J Mol Evol (2010) 71:128–140
DOI 10.1007/s00239-010-9369-4

# De Novo Evolution of Complex, Global and Hierarchical Gene Regulatory Mechanisms

**Dafyd J. Jenkins · Dov J. Stekel**

Received: 13 May 2010 / Accepted: 12 July 2010 / Published online: 3 August 2010


**Abstract** Gene regulatory networks exhibit complex, hierarchical features such as global regulation and network motifs. There is much debate about whether the evolutionary origins of such features are the results of adaptation, or the by-products of non-adaptive processes of DNA replication. The lack of availability of gene regulatory networks of ancestor species on evolutionary timescales makes this a particularly difficult problem to resolve. Digital organisms, however, can be used to provide a complete evolutionary record of lineages. We use a biologically realistic evolutionary model that includes gene expression, regulation, metabolism and biosynthesis, to investigate the evolution of complex function in gene regulatory networks. We discover that: (i) network architecture and complexity evolve in response to environmental complexity, (ii) global gene regulation is selected for in complex environments, (iii) complex, inter-connected, hierarchical structures evolve in stages, with energy regulation preceding stress responses, and stress responses preceding growth rate adaptations and (iv) robustness of evolved models to mutations depends on hierarchical level: energy regulation and stress responses tend not to be robust to mutations, whereas growth rate adaptations are more robust and non-lethal when mutated. These results highlight the adaptive and incremental evolution of complex biological networks, and the value and potential of studying realistic in silico evolutionary systems as a way of understanding living systems.

**Keywords** Complexity · Gene regulatory network · Evolution · Hierarchical · Computer model · In silico

## Introduction

Biology is, at its core, the study of systems that evolve by natural selection. The complexity of biological systems, and the evolution of such complexity, has posed many problems for evolutionary biologists and theoreticians since Charles Darwin first published the theory of evolution by natural selection. One such hurdle is the lack of a complete fossil record from ancestor to present-day organisms, complete with genomic DNA. One way to overcome this is with computer models of evolution that can track the entire evolutionary history and lineages of the digital organisms, providing a complete 'fossil record' from the 'ancestor' of each organism. For example, Richard Lenski and colleagues' work with the Avida model (Ofria and Wilke 2004) has shown the adaptive evolution of complex features within 'digital organisms' arises in an incremental fashion (Lenski et al. 2003). These results show that increasingly complex functions can be built from simpler functions, and in many cases functionality, or lack of, differed by a single mutation between parent and offspring. In several cases deleterious

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-010-9369-4) contains supplementary material, which is available to authorized users.

D. J. Jenkins · D. J. Stekel
Centre for Systems Biology, School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

D. J. Jenkins
Warwick Systems Biology Centre, University of Warwick, Coventry CV4 7AL, UK
e-mail: Dafyd.jenkins@warwick.ac.uk

D. J. Stekel (✉)
Centre for Integrative Systems Biology, School of Biosciences, University of Nottingham, Sutton Bonington, Loughborough LE12 5RD, UK
e-mail: dov.stekel@nottingham.ac.uk

mutations provided stepping-stones for further complex functionality to evolve.

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

In considering the evolution of gene regulatory networks (GRNs), there is much debate as to whether the observed architectures result from adaptive pressures or non-adaptive processes. Recent evolutionary work using *Escherichia coli* has shown the de novo evolution of global regulatory networks governing DNA superhelicity and stringent response in multiple, independent populations, indicating adaptive selection (Philippe et al. 2007). Further key innovations such as the evolution of a population capable of using citrate as a food source in a glucose-limited environment indicate adaptive selection for such functionality (Blount et al. 2008). Recent theoretical and computational works have also indicated adaptive forces drive and shape network evolution (Crombach and Hogeweg 2008; Jenkins and Stekel 2010). One aspect debated is that of network motif abundance. Proponents of network motifs claim that the over-abundance of several motifs, such as the Feed-Forward Loop (FFL), is evidence of adaptive selection (Milo et al. 2002). This argument is further strengthened by research indicating that FFLs have specific function with GRNs and are therefore adaptively selected for on this functionality (Shen-Orr et al. 2002; Mangan and Alon 2003; Dekel et al. 2005; Kaplan et al. 2008) and specific motifs are more conserved in organisms sharing similar lifestyles (Babu et al. 2006). Yet, evidence also suggests that network motif structure does not determine function (Mazurie et al. 2005; Ingram et al. 2006; Meshi et al. 2007). Moreover, the over-abundance of specific motifs is based on a flawed argument using Erdos–Renyi random graphs. When realistic replication process of GRNs is taken into account the over-abundance of network motifs is easily explained as a result of non-adaptive processes (Teichmann and Babu 2004; Babu et al. 2004; Banzhaf and Kuo 2004; Cases and de Lorenzo 2005; Cordero and Hogeweg 2006). Other aspects are global regulators (Gottesman 1984; Martínez-Antonio and Collado-Vides 2003) and 'scale-free' network properties (Barabasi and Albert 1999). Cases and de Lorenzo (2005) describe a process of non-adaptive genome evolution that will result in the occurrence of global regulators through purely non-adaptive gene duplications. Lynch (2007a, b) takes this argument further showing that "many of the qualitative features of known transcriptional networks can arise readily through the non-adaptive processes of genetic drift, mutation and recombination". Without being able to repeat evolution of life on Earth, it is difficult to resolve these questions.

Global regulators have previously been defined by Gottesman (1984) and further defined by Martínez-Antonio and Collado-Vides (2003), as transcription factors (TF) that: (i) regulate several metabolic pathways, or responses to environmental stimuli, (ii) regulate large numbers of genes and operons, (iii) will form regulation cascades, providing a hierarchy of regulation, (iv) are likely to co-regulate with other TFs or global regulators and (v) regulate operons which are transcribed by different $\sigma$ factors. The clear and quantitative definition of global regulator structure therefore makes this complex structure an ideal 'motif' to search for within gene regulatory networks and their evolution.

Once specific features have been identified within gene regulatory networks, the adaptive and non-adaptive hypotheses for network evolution can be addressed. If the non-adaptive hypothesis, supported by Lynch and others, were correct, then we would expect to find similar structures, such as global regulation and hierarchy in networks evolved both adaptively and non-adaptively. Whereas, if the adaptive hypothesis, supported by Lenski and proponents of network motifs, is correct we would expect to find statistically significant differences between network structures and architecture of networks evolved adaptively when compared to those evolved using only non-adaptive processes.

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

In posing these questions, we are relying on a concept of complexity; however, measurement of the complexity of biological systems is inherently a very difficult task. Adami et al. (2000) discuss measures of biological complexity at great length and point out that complexity can be defined in physical, structural or functional ways. We use a loose, qualitative view of complexity, considering the numbers of and interactions between functional sub-systems.

