On the Evolution of Scale-Free Topologies with a Gene Regulatory Network Model

Miguel Nicolau, Marc Schoenauer

On the Evolution of Scale-Free Topologies with a Gene Regulatory Network Model

Miguel Nicolau and Marc Schoenauer
Projet TAO - INRIA Saclay - Île-de-France
LRI - Université Paris-Sud, FRANCE
{Miguel.Nicolau,Marc.Schoenauer}@inria.fr

Abstract

A novel approach to generating scale-free network topologies is introduced, based on an existing artificial Gene Regulatory Network model. From this model, different interaction networks can be extracted, based on an activation threshold. By using an evolutionary computation approach, the model is allowed to evolve, in order to reach specific network statistical measures. The results obtained show that, when the model uses a duplication and divergence initialisation, such as seen in nature, the resulting regulation networks not only are closer in topology to scale-free networks, but also require only a few evolutionary cycles to achieve a satisfactory error value.

1 Introduction

Scale-Free networks are complex networks which have a few highly connected nodes, while most nodes are poorly connected (Barabási and Albert, 1999). More precisely, in such networks, the connectivity of the nodes follows a power law: the proportion $P(k)$ of nodes with degree k (i.e. that are connected to k other nodes) is roughly proportional to $k^{-\gamma}$, for a positive real number γ, at least above a given k value.

This network topology has been shown to exist in a variety of real-world, artificial and biological systems (Albert and Barabási, 2002; Wuchty, 2001; Jeong et al., 2000; Guelzim et al., 2002; van Noort et al., 2004; Babu et al., 2004), and has been widely studied because of its high resistance to random failures. Different generative models have been shown to create scale-free networks: in the original "preferential attachment" model, the network is built gradually, and new nodes attach preferentially to highly connected nodes (Barabási and Albert, 1999); the growing random network model (Krapivsky et al., 2000) extends this idea, by adapting the connection probability of a new node through a connection kernel. This topology can also occur as a consequence of optimization processes, such as the wiring cost to existing software components (see (Valverde et al., 2002) and references therein), and through evolutionary processes applied to cellular automata (Tomassini et al., 2004); finally, some artificial genome models, created through duplication and divergence, have been shown

to generate networks with a power-law degree distribution (Pastor-Satorras et al., 2003; Kuo and Banzhaf, 2004). However, all these models use rules that are not directly connected to the topology of the resulting network, and in particular do not offer an easy tuning of the statistical properties of the network they build. Using the last type of generative model – the generation of genomes through duplication and divergence – the current work investigates the possibility of designing scale-free networks with a given exponent for its power-law tail.

Genetic Regulatory Networks (GRNs) are biological interaction networks among the genes in a chromosome and the proteins they produce: each gene encodes a specific type of protein, and some of those, termed *Transcription Factors*, regulate (either enhance or inhibit) the expression of other genes, and hence the generation of the protein those genes encode. The study of such networks of interactions provides many interdisciplinary research opportunities, and as a result, GRNs have become an exciting and fast evolving field of research.

In order to study the characteristics of GRNs, many artificial systems have been designed, either through the modeling of biological data, or purely artificially; de Jong (2002) provides a relatively recent overview of this area of research.

One interesting research direction regarding the use of GRNs is the extraction and analysis (static or dynamic) of their regulation network. Previous work on the structural analysis of GRNs has provided many insights, of which the following are but a few examples. It has been shown that these networks can be grown through a process of duplication and divergence (Wolfe and Shields, 1997; Kellis et al., 2004); that they can exhibit scale-free and small-world topologies (Bhan et al., 2002; van Noort et al., 2004; Babu et al., 2004; Kuo et al., 2006); that some specific network motifs, resembling those identified by biologists as building-blocks, are present within these artificial networks (Wuchty et al., 2003; Milo et al., 2004); that different regulatory behaviours are detectable in these models, such as perfectly ordered, chaotic, or cyclic (Reil, 1999); that it is possible to apply evolutionary approaches leading to stable regulatory patterns, under different random starting conditions (Rohlf and Winkler, 2008); and that in response to diverse stimuli, the wiring of these networks changes over time, with a few transcription factors acting as permanent hubs, but most adapting their role as an answer to the changing environment (Luscombe et al., 2004).

The generation of specific network topologies allows their incorporation in a variety of systems. Although regular and random topologies have been shown to work well in several fields, such as parallel and distributed computing (Leighton, 1992) or simple automata (Garzon, 1995), other systems have been shown to profit from specific topologies, such as large cellular automata systems (Watts, 1999), and evolutionary algorithms of different classes (Giacobini et al., 2005, 2006; Payne and Eppstein, 2007, 2008). The objective of the current work is therefore to generate topologies that are to be tested in other optimisation algorithms, such as Echo-State Networks (Jaeger, 2001; Jiang et al., 2008).

The present work focuses on the analysis of the underlying network topologies of one artificial GRN model (Banzhaf, 2003), and of its use as a generative model for scale-free topologies. Both random genomes and genomes initialised through a duplication and divergence method are first analysed with respect to statistical properties of the topology of the resulting interaction network. Then, the inverse problem

is addressed: an Evolutionary Algorithm is used to evolve artificial GRNs so that the topology of the resulting network has specific statistical properties – more precisely, a scale-free topology with a given exponent. The results obtained show that genomes created through duplication and divergence are better suited for evolution, and generate networks exhibiting power-law tails, a clear sign of a scale-free topology.

This paper is structured as follows: Section 2 presents the GRN model used in the simulations, including the description and analysis of the duplication/divergence process used to initialise the genomes. Section 3 introduces the statistical tools used to assess the scale-free properties of the networks, along with the techniques to actually compute them. Section 4 describes the experimental setup, the error measure and the results obtained when evolving GRNs to obtain scale-free network topologies. Section 5 then analyses the reasons behind the success of the initialisation procedure in generating suitable topologies, and finally Section 6 discusses those results and sketches some hints for future research directions.

2 The GRN Model

2.1 Representation and dynamics

The artificial model described here is that proposed by Banzhaf (2003). It is built over a genome represented as a bit string, and assumes that each gene produces a single protein, with all proteins regulating all genes (including the gene that produced it).

A gene is identified within the genome by an *Activator* (or *Promoter*) site, that consists of an arbitrarily selected bit pattern: in this work, a 32 bits sequence whose last 8 bits are the pattern `01010101`.

The 160 (5×32) bits immediately following a promoter sequence represent the gene itself, and are used to determine type of the protein this gene produces. This protein (like all proteins within the model) is a 32 bit sequence, resulting from a many-to-one mapping of the 160-bits gene sequence: each of the 32 bits of the protein results from the application of a majority rule for each of the five sets of 32 bits taken from the 5×32 bits of the gene (see Fig. 1).

Upstream from the promoter site are two additional 32 bit segments, representing the *Enhancer* and *Inhibitor* sites: these are used for the regulation of the protein production of the associated gene.

The binding of proteins to the enhancer or inhibitor sites is calculated through the use of the `XOR` operation, which returns the degree of match as the number of bits set to one (that is, the number of complementary bits in both bit patterns). In general, a Normal distribution results from measuring the match between proteins and these sites, in a randomly generated genome (Banzhaf, 2003).

The enhancing and inhibiting signals regulating the production of protein p_i are then calculated as:

$$e_i, h_i = \frac{1}{N} \sum_{j=1}^{N} c_j \exp(\beta(u_{i,j} - u_{i,max})) \tag{1}$$

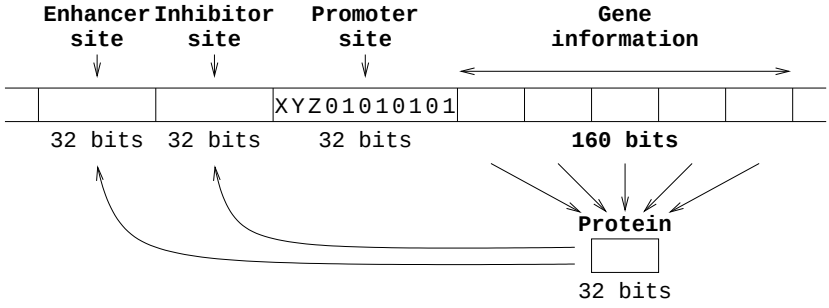

Figure 1: Bit string encoding of a gene. If a promoter site is found, the gene information is used to create a protein, whose quantity is regulated by the attachment of proteins to the enhancer and inhibitor sites.

where N is the number of existing proteins, c_j is the concentration of protein j, $u_{i,j}$ is the number of matching bits between the regulating site of gene i and protein j, $u_{i,max}$ is the maximum match achieved for gene i, and β is a positive scaling factor.

