Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 238 (2006) 737–753
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Modelling the evolution of genetic regulatory networks

A.P. Quayle[a,*], S. Bullock[b]

[a]Genome Sciences Centre, BC Cancer Agency, Suite 100, 570 West 7th Avenue, Vancouver, BC, Canada V5Z 4S6
[b]School of Computing, University of Leeds, LS2 9JT, UK

Received 4 October 2004; received in revised form 6 June 2005; accepted 22 June 2005
Available online 10 August 2005

Abstract

An evolutionary model of genetic regulatory networks is developed, based on a model of network encoding and dynamics called the Artificial Genome (AG). This model derives a number of specific genes and their interactions from a string of (initially random) bases in an idealized manner analogous to that employed by natural DNA. The gene expression dynamics are determined by updating the gene network as if it were a simple Boolean network.

The generic behaviour of the AG model is investigated in detail. In particular, we explore the characteristic network topologies generated by the model, their dynamical behaviours, and the typical variance of network connectivities and network structures. These properties are demonstrated to agree with a probabilistic analysis of the model, and the typical network structures generated by the model are shown to lie between those of random networks and scale-free networks in terms of their degree distribution. Evolutionary processes are simulated using a genetic algorithm, with selection acting on a range of properties from gene number and degree of connectivity through periodic behaviour to specific patterns of gene expression. The evolvability of increasingly complex patterns of gene expression is examined in detail. When a degree of redundancy is introduced, the average number of generations required to evolve given targets is reduced, but limits on evolution of complex gene expression patterns remain. In addition, cyclic gene expression patterns with periods that are multiples of shorter expression patterns are shown to be inherently easier to evolve than others. Constraints imposed by the template-matching nature of the AG model generate similar biases towards such expression patterns in networks in initial populations, in addition to the somewhat scale-free nature of these networks. The significance of these results on current understanding of biological evolution is discussed.

Keywords: Evolution; Genetic regulatory networks; Gene expression; Artificial genome; Boolean network model; Genetic algorithm

1. Introduction

Recent advances in the area of functional genomics have developed a range of high-throughput technologies for the generation of gene expression data. Technologies such as microarrays, Affymetrix chips, and SAGE enable a detailed gene expression profile to be determined from cells from different tissues and organs in the body, under different conditions (Harrington et al., 2000; Madden et al., 2000). One of the biggest challenges in this area is to analyse the data being generated, in order to understand how genes and proteins interact with each other in performing different biological processes and functions. In order to develop this understanding, we need to integrate our knowledge of individual gene interactions into network level models of gene/protein dynamics.

In this study, we investigate the behaviour of genetic regulatory networks, and in particular, their evolution. The model used in this study can be broken down conceptually into three different levels; a genome level, a network level, and a finally a population of networks and their evolution. An existing model called the Artificial Genome (AG) (Reil, 1999) is used to encode network structures from the genome level. The use of this model, and the genome level, enables a set of constraints to be associated with the networks in the

*Corresponding author.
E-mail address: aquayle@bcgsc.ca (A.P. Quayle).

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.06.020

model, to produce networks with characteristic structures, prior to any influence of evolution on the network structures. The motivation for using this approach is to model the way in which nature imposes constraints on the genetic networks which can be produced, due to the encoding of the network from the genome level. The dynamics of the resulting networks are modelled using Boolean logic, and a genetic algorithm is used to model the evolution of these networks, generally based on the dynamical properties of the networks.

Prior to running evolutionary simulations, the underlying model of the networks and their dynamics was investigated in detail, in order to understand the results of evolutionary simulations in more detail. The set of constraints on the network structures which results from the AG model was investigated, both analytically (see Appendix A), and numerically, as described in Section 3.1. The behaviour of the evolutionary model was investigated by running simulations towards increasingly complex target functions, to test the power of selection in evolutionary processes to evolve complex target systems. The structures of the evolved AG networks were then investigated, and compared to the "random" AG networks which comprise the initial population in the model.

The networks are modelled as Boolean networks in order to make the investigations and simulations realistic on a global network scale. A Boolean network model is a type of logical network model, where the set of genetic interactions can be represented as a type of directed graph, consisting of a set of vertices (genes), and a corresponding set of directed edges (gene–gene interactions). The expression level of any gene is assumed to be either on or off and is determined by the states of the genes that interact with it. The dynamics of the resulting network can then be determined from the starting states of the genes in a network, given a set of rules to describe how the state of each gene depends on the states of the genes that influence it.

Although a Boolean description of the expression of a gene is a simplification, there is evidence of multistationarity in real biological networks, with switch-like transitions between the corresponding steady states (Keller, 1995). The use of Boolean network theory to describe the dynamics of genetic networks was first applied to relatively simple biological networks (Kauffman, 1974). More recently, these ideas have been used to study the global properties of large-scale regulatory systems (Somogyi and Sniegoski, 1996).

A number of alternative approaches have been used to model the behaviour of genetic regulatory networks, including networks based on Bayesian probability (Friedman et al., 2000), generalized logical network models (Thomas, 1991), and systems of differential equations (Weber, 1965; Cherry and Adler, 2000), which are all deterministic in nature. An alternative group of models called stochastic models attempt to incorporate the intrinsic variability and probabilistic nature of gene expression, which are used for smaller systems when a more detailed analysis is possible or required (McAdams and Arkin, 1998; Gillespie, 2000). The use of a Boolean network model in the current study simplifies the network dynamics, allowing us to focus on the evolution of genetic network dynamics.

The AG model used in this study describes the encoding of network structures from the genome level (Reil, 1999). This model, described in full below, uses a method of synchronous update, which is a simplification of changes in gene expression in nature. Some studies have investigated the use of an asynchronous method of update for the AG model. A comparison of the two methods of update showed that some of the order and cyclic properties of the networks generated using synchronous updating are lost when asynchronous updating is used, or are at least less clearly defined (Tong, 2002).

There has been previous work investigating various aspects of the evolution of networks, with some studies focusing on regulatory and developmental networks. A range of theories relating to the evolution of regulatory networks have been developed (Wilkins, 2002), and many recent approaches have combined theoretical approaches with large gene expression datasets. Theoretical advances in this type of modelling include the development of the theory of scale-free networks, which provides a new framework for modelling the evolution of regulatory systems (Gibson and Honeycutt, 2002). More general approaches to network modelling have also been undertaken, with simple logical rules determining the growth or loss of connections (Bornholdt and Rohlf, 2000).

The resulting dynamical behaviour of networks from the AG model is a central part of this study, since many of the fitness functions used in the genetic algorithms act on properties associated with the network dynamics. The state of a network in the model is defined by the set of gene expression levels of all the genes in the network. The use of a synchronous method of update means that the resulting dynamics are fully deterministic. Hence, the network dynamics are determined explicitly by the network architecture and an initial state, and follow a unique trajectory.

Dynamical systems theory organizes the state space of a system into regions called basins of attraction, which contain attractor states (Wuensche, 1998). Attractors can be thought of as a set of equilibrium states of a system, and can exist as either point attractors or cyclic attractors. A deterministic system will converge to an attractor state of the system, and will then remain in this state unless an external force is applied. An example of a basin of attraction containing a cyclic attractor is shown in Fig. 1. The basin of attraction is represented as a

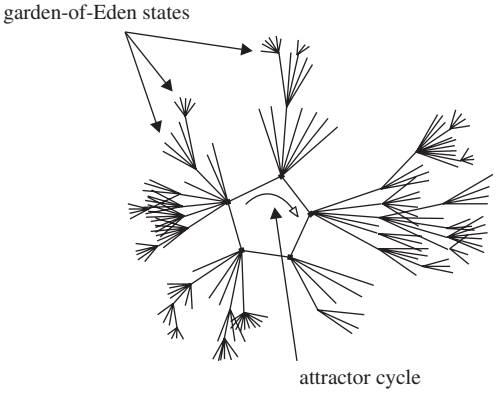

Fig. 1. Example of a basin of attraction containing a cyclic attractor.

particular type of graph, where each vertex represents a single state of the system, and the edges correspond to discrete time steps.

The path or trajectory that a system takes towards an attractor state is called a transient, and the period between the initial state and the start of the attractor is the transient length. An attractor state also has an associated length, which is the time period of the cycle (one for a point attractor). The length of an attractor is a useful classification measure, although an additional measure is required for a unique classification of attractors (Bagley and Glass, 1996).

Random genomes of sufficient length will typically give rise to multiple attractor states in the AG model, which is consistent with the notion that an attractor state of gene expression in a biological network model represents a cell type, as first put forward by Kauffman (1974). In addition, the system in the model is able to differentiate between attractor states, in analogy to the differentiation observed between biological cell types. The cell type and the gene expression pattern can be thought of as the phenotype, illustrating the importance of the resulting attractor states in the model. For this reason, the fitness functions in the genetic algorithm are generally based on properties of the attractor states produced from the network dynamics.