Unlike much previous work, we use a biologically informed GRN model that includes the processes of transcription, translation, metabolism, biosynthesis and detoxification, which is fairer to compare with biological

systems. To achieve such a comparison, the single fitness function within the model, rather than a specific goal, is equivalent to the 'fitness' of any biological organism: survival and replication. We investigate the evolution of GRNs and their functional complexity using several simulated environments of varying complexity, with the single goal of surviving and replicating.

We show that network structure and function is directly related to environmental conditions. Further, network complexity also evolves in a number of stages, generating a functional hierarchy of inter-connected systems, in which specific systems require others as a prerequisite to their evolution. Global regulators are also strongly selected for under specific conditions. The resultant selected global regulators have a structure different to those found non-adaptively, indicating an adaptive influence. Perturbations to this hierarchy produce effects of differing severity based on the level perturbed; lethal in low-level, 'essential' systems, to negligible in high-level systems, indicating varying levels of robustness.

## Materials and Methods

### Model Components

The gene regulatory network model used is an extension of the adaptive model described by Jenkins and Stekel (2010). Briefly, the model consists of a gene regulatory network (GRN) which contains three types of genes: (1) input genes, which represent the state of the internal energy, detection of external food environments of the model or detection of internal stress conditions, (2) regulatory genes, which act solely as transcription factors (TFs) and (3) output genes, which represent the production of biomass, or response by a stress pathway, and a number of molecular species representing proteins, energy and stress.

The first type of input gene, *nrg1* and *nrg2*, represents internal energy signalling systems, responding to the level of energy within the model, and is activated if level of energy is above a threshold, $T_1^{\text{energy}}$ or $T_2^{\text{energy}}$, for *nrg1* and *nrg2*, respectively (see Table 1 for parameter values and how they are derived). This process is analogous to unbound catabolite repressor protein (CRP) in *E. coli* reflecting a high-energy state, and bound CRP–cAMP complexes reflecting a low-energy state. A second type of input gene, *fod1* to *fod9*, represents a generic enzyme within a metabolic pathway associated with a particular food and therefore reflects the activation of that pathway when that food is detected in the external environment. Food availability is determined by a predefined function, such that each food will be available for approximately 12% of the time, and when food is available a fixed amount of energy is produced within the model, $P_{1-9}^{\text{energy}}$, for *fod1* to *fod9*, due to metabolism (not modelled). The final type of input gene, *rcp1* and *rcp2*, represents an intra-cellular stress receptor and signalling system, and is activated if any amount of a specific stress is present within the model.

The first type of output genes, consisting of *syn1* to *syn4,* represents generic proteins within four different biosynthetic pathways and therefore represents activation of each of those pathways as a whole. A fixed amount of biomass is generated with each activation event, $P_{1-4}^{\text{bio}}$ (for *syn1* to *syn4*, respectively). Similarly, a fixed cost is associated with each activation event representing the production of additional proteins (not modelled) within the pathway, $C_{1-4}^{\text{bio}}$ (for *syn1* to *syn4*, respectively). The production and cost values for each biosynthesis gene are different meaning that the biosynthesis pathways have varying energetic properties. The final type of output genes, *rsp1* and *rsp2*, represents stress response pathways, and with each activation event will remove a fixed amount of specific stress molecules to the receptor/response pair of genes, $R_1^{\text{stress}}$, $R_2^{\text{stress}}$ (for *rsp1* and *rsp2*, respectively). Similarly, a fixed cost is associated with each response event representing the additional energy required to degrade or remove the stress molecules from the model, $C_1^{\text{stress}}$, $C_2^{\text{stress}}$ (for *rsp1* and *rsp2*, respectively).

Each type of gene (input, regulatory and output) produces a protein product, which can interact with specific binding sites on the DNA and influence expression of other genes. A cost is associated with each gene expression event, consisting of cost for transcription ($C^{\text{mRNA}}$) and translation ($C^{\text{protein}}$) of $n$ proteins, given by Eq. 1:

$$C^{\text{n}} = C^{\text{mRNA}} + nC^{\text{protein}} \quad (1)$$

The model abstractly represents the energy-containing potential of molecules such as ATP, nucleotides and amino acids as a single 'energy' molecule, represented as an integer value within the model. Energy, as in biological organisms, is essential for fuelling cellular processes, such as transcription and translation, and if the energy level falls to, or below, 0, then the model dies. Stresses are modelled as integer values representing how many stress molecules are present within the model, and if either threshold, $T_1^{\text{stress}}$ or $T_2^{\text{stress}}$ (for stress 1 and 2, respectively) is exceeded, then the model dies. Biomass is modelled as an integer value representing how much biomass has been produced. Biomass, or yield, is an indicator of growth and as such is the primary measure of fitness of the models.

### Genes

Each gene $i$, with the exception of input genes, has an associated regulatory region consisting of a set, $J$, of binding sites, $j$. Each binding site $j$ can be either activating, $r_{ij} = 1$, or inhibitory, $r_{ij} = -1$ and has an occupancy

**Table 1** Model and evolution parameters

| Parameter | Value | Note |
|---|---|---|
| $s^{max}$ | 128 | |
| $d^{max}$ | 3 | |
| $C^{mRNA}$ | 3 | ~2000 ATP molecules required to transcribe 1080 nt sequence (Sunderaraj et al. 2004) |
| $C^{protein}$ | 2 | ~1500 ATP molecules required to translate 360 aa sequence (Sunderaraj et al. 2004) |
| $K^{basal}$ | $1 \times 10^{-2}$ | Derived from model analysis (Jenkins and Stekel 2010) |
| $P^{energy}_{1-9}$ | 5, 5, 10, 10, 15, 15, 20, 20, 25 | Single 'energy' molecule is approximately equal to 700 ATP molecules |
| $P^{bio}_1$, $P^{bio}_3$ | 50 | High production biosynthesis pathways |
| $P^{bio}_2$, $P^{bio}_4$ | 10 | Low production biosynthesis pathways |
| $C^{bio}_1$, $C^{bio}_2$ | 75 | High cost biosynthesis pathways |
| $C^{bio}_3$, $C^{bio}_4$ | 5 | Low cost biosynthesis pathways |
| $P^{stress}_1$, $P^{stress}_2$ | 25 | Stress molecules removed per stress pathway activation |
| $C^{stress}_1$, $C^{stress}_2$ | 100 | Energetic cost per stress pathway activation |
| $T^{stress}_1$, $T^{stress}_2$ | 100 | Stress threshold |
| $T^{energy}_1$ | 500 | High-energy signal threshold |
| $T^{energy}_2$ | 333 | Low-energy signal threshold |
| Initial genome size | 32 | Derived empirically from (Jenkins and Stekel 2008) |
| Starting energy | 1000 | |
| Simulation time-steps | 2000 | Each time-step approximates 1 min |
| Population size | 100 | |
| Generations (stress-free and non-adaptive) | 1000 | Represents approximately $5 \times 10^6$ bacterial generations |
| Generations ('stressed') | 10000 | Represents approximately $5 \times 10^7$ bacterial generations |