Given these signals, the production of protein i is calculated via the following differential equation:

$$\frac{dc_i}{dt} = \delta(e_i - h_i)c_i - \Phi \tag{2}$$

where δ is a positive scaling factor (representing a time unit), and Φ is a term that proportionally scales protein production, ensuring that $\sum_i c_i = 1$, which results in competition between binding sites for proteins.

Note that this model makes some simplifying hypotheses regarding some of the known characteristics of the biological regulatory process: all proteins are assumed to be *Transcription Factors*, i.e., all proteins are used to regulate the expression of all genes: in other words, the model is a closed world. Also, the model uses only one enhancing and inhibiting site per gene. However, it captures interesting properties of actual GRNs, in particular through the genome construction technique. One should nevertheless be careful when translating to real GRNs the results that are obtained using this model.

2.2 Genome Construction

The technique of duplication and mutation proposed (Banzhaf, 2003) consists in creating a random 32 bit sequence, followed by a series of length duplications associated with a (typically low) mutation rate. It has been shown (Wolfe and Shields, 1997; Kellis et al., 2004) that such evolution through genome duplication and subsequent divergence (mostly deletion) and specialisation occurs in nature.

Number of genes

An analysis of the resulting number of genes in a genome was first presented by Kuo and Banzhaf (2004). A similar technique was used here to investigate the influence of the initialisation mutation rate on the number of genes per genome: 1000 genomes were created using 14 duplication and divergence events, giving a genome length of $L_G = 32 \times 2^{14} = 524288$. The histogram for the resulting numbers of genes is shown in Fig. 2: if little or no mutation is used, a large proportion of genomes have no genes at all, but a few genomes have a large amount of genes. This was to be expected: if the original random sequence contains the promoter pattern, or if it appears early in the sequence of duplications thanks to a lucky mutation, then a large number of genes will be created by the duplication process. Otherwise, little or no genes will be created.

When the mutation rate increases, the number of genes rapidly converges towards a stable average range: with rates higher than 15%, the duplication technique becomes sufficiently randomised to roughly lead to the same number of genes per genome (around 900 here) as if using randomised genome bit-strings (or, equivalently, if using a mutation rate of 50% with the duplication/divergence process), as the genome sequence becomes randomised after just a few duplication steps.

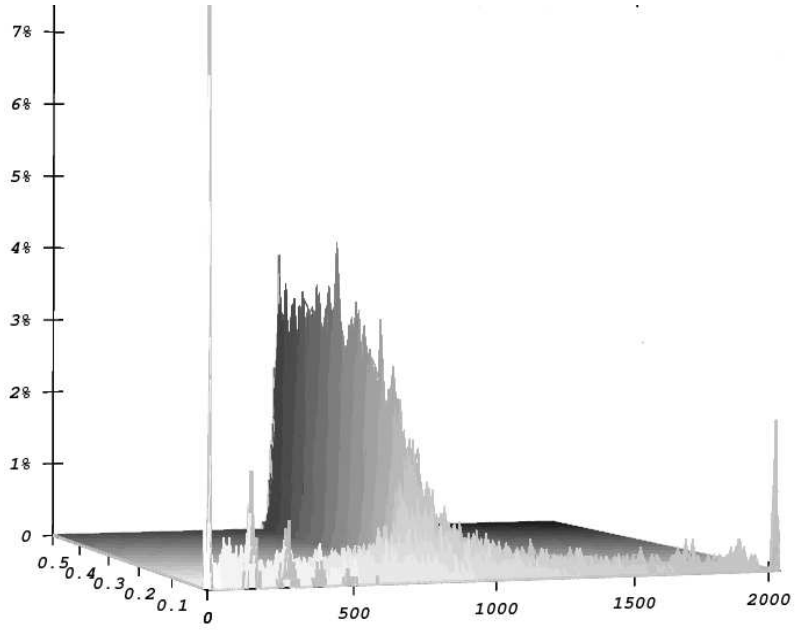

Figure 2: Histogram of the number of genes per genome, computed over 1000 genomes per mutation rate. The x-axis plots the number of genes, the y-axis (depth) the mutation rate, and the z-axis (height) the percentage of genomes having this particular number of genes. Genes were not allowed to overlap.

2.3 GRN Topologies

As seen before, all proteins within the model regulate the expression of all genes. The strength of this regulation is determined by the binary match between the protein pattern and the regulating sites of the destination gene (Eq. 1).

The resulting network of gene interactions can be drawn as a directed graph, with vertices connecting genes producing transcription factors to the genes they regulate (Banzhaf, 2003). As all genes produce transcription factors, the graph of the resulting interaction network is a complete graph, where all nodes are linked together. However, because of the exponential nature of the interactions given by Eq. 1, small interactions will have almost no effect on the production of a given protein. It is hence natural to establish a minimum matching strength (*threshold*) and to remove weaker regulation relationships.

Moreover, by using different thresholds, different network topologies can be obtained. For instance, Figs. 3 and 4 show the graphs of the same completely random genome for two slightly different thresholds (respectively 23 and 24). While almost all nodes are still connected on Fig. 3, increasing the threshold by one removes many connections, and the graph on Fig. 4 is only a small sub-graph of the previous one (nodes which become isolated are not shown, which explains the smaller number of genes). Note also how the increase of the threshold creates unconnected (independent) sub-graphs.

A completely different picture is that of genomes initialised through the duplication/divergence process (hereafter called DM-genomes), described in Section 2.2. Fig. 5 is an example of the topology of the interaction graph for such a genome, initialized with 1% mutation rate, using 16 as the connection threshold.

The use of a low mutation rate results in a much shallower hierarchy of nodes, with a few master genes being connected to most of the other genes, regulating and/or being regulated by them. Varying the threshold results in networks with similar dynamics: Figs. 6 and 7 depict the same DM-genome, with higher connectivity thresholds (17 and 18, respectively). The presence of master genes is still clear, but their connectivity is obviously lower. Note also how some master genes disappear if the threshold parameter is increased.

Another observation is that the thresholds generating "interesting" topologies for randomly created genomes are higher than those for genomes created with duplication and low mutation. This is because the latter exhibit a high degree of similarity in their bit patterns, leading to a lower value of $u_{i,j}$, when applying the `XOR` operator (see Equation 1).

2.4 Connectivity variance

In order to generalize the observations made on the graphs above, an approach similar to that of Kuo et al. (2006) has been used here to analyse the relationship between the number of edges and the threshold: 100 genomes have been generated, using 14 duplication events, and the network connectivity (fraction of edges) has been computed for each threshold.

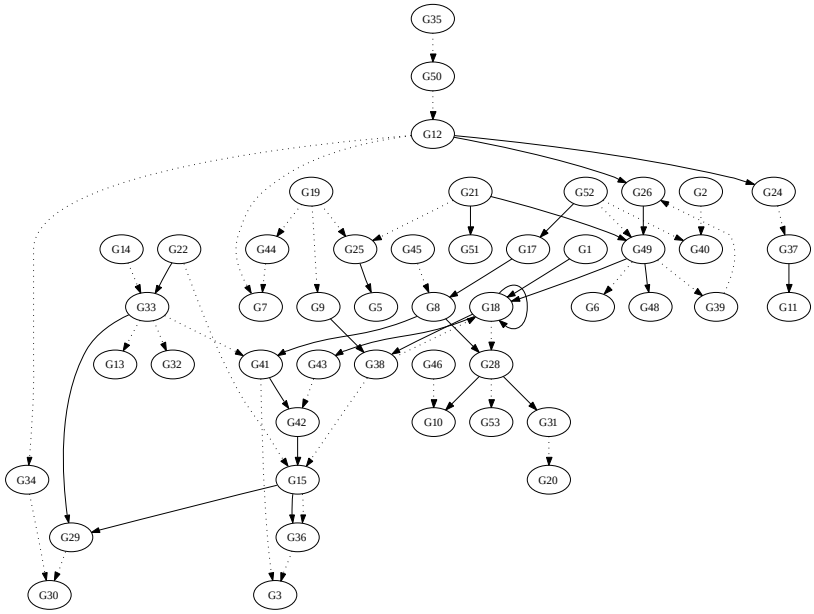

Figure 3: Gene regulatory network for a random genome of length $L_G = 32768$, i.e. created using 10 duplication events and a mutation rate of 50%, at a threshold of 23 bits. Solid edges indicate enhancing interactions, dotted edges indicate inhibiting interactions.

