Evolving Scale-Free Topologies using a Gene Regulatory Network Model

Miguel Nicolau, Marc Schoenauer

Evolving Scale-Free Topologies using a Gene Regulatory Network Model

Miguel Nicolau
Marc Schoenauer

***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 exhibit a much higher potential for evolution.**

I. Introduction

Scale-Free networks are complex networks which have a few highly connected nodes, while most nodes are poorly connected [1]. 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 some positive real number γ, at least above a k given value.

Such network topology has been shown to exist in a variety of both artificial and biological systems [2], [3], [4], [5], [6], and has been widely studied because of its high resistance to random failure. Different generative models have been shown to create scale-free networks: in the original "preferential attachment" model, the network is gradually built, and new nodes attach preferentially to highly connected nodes [1]; however, this topology can also occur as a consequence of optimization processes, such as the wiring cost to existing software components (see [7] and references therein); finally, some artificial genome models, created through duplication and divergence, have been shown to generate networks with a power-law degree distribution [8], [9]. 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 – this paper 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 inter-disciplinary research opportunities, and as a result, GRNs provide 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 [10] provides a relatively recent overview of such researches.

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 [11], [12]; that they can exhibit scale-free and small-world topologies [13], [5], [6], [14]; that some specific network motifs, resembling those identified by biologists as building-blocks, are present within these artificial networks [15], [16]; 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 [17].

The present work focuses on the analysis of the underlying network topologies of one artificial GRN model [18], 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 analyzed 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 some given 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 II presents the GRN model used in the simulations, including the description and analysis of the duplication/divergence process used to initialize the genomes. Section III introduces the statistical tools used to assess the scale-free properties of the networks, along with the techniques to actually compute them. Section IV describes the experimental setup, the fitness measure and the results obtained when evolving GRNs to obtain scale-free network topologies. Finally Section V discusses those results and sketches some hints for future research directions.

Miguel Nicolau and Marc Schoenauer are with the INRIA Saclay - Île-de-France, TAO project-team, LRI - Bâtiment 490, Université Paris Sud, 91405 Orsay, France; email: [miguel.nicolau,marc.schoenauer]@inria.fr).

II. The GRN Model

A. Representation and dynamics

The artificial model described here is that proposed by W. Banzhaf [18]. 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 bits (5×32) immediately following a promoter sequence represent the gene itself, and are used to determine 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 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.

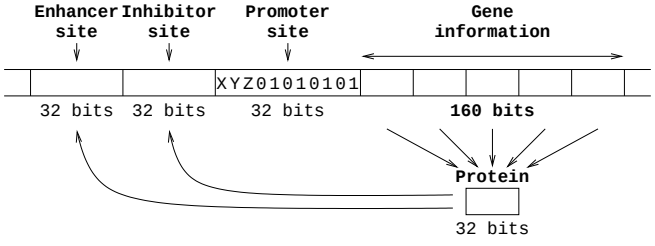

Fig. 1. Bit string encoding of a gene. If a promoter site is found, the gene information is used to create a protein, whose quantity is regulated by the attachment of proteins to the enhancer and inhibitor sites.

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 [18].

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

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

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:

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

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 simplifies some of the known characteristics of the biological regulatory process: all proteins are assumed to be *Transcription Factors*, that is, 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.

B. Genome Construction

The technique of duplication and mutation proposed [18] 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 [11], [12] 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 & Banzhaf [9]. For the sake of completeness, a similar technique has been used here to investigate the influence of the mutation rate on the number of genes per genome: 1000 genomes have been created using 14 duplication and divergence events, giving a genome length of $L_G = 32 \times 2^{14} = 524288$. The resulting number 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 indeed 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 ever exist in the genome sequence.

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).

C. 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 [18]. As all

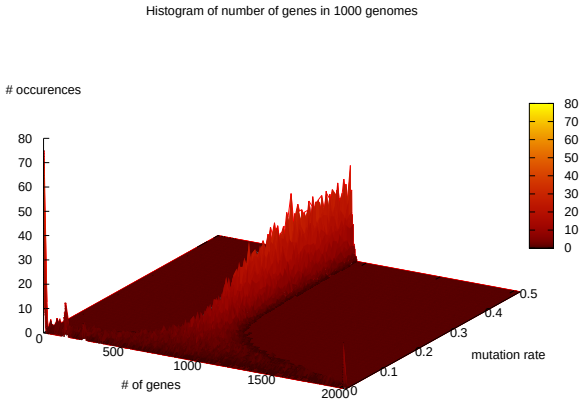

Fig. 2. Histogram of the number of genes per genome, over 1000 genomes. The x-axis plots the number of genes, the y-axis (height) the number of genomes having a particular number of genes, and the z-axis (depth) the mutation rate used. Genes were not allowed to overlap.