value, $o_{ij}$, which is 1 if bound and 0 otherwise, where $i$ is the index of the gene, and $j$ is the index of the binding site within gene $i$. Regulation state, $a_i$, of a gene is dependent on activating and inhibitory binding site occupation according to the Eq. 2:

$$a_i = \sum_{j \in J} r_{ij} o_{ij} \tag{2}$$

Input gene activation is dependent on other cellular or environmental states (as defined above). Each gene encodes a protein product which has a number of parameters: shape, $s_i$, an integer value from a 1D circular shape space (of size $s^{max}$); protein production, $\text{prod}_i$, a non-negative integer value determining the mean number of proteins produced per gene expression event; protein degradation rate, $\text{deg}_i$, a non-negative integer value determining the mean number of simulation time-steps before the protein passively degrades. Each binding site also has a shape parameter drawn from the same shape space as protein shape.

## Protein–DNA Binding and Affinity

Each gene and binding site has a shape, and the complementarity between the two shapes determines the binding affinity between a protein and a binding site. The 1D circular shape space was introduced by Cordero and Hogeweg (2006), and provides an abstracted representation of protein binding domains and binding site structure. The binding affinity, $b_{ij}$, between two integer shapes, $s_i$ and $s_j$, will take a discrete value given by Eq. 3:

$$b_{ij} = \begin{cases} \frac{1}{d_{ij}+1} & \text{if } d_{ij} \leq d^{max} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$d_{ij} = \left\| s_i - s_j \right\| \tag{4}$$

where $d^{max}$ is the largest integer Euclidean distance which two shapes can bind. A 3-gene regulatory network example showing protein–DNA interactions and binding affinity is shown in Fig. 1.

## Transcription, Translation and Basal Expression

Transcription and translation are modelled as bursts, in which several mRNA or protein molecules are synthesised simultaneously, reflecting the experimental work of Cai et al. (2006) with an energetic cost associated with the transcription ($C^{mRNA}$) and translation ($C^{protein}$) events. Any gene $i$ whose regulation state, $a_i$, is $\geq 1$ (more positive bindings) will be expressed. Additionally, any gene with a regulation state of 0 (equal positive and inhibitory binding)

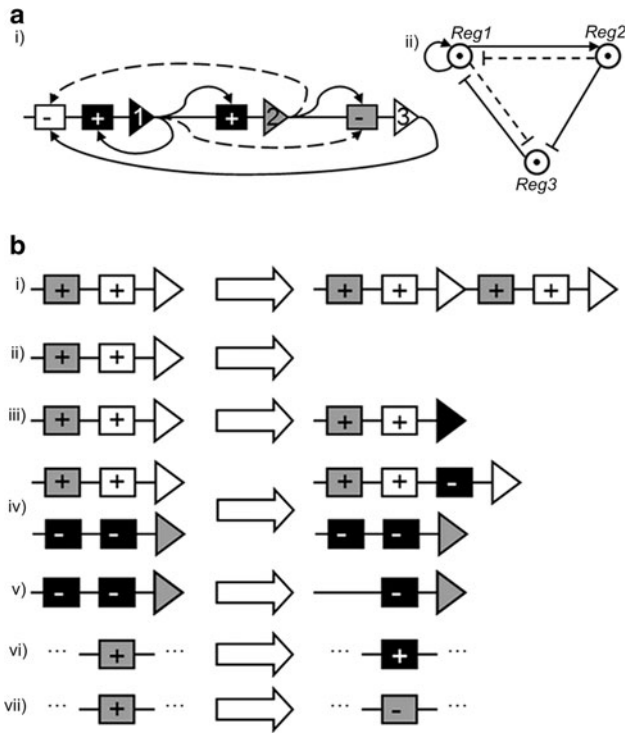


**Fig. 1** Network interaction examples and mutational operators. **a**(i) shows an example 3-gene regulatory network, indicating protein–DNA interactions, where *rectangles* represent binding sites (+ activating; − inhibitory), *triangles* represent genes. The *greyscale* fill of each binding site and gene represents its 'shape'. *Unbroken lines* represent strongest binding affinity, whilst *dashed lines* represent a weaker binding affinity. **a**(ii) shows the same 3-gene network visualised as a graph. Nodes represent genes; → are positive interaction; —| are negative interaction and weighting of line again represents binding affinity. **b** shows several of the mutational operators in diagrammatic form: (i) gene duplication, (ii) gene loss, (iii) protein shape mutation, (iv) binding site duplication, (v) binding site loss, (vi) binding site shape mutation and (vii) binding site regulation type 'flip'

can randomly express with a given fixed probability, $K^{\text{basal}}$, representing random RNA polymerase binding events. The mean number of proteins produced per transcription event (and associated energetic cost, see Eq. 1) is dependent on the protein production value.

## Protein Degradation

Protein degradation is a passive process, which is determined by the degradation value of a protein, $\text{deg}_i$, for the protein corresponding to gene $i$. This value represents the number of time-steps the protein remains stable and functional. Once a protein has degraded, it is removed from the model.

## Model Initialisation

All models are initialised with a linear genome, consisting of two energy signalling genes, nine food genes, two stress receptor genes, 32 regulatory genes, four biosynthesis genes and two stress response genes in this order. Each gene parameter is randomly assigned with the protein shape selected from the 1D integer shape space, protein production value between 0 and 8 (with equal probabilities) and protein degradation value between 1 and 3 (with equal probabilities). Further, each regulatory gene and output gene is assigned a random regulatory region consisting of between 0 and 3 binding sites (with equal probability), each with the binding site shape selected from the 1D integer shape space, and regulatory type either activating or repressing (with equal probability). All parameters for energy and stress thresholds, energy production, biosynthetic cost and production and basal expression values are identical and fixed for all models.

