Effects of Signalling on the Evolution of Gene Regulatory Networks

Dafyd J. Jenkins and Dov J. Stekel

Centre for Systems Biology, School of Biosciences, University of Birmingham, U.K., B15 2TT
djj134@bham.ac.uk d.j.stekel@bham.ac.uk

Abstract

We investigate whether observed transcription network structures and network motifs are a byproduct of the mechanisms by which DNA strands evolve, or if they are fundamental to the function of the network. We explore this with an evolutionary model with stochastic Boolean network simulation. Structurally distinct regulation strategies are observed in populations evolved with and without internal energy signalling. However, food signalling is not used in either population in the case when the food supply itself is constant. Parallels between the evolved networks and CRP-cAMP regulation in *Escherichia coli* and the endosymbiont *Buchnera aphidcola* are presented and discussed. Comparing the evolved networks with neutrally evolved populations indicates that networks evolve to lose most regulatory activity, due to loss of binding sites and transcription factor activity, including losing global regulation mechanisms.

Introduction

Transcription regulation and cell-signalling networks have been studied extensively; recently high-throughput 'omics' technologies have provided a wealth of data, including genome sequences, gene-expression, metabolism and protein-protein interaction profiles. Systems biology attempts to use these data to develop models for reconstructing and analysing transcription, signalling and other networks. Local analysis determines the function of over-abundant network motifs (Milo et al. (2002)), such as the 'feed-forward loop' (FFL), which can function as a low-pass filter (Mangan and Alon (2003)). Network motifs are a valuable tool for analysing transcription regulation networks, but do not always indicate dynamical behaviour: the bi-fan motif exhibits wide ranging and non-characteristic behaviour when modelled using biologically plausible parameters (Ingram et al. (2006)) and the process of DNA replication can also cause the over-abundance of motifs (van Noort et al. (2004)). Global analysis, such as node degree, of entire transcriptional networks has indicated an approximately scale-free out-degree distribution and an exponential in-degree distribution in both prokaryotic and eukaryotic organisms (Albert (2005)).

The use of energy signals in biological regulatory networks is well studied. The transcriptional regulator complex CRP-cAMP is one of *Escherichia coli*'s global regulators, known to regulate several hundred genes as listed in the EcoCyc database (Karp et al. (2007)). The large number of positive interactions by CRP-cAMP in biosynthesis pathways indicates that energy signals are used for growth by cells (Zheng et al. (2004), Hardiman et al. (2007)). A subunit of the CRP-cAMP complex, cAMP, is a signalling molecule derived from ATP; ATP concentration indicates 'energy' within the cell. When the concentrations of CRP and cAMP reach sufficient levels, the activated transcription factor complex forms. Whilst CRP-cAMP is a dual-regulator (activation and repression), 142 of the 173 known and predicted interactions in the EcoCyc database are identified as activating interactions.

Organisms without energy signalling are also prevalent in nature. *Buchnera aphidcola* is a bacterium related to *E. coli*, having a common ancestor diverging 250 million years ago (Moran and Mira (2001), Shigenobu et al. (2000)). *B. aphidcola* has a different lifestyle to *E. coli*; it has evolved an endosymbiotic relation with aphids, while *E. coli* exists as a free-living bacterium. *B. aphidcola* cells live in an environment of sufficient food, which is simpler than many other bacterial environments. *B. aphidcola* strains have lost most of their genome and regulatory network, retaining around 600 genes, representing a subset of *E. coli* genomes (Shigenobu et al. (2000), Wilcox et al. (2003)). This lack of regulation allows the over production of several amino acids, which are excreted and subsequently used by the aphid. The lack of an 'energy signal' observed in *B. aphidcola* is due to the absence of *crp* and *cycA*, the genes responsible for the CRP-cAMP transcription factor (Shigenobu et al. (2000)).

Many computational models exist for evolving transcription network structure, such as the Artificial Genome (Quayle and Bullock (2006)) and Artificial Regulatory Network (Kuo et al. (2006)), which are capable of evolving very realistic structure. However, the behaviours evolved are often arbitrary and non-realistic, such as matching a specific pattern of expression. Models also typically omit en-

ergy usage, which is a fundamental requirement for transcription regulation. Stochasticity, whilst has been shown to have substantial effects on gene expression in biological cells (Elowitz et al. (2002)), is also often omitted from models of gene regulation networks.