2. Model and methods

We adopted Reil's artificial genome model as a representation of how genetic encoding constrains the structure of gene regulatory networks (Reil, 1999). A genome is represented by a linear sequence of "bases" drawn from the set $\{0,1,2,3\}$ (analogous to the four bases, A, C, G, and T in DNA). Within this genome, every occurrence of the sequence $\{0\,1\,0\,1\}$ is identified as a promoter (analogous to the "TATA" sequence in biological genomes). The g bases immediately downstream of every promoter are identified as a gene ($g = 6$, by default). Every gene encodes for a regulatory protein, also represented by a string of g characters drawn from the set $\{0,1,2,3\}$. Each element of a protein is constructed by incrementing (modulo 4) the value at the corresponding locus of the gene (e.g. $\{0\,1\,2\,3\,3\,3\} \rightarrow \{1\,2\,3\,0\,0\,0\}$). Upon expression, these proteins are free to bind to any matching sequence within the genome and are assumed to do so instantaneously and for the duration of one clock tick. We allow binding sites to overlap (and be bound to simultaneously), but do not allow promoter sites or genes to overlap with one another. Once bound, a protein regulates the closest downstream gene. The nature of this regulation is determined by the value of the character immediately following the binding site. By default, regulation will *activate* gene expression unless a $\{0\}$ occupies this location, in which case it is inhibited (hence, on average, 25% of regulation is inhibitory). Where simultaneous, conflicting regulatory influences occur, inhibition vetoes activation.

Fig. 2(a) depicts the model in schematic and slightly simplified form. In order to encode a simple network in a very short sequence of bases, shorter genes and promoter sequences have been employed, and we include less non-functional "junk" DNA than would tend to be present if the genome had been generated at random.

It is simple to derive a network of gene regulatory interactions from an AG (see Fig. 2(b)). Since, given an appropriate arrangement of binding sites, any gene can be regulated by an arbitrary number of other genes and can itself regulate multiple genes, any gene regulatory network may be encoded in this way. However, since there is not infinite space on a genome or between consecutive genes on a genome, in practice there are constraints on genetic encoding which favour some kinds of gene regulatory network over others.

Given a particular gene regulatory network, patterns of gene expression may be generated by imposing some update scheme. Here, these "network dynamics" are produced (from a given starting state of the system, say one random gene activated) by interpreting the network as a synchronous, discrete time, binary automaton (Kauffman, 1993). At each time step, each gene may either be active (1), i.e. expressing a protein, or inactive (0). The state of each gene at the next clock tick is determined by how it is currently being regulated. By default, genes are inactive. If a gene is being activated by a currently bound regulatory protein then it will become active, unless it is being inhibited by any currently bound regulatory protein. Over a series of time steps, this scheme produces a "piano roll" representation of the expression pattern generated by the network of genes (see Fig. 2(c)).

Systems such as the one described here are typically capable of exhibiting three regimes of dynamical

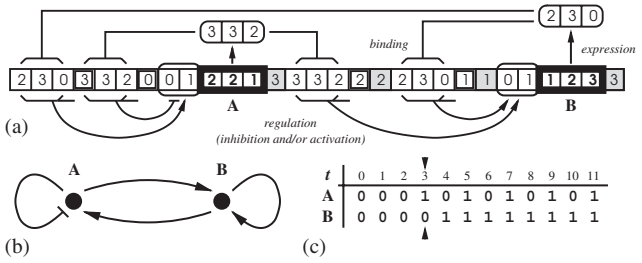

Fig. 2. A schematic representation of the AG model (with a simplified promoter and short gene length). (a) A simple hand-constructed genome consisting of a string of digits drawn from the set {0, 1, 2, 3}. A simple "promoter sequence", {01}, occurs twice within this string. The three characters immediately downstream of each promoter are "genes" (labelled **A** & **B**) which, when expressed, give rise to regulatory proteins (depicted as two lozenges above the genome). Each element of a protein is constructed by incrementing (modulo 4) the value at the corresponding locus of the gene. Hence **A**: {2 2 1} → {3 3 2} and **B**: {1 2 3} → {2 3 0}. Upon expression, these proteins are free to bind to any matching sequence within the genome and are assumed to do so instantaneously and for the duration of one clock tick. Once bound, they regulate the closest downstream gene. The nature of this regulation is determined by the character immediately following the binding site (double boxed in the diagram). By default, regulation will *activate* gene expression unless a zero occupies this location, in which case it is inhibited. Where simultaneous, conflicting regulatory influences occur, inhibition vetoes activation. Bases that are not involved in any gene regulation/expression/binding, etc. are depicted in grey. (b) The gene regulatory network resulting from the genome depicted in (a). **A** activates the expression of **B** and inhibits its own expression, and **B** activates the expression of both itself and **A**. (c) The pattern of gene expression over time generated by the network depicted in (b). At $t = 0$ neither gene is being expressed. This state persists until $t = 3$, when some form of external stimulation (depicted by arrows) triggers the expression of **A**. At the next time step, **A** is inhibited, while **B**'s expression is activated. Subsequently, a cyclic pattern is generated. Note: In this paper we explore genomes that are considerably more complex than the one depicted here.

behaviour: static, cyclic and chaotic. Previous work on the AG model demonstrated these three regimes, and determined the parameter settings in the model that produce them (Reil, 1999). Here we explore parameter settings that tend to produce networks that exhibit cyclic behaviour—a genome length of 10,000 bases and a gene length of 6. This results in a mean network connectivity, K, in the region $3 < K < 7$.

We employ a genetic algorithm to discover genomes that encode particular gene regulatory network structures or exhibit particular kinds of dynamic. An initial population of 50 random genomes is assessed, with each genome assigned some score dependent on what properties we are interested in. High scoring genomes are selected (with replacement) to sexually "reproduce" by combining genetic material to form an offspring genome. Every generation, 50 offspring are formed in this way. This process is not error-free. With probability 0.0002, a base is mis-copied (i.e. on average 2 bases are mutated per offspring genome). Over many generations (we typically run between 1000 and 10,000), the population of genomes will tend to comprise individuals tailored to the particular selection pressures being imposed, e.g. genomes encoding genetic regulatory networks capable of exhibiting a particular pattern of dynamic gene expression.

3. Results and discussion

3.1. Network constraints: dynamics and structure

The dynamical properties of the underlying AG model were investigated, to test the behaviour of the basic model. The networks encoded from random genomes were shown to display the three possible regimes of dynamical behaviour, of static, cyclic and chaotic behaviour. The average degree of a network, K, is the dominating parameter determining the resulting dynamical behaviour for a given network. The regimes of behaviour were found to lie approximately in the regions defined by $K < 2$ for static behaviour, $2 < K < 7$ for cyclic or static behaviour, and $K > 7$ for chaotic behaviour. In the region where cyclic or static behaviour is observed, static behaviour becomes increasingly rare for increasing K.

A probabilistic analysis of the dependency of AG network properties on parameters in the model was developed (see Appendix A). Characterizing the network structures that result from random genomes is important, since the dynamical behaviours of networks are strongly dependent on these structures. In particular, the variation of network structures is important, since this provides the initial variation in the population for evolutionary simulations.

A simplistic measure of the variation of network structures from the model can be obtained from the variation of the average degree of the networks. It can be shown that the variation of the number of genes in the model is expected to follow a Poisson distribution for a discrete variable (see Appendix A). Similarly, the variation of the average degree of networks is expected to follow a Normal distribution, since this is a continuous variable. Fig. 3 shows the variation of the network degree over 500 runs of the program, where each simulation run uses an independently generated random genome.

The mean degree over 500 runs was 2.24, with a standard deviation of 0.24 and a total range of 1.52. Fig. 3 shows the Normal curve corresponding to this mean and standard deviation, and demonstrates that the variation of network degree is very close to a Normal distribution, as expected. The variation of the network degree indicates that an initial population of networks created from random genomes of a fixed length will contain networks with many different average degree values, and correspondingly variable network structures.

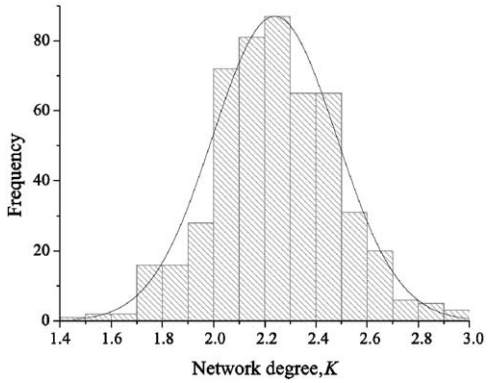

Fig. 3. Variation of network degree, K, from 500 different runs, with genome length = 10,000, gene length = 6, base = 4 and inhibition = 1, against the corresponding Normal distribution.

A derivation of the expected variance of the resulting network structures produced by the AG model is given in Appendix A. Comparing the variance of the vertex (gene) degree, k, for networks from the AG model with that of random networks shows that for random networks, $\sigma_k = \sqrt{\mu_K}$, while for AG networks, $\sigma_k \sim \mu_K$, where σ_k is the standard deviation of the vertex degree, and μ_K is the mean or expected network degree (and $\mu_k = \mu_K$). We distinguish the gene degree, k, from the average degree of a network (or network degree), K, since although the mean of these parameters is clearly equivalent, the standard deviation of these parameters is different. Note that although AG networks are generated from a random genome, the typical network structures produced are significantly different from a typical random network (e.g. Erdös and Rényi, 1960), as shown in Appendix A.