## Model Simulation

The models are simulated using discrete Boolean networks. To facilitate the evolution of realistic network structures and mechanisms, a stochastic formulation of Boolean networks is used, capturing the essential inherent noise in the molecular processes (Jenkins and Stekel 2010). Simulation of each model consists of a fixed number of time-steps, consisting of a sequential ordering of sub-steps, and an equal starting energy level:

1. Determine food availability and stress levels from time-dependent functions.
2. *Determine ordering of protein and binding site interactions* Each species of protein within the model is selected to interact with binding sites in a specific order. The ordering of the binding sites is also specified each time-step. Each protein species will attempt to interact with unoccupied binding sites, until either no free protein molecules are available, or all binding sites have been selected. This ordering of protein and binding sites is randomised each time-step.
3. *Determine regulation state of input genes nrg* genes are activated based on specific energy levels, *fod* genes are activated based on environmental food availability and *rcp* genes are activated based on the presence of intra-cellular stress molecules.
4. *Transcribe and translate input genes* Each input gene that was activated in step 3 is expressed, producing protein. The number of proteins generated is a random normal, with $\mu = \text{prod}_i$ and $\sigma = 0.5$, rounded to the nearest non-negative integer. The energetic costs for the input genes transcription and translation events are calculated using Eq. 1.
5. *Determine Protein–DNA interactions* Using the protein and binding site order from step 2, protein–DNA

interactions are determined. The binding affinity, $b_{ij}$, is used to determine if binding occurs between protein $i$ and binding site $j$. A protein will bind if $b_{ij} > [0,1)$ (where [0,1) is a random uniform variable between 0 (inclusive) and 1 (exclusive)).

6. *Determine regulation state of regulatory and output genes* The regulation state, $a_i$, is calculated based on binding site occupancy, $o_{ij}$.
7. *Transcribe and translate regulatory and output genes* Activated genes ($a_i > 0$) and basally expressed genes ($a_i = 0$ and $K^{\text{basal}} > [0,1)$) are transcribed and translated (including output processes).
8. All proteins unbind from the DNA ($o_{ij} = 0$) and all genes are inactivated ($a_i = 0$).
9. *Determine protein degradation* Proteins are degraded with the probability of $1/\deg_i$.
10. *Check simulation termination criteria* (1) required number of time-steps completed (model survives), (2) energy level falls to, or below, 0 (model dies) and (3) any stress threshold is exceeded (model dies).

## Evolutionary Framework

The evolutionary environment is a genetic algorithm (Holland 1992), which consists of a fixed-size population of models. The initial generation, consisting of randomly generated models, is non-adaptively evolved for 10 generations to generate a more biologically realistic network (for evolutionary operators see below). A single offspring from each non-adaptively evolved initial network is generated, creating a population with twice the initial number of models. Each model within the population is then simulated independently, but with identical environmental conditions for 2000 time-steps. After simulation, each model is assigned a fitness value, $f_i$:

$$f_i = \begin{cases} \text{biomass generated} + \text{time steps} & \text{if model replicates} \\ \text{time steps survived} & \text{otherwise} \end{cases} \quad (5)$$

Each successive generation consists of the fittest 50% of the surviving population (elitist selection). The selected models then each replicate once, creating a new generation of models. In each generation of the evolutionary algorithm, the models are reset, so have the same initial energy level of 1,000 molecules, and 0 protein. During the replication process both the parent and the daughter model can mutate (see below). The evolutionary process is then repeated with the new population. Whilst no direct competition between models is present (such as competition for food), the number of models in each generation is constant, and so is a limiting factor and as such introduces competition between the models, generating evolutionary pressure.

## Evolutionary Operators

A number of evolutionary operators are defined at the individual gene level and genome level: (i) *Gene duplication*, in which the entire gene (including protein parameters) and its regulatory region is duplicated and added to the genome with probability $M_{\text{gDup}}$. If either an input or output gene is duplicated then the gene and its associated parameters and regulatory region are duplicated, however, the duplicate gene does not function as an input/output gene. When a gene is duplicated, for the purpose of gene/protein ordering during simulation, they are placed after the existing regulatory genes in a linear genome, but before output genes; (ii) *Gene loss*, where the entire gene and its regulatory region is removed from the genome with probability $M_{gLoss}$. Input and output genes cannot be lost. This ensures a 'minimal' genome will always exist consisting of the initially defined number of input and output genes; (iii) *Protein shape mutation*, in which the protein shape is mutated by a random normal, with $\mu = 0$ and $\sigma = \log_{10} s^{\max}$, with probability $M_{\text{pShape}}$; (iv) *Protein production mutation*, where the protein production value is mutated by a random normal, with $\mu = 0$ and $\sigma = 2$, with probability $M_{\text{pProd}}$; (v) *Protein degradation mutation*, where the protein degradation value is mutated by a random normal, with $\mu = 0$ and $\sigma = 2$, with probability $M_{\text{pStab}}$; (vi) *Binding site duplication*, where a random binding site from the genome is duplicated with probability $M_{\text{bsDup}}$; (vii) *Binding site loss*, where a binding site is lost with probability $M_{\text{bsLoss}}$; (viii) *Binding site shape mutation*, where the shape of the binding site is mutated by a random normal, with $\mu = 0$ and $\sigma = \log_{10} s^{\max}$ with probability $M_{\text{bsShape}}$; (ix) *Binding site regulation flip*, in which the regulation type (activating or inhibitory) is flipped with probability $M_{\text{bsFlip}}$. Mutation probabilities are defined per gene for gene duplication/loss, protein shape/production/degradation mutation operators and per binding site for binding site loss/duplication/shape/regulation mutation operators. All mutation parameter values are shown in Table 2.

## Experimental and Environmental Conditions

Two sets of internal and external environments of the 100 models were simulated and evolved: (1) a stress-free, base environment, and (2) a 'stressed' environment, which introduces a number of stresses to the base environment. The base environment consists of nine food sources, each providing 5, 10, 15, 20 (two sources of each) or 25 molecules of energy, which are always available. In the 'stressed' environment, each food source is randomly available for approximately 12% of the simulation. Four biosynthesis pathways, with combinations of high and low yield, high

**Table 2** Mutation parameter values

| Parameter | Description | Value |
|---|---|---|
| $M_{\text{gDup}}$ | Probability of gene duplication mutation event per gene | $1 \times 10^{-3}$ |
| $M_{\text{gLoss}}$ | Probability of gene loss mutation event per gene | $1 \times 10^{-3}$ |
| $M_{\text{pShape}}$ | Probability of protein shape mutation event per gene | $5 \times 10^{-3}$ |
| $M_{\text{pProd}}$ | Probability of protein production mutation event per gene | $5 \times 10^{-3}$ |
| $M_{\text{pStab}}$ | Probability of protein stability mutation event per gene | $5 \times 10^{-3}$ |
| $M_{\text{bsDup}}$ | Probability of binding site duplication mutation event per binding site | $8 \times 10^{-3}$ |
| $M_{\text{bsLoss}}$ | Probability of binding site loss mutation event per binding site | $8 \times 10^{-3}$ |
| $M_{\text{bsShape}}$ | Probability of binding site shape mutation event per binding site | $8 \times 10^{-4}$ |
| $M_{\text{bsFlip}}$ | Probability of binding site regulation flip mutation event per binding site | $8 \times 10^{-4}$ |

and low cost, are represented in the genome. Two energy signalling genes, one detecting high energy concentrations (500 molecules) and the other detecting low energy concentrations (333 molecules) are also represented. The 'stressed' environment consists of two further stresses (representing denatured proteins within the cell) and two corresponding stress response pathways. Every 25 simulation time-steps, 25 stress molecules enter the cell, generating a fixed cyclical stress regime. Each activation event of a stress response pathway removes 25 molecules of the associated stress. If the number of a specific type of stress molecules reaches the given lethal stress threshold (100), then the cell dies. Model parameters are given in Table 1, and a discussion of parameter value selection can be found in Jenkins and Stekel (2010).