We investigate the effects of dynamics in the evolution of transcription network structure using models with and without energy signalling. We introduce a model that evolves networks using realistic evolutionary operators and is simulated with simple inputs and output to determine fitness. Our model introduces regulation type for binding sites, new evolutionary operators, signalling mechanisms as inputs and biosynthesis as output. We simulate the networks using a stochastic Boolean network paradigm, representing simplified transcriptional network dynamics. The results of these evolutions are presented and analysed, and relevance to biological systems is discussed. Graph theoretic approaches are used to compare the directed evolution and networks that have evolved neutrally over the same time period, highlighting the effects of the directed evolution.

The exploratory results presented in this paper highlight the potential insights into evolutionary behaviour that can be obtained using simple, yet biologically realistic models.

Method

The model has two distinct components: 1) network generation and static structure and 2) network simulation, dynamics and evolution.

Network Generation

To generate the gene regulatory network, we use the model introduced by van Noort et al. (2004) and extended by Cordero and Hogeweg (2006). This model produces a network with realistic connectivity and structure of specific protein-DNA binding interactions when evolved without a fitness function, "neutral evolution". A genome initially consists of N regulatory genes, where each gene has a regulatory region with between 0 and I binding sites, bs, and a protein, p. Each binding site and protein has a specific shape, S, represented by an integer drawn from a discrete circular space $\{0, 1, 2, ..., S_{max} - 1\}$ (with $S_{max} - 1$ adjacent to 0). The binding strength, B_{ij}, between two shapes, S_i and S_j is defined as:

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

where D_{ij} is the shortest integer distance between the shape of the protein, S_i, and the binding site, S_j. A binding distance, D_{max}, is defined as the maximum distance between two shapes that will interact. A matrix, M, is created where M_{ij} is the strength of binding B between protein i and binding site j. From this matrix, the network connectivity can be visualised and analysed. The binding strength, B, between a protein and binding site is used during network simulation.

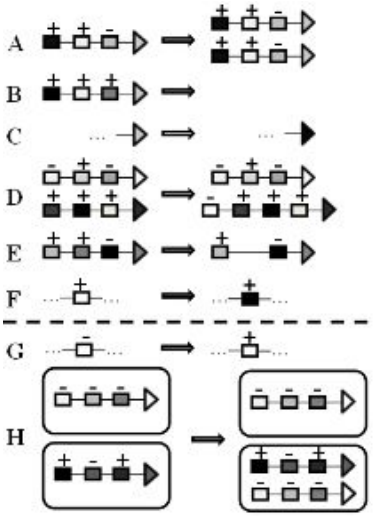

Figure 1: Evolutionary operators. Rectangles represent binding sites (+ activating; - repressing), triangles represent gene/protein product. Shape is represented by greyscale colour. Original operators defined in Cordero and Hogeweg (2006) are shown in parts A - F. A) gene duplication, B) gene loss, C) protein mutation, D) binding site duplication, E) binding site loss and F) binding site mutation. The two evolutionary operators introduced in this work, G) binding site regulation 'flip' and H) horizontal gene transfer. Operators A - G apply to each gene or regulatory region within a genome, whereas H only applies to the whole genome level.

In addition, our model introduces two types of regulation for each binding site: positive, bs^+; and negative, bs^-. Thus, as in real gene regulatory networks, binding sites can either increase the rate of a gene's transcription (positive regulation) or decrease transcription (negative regulation). Figure 2 shows an example network and interactions.

Specialised Genes In addition to the regulatory genes in the original models, we introduce three new types of genes:
Energy signal genes: these genes have a protein product, but no regulatory region. The expression status is based on the amount of energy within the cell. Energy in the model abstractly represents the ATP, amino acids and other molecules a biological cell requires to grow, transcribe mRNA molecules and translate them into protein molecules and other processes.
Food signal genes: these genes represent the food available to the cell and are used as the input into the model when it is simulated. they have a protein product, but no regulatory region. The energy level of the model increases whenever a *food signal gene* is activated. Each *food signal gene* has an energy value associated with it, which is the amount of energy added to the model when the gene is activated.
Biomass pathway genes: these genes have a regulatory re-

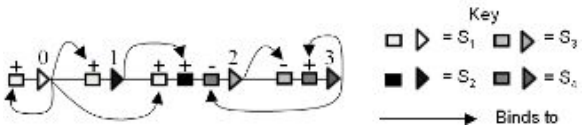

Figure 2: An example 4-gene network showing protein-DNA interactions. Genes 0,1 and 2 form a type-1 coherent feed-forward loop (FFL). Additionally, gene 0 has an activating self-regulating connection. A fourth gene in the circuit acts as an AND gate in the FFL, by negatively regulating gene 2. If gene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

gion, and generate *biomass* when expressed. They are used as the output of the model when it is simulated, represent cell growth, and have both an energy consumption (amount of energy used when gene is activated) and biomass production (amount of biomass added when activated) value associated with them.