The expression for AG networks cannot be derived exactly (see Appendix A), but an approximation can be made to produce a good estimate. This result clearly shows a greater variance in gene degree in AG networks compared to random networks. This results from the template-matching scheme used to determine which genes interact. The predictions from this analysis were tested by running the AG model 500 times, for different promoter lengths, and appropriate mean and variance values were compared for each network. Table 1 presents the results of these calculations against values predicted by the analysis presented in Appendix A.

The data shows a close similarity to the values predicted by the analysis, although there is some variation between data and theory values. This variation is up to around 10% of the predicted values, and a certain amount of inherent variation is expected due to the random nature of genome generation. The variance of the gene degree is of particular interest, and the data matches the predicted values from the analysis, to clarify the difference between the structure of AG networks and random networks.

Table 1
Mean and standard deviation values of the distribution of the gene distance (μ_d and σ_d) and gene degree (μ_k and σ_k) in AG networks, for different promoter lengths

Promoter length	2	3	4
μ_d (Data)	21.49	74.49	284.11
μ_d (Theory)	23.00	72.00	265.00
σ_d (Data)	14.91	64.82	235.75
σ_d (Theory)	16.00	64.00	256.00
μ_k (Data)	2.42	2.38	2.23
μ_k (Theory)	2.43	2.42	2.36
σ_k (Data)	2.08	2.44	2.24
σ_k (Theory)	2.43	2.42	2.36

Since this prediction is validated, there is clearly some degree of inherent order in the networks produced by the AG model. The variance of the gene degree gives a scalar measure of the network structure, but does not provide a complete description of the network structure. An analysis of the degree distributions of the resulting networks provides a more detailed account of the variation of network structures. Two distinctive degree distributions that have received significant attention are those of random networks (Erdös and Rényi, 1960) and scale-free networks (Barabási and Albert, 1999). A characteristic bell-shaped curve describes the degree distribution of a random network, as shown in Fig. 4, where the degree of most vertices is close to the average degree, and the frequency decreases exponentially on either side of the maximum. Conversely, scale-free networks exhibit a power-law degree distribution as shown in Fig. 4. In this case it can be seen that most vertices have very few connections, while some vertices act as hubs, and are highly connected. The degree distributions of networks produced from the AG model were calculated, and averaged to obtain a characteristic distribution, as shown in Fig. 5.

The distributions appear to be similar to that of scale-free networks, with many genes having no or very few connections, and a few genes having many connections. However, the distribution of AG networks does not fall off as rapidly as the power-law of scale-free networks, and is closer to the exponential decrease of the random networks. Additionally, the distribution for a genome length of 20,000 has a maximum at $x = 1$, which is not produced by a power-law distribution. For larger genome lengths, the maximum of the distribution increases, showing that the distribution of AG networks is not described by a power-law. This is confirmed by a log–log plot of the frequency against the gene degree. Such a plot for scale-free networks produces a straight line with a gradient of the corresponding power ($f(x) = x^{\alpha} \Rightarrow \log(f(x)) = \alpha \times \log(x)$), but Fig. 6 clearly shows that this is not true for AG networks.

These results show that AG networks in fact have degree distributions between those of random networks

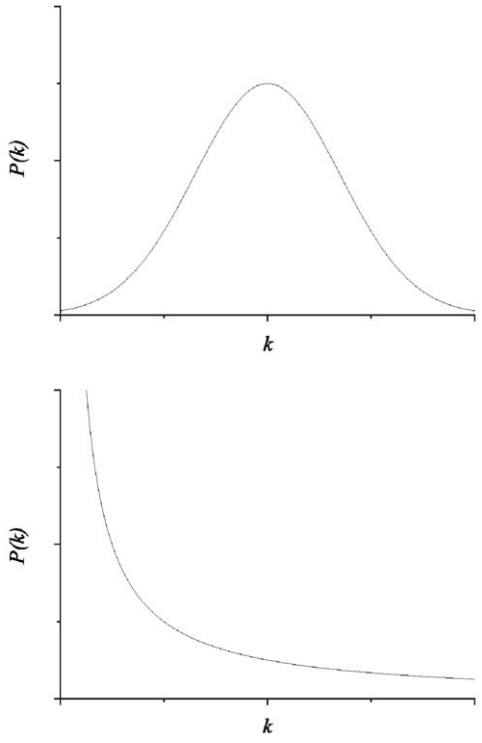

Fig. 4. Characteristic degree distributions of random networks (top) and scale-free networks (bottom).

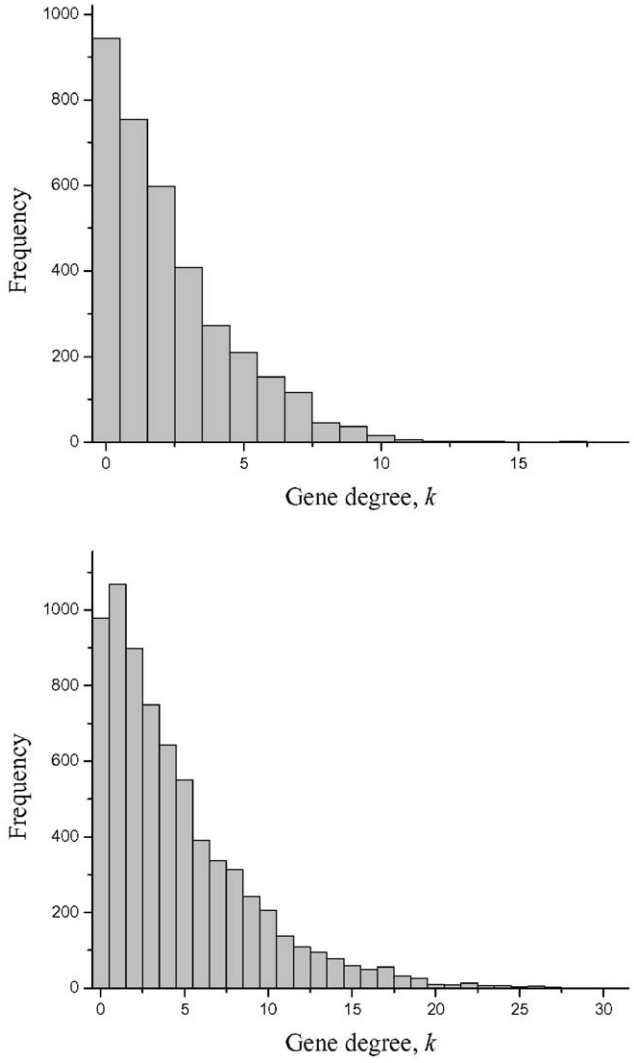

Fig. 5. Degree distribution of AG networks with genome length = 10,000 (top) and genome length = 20,000 (bottom), generated from 100 random genomes for each plot.

and scale-free networks. An explanation of the specific degree distribution for these networks relates to the analysis presented in Appendix A in which it was shown that the distance between adjacent genes on random genomes varies as a geometric distribution, and the gene degree varies as a binomial distribution, which can be approximated by a Poisson distribution. Hence the resulting degree distribution is a combination of the geometic and Poisson distributions, which can be seen in the results presented in Fig. 5. As explained in Appendix A, an exact expression for the distribution is not possible, but the interaction of the two distributions can be seen in the results. The distribution for a genome length of 10,000 is very similar to a geometric distribution, and for a higher genome length, the influence of the Poisson variation becomes visible in the position of the maximum, and the decrease either side of the maximum. Details of network structures which are evolved in later simulations can now be compared more accurately to the structures of networks produced by the AG model, which are present in the initial population of evolutionary simulations.

Since we assume that an attractor in the model corresponds to a cell type (as discussed in Section 1), we principally study the properties of these attractors for their influence on selection. The different types of attractor produced from networks in the model are important, since a given network can have many different attractor states, but only one type of attractor may correspond to the target cell-type. The typical range of attractor lengths produced by the model was investigated by running the program for 500 different random genomes. For each genome, the program was run 2000 times in order to find as many attractor states of that genome as possible. This is a sufficient number of repetitions such that many simulations discover the same attractors, which will typically correspond to the largest basins of attraction, while keeping the required computational time to a reasonable level. It would be impossible to guarantee the discovery of every possible attractor of a given genome, due to the huge number of possible initial states, n_S, since,

$$n_S = \sum_{r=1}^{n} \frac{n!}{(n-r)!}, \tag{1}$$

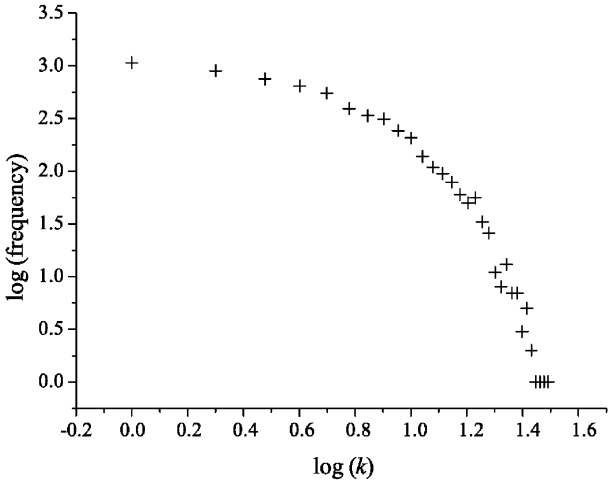

Fig. 6. Log–log plot of frequency against gene degree for AG networks with genome length = 20,000.