### Functional Complexity

Gene regulatory networks can be viewed as consisting of a number of integrated components or systems. In this model, we define three functional systems, each corresponding to a specific biological function:

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

Therefore in this model, a qualitative representation of complexity is the number of and interactions between these functional systems.

### In Silico Global Regulators

In this model, the primary criteria for global regulation classification is based on global regulator properties (i) regulation of several metabolic pathways and (ii) regulation of large numbers of genes and operons. Thus, as the model consists of two types of outputs (biosynthesis and stress response), and each type of output can consist of a number of individual pathways, a global regulator is defined as regulating not only intra-type pathways (biosynthesis or stress response), but also inter-type pathways.

To test percentage of genome regulated, percentage of network edges regulated and percentage of positive edges regulated by the regulators, a $\chi^2$ test could not be used as the genome and regulated edges sizes of the evolved models were so small that the expected values were much less than 5 for each network (see Supplementary Tables S1 and S2). Therefore, the non-parametric Wilcoxon rank-sum test was used to test the proportions (computed using *R*).

## Results

### Complexity of Evolved Network is Strongly Influenced by Environmental Complexity

We evolved a number of randomly initialised model populations under 'stress-free', $S^-$, and 'stressed', $S^+$, environmental conditions, reflecting increasing complexity. The network architectures were dramatically different between the two types of populations (Fig. 2). Networks evolved under stressed conditions had a large and dominant energy regulation system (red box), which mainly consisted of one or several co-regulating global regulators (Fig. 2a). Two stress response systems (yellow boxes) were observed in all final models. The double activation by the associated stress receptor signal is a mechanism to over-ride the global energy regulation system, and thus indicates a co-evolving relationship between the two systems. The presence or absence of growth systems (blue boxes) within the network is highly dependent on whether the network can replicate.

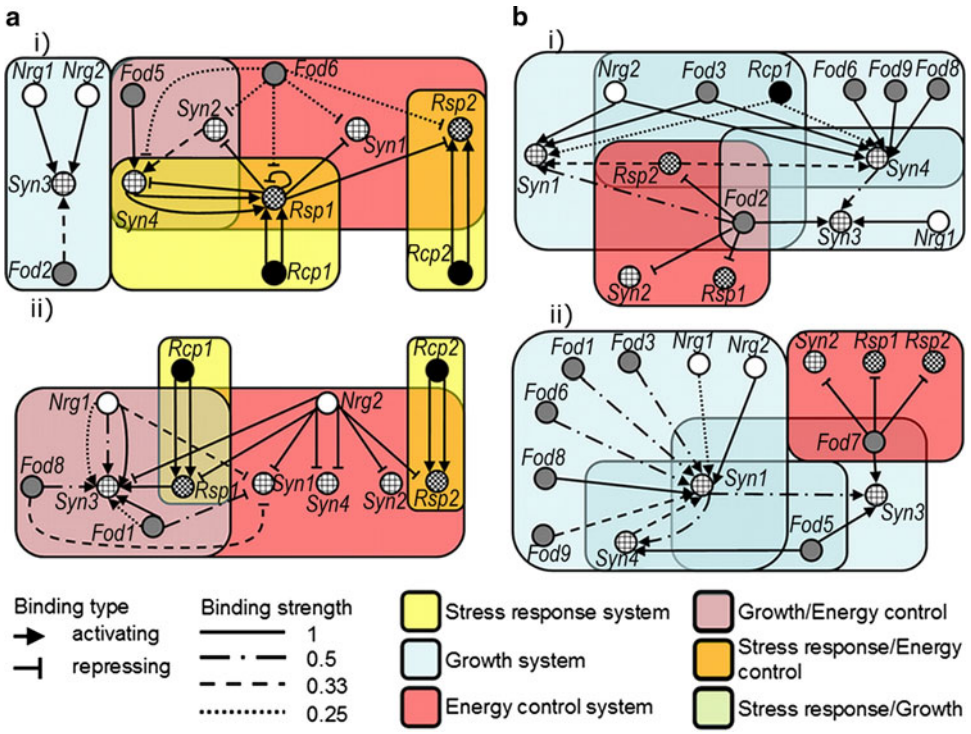


**Fig. 2** Different network topologies evolve in different environments. Populations evolved in environments with stresses (starvation and heat-shock-like stresses) have networks with a wide variety of all possible sub-systems **a**(i–ii). A similar functional structure is observed, consisting of a large energy regulation system (*pink*), two stress response systems (*yellow*) and growth systems (*blue*). Global regulators are evident in each example (*Rsp1* and *Fod6* in **a**(i); *Nrg2* in **a**(ii)), each performing only repression. Populations evolved in a stress-less environment do not show as much variety of sub-systems **b**(i–ii). Energy regulation is on a smaller scale than in stressed populations, but growth systems are more heavily utilised, usually consisting of at least three growth systems. Global regulators are also present in all examples, but have a different structure to those found in the stressed populations. The global regulators perform the energy regulation, but are also incorporated into the growth systems, meaning the global regulators are dual-function

Networks a(i) and a(ii) each contain at least one growth system. Networks evolved under stress-free conditions have a much smaller energy regulation system, although still usually consisting of a single global regulator (Fig. 2b). However, the global regulators often perform both activation and repression of different output pathways. The growth systems are the more dominant systems (examples are shown in networks b(i, ii)). The systems are largely interconnected, with many input pathways activating several biosynthesis pathways.