Neutral Evolution and Evolutionary Operators

Once the network has been initialised, it is neutrally evolved for a given number of steps by randomly selecting a gene from the genome and applying a mutation operator. Cordero and Hogeweg define six mutation operators which operate at either the gene or binding site level: 1) gene duplication: the entire gene (protein product and regulatory region) is copied and added to the genome, producing an exact replica of the original gene, 2) gene loss: the entire gene is removed from the genome, 3) protein mutation: the protein shape is changed, 4) binding site duplication: a binding site from another gene is randomly copied into the regulatory region, 5) binding site loss: the binding site is removed from the regulatory region, 6) binding site mutation: the binding site shape is changed.

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

We define two new evolutionary operators: 7) binding site regulation 'flip': the binding site 'flips' its regulation type from positive to negative or vice versa, 8) horizontal gene transfer (HGT): a portion of another genome is horizontally transferred and copied into the genome (corresponding to DNA-uptake or plasmid transfer). This operator is applied at the genome level only.

Due to the specific function of the *energy*, *food* and *biomass* genes, not all evolutionary operators are applied to them. The evolutionary operators applied to *energy signal genes* and *food signal genes* are 1) gene duplication (however, the duplicated gene loses the specialised functionality as the original) and 3) protein mutation. The evolutionary operators applied to *biomass pathway genes* are 1) gene duplication (functionality not duplicated), 3) protein mutation, 4) binding site duplication, 5) binding site loss, 6) binding site mutation and 7) binding site regulation 'flip'.
All mutation rates (and other model parameters) are given in Table 1.

Network Simulation and Dynamics

In order to further investigate the structure of the networks evolved using realistic evolutionary operators, we introduce a simulation system for examining the dynamics of the networks. We use a Boolean network model (Kauffman (1969)) to simulate the dynamics of the network over a number of discrete time steps. Stochasticity is added to the simulation with random, basal levels of transcription. At each time step a number of steps takes place in order:

1. Energy signal gene status (ON if energy threshold is exceeded, OFF otherwise) and food signal gene status (ON if food available this time step, OFF otherwise) is determined.
2. Determine protein-DNA interactions for all ON genes.
3. Determine gene activation status.
4. Update energy and biomass levels.
5. All bound binding sites unbind (all binding sites are OFF).
6. Check model has energy remaining - if the energy level is ≤ 0 then the model 'dies' due to lack of energy, and simulation terminates.

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

Protein-DNA Interaction Protein-DNA interactions are determined by the following logic equation:

$$binding\_status_{ij} = (B_{ij} \times gene\_status_i) > R$$

where B_{ij} is the binding strength between protein i and binding site j, $gene\_status_i$ is the activated status of gene i (is the gene transcribing/translating) and R is a random number between 0 (inclusive) and 1 (exclusive).
The resultant matrix indicates binding site occupancy.

Gene Activation Status Gene activation status is determined by:

$$gene\_status_i = \begin{cases} 1 & if\ f(x) > 1 \\ & or\ f(x) = 0\ \&\ K_{basal} > R \\ 0 & otherwise \end{cases}$$

$$f(x) = \sum_{a=1}^{A} G_i bs_a^+ - \sum_{b=1}^{B} G_i bs_b^-$$

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

Table 1: Model and evolution parameters

where A is the number of occupied positive binding sites of gene G_i, B is the number of occupied negative binding sites of gene G_i and R is a random number between 0 (inclusive) and 1 (exclusive). Binding site occupation is determined by the $binding\_status$ matrix.

Molecular Production Costs Transcription and translation are not free processes: energy is used whenever they take place. When an activated gene's protein binds to a binding site, the energy value of the model is decreased by 1, representing the cost of transcribing and translating the transcription factor.