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, Fig. 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.

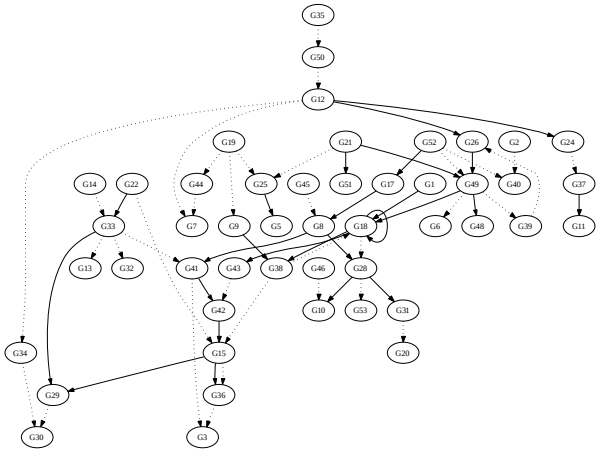

Fig. 3. Gene regulatory network for a genome of length $L_G = 32768$, created using 10 duplication events and a mutation rate of 50%, at a threshold of 23 bits. Solid edges indicate enhancing interactions, dotted edges indicate inhibiting interactions.

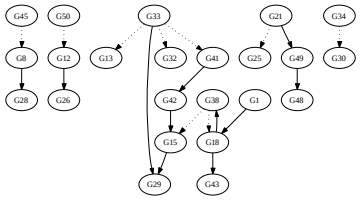

Fig. 4. Gene regulatory network for the same genome as in Fig. 3, at a threshold of 24 bits.

A completely different picture is that of genomes initialised through the duplication/divergence process described in Section II-B. 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.

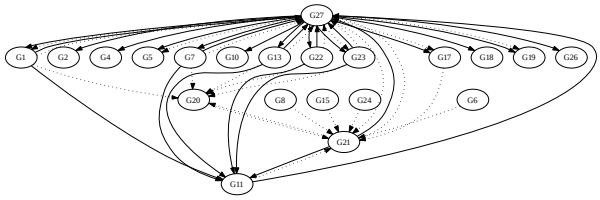

Fig. 5. Gene regulatory network for a genome of length $L_G = 32768$, created using 10 duplication events and a mutation rate of 1%, at a threshold of 16 bits.

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 used results in networks with similar dynamics: Fig. 6 and 7 depict the same 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.

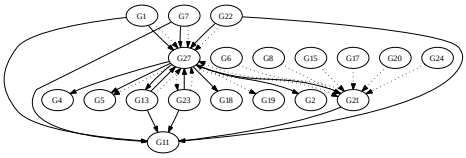

Fig. 6. Gene regulatory network for the same genome as in Fig 5, at a threshold of 17 bits.

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

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).

D. Connectivity variance

In order to generalize the observations made on the graphs above, an approach similar to that of Kuo et al. [14] 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.

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).

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.

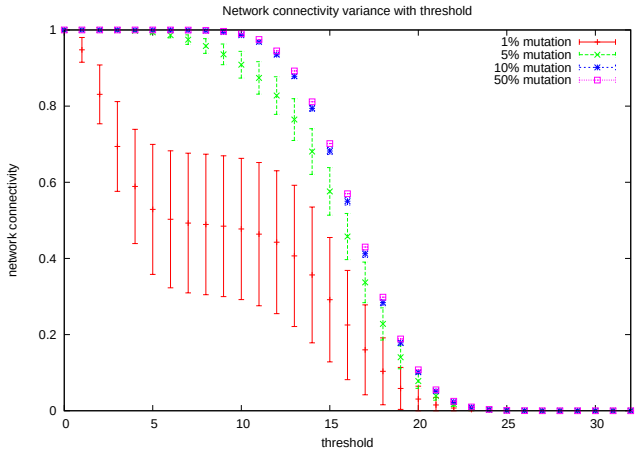

Fig. 8. Fraction of edges in a graph as compared to a fully connected network (and standard deviations), versus threshold parameter, based on samples of 100 genomes, created using 14 duplication events, and mutation rates of 1%,5%, 10%, and 50%.

III. Scale-free Topologies

Even though the model used is overly simplified compared to what is known of biological GRNs (as discussed in Section II-A), 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 [2], [3], [4], [5], [6], 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 some power law. But assessing such a distribution is not as obvious as it seems.

A. Measurement of Power Laws

Given a sample of some quantity, the typical method for measuring whether or not this quantity follows some 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 [19].

