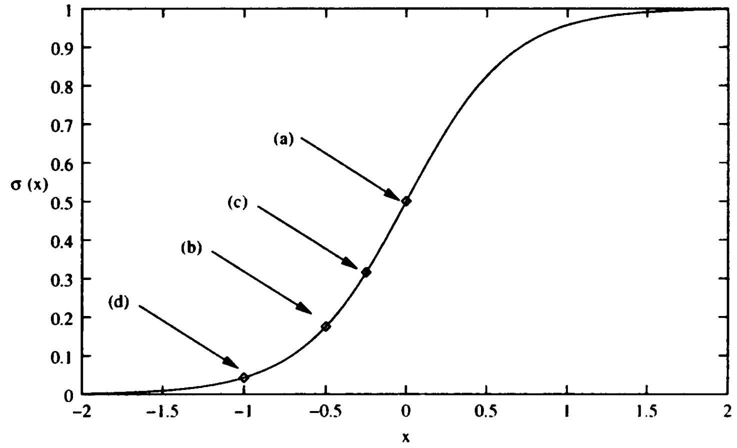


Figure 4.1: The graph shows the sigmoid curve $\sigma(x)$ with default steepness $a = 3.1$. Four points on the curve are highlighted: (a) output $S_i = \sigma(0) = 1/2$ when $T_i = 0$. Output $S_i = \sigma(-k_T)$ when $T_i = 1$, (b) with the default value $k_T = 1/2$, (c) the low value $k_T = 1/4$, and (d) the high value $k_T = 1$.

Default values used are $a = 3.1$ and $k_T = 0.5$. Since the value of these parameters may significantly affect results, high and low values for each are also considered. For a, we chose two extremes from (dimensionless) reported measurements for steepness [6]: $a = 1.6$ (low), and $a = 4.6$ (high), with the default ($a = 3.1$) at the midpoint of the two extremes. Low and high values for k_T were chosen above and below the default value at $k_T = 1/4$ and $k_T = 1$, as shown in figure 4.1.

The reducing term $-k_T T_i$ is intended to encompass a range of mutations which might quantitatively reduce output: for example, (a) a mutation which reduces the binding efficiency of a transcription factor in a cis-regulatory element, or (b) a mutation at a cryptic splice acceptor site which increases the probability of a stop codon being included in the transcripts. Crucially, the reduced-output effect of mutations such as these can often be counteracted by an appropriate input: for example, a mutation in the corresponding transcription factor or RNA-binding protein.

4.2.2 Sex determination

Each diploid genotype determines a sex (male or female). This depends on the combined expression level of the two alleles (*f1* and *f2*) at the F locus, $\hat{S}_F = \hat{S}_{f1} + \hat{S}_{f2}$. We assume that if $\hat{S}_F > \theta$, where θ is an expression threshold, then the genotype is female, otherwise the genotype is male. This system mirrors the way *doublesex* expression is used to determine somatic sex in *Drosophila* and a number of other insects [131] (in *Drosophila*, activation of the sex determining pathway leads to a female phenotype as a consequence of high expression of *dsxF*). Other genetic loci do not directly affect sex determination, though they can alter gene expression at the F locus and so indirectly alter sex.

4.2.3 Fitness

The overall fitness of the genotype is calculated by amalgamating contributions from each allele. Fitness in this case is understood to be proportional to the probability of survival and rate of reproduction of the genotype. The fitness effect of an allele i at a locus $L(i)$ is the product $\hat{S}_i w_{L(i)}$, where $w_{L(i)}$

defines the locus-specific contribution to fitness. Fitness is calculated separately for the two sexes, $w^{♂}_{L(i)}$ and $w^{♀}_{L(i)}$, which allows alleles to have different fitness in each sex, for example to be beneficial in one but deleterious in the other [130]. Since the parameters $w^{♂}_{L(i)}$ and $w^{♀}_{L(i)}$ will not usually be known, the fitness contribution of each locus will be sampled from a Gaussian distribution $N(0,1)$ (i.e. with mean 0 and standard deviation 1) unless stated otherwise. Alleles also have indirect fitness effects via the gene network by altering the expression level of other alleles. The fitness of a genotype $W(g)$ is the sum of the $2l$ allelic contributions,

$$\text{If} \quad w(g) = \sum_{i}^{2l} \hat{S}_i w_{L(i)} \qquad \text{, then} \qquad W(g) = \begin{cases} w(g) & \text{if} \quad w(g) \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

A non-negative $W(g)$ value is required, as negative values would lead to negative populations in the evolutionary dynamics (see below). Note that, since $\hat{S}_i$ is positive, at least one of the $w_{L(i)}$ values must be positive for $W(g)$ to be positive.

A slight complication must be introduced to deal with the F locus, which represents the *dsx* locus in our model. This gene undergoes alternative splicing into female *dsxF* or male *dsxM* mRNA forms. High levels of *dsxF* and low levels of *dsxM* contribute positively to female fitness, whereas the reverse is true for male fitness ([76]). To model this we simplify, and measure the effective gene expression for an allele i at the locus F, as $\hat{S}_i$ if the sex is female, and 1-$\hat{S}_i$ if the sex is male.

4.2.4 Mutation

Network mutations alter the interactions between genes in the sex determination system. We use the concept of a network *neighbour* to define possible network mutations [119]. Let i^* be a *neighbour* of an allele i if it fulfils exactly one of the following three conditions: a) there is a single difference in the *input* vector I_{i^*j}, which can be either a link deletion (from $-1 \rightarrow 0$, or $+1 \rightarrow 0$) or an insertion (from $0 \rightarrow -1$, or $0 \rightarrow +1$); b) there is a change in *regulation* R_{i^*}, which can be either a loss (from $+1 \rightarrow 0$) or gain (from $0 \rightarrow +1$) of a regulatory domain; or c) there is a change in the reduction parameter T_{i^*} $(0 \leftrightarrow +1)$. Since these parameters are discrete, there are a finite number of *neighbours* for each allele. Limiting mutation to *neighbours* makes the mutational process incremental and precludes macromutations which change many interactions in one event or reversals in the *input* vector ($-1 \leftrightarrow +1$ not allowed in I_{i^*j}). For the purpose of this study we did not consider mutations leading to autoregulatory interactions $I_{ii} = \pm 1$ (see discussion).

4.2.5 Evolutionary dynamics

We start from the simplest-possible "ancestral" population, in which sex determination is controlled by a single locus F. The population contains one male and one female genotype, determined by two alleles at the F locus: f, the normal allele ($T_f = 0$), and m, a reduced-output allele ($T_m = 1$). There are only two possible "ancestral" populations using these alleles. In the first (*ff*/*mf* ancestral heterogamety), m is a dominant masculinizing allele, males are heterozygous *mf* and females are homozygous *ff* . In the second (*mm*/*mf* ancestral heterogamety), *f* is a dominant feminizing allele, females are heterozygous *mf* and males are homozygous *mm* . If the initial male and female expression levels at the F locus are $\hat{S}_F^{\male}$ and $\hat{S}_F^{\female}$ respectively, then the sex threshold θ for each experiment is chosen randomly from a Gaussian distribution with mean $\bar{S}_F = (\hat{S}_F^{\male} + \hat{S}_F^{\female})/2$, and standard deviation $(\hat{S}_F^{\female} - \hat{S}_F^{\male})/8$, sufficiently small that θ rarely occurs above $\hat{S}_F^{\female}$ or below $\hat{S}_F^{\male}$ (note that $\hat{S}_F^{\female} > \hat{S}_F^{\male}$ is a requirement for all ancestral populations).

Setting the mean value for θ to the mid-point $\bar{S}_F$ between $\hat{S}_F^{\female}$ and $\hat{S}_F^{\male}$, is consistent with our assumption of an unbiased (1:1) sex ratio. The reason for this is that there is likely to be certain variability in the expression level of both $\hat{S}_F^{\female}$ and $\hat{S}_F^{\male}$. Although in this model, the steady-state levels of $\hat{S}_F^{\female}$ and $\hat{S}_F^{\male}$ are deterministic, in reality these levels will contain some degree of stochastic variability. As a consequence, any mutation moving θ away from $\bar{S}_F$ (and closer to $\hat{S}_F^{\female}$ or $\hat{S}_F^{\male}$) would probably cause an imbalance in the sex ratio. According to Fisher's classic theory for the unbiased sex ratio [132], a counteracting mutation would then be expected to move θ back towards $\bar{S}_F$.

Therefore, it is reasonable to assume that the mean value for θ will stabilise around $\bar{S}_F$. Initial populations where $\theta > \hat{S}_F^{\text{♀}}$ or $\theta < \hat{S}_F^{\text{♂}}$ are discarded (i.e. they not considered *viable* - see definition below).

Note that the two forms of ancestral heterogamety *ff*/*mf* and *mm*/*mf* are not symmetrical, as a consequence of the fact that the *f* and m alleles have mid-level ($T_f = 0$) and low-level ($T_m = 1$) outputs respectively (rather than high and low levels). As discussed above in section 4.2.1, our intention here is to model particular mutant alleles which may quantitatively reduce output (e.g. a mutation at a cryptic splice acceptor site), and which have been suggested to play an important role in the evolution of sex determination [92]. This class of mutant allele is represented here by the allele m, The "normal" (non-mutant) case is represented by the allele f, which can upregulated or downregulated in equal measure by a single interaction (see figure 4.1).

The "ancestral" population also contains an additional locus A, that is monomorphic for the allele a . This allele has no regulatory interactions with alleles at the F locus. With the "ancestral" population as a starting point, we proceed by introducing random mutations (mutant alleles), and evaluate the consequences. For purposes of clarity, the procedure will be explained in terms of two nested loops, as shown in figure 4.2: (a) an "outer loop" which determines which mutations are tested as well as when the process should terminate, and (b) an "inner loop" which follows the fate of each mutation.

The outer loop

The outer loop generates the mutations, tests the effect of inserting each mutation into the population, and determines when the procedure should terminate. The outer loop is delineated in figure 4.2(a). The starting point of the outermost loop is an input population, which in the first instance will be the *ff*/*mf* and *mm*/*mf* populations described above.

Mutations are generated and chosen as follows: First of all, a "parent" allele needs to be selected for mutation. This is done by: (1) choosing a random locus, (2) choosing a "parent" genotype g in proportion to its frequency in the population, (3) selecting an allele (if heterozygous) at random. With

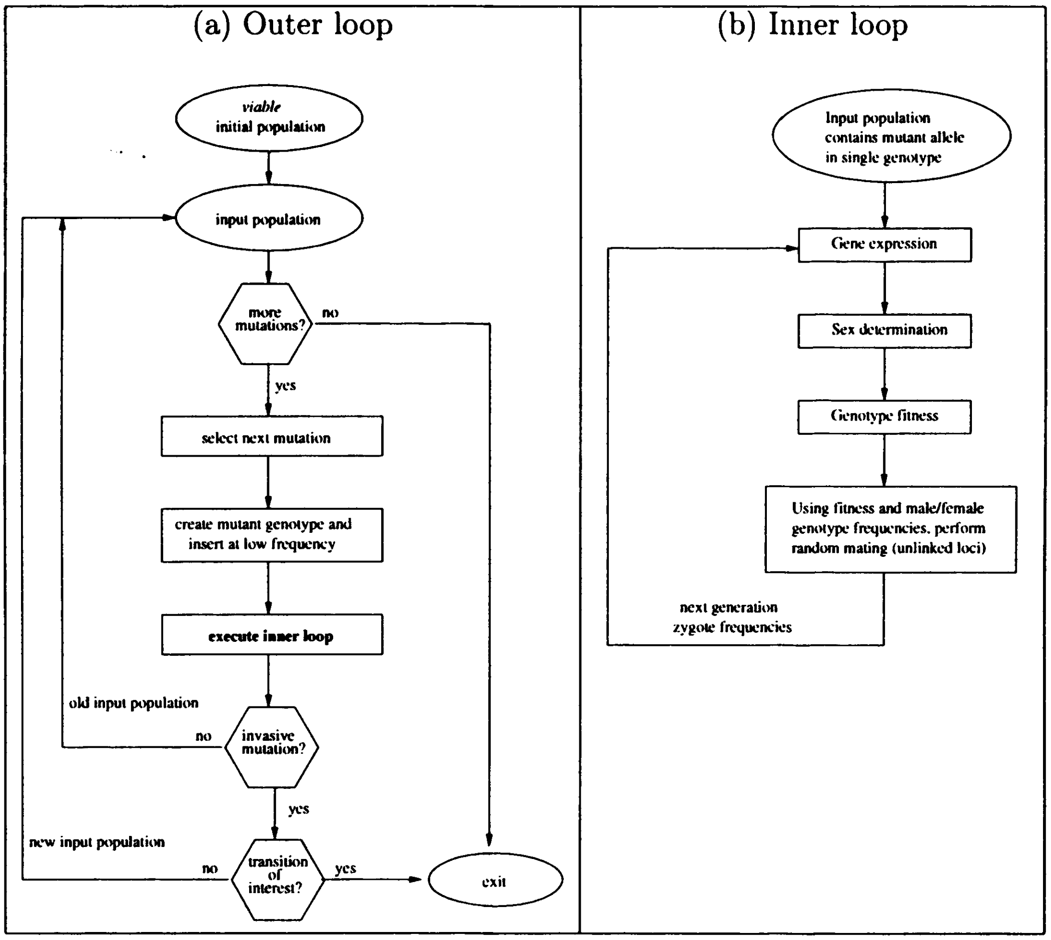

(a) Outer loop
viable initial population
input population
more mutations?
no
yes
select next mutation
create mutant genotype and insert at low frequency
execute inner loop
old input population
no
invasive mutation?
yes
new input population
no
transition of interest?
yes
exit
(b) Inner loop
Input population contains mutant allele in single genotype
Gene expression
Sex determination
Genotype fitness
Using fitness and male/female genotype frequencies, perform random mating (unlinked loci)
next generation zygote frequencies

Figure 4.2: Overview of simulation procedure: (a) Outer loop handles mutation generation and testing, and (b) Inner loop follows the fate of each mutation in the input population.

the "parent" allele selected, (4) a mutation of this allele is chosen at random. The selections for steps (1) and (4), i.e. loci and mutations, are taken from randomly shuffled lists, so that they can be evaluated iteratively until all are exhausted (this is represented by the decision block "more mutations?" in figure 4.2(a))

Once the mutant allele is generated, it is introduced into the population in the form of a mutant genotype. The mutant genotype will contain the mutant allele, but is otherwise identical to the "parent" genotype. If the selected allele was homozygous, the mutant allele is introduced in heterozygous form. The mutant genotype is introduced at low frequency (1% of "parent" genotype frequency).

The next step is the inner loop simulation, which will be explained in more detail below. It is sufficient to say at this point that the inner loop will evaluate the effect of inserting the mutant genotype into the population. The inner loop simply decides whether the mutant allele is invasive or not. Non-invasive mutations are of little interest and are discarded. However, if the mutation was invasive, then the resulting population may represent a transition of interest (a change in sex determination locus or a change in heterogamety), in which case, it is recorded and the process terminates. Populations containing other types of invasive mutations are reintroduced as input populations recursively to maximum depth of 3 (i.e. up to 3 invasive mutations in succession can be tested). The process of testing successive mutations is represented by the outermost loop of figure 4.2(a). Tests with recursion depth greater than 3 did not qualitatively change the results. Note that the global random parameters (θ, $w_{L(i)}$) are generated together with the initial population, remaining constant throughout (the outer loop process).

This procedure attempts to maximise the probability of finding an invasive mutation without introducing bias in the choice of locus, genotype or allele, and was adopted for two reasons of efficiency. Firstly, given that in practice most mutations are non-invasive, and that there are a finite number of mutations for any given allele, all allele mutations can be generated once, then tested without repetition. Secondly, because the procedure increases the probability of finding an invasive mutation, it allows efficient evaluation of successive invasions (see below). Since only invasive mutations are considered (non-invasive mutations are discarded), we should emphasize that the results do not describe the relative probability of invasive vs. non-invasive mutants, but they do allow a comparison of the different starting conditions, which is the objective.

The inner loop

The "inner loop" follows the fate of each mutation, and takes the form of a standard population genetics simulation (see Appendix A), as used in [92]. The model assumes an infinite diploid population with nonoverlapping generations, as well as a constant 1:1 (male:female) sex ratio.

As explained above, the mutant genotype is introduced at low frequency (1% of "parent" genotype frequency). The initial zygote genotype frequencies are therefore given. At each generation the following steps are taken: gene expression ($\hat{S}_i$) is calculated for each genotype (see section 4.2.1), sex determination (male or female - see section 4.2.2), genotype fitness (see section 4.2.3). With the adult male and female genotype frequencies, as well as the corresponding fitness of each genotype, random mating amongst adults assuming unlinked loci is used to define the zygote frequencies in the following generation.

The outcome of the inner loop process is a population which has reached equilibrium (see Appendix A). The frequency of the mutant allele is then calculated and classified as "invasive" if it has grown in frequency, "non-invasive" if it has not.

The choice of inserting the mutant genotype at a "low frequency" corresponding to 1% of "parental" genotype frequency may seem arbitrary. However, as explained in appendix A, the outcome of the inner loop simulation process is to discover whether the mutant allele is invasive or non-invasive *relative to the frequency at which it was inserted.* If the mutant genotype (together with any other genotypes which may be generated) have either neutral or deleterious fitness relative to the parent genotype, the mutant allele frequency will not increase and the population will be discarded. For invasive mutant alleles, the same equilibrium will be reached for any reasonable choice of "low frequency". The particular choice of 1% is therefore of no consequence.