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

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

$$binding\_status'_{ij} = (B'_{ij} \times gene\_status_i)$$

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

[1] Values taken from Cordero and Hogeweg (2006)

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

Evolution Framework

The evolution framework used in the model is a standard genetic algorithm, with a fixed population size, and a purely elitist strategy that emulates the spatial constraints on a bacterial population, in, for example, a chemostat, where the fittest cells are ones that replicate fastest. A daughter cell is generated at each generation representing a simplified bacterial asexual replication. Due to the nature of DNA replication both the daughter parent cells are subjected to possible mutation. During replication, each gene in the genome can be affected by one of the evolutionary operators (#1-7). HGT (#8) is applied after genome replication and mutation. If HGT takes place, a donor genome from the population is selected at random, and a randomly selected number of genes are copied from the donor genome.

Fitness of an individual model is based solely on the level of biomass production after the defined number of time steps. If the simulation terminates due to lack of energy, the model has died and has a fitness of -1. In the neutrally evolved populations, the fitness function is a random number between 0 and 1, implying no selection pressure.

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

Results and Discussion

Model and Environment Regimes

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

With an energy signal gene and a small genome, a final population evolves with a very simple regulatory network (Table 2). The main component of this network is a strong positive regulation of one of the *biomass genes* from the *energy signal gene*, but also has some residual connectivity between regulatory genes. However, no regulation (positive or negative) due to the input food genes was evolved. This is to be expected, as the environment remains constant, and so provides no useful information to be exploited. This regulation network is a simple, but effective system; whenever the model has sufficient energy, the energy signal is present, and it strongly activates the biosynthesis pathway gene; when the energy drops below this level activation of the biosynthesis pathway ceases. Only one of the biomass genes is activated, so whilst the system may not be maximally efficient at generating biomass, the model is far less likely to overexpress genes, in particular the energy-expensive biomass genes, and so is far more likely to survive to the end of the

simulation. This network also allows a far more robust regulation of biosynthesis, as the energy signal gene is not affected by noise. The use of an energy signal for activating growth parallels many organisms such as *E. coli*, with its use of CRP-cAMP.

Regulator type		Population			
		E	**N**	**NI**	**R**
No ES	Activator	31.85	43.08	46.48	45.92
	Repressor	17.27	43.58	46.13	46.05
	Dual	1.73	3.02	2.37	1.77
ES	Activator	6.17	43.45	47.39	47.41
	Repressor	1.66	51.64	46.94	47.70
	Dual	0.05	3.48	2.32	1.75

Table 2: Mean distribution of connection type in different populations of both energy signal and no energy signal. E is evolved, N is neutral, I is initial and R random population

With no energy signal gene and a small genome, a very different regulation network is evolved. In this population, the most successful models again consisted of no regulation due to the input food genes, and so no input stimuli at all were available (as the energy threshold gene is regulated by the model itself, it can be classed as an input). Thus the models rely solely on stochasticity for transcription and translation of random genes. The model did however evolve some positive regulation from a small number of standard genes to the biomass genes (Table 2); this increases the probability that the biomass genes will be activated at a given time step, and so the efficiency of generating biomass. Whilst this network is not very efficient at generating biomass, or robust to noise due to the reliance on stochasticity, it is well adapted for survival. The lack of energy signalling used in the evolution of this population of models shares several parallels with the lack of signalling in *B. aphidcola* cells, and a similar, simple regulatory network in observed in both. The exploitation of stochastic gene expression seems to be a robust sub-optimal solution for survival without environmental information. This solution may provide a mechanism for survival in early gene regulatory networks, until more precise signalling networks evolve, or could itself be the basis for a signalling network.

With a much larger genome, with or without an energy signal, we observe very different results. Network connectivity is necessarily high because of the number of genes and small shape space. Models are unable to survive because they very quickly over-express many genes and use up all energy. Even under more energetically favourable conditions (energy from food = 40; starting energy = 4000) the models are still unable to survive. This indicates the importance of repressors within biological networks to tightly regulate the processes of transcription and translation, as are not 'free' (they require energy sources e.g. ATP). Other com-