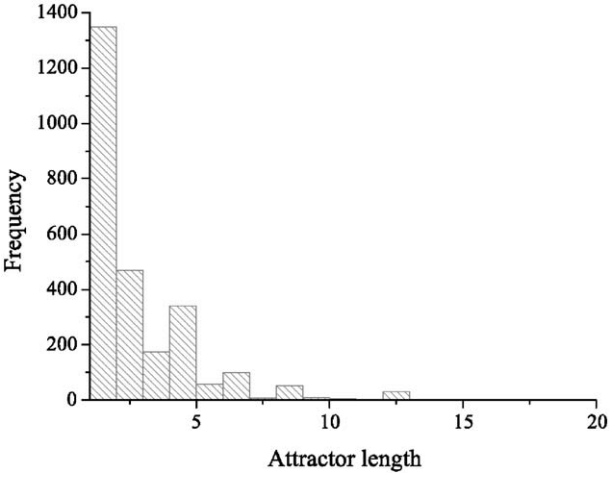

Fig. 7. Distribution of attractor lengths, with genome length 10,000, gene length = 6, base = 4, inhibition = 1, over 2000 runs, averaged over 500 different genomes.

where n is the number of genes. We used initial states where 5 randomly selected genes are expressed, after running some initial tests to investigate the optimum number of genes expressed in order to discover the most attractors of a given genome. A histogram of the resulting attractor lengths is shown in Fig. 7.

The most common attractor length is 1, corresponding to static gene expression. Many different cyclic attractor states were found, where in general the frequency of the attractors decreases with increasing attractor length, and the largest attractor length was 33. The overall mean attractor length was 2.40, and the mean number of different attractors per genome was 5.22. This result confirms the existence of multiple attractor states for each random genome and associated network. For shorter genomes (i.e. simpler networks) more static attractor states are observed, whereas longer genomes exhibit more cyclic attractors and a greater mean attractor length. The detailed distribution of attractor lengths is particularly important in the context of the evolvability of gene expression patterns with particular periods presented in Section 3.2.2.

3.2. Network evolution

3.2.1. Simple network properties

The evolutionary model of genetic regulatory networks was initially tested by using fitness functions that select for some simple network properties. These simulations act as baseline conditions, which can be used to determine appropriate parameter settings for use in later simulations. Research into the optimum parameter settings to be used in GAs has shown that there is generally no way to derive these a priori, and that the best values are usually problem specific.

Sets of simulations where fitness is proportional to the size of the networks (i.e., the number of genes in the underlying genomes) and the network average degree were used to determine the optimum mutation rate. Although the simulations described in Sections 3.2.2–3.2.4 use targets of much greater complexity, the presence or absence of genes and the network degree are fundamental to acheiving these more complex targets. Generally a high mutation rate gives a high initial rate of evolution, but does not produce the maximum extent of evolution after many generations, due to the mutation/selection balance. In terms of the associated search space and fitness landscape for a given problem, a high mutation rate enables the system to escape from local maxima, but may often prevent the system from reaching the global maximum, since it cannot sustain fitness as effectively as a simulation involving a lower mutation rate.

By averaging fitness plots over 50 simulations for a given mutation rate, it was found in both sets of simulations that a mutation rate of 0.0002 in the model optimizes the extent of evolution. The fitness plots resulting from simulations where the fitness function is proportional to the network average degree were significantly more rugged than those with fitness functions proportional to the network size, reflecting the increased complexity of the target system. The optimum mutation rate was found to be the same in both cases, since although a higher mutation rate would allow a greater chance of escaping local maxima in the more rugged fitness landscape, the system cannot maintain fitness as effectively in order to reach the global maximum. The extent of evolution was measured by the average individual fitness in the population (population fitness), where the fitness of an individual is equal to either the number of genes or network degree, as appropriate.

Many evolutionary studies consider only the individual fitness and not the population fitness as the chosen method of assessing evolutionary fitness. We use the population fitness in this case since it provides a more sensitive measure when evolving towards relatively simple target systems, such that the time to reach maximum fitness covers a greater range, and hence allows a greater differentiation between the ability of selection to evolve these targets. Maximum population fitness occurs only when all individuals in the population have maximum fitness, for example, each individual has a fitness of 1.0 when fitness measures are normalized between 0.0 and 1.0. In later simulations towards relatively complex targets, we compare simulations based on population fitness with those based on individual fitness.

Generally in simulations attempting to model the evolution of genetic networks we are interested in modelling the evolution of particular patterns of gene expression rather than an arbitrarily large number of genes or arbitrarily large network degree. Before attempting to evolve gene networks capable of generating particular patterns of gene expression (described in Sections 3.2.2–3.2.4, below), we first examine the extent to which particular kinds of network structure can be selected for. For genomes of a certain fixed length, we derive the expected number of genes and the corresponding expected network degree. By selecting for networks with a greater or lesser number of genes, or for networks with increased or decreased degree compared with this baseline, we can assess the extent to which selection is able to influence gross network structure.

From the derivation in Appendix A, the mean number of genes, μ_G, and mean network degree, μ_K, are given by,

$$\mu_G = \frac{l}{(B^p + g + p - 1)}, \tag{2}$$

$$\mu_K = \frac{l}{(B^p + g + p - 1)}\left(1 - \exp\left(-\frac{B^p + p + g - 1}{B^g}\right)\right), \tag{3}$$

where B is the number of possible bases at each locus, p is the length of the promoter sequence, and g is the length of each gene sequence. From Eqs. (2) and (3), the expected number of genes and network degree for a genome length of 10,000 are given by

$$\mu_G = 37.74, \quad \mu_K = 2.36. \tag{4}$$

Targets were chosen at equal spacing from these expected values, in order to compare the rate of evolution in either direction. Target values for the number of genes of $G = 20$ and 55 were selected. Since the network degree is correlated with the gene number, we were able to calculate equivalent values for the network degree, $K = 1.25$ and 3.45, from Eq. (24)

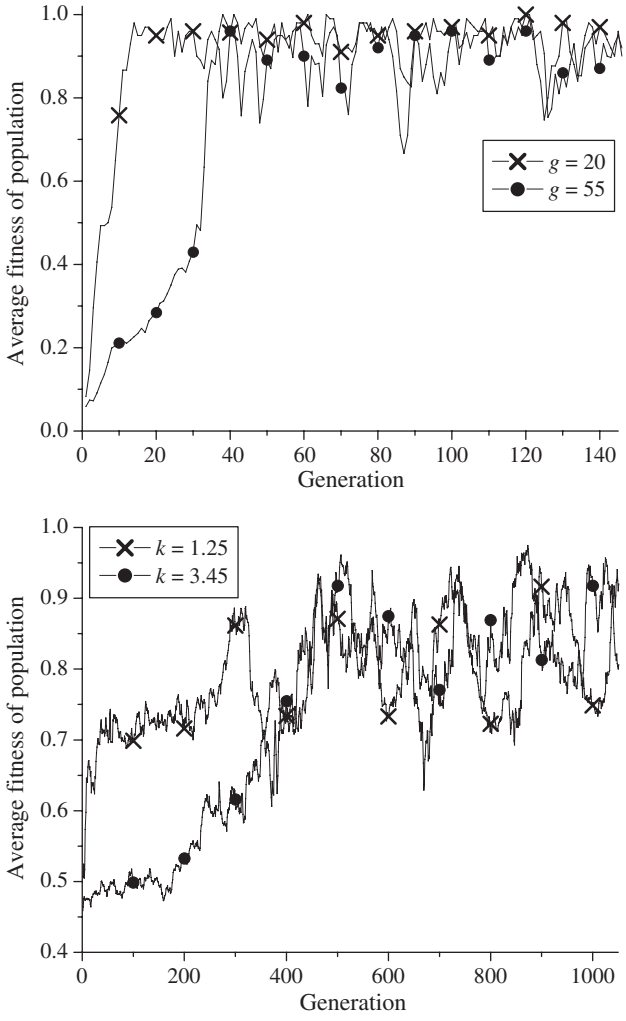

Fig. 8. Plot of the average fitness of the population against the generation number, with selection criteria of the number of genes (top), and the network degree (bottom), and target values as illustrated.