Iterations for global random parameters

By iterating the entire process (outer/inner loops), we measure the frequency and average parameter values required for alterations in the sex determination system. At each iteration new random values for the global random parameters (θ, $w_{L(i)}$) are generated. We are interested in measuring the frequency of: (a) transitions to a two locus system from a single locus system, (b) changes in heterogamety at the ancestral discriminatory locus (e.g. from *mf* to *ff* females) (c) the existence and importance of fitness bias underlying the preceding outcomes. The frequency of events (a) and (b) is defined relative to the total number of *viable* initial fitness conditions. That is, we do not count runs in which the fitness contributions of each locus resulted in $W(g) = 0$ in either sex in the initial population. Fitness bias refers to the statistical differences in w values leading to events (a) and (b), relative to those which are simply *viable*.

4.2.6 Evolution from single locus sex determination

For sex determination to involve the A locus, a new regulatory connection must be formed with the F locus. This connection can be formed by a single mutational step in a limited number of ways (figure 4.3):

1) A pre-existing *regulatory* domain is present at the A locus ($R_a = 1$), to which a mutant f or m allele subsequently gains a positive (or negative) *inbound* link, which can happen in four ways: if $I_{fa} \rightarrow +1(-1)$ creating a mutant allele ${}^{+}f({}^{-}f)$, or if $I_{ma} \rightarrow +1(-1)$ creating a mutant allele ${}^{+}m({}^{-}m)$.

2) A pre-existing *inbound* site is present at the F locus (four possibilities: $I_{fa} = \pm 1$, or $I_{ma} = \pm 1$) to which a mutant allele a^{+} subsequently creates the corresponding *regulatory* link (i.e. $R_a \rightarrow 1$).

These five pre-existing cases will be denoted R_a^+,I_{fa}^+,I_{fa}^-,I_{ma}^+, and I_{ma}^- as in figure 4.3.

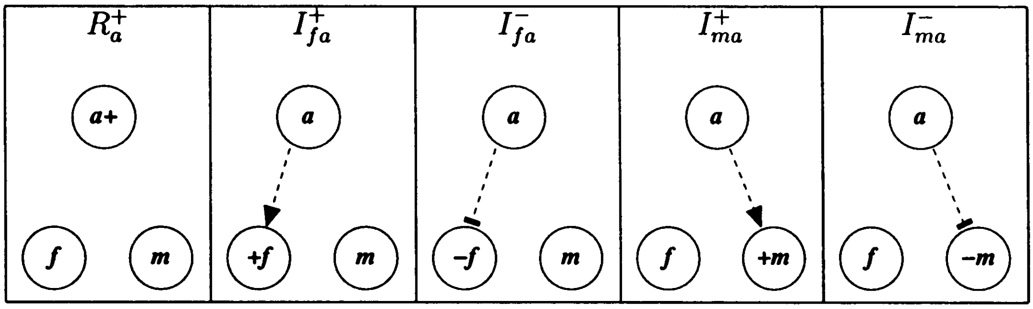


Figure 4.3: Five possible cases of pre-existing conditions which might permit a single-locus sex determination system to evolve to a two-locus system. At the top of each box, the pre-existing condition is indicated with a graphical representation below. Since *inbound* links are both signed and allele-specific, these are indicated by the dotted lines.

Separating connections into *regulatory* domains and *inbound* sites in this way is intended to reflect key aspects of regulatory changes suggested in [92]. Given the importance of regulatory changes mediated by alternative splicing, we have ensured this class of regulatory changes are properly represented in the model. One such mutation suggested in [92], causes an ancestral form of *Sxl* to be recruited into the pathway. This hypothetical mutation turns a non-functional *tra* allele (*traS*) to a functional form by causing a stop-codon-containing exon to be spliced out of the final transcripts. In our model, this mutation would be equivalent to a mutation which activates a *regulatory* domain (e.g. $R_a : 0 \longrightarrow 1$), creating a link to a pre-existing positive *inbound* link (e.g. I_{fa}^+). It is straightforward to imagine a negative counterpart of this mutation which creates a link to a negative *inbound* site, thus increasing the probability of the stop-codon-containing exon being *included.*

In addition to the pre-existing cases concerning regulatory connections, two other conditions regarding the *a priori* condition of the A locus will also be tested:

The first condition relates to the signal strength of the *a* allele. We recall that, in the absence of inputs, the expression level S_a will be stronger if $T_a = 0$ than if $T_a = 1$. In order to assess the importance of *a priori* expression level for recruitment, both the strong ($T_a = 0$) and weak ($T_a = 1$) forms will be evaluated, with strong being the default used.

The second condition relates to the fitness parameters at the A locus ($w_A^{\male}$ and $w_A^{\female}$). From a biological standpoint, it is reasonable to assume that genes which are not yet involved in the sex determination pathway (the vast majority of genes) have no *a priori* reproductive fitness bias. We are therefore interested in analyzing the situation for which the fitness bias at the A locus is "clamped" to zero ($w_A^{\male} = w_A^{\female} = 0$). This condition will be denoted as "clamped A", and will be the default condition for the A locus. Under these conditions, A does not make a direct contribution to fitness, but can only do so indirectly through regulation of the F locus. Note also that both $w_F^{\male} > 0$ and $w_F^{\female} > 0$ are needed for a viable initial population. The alternative condition, where $w_A^{\male}$ and $w_A^{\female}$ are sampled from a Gaussian distribution $N(0, 1)$ will be referred to as "unclamped A".

20,000 trials were attempted for each pre-existing regulatory case (R_a^+, I_{fa}^+, I_{fa}^-, I_{ma}^+, and I_{ma}^-), across different parameter values and *a priori* conditions, both for male (*ff*/*mf*) and female (*mm*/*mf*) ancestral heterogamety. We record the trials resulting in a transition of interest (either a change in the sex determination locus or a heterogamety switch) and compare to those which were simply viable for each ancestral heterogamety and case combination. Parameter values used are default unless otherwise stated. Values used for all parameters are summarised in table 4.1.

par.	description	default value	alternative value(s)
a	sigmoid slope	3.1	1.6 (low), 4.6 (high)
k_T	reduction factor	1/2	1/4 (low), 1 (high)
T_a	*a* allele output	0 (strong)	1 (weak)
w_A	A locus fitness	clamped	unclamped (random)
w_F	F locus fitness	random N(0,1)	
θ	sex-det threshold	random N($(\hat{S}_F^{\male} + \hat{S}_F^{\female})/2$, $(\hat{S}_F^{\female} - \hat{S}_F^{\male})/8$)	

Table 4.1: Summary of parameter values described. The first four parameters are shown with default and alternative values, whereas the last two parameters are always random and therefore do not have default/alternative values.

Because, as discussed above, we have chosen a method which attempts to maximise the probability of finding an invasive mutation, we will refer to the relative frequency of transitions of interest as a “transition probability” (in quotes). As used here, the concept of “transition probability” is therefore closer to “the probability that the initial population is such that, when all possible mutations are exhaustively tested, at least one of these will be found to be capable of changing the population to a transition of interest, understood as a change in the sex determination locus or a heterogamety switch”.

4.3 Results

4.3.1 Mutation at A causes locus transition

The “transition probabilities” to a two locus sex determination system, in which control has passed to the new locus A, are shown in table 4.2. The maximum “transition probability” of 1 indicates that for every *viable* ancestral population, every time the mutation selection method is applied (potentially for successive mutations), it provokes a locus transition. Note that the "transition probabilities” may in turn represent (a) one or more transitions where the final evolved population is the same, but where the intermediate steps were different or, (b) one or more transitions where the final evolved populations were different. However, we find that most "transition probabilities” are represented overwhelmingly by a single transition, which should be assumed to be the case where not stated otherwise.

We only observe high “transition probabilities” for certain cases where a pre-existing *inbound* link is present at the F locus (in particular pre-existing cases I_{fa}^{-} and I_{ma}^{+}, see figure 4.3). These cases require a mutation at the A locus (in the form of the mutant allele a^{+}) to create the necessary *regulatory* link. In the vast majority of cases, this single mutation is sufficient for the transition to occur.

In contrast, if we look down the column for case R_{a}^{+} (i.e. when a pre-existing *regulatory* domain is present at the A locus), the results show only low "transition probabilities". There is a fairly obvious reason of parsimony

Anc. het.	case:	R_a^+	I_{fa}^+	I_{fa}^-	I_{ma}^+	I_{ma}^-
ff/*mf*	defaults	0.04		1.00	1.00	0.08
	low a			1.00		
	high a	0.04		1.00	1.00	
	low k_T	0.04		1.00	1.00	0.08
	high k_T			1.00		
	weak ($T_a = 1$)			0.14		
	unclamped A	0.05		0.69	0.26	0.01
mm/*mf*	defaults			1.00	1.00	
	low a			1.00	1.00	
	high a	0.08		1.00	1.00	
	low k_T	0.03		1.00	1.00	
	high k_T				0.28	
	weak ($T_a = 1$)	0.01				
	unclamped A	0.06	0.01	0.33	0.60	0.05

Table 4.2: "transition probabilities" for changes in sex determination locus for each ancestral heterogamety and case. The rows show results for clamped A (default parameter values), with low/high values for parameters a (sigmoid slope) and k_T (reduction factor), with weak signal strength ($T_a = 1$), as well as for unclamped A. Blanks indicate zero.

for this deficit, because mutations in both loci are required. To see this, consider a mutation that occurs at the F locus recruiting a^+. This must be followed by a second mutation in a^+ to create the discriminatory signal, before f can become homozygous in both sexes (by driving out m). This outcome is possible, but the choice of mutation in a^+ is very limited, as only a change in the signal strength of a^+ has any effect on the system, i.e. strong ($T_{a^+} = 0$) → weak ($T_{a^+} = 1$).

Together, these results suggest that the mutations leading to expansion of the sex determination system are far more likely to have been *upstream* (i.e. in the coding region of the recruited gene A), rather than *downstream* (i.e. in the promoter or splicing sites of the F locus). It is quite reasonable to assume such mutations are common. For example, it has been suggested that recruitment of *Sxl* into the *Drosophila* sex determination pathway was due to the occurrence of a mutant allele at the *Sxl* locus [92]. By blocking a splice acceptor site in its downstream target *tra*, the hypothetical mutant transformed a nonfunctional *tra* allele into a functional form.

4.3.2 Locus transition requires strong pre-existing signal at A

In table 4.2, we observe that when the pre-existing signal strength for the allele a is weak, i.e. the entry labelled "weak ($T_a = 1$)", "transition probabilities" are also low. The strong ($T_a = 0$) signal for a is the default value used for all other rows. This shows that mutations which initially have a weak effect on F are less likely to be recruited, suggesting that particular kinds of mutations representing strong-signal mutants (e.g. wholesale acquisition of an appropriate RNA-binding domain), are more likely to induce recruitment than weak-signal mutants (e.g. small-effect point mutations).

The reason for this is clear. Consider the successful locus transitions shown in figure 4.4(a) and (b). In both cases the evolved male pathway requires a suppression of the ^-f allele such that both (1) the new phenotype be male, and (2) it has higher fitness than the previous male. This dual requirement suggests a weak interaction from a^+ will not be sufficient, and this is indeed what the simulations confirm. Similar arguments apply for the transitions shown in figure 4.4(c) and (d).

4.3.3 Mutations at A which downregulate f or upregulate m are most likely to provoke a locus change

Two specific pre-existing cases are most likely to provoke a change in sex determination locus:

1) Case I^-_{fa}, where the a^+ mutation at A downregulates f.
The two transitions observed for this case, under default conditions, are shown in figures 4.4(a) and (b) for male (*ff*/*mf*) and female (*mm*/*mf*) ancestral heterogamety respectively. Both transitions lead, in a single step, to the same population on the right hand side in which males take over the pathway by downregulation of the pre-existing allele ^-f. Female becomes the default sex.

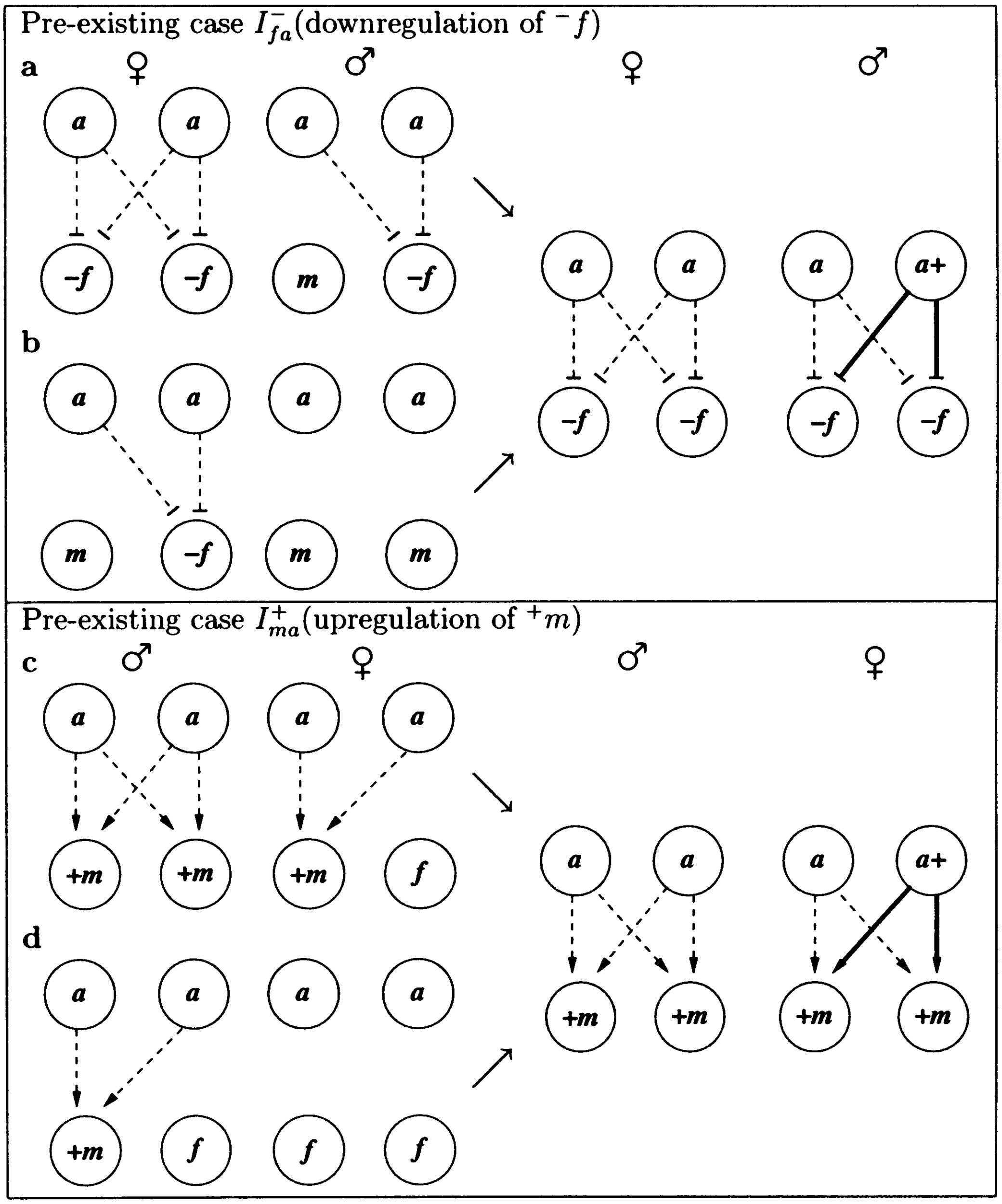
2) Case I^+_{ma}, where the a^+ mutation at A upregulates m.
Similarly, the two transitions observed for this case, under default conditions, are shown in figures 4.4(c) and (d) for ancestral heterogamety *mm*/*mf* and *ff*/*mf* respectively. Again, these are single-step transitions leading to a single population, though now females acquire the pathway by upregulating the ^+m allele, with male becoming the default sex.

As can be seen in table 4.2, for certain non-default parameter values, such as high k_T, the transitions in figures 4.4(b) and (d) are not observed, though the others, shown in figures 4.4(a) and (c), are. However, where nonzero "transition probabilities" appear in columns I^-_{fa} and I^+_{ma}, the observed transitions were qualitatively equivalent to the corresponding transitions shown in figure 4.4.

4.3.4 Other mutations at A are unlikely to provoke a locus change

As can be seen in table 4.2, cases I^+_{fa}, (where the mutation a^+ at A will *upregulate* f) and I^-_{ma} (where the same mutation will *downregulate* m) exhibit low "transition probabilities". There is a good parsimony explanation as to why these pre-existing states do not permit a change in control to the A locus in a single step, as follows:

Consider the pre-existing positive inbound site under case I^+_{fa}, which is then upregulated by a mutant a^+ allele. In a female heterogametic population (figure 4.5a), the a^+ allele has no effect on sex determination as f is a dominant feminizer. In a male heterogametic population (figure 4.5b), the a^+ allele is likely to change the sex of the double heterozygote $(aa^+; mf)$ from male to female. However, both F locus homozygotes will be the same sex, as follows: $aa; ff$ and $aa^+; ff$ are female, and $aa; mm$ and $aa^+; mm$ are male. So in both cases, even if selection at the A locus favours the a^+ allele, it can't drive one of the F alleles to fixation. Secondary mutations are required for a transition of the sex determination system. The iterated results show that these intuitive arguments are correct, since neither pre-existing case (I^+_{fa}or I^-_{ma}) gave a significant transition rate.



Pre-existing case I^{-}_{fa}(downregulation of ^{-}f)
a
♀
♂
♀
♂
a
-f
m
a+
b
Pre-existing case I^{+}_{ma}(upregulation of ^{+}m)
c
♂
♀
♂
♀
+m
f
d

Figure 4.4: Most commonly observed transitions in which the discriminatory sex determination signal changes from the F locus to the A locus.