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 [20], [21]). 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 address the limitations of the above methods, a specific method, called *Maximum Likelihood Estimation* (MLE), has been proposed [19], [22], [23]. It seems to be one of the most stable methods for the approximation of the exponent of a power-law, and has been used here. The MLE method computes the exponent of the power-law as:

$$\gamma = 1 + n \left[\sum_{k=1}^{n} \ln \frac{P(k)}{P_{min}} \right]^{-1} \tag{4}$$

where $P(k)$ is the proportion of nodes with degree k, P_{min} the minimum value of all $P(k)$ in the sample, and n the number of samples. Note that values of k for which $P(k) = 0$ are not taken into account. Furthermore, the standard error of this estimation can also be easily computed as being:

$$\sigma = \sqrt{n} \left[\sum_{k=1}^{n} \ln \frac{P(k)}{P_{min}} \right]^{-1} = \frac{\gamma - 1}{\sqrt{n}} \tag{5}$$

B. Data Quality

Newman [22] observed that quite often, the tail of power-law distributions tends to be quite noisy, because of sampling errors: this is due to 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.

To tackle this problem, a technique known as *logarithmic binning* can be used [24]. It smoothes the histogram by grouping the distribution data per ranges of k values with exponentially increasing sizes (e.g. 1, $[2, 3]$, $[4, 7]$, $\dots$). This technique is also used in the present work.

C. 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 $\gamma = 1.59$. Using logarithmic binning does not help: due to the proximity of all values, there are only two points left in the distribution, leading to an MLE estimation of $\gamma = 2.219$, but with a high error for the estimation (a small n leads to a high error, as seen in Equation 5).

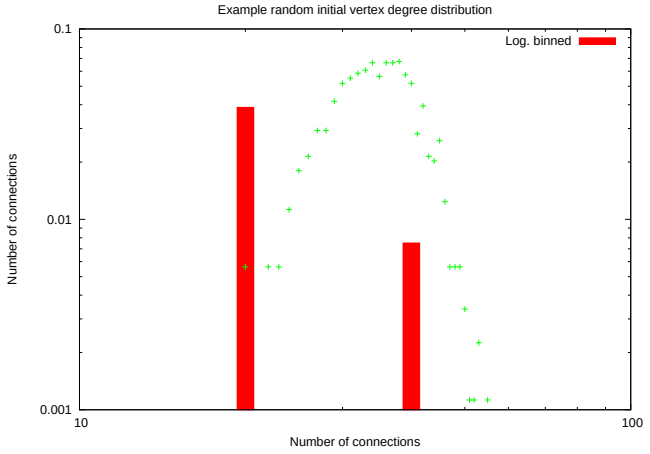

Fig. 9. Vertex degree distribution for the best network of a random genome, before (+) and after (vertical bars) logarithmic binning.

Networks extracted from initialised 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; a least-squares regression gives the value $\gamma = 0.767$, as a result. However, by using logarithmic binning, the values towards the end are somewhat normalised, resulting in an MLE estimation of $\gamma = 1.370$.

The occurrence of misleading γ values with random genomes can be further observed in Fig. 11: the size of logarithmic bins with random genomes is much smaller, giving rise to misleading 'good' (high) γ values. Initialised 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 [9], their degree distribution is generally quite far from a true power law. Nevertheless, while random graphs, because of the poor spread of their degree distribution, seem to be difficult to modify toward more scale-free topologies, initialised ones are more promising as seed topologies for the evolution of scale-free topologies. The 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.

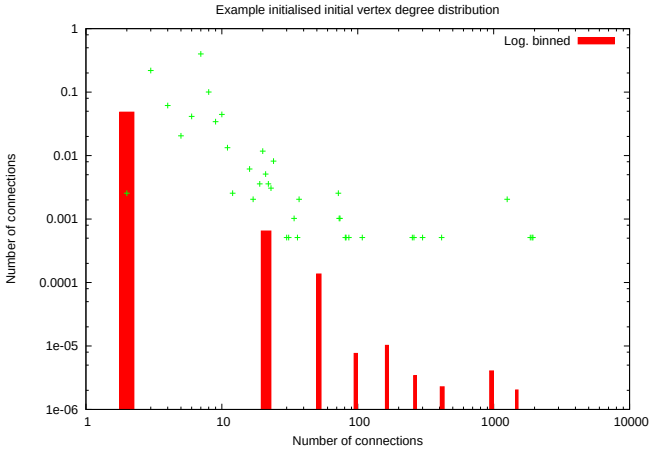

Fig. 10. Vertex degree distribution for the best network of an initialised genome, before (+) and after (vertical bars) logarithmic binning.