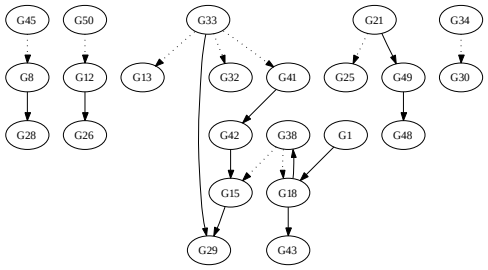

Figure 4: Gene regulatory network for the same genome as in Fig. 3, at a threshold of 24 bits.

The network connectivity is defined as:

$$NC = \frac{\#edges}{2n^2} \tag{3}$$

where $\#edges$ is the number of edges in the network, and n is the number of nodes, or genes ($2n^2$ is hence the maximum number of possible edges, as each node can be connected twice to any other node, including itself).

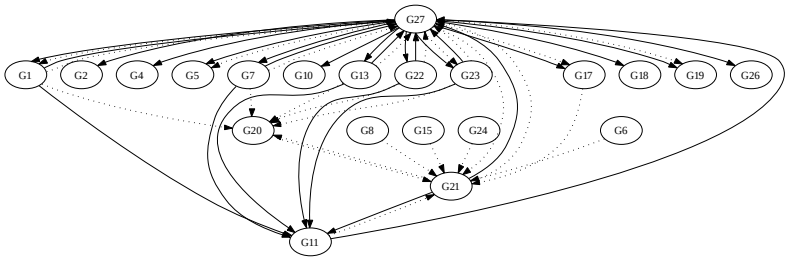

Figure 5: Gene regulatory network for a genome of length $L_G = 32768$, created using 10 duplication events and a mutation rate of 1%, at a threshold of 16 bits.

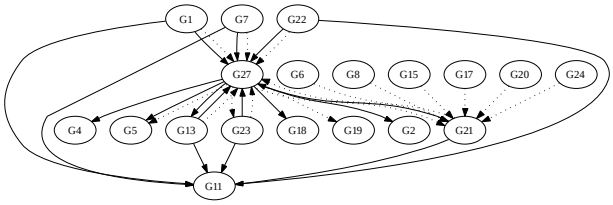

Figure 6: Gene regulatory network for the same genome as in Fig 5, at a threshold of 17 bits.

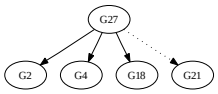

Figure 7: Gene regulatory network for the same genome as in Fig 5, at a threshold of 18 bits.

Fig. 8 shows the connectivity as a function of the threshold, for mutation rates of 1%, 5%, 10%, and 50%. It is a clear illustration of the very different behaviors with respect to connectivity depending on the mutation rate used during the duplication/divergence process:

- A high mutation rate (or, equivalently, the completely random generation of the genome) creates a network which stays fully connected with a wide range of threshold values; then, there is a sharp transition from full connectivity to no connectivity (see also Fig. 9). Moreover, there is a very small variance between different networks.
- A low mutation rate creates a network which quickly loses full connectivity; however, its transition from full connectivity to no connectivity is much smoother than that of random networks. Moreover, there is very large variance between different networks generated with the same mutation rate.

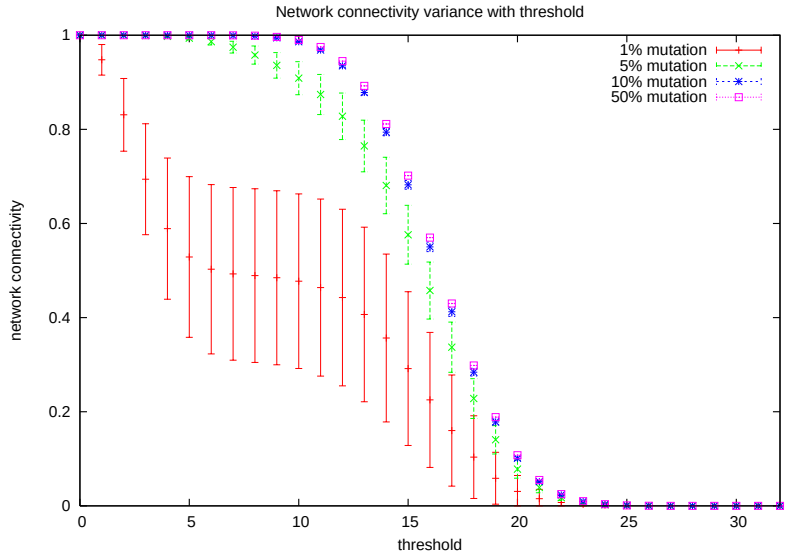

Figure 8: Fraction of edges in a graph as compared to a fully connected network (and standard deviations), versus threshold parameter, based on samples of 100 genomes, created using 14 duplication events, and mutation rates of 1%,5%, 10%, and 50%.

3 Scale-free Topologies

Even though the model used is overly simplified compared to what is known of biological GRNs (as discussed in Section 2.1), an interesting issue is to find out whether or not the resulting interaction network exhibits particular properties resembling those found in natural networks, such as being Scale-Free (Wuchty, 2001; Jeong et al., 2000; Guelzim et al., 2002; van Noort et al., 2004; Babu et al., 2004), at least for certain values of the threshold used to prune the graph of connections. A characteristic feature of Scale-Free graphs is that the distribution of the degrees approximately follows a power law. However, assessing such a distribution is not as obvious as it seems.

3.1 Measurement of Power Laws

Given a sample of a quantity, the typical method for measuring whether or not this quantity follows a power law consists in measuring whether the histogram of the sampled quantity at hand is roughly linear on logarithmic scales. A linear regression (using e.g. any *Least-Squares* method) can be used, and the slope of the best linear approximation will be the exponent $-\gamma$ of the power-law. This method, however, has been shown to introduce systematic biases into the value of the exponent (Goldstein et al., 2004; Newman, 2005), because quite often, the tail of power-law distributions tends to be noisy, due to sampling errors: this arises from the fact that very few samples

exist towards the high end of the distribution. This is certainly the case with the vertex degree distributions analysed here.

Another option is to work directly on the sample itself, rather than on the logarithms, and to use a non-linear curve-fitting method, such as the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963). In this case, however, the difficulty lies in choosing the correct parameters for the optimization method, and taking into account all points of the histogram, despite their very different orders of magnitude.

To tackle this problem, a technique known as *logarithmic binning* can be used (Albert et al., 2000). It smoothes the histogram by grouping the distribution data per ranges of k values with exponentially increasing sizes (e.g. 1, $[2, 3]$, $[4, 7]$, $\ldots$). This technique, combined with the linear least-squares approximation, tends to give good results (Albert et al., 2000), and is the technique of choice for this work.

Finally, another question arises as to where to measure the power-law behaviour; indeed, it has been shown that many distributions follow a power-law only in the tail (Newman, 2005; Clauset et al., 2007). After observation of a few samples, it was noted that the topologies generated by the current technique also tend to produce a power-law behaviour only in the tail, and therefore the error function was adapted to optimise this behaviour (see Section 3.2). The measurement adopted is quite simple: the tail of a distribution starts when the proportion of nodes having a certain number of connections is lower than the preceding one (Fig 9 and 10 provide examples).

3.2 Are GRNs Topologies Scale-Free?

Random genomes are, in terms of degree distribution, highly regular, in that their degree distribution is highly peaked; this in turn leads to potentially misleading good γ values (linear regression of 2 points is always perfect!). This can be seen in Fig. 9, which shows an example network extracted from a random genome. The vertex degree distribution is clearly Gaussian, even when plotted in a log/log graph; however, a least-squares regression gives the value $\gamma = 3.25$. Using logarithmic binning does not help: due to the proximity of all values, there are only three points left in the distribution tail, leading to a good γ value, but with a high approximation error.

Networks extracted from DM-genomes give a completely different picture, as seen in Fig. 10. The initial distribution has a clear linear trend in a log/log scale, but is affected by noise towards the end of the distribution, and by initial low proportion values; by using logarithmic binning, the values towards the end are grouped and therefore smoothed, whereas the initial values of the distribution are discarded after detection of the distribution tail.

The occurrence of misleading γ values with random genomes can be further observed in Fig. 11: the number of logarithmic bins with random genomes is much smaller, making the task of measuring the scale-free behaviour of the resulting distribution very difficult and error prone. DM-genomes, on the other hand, give a wider spread of distribution sizes, with γ values typically in the range $[1, 2]$.