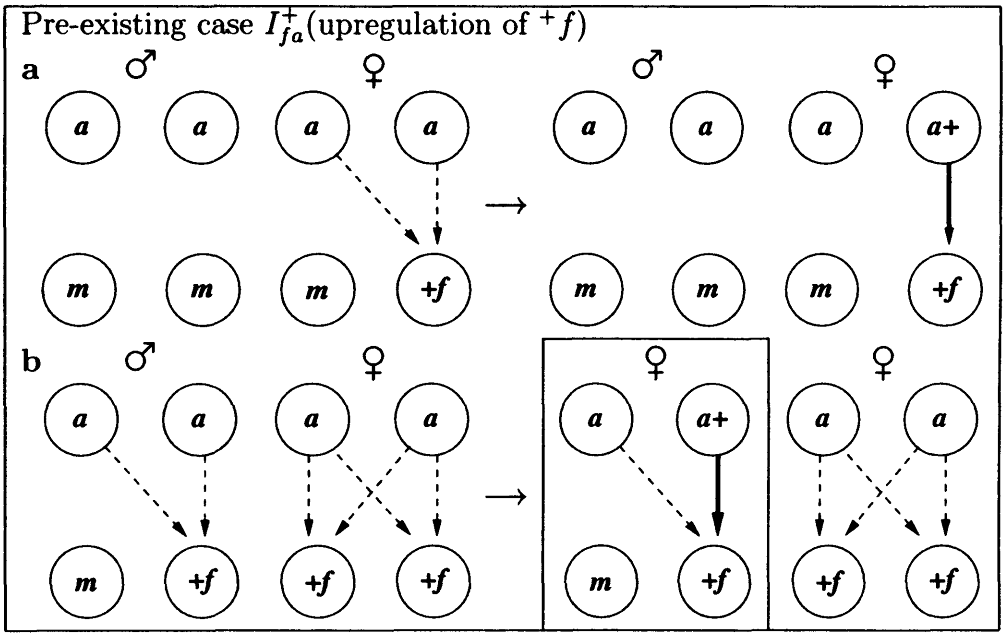


Figure 4.5: Two examples of mutations for pre-existing case I^+_{fa} illustrating why this pre-existing case is unlikely to lead to a change in sex-determining locus. Note that the highlighted mutant is likely to be phenotypically *female*, whereas the parental genotype is *male*

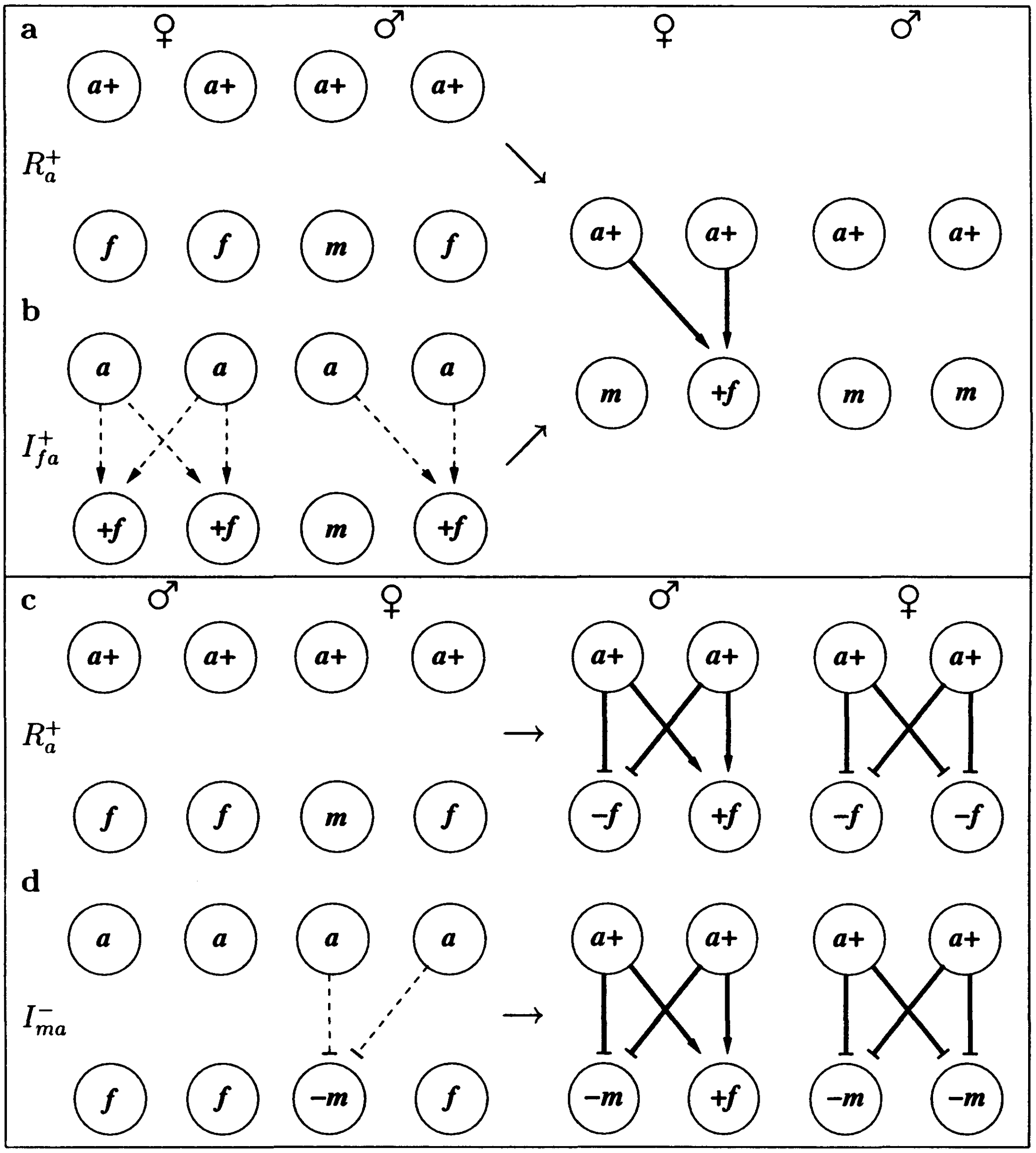


Figure 4.6: Most commonly observed transitions in which there is a change in heterogamety at the F locus due to recruitment of a gene at locus A, under ancestral male heterogamety (*ff*/*mf*). Transitions (a) and (c) are for pre-existing case R_a^+. Transitions (b) and (d) are for pre-existing cases I_{fa}^+ and I_{fa}^- respectively. Transitions (c) and (d) require one or more intermediate states (not shown), since they are at least two mutations removed from the ancestral population.

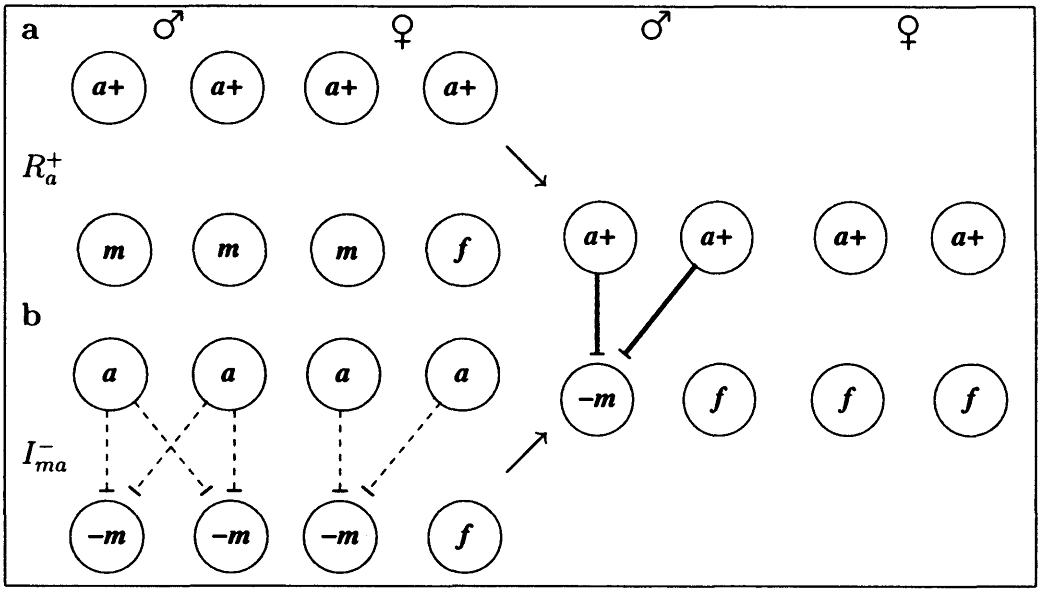


Figure 4.7: Most commonly observed transitions in which there is a change in heterogamety at the F locus due to recruitment of a gene at locus A, under ancestral female heterogamety *mm*/*mf*.

final population in both cases.

The transition shown in figure 4.6(a) only accounts for ~60% of observed transitions for pre-existing case R_a^+ (default values). The second most commonly observed transition for this case (accounting for ~30%) is that shown in figure 4.6(c). Whether the former or latter transition occurs, is dependent upon the order of the mutations: 1) if the first mutation to occur is ^+f, creating a positive *inbound* link from a^+, then the transition of figure 4.6(a) occurs, 2) if the first mutation is ^-f, creating a negative *inbound* link from a^+, then the transition of figure 4.6(c) is observed via at least one intermediate step in which a ^+f mutation drives m out. The latter transition is similar, though not identical, in its final population to the transition shown in figure 4.6(d) (for pre-existing case I_{ma}^-) which also occurs via at least one intermediate step.

Single-step transitions are observed for ancestral female (*mm*/*mf*) heterogamety, as shown in figure 4.7. Again, the these transitions represent two distinct pre-existing cases: R_a^+ as before, and I_{ma}^- (a pre-existing negative *inbound* link from a is present at ^{-}m). When the complementary mutation (^{+}f or a^{+} respectively) appears, a^{+} fixes at the A locus and heterogamety changes to male (*ff/mf*), leading to the same final population in both cases. Although this outcome is comparable to that of figure 4.6(a) and (b), no multi-step transitions, comparable to those shown in figure 4.6(c) and (d), are observed for ancestral female (*mm/mf*) heterogamety.

4.3.6 Unbiased fitness of the recruited gene in locus transition

We now consider whether the fitness values $w_{L(i)}$ are biased in the cases leading to transitions in the sex determining locus. Fitness bias (Δw) was calculated for each sex separately, as the difference of mean fitness values between: (1) the cases resulting in a transition, and (2) those which were simply viable.

Where we observe "transition probabilities" close to, or equal to, 1 (in tables 4.2 and 4.3), we will not, by definition, observe any significant bias, since all viable cases result in a transition. Such high "transition probabilities" were only observed for clamped A. For clamped A (default conditions), we observe "transition probabilities" equal to 1 under both *ff/mf* and *mm/mf* ancestral heterogamety. This tells us that no *a priori* bias (in $w_F^{♂}$ or $w_F^{♀}$) is required for the transition to occur.

On the other hand, with unclamped A (random $w_A^{♂}$ and $w_A^{♀}$), the "transition probability" is lower than 1. This suggests that when a trade-off is allowed between the two loci F and A, the certainty of a transition which existed before (for these conditions) is removed. As a consequence, we will observe transitions only when particular bias requirements are fulfilled at both loci (shown in table 4.4).

Table 4.4 contains several clear patterns. Firstly, for any particular locus, male ($\Delta w^{♂}$) and female ($\Delta w^{♀}$) bias values are always of opposite sign. Secondly, for any particular sex, the bias values for the different loci are also of opposite sign. Together, these constraints allow only two possible combinations for each case: (1) $\Delta w_F^{♂}$, $\Delta w_A^{♀}$ positive and $\Delta w_A^{♂}$, $\Delta w_F^{♀}$ negative, or

Anc. het.	case	locus	$\Delta w^{♂}$	$\Delta w^{♀}$
ff/*mf*	I^-_{fa}	F	0.27	-0.071
		A	-0.096	0.11
	I^+_{ma}	F	0.5	-0.095
		A	-0.87	0.13
mm/mf	I^-_{fa}	F	-0.051	0.69
		A	0.11	-0.52
	I^+_{ma}	F	-0.022 (n/s)	0.45
		A	0.12	-0.11

Table 4.4: Summary of fitness bias in evolution from single-locus to two-locus sex determination ("unclamped A") for both *ff*/*mf* and mm/mf ancestral heterogamety, and both cases (I^-_{fa} and I^+_{ma}). Each line contains the locus and the difference of means (Δw) for successful cases relative to the viable cases. P values (t-test) for each Δw were calculated for all entries. The single entry indicated (n/s) was not significant at the 95% level.

(2) $\Delta w^{♂}_{F}$, $\Delta w^{♀}_{A}$ negative and $\Delta w^{♂}_{A}$, $\Delta w^{♀}_{F}$ positive. Clearly, combination (1) occurs when the ancestral heterogamety is *ff*/*mf*, and combination (2) occurs when the ancestral heterogamety is mm/mf,

As far as the magnitudes ($|\Delta w|$) are concerned, it appears that the largest magnitude entry for each case occurrs in male ($\Delta w^{♂}$) for ancestral *ff*/*mf* heterogamety, and female ($\Delta w^{♀}$) for ancestral mm/mf heterogamety, although there is no clear pattern of consistency as to which locus it occurs in. One biologically notable feature is that the highest magnitude entry for case I^+_{ma}transitions (comparable to *D. melanogaster*) with ancestral *ff*/*mf* heterogamety, occurs at A making this gene highly deleterious in males. The corresponding gene *transformer* is indeed highly deleterious in males, since its net effect is to suppress male courtship.

4.4 Discussion

We have presented a model which integrates two previously separate techniques: a) a standard model from population genetics, and b) a network model using ODEs [32]. In order to do this, it was necessary to extend the concept of *allele* to include network-specific features such as interactions.

This definition also allows us to represent mutations as discrete changes in the allele parameters [119]. It was our aim to study this system at a coarse-grained level, representing interactions qualitatively, and the corresponding mutations as creation or deletion of these interactions. This coarse-grained level definition for mutation was chosen because it appears to be the most important class of interaction-related mutation in the evolution of sex determination [92]. Since quantitative changes appear to be less important, these are ignored.

Wilkins hypothesis [86] suggests a simplest-possible ancestral sex determination system involving a single locus. Specifically, we consider two single locus sex determination systems: male heterogamety *ff*/*mf* (*m* , a dominant masculinizing allele), and female heterogamety *mm*/*mf* (*f* , a dominant feminizing allele). We have used this as our starting point to examine the conditions under which a more complex sex determination system, involving two loci may have evolved. We did not consider dosage compensation in our model. In model organisms for which the process is best understood, dosage compensation initiates after the pathway has been activated, and as a first approximation can be considered a separate process to sex determination. Also, for simplicity we did not consider autoregulatory interactions in the network model. Although autoregulation is necessary for sex determination in two known cases, a recent study [92] suggests that in both these cases, the autoregulating gene (*Sxl* in the case of *D. melanogaster*, and *transformer* in *C. capitata*) became autoregulatory *after* recruitment into the pathway, we therefore consider this omission to be a reasonable simplification.

In this study, the two loci A and F were considered to be unlinked. Linkage may be an important factor, particularly for more complex cases: for example, it has been suggested that recruitment of the *sis* alleles in *D. melanogaster* (involved in primary signal generation), may have depended on (1) linkage to *Sxl* and (2) the absence of male recombination [92]. However, at the level relevant to this analysis, the same study found that linkage of *dsx* and *tra* did not qualitatively affect the relevant evolutionary transitions. Linkage patterns in *C. elegans* are more difficult to assess, since it is not clear which gene corresponds to the hypothetical F locus: *tra-1* is closer to being the "switch" gene, while its target *mab-3* is the true *dsx* homolog. While *mab-3* (chromosome II) is unlinked to its regulator *tra-1* (chromosome III), in the case of *tra-1*, the pattern is complicated by it having three upstream signals (*fem-1,2,3*), one of which (*fem-2*) is on the same chromosome. A more sophisticated model might therefore include extra features such as linkage, absence of recombination in one sex, and multiple-signal nodes (e.g. molecular complex formation), but these are left for future work.

In the model, genotype fitness $W(g)$ is a linear function of the expression levels $\hat{S}_i$. At a first glance, it would appear this represents a weakness of the model, since common biological features such as heterozygote advantage are inherently nonlinear. However, because $W(g)$ depends on the network parameters in a nonlinear way, these effects can indeed be represented. Consider, for example, the evolved female network on the right of figure 4.6a-b. If we now imagine a polymorphism involving strong ($T_{a^+} = 0$) and weak ($T_{a^+} = 1$) forms of the a^+ allele, we found it straightforward to determine appropriate parameter values (with $w_A^♀ < 0$ and $w_F^♀ > 0$) leading to heterozygote advantage at the A locus.

Each expression level $\hat{S}_i$ is multiplied by a random variable $w_{L(i)}$. If the resulting value $w(g)$ is negative, the genotype is considered to be lethal (i.e. receives zero fitness) to avoid negative population levels. Under condition "unclamped A", a certain trade-off can occur between the two loci (since one of the two can make a negative contribution to fitness). There are a number of ways one could define fitness in a model such as this. An alternative, perhaps more conventional, approach might have construed the fitness value in such a way that the ancestral values were equal to 1, with mutant genotypes producing values higher or lower than 1. However, changing fitness to a relative scale such as this creates new problems. For example, what to do when there is more than one male (or female) genotype (this situation may occur when evaluating successive mutations - see section 4.2.5, under heading "outer loop"). The problem here is to decide which genotype is assigned a fitness of 1. The experimental outcome may well depend on this choice. It is far from clear whether an alternative model such as this would qualitatively change the results seen in this study. The evaluation of alternative fitness definitions is a fairly complex issue, and is left for future work.