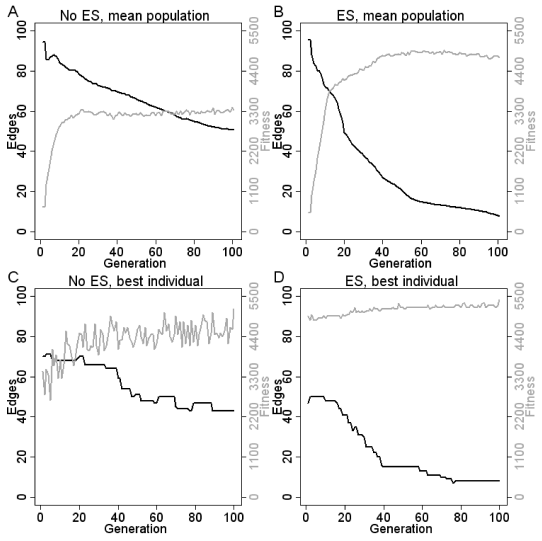

Figure 3: Evolutionary history of mean population fitness and number of regulatory connections in a small genome. A) mean fitness and number of regulatory connections in the population with no energy signal. B) mean fitness and number of regulatory connections in the population with an energy signal. C) best individual in the final population without an energy signal. D) best individual in the final population with an energy signal. The decrease in network connectivity and increased fitness can be seen in all plots.

putational models have obtained the evolution of repressor systems, even in constant environments (Jenkins and Stekel (2008)). The regulatory network of *E. coli* displays a preference for negative regulation by transcription factors in many different systems (Karp et al. (2007)). This may indicate a further use of negative regulation as an adaption for efficiency, as well as enabling large scale switching of regulatory systems, fast responses and maintaining homeostasis. Indeed, strong negative self-regulation has been shown to decrease the amount of mRNA needed to express a protein at a set level, thus reducing the use of energy expensive processes (Stekel and Jenkins (2008)). One possible explanation for the lack of large global repressors evolving in the current implementation of the model is the energy cost of maintaining sufficient numbers of repressor proteins. Protein stability is fixed to one timestep, so proteins must be produced each timestep, using up large amounts of energy. In biological systems, protein stabilities ranging from minutes to many hours are observed (Nath and Koch (1970)). The stability of a protein is often associated with function: signalling proteins are typically short-lived; metabolic proteins are often more stable. Modifying the model to allow proteins to evolve their stability may allow the evolution of global regulators. In addition, real biological molecules have a large shape space, due to the very high dimensionality of protein shape. Increasing the shape space in the model

could help alleviate the high network connectivity.

Effects of Stochasticity Removing stochasticity dramatically alters the networks evolved. Whilst a similar regulation mechanism is observed in populations with an energy signal, the number of connections does not rapidly decrease. Network connectivity remains high, with the exception of input genes. This occurs as the regulatory genes will only be transcribed if activated by another gene, leading to large parts of the network which are highly intra-connected with no external inputs. There is no pressure to reduce this connectivity, provided no input genes connect into the large highly connected parts. The high connectivity may appear to be a complex solution, however, the increased connectivity may merely mask the underlying core functionality of the model.

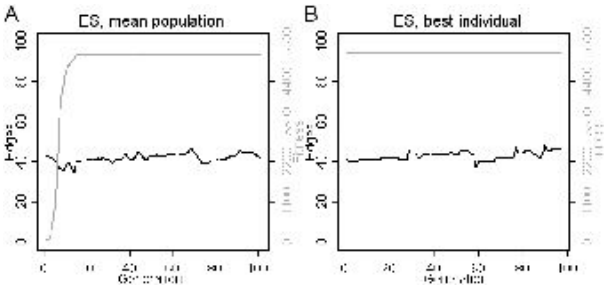

Figure 4: Evolutionary history of deterministically simulated small genome population with energy signal. A) mean fitness and number of regulatory connections in the population. B) best individual in the final generation of the population. Network connectivity remains high in both plots, unlike in the stochastic simulation populations.

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

Neutral Evolution and Comparison

To compare the effects of directed evolution, we evolved populations under the same four conditions but used a random fitness function to simulate 'neutral' evolution. We examined the model networks at three points during the evolution: 1) after random model initialisation (R), 2) after initial period of evolutions, creating a 'realistic' network (NI) and 3) after a given number of generations (N).