Though some graphs built from the artificial GRNs considered here exhibit some characteristics of scale-free networks (Kuo and Banzhaf, 2004) when the described initialisation process is applied, their degree distribution is generally quite far from a true power law. Nevertheless, while random graphs, because of the poor spread of

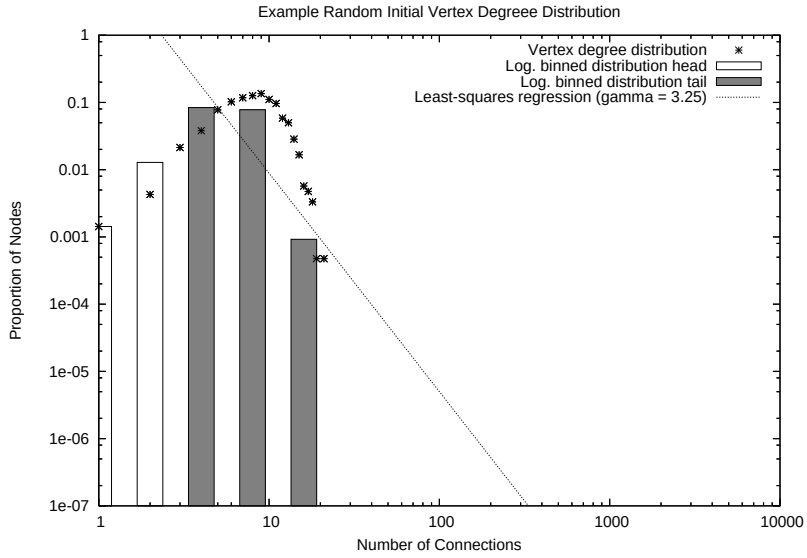

Figure 9: Vertex degree distribution for the best network of a random genome, before (stars) and after (vertical bars) logarithmic binning. Grey vertical bars represent the tail of the distribution, and the line drawn represents the least-squares regression calculated using the distribution tail.

their degree distribution, seem to be difficult to modify toward more scale-free topologies, graphs that have been initialized through the duplication-divergence mechanism are more promising as seed topologies for the evolution of scale-free topologies. Next section demonstrates that evolving networks created with the duplication/divergence process described is indeed possible, resulting in yet another method to construct networks with scale-free properties.

4 Evolution of Topologies

The objective of this section is to evolve GRN genomes, using a simple evolutionary algorithm, so that the resulting interaction network gets as close as possible to a scale-free topology with given exponent γ value. In these experiments, genomes of length $L_G = 32 \times 2^{15} = 1048576$ bits were used, i.e. obtained using 15 duplication steps. Furthermore, both random genomes and DM-genomes are used; to keep the comparison fair, in both cases, only genomes with network sizes between 2000 and 3000 were considered, the others being simply discarded.

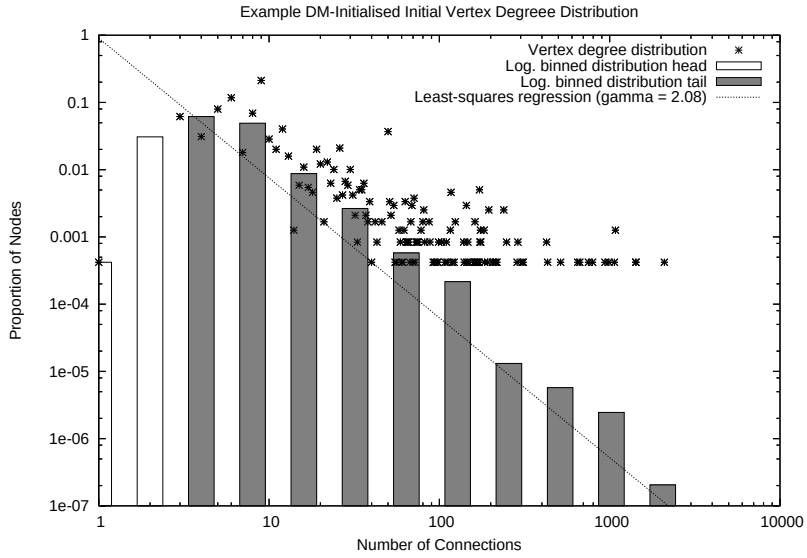

Figure 10: Vertex degree distribution for the best network of a DM-genome, before (stars) and after (vertical bars) logarithmic binning. Grey vertical bars represent the tail of the distribution, and the line drawn represents the least-squares regression calculated using the distribution tail.

4.1 The Evolutionary Algorithm

The Evolutionary Algorithm that was used to evolve populations of bit-string genomes, such as the ones described in Section 2.1, is base on only one variation operator, the standard bit-flip mutation. Furthermore, this algorithm uses a straightforward $(25 + 25) - ES$ Evolution Engine, i.e., 25 parents give birth to 25 offspring, and the best 25 of the 50 parents+offspring become the parents of next generation. The only tricky part lies in the adaptive way to modify the mutation rate along evolution: its rate per bit is initially set to 1%, and is adapted in a way that is similar to the well-known $1/5$ rule of Evolution Strategies (Rechenberg, 1994): when the rate of successful mutations is higher than $1/5$ (i.e. when more than 20% mutation events result in a reduction of the error measure), the mutation rate is doubled; it is halved in the opposite case[1].

In order to compare the initialisation method presented in Section 2.2 (with mutation rate 1%) and completely random populations (or, equivalently, populations built with the same method and mutation rate 50%), 50 independent runs of 50 generations were performed with each of those initialisation procedures.

[1]Note that, because of the possibility of neutral mutations (especially with low mutation rates), if there were more than 50% neutral mutations, the rate was doubled, too.

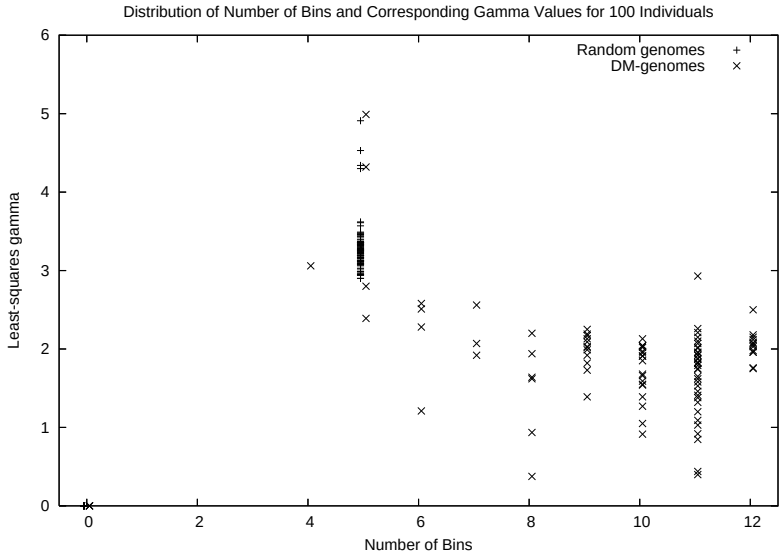

Figure 11: Number of logarithmic bins and corresponding γ values, for random (+) and DM (x) genomes, based on a sample of 100 genomes (best threshold found).

4.2 Error Function

Cohen and Havlin (2004) have shown that a large proportion of networks displaying scale-free behaviour exhibit values of $\gamma \in [2, 3]$, with some emphasis on the central value. In this work, a narrow interval around 2.5 was used, and values of γ in $[2.4, 2.6]$ were considered ideal. The least-squares method on the logarithmic binned distributions was used to compute an estimation of γ as described in Section 3.1. The error function (to minimise) was therefore:

$$F(x) = \left\{ \begin{array}{ll} 2.4 - \gamma & \text{if } \gamma < 2.4 \\ 0 & \text{if } 2.4 \leq \gamma \leq 2.6 \\ \gamma - 2.6 & \text{if } \gamma > 2.6 \end{array} \right\} + \sigma + \frac{1}{n} \tag{4}$$

The mean squared error (σ) between the least-squares regression and the actual distribution was added to the absolute difference to the target γ values, as an estimate of the quality of the measurement. The number (n) of points in the logarithmic binned vertex degree distribution tail was used as a penalisation of "regular" distributions where only a few data points remain after the logarithmic binning (as seen in Section 3.2), and as an incentive for the evolution of more diverse distributions.

From each GRN individual, several networks were extracted, by varying the threshold value; only the threshold giving the lowest error was kept.

4.3 Experimental Results

Fig. 12 shows the best (lowest) error in the population, averaged over the 50 runs, for both random and DM-genomes.