Starting from the simplest-possible system determined at a single locus F, we examine the conditions which allow control to pass to a new locus A, with the F locus becoming homozygous in both sexes. For this to occur, at least one new interaction between an allele at locus A and another at locus F must be formed. It was shown that this is most likely to occur when (a) a pre-existing *inbound* connection exists at the F locus, and (b) this is followed by a strong-effect mutation at A creating the connection with F, rather than the other way around. Two main outcomes leading to two-locus systems were observed. In the first, a male-defining mutant allele $a+$ at the A locus works by downregulating the existing allele ^{-}f at F (case I_{fa}^{-}, see figure 4.3), whereas in the second, a female-defining allele $a+$ at A upregulates the existing allele ^{+}m at F (case I_{ma}^{+}), as shown in figure 4.4. Both outcomes are only observed when $T_a = 0$, suggesting they would be caused by a strong-signal mutation (e.g. wholesale acquisition of an appropriate RNA-binding domain), rather than a weak-signal mutation (e.g. small-effect point mutation).

Particular conditions, such as parameter values, affect whether these outcomes originate from both male and female ancestral heterogamety, or from one in particular, see figure 4.4(a),(c). In both ancestral populations, the parameter k_T affects the difference in F expression level between the two sexes. Given (1) the low fitness of intersex phenotypes, and (2) the expectation that male and female F (*doublesex*) expression levels would evolve away from each other [92], we might expect k_T to be relatively high, and therefore only the transitions shown in figure 4.4(a) and (c) would be observed.

The results suggest that, if the recruited A locus gene had no *a priori* fitness bias (clamped A), recruitment was more likely since it imposes little or no bias requirements at the existing F locus. We make the biologically realistic assumption that a random gene choice (from any gene in the genome) is unlikely *a priori* to be involved in a process related to reproductive fitness (i.e. in terms of the model, has zero fitness). This assumption makes the unbiased transition more likely from a biological perspective than a potentially biased one, since the latter imposes bias requirements at both A and F before the transition can occur. These results suggest that the *D. melanogaster* gene *transformer* (*tra*) was recruited due to a strong gain-of-function mutation at the *tra* locus, and that it acquired its sex-specific role (in male courtship [80]) after recruitment. However, even if we do allow for the possibility that *a priori* bias may have existed at A (unclamped A), the results appear, for one case at least, to be biologically consistent with the known role of *tra*, namely that it is highly deleterious to males. Interestingly, the fitness bias results for the potentially biased case (table 4.4) also illustrate the capacity for opposing effects, both between loci in the same sex and between sexes at the same locus, in shaping sex determination.

"Transition probabilities" for the clamped A case were, in many cases, equal to 1, which suggests these transitions are certain to occur. An analytical approach would help in understanding why this is so. However, several factors make a rigorous mathematical analysis of the simulation results an unusually complicated task, in particular, (1) the fact that all possible mutations may need to be evaluated before an invasive mutation is found, and (2) the appearance in the population of additional genotypes (generated by mating and recombination) following insertion of a mutant genotype. Clearly though, particular parameter values are important in determining whether a transition is observed or not. For example, the results for *ff*/*mf*, case I_{fa}^{-} show "transition probabilities" equal to 1 for all parameter values tested, which is of particular concern, since it suggests that the model is constructed in such a way that this particular transition will always occur. However, we have found that for certain other parameter values (e.g. $a = k_T = 2$), the "transition probability" becomes 0, which shows this is not the case. In this study we have made a number of fairly crude assumptions (for example, the fact that connections can only be a qualitative ±1) and, have consequently tried to interpret the results in a suitably broad manner. Accordingly, we have emphasised the notable differences between the qualitatively distinct pre-existing cases, while giving less importance to the differences in outcome for different parameter values within each pre-existing case.

Transitions leading to a change in heterogamety at the F locus were also considered. It was found that heterogamety changes occur mainly under the three pre-existing cases (R_a^+, I_{fa}^+, I_{ma}^- , see figure 4.3) where a locus transition did not occur. In this case, the transition is likely to occur following a mutation at either locus (A or F), leading to one of the three qualitatively distinct populations shown in figures 4.6 and 4.7. These evolved populations were subsequently tested as ancestral populations for change in sex determination locus, but the "transition probability" was found to be fairly low (< 0.15 for all evolved populations).

Together these results suggest that mutations causing new genes to be recruited into the sex determination pathway will have specific consequences depending on how the ancestral discriminatory locus is affected. Some mutations in the recruited gene (pre-existing cases I_{fa}^- and I_{ma}^+) may cause a transition to a new discriminatory locus at A. Transitions to a two locus system requires a strong signal from the recruited gene suggesting only non-gradual mutations (e.g. wholesale acquisition of a regulatory domain) will provoke the change. Other mutations, both in the ancestral discriminatory locus (case R_a^+) or in recruited gene (cases I_{fa}^+ and I_{ma}^-), cause recruitment of A (in a non-discriminatory role), while at the same time inducing a change in heterogamety at the F locus (*ff*/*mf* ↔ *mm*/*mf*). Once a change in heterogamety has occurred, a further change in sex determination locus (involving the newly recruited gene) is unlikely.

4.5 Appendix A: Stability criteria

For the population simulations, we consider the dynamics of a discrete time series $(p(t), p(t+1), ...)$ of genotype frequencies. During the first steps following the introduction of a mutant, the length N (N is the number of genotypes) of $p(t)$ may increase, as crosses generate new genotypes. Once N has stabilised, we consider the maximum (over all genotypes g) difference between successive timesteps, $e(t) = \max_g |p_g(t) - p_g(t-1)|$. The system is considered stable when $e(t) < \epsilon_p$, with $\epsilon_p = 10^-6$. This simple method was found to be more efficient than more elaborate methods such as that described in [43]. Recall that the mutant is inserted at 1% of "parent" genotype frequency(see section 4.2.5). If the equilibrium mutant frequency is $\leq$1.05%

(slightly above 1% to allow for rounding errors), the mutation is classified as non-invasive, which includes deleterious and neutral mutations. Following invasive mutations (mutant frequency >1.05%), we look for two exit criteria: 1) a change in sex determining locus: (i.e. the F locus is homozygous in both sexes), or 2) a change in F locus heterogamety. If these criteria are fulfilled, and additionally the population consists only of a male and a female genotype, we exit and record the transition (including the initial conditions such as fitness), otherwise the mutation process is again repeated with the derived population. Successive invasive mutations (not fulfilling the exit criteria) are assessed in this way to a maximum recursion depth of 3.

For the network simulations, numerical integration proceeds until two conditions are met: a) dS_i/dt is below a threshold $\epsilon_s = 10^-5$ for all i, and b) $Re(\lambda) \leq 0$ for all eigenvalues λ of the Jacobian matrix. If these two conditions are not met by $t = 800$ (higher values were found to not qualitatively change the results), the system is considered unstable and the genotype is deleted from the population. This assumption here is that functional genetic networks will reach a stable equilibrium gene-expression state, and that unstable networks reflect, in a sense, the failure of development [43, 51]. The Jacobian matrix of the system can be computed very efficiently, since the function $\sigma(x) = 1/(1+e^{-ax})$ has a straightforward derivative: $\sigma' = a\sigma(1-\sigma)$. Therefore, the elements J_{ij} of the $(2l \times 2l)$ Jacobian matrix are:

$$J_{ij} = \frac{\partial[\sigma(u_i) - S_i]}{\partial S_j} = a\sigma(u_i)(1 - \sigma(u_i))I_{ij}R_j - \delta_{ij}$$

where δ_{ij} is the Kronecker delta, $\delta_{ij} = \begin{cases} 1 & \text{for} \quad i = j \\ 0 & \text{otherwise} \end{cases}$
For numerical integration of the ODE system, we use the Runge-Kutta-Fehlberg method as implemented in the Gnu Scientific Library (GSL) v1.3. Eigenvalues for the Jacobian are calculated using the *dgeev* function implemented in LAPACK v3.0.

4.6 Appendix B: Definitions

Term	Description
regulatory domain	a binary value defining whether or not a particular allele can influence the expression of other alleles
inbound link	a vector representing the inputs a particular allele receives from the other alleles
viable	an initial population is *viable* if, for both the male and female genotypes, the fitness $W(g) > 0$
“clamped”	the locus A is “clamped” when both fitness contributions for the locus are zero ($w_A^{♂} = w_A^{♀} = 0$)
“unclamped”	the locus A is “unclamped” when fitness contributions ($w_A^{♂}$,$w_A^{♀}$) for the locus are sampled from a Gaussian distribution $\mathcal{N}(0,1)$

Chapter 5

Evolution of a two-locus sex determination system to a system involving three loci

5.1 Background

The discovery that the key Drosophila gene *doublesex* was highly conserved, has added support to a hypothesis proposed by Wilkins [86], that sex determination networks have evolved in a retrograde manner from bottom to top. Following this hypothesis, it is reasonable to make the assumption that the simplest-possible ancestral sex determination systems would be based on a single locus. In the previous chapter, a diploid hierarchical model is presented which integrates techniques from standard population genetics with network dynamics (ODEs). Two possible systems (male and female heterogamety) based at a single ancestral sex determining locus F, were considered as starting points, then the conditions under which a new and previously independent locus A becomes the new discriminatory locus were studied, with the ancestral sex determining locus becoming homozygous in both sexes. It was found that recruitment of a new discriminatory gene into a pathway is most likely due to an adaptive mutation in the recruited gene rather than in a gene from the existing pathway. Furthermore, the mutation must have a strong ($T_a = 0$) effect on its target in order to provoke a transition. Transitions resulting in a locus change predominantly lead to one of two outcomes: (a) a dominant masculinizer at A downregulates F locus expression in males, or (b) a dominant feminizer at A upregulates F locus expression in females. These two evolved populations are shown in figure 5.1. Although it is argued that the recruited gene was probably neutral with respect to reproductive fitness, if we allow for the contrary possibility that the recruited gene did make an independent *a priori* contribution to fitness, conflicting patterns in fitness bias emerge between loci and between sexes (e.g. the same gene has a beneficial effect in males and a deleterious effect in females).

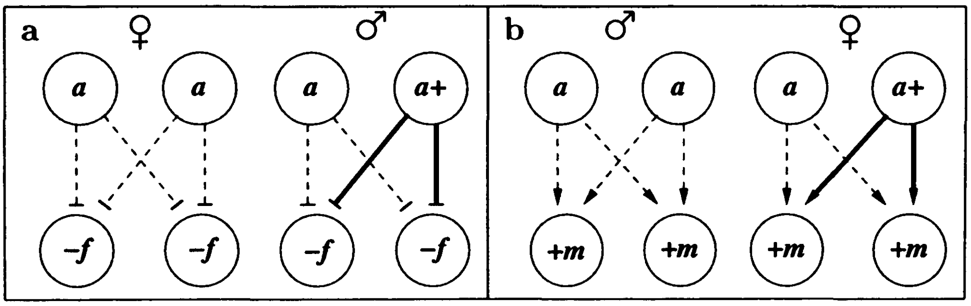


Figure 5.1: These two ancestral populations serve as starting points for analysis of further evolution in this chapter. Recall that the ^{+}m allele is an F locus allele for which $T_{+m} = 1$.

In this chapter we consider further evolution of the two-locus sex determination networks derived in the previous chapter 4 (figure 5.1). Specifically, we will consider the conditions allowing recruitment of a new locus B into the pathway.

5.1.1 Pre-existing cases

Taking the previously evolved two-locus sex determination networks as our starting point, two types of transition are again considered: (a) transitions in sex determination locus such that the discriminatory signal moves to a new locus B, and (b) transitions in heterogamety at the A locus following B locus recruitment. Also, as before, we can consider the possible pre-existing cases.

There are now five cases to consider for each ancestral two-locus network (shown in figure 5.2), which will be labelled as cases R_b^+, $I_{a^+b}^+$, $I_{a^+b}^-$, I_{ab}^+ and I_{ab}^- for both male and female ancestral heterogamety, as shown. So, a pre-existing case in which a positive *inbound* link exists at the the allele ${}^+a^+$ is denoted $I_{a^+b}^+$ /♂ for ancestral male heterogamety, and $I_{a^+b}^+$ /♀ for ancestral female heterogamety.

5.1.2 Experimental procedure

Repeating the experimental procedure from the previous chapter, two types of transition are considered: locus transitions and heterogamety transitions. Transitions in sex determination locus occur when the discriminatory sex determination locus passes from the A locus (the ancestral state) to the B locus, with the A locus becoming homozygous in both sexes. Heterogamety transitions occur when there is a switch between male and female heterogamety at the A locus as a consequence of B locus recruitment. Recall from chapter 4 that transitions are reported relative to *viable* initial populations, and that the measured "transition probabilities" (see section 4.2.5) are best considered in terms of bias requirements in the fitness contributions. Sex determination locus "transition probabilities" are shown in table 5.1, heterogamety "transition probabilities" are shown in table 5.3.

Note that, as before, the *general*"transition probabilities" shown in tables 5.1 and 5.3 may in turn represent one or more qualitatively *distinct* transitions, i.e. distinct final evolved populations. For example, consider briefly the transition shown in figure 5.3a. Most often, this transition is observed ocurring in a single step following a single mutation b^+, as shown. Sometimes though, the same transition is also observed with one or more intermediate states (e.g. following the successful invasion of a mutant of a which receives a positive input from a^+). These intermediate states turn out to be irrelevant since the final population is the same as the evolved population on the right of figure 5.3a (i.e. the first mutant was driven out by the second so it is as if the first mutant had never occurred). The discussion below will therefore treat shortest-possible transitions (e.g. the single-step transition of figure

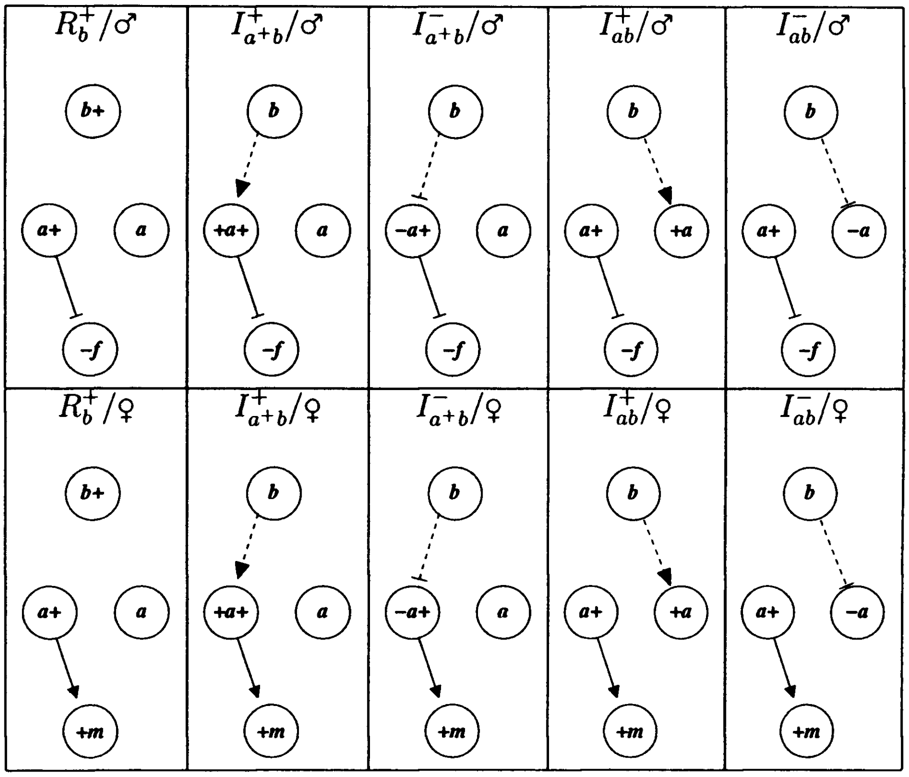


Figure 5.2: Ten possible cases of prior conditions which might permit the two-locus sex determination systems from figure 5.1 to evolve. At the top of each box, the pre-existing condition is indicated with a graphical representation below. Since *inbound* links are both signed and allele-specific, these are indicated by the dotted lines. The top row represent ancestral male heterogamety, where the dominant masculinizer a^+ represses the ^-f allele. The bottom row represents ancestral female heterogamety, where the dominant feminizer a^+ activates the ^+m allele.

5.3a) as equivalent to cases with one or more intermediate steps, but where the final population was the same. These shortest-possible transitions will be referred to as *distinct* transitions.

To summarise then, for a particular case, ancestral heterogamety and conditions (parameter values), we have a *general* "transition probability", which is simply the value corresponding to a change of locus (or of heterogamety), irrespective of how many particular transitions were observed for that category. Since several particular transitions can have the same final population, these transitions are grouped into qualitatively *distinct* transitions, which will be represented in each case by the shortest path found.

Since a large number of transitions (both of sex determination locus and of heterogamety), were observed at very low frequencies, only the most important *distinct* transitions, observed at "transition probabilities" greater than 2% in any of the experiments will be considered in detail. This approach represents a different strategy from chapter 4, where *general* "transition probabilities" usually represented only one or two *distinct* transitions.

5.1.3 Ancestral fitness at the B locus

As before, we will consider "clamped B" ($w_B^{♂} = w_B^{♀} = 0$), i.e. the biologically realistic hypothesis that there is no *a priori* fitness bias in the recruited gene, separately from "unclamped B"($w_B^{♂}$ and $w_B^{♀}$ drawn from a random Gaussian distribution) for locus B. Since the results for clamped and unclamped B are now often of similar magnitude (whereas in the previous chapter the "transition probabilities" for the unclamped case were low relative to the clamped case) the two conditions are now considered together.

Since the A locus is now taking the place previously occupied by the F locus, only unclamped A will be considered here (in chapter 4, both clamped and unclamped A were analysed).

5.1.4 Known sex determination networks

The best understood sex determination networks are probably *D. melanogaster* and *C. elegans* [91]; however, interest in sex determination evolution has led to studies in other insects, such as *C. capitata*. What is known of the structures of these three networks is shown in figure 1.4. There are two common features: a) the shared homolog (*dsx*, *mab-3*) at the end of all three pathways, and b) *D. melanogaster* and *C. capitata* both share *transformer* (*tra*). Apart from this, all three networks are different. In *C. elegans* and *D. melanogaster*, the primary signal is defined by the X:A ratio (albeit by different genes in each), whereas in *C. capitata* it is defined by a male determining factor (as it is in mammals).