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

Binding Sites A general trend for loss of binding sites can be see in Figure 5. In the directed evolution populations, a larger number of genes in both populations have no binding sites, and have a much smaller distribution of maximum binding sites per gene. This shows how the model has evolved to optimise its regulatory network, by reducing it. There was no significant bias to regulation type in each population (Table 3), however, a clear trend for activating connections in the evolved populations is shown in Table 2. This may be linked to the lack of the evolution of global repressors as discussed above. Without a global regulatory mechanism, the model is unable to effectively regulate the expression of the genes, and so the alternative solution is to reduce the probability of transcription factor activity by losing binding sites. Whilst this solution does not prevent transcription, it does reduce it. In fact this mechanism is exploited in the populations without an energy signal.

	Binding site type	Population			
		E	N	NI	R
NoES	Activator	17.06	24.37	25.84	25.42
	Repressor	18.55	25.27	25.73	25.51
ES	Activator	12.38	24.70	25.76	25.54
	Repressor	13.69	26.56	25.67	25.64

Table 3: Mean distribution of binding site regulation type in different populations of both energy signal and no energy signal. E is evolved, N is neutral, I is initial and R is random population

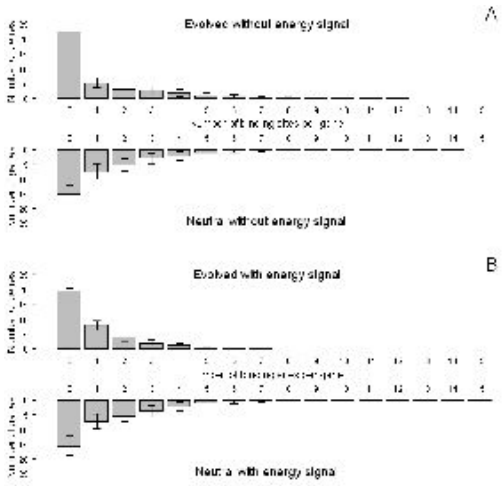

Figure 5: Mean number of binding site for each gene in each model in small genome populations. Error bars are 1 s.d. of the population. A) evolved and neutral populations without energy signal, B) evolved and neutral populations with energy signal. The loss of binding sites in the non-neutral populations can be seem in both panels; the evolved populations have a larger number of genes without any binding sites, and have a lower maximum number of binding sites per gene.

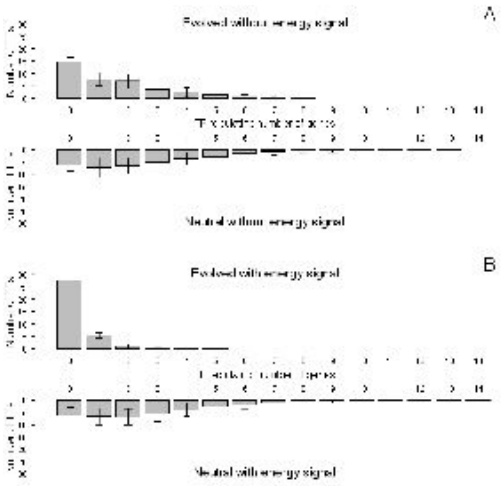

Figure 6: Mean gene 'out' degrees for each gene in each model in small genome population. Error bars are 1 s.d. of the population. A) evolved and neutral populations without energy signal, B) evolved and neutral populations with energy signal. The loss of connectivity in the evolved populations is indicated by a larger number of genes which do not act as transcription factors and a reduced number of 'global' transcription factors.

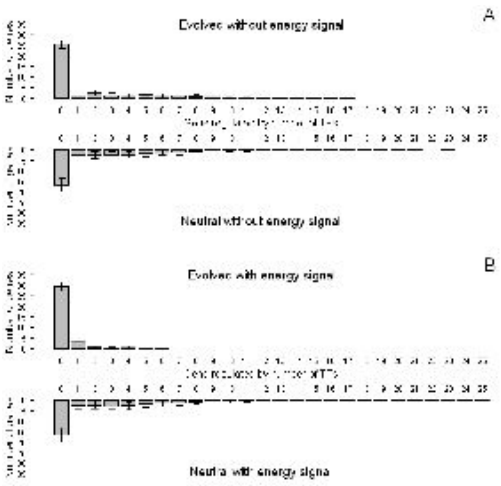

Figure 7: Mean gene 'in' degrees for each gene in each model in small genome populations. Error bars are 1 s.d. of the population. A) evolved and neutral populations without energy signal, B) evolved and neutral populations with energy signal. The loss of connectivity in the evolved populations is indicated by a larger number of genes without any transcription factor interaction and a smaller distribution of interactions.