(Appendix A). Simulations were run using these target values, with all other parameters set to the values discussed, and the results are presented in Fig. 8. Unless otherwise stated, all simulations described use parameter values of genome length 10,000, population size 50, and mutation rate 0.0002. The function used to calculate individual fitness for simulations evolving towards target values of a specified parameter is given by

$$f_i = \frac{1}{1 + |x_i - x_t|}, \tag{5}$$

where f_i is the individual fitness, x_i is the current parameter value for individual i, and x_t is the target parameter value. The population fitness is the average individual fitness in the population, and hence,

$$f_p = \frac{1}{n}\sum_{i=1}^{n} f_i, \tag{6}$$

where f_p is the population fitness and n is the population size. The fitness function defined in Eq. (5) is used in the simulations described above, and also in simulations evolving towards target attractor lengths (see Section 3.2.2).

This fitness function normalizes individual fitness between 0.0 and 1.0, and hence the maximum possible population fitness is also 1.0. Comparing the four different simulations plotted in Fig. 8, a number of observations can be made from the results. Firstly, the number of generations taken to reach the fittest population is much greater if the selection criteria is the network degree, which requires around 500 generations to reach the fittest population, compared with around 50 generations for the number of genes. This represents the population's increased difficulty in satisfying the fitness function, presumably due to increased ruggedness in the fitness landscape for network degree, compared to that for gene number.

Second, both sets of data show that it is more difficult for evolution to bring about an increase in either the number of genes or network degree than a decrease. This is because it is easier to destroy rather than create a promoter sequence or binding site through random mutations.

The final most important result is the fluctuation of the population fitness once the fittest, or close to fittest population has been reached, both in terms of the extent of fluctuation and the typical oscillation period of the fluctuations. For simulations where the selection criteria is the number of genes, fluctuations are smaller, and the oscillation period is much less than in the case of the network degree. Since it is more difficult to evolve to target values for the network degree, it is similarly more difficult to maintain values which are close to the target. Hence when random mutations reduce the population fitness, a greater number of generations is required to compensate for this loss in fitness. This also has the effect that the population fitness is on average less for simulations evolving to target degree values than to a target number of genes. Under selection for gene number, the population fitness is often close to 1.0, but this is rarely observed during selection for network degree.

These results illustrate the ability of the genetic algorithm to evolve specific gross network properties, and also the effect of the complexity of the target property which is being selected for. These act as a basis for understanding the results of evolutionary simulations where populations are evolved towards more specific behavioural properties.

3.2.2. Cyclic behaviour

Since we are more interested in the behaviour that genetic networks are capable of than their structural properties (e.g. size, degree, etc.), we now consider the extent to which selection can bring about particular patterns of gene expression. In particular, we are interested in the cyclic patterns of gene expression that are exhibited when a network falls into a periodic attractor. The simplest property of such an attractor is its length (the characteristic period of its cyclic behaviour). First, simulations were run in which selection favoured networks able to generate cyclic behaviour of a specified period. Selected results are presented in Fig. 9.

The fitness plots for each of the target attractor lengths are qualitatively similar to those achieved under selection for particular network structural properties (see above). However, there are a number of differences between these two types of simulations, and also between simulations with different target attractor lengths. The fitness plots shown in Fig. 9 display short periods of rapid evolution, inbetween longer periods of stasis, which occur due to the population becoming trapped in local maxima in the fitness landscape. This "trapping" behaviour becomes more likely as the target complexity increases and the fitness landscape becomes correspondingly more rugged.

The level of fluctuation in the population fitness also varies with the target attractor length, with greater fluctuations for longer attractor lengths, again indicating the increasing ruggedness of the fitness landscape for increasing target complexity. However, in each case a population fitness of 1.0 was observed at some generation in the simulation, and hence it is clear that selection has the power to evolve to particular attractor lengths (see Eqs. (5) and (6) for fitness definitions).

To investigate in more detail the relationship between the target attractor length and the associated evolvability, simulations were run for each attractor length between 1 and 15. In this case and in all simulations

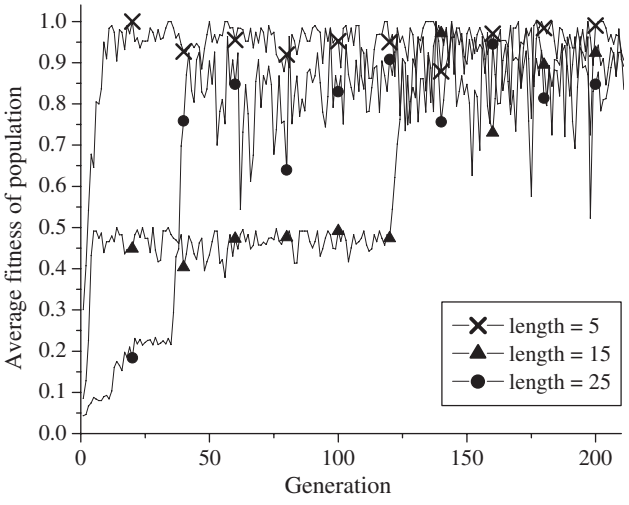

Fig. 9. Plot of the average fitness of the population against generation number, for a range of target attractor lengths.

described later, we calculate average evolvabilities over 20 simulation runs for a given target. Simulations towards target expression patterns are computationally intensive, and hence this number of runs was chosen to allow simulations to be completed in a reasonable time, while providing a sufficient number of runs to calculate an average evolvability in each case. The associated evolvability was measured by the number of generations taken for the population to reach a fitness of 1.0 (where every individual in the population has reached the required attractor length). Fig. 10 illustrates the results of these simulations, where the mean number of generations to reach a population fitness of 1.0 is plotted for each attractor length.

These results show that there is a general trend where longer attractor lengths are more difficult to evolve, but there is a significant variation for particular attractor lengths. It can be seen that some attractor lengths are inherently easier to evolve than others, and that attractor lengths with a greater evolvability are multiples of each other. For example, attractor lengths of 2, 4, 6, 8 etc. are easier to evolve than 3, 5 and 7. This is an important result, since it indicates that selection and evolution build up more complicated expression patterns from building blocks of simpler patterns. A gene expression pattern with an attractor length of 4 may be the same as another with an attractor length of 2, except for the expression pattern of a particular gene which gives the pattern a length of 4. It seems possible that real biological evolution could work in a similar way, by evolving complex and intricate expression patterns from systems which have simpler dynamical behaviours.

These principles can be visualized using an expression graph or "piano roll" representation, which plots the expression of a set of genes against time, in order to show the associated expression pattern. Fig. 11 shows an example expression graph for a genome of length 20,000, where the system falls into a cyclic attractor state. Although the resulting attractor state has a length of 12, many genes exhibit expression patterns of length 6. The existence of expression patterns whose length is an exact sub-division of the attractor length is regularly observed in cyclic attractor states. The expression graph also illustrates how some genes typically exhibit static expression patterns, even when a cyclic attractor state is reached. In this case, the genes may or may not be explicitly involved in the regulation cycle which produces the cyclic expression pattern.

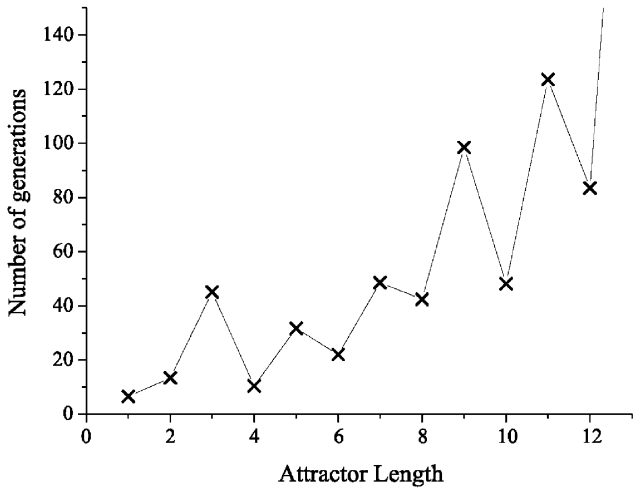

Fig. 10. Plot of the mean number of generations to reach a population fitness of 1.0, for target attractor lengths between 1 and 15, each averaged over 20 runs.

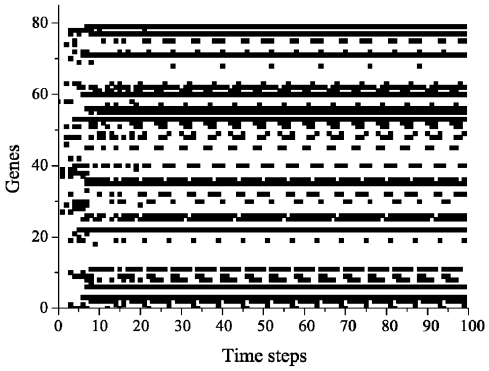

Fig. 11. An example expression graph where the system falls into a limit cycle attractor of length 12.

In addition to the influence of evolution on a population of AG networks, the genetic encoding in the AG model constrains the resulting network structures. As a direct result of these constraints, some attractor lengths are produced with greater frequency than others in the initial population of AG networks. Fig. 7 in Section 3.1 shows that the AG model is more likely to exhibit attractor lengths of 2, 4, 6, 8, etc., from randomly generated genomes. An analysis of the associated expression patterns reveals the presence of expression patterns with exact sub-division lengths of the attractor, as also observed in the expression patterns of evolved AG networks.