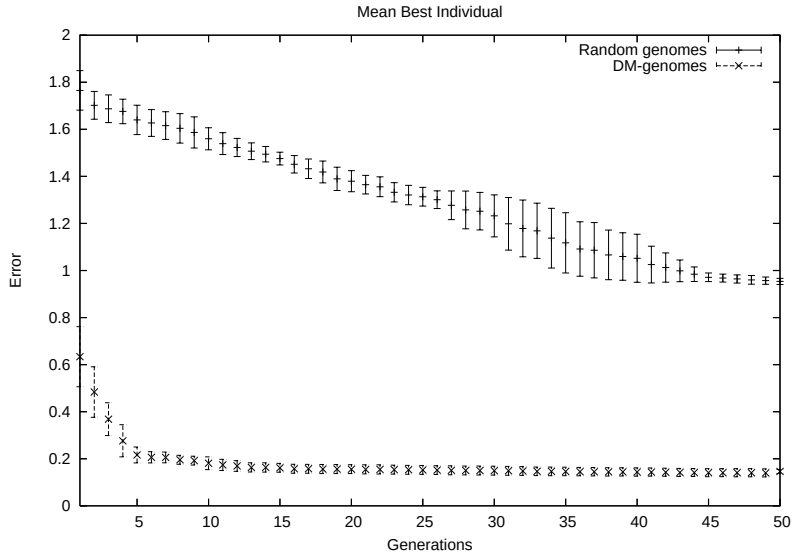

Figure 12: Average lowest error value per generation across 50 independent runs, for random genomes and 1% DM-genomes. Error bars plot the standard deviation across runs.

As can be seen from the figure, not only do DM-genomes start with a much smaller error, but they also converge within just a few iterations to a very low value, with small variance between runs. Random genomes, on the other hand, start with a worse error value, and although this is improved over time, it never reaches the same level as DM-genomes. In fact, random genomes were allowed to evolve for an extra 50 generation (making it 100 generations in total), and yet the results remained quite poor; the example on Fig 15 is extracted from a genome evolved for 100 generations.

One of the reasons for the success of DM-genomes lies in the wider span of their initial vertex degree distribution, which after logarithmic binning still keeps many values. This, along with the fact that a larger set of potentially fit networks can be extracted from each genome (as per Fig. 8), creates potential for evolution, through changes to the number of connections of each node. Random genomes, on the other hand, have their vertex degree distribution concentrated around a very small value range for their best threshold values (as in Fig. 15); this means that it is very hard to change the binned distribution in a meaningful way, through small gene connectivity changes.

Another reason for the greater efficiency of the use of DM-genomes as initial population for evolution is their ability to minimise the error value by varying the size of the genomes. This is illustrated on Fig. 13, which shows that random genomes

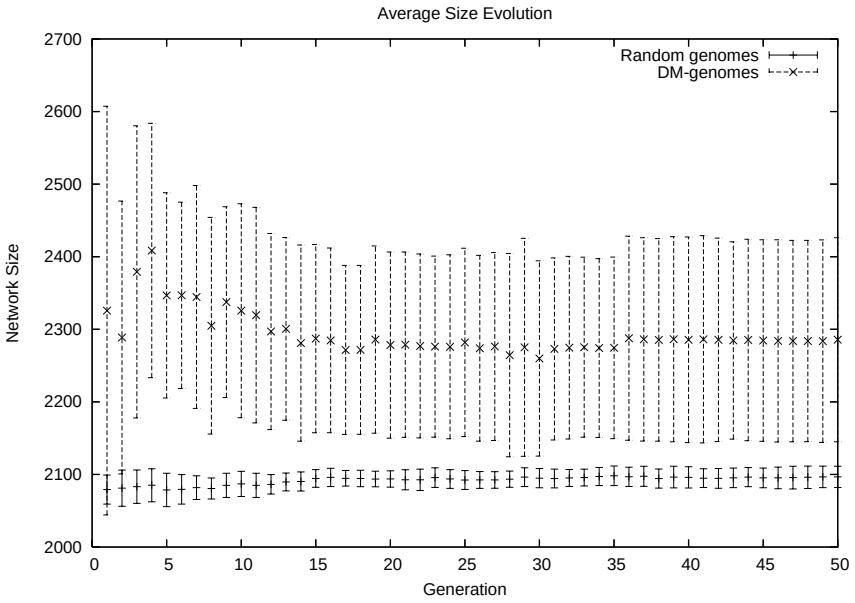

Figure 13: Average population genome size across 50 independent runs, for random genomes and 1% DM-genomes. Error bars plot the standard deviation.

keep roughly the same size for all genomes in the population across evolution, with very small variance across runs; DM-genomes, on the other hand, vary their size much more, with a much higher variance across runs. Even though a mutation event was equally likely to add or remove a gene during evolution (by creating or deleting the 01010101 promoter pattern somewhere on the genome), such operations rarely reduced the error value for random genomes, because of the small number of sample points that remained after binning for random genomes. These findings correlate well with the results already seen in Fig. 2.

The difference in terms of evolution potential with regard to scale-freeness can further be seen in Fig. 14, which plots the mutation rate across time (adapted with the $1/5$ rule, explained in Section 4.1), averaged across the 50 runs. It shows that networks based on DM-genomes are more resilient, in that they accept higher mutation rates and yet progress on the error landscape, whereas random genomes require very small mutation rates, which result in very small improvements over time. The difference between the averaged mutation rates is very large: it reaches a mean value of 16% at generation 4 with DM-genomes, whereas it only reaches 3% with random genomes, at generation 2.

Fig. 15 shows an example of an evolved network, extracted from a random genome. It clearly shows that the reduction of error was due to finding a suitable γ value (2.48 in this example), and by reducing the least-squares regression error (which however stays fairly high). It can be seen that the distribution still exhibits a Gaussian behaviour,

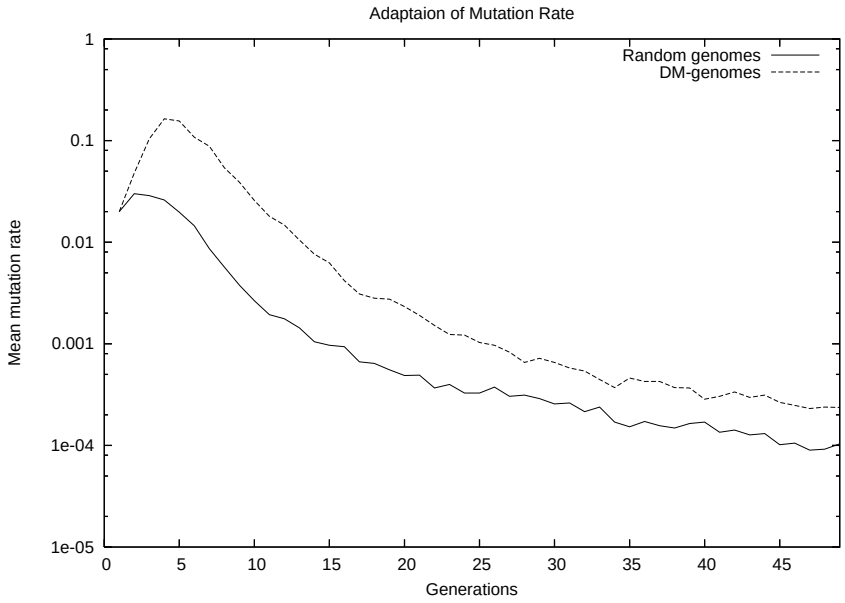

Figure 14: Evolution of mutation rate for random genomes and 1% DM-genomes, averaged across 50 independent runs.

with most distribution points very close to each other, resulting in very few logarithmic binned points.

This is in stark contrast with Fig. 16, which shows two examples of typical evolved networks extracted from DM-genomes, in a log/log plot, after binning. It can be seen that not all points follow a perfect line, but the distribution clearly has a power-law tail. Similar plots were obtained for most evolved networks.

4.4 Other topologies

Although a value of $\gamma = 2.5$ is a typical value for scale-free networks, one does not need to rely solely on it. In fact, experiments with other γ values yield equally satisfying results; Fig. 17 shows example networks evolved with target γ values of 1.5 and 2.0. Obviously, the further the target γ values are from the average values obtained just after initialisation (around 2.0, as seen in Fig. 11), the slower evolution is.

The method proposed is also able to evolve other similar topologies, such as small-world (Watts, 1999), as was shown recently (Nicolau and Schoenauer, 2009). The methodology is the same, except that different statistical measurements are targeted.