Transcription Factor Activity The loss of large amounts of regulation can be seen in the interactions between transcription factors (TFs) and genes. This is indicated by the 'out' degree for each transcription factor (Figure 6), and the 'in' degree for each gene (Figure 7). We observe an increase in the number of proteins that do not act as TFs and the number of genes which are not regulated by any TFs. The maximum number of genes regulated by a TF is also significantly reduced in the evolved populations, in particular the population with an energy signal. The maximum number of TF's regulating a gene is also significantly reduced in the evolved populations.

The number of self-regulating genes were separated into: activating only, repressing only, and dual regulation. Again, a clear trend can be observed from the directed populations from Table 4. The two evolved populations have lost nearly all of their self-activating connections, and a large proportion of their self-repressing connections. Whilst more activating connections in total are conserved (Table 2), a larger number of negatively self-regulating connections are conserved, indicating the importance of negative self-regulation in transcription networks.

These results indicate the loss of interaction within the network, and highlight that complex regulatory networks are unnecessary to survive within a stable environment. The preference for losing self-activating connections and preserving more self-repressing connections shows that the network attempts to optimise its energy usage by preventing further transcription of unrequired genes.

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

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

These rather surprising results highlight the complexity of regulation networks even in the most simple of environments. They also show the ingenious mechanisms which natural selection, and the evolutionary operators it uses, have discovered and optimised in both the model networks presented here and the real biological systems.

Regulator type		Population			
		E	N	NI	R
No ES	Activator	0.0165	1.0535	1.319	1.3295
	Repressor	0.4295	1.1645	1.3125	1.3060
	Dual	0	0.0785	0.0600	0.0435
ES	Activator	0.0005	1.0070	1.3125	1.3205
	Repressor	0.2830	1.4440	1.3130	1.3395
	Dual	0	0.1060	0.0575	0.0520

Table 4: Mean number of activating, repressing and dual-regulating self-regulating connections per model within the energy signal and no energy signal populations. The loss of connectivity can be seen in the evolved populations. The evolved populations show a significantly smaller number of activating and repressing and no dual interactions compared with the neutral and random populations. E is evolved, N is neutral, I is initial and R is random population

Summary and Conclusions

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

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

Acknowledgments

We thank Gavin Thomas for information on *B. aphidcola*. DJJ is funded by BBSRC Studentship BBS/S/S/2005/12006. Simulations were on High Performance Compute Cluster funded by BBSRC grant BB/D524624/1.

References

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

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

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

Hardiman, T., Lemuth, K., Keller, M. A., Reuss, M., and Siemann-Herzberg, M. (2007). Topology of the global regulatory network of carbon limitation in *Escherichia coli*. *Journal of Biotechnology*, 132:359–374.

Ingram, P. J., Stumpf, M. P. H., and Stark, J. (2006). Network motifs: structure does not determine function. *BMC Genomics*, 7(108).

Jenkins, D. J. and Stekel, D. J. (2008). A new model for investigating the evolution of transcription control networks. *Artificial Life*. In press.

Karp, P. D., Keseler, I. M., Shearer, A., Latendresse, M., Krummenacker, M., Paley, S. M., Paulsen, I., Collado-Vides, J., Gama-Castro, S., Peralta-Gil, M., Santos-Zavaleta, A., naloza Spínola, M. I. P., Bonavides-Martinez, C., and Ingraham, J. (2007). Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Research*, 37:7577–7590.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467.

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

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

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

Moran, N. A. and Mira, A. (2001). The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biology*, 2(12).

Nath, K. and Koch, A. L. (1970). Protein degradation in *Escherichia coli*: 1. measurement of rapidly and slowly decaying components. *The Journal of Biological Chemistry*, 245(11):2889–2900.

Quayle, A. P. and Bullock, S. (2006). Modelling the evolution of genetic regulatory networks. *Journal of Theoretical Biology*, 238(4):737–753.

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

Stekel, D. J. and Jenkins, D. J. (2008). Strong negative self regulation of prokaryotic transcription factors increases the intrinsic noise of protein expression. *BMC Systems Biology*, 2(6).

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

Wilcox, J. L., Dunbar, H. E., Wolfinger, R. D., and Moran, N. A. (2003). Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Molecular Microbiology*, 48(6):1491–1500.

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