In this chapter, we will use the model to explain some features of the known biological networks. In the known networks, the new B locus would be equivalent to the *C. elegans* genes *fem-1,2,3* (if *mab-3* is considered to be the equivalent of F in the model), the *D. melanogaster* gene *Sex-lethal*, or the, as yet unidentified *M* gene in *C. capitata*.

It has been argued by Wilkins [86], that the component genes of the *C. elegans* sex determination pathway may have evolved as a consequence of successive imbalances in the sex ratio, with each new gene being recruited as a determiner via frequency dependent selection. In the case of *D. melanogaster*, a more specific hypothesis was proposed for the recruitment of the gene *Sex-lethal* in [92]. Here, it was suggested that *Sxl* was recruited as a splicing regulator of *tra* (removing the stop codon from *tra* mRNA transcripts), which in turn provoked three further changes in *Sxl*, which we observe today: autoregulation, the exon containing a stop-codon, and recruitment of *sis* as an activator of the early promoter Sxl_{Pe}.

5.2 Results

5.2.1 "Transition probabilities" are lower than single locus case

Sex determination locus "transition probabilities" are shown in table 5.1, heterogamety "transition probabilities" are shown in table 5.3. Whereas for the single locus to two locus transitions, there were many "transition probabilities" close to or equal to 1, the highest is now 0.16, suggesting stronger bias requirements.

As discussed in chapter 4, "transition probabilities" and fitness bias are closely related. For example, the transition shown in figure 5.3a requires a strong negative bias at the A locus for males ($\Delta w_A^{\male} \simeq -1$), suggesting A must be strongly deleterious for the transition to occur. The majority of ancestral populations which are simply *viable* (mean $\Delta w_A^{\male}$ is zero) will not therefore provoke a transition. This fact translates into a lower "transition probability", which can be seen clearly by considering the tables showing fitness bias for both locus (table 5.2) and heterogamety (table 5.4) transitions.

5.2.2 Locus transitions are most likely due to a mutation at the B locus

Looking at the R_b^+ (pre-existing regulatory output at B) column in table 5.3, we can see that the "transition probabilities" are relatively low, which seems to indicate that pre-existing cases R_b^+ do not significantly provoke a change of sex determination locus. Although under certain conditions we do observe *general* "transition probabilities" above 2%, a closer analysis shows that no *distinct* transition is observed above 2% in this case. This suggests that mutations causing a change in the sex determination locus are most likely to occur in the recruited gene B, though the difference is not as great in relative terms as in chapter 4.

5.2.3 Locus transition requires strong pre-existing signal at B

In table 5.1, we observe that when the pre-existing signal strength for the allele b is weak ($T_b = 1$), "transition probabilities" are also low. The same effect was also observed for the single locus ancestral system considered in chapter 4. The strong ($T_b = 0$) signal for b is the default value used for all other rows. This shows that mutations which initially have a weak effect on A are less likely to be recruited, suggesting that particular kinds of mutations representing strong-signal mutants (e.g. wholesale acquisition of a DNA-

Anc. het.			/♀					/♂		
case:	R_b^+	I_{a+b}^+	I_{a+b}^-	I_{ab}^+	I_{ab}^-	R_b^+	I_{a+b}^+	I_{a+b}^-	I_{ab}^+	I_{ab}^-
clamped B										
defaults	0.01		0.06	0.01	0.01	0.01	0.01	0.02		0.01
low a										0.01
high a	0.01	0.02	0.16	0.01	0.01			0.05		0.01
low k_T		0.01	0.05		0.01		0.01	0.02		0.01
high k_T			0.06			0.01	0.01	0.01		0.01
$T_b = 1$							0.01			
unclamped B										
defaults	0.03		0.08			0.01	0.01	0.06		
low a	0.01		0.01			0.01		0.01		
high a	0.04	0.01	0.11			0.01	0.01	0.08		
low k_T	0.02		0.07			0.01	0.01	0.06		
high k_T	0.02		0.08			0.02	0.01	0.05		
$T_b = 1$	0.01		0.01			0.01		0.01		

Table 5.1: *General* sex determination locus "transition probabilities". Blanks indicate zero.

binding domain), are more likely to induce recruitment than weak-signal mutants (e.g. small-effect point mutations).

5.2.4 The most common locus transition resembles *C. capitata* network

The highest *general* "transition probability" in table 5.1 is 0.16 for case I_{a+b}^- /♀ (see figure 5.2). This high "transition probability" is only observed for particular conditions: clamped B (the fitness parameters $w_B^♂ = 0$) and high a (the sigmoid slope parameter). For this particular case, the most common, by far, *distinct* transition (representing over 98% of observed transitions) is that shown in figure 5.3a. Here, ancestral female heterogamety at the A locus has evolved into male heterogamety at the B locus via recruitment of the b^+ allele, with female becoming the default pathway in the evolved population.

All transitions apart from this one (including A locus heterogamety changes) occur at frequencies lower than ~5%. Since only those transitions above 2% frequency will be described, this represents a fairly narrow range of frequencies in which many transitions are observed.

It is interesting to note the similarity of the evolved pathway with the sex determination network of *Ceratitis capitata* shown in figure 1.4c, where the male pathway consists of a dominant masculinizing allele which downregulates *transformer* (*tra*, equivalent to a). The main difference consists in the additional *tra* autoregulation of the real *C. capitata* network, which is not considered as a possibility in our analysis.

Under case $I^{-}_{a+b}/♀$ (see figure 5.2) with default conditions we also observe a second *distinct* single-step transition resulting in a similar (in the sign of the connections between loci) pathway, shown in figure 5.3b. Comparing this transition to the one described before (figure 5.3a), we can immediately see that the former *male* phenotype is the same as the latter *female* phenotype.

Such a transition is possible because phenotypic sex is strongly parameter-dependent, in particular on the two parameters a (sigmoid slope) and θ (sex determining threshold). Since the transition of figure 5.3a is also observed for default values of a (albeit at lower frequency), this suggests it is the random value of θ which is more important in deciding which transition occurs.

5.2.5 Locus transitions leading to serial negative interactions

Another class of locus transition which appears to be of particular biological relevance is shown in figure 5.3c, and is observed under case $I^{-}_{a+b}/♂$ (see figure 5.2) with unclamped B. Here, the b^{+} allele again forces a change in sex determination locus to B, with a new negative interaction forming.

The hypothesis proposed by Wilkins [86], suggested that the *C. elegans* sex determination pathway evolved in a retrograde manner from bottom to top, adding a new negative discriminator at each step. Adopting this hypothesis, it is reasonable to suppose that a system such as this, with two serial negative connections, was ancestral to the sex determination system we see today in *C. elegans* [77], shown in figure 5.3a. As mentioned above, Wilkins also proposed an evolutionary mechanism based on successive skewed sex ratios. The results here show clearly how a mechanism *not* based on sex

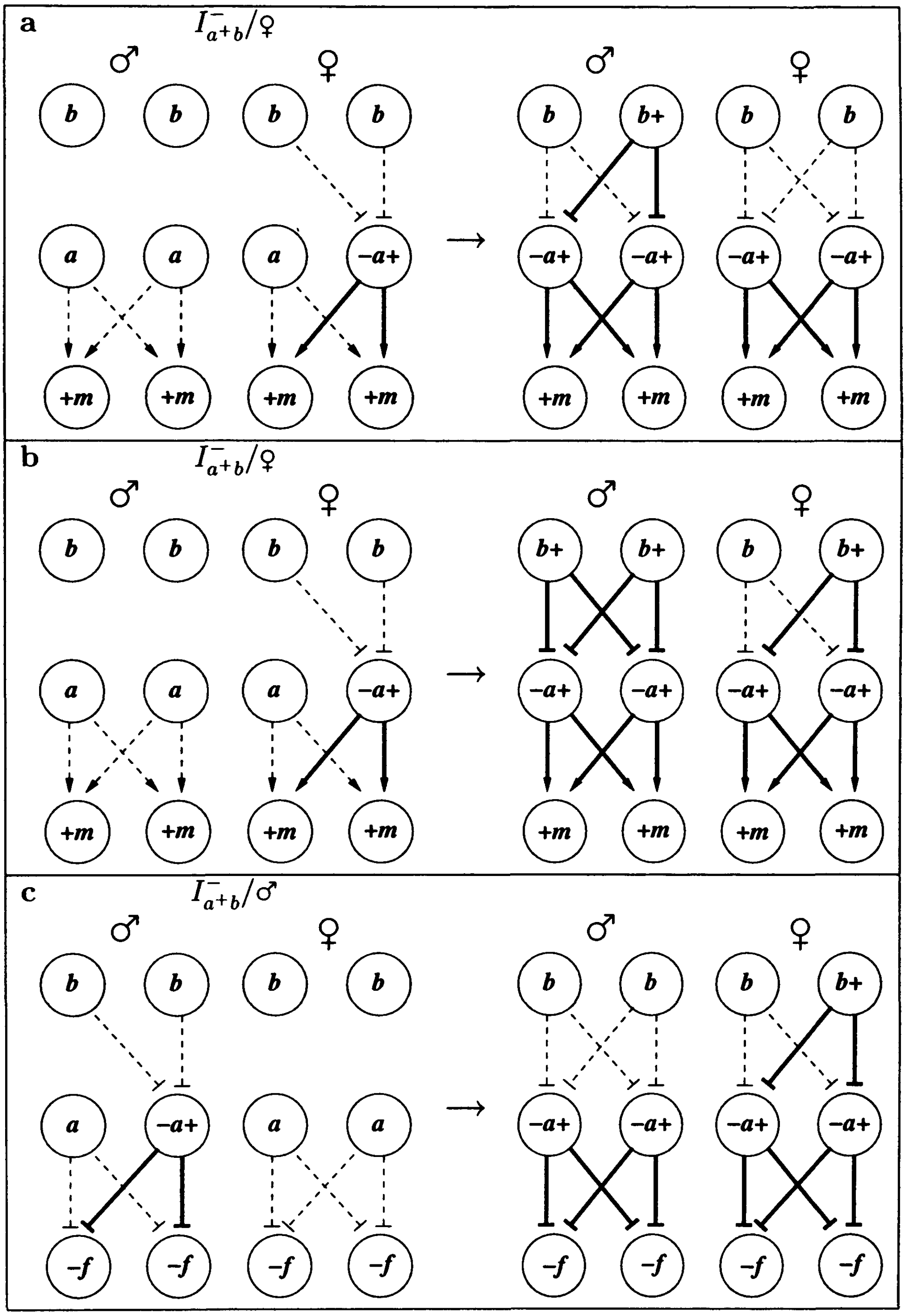


Figure 5.3: Commonly observed single-step transitions leading to recruitment of B as discriminatory locus.

ratios (the sex ratio is always 50:50 in these simulations), can also produce such a pattern.

5.2.6 Locus transition involving evolution of a negative feedback loop

A more complex class of *distinct* transition, requiring at least one intermediate step, was also observed. The most common of these occurs under case I^{-}_{a+b}/♀/high a (sigmoid slope parameter) with unclamped B, where we observe the transition described in figure 5.4(a). Here the evolved male genotype contains a feedback inhibition loop which effectively shuts off the activity of the allele ${}^{-}a^{+}$.

A similar transition (also observed at greater than 2% frequency) in which a negative feedback loop evolves is shown in figure 5.4(b), though this time for I^{-}_{a+b}/♂ (ancestral male heterogamety). Here the final female (rather than male) genotype contains the feedback inhibition loop. Although both feedback inhibition loops shown in figure 5.4 were found to be stable, in reality (i.e. under other conditions), such circuits are prone to instability due to possible oscillations [36]. Since oscillatory behaviour in developmental networks would probably be purged by negative selection [43, 51], transitions leading to feedback inhibition loops are unlikely to be common in sex determination.

5.2.7 Locus transitions show strongly negative fitness bias at A locus

The fitness parameter bias results for the single-step transitions of figure 5.3 are shown in table 5.2. In all cases, the strongest bias (i.e. largest $|\Delta w|$) occurs at the A locus and is negative, indicating that expression of A will have a deleterious effect on fitness in both sexes. Furthermore, this negative effect is strongest in male for ancestral female heterogamety (transitions 5.3a and b), and strongest in female for ancestral male heterogamety (transition 5.3c), as the highlighted entries show. More generally, we can say that the strongest bias (which always negative) occurs at the A locus in the ancestral

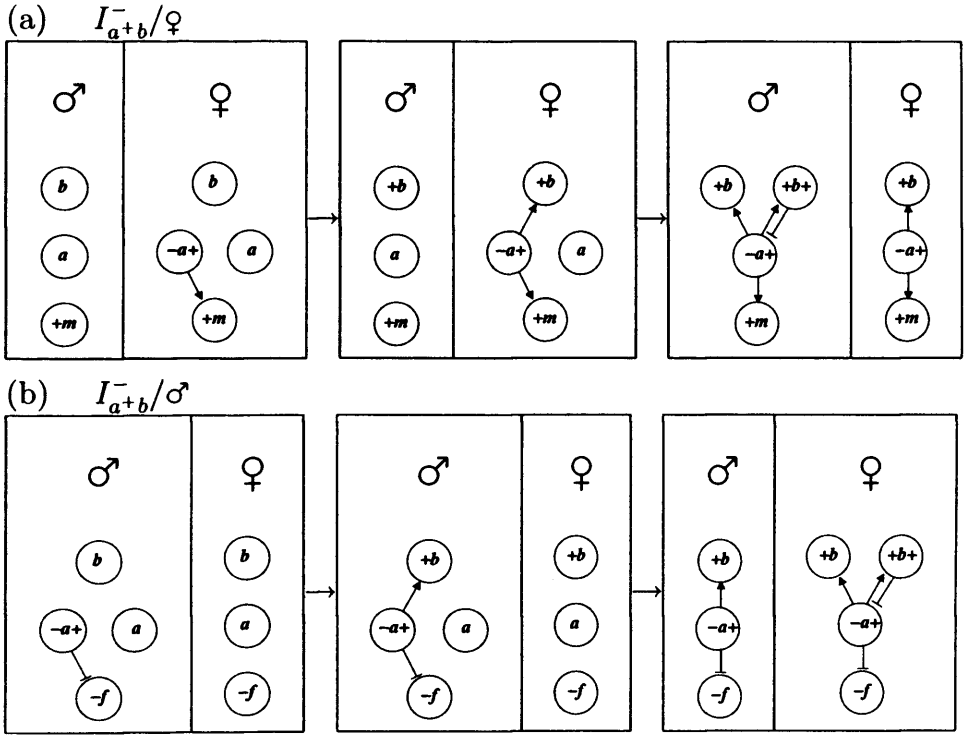


Figure 5.4: Two commonly observed complex transitions in the evolution from a two locus system to a three locus system. The diagrams show in each case how the population evolves in two steps from left to right. For clarity, homozygous loci are shown as a single node and potential connections are not shown, as they were previously in figure 5.3. The ancestral population (left) is the two locus system, which evolves to an intermediate population (middle), through to a final evolved population in which the B locus is discriminating. In both cases the transition is such that the negative connection from B to A does not appear until after an intermediate stage in which a mutant allele b' is upregulated by the existing allele a. Transition (a) is observed under case $I^-_{a+b}/♀$ (ancestral female heterogamety), whereas transition (b) is observed under case $I^-_{a+b}/♂$ (ancestral male heterogamety).

homogametic sex.

The bias effect at the F locus is positive in all clamped B ($w_B^{\male} = w_B^{\female} = 0$) cases. Under clamped B, overall fitness depends on the contributions of the F and A loci alone (B only contributes via regulation). If A has a strongly negative bias, then the F locus contribution ($w_F^{\male}, w_F^{\female}$) must be positive to counteract this, since overall fitness needs to be positive. Under unclamped B ($w_B^{\male}, w_B^{\female}$ random), the negative contribution at the B locus can be counteracted by a positive contribution at either remaining locus (F or A). Other than this, there is no clear pattern for F and B locus bias under unclamped B.

transition	w_B	locus	$\Delta w^{\male}$	$\Delta w^{\female}$
5.3a /♀	CL	F	0.29	0.33
		A	**-1**	-0.27
		B	N/A	N/A
	UN	F	0.087	0.41
		A	**-1.0**	-0.11
		B	0.26	-0.096
5.3b /♀	CL	F	0.13	0.45
		A	**-1.3**	-0.62
		B	N/A	N/A
	UN	F	-0.17	0.22
		A	**-1.3**	-0.57
		B	0.54	0.42
5.3c /♂	CL	F	0.23	0.47
		A	-0.25	**-1.4**
		B	N/A	N/A
	UN	F	0.23	-0.23
		A	-0.14 (n/s)	**-1.1**
		B	0.0079	0.6

Table 5.2: Fitness bias observed for single-step locus transitions. The largest bias ($|\Delta w|$) is highlighted for each transition and corresponds to the ancestral homogametic sex. Transitions are represented by the corresponding figure number, and the sign of the interaction between the A and F loci in the ancestral population. The w_B column describes whether the entry is for clamped B (CL), or unclamped B (UN). All bias values are statistically significant (t-test) at the 95% level, except where indicated (n/s).

In all single-step cases, we observe transitions for both clamped and unclamped B. These results tell us that *a priori* fitness bias at the B locus is not required for the transition to occur, though a particular bias may exist at B (as observed in the table) without significantly reducing the chances of a transition.

The complex two-step transitions (figure 5.4) are only observed for unclamped B. Given the arguments presented in chapter 4 concerning the low probability of fitness bias in a newly recruited gene, the fact that this transition is only observed for unclamped B gives us a second reason for rejecting this particular transition (the first reason being susceptibility to oscillations).