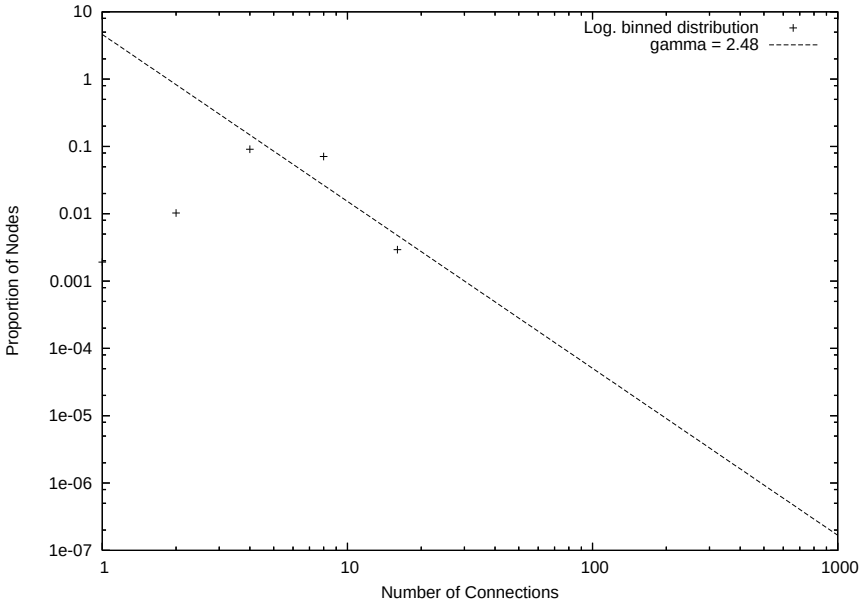

Figure 15: Example network extracted from the best genome, after evolutionary process based on random genomes. Vertex degree distribution was logarithmically binned, and plotted on a log/log scale. The least-squares regression of the (tail of the) binned distribution is also plotted.

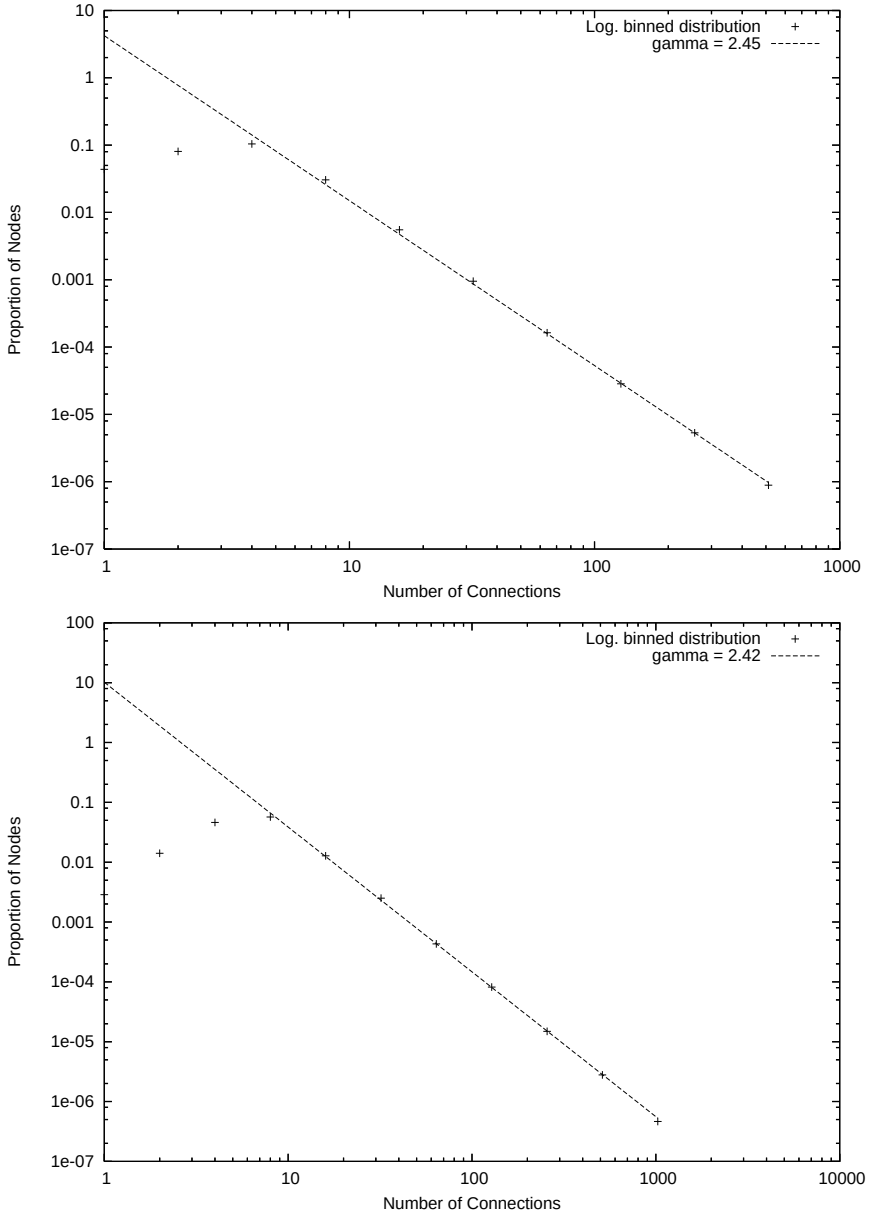

Figure 16: Example networks extracted from best genomes, after evolutionary process based on DM-genomes. Vertex degree distribution was logarithmically binned, and plotted on a log/log scale. The least-squares regression of the (tail of the) binned distribution is also plotted.

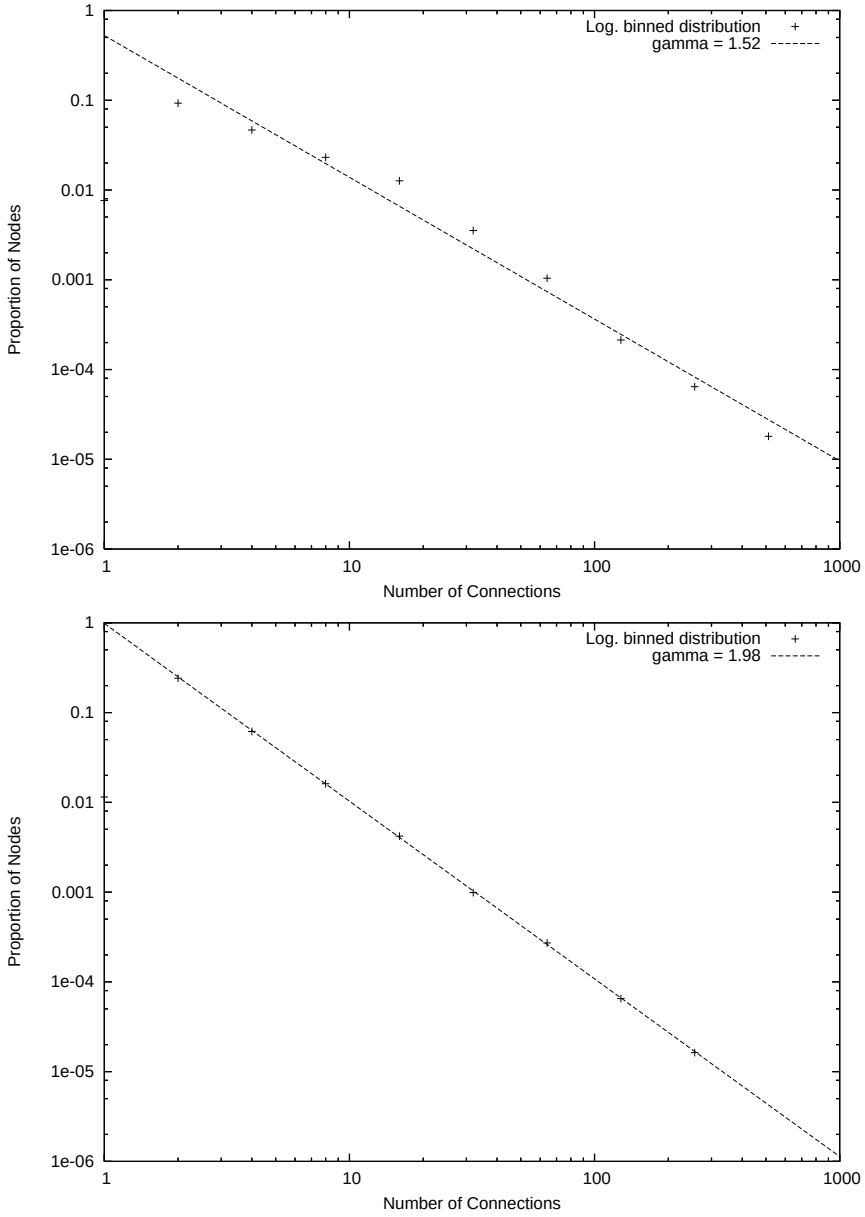

Figure 17: Example networks extracted from best genomes targeting values of $\gamma = 1.5$ and 2.0 respectively, after evolutionary process based on DM-genomes. Vertex degree distribution was logarithmically binned, and plotted on a log/log scale. The least-squares regression of the (tail of the) binned distribution is also plotted.