## The Evolution of Global Regulators is Adaptive

We examined the largest regulator (gene with most outgoing connections) within each population, evolved under stressed ($S^+$), stress-free ($S^-$) or non-adaptive (*NA*; random fitness function) conditions. Table 3 shows the mean percentage of genome regulated and percentage of all network interactions regulated by the largest regulator in each of the 10 replicate populations (networks with multiple different regulators were excluded. Full data available in Supplementary Table S1). Under stressed conditions 62.9% of all genes were regulated by the largest regulator, and a similar proportion (57.1%) of genes was regulated under stress-free conditions. In non-adaptively evolved populations, the percentage of genes regulated by the largest regulator was significantly smaller, with an average of 11.2% of genes regulated (stressed: $P = 1.717 \times 10^{-4}$; stress-free: $P = 1.817 \times 10^{-4}$). The total proportion of network interactions regulated by the largest regulator was also significantly higher in the stress (22.8%) and stress-free (19.6%) populations, than the 1% in non-adaptive populations (stressed: $P = 1.083 \times 10^{-5}$; stress-free: $P = 1.083 \times 10^{-5}$). The numbers of positive and negative interactions by the largest regulator in the non-adaptive populations were statistically equivalent ($P = 0.23$), whereas stressed populations had a significant bias towards negative interactions (100%; $P = 6.386 \times 10^{-5}$), and the stress-free populations had a less significant bias towards negative interactions (66.2%; $P = 2.305 \times 10^{-3}$). None of the largest regulators within the non-adaptively evolved populations were classed as global regulators, but in stressed populations 100% and in stress-less populations 90% of the largest regulators were classed as global regulators. Global regulation was therefore strongly selected for under both stressed and stress-free environmental conditions, and adaptive regulator structure

**Table 3** Largest regulator statistics in stressed ($S^+$), stress-free ($S^-$) and non-adaptive (NA) populations

| Population | % genome regulated[a] | % interactions regulated | % activating interactions | % of networks where largest regulator is 'global' |
|---|---|---|---|---|
| NA | 11.2 ± 2.1 | 1.0 ± 0.4 | 53.4 ± 8.0 | 0 |
| $S^+$ | 62.9 ± 20.1 | 23.4 ± 7.8 | 0 ± 0 | 100 |
| $S^-$ | 57.1 ± 29.7 | 19.6 ± 12.3 | 33.8 ± 15.2 | 90 |

[a] Input genes are excluded from genome size

is significantly different from non-adaptive regulator structure. Analysis of the largest regulators in the randomly generated networks is in Supplementary Results and Supplementary Table S2.

## Complexity of a Network Arises in Stages

We examined the entire evolutionary history of a number of models from each type of population. Figure 3 shows selected 'snapshots' of the best performing model in one of the stressed populations over its 10,000 generation lineage. The initial network (randomly generated) consists of a small energy regulation system regulating a single biosynthesis pathway, and an inter-connected growth system (Fig. 3a). Additionally, a stress response system is present. This network is non-viable (able to survive no more than 50 time-steps) and with the minimal system architecture is not complex.

The network evolution progresses in two broad phases; this is related to the two components of the fitness function. In the first phase (Fig. 3a–d), the networks are unable to be replicated, and selection is for longevity. In the second phase (Fig. 3e–h), the networks are able to replicate, and selection is for rate of growth. After 100 generations the network has substantially changed from its initial state (Fig. 3b). The complexity of the energy regulation system is increased, with three biosynthesis pathways regulated. A second stress response system has evolved, whilst the complexity and efficiency of the original response system has increased, with the addition of activation by its associated receptor. The specific growth system has been lost, with increasing efficiency of energy regulation. This together with the stress response systems allow the network to survive around 100 time-steps. The network after 250 generations has increased the efficiency of the second stress response system, also evolving activation by its receptor (Fig. 3c). The increased efficiency of the stress response systems now prevents the cell from dying due to the lethal stress levels. The efficiency of the energy regulation systems has further increased, with additional input pathways regulating the biosynthesis pathways. This network is able to survive between 150 and 250 time-steps. The network after 500 generations again has an increasingly complex energy regulation system, with three co-regulating global regulators providing an efficient, but redundant, energy saving mechanism (Fig. 3d). A different growth mechanism has also reappeared, but the energetic cost is still not sustained by the network. These adaptations increase the survival to between 300 and 400 time-steps.

The second phase begins at generations 1166 with the emergence of the first replicating network. This network shows the first appearance of what becomes the primary global regulator, *Rsp1*, in the energy regulation system (Fig. 3e). A co-regulating global regulator, *Fod2*, is also present in this system. The complexity of the stress response systems has increased, with each receptor binding to multiple binding sites of each response gene. This additional complexity has evolved in response to the incorporation of both stress response pathways into the energy regulation system, indicating an adaptive response to the other systems. By 1,500 generations, the network whilst maintaining similar stress response systems, has lost the co-regulating global regulator *Fod2* (Fig. 3f). The growth system, *Syn3*, has been modified to be more efficient, using a food signal. Moreover, the network is now able to sustain this system and biomass production (now the measure of fitness) has increased fitness from around 3,500 to over 5,500. The network after 5,000 generations has evolved a new co-regulator, *Fod6*, in the energy regulation system (Fig. 3g), possibly indicating an adaptation for robustness or role as a secondary regulator. The functional systems are increasingly inter-connected, producing an increasingly complex network. The growth system remains and is increasingly more efficient, with regulation from the energy signalling pathways (*Nrg1* and *Nrg2*), resulting in an increase of biomass production and fitness to around 19,500. At the end of the evolution, 10,000 generations, the network structure is similar to the previous network. The energy-regulating system still consists of two global regulators, but with weakened interactions from the secondary *Fod6* regulator (Fig. 3h). The main growth system has become decoupled from the energy regulation system. Protein production and stability rates have also been modified, leading to an increase in biomass production and fitness of around 22,000 (Supplementary Table S3). Analysis of fitness and modularity can be found in Supplementary Results and Supplementary Figures S1 and S2.

## Highly Adapted Models Consist of Inter-Connected Systems with Essential and Non-Essential Components

The sub-systems of the evolved networks are mostly overlapping throughout the evolutionary simulation (Fig. 3).