3.2.3. *Patterns of gene expression*

The next step is to select for particular patterns of gene expression, rather than simple periodicity. If a target gene expression pattern for a single gene is specified, this inherently specifies a target attractor length as before, but also the required expression pattern for a particular gene. For example, the target pattern of "11" requires that the specified gene is expressed at each time step, and also that the network as a whole exhibits a cyclic gene expression pattern with period 2 (see Section 2 for a more detailed explanation of notation). Simulations were run where the target expression pattern was incrementally increased in complexity, starting with the simplest

Table 2
Number of generations to reach maximum population fitness, averaged over 20 runs, for each distinct gene expression pattern for small attractor lengths

Pattern	Mean generations	Median generations
0	6	5
1	10	7
00	19	13
01	143	50
11	67	32
000	98	56
001	267	130
011	452	396
111	195	114

possible expression patterns, with an attractor length of one. As for the simulations in which attractor lengths were selected for, the number of generations for the population to reach a fitness of 1.0 was used to measure the evolvability of each expression pattern. The results of these simulations are summarized in Table 2.

These results clearly show that selection can successfully evolve simple target expression patterns, since in each simulation presented in Table 2, the population eventually reached a fitness of 1.0. However, the results also show that as the complexity of the target expression pattern is incrementally increased, the associated evolvability is significantly reduced, since the number of generations required to reach maximum fitness increases. In general, the evolvability is decreased as the attractor length is increased, in line with results presented in Fig. 10, but there is also variation in the evolvability for different expression patterns with the same attractor length. The mean and median were calculated in each case, and both averages follow the same qualitative variation, but the median number of generations is typically less than the mean. This reflects how a few simulations take much longer than the mean, due to trapping in local maxima, and hence the median may be considered a more appropriate measure of the average in this case.

Expression patterns where the gene is off for all times steps in the attractor are the easiest to evolve, since in this case the gene is effectively not involved in the expression pattern produced by the network. The gene must still be present to be labelled as "off", and hence this type of expression pattern cannot be produced by the mutation of a gene during evolution, which removes the gene. The next easiest pattern to evolve features a gene that is on for all time steps, which is similar to a static attractor, except that the interactions of other genes determine the attractor length. Expression patterns where the gene must be turned on and off in a specified pattern are the most difficult to evolve.

Table 2 shows that the evolvability of systems decreases rapidly as the attractor length of the target expression pattern is increased. For example, an expression pattern of 1 requires on average 10 generations to reach maximum population fitness, a pattern of 01 requires 143 generations, and a pattern of 001 requires 267 generations. The same problem was also considered where evolvability is measured instead by the number of generations to obtain a single individual of fitness 1.0, which gives a similar qualitative variation of evolvability, but over a much narrower range. Simulations using fitness functions based on population fitness provide a more sensitive measure for the evolvability of simple target systems, but are less suitable for the evolution of complex target systems. Maximum population fitness is only satisfied if every individual in the population has maximum fitness, which is artificially restrictive and unrealistic for the evolution of complex target systems. For this reason we use evolvability measures based on individual fitness when studying the evolution of complex target systems below.

As the target complexity is increased, the inverse correlation between the target complexity and associated evolvability continues, which suggests that at some point, as the ruggedness of the fitness landscapes increases, the system becomes unable to reach the required global maximum. This hypothesis was investigated by continually incrementing the length and complexity of target single gene expression patterns. The most complex targets to evolve for a given length are those where the gene is required to switch on and off frequently and at non-regular intervals.

It was found that some simulations towards target expression patterns of length 9 were unable to reach the global maximum (where an individual has fitness 1.0) after as many as 10,000 generations. Due to the random nature of mutations and the generation of initial populations in the simulations, some simulations reach the required global maximum while other simulations do not. This can be visualized in terms of two simulations taking different paths across the associated fitness landscape, where one path reaches the global maximum, and another gets trapped in some localized area of the landscape. Hence, this "loss of power" does not occur at a particular point, but rather over a range of target expression patterns.

Table 3 shows the fraction of simulations (out of 20 runs in each case) which reached the global maximum, for the least evolvable, or most complex target expression pattern of a given length. This table also presents results for simulations featuring neutrality in the fitness function which are described in Section 3.2.4. The fraction of simulations evolved to the global maximum decreases rapidly as the pattern length increases over the range 8–16. Again, this decrease in evolvability is not linear, as some pattern lengths are easier to evolve than

Table 3
Fraction of simulations generating an individual with fitness 1.0 over 20 runs, for the most complex target expression pattern of a given length, comparing simulations with and without neutrality

Pattern length	Fraction evolved (without neutrality)	Fraction evolved (with neutrality)
8	1.00	1.00
9	0.60	0.85
10	0.65	0.85
11	0.25	0.60
12	0.35	0.65
13	0.10	0.40
14	0.05	0.45
15	0.10	0.50
16	0.05	0.40
17	0.00	0.15
18	0.05	0.20

others, as previously demonstrated in Section 3.2.2. In fact, no simulations evolved to the most complex expression pattern of length 17. This pattern length is intrinsically difficult to evolve since 17 is a prime number, and hence there are no other pattern lengths which are exact sub-divisions of this length.

3.2.4. Cumulative patterns

The results presented in Table 3 related to the most complex expression pattern of a given length, but there is a wide range in the evolvability of different patterns of the same length, as demonstrated in Table 2. However, it may sometimes be the case that the exact timing of a particular gene's expression is less critical than the amount of time that the gene spends expressing some product during one cycle of gene expression. A cell cycle produces certain amounts of different proteins which interact on various levels, and so it may be that the quantities of the proteins are more important than the detailed dynamics describing how these proteins were produced. This can be interpreted in the model as the number of time steps in an attractor for which each gene is expressed.

From this perspective injecting a certain type of "neutrality" into the selection criterion is justified, since multiple attractor states may give rise to the same levels of gene activity, if not the same detailed pattern of expression. For example, if we require a given gene to be on for two time steps in a cycle of five time steps, both the attractor states 00011 and 00101 satisfy this condition. Introducing this type of redundancy could enable selection to evolve more complex expression patterns. This type of neutrality was included into the model, and the resulting simulations are also presented in Table 3, in order to compare directly with those without neutrality. As for the results without neutrality, data is presented for the most complex or most difficult to evolve target for a given pattern length.

These results show that the introduction of neutrality into the model has no discernable effect on the target complexity which can be consistently fully evolved. An increased neutrality means that an individual simulation has an increased likelihood of reaching the global maximum, but the complexity of the problem is still effectively the same. Table 3 shows that the difference between the simulations with or without neutrality is that on average, a greater fraction of simulations are fully evolved with increased neutrality, reflecting the increased likelihood discussed.

Simulations were also run for more complex target expression patterns involving more than one gene, in order to compare the evolution of single gene target expression patterns and multiple gene expression patterns. The simulations and associated fitness plots display similar behaviour to those for single gene expression patterns, where the system is unable to evolve to the global maximum. The only significant difference is that the maximum population fitness reached is typically lower for simulations evolving towards multiple gene expression patterns. This reflects the increased complexity of the target system when multiple gene expression patterns are specified.

In addition to investigating the ability of selection to evolve particular target dynamics, the structures of the evolved AG networks were characterized. The structures of random AG networks (produced from random genomes) were investigated in detail in Section 3.1, and hence the structure of evolved AG networks can be compared directly to that of random AG networks.

Network parameters were extracted from a range of selected simulations in which target patterns up to length 7 were selected for, and were averaged over 20 runs for each selected simulation. The network structures can be characterized on a number of levels, as in Section 3.1, by investigating parameters such as the number of genes, network degree, standard deviation of the network degree and the degree distribution. Averaged over all selected simulations, it was found that the number of genes in the evolved AG networks was 36.80, compared to 37.74 in random AG networks. Since it is much easier to remove genes than to create genes by random mutations (see Section 3.2.1), it may be expected that the number of genes in the networks would decrease after evolution (unless a highly complex target is being evolved). However, this result shows no significant decrease in the number of genes in the networks.

The degree distribution of the evolved AG networks was plotted, as a direct comparison to that of random AG networks, shown in Fig. 3. The degree distribution was virtually unchanged, again following a Normal distribution, and the evolved AG networks had a mean degree of 2.19, standard deviation of 0.27, and a total range of 1.36. This is in comparison to random AG

networks with a mean degree of 2.24, standard deviation of 0.24 and a total range of 1.52.

The same equality of network parameters was also observed for the variance (or standard deviation) of the gene degree within networks. It was shown in Section 3.1 that the standard deviation of the gene degree, σ_k, in random AG networks is 2.24. After evolution, the networks were found to also have $\sigma_k = 2.24$. The distribution of σ_k can also be calculated, and in both cases the distribution approximates a Normal distribution, with a standard deviation in the distribution of 0.4 for random AG networks, and 0.36 for evolved AG networks.