5.2.8 Heterogamety transitions are derived from recruitment of an unbiased gene

Transitions in heterogamety at the A locus are now considered. In this case, the B locus may be recruited in a non-discriminatory role, while provoking a change in heterogamety at the A locus. The *general*"transition probabilities" for heterogamety are shown in table 5.3. Again, the *general* "transition probabilities" have only coarse-grained informative value since each entry may represent more than one *distinct* transition. We again consider only the most important distinct transitions observed at "transition probabilities" greater than 2% in any of the experiments. One common feature of the *distinct* transitions is that they are only ever observed for clamped B ($w_B^{♂} = w_B^{♀} = 0$). This indicates that heterogamety transitions are most likely to arise through recruitment of an unbiased (with respect to reproductive fitness) locus B.

5.2.9 Heterogamety transitions often involve a double interaction

The most common of the A locus heterogamety transitions is shown in figure 5.5a, and is observed for case $I^{+}_{a+b}/♂$ (with low sigmoid slope a, among other conditions). Here, the heterogamety transition occurs in (minimum)

Anc. het.			/♀					/♂		
case:	R_b^+	I_{a+b}^+	I_{a+b}^-	I_{ab}^+	I_{ab}^-	R_b^+	I_{a+b}^+	I_{a+b}^-	I_{ab}^+	I_{ab}^-
clamped B										
defaults	0.03	0.03	0.01	0.01		0.05	0.09	0.01	0.02	0.01
low a	0.04	0.06	0.02	0.01	0.01	0.05	0.07	0.04		
high a	0.01	0.02				0.04	0.08	0.01	0.01	
low k_T	0.03	0.05				0.05	0.07			
high k_T	0.01	0.02				0.04	0.07			
$T_b = 1$	0.01	0.01		0.01		0.01	0.03	0.02	0.02	0.01
unclamped B										
defaults	0.02	0.02	0.02	0.02	0.01	0.03	0.04	0.02	0.02	0.01
low a	0.03	0.03	0.03	0.01	0.01	0.03	0.03	0.03	0.01	0.01
high a	0.01	0.01				0.02	0.03	0.01	0.01	0.01
low k_T	0.02	0.02				0.03	0.02			
high k_T	0.01	0.01				0.03	0.02			
$T_b = 1$	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01

Table 5.3: *General* heterogamety "transition probabilities". Blanks indicate zero.

two steps. In the first step, a mutant allele b^+ drives out the previously monomorphic b. The allele b^+ upregulates a in males, leaving female fitness unchanged. A second mutant allele f', at the F locus, inherits its negative input from a and adds a positive input from b^+, with b^+ now upregulating both the A and F loci. Both steps are facilitated by the strong positive bias at the A locus (see table 5.4), since both steps involve the addition of at least one positive link upregulating the ${}^+a^+$ allele. This general observation applies to all three transitions shown in figure 5.5.

The second most commonly observed heterogamety transition is observed for case $I_{a+b}^+/$♀, again with low sigmoid slope a, among other conditions. This transition also occurs in (minimum) two steps, and is shown in figure 5.5b. Here, the changes are similar to the previous transition, but for ancestral /♀. As before, in the first step, b^+ drives out the b allele, but in a nondiscriminatory role. In the second step, the mutant allele m' adds a negative input to the F locus. In the final population, b^+ *upregulates* a, and at the same time *downregulates* the monomorphic allele m'.

A third heterogamety transition involves a fairly complex series of tran-

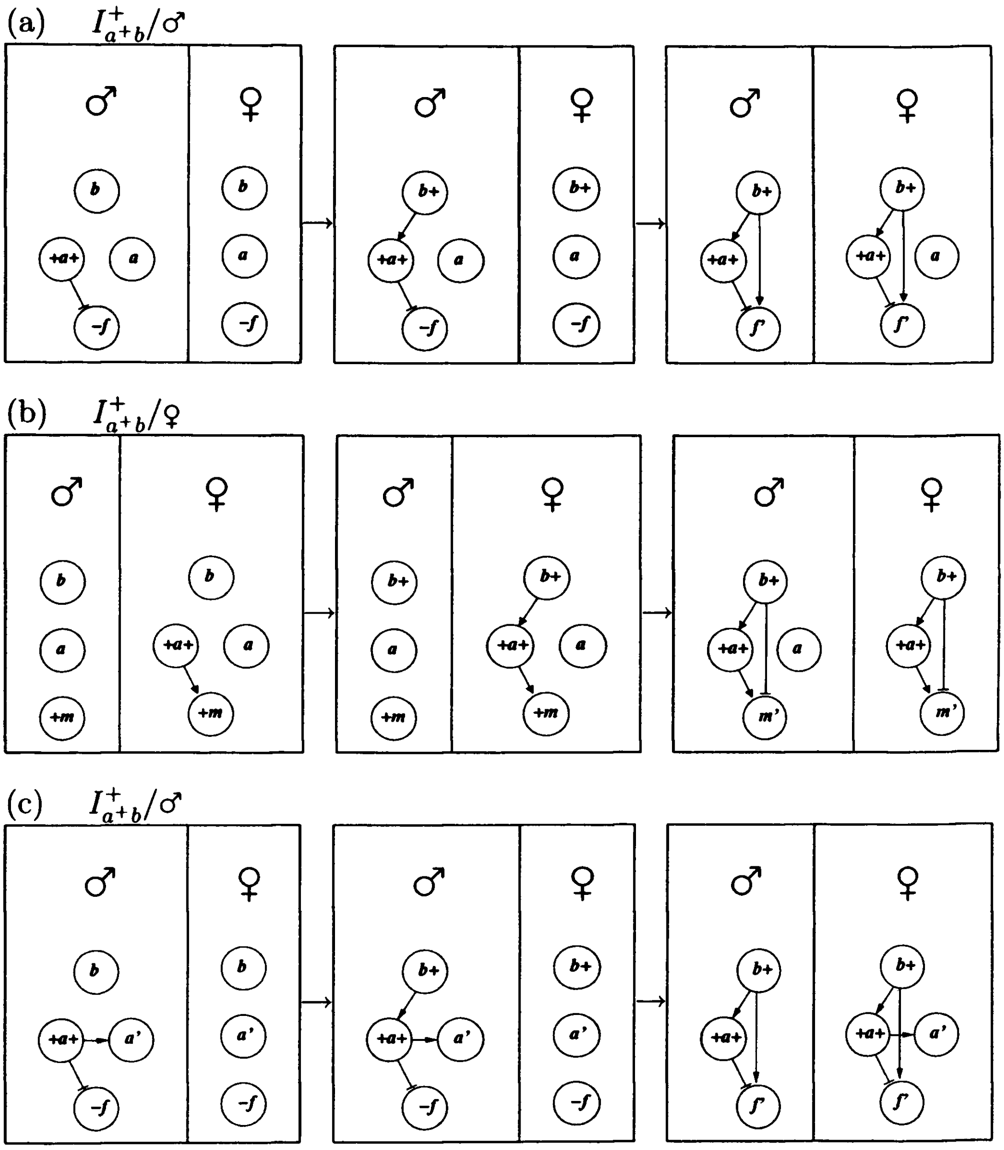


Figure 5.5: Three commonly observed sequences in which a change in heterogamety at the A locus is observed, leading to a double interaction. (a) Observed under case $I^{+}_{a+b}/♂$, in which there has been a change to female heterogamety from male heterogamety. (b) Observed under case $I^{+}_{a+b}/♀$. In the final population, there has been a change to male heterogamety from female heterogamety. (c) This transition actually occurs in four stages. The ancestral stage is identical to that of (a), with only the remaining three steps shown.

sitions (with two intermediate states), and is observed for $I^{-}_{a+b}/♂$ with low a and clamped B. This transition is shown in figure 5.5c. In the figure, the first (ancestral) stage is not shown since it is identical to the ancestral stage of figure 5.5a. In the first step, an (upregulated by a) allele A', drives out A. Next, the mutant b+ displaces b by upregulating a. Finally, the mutant allele f' replaces f, adding a positive interaction from b, changing the final heterogamety at A from male to female.

What these three complex heterogamety transitions have in common is the evolution of a double interaction from the b^{+} allele to alleles at both other loci (A and F). All three evolved networks are instances of a "network motif" (patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks) called the feedforward loop [21].

5.2.10 Only one single-step heterogamety transition is observed

Only one single-step heterogamety transition was observed, and this was for case $I^{-}_{a+b}/♂$, with low sigmoid slope a. This transition occurs in a single step and is shown in figure 5.6. Here, the B locus is recruited and downregulates the allele a. However, unlike the comparable sex determination transitions of figures 5.3a and b, the change is such that only a heterogamety change at A takes place, with the mutant b^{+} becoming homozygous in both sexes.

5.2.11 Strongest fitness bias for heterogamety transitions is also at A locus, but in ancestral heterogametic sex

Table 5.4 shows the bias (Δw) in the fitness parameters for the heterogamety transitions shown in figures 5.5 and 5.6. As with the sex determination locus transitions, the largest bias in magnitude is at the A locus, though in contrast to before, in most cases it now has a *positive* bias which is strongest in the ancestral heterogametic sex. The exception to the positive bias rule,

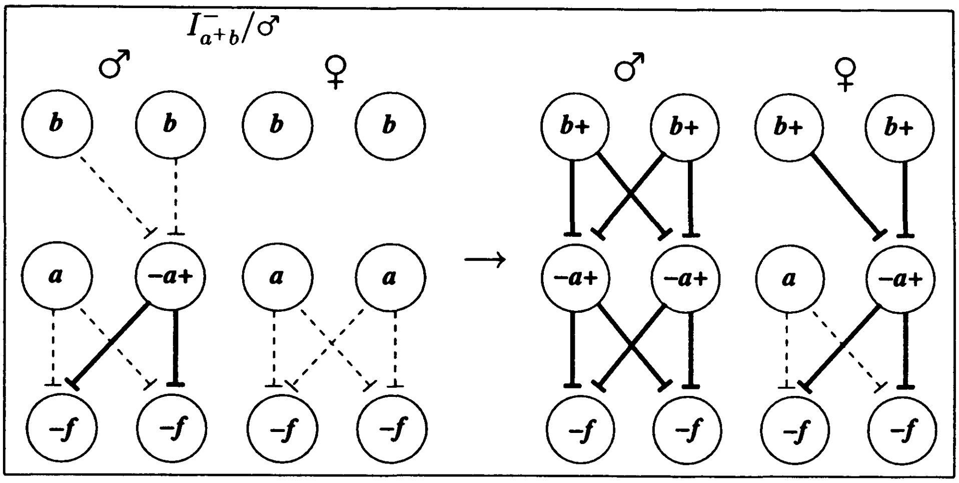


Figure 5.6: One commonly observed single-step transition leading to a heterogamety change at the A locus.

is the single-step heterogamety transition (figure 5.6), which appears with a negative bias.

At the same time, there often appears to be a bias at the F locus of comparable magnitude, making the bias results for heterogamety transitions less consistent than for locus transitions.

5.3 Discussion

Starting from the two locus systems which were observed as evolved outcomes in the previous chapter, we now consider recruitment of a third locus into the pathway. We found that *general* "transition probabilities" (table 5.1) were lower than for the single-locus case. The low "transition probabilities" observed, in turn imply significant bias requirements. Fitness bias requirements were largest at the A locus for all transitions considered. The largest bias values tended to be negative for sex determination transitions, and positive for heterogamety transitions.

The *general* "transition probabilities" do not give as clear a picture as for the ancestral single locus system, since here each "transition probability" usually represents several qualitatively *distinct* transitions. Since many

transition	locus	$\Delta w^{♂}$	$\Delta w^{♀}$
	F	0.24	-0.64
5.5b	A	0.44	**0.66**
/♀	B	N/A	N/A
	F	-0.13	0.35
5.5a	A	**0.5**	0.27
/♂	B	N/A	N/A
	F	0.11	0.37
5.5c	A	**0.42**	0.16
/♂	B	N/A	N/A
	F	0.14	0.20
5.6	A	**-1.4**	-0.23
/♂	B	N/A	N/A

Table 5.4: Fitness bias statistics for all heterogamety transitions. These are shown in figures 5.5 and 5.6. The highest magnitude bias ($|\Delta w|$) is highlighted for each transition. All bias values were found to be statistically significant (t-test) at the 95% level.

transitions were observed at low frequency, only transitions observed at frequencies above 2% were considered. A number of locus transitions (where the B locus becomes discriminatory), and of heterogamety (where the B locus is recruited in a non-discriminatory role causing a change in heterogamety at the A locus), were observed in this range.

As was the case with the single locus ancestral system, it is clear from the results that the key mutation leading to recruitment will occur at the B locus, not at A. Also, as before, the low "transition probability" for initial low output alleles at B ($T_b = 1$), indicate that a strong-effect mutation (e.g. wholesale acquisition of an RNA-binding domain), is more likely to induce recruitment than weak-signal mutants (e.g. small-effect point mutations).

Considering locus transitions, it is clear from the results that if expression of A is heavily deleterious in one sex, then an incentive exists to suppress it in that sex. This is most obvious in the single-step transitions shown in figure 5.3, where the effect of the allele a is suppressed by the transition in the sex for which it is most deleterious. One of the resulting single-step transitions, shown in figure 5.3 (a), leads to a system similar to that of *C.*

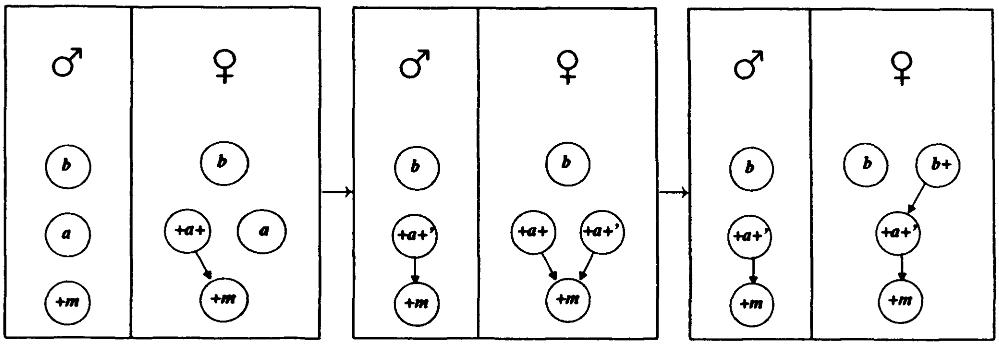


Figure 5.7: Transition similar to that which may have occurred in the recruitment of *Sex-lethal* into the *Drosophila* sex determination pathway.

capitata. Parameter sensitivity (specifically to the sex-determining threshold parameter θ) meant that for the same pre-existing case (I^{-}_{a+b}/♀) we also observe the transition of figure 5.3 (b). Another locus transition, shown in figure 5.3 (c), leads to a system which might reasonably be an ancestral form of the *C. elegans* system.

It was suggested that the remaining locus transitions observed at high frequency (shown in figure 5.4) are biologically unrealistic for two reasons: 1) the inhibitory feedback loop, although stable in this particular case, is unlikely to be generally robust against oscillations, 2) the transitions were only observed for unclamped B which, as was argued previously, is probably less biologically realistic than clamped B, since a gene chosen at random is more likely to be involved in some process neutral with respect to reproductive fitness.

All locus transitions apart from the one just mentioned were observed for unclamped B as well as clamped B. It appears therefore that bias in the recruited gene is not in any way indispensable for recruitment to occur. However, for A locus heterogamety transitions, only clamped B cases are observed at frequency higher than 2%. This result suggests it is more likely that, at least for heterogamety transitions, the recruited gene was unbiased with respect to reproductive fitness.

One notable absence from the described locus transitions is a transition which in any way resembles the recruitment of *Sex-lethal* in *Drosophila*. However, such a transition was indeed observed at a lower frequency of ~1.6%. In the transition shown in figure 5.7, the incorporation of a low output allele $+a^{+'}$ (i.e. $T_{+a^{+'}} = 1$) serves to reduce the expression level of a heavily deleterious A locus, particularly in males, who now become homozygous $+a^{+'}$. This change permits recruitment of the mutant allele b^+ as discriminatory allele, now at the B locus. The second change only affects females, and is possible only if the benefit of upregulating m outweighs upregulation of the deleterious allele $+a^{+'}$. This two-step transition is very similar to that proposed in [92], in which a null allele of *transformer* is incorporated in a similar way as a precursor to recruitment of *Sex-lethal.*

The fitness bias requirements for locus transitions are clearly different from those discovered in the previous chapter. A significant bias at both existing loci (F and A) appears to be necessary for recruitment of B as a discriminatory locus. A strong negative bias at the A locus occurs in the ancestral homogametic sex. The proposed network evolution of [92] made the assumption that differential expression of *doublesex* (F locus) is the primary force driving changes in the network. The results here extend the results of [92], by indicating that differential expression at intermediate loci (in this case the A locus) may also play an important role.

In chapter 4 it was suggested that recruitment of the A locus as a discriminatory locus was most likely to occur if it was *a priori* unbiased, i.e. for clamped A ($w_A^{♂} = w_A^{♀} = 0$). However, in this chapter we have found that a strong bias at A is usually needed for recruitment of locus B. Although it is beyond the scope of this study to consider why such a change in fitness might occur, we can look for transitions assuming clamped A. Under the conditions clamped A and B ($w_A^{♂} = w_A^{♀} = w_B^{♂} = w_B^{♀} = 0$), no high frequency transitions were observed except for one complex transition seen at ~8% under pre-existing case $R_b^+/♀$, which is shown in figure 5.8. Since this transition involves two intermediate states involving populations with more than two genotypes, only the ancestral and final evolved populations are shown. Interestingly, there is a locus transition to B, but in which the B locus has *replaced* the A locus as discriminatory, rather than extend the cascade. The ancestral dominant feminizer allele a^+ evolves into a co-regulator

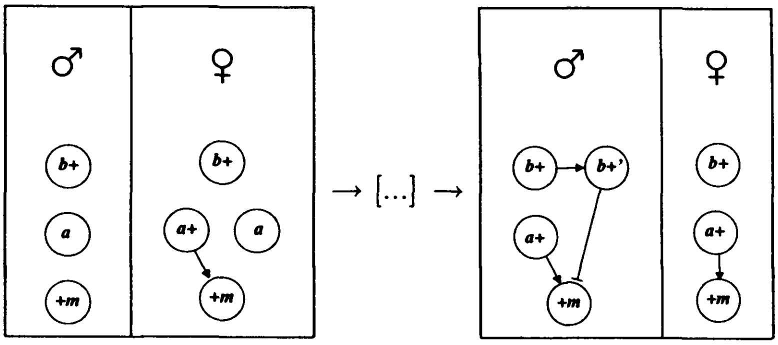


Figure 5.8: Transition observed for clamped A and B ($w_A^{♂} = w_A^{♀} = w_B^{♂} = w_B^{♀} = 0$), under pre-existing case R_b^+. B replaces A as the discriminatory locus.

which is homozygous in both sexes. Co-regulators which are not differentially expressed are common features in known sex determination networks (e.g. *transformer-2* in *D. melanogaster*).