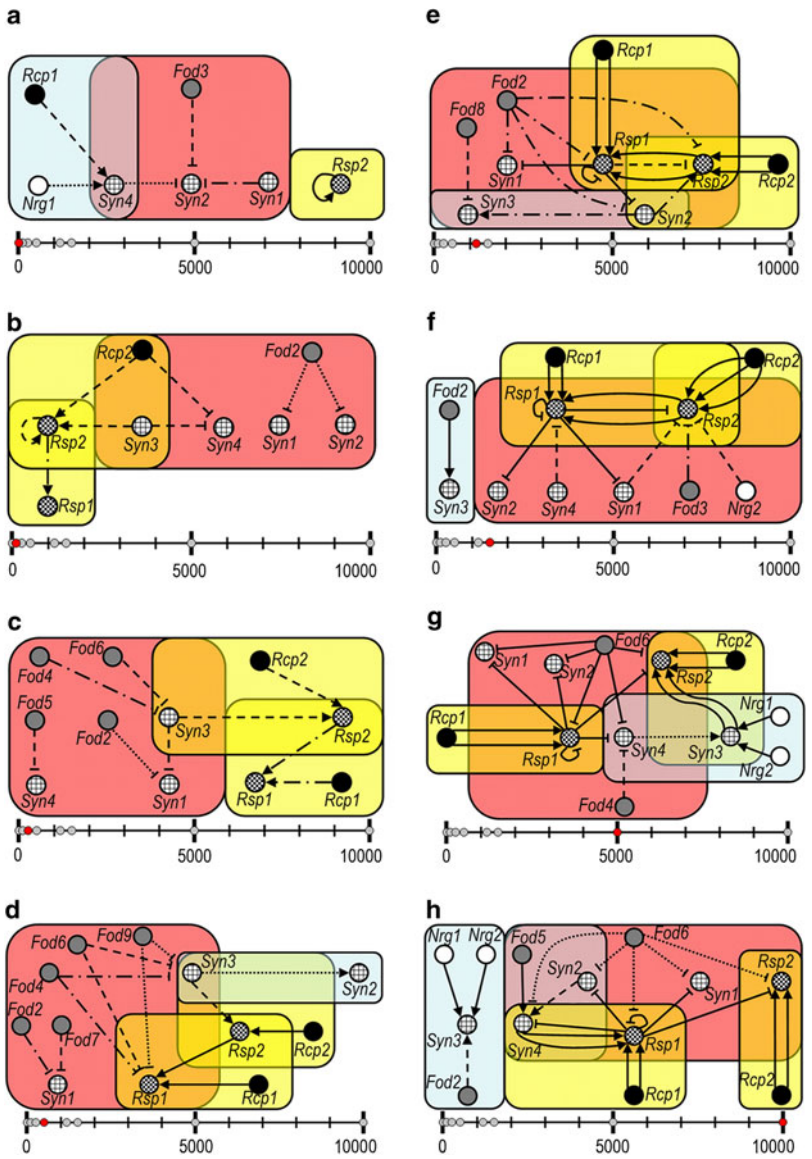


**Fig. 3** Incremental evolution of a functionally complex gene regulatory network. The network initially starts with each type sub-system, yet is unable to survive more than a few tens of time-steps (**a**). After 100 generations the network structure has changed dramatically, losing the growth system, but gaining an efficient stress response system (**b**). By 250 generations a second stress response system has evolved, and also the energy regulation system has continued to grow (**c**). By 500 generations a number of small global regulators have evolved, further increasing the energy regulation system (**d**). The network also has a number of redundant regulators each performing identical roles. A growth system has also re-emerged which is interacting with stress and energy regulation systems. The first replication event after 1,166 generations shows the network has a very efficient set of stress response systems, and also the emergence of just two co-regulating global regulators performing the majority of the energy regulation (**e**). The different systems have become increasingly interconnected. After 1,500 generations only a single global regulator now performs the key role in global regulation (**f**). An independent growth system has also emerged, which is now viable due to efficient energy regulation and stress systems. Network functionality remains similar after 5,000 generations, with the global regulator increasing the number of pathways regulated, and recruitment of another gene as a transient global regulator (**g**). The growth system has increased in efficiency, now utilising the energy signals. After 10,000 generations network structure and function is again similar (**h**). The energy regulation system is still controlled by the same global regulator, and a secondary weaker connected regulator. The main growth system is now independent of other systems, and a second has evolved within the energy regulation and stress systems. The network functionality is evolved in stages, with certain systems as prerequisites for the sustainability of others

Several of these sub-systems are very sensitive to mutation (Fig. 4 and Supplementary Table S4). Removal of either stress response system, by disabling the interactions from the proteins Rcp1 (*str1KO* in Fig. 4), or Rcp2 (*str2KO*) is completely lethal. Removal of the regulation interactions from the global regulator Rsp1 (*nrgKO*) is also completely

lethal; removal of the interactions from the global regulator Fod6 has only a minor impact on survival (data not shown). This suggests that the additional global regulation by *Fod6*, rather than acting as either a back-up regulator or co-regulator, may be non-adaptive. In Fig. 3g and h it can be seen that the strengths of the interactions from Fod6 have decreased over the 5,000 generations shown and it is plausible that this global regulatory system is in the process of being lost. Major reductions (90% or greater) in the production or stability rates of the global regulator, *Rsp1*, are also detrimental to survival rate. However, the network is also robust to other mutations and is able to withstand, to varying degrees, entire removal of some systems. Removal of the decoupled growth system (*grwKO*) severely reduces the growth rate of the model, but increases its survival to almost 100%. Small reductions in production or stability rates (<50%) are mostly non-lethal.

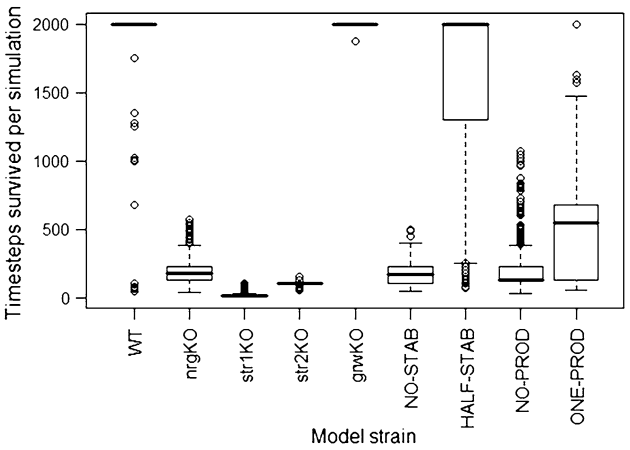


**Fig. 4** Robustness and fragility to mutations in network components. Wild-type (WT) model is evolved network from Fig. 3h and each strain is simulated 1000 times. The WT consistently reaches the termination criteria of 2,000 simulation steps, indicating a robust and efficient network. Removing the global regulators (*nrgKO*) governing the energy regulation system reduces survival rate to 0, but is able to survive around 150 time-steps. Removal of the first stress response system (*str1KO*) also reduces the survival rate to 0, and can survive only tens of time-steps. A similar result is observed removing the second stress response system (*str2KO*), but survives around 100 time-steps. The mutants, *str1KO* and *str2KO*, cause the network to die at different points in simulation, due to the additional global regulator activity of elements of the *str1* system. Removal of the independent growth system (*grwKO*) has a positive effect on survival rate, reaching nearly 100%. Therefore, certain sub-systems are pre-requisites for survival, whilst others can be lost with little effect on survival rate. Perturbing the global regulator, *Rsp1*, also dramatically affects survival rate. Halving the protein stability (*HALF-STAB*) causes the network to die at any point, but mostly replicates. Reducing the protein production rate also has a large impact on survival, indicating the highly tuned state of the network