The extent of the similarity between the network structures after evolution and in the initial population is due to the fact that a number of different network structures can satisfy any particular target expression pattern. This reflects the inherent neutrality in the problem and the associated fitness landscapes. For a given target expression pattern, a range of values of the average degree and standard deviation of the gene degree is found in the final population, over all runs of the program. However, for any given run, the individuals in a population evolve towards the same average degree, and the same network structure, when the population reaches maximum fitness. Hence in the final population for a given run which reaches maximum population fitness, the standard deviation of the distribution of network degrees is zero, when all individuals have reached maximum fitness, and in fact all individuals are found to share the same network topology. So although a number of different network structures typically satisfy a particular expression pattern, the structures in a fully evolved population are found to be identical. This can be visualized in terms of a fitness landscape for the network structure, with a number of maximums, and the population evolves towards one of these maximums.

The degree distribution in evolved AG networks was also investigated, and the characteristics of the distribution were found to be similar to those for random AG networks. A log–log plot of the distribution for a typical network gives a similar profile to that for random AG networks, as illustrated in Fig. 6. It may have been expected that the networks would evolve to become more scale-free in nature, since biological networks in simple organisms such as *S. cerevisiae* and *C. elegans* have been shown to have a scale-free structure (Li et al., 2004). However, these simulations typically only evolved over a few hundred or thousand generations, and the evolution towards a scale-free structure may require many more generations, and possibly a more complex target.

The results presented indicate some measure of the ruggedness of the fitness landscapes which underlie these simulations. As the required targets become more complex, the associated fitness landscapes become more rugged, with many more local maxima and variations, and also become much larger. Kauffman predicted that selection has insufficient power to evolve towards a general target gene expression pattern, and hence that self organization must play a significant role in the inherent order in complex biological structures (Kauffman, 1993). We have shown that selection loses its power to evolve exact gene expression patterns as the pattern complexity is increased, where the AG model is used to model the transition from the genome level to dynamic gene expression. Within this model, selection is able to evolve relatively simple gene expression patterns.

We have also demonstrated some of the inherent biases and constraints of the AG model, which influence the resulting genetic network structures prior to the influence of selection, and also influence the evolvability of certain network types and expression patterns. Since the template-matching scheme of the AG model is designed to represent aspects of genetic encoding, it may be that similar biases and constraints are present in real genetic systems, as Kauffman predicted in terms of self-organization. However, the AG model is a highly streamlined and simplistic model of real genetic systems, and more work is required to determine if the biases and constraints observed in this model correspond to inherent properties of biological systems.

4. Summary and conclusion

The evolution of genetic regulatory networks has been investigated using a combination of several simulation techniques. The genetic networks and their gene expression dynamics were modelled using the AG model, and the behaviour of this model was investigated in detail. The expected network structures and the variability of network size and network degree were characterized. The model was shown to produce a wide range of gene network topologies, which typically lie between those of random networks and scale-free networks in terms of their degree distributions.

Simulations of network evolution were performed using genetic algorithms to model selection on a number of properties ranging from network size and degree to specific patterns of gene expression. The ability of selection to achieve increasingly complex target behaviours was assessed. These results showed that while simple patterns of gene expression were achievable, more complex behaviours were not, even with the injection of additional redundancy into the selection criteria.

It has been shown that attractor lengths which are multiples of other possible attractor lengths are inherently easier to evolve than others. Similar biases towards

certain attractor lengths are also observed in initial populations, reflecting the genetic encoding constraints introduced in the AG model. The existence of expression patterns whose length is an exact sub-division of the corresponding attractor length has been demonstrated, and is common to cyclic attractors of these lengths. These results demonstrate the influence of the inherent constraints and biases present in the AG model on the associated evolvability of the networks.

Acknowledgements

We thank the Medical Research Council (MRC) for funding.

Appendix A. Theory on artificial genome

A derivation of expressions for the calculation of the expected number of genes present in a random genome and the associated network degree is presented below. Note: this derivation takes into account the length of the promoter sequence and of each gene, such that if a second promoter sequence is contained between the start point of the original promoter and the end point of the corresponding gene, it is not considered as a promoter sequence. This is illustrated in Fig. 12.

For a given randomly generated genome, let l be the length of the genome sequence, p the length of the promoter sequence, g the length of each gene sequence, and B the number of possible bases at each locus. For example, in the model used in this study, the allowed bases are 0, 1, 2 and 3, and hence $B = 4$. The chance of a promoter sequence starting at the ith base in the genome is then given by $(1/B)^p$. Hence an estimate for the expected number of genes, μ_G in a randomly generated genome is given by

$$\mu_G = l(1/B)^p. \tag{7}$$

However, this does not account for the length of the promoter and gene sequence, as mentioned above. If a promoter sequence starts at the ith position in a genome, then the $(i+1)$th to the $(i+p+g-1)$th positions should not be considered as potential promoter sequence start sites. A modified expression to account for this is given by Eq. (8).

$$\mu_G = (l - \mu_G(g+p-1))(1/B)^p. \tag{8}$$

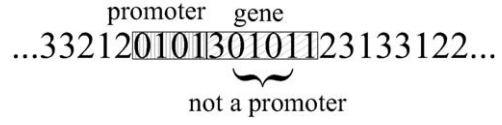

Fig. 12. Accounting for the promoter and gene length.

The mean number of genes can then be found by rearranging Eq. (8), to give the expression below.

$$\mu_G = \frac{l}{(B^p + g + p - 1)}. \tag{9}$$

If many random genomes are generated, then the number of genes in each genome would be expected to vary according to a Poisson distribution. This distribution is appropriate for discrete random variables following a Poisson process, where events occur independently and at random. This is true in this case, since the chance of a promoter sequence starting at the ith position is random, and independent of the chance of a promoter sequence starting at the jth position in the same genome, ignoring the promoter and gene lengths, which have been accounted for.

For a discrete random variable X with a Poisson distribution, the probability that $X = k$ is given by

$$p_x(k) = P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \dots, \tag{10}$$

where λ is a positive constant, and is the mean value of X (Larsen and Marx, 2001). The variance of X is also λ, and hence for this distribution the mean is equal to the variance. In this way, the distribution of the number of genes from randomly generated genomes can be compared with the modelled distribution, given by Eqs. (9) and (10).

An expression for the average degree of networks produced from randomly generated genomes can be derived, based on the expected number of genes. The model used allows a given transcription factor sequence to bind anywhere, which then interacts with the next promoter and gene downstream from the binding site. If two or more binding sites for a given sequence are present upstream of the promoter, the effect is the same as one binding site, since we are considering a Boolean model where a gene is considered to be only on or off. Hence, the gene degree is effectively only one, even if there is more than one binding site present upstream of the promoter.

The expected distance between genes, d, (between the start of promoter sequences) can be calculated from the genome length and the expected number of genes, as given by Eq. (11).

$$d = \frac{l}{\mu_G} = B^p + p + g - 1. \tag{11}$$

The chance of a matching binding site for a given transcription factor being present at the ith position in the non-coding region d is given by $(1/B)^g$. Hence across the distance d, the expected number of binding sites for a given transcription factor, μ_S, is given by

$$\mu_S = \frac{B^p + p + g - 1}{B^g}. \tag{12}$$

The number of binding sites in this region follows the Poisson distribution for the same reasons as described above for the number of genes. To calculate the expected network degree, an expression for the probability of one or more binding sites being present in the distance d can obtained from Eqs. (11) and (12). If the number of binding sites is given by the random variable S, then

$$p(S \geqslant 1) = 1 - p(S = 0) = 1 - \exp\left(-\frac{B^p + p + g - 1}{B^g}\right). \tag{13}$$

The expected degree, μ_K is then given by this probability multiplied by the number of genes.

$$\mu_K = \frac{l}{(B^p + g + p - 1)}\left(1 - exp\left(-\frac{B^p + p + g - 1}{B^g}\right)\right). \tag{14}$$

The mean number of genes and the mean degree of the networks are important parameters to describe the network structures produced by the model. However, these parameters do not provide any measure of the variation of structure throughout the network, and only measure global network properties. Details of the network structures produced by the AG model have so far not been investigated, and may be important in understanding their evolution. Since the model uses an underlying genome and template-matching system to produce network structures, the resulting networks are expected to differ from purely randomly generated networks. Also, an initial population of network structures in an evolutionary model may contain a surprisingly high degree of variation, and may account for much of the variation in a population for selection to act on, without the effects of crossover and mutation.

In order to quantify the network structures produced by the model in some way, a useful parameter to measure is the variance of the gene degree. This measurement effectively describes whether most genes are regulated by the same (mean) number of genes, or whether some genes are regulated by many genes, and others are regulated by none, or relatively few. This variance depends on the variance of the distance between genes in the random genomes, since regulatory connections are determined by the presence of binding sites on the genome. More regulatory inputs are expected for a gene where the distance upstream to the next gene is large, since in this case there are more possible sites for transcription factors to bind to, as illustrated in Fig. 13.