To better understand the issue of fitness bias, it might be useful to consider an extended model in which the fitness contributions $w_{L(i)}$ were defined at the finer-grained allele level rather than at the locus level. The reason for doing this would be to consider certain biologically realistic situations in greater detail. For example, consider ancestral female heterogamety. Here, the dominant feminizer allele $a+$ is not expressed in males, and therefore might evolve independently of a to be strongly beneficial in females, and potentially deleterious in males. This is left for future work.

The three transitions in A locus heterogamety shown in figure 5.5 introduce the possibility that a mutant b^+ allele evolves a double interaction which regulates both the A and the F locus. However, no such pattern has been observed in known sex determination systems. Although it might seem unlikely that one allele (b^+) would be able to regulate another (m') when it already regulates $^+a^+$, this "network motif" has been shown, not only to occur often in gene networks generally [21], but also to have frequently evolved *de novo* [27]. An alternative explanation for the deficit may lie in the fitness bias requirements (table 5.4). Although again the largest $|\Delta w|$ values are to be found at the A locus for the ancestral heterogametic sex, the principal difference is that the bias is positive rather than negative. If it were discovered that the bias at A tended to evolve a negative rather than a positive bias, then this would represent a clear argument as to why we only observe certain evolved structures in nature. For example, in nature we only observe transitions such as those in figures 5.3a (resembling *C. capitata*) and 5.3c (resembling an ancestral form of the *C. elegans* pathway), but not those corresponding to the heterogamety changes of figure 5.5.

The heterogamety transition shown in figure 5.6 illustrates how a mutation which in one case can lead to B becoming the discriminatory sex determination locus (as in figure 5.3), under different conditions can lead to fixation of the new allele b^+ and a change in heterogamety at A.

5.4 Appendix A: Definitions

See also previous definitions for chapter 4.

Term	Description
distinct transition	each *distinct* transition corresponds to a different final evolved population
general "transition probability"	a "transition probability" which represents one or more qualitatively *distinct* transitions for a particular pre-existing case

Chapter 6

Conclusions and Future work

6.1 Conclusions

The study of networks is an increasingly important part of complex systems research across many scientific disciplines. Following the discovery of transcription regulatory (gene) networks in the 1960s, early theoretical work focussed either on simple models of small networks (*microscopic*), or ensemble behaviour of large networks (*macroscopic*). However, more recently, advances in mathematical and computational techniques have meant that more detailed models are possible, particularly of systems studied in developmental biology. If we are to address the evolution of (developmental) gene networks, sex determination networks are a good choice, since, among other things, they evolve rapidly. Of the well characterised sex determination networks, the most convenient for study in an evolutionary context is that of *D. melanogaster*, since the sex determination networks of a number of other insects are being elucidated (e.g. *C. capitata*). As more experimental data become available from high-throughput techniques such as microarrays, it has become possible to use these new data to refine existing models. The successes of topological feature analysis (such as scale-free networks and "network motifs") give a clear indication of such uses.

The aim of this thesis has been to study the evolution of gene networks in sex determination using a modelling approach which takes into account the dynamic behaviour of the network, as well as evolution. The growing availability of molecular data for *D. melanogaster* and related insects means it is now possible to do such an analysis using real networks as reference, rather than hypothetical ones. The thesis is organised around two models: a simpler synchronous logic model (chapters 2 and 3), and a more complex hierachical model (chapters 4 and 5).

In chapter 2, a simple theoretical framework was proposed which combines gene expression dynamics (the synchronous logic model), together with evolution (the concepts of *neighbour* and *neighbourhood*). The synchronous logic model permits us to define equivalence between two networks - defined as reproducing a given dynamics provided we start with the same initial state(s), which in turn allows us to find *neighbourhoods* which can perform the sex determination task with the same dynamics as the known network. One main result from chapter 2 shows that, not only is the set of networks able to perform the sex determination task large, but also that *neighbourhoods* within this set are large, suggesting that a high degree of flexibility is available without compromising the core functionality. A second important result is that the known network has relatively high *local dynamic diversity*, another indication of high flexibility. Whether or not a particular gene network can produce a novel dynamic pattern will depend on two factors prior to a "network mutation" Firstly, the dynamics the network is required to produce, and secondly, the network architecture. From the analysis, we can conclude that evolution can overcome this dual constraint by taking advantage of the combinatorial nature of networks, which lends itself to creating flexibility.

Since the known *Drosophila* network was sufficiently small, it was possible to reconstruct the entire set of equivalent networks, $\mathcal{G}$, using an exhaustive algorithm. Although for larger systems reconstructing $\mathcal{G}$ is not feasible, a recent analysis of the yeast gene network [18] suggests the number of inputs to each gene will in practice be small (93% of genes in yeast were found to have between 1 and 4 known interactions). As a consequence, partially reconstructing (within computationally feasible limits) only the set of most parsimonious networks $\mathcal{M}$ may yield useful results. This assumption was put to use in chapter 3, which suggests how experiments involving large-scale perturbations can be designed to obtain reasonably accurate reconstructions. Keeping in mind that many proposed methods for gene network reconstruction require as many experiments as there are genes, the main result from chapter 3 is that a much lower number of experimental perturbations will improve accuracy substantially, particularly for low-order inputs, as long as the perturbations themselves alter the expression level of approximately half the genes in the network.

In chapter 4, a diploid hierarchical model is presented which integrates techniques from standard population genetics with network dynamics (using Ordinary Differential Equations). The model retains a similar, though slightly extended, model of network evolution to that introduced in chapter 2. The discovery that the key Drosophila gene *doublesex* was highly conserved has added support to a hypothesis proposed by Wilkins [86], that sex determination networks have evolved in a retrograde manner from bottom to top. Following this hypothesis, it is reasonable to make the assumption that the simplest-possible ancestral sex determination systems would be based on a single locus. The transition from a single locus system to a two locus system is analysed in chapter 4, and the further transition from a two locus system to a three locus system is considered in chapter 5.

In chapter 4, the two possible systems (male and female heterogamety) based on a single ancestral sex determining locus F were considered as starting points. The conditions under which a new locus A becomes the new discriminatory locus were studied, with the ancestral sex determining locus becoming homozygous in both sexes. It was found that recruitment of a new discriminatory gene into a pathway is most likely due to a mutation in the recruited gene rather than in a gene from the existing pathway. Furthermore, the mutation must have a strong effect (as opposed to a small-effect point mutation, for example) on its target in order to provoke a transition. Transitions resulting in a locus change predominantly lead to one of two outcomes: (a) a dominant masculinizer at A downregulates F locus expression in males, or (b) a dominant feminizer at A upregulates F locus expression in females. Transitions not leading to a locus change, nonetheless lead to A locus recruitment in such a way that the ancestral discriminatory locus is switched between male and female heterogamety. It is argued that the recruited gene was probably neutral with respect to reproductive fitness. However, if we allow for the contrary possibility that the recruited gene may indeed have made an independent *a priori* contribution to fitness, conflicting patterns in fitness bias emerge between loci and between sexes (e.g. the same gene has a beneficial effect in males, a deleterious effect in females).

In chapter 5, the findings of the previous chapter 4 are extended to consider the incorporation of a third locus (B) into the pathway. Again we consider locus transitions in which the new B locus becomes discriminatory, using the evolved networks from chapter 4 (in which the A locus is discriminatory) as the starting point. As before, heterogamety transitions at the previously discriminatory (A) locus are also considered. The resulting transition rates turn out to be much lower than with one to two locus evolution. These low transition rates are in turn caused by strong requirements in the fitness contribution from certain loci, particularly the A locus. This suggests that differential expression at the A locus (rather than the original discriminatory locus F) is the most important factor affecting recruitment of B as a discriminatory signal. Interestingly, the bias at A was again found to be negative (deleterious) for locus transitions, but in most cases positive (beneficial) for heterogamety transitions. For locus transitions, the strongest bias affected the ancestral homogametic sex, whereas for heterogamety transitions it mostly affected the ancestral heterogametic sex.

As in chapter 4, it was found that mutations leading to a locus transition are most likely due to a strong-effect mutation in the recruited gene B. Among the evolved locus transition networks we observe certain similarities with known sex determination networks. For example, the most commonly observed locus transition led to a network which is qualitatively similar to that of *C. capitata*. A network we might reasonably expect to be ancestral to that of *C. elegans* was also observed at high frequency. In contrast, heterogamety transitions led to networks which have not been observed in nature, though interestingly, many of these networks developed a double interaction to form a common "network motif" known as the feedforward loop

[20].

The results of chapters 4 and 5 rely exclusively on computer simulations. The recent growth in computational power has allowed problems of ever-increasing complexity to be addressed through simulation. However, this approach usually comes with a drawback in that often the results cannot be properly explained, as they would be using a purely mathematical analysis. This drawback appears to relate in particular to the results of chapters 4 and 5, since for certain important model parameters, only three values (high/default/low) were evaluated. Two points should be mentioned here. Firstly, we should consider the research context in which such a study surfaces. It is common that, when a particular scientific question is first addressed (as is the case here), a simulation-based study will suggest a general result which is later developed in greater detail. This has occurred, for example, with the issue of robustness in gene networks. A pioneering simulation-based study [39] has suggested that the *Drosophila* segment polarity network is both modular (in that its inputs can be rearranged without changing the intrinsic behaviour), and robust to perturbations in the system parameters. A subsequent study [133] undertaken by a different research group, has used mathematical techniques to explain certain aspects of the observed behaviour, vastly improving the understanding of the problem. Secondly, it is important to keep in mind that when a model contains crude assumptions, as this one does, the results must also be interpreted at a suitably coarse-grained level. Accordingly, one of the main aims of the study has been to evaluate the qualitatively distinct pre-existing cases (e.g. was the mutation likely to have occurred at the A locus?), and observing the outcomes in broadest possible way, with less importance given to the differences in outcome for different parameter values within each pre-existing case. From the observations we also have been able to derive intuitive explanations for many of the results, which, though fairly obvious in retrospect, had previously proved elusive.

Comparing the network evolution model introduced in chapter 2 with the more sophisticated model presented in chapter 4, a fundamental difference between them is that while the first model treated network transitions as being neutral, the second model deals with directed, and irreversible, network transitions (of sex determination locus or heterogamety). In this sense evolution is "directed" in the second model, where it was not in the first model. We can therefore consider that the second model in a way extends the first model by introducing a directionality to the *neighbourhood.*

6.2 Future work

In the last decade, the wide availability of DNA data has allowed the creation of accurate molecular phylogenies, with useful quantitative estimates of properties like branch lengths, which were previously unavailable. Molecular evolution techniques have also helped researchers infer ancestral gene sequences, and some have gone as far as to synthesize (or "resurrect") the purported ancestral proteins [134]. We propose that phylogenies can be used in a similar way with gene network models to generate hypotheses about network structures and their associated parameters. This evolutionary approach to network inference will help in understanding the order in which genes were added to networks, whether there are single or multiple origins and convergence, and aid in the reconstruction of ancestral networks. Taking advantage of these data, the model introduced in chapter 2 should now be extended in order to consider how networks evolve between species.

Using similar models to those of chapter 2, we have since done some preliminary investigations, and been able to find (by random sampling) evolutionary paths of neighbours between two species, with the constraint that the networks maintain *viability* at each step. By assigning a symbol to every possible network mutation, it is possible to represent each evolutionary path as a string of symbols. Since a large number of evolutionary paths were found, we have analysed the corresponding strings using Hidden Markov Models (HMMs) to look for patterns. Applying a standard first-order HMM method [135], some constraints were found. The method works by inferring a matrix p_{ij}, indicating the probability a mutation j will occur at timestep i, taking only the previous state (timestep $i-1$) into account. Although easily implemented, this method is suboptimal as the following simple example illustrates: Assume there are three nonrepeatable mutations (A, B, C) and we are given the valid strings (BAC, ABC, ACB). The task is to discover the (in this case, obvious) pattern that A always precedes C. The HMM method will show, for example, that C has zero probability of occurring at the first timestep, but that it can occur at later timesteps directly after either A or B. Although helpful (in that a less specific pattern has been discovered), this output makes it difficult to discover the underlying relationships, particularly if they are complex. Given the limitations of this approach, a novel method, perhaps based on Bayesian inference, might be developed.

The work so far has concentrated on evolutionary paths between just two species (*D. melanogaster* and *C. capitata*, the Mediterranean fruitfly). The inference of the evolutionary paths is constrained much further by introducing a multi-species phylogeny, since all branches now participate in more than one path. The first approach would be to assume network mutations (applicable to both models) occur along the branches with frequency equal to the (scaled) branch length. This would require the development of an algorithm for generating random evolutionary paths which are statistically consistent with the branch lengths. The need for such an algorithm derives from the requirement that the number of mutations along a particular path be of integer length (since for both models, we define mutations as discrete events), whereas on a phylogeny inferred from genetic data, they may be real-valued. For this, each integer path length would be taken from a discrete probability distribution with mean equal to the (scaled) real-valued phylogeny branch length. Before we can start though, we need to know the network structures for the leaf nodes (existing species) in the phylogeny, and this knowledge is at present extremely limited. However, a detailed re-creation of the evolution of the Drosophila network based on existing molecular data does exist [92], and this study along with emerging understanding in related species should be sufficient to build a phylogeny consisting of at least five species (using the species above and the domestic house fly [89], honeybee [49] and silkworm [50]). Generating the phylogenetic tree itself (with branch lengths) is fairly straightforward, and could be achieved by reproducing and consolidating results from published phylogenetic analyses [136, 137, 138]

Bibliography

[1] SH. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[2] F. Jacob, D. Perrin, C. Sanchez, and J. Monod. Operon: A group of genes whose expression is coordinated by an operator. *C R Hebd Seances Acad Sci.*, 250:1727–1729, Feb 1960.

[3] TS. Gardner, CR. Cantor, and JJ. Collins. Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403(6767):339–342, 2000.

[4] A. Becskei and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–593, 2000.

[5] MB. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[6] Y. Setty, AE. Mayo, MG. Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci U S A*, 100(13):7702–7707, 2003.

[7] J. Hasty, D. McMillen, and JJ. Collins. Engineered gene circuits. *Nature*, 420(6912):224–230, 2002.

[8] S. Kauffman. The large scale structure and dynamics of gene control circuits: an ensemble approach. *J Theor Biol*, 44(1):167–190, 1974.

[9] R. Thomas. Boolean formalization of genetic control circuits. *J Theor Biol*, 42(3):563–585, 1973.

[10] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–372, 2001.

[11] AL. Barabási and ZN. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, 2004.

[12] E. Szathmáry, F. Jordán, and C. Pál. Molecular biology and evolution. Can genes explain biological complexity? *Science*, 292(5520):1315–1316, 2001.

[13] H. Jeong, SP. Mason, AL. Barabási, and ZN. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[14] P. Uetz, L. Giot, G. Cagney, TA. Mansfield, RS. Judson, JR. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and JM. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, Feb 2000.

[15] I. Xenarios, L. Salwínski, XJ. Duan, P. Higney, SM. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–305, Jan 2002.

[16] R. Mrowka, A. Patzak, and H. Herzel. Is there a bias in proteome research? *Genome Res*, 11(12):1971–1973, Dec 2001.

[17] H. Jeong, B. Tombor, R. Albert, ZN. Oltvai, and AL. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[18] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–63, 2002.

[19] A. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

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

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

[22] TI. Lee, NJ. Rinaldi, F. Robert, DT. Odom, Z. Bar-Joseph, GK. Gerber, NM. Hannett, CT. Harbison, CM. Thompson, I. Simon, J. Zeitlinger, EG. Jennings, HL. Murray, DB. Gordon, B. Ren, JJ. Wyrick, JB. Tagne, TL. Volkert, E. Fraenkel, DK. Gifford, and RA. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[23] Y. Artzy-Randrup, SJ. Fleishman, N. Ben-Tal, and L. Stone. Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, 305(5687):1107–1107, 2004.

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

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

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

[27] GC. Conant and A. Wagner. Convergent evolution of gene circuits. *Nat Genet*, 34(3):264–266, 2003.

[28] H. Salgado, S. Gama-Castro, A. Martínez-Antonio, E. Díaz-Peredo, F. Sánchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jiménez-

Jacinto, A. Santos-Zavaleta, C. Bonavides-Martínez, and J. Collado-Vides. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res*, 32(Database issue):303–306, Jan 2004.

[29] HW. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer, and AP. Zeng. An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res*, 32(22):6643–6649, 2004.

[30] U. Alon. Biological networks: the tinkerer as an engineer. *Science*, 301(5641):1866–1867, 2003.

[31] JD. Han, N. Bertin, T. Hao, DS. Goldberg, GF. Berriz, LV. Zhang, D. Dupuy, AJ. Walhout, ME. Cusick, FP. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, 2004.

[32] J. Stark, D. Brewer, M. Barenco, D. Tomescu, R. Callard, and M. Hubank. Reconstructing gene networks: what are the limits. *Biochem Soc Trans*, 31(Pt 6):1519–1525, 2003.