5 Analysis

The shallow hierarchies observed in DM-genomes exhibit characteristics similar to those of scale-free topologies, leading to the results observed. To analyse the reasons leading to such a difference in the extracted network topologies, a sequence of duplication/mutation steps (DM events) was analysed (as explained in Section 2.2), as it took place.

The original random 32 bit sequence was as follows:

$$10001011110000111111011110110101$$

This sequence was then subjected to a series of DM events, with a probability of mutation of 1% per bit. After 6 DM events, the first gene appeared, and after 7 events, there are already four genes. The resulting networks were extracted (Figs. 18 and 19) using 13 as the connection threshold[2].

Figure 18: Gene network after 6 duplication events.

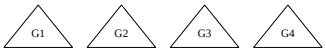

Figure 19: Gene network after 7 duplication events.

The starting location of a gene, when it is mapped to the original 32 bit sequence, determines the shape that is used to represent it in the following. For example, for the genome in Fig. 19, the starting locations for its genes were bits 905, 1929, 2377 and 2761, respectively. If we divide these by 32 and take the remainder, we see that they all start at the 9^{th} bit of a duplication of the original sequence, so they are all represented by the same (triangular) shape.

This also explains why there are no connections between genes in Fig. 19. As all genes originate from the same initial sequence of bits, the few mutations that occurred during the 7 DM steps did not create enough differences between regulating sites and produced proteins, to trigger a connection at threshold 13.

After the 8^{th} DM event, the network takes on a different topology (Fig. 20). Most genes are still duplications of the 9^{th} bit of the original sequence; however, **G7** starts at a different location, and is thus represented by a different (rectangular) shape.

As the connectivity between genes is established by the difference between regulation sites and proteins (see Section 2.1), genes originating from different locations are

[2]This value was chosen deliberately, based on the resulting network after all DM events, to illustrate the discussed process.

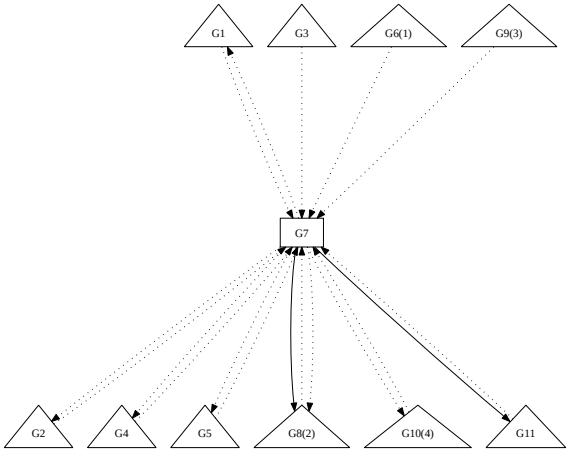

Figure 20: Gene network after 8 duplication events.

more likely to be connected, even using lower threshold values. This can be seen in Fig. 20: genes labelled with equal shapes do not connect to each other.

In this DM step one can also see *pure* duplications of genes, that is, genes that are created as duplications of other genes appearing upstream in the genome sequence: in those cases, the genes are labelled with their originating gene between brackets (e.g. **G6(1)**). But even *pure* duplications can generate slightly different genes, because mutation events can occur during the duplication process. **G6(1)** is an example: it only has an outward inhibiting connection to **G7**, whereas **G1** also has an inward inhibiting connection originating from the same gene.

With 9 DM steps, the network becomes a lot more complex (Fig. 21). There are still only two relative gene origins (triangles and rectangles), but either through pure duplications or discovery of new genes, there are now 25 genes.

One can see that triangles still connect only with rectangles (due to the threshold value chosen). Therefore, since there are a lot more triangles than rectangles, the latter become highly connected, and can be seen acting as *connection hubs*.

Finally, a last DM step is performed (Fig. 22), creating a network with 50 genes. Although hard to analyse for the naked eye, one can clearly see its shallow hierarchy, with a few highly connected nodes, to which most other nodes connect. One can also see the appearance of a third type of gene, labelled with a pentagon shape, which becomes the most connected gene. Table 1 shows a list of the gene families, along with their count, initial location, corresponding initial bit, and average number of inward, outward, and total connections.

Although this network is just an example, many networks were found to follow

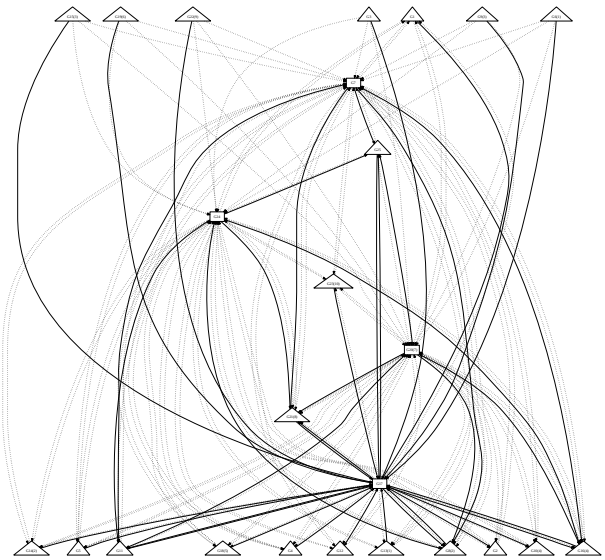

Figure 21: Gene network after 9 duplication events.

Table 1: List of all genes after 10 duplication events.

Family	# genes	1^{st} loc.	1^{st} seq. bit	Avg. in	Avg. out	Avg. total
Triangle	39	905	9	9	8.8	17.8
Quadrangle	10	5713	17	33	31.6	64.6
Pentagon	1	27872	0	37	59	96

the same mechanics while being extracted from genomes grown with DM steps. It shows that the tendency of DM-genomes to generate shallow hierarchies comes from the fact that genes starting at the same bit from the duplicated initial sequence tend not to connect, due to the use of the XOR operator (see Section 2). As duplications of the first gene(s) represent the majority of the genes present in the genome, they will not be connected (when choosing an appropriate threshold value), and genes discovered in later DM steps (in smaller numbers) will be highly connected to those ealier genes.

6 Conclusions

The experiments presented in this paper demonstrate that it is possible to evolve some networks so they approach a scale-free topology with a given exponent, by minimising an error measure that is directly connected to the topological property of the network, as opposed to the more classical generative methods where the scale-free prop-

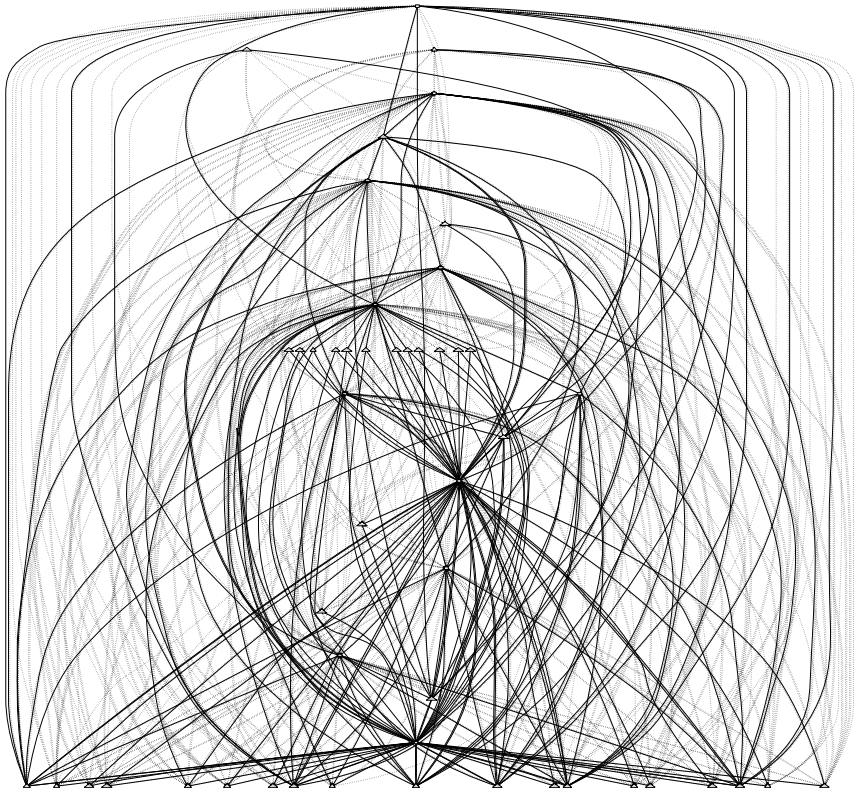

Figure 22: Gene network after 10 duplication events.

erty emerges from the rules that are used (or known to be used) to build the scale-free networks both in the biological and the artificial world. The long term result of such research can be to design a methodology for building artificial networks with precisely specified characteristics – motivated by known properties related to such statistical characteristics, e.g. the high resistance to random failure of scale-free networks.