Network structures produced from genomes similar to Genome 2 in Fig. 13 will, on average, have much greater structural variation than the random networks produced from genomes similar to Genome 1. The variance of the distance between genes in genomes

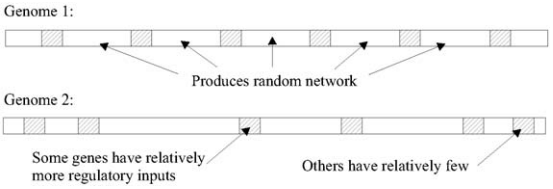

Fig. 13. The effect of variation in the distance between genes.

produced from the AG model can be calculated if the statistical distribution of the distance between genes is known. This distance variable, d, can be shown to follow a geometic distribution, which considers the success or failure of a series of trials. In a geometric distribution, the random variable X represents the number of trials needed to obtain the first success (Milton and Arnold, 1995). The distance between genes, can be thought of as the number of trials, moving along the genome, to obtain a success of finding a promoter sequence, and hence the presence of a gene. The trials are identical and independent, and the probability of success, P, is a constant. The geometric distribution is defined by Eq. (15) for the random variable X, by its density function, $f(x)$;

$$f(x) = (1 - P)^{x-1}P, \tag{15}$$

where $x = 1, 2, 3, \ldots$ and $0 < P < 1$. It can be shown, using moment generating functions, that the mean and variance of the geometric distribution are given by

$$E(X) = 1/P,$$

$$Var(X) = Q/P^2, \tag{16}$$

where $Q = 1 - P$. In the AG model, the probability of success of finding a promoter sequence is $1/B^p$, and hence, according to the geometric distribution, the mean distance between genes is given by

$$d = B^p. \tag{17}$$

However, genes are not allowed to overlap and hence the length of the gene should be taken into account as before, such that the mean distance between genes is given by

$$d = B^p + p + g - 1. \tag{18}$$

This agrees with the previous derivation of d from the expression for the mean number of genes in a random genome. The expression for the variance can be approximated and simplified, since the chance of success is very small, and hence $Q \sim 1$. This means that the standard deviation of the distance between genes, σ_d, can be expressed as

$$\sigma_d = 1/P$$
$$= B^p. \tag{19}$$

No correction for the gene length is required for the variance or standard deviation, since this is a fixed correction to each distance, and hence the variance is unaffected. Given expressions for the mean and standard deviation of the distance between genes, information about the variation of network structure can be derived by considering the distribution of the number of binding sites and regulatory inputs per gene.

For a given gene, if the distance upstream to the next gene is known to be d, then the expected or mean degree for that gene, μ_k, is given by

$$\mu_k = \frac{Gd}{B^g}, \tag{20}$$

where G is the number of genes in that particular genome, and $1/B^g$ is the probability of finding a particular binding sequence at each upstream position. This expression describes the mean gene degree and not the exact gene degree, since the element of probability in finding binding sites introduces a variance. This illustrates why it is in fact not possible to produce an exact expression for the variance of gene degree, since it is due to the combination of variance from the binding probability, and the variance in the distance between genes, which are independent. However, some analysis of the extent of variance is possible. In Eq. (20), the number of genes, G, is a constant since we are considering the type of network structures produced and the variance of gene degree within networks, and not the variance of the network degree. However, the variance of the network degree can also be considered, and is important in relation to the variation of structures in populations. This is covered in Section 3.1, and relevant results are illustrated in Fig. 3.

A useful approach to this analysis is to consider the variance of the binding and the variance in the gene distance individually, and assess which factor contributes more to the overall variance in gene degree. If the variance in binding is neglected, then Eq. (20) can be rewritten as

$$k = \frac{Gd}{B^g}. \tag{21}$$

For a general network where the number of genes is μ_G, the variance of the gene degree can now be related exactly to the variance of the gene distance, as follows:

$$Var(aX) = a^2 Var(X) \Rightarrow$$

$$\sigma_k = \frac{\mu_G}{B^g}\sigma_d \tag{22}$$

$$= \frac{\mu_G B^p}{B^g}. \tag{23}$$

This can be related back to the mean degree, μ_K, since from Eq. (21),

$$\mu_k = \frac{\mu_G \mu_d}{B^g}. \tag{24}$$

If the correction for the gene length is neglected, the mean gene distance can be approximated as $\mu_d \sim \sigma_d$, and hence combining Eqs. (22) and (24) gives,

$$\sigma_k \sim \mu_k. \tag{25}$$

But the mean gene degree is equal to the mean degree of the network, i.e. $\mu_k = \mu_K$, and hence we can write,

$$\sigma_k \sim \mu_K. \tag{26}$$

This gives a measure of the variance of the gene degree based on the variance of the gene distance. A measure of the variance of the gene degree due to the binding probability can be obtained from considering the variance of gene degree in random networks. As described in Fig. 13, random networks would be produced from genomes with a fixed gene distance. Hence, the variance of the gene degree in random networks approximates the variance of the gene degree in AG networks due to the binding probability, if the variance in gene distance is removed.

In a random network, the number of input connections for each gene follows a binomial distribution, where each trial corresponds to each gene in the network, and a success is the presence of an input connection. The trials are identical and independent, and the variable is the number of successes in n trials, where n is the number of vertices. This underlying binomial distribution can be approximated as a Poisson distribution, since in general, n is large, and P, the probability of success is small.

As mentioned previously, the variance of a Poisson distribution is equal to the mean, and hence for randomly generated networks with large n and small P, the variance of the vertex degree is equal to the mean vertex degree, which is the mean network degree. Hence for random networks,

$$\sigma_k = \sqrt{\mu_K}. \tag{27}$$

Eqs. (24) and (27) indicate the different influences of the two types of variance on the variance of the gene degree, and indicate clearly that the variance due to the gene distance dominates the variance due to the binding probability. Hence a best estimate for the variance of the gene degree within AG networks is given by $\sigma_k \sim \mu_K$. A comparison is now possible between the network structures produced by the AG model and random networks. In summary,

For random networks, $\sigma_k = \sqrt{\mu_K}$, and for AG networks, $\sigma_k \sim \mu_K$.

This is an important result, since it clearly shows the difference between the two types of networks structures, which may well be significant in the model, in terms of the initial population of networks in the evolutionary model and the tendency for random mutations to bring about particular kinds of network topology during evolution.

References

Bagley, R.J., Glass, L., 1996. Counting and classifying attractors in high dimensional dynamical systems. J. Theor. Biol. 183, 269–284.

Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Bornholdt, S., Rohlf, T., 2000. Topological evolution of dynamical networks: Global criticality from local dynamics. Phys. Rev. Lett. 84, 6114–6117.

Cherry, J.L., Adler, F.R., 2000. How to make a biological switch. J. Theor. Biol. 203, 117–133.

Erdös, P., Rényi, A., 1960. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. 5, 17–67.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. J. Comput. Biol. 7, 601–620.

Gibson, G., Honeycutt, E., 2002. The evolution of developmental regulatory pathways. Curr. Opin. Genet. Develop. 12, 695–700.

Gillespie, D.T., 2000. The chemical langevin equation. J. Chem. Phys. 113 (1), 297–306.

Harrington, C., Rosenow, C., Retief, J., 2000. Monitoring gene expression using DNA microarrays. Curr. Opin. Microbiol. 3 (3), 285–291.

Kauffman, S., 1974. The large scale structure and dynamics of gene control circuits: An ensemble approach. J. Theor. Biol. 44, 167–190.

Kauffman, S., 1993. The Origins of Order, first ed. Oxford University Press, Oxford, UK.

Keller, A., 1995. Model genetic circuits encoding autoregulatory transcription factors. J. Theor. Biol. 172, 169–185.

Larsen, R.J., Marx, M.L., 2001. An Introduction to Mathematical Statistics and its Applications, third ed. Prentice-Hall, New Jersey, US.

Li, S., et al., 2004. A map of the interactome network of the metazoan *C. elegans*. Science 303, 540–543.

Madden, S., Wang, C., Landes, G., 2000. Serial analysis of gene expression: from gene discovery to target identification. Drug Discov. Today 5 (9), 415–425.

McAdams, H.H., Arkin, A., 1998. Simulation of prokaryotic genetic circuits. Annu. Rev. Biophys. Biomol. Struct. 27, 199–224.

Milton, J.S., Arnold, J.C., 1995. Introduction to Probability and Statistics, third ed. McGraw-Hill, New York.

Reil, T., 1999. Dynamics of gene expression in an artificial genome—implications for biological and artificial ontogeny. Advances in Artificial Life, Proceedings, Lecture Notes in Artificial Intelligence, vol. 1674, pp. 457–466.

Somogyi, R., Sniegoski, C.A., 1996. Modelling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. Complexity 1 (6), 45–63.

Thomas, R., 1991. Regulatory networks seen as asynchronous automata: a logical description. J. Theor. Biol. 153, 1–23.

Tong, W.C., 2002. Artificial genome—a model on genetic regulation network, the asynchronous version. MRes, University of Leeds.

Weber, G. (Ed.), 1965. Oscillatory Behaviour in Enzymatic Control Processes. Pergamon Press, Oxford, pp. 425–438.

Wilkins, A.S., 2002. The Evolution of Developmental Pathways. Sinauer Associates, Sunderland, MA.

Wuensche, A., 1998. Genomic regulation modelled as a network with basins of attraction. Pacific Symp. Biocomput. 3, 89–102.