[33] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[34] H. DeJong. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol*, 9(1):67–103, 2002.

[35] Stuart Kauffman. *The Origins of Order*. Oxford University Press, Inc, USA, New York, NY, 1993.

[36] NA. Monk. Oscillatory expression of Hesl, p53, and NF-kappaB driven by transcriptional time delays. *Curr Biol*, 13(16):1409–1413, 2003.

[37] WJ. Blake, M. KAErn, CR. Cantor, and JJ. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[38] J. Lewis. Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator. *Curr Biol*, 13(16):1398–1408, 2003.

[39] G. von Dassow, E. Meir, EM. Munro, and GM. Odell. The segment polarity network is a robust developmental module. *Nature*, 406(6792):188–192, 2000.

[40] E. Meir, G. von, Dassow, E. Munro, and GM. Odell. Robustness, flexibility, and the role of lateral inhibition in the neurogenic network. *Curr Biol*, 12(10):778–786, 2002.

[41] M. Kerszberg. Noise, delays, robustness, canalization and all that. *Curr Opin Genet Dev*, 14(4):440–445, 2004.

[42] CH. Waddington. Canalization of development and genetic assimilation of acquired characters. *Nature*, 183(4676):1654–1655, 1959.

[43] ML. Siegal and A. Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proc Natl Acad Sci U S A*, 99(16):10528–10532, 2002.

[44] LM. Steinmetz, C. Scharfe, AM. Deutschbauer, D. Mokranjac, ZS. Herman, T. Jones, AM. Chu, G. Giaever, H. Prokisch, PJ. Oefner, and RW. Davis. Systematic screen for human disease genes in yeast. *Nat Genet*, 31(4):400–404, 2002.

[45] Z. Gu, LM. Steinmetz, X. Gu, C. Scharfe, RW. Davis, and WH. Li. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421(6918):63–66, 2003.

[46] A. Pane, M. Salvemini, P. Delli, Bovi, C. Polito, and G. Saccone. The transformer gene in Ceratitis capitata provides a genetic basis for selecting and remembering the sexual fate. *Development*, 129(15):3715–3725, 2002.

[47] M. Hediger, G. Burghardt, C. Siegenthaler, N. Buser, D. Hilfiker-Kleiner, A. Dübendorfer, and D. Bopp. Sex determination in Drosophila melanogaster and Musca domestica converges at the level of the terminal regulator doublesex. *Dev Genes Evol*, 214(1):29–42, 2004.

[48] V. Sievert, S. Kuhn, A. Paululat, and W. Traut. Sequence conservation and expression of the sex-lethal homologue in the fly Megaselia scalaris. *Genome*, 43(2):382–390, 2000.

[49] M. Beye, M. Hasselmann, MK. Fondrk, RE. Page, and SW. Omholt. The gene csd is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell*, 114(4):419–429, 2003.

[50] F. Ohbayashi, MG. Suzuki, K. Mita, K. Okano, and T. Shimada. A homologue of the Drosophila doublesex gene is transcribed into sex-specific mRNA isoforms in the silkworm, Bombyx mori. *Comp Biochem Physiol B Biochem Mol Biol*, 128(1):145–158, 2001.

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

[52] J. Masel. Genetic assimilation can occur in the absence of selection for the assimilating phenotype, suggesting a role for the canalization heuristic. *J Evol Biol*, 17(5):1106–1110, 2004.

[53] CH. Waddington. Genetic assimilation. *Adv Genet*, 10:257–293, 1961.

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

[55] LJ. Johnson and JF. Brookfield. Evolution of spatial expression pattern. *Evol Dev*, 5(6):593–599, 2003.

[56] RV. Solé, P. Fernández, and SA. Kauffman. Adaptive walks in a gene network model of morphogenesis: insights into the Cambrian explosion. *Int J Dev Biol*, 47(7-8):685–693, 2003.

[57] I. Salazar-Ciudad, SA. Newman, and RV. Solé. Phenotypic and dynamical transitions in model genetic networks. I. Emergence of patterns and genotype-phenotype relationships. *Evol Dev*, 3(2):84–94, 2001.

[58] P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19 Suppl 2:122–122, 2003.

[59] Daniel L. Hartl and Andrew G. Clark. *Principles of population genetics, 3rd edition.* Sinauer Associates, Inc, Sunderland, MA, 1997.

[60] F. Galis, TJ. van Dooren, and JA. Metz. Conservation of the segmented germband stage: robustness or pleiotropy? *Trends Genet*, 18(10):504–509, 2002.

[61] JA. de Visser, J. Hermisson, GP. Wagner, L. Ancel Meyers, H. Bagheri-Chaichian, JL. Blanchard, L. Chao, JM. Cheverud, SF. Elena, W. Fontana, G. Gibson, TF. Hansen, D. Krakauer, RC. Lewontin, C. Ofria, SH. Rice, G. von Dassow, A. Wagner, and MC. Whitlock. Perspective: Evolution and detection of genetic robustness. *Evolution Int J Org Evolution*, 57(9):1959–1972, 2003.

[62] H. Lipson, JB. Pollack, and NP. Suh. On the origin of modular variation. *Evolution Int J Org Evolution*, 56(8):1549–1556, 2002.

[63] A. Gardner and W. Zuidema. Is evolvability involved in the origin of modular variation? *Evolution Int J Org Evolution*, 57(6):1448–1450, 2003.

[64] P. Hines and E. Culotta. The evolution of sex. *Science*, 281(5385):1979–1979, 1998.

[65] L. Wolpert, R. Beddington, T. Jessell, P. Lawrence, E. Meyerowitz, and J. Smith. *Principles of development.* Oxford University Press, Oxford, UK, 2001.

[66] D. Zarkower. Establishing sexual dimorphism: conservation amidst diversity? *Nat Rev Genet*, 2(3):175–185, 2001.

[67] JA. Graves. Mammals that break the rules: genetics of marsupials and monotremes. *Annu Rev Genet*, 30:233–260, 1996.

[68] AP. Bird. Gene number, noise reduction and biological complexity. *Trends Genet*, 11(3):94–100, Mar 1995.

[69] P. Bork and R. Copley. The draft sequences. Filling in the gaps. *Nature*, 409(6822):818–820, Feb 2001.

[70] C. Chothia, J. Gough, C. Vogel, and SA. Teichmann. Evolution of the protein repertoire. *Science*, 300(5626):1701–1703, 2003.

[71] L. Patthy. Modular assembly of genes and the evolution of new functions. *Genetica*, 118(2-3):217–231, 2003.

[72] M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, 2003.

[73] Eric H. Davidson. *Genomic Regulatory Systems, 1st edition.* Academic Press, 2001.

[74] Sean B. Carroll, Jennifer K. Grenier, and Scott D. Weatherbee. *From DNA to diversity: molecular genetics and the evolution of animal design.* Blackwell Science, Malden, MA, 2001.

[75] I. Marin and BS. Baker. The evolutionary dynamics of sex determination. *Science*, 281(5385):1990–1994, 1998.

[76] C. Schütt and R. Nöthiger. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development*, 127(4):667–677, 2000.

[77] P. Stothard and D. Pilgrim. Sex-determination gene and pathway evolution in nematodes. *Bioessays*, 25(3):221–231, 2003.

[78] M. Louis, L. Holm, L. Sánchez, and M. Kaufman. A theoretical model for the regulation of Sex-lethal, a gene that controls sex determination and dosage compensation in Drosophila melanogaster. *Genetics*, 165(3):1355–1384, 2003.

[79] V. Heinrichs, LC. Ryner, and BS. Baker. Regulation of sex-specific selection of fruitless 5' splice sites by transformer and transformer-2. *Mol Cell Biol*, 18(1):450–458, 1998.

[80] RJ. Greenspan and JF. Ferveur. Courtship in Drosophila. *Annu Rev Genet*, 34:205–232, 2000.

[81] M. Bernstein and TW. Cline. Differential effects of Sex-lethal mutations on dosage compensation early in Drosophila development. *Genetics*, 136(3):1051–1061, 1994.

[82] JB. Skeath and CQ. Doe. The achaete-scute complex proneural genes contribute to neural precursor specification in the Drosophila CNS. *Curr Biol*, 6(9):1146–1152, 1996.

[83] I. Salazar-Ciudad, J. Jernvall, and SA. Newman. Mechanisms of pattern formation in development and evolution. *Development*, 130(10):2027–2037, 2003.

[84] CS. Raymond, CE. Shamu, MM. Shen, KJ. Seifert, B. Hirsch, J. Hodgkin, and D. Zarkower. Evidence for evolutionary conservation of sex determining genes. *Nature*, 391(6668):691–695, 1998.

[85] Z. Shan, I. Nanda, Y. Wang, M. Schmid, A. Vortkamp, and T. Haaf. Sex-specific expression of an evolutionarily conserved male regulatory gene, DMRT1, in birds. *Cytogenet Cell Genet*, 89(3-4):252–257, 2000.

[86] AS. Wilkins. Moving up the hierarchy: a hypothesis on the evolution of a genetic sex determination pathway. *Bioessays*, 17(1):71–77, 1995.

[87] D. Bopp, G. Calhoun, JI. Horabin, M. Samuels, and P. Schedl. Sex-specific control of Sex-lethal is a conserved mechanism for sex determination in the genus Drosophila. *Development*, 122(3):971–982, 1996.

[88] G. Saccone, I. Peluso, D. Artiaco, E. Giordano, D. Bopp, and LC. Polito. The Ceratitis capitata homologue of the Drosophila sex-determining gene sex-lethal is structurally conserved, but not sex-specifically regulated. *Development*, 125(8):1495–1500, 1998.

[89] M. Meise, D. Hilfiker-Kleiner, A. Dübendorfer, C. Brunner, R. Nöthiger, and D. Bopp. Sex-lethal, the master sex-determining gene in Drosophila, is not sex-specifically regulated in Musca domestica. *Development*, 125(8):1487–1494, 1998.

[90] E. Abouheif and GA. Wray. Evolution of the gene network underlying wing polyphenism in ants. *Science*, 297(5579):249–252, 2002.

[91] T. W. Cline and B. J. Meyer. Vive la difference: males vs females in flies vs worms. *Annu Rev Genet*, 30:637–702, 1996.

[92] A. Pomiankowski, R. Nöthiger, and A. Wilkins. The evolution of the Drosophila sex-determination pathway. *Genetics*, 166(4):1761–1773, 2004.

[93] H. Amrein, M. Gorman, and R. Nthiger. The sex-determining gene tra-2 of Drosophila encodes a putative RNA binding protein. *Cell*, 55(6):1025–1035, 1988.

[94] AE. Christiansen, EL. Keisman, SM. Ahmad, and BS. Baker. Sex comes in from the cold: the integration of sex and pattern. *Trends Genet*, 18(10):510–516, 2002.

[95] EL. Keisman, AE. Christiansen, and BS. Baker. The sex determination gene doublesex regulates the A/P organizer to direct sex-specific patterns of growth in the Drosophila genital imaginal disc. *Dev Cell*, 1(2):215–225, 2001.

[96] A. Kopp, I. Duncan, D. Godt, and SB. Carroll. Genetic control and evolution of sexually dimorphic characters in Drosophila. *Nature*, 408(6812):553–559, 2000.

[97] D. Bopp, JI. Horabin, RA. Lersch, TW. Cline, and P. Schedl. Expression of the Sex-lethal gene is controlled at multiple levels during Drosophila oogenesis. *Development*, 118(3):797–812, 1993.

[98] M. Steinmann-Zwicky, H. Schmid, and R. Nthiger. Cell-autonomous and inductive signals can determine the sex of the germ line of drosophila by regulating the gene Sxl. *Cell*, 57(1):157–166, 1989.

[99] I. Carmi and BJ. Meyer. The primary sex determination signal of Caenorhabditis elegans. *Genetics*, 152(3):999–1015, 1999.

[100] J. Hodgkin. Two types of sex determination in a nematode. *Nature*, 304(5923):267–268, 1983.

[101] J. Hodgkin. Genetic sex determination mechanisms and evolution. *Bioessays*, 14(4):253–261, 1992.

[102] P. Berta, JR. Hawkins, AH. Sinclair, A. Taylor, BL. Griffiths, PN. Goodfellow, and M. Fellous. Genetic evidence equating SRY and the testis-determining factor. *Nature*, 348(6300):448–450, 1990.

[103] MW. Nachtigal, Y. Hirokawa, DL. Enyeart-VanHouten, JN. Flanagan, GD. Hammer, and HA. Ingraham. Wilms' tumor 1 and Dax-1 modulate the orphan nuclear receptor SF-1 in sex-specific gene expression. *Cell*, 93(3):445–454, 1998.

[104] A. De Grandi, V. Calvari, V. Bertini, A. Bulfone, G. Peverali, G. Camerino, G. Borsani, and S. Guioli. The expression pattern of a mouse doublesex-related gene is consistent with a role in gonadal differentiation. *Mech Dev*, 90(2):323–326, 2000.

[105] Frank Rosenblatt. *Principles of Neurodynamics.* Spartan Books, New York, 1962.

[106] K. E. Kurten. Correspondence between neural threshold networks and kauffman boolean cellular automata. *J Phys A*, 21(11):L615–L619, 1988.

[107] S. Bornholdt and T. Rohlf. Topological evolution of dynamical networks: global criticality from local dynamics. *Phys Rev Lett*, 84(26 Pt 1):6114–7, 2000.

[108] S. Bornholdt and K. Sneppen. Robustness as an evolutionary principle. *Proc R Soc Lond B Biol Sci*, 267(1459):2281–6, 2000.

[109] S. B. Carroll. Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, 101(6):577–580, Jun 2000.

[110] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: a case study in rna secondary structures. *Proc R Soc Lond B Biol Sci*, 255(1344):279–84, 1994.

[111] J. Tegner, MK. Yeung, J. Hasty, and JJ. Collins. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc Natl Acad Sci U S A*, 100(10):5944–5949, May 2003.

[112] TS. Gardner, D. di Bernardo, D. Lorenz, and JJ. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, Jul 2003.

[113] S.G. Tringe, A. Wagner, and S.W. Ruby. Enriching for direct regulatory targets in perturbed gene-expression profiles. *Genome Biol*, 5(4):–60, 2004.

[114] A. Wagner. Reconstructing pathways in large genetic networks from genetic perturbations. *J Comput Biol*, 11(1):53–60, 2004.

[115] BN. Kholodenko, A. Kiyatkin, FJ. Bruggeman, E. Sontag, HV. Westerhoff, and JB. Hoek. Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc Natl Acad Sci U S A*, 99(20):12841–12846, Oct 2002.

[116] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput*, pages 17–28, 1999.

[117] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *Proc Natl Acad Sci U S A*, 101(14):4781–4786, 2004.

[118] R. Serra, M. Villani, and A. Semeria. Genetic network models and statistical properties of gene expression data in knock-out experiments. *J Theor Biol*, 227(1):149–157, 2004.

[119] T. MacCarthy, R. Seymour, and A. Pomiankowski. The evolutionary potential of the Drosophila sex determination gene network. *J Theor Biol*, 225(4):461–468, 2003.

[120] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, pages 18–29, 1998.

[121] MK. Yeung, J. Tegnér, and JJ. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*, 99(9):6163–6168, 2002.

[122] ME. Newman, SH. Strogatz, and DJ. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(2 Pt 2):026118–026118, 2001.

[123] RJ. Cho, MJ. Campbell, EA. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, TG. Wolfsberg, AE. Gabrielian, D. Landsman, DJ. Lockhart, and RW. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2(1):65–73, 1998.

[124] PT. Spellman, G. Sherlock, MQ. Zhang, VR. Iyer, K. Anders, MB. Eisen, PO. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, 1998.

[125] H. Yu, NM. Luscombe, J. Qian, and M. Gerstein. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet*, 19(8):422–427, 2003.

[126] P. Sudarsanam, Y. Pilpel, and GM. Church. Genome-wide cooccurrence of promoter elements reveals a cis-regulatory cassette of

rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res*, 12(11):1723–1731, 2002.

[127] MW. Covert, EM. Knight, JL. Reed, MJ. Herrgard, and BO. Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, 2004.

[128] GP. Wagner. What is the promise of developmental evolution? Part I: why is developmental biology necessary to explain evolutionary innovations. *J Exp Zool*, 288(2):95–98, 2000.

[129] R. Nothiger and M. Steinmann-Zwicky. A single principle for sex determination in insects. *Cold Spring Harb Symp Quant Biol*, 50:615–621, 1985.

[130] WR. Rice. Sexually antagonistic genes: experimental evidence. *Science*, 256(5062):1436–1439, 1992.

[131] G. Saccone, A. Pane, and LC. Polito. Sex determination in flies, fruitflies and butterflies. *Genetica*, 116(1):15–23, 2002.

[132] Ronald A. Fisher. *The Genetical Theory of Natural Selection.* Clarendon Press, London, UK, 1930.

[133] NT. Ingolia. Topology and robustness in the Drosophila segment polarity network. *PLoS Biol*, 2(6), Jun 2004.

[134] EA. Gaucher, JM. Thomson, MF. Burgan, and SA. Benner. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature*, 425(6955):285–288, 2003.

[135] L.R. Rabiner and BH. Huang. *Fundamentals of Speech Recognition.* Prentice Hall, New Jersey, USA, 1993.

[136] J. Remsen and P. O'Grady. Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support. *Mol Phylogenet Evol*, 24(2):249–264, 2002.

[137] MV. Bernasconi, C. Valsangiacomo, JC. Piffaretti, and PI. Ward. Phylogenetic relationships among muscoidea (Diptera: calyptratae) based on mitochondrial DNA sequences. *Insect Mol Biol*, 9(1):67–74, 2000.

[138] MC. Arias and WS. Sheppard. Molecular phylogenetics of honey bee subspecies (Apis mellifera L.) inferred from mitochondrial DNA sequence. *Mol Phylogenet Evol*, 5(3):557–566, 1996.