## Discussion

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

The evolution of increasingly complex gene regulatory mechanisms has also been observed in other in silico bacterial models, for instance, in evolving chemotaxis dynamics, simple mechanisms were observed in environments of constant stimuli, whereas under fluctuating stimuli environmental conditions more complex mechanisms were observed (Goldstein and Soyer 2008). This further implies a strong connection between environmental and network complexity. The incremental functional evolution observed during our experiments is also an exciting result. Randomly generated networks are generally non-viable due to the energetic cost of over-expression of one of the components and/or lethal stress levels. Therefore, the solution is to remove the energetic requirements, which can be achieved in a number of ways: (1) remove non-essential/non-functional genes, (2) reduce the basal expression rate or (3) regulate the expression of genes. It is evident that all three actions are utilised, as genome size very quickly reaches a small size, and many gene expression rates are also reduced. Global regulation of gene expression was a selected mechanism and the evolution of similar global regulatory structures was observed in many populations. The global regulation mechanism is a very energy efficient solution, requiring expression of only a single gene to regulate many. The relative ease, in living systems, of adaptive evolution of a binding site via point mutations to a specific transcription factor, in a reasonable evolutionary timescale (Berg et al. 2004), would further strengthen the selection of such a regulatory mechanism in the model. This energy efficiency, and ease of evolving new regulatory interactions, along with the similar structure observed in many populations, may be strong evidence for the adaptive selection of global

regulation mechanisms observed in many biological networks, in contrast to the non-adaptive mechanisms proposed (Cases and de Lorenzo 2005; Lynch 2007a, b). Once energy regulation is resolved, the models adapt to counter lethal stress levels, which are only encountered once energy regulation is in place. When both energy regulation and lethal stress levels are resolved, the next adaptation is for speed of growth. Growth systems were observed at multiple points during the network evolution, however, it is only once the 'core' systems are in place that the growth systems become fixed. Thus, a reasonable hypothesis is that, early in evolution the 'core' survival systems of energy regulation and detoxification might have evolved prior to efficient growth and replication systems.

Biological networks are often thought of consisting of modular, independent units. Indeed, other in silico experiments have found modularity to increase with network complexity (Kashtan and Alon 2005; Hintze and Adami 2008). However, the observed network structures, whilst displaying some clearly modular functional systems, were not independent with many cases of inter-connected systems, with overlapping modules sharing genes. Examining biological networks in more detail we see a similar inter-connected functional structure. For instance, the global regulator *CRP* regulates the central carbon metabolism of *E. coli*. Yet, it also regulates many other metabolic and stress response pathways, creating a centrally connected hub structure, rather than independent functional modules (Keseler et al. 2009). Although it is convenient to attempt to separate a biological network into smaller independent sub-graphs, such as the network motif approach, it is also possible to 'lose the bigger picture'. Such an approach may yield some dynamical or functional behaviour from a network, but without taking all other interactions and connections into account, not all behaviours will be identified. As such, we suggest that a true biological complexity measure should not only take structural information, such as modularity, into account, but also necessarily requires functional information, such as the functional systems approach taken in this study.

**Acknowledgements** The authors thank Francesco Falciani, Chrisantha Fernando, Richard Goldstein, Helen Parsons, Charles Penn and Jon Rowe for helpful discussion and comments. This work also benefitted from discussions enabled by the StoMP Network, BBSRC Grant Reference BBF0037651. DJJ was funded by BBSRC Strategic Research Studentship BBS/S/S/2005/12006. Computer simulations were carried out using the Birmingham Biosciences High Performance Compute Cluster, funded by BBSRC REI grant BB/D524624/1. Warwick Systems Biology Centre and the StoMP network funded the colour figure publication cost. Open Access publication was funded by The University of Nottingham Open Access Publishing Fund.



## References

Adami C, Ofria C, Collier TC (2000) Evolution of biological complexity. Proc Natl Acad Sci USA 97:4463–4468

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14:283–291

Babu MM, Teichmann SA, Aravind L (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 358:614–633

Banzhaf W, Kuo PD (2004) Network motifs in natural and artificial transcriptional regulatory networks. J Biol Phys Chem 4:85–92

Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Berg J, Willman S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. BMC Evol Biol 4:42

Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proc Natl Acad Sci USA 105:7899–7906

Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. Nature 440:358–362

Cases I, de Lorenzo V (2005) Promoters in the environment: transcriptional regulation in its natural context. Nat Rev Microbiol 3:105–118

Cordero OX, Hogeweg P (2006) Feed-forward loop circuits as a side effect of genome evolution. Mol Biol Evol 23:1931–1936

Crombach A, Hogeweg P (2008) Evolution of evolvability in gene regulatory networks. PLoS Comput Biol 4:e1000112

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

Goldstein RA, Soyer OS (2008) Evolution of taxis responses in virtual bacteria: Non-adaptive dynamics. PLoS Comput Biol 4:e1000084

Gottesman S (1984) Bacterial regulation: global regulatory networks. Ann Rev Genet 18:415–441

Hintze A, Adami C (2008) Evolution of complex modular biological networks. PLoS Comput Biol 4:e23

Holland J (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology control and artificial intelligence. MIT Press, Cambridge

Ingram PJ, Stumpf MPH, Stark J (2006) Network motifs: structure does not determine function. BMC Genomics 7:108

Jenkins DJ, Stekel DJ (2008) Effects of signalling on the evolution of gene regulatory networks. In: Bullock S, Noble J, Watson R, Bedau MA (eds) Artificial life XI – proceedings of the eleventh international conference on the simulation and synthesis of living systems. The MIT Press, Cambridge, MA, pp 289–296

Jenkins DJ, Stekel DJ (2010) Stochasticity versus determinism: consequences for realistic gene regulatory network modelling and evolution. J Mol Evol 70:215–231

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

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

Keseler IM, Bonavides-Martínez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, Perata-Gil M, Santos-Zvaleta A, Shearer AG, Karp PD (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. Nucleic Acids Res 37:D464–D470

Lenski RE, Ofria C, Pennock RT, Adami C (2003) The evolutionary origin of complex features. Nature 423:139–144

Lynch M (2007a) The evolution of genetic networks by non-adaptive processes. Nat Rev Genet 8:803–813

Lynch M (2007b) The frailty of adaptive hypothesis for the origins of organismal complexity. Proc Natl Acad Sci USA 108(suppl 1):S8597–S8604

Mangan S, Alon U (2003) Structure and function of the fee-forward loop network motif. Proc Natl Acad Sci USA 100:11980–11985

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

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

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

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

Ofria C, Wilke CO (2004) Avida: a software platform for research in computational evolutionary biology. Artif Life 10:191–229

Philippe N, Crozat E, Lenski RE, Schneider D (2007) Evolution of global regulatory networks during a long-term experiment with *Escherichia coli*. BioEssays 29:846–860

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

Sunderaraj S, Guo A, Habibi-Nazhad B, Rouani M, Stothard P, Ellison M, Wishart DS (2004) The CyberCell Database (CDDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. Nucleic Acids Res 32:D293–D295

Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. Nat Genet 36:492–496