The results presented in this paper also show that genomes created using the duplication and divergence method (with small mutation rate) described in the artificial GRN model proposed by Banzhaf (2003) can be used as starting points to generate network topologies that are typical of scale-free networks. Indeed, these initialised genomes are far better suited for evolution than purely random networks, due to the larger range of degrees in the networks they encode, as well as to the wider choice of resulting networks they can provide by varying the threshold parameter that decides of the existence of an edge between nodes.

There are still a few issues that need to be addressed with the current approach. The use of logarithmic binning, for example, results in a distribution with a small number of points. A possible solution to this problem can be to use overlapping bins, in order to

artificially increase the size of the sample. This is however a custom approach, and was not used in order to keep the methodology as standard as possible. Another possibility could be to use bin sizes that increase less rapidly (for example, increasing bin sizes by a factor of 1.5). But again, that would be a measure that is not standard.

More generally, much larger networks should be built to assess the statistical properties with more confidence. However, whereas it is not a problem to do more duplications in the initial phase of duplication/divergence, the issue when tackling larger networks will rapidly be that of CPU time: with the current size of network (between 2000 and 3000 nodes), a single evaluation takes approximately 5 minutes of a recent Pentium computer (3.6GHz) for random networks, and 8 minutes for duplication/divergence initialized networks, due to the higher number of threshold values that need to be checked for power-law distribution – and the main source of computational cost is the need to try several thresholds per genome. A possible solution might be to devise a heuristic in order to only evaluate promising threshold values. Another possible extension of this work would be to use localised mutations at gene encoding sections of the genomes only (or, equivalently, to remove all non-coding parts of the genome). While this will potentially increase the speed of evolution (by removing most neutral mutations), it will also remove the potential to add (or remove) genes. Though the number of genes did not vary greatly during the experiments presented here (Section 4.3), the influence of fixing the number of genes remains to be studied in more detail. Finally, a potential solution might be to devise a heuristic method which will replace using the full statistical analysis of each topology produced by a threshold.

Acknowledgment

The authors would like to thank Wolfgang Banzhaf for his helpful suggestions. This work was supported by the Sixth European Research Framework (proposal number 034952, GENNETEC project).

References

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406:378–382.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14:283–292.

Banzhaf, W. (2003). Artificial regulatory networks and genetic programming. In Riolo, R. and Worzel, B., editors, *Genetic Programming Theory and Practice*, chapter 4, pages 43–62. Kluwer Publishers, Boston, MA, USA.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Bhan, A., Galas, D. J., and Dewey, T. G. (2002). A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2007). Power-law distributions in empirical data. *arXiv:0706.1062 (E-print)*.

Cohen, R. and Havlin, S. (2004). Scale-free networks are ultrasmall. *Physical Review Letters*, 90(5):058701.1–058701.4.

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

Garzon, M. (1995). *Models of massive parallelism: analysis of cellular automata and neural networks*. Springer-Verlag, London, UK.

Giacobini, M., Preuss, M., and Tomassini, M. (2006). Effects of scale-free and small-world topologies on binary coded self-adaptive cea. In Gottlieb, J. and Raidl, G. R., editors, *Evolutionary Computation in Combinatorial Optimization, 6th European Conference, EvoCOP 2006, Budapest, Hungary, April 10-12, 2006, Proceedings*, volume 3906 of *Lecture Notes in Computer Science*, pages 86–98. Springer.

Giacobini, M., Tomassini, M., and Tettamanzi, A. (2005). Takeover time curves in random and small-world structured populations. In Hans-Georg Beyer et al., editor, *Genetic and Evolutionary Computation - GECCO 2005, Genetic and Evolutionary Computation Conference, Washington, D.C., USA, June 25-29, 2005, Proceedings*, volume 2, pages 1333–1340. ACM.

Goldstein, M. L., Morris, S. A., and Yen, G. G. (2004). Problems with fitting to the power-law distribution. *European Physical Journal B*, 41(2):255–258.

Guelzim, N., Bottani, S., Bourgine, P., and Képès, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31:60–63.

Jaeger, H. (2001). The echo state approach to analysing and training recurrent neural networks. Technical Report GMD Report 148, German National Research Center for Information Technology.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407:651–654.

Jiang, F., Berry, H., and Schoenauer, M. (2008). Supervised and evolutionary learning of echo state networks. In Günter Rudolph et al., editor, *Parallel Problem Solving from Nature - PPSN X, 10th International Conference, Dortmund, Germany, September 13-17, 2008, Proceedings*, volume 866 of *Lecture Notes in Computer Science*, pages 215–224. Springer-Verlag.

Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428:617–624.

Krapivsky, P. L., Redner, S., and Leyvraz, F. (2000). Connectivity of growing random networks. *Physical Review Letters*, 85(21):4629–4632.

Kuo, P. D. and Banzhaf, W. (2004). Small world and scale-free network topologies in an artificial regulatory network model. In J. Pollack et al., editor, *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, pages 404–409. Bradford Books, USA.

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

Leighton, F. T. (1992). *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*. Morgan Kaufmann, San Mateo, CA, USA.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quart. Appl. Math.*, 2:164–168.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441.

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

Newman, M. E. J. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351.

Nicolau, M. and Schoenauer, M. (2009). Evolving specific network statistical properties using a gene regulatory network model. In G. Raidl et al., editor, *Genetic and Evolutionary Computation - GECCO 2009, Genetic and Evolutionary Computation Conference, Montreal, Canada, July 8-12, 2009, Proceedings*. ACM.

Pastor-Satorras, R., Smith, E., and Solé, R. V. (2003). Evolving protein interaction networks through gene duplication. *Theoretical Biology*, 222:199–210.

Payne, J. L. and Eppstein, M. J. (2007). Takeover times on scale-free topologies. In Dirk Thierens et al., editor, *Genetic and Evolutionary Computation - GECCO 2007, Genetic and Evolutionary Computation Conference, London, UK, July 7-11, 2008, Proceedings*, pages 308–315. ACM.

Payne, J. L. and Eppstein, M. J. (2008). The influence of scaling and assortativity on takeover times in scale-free topologies. In Maarten Keijzer et al., editor, *Genetic and Evolutionary Computation - GECCO 2008, Genetic and Evolutionary Computation Conference, Atlanta, GA, USA, July 12-16, 2008, Proceedings*, pages 241–248. ACM.

Rechenberg, I. (1994). *Evolutionsstrategie '94*. Frommann-Holzboog, Stuttgart.

Reil, T. (1999). Dynamics of gene expression in an artificial genome - implications for biological and artificial ontogeny. In Floreano, D., Nicoud, J.-D., and Mondada, F., editors, *Advances in Artificial Life, 5th European Conference, ECAL'99, Lausanne, Switzerland, September 13-17, 1999, Proceedings*, volume 1674 of *Lecture Notes in Artificial Life*, pages 457–466. Springer.

Rohlf, T. and Winkler, C. (2008). Network structure and dynamics, and emergence of robustness by stabilizing selection in an artificial genome. In *8th German Workshop on Artificial Life, GWAL-8, Leipzig, Germany, July 30 - August 1, 2008, Proceedings*. Universitat Leipzig.

Tomassini, M., Giacobini, M., and Darabos, C. (2004). Evolution of small-world networks of automata for computation. In Xin Yao et al., editor, *Parallel Problem Solving from Nature - PPSN VIII, 8th International Conference, Birmingham, UK, September 18-22, 2004, Proceedings*, volume 3242 of *Lecture Notes in Computer Science*, pages 672–681. Springer.

Valverde, S., Cancho, R. F., and Solé, R. V. (2002). Scale-free networks from optimal design. *Europhysics Letters*, 60(4):512–517.

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.

Watts, D. J. (1999). *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press.

Wolfe, K. and Shields, D. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, 18:1694–1702.

Wuchty, S., Oltvai, Z. N., and Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35:176–179.