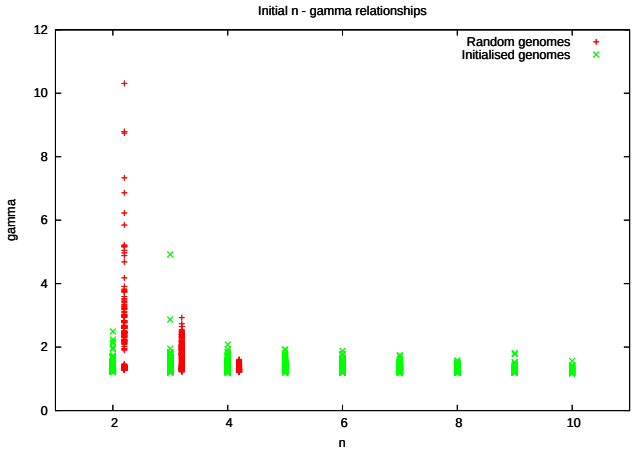

Fig. 11. Size of logarithmic bins and corresponding γ values, for random (+) and initialised (x) genomes, based on a sample of 100 genomes.

IV. Evolution of Topologies

The objective of this section is to evolve GRN genomes, so that the resulting interaction network gets as close as possible to a target γ value, using a simple Evolutionary Algorithm.

A. The Evolutionary Algorithm

A population of bit-string genomes, such as the ones described in Section II-A is evolved using the simple bit-flip mutation as the only variation operator. The evolution is a straightforward $(25+25)-ES$, 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 is initially set to 1% (per bit), and adapted in a way that is similar to the well-known $1/5$ rule of Evolution Strategies [25]: when the rate of successful mutations is higher than $1/5$ (i.e. when more than 20% mutation events result in an increase of fitness), the mutation rate is doubled; it is halved

in the opposite case[1].

In order to compare the evolvability of populations generated by the duplication/divergence method presented in Section II-B (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 have been performed with each of those initialisation procedures.

B. Fitness Function

Cohen and Havlin [26] 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 is used, and values of γ in $[2.4, 2.6]$ are considered ideal. The MLE method is used to compute an estimation of γ as described by Equation 4. The fitness function (to minimise) is therefore:

$$F(x) = \begin{cases} 0 & \text{if } 2.4 \leq \gamma \leq 2.6 \\ 2.4 - \gamma & \text{if } \gamma < 2.4 \\ \gamma - 2.6 & \text{if } \gamma > 2.6 \end{cases} + \frac{\sigma}{n} \qquad (6)$$

The statistical error of MLE σ (see Equation 5) is added to the absolute difference to the target γ values, as an estimate of the quality of the measurement. It is divided by the number n of points in the logarithmic binned vertex degree distribution, in order to penalize even more highly "regular" distributions where only a few data points would remain after the logarithmic binning (as seen in Section III-C).

From each GRN individual, several networks are extracted, by varying the threshold value; only the threshold giving the best fitness score is kept.

C. Experimental Results

Fig. 12 shows the best fitness in the population averaged over the 50 runs, for both initialisations procedures.

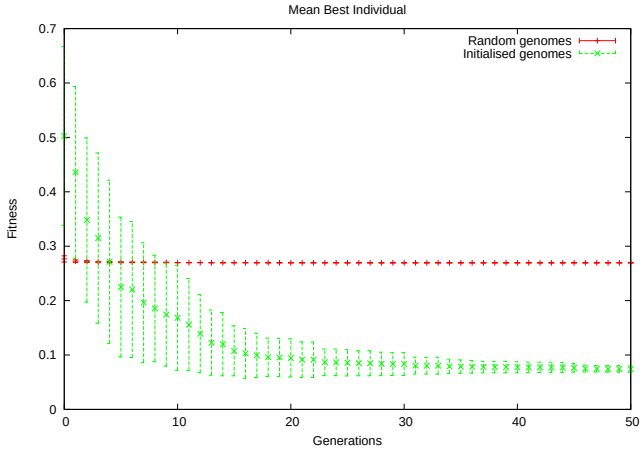

Fig. 12. Average best fitness per generation across 50 independent runs, for random genomes and 1% duplication/divergence initialised genomes. Error bars plot the standard deviation across runs.

[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 in any case.

As can be seen from the figure, though random genomes start with a much lower fitness, they hardly improve it during evolution (there is a slight improvement of fitness, although not visible at the scale of the plot). As already mentioned, this is due to the highly regular degree distribution in random networks: this gives very few points after logarithmic binning, and leads to fakedly good γ values. But it then makes it hard to vary the number of connections between nodes, even by adjusting the threshold parameter (see Fig. 8).

On the other hand, duplication/divergence initialised genomes do start with a worse fitness (smaller γ values), but are able to evolve to much better fitness values. The widely spread degree distribution results in more data points in the vertex degree distribution after binning. Although resulting in a higher initial error when estimating γ, it also creates a larger set of potentially fit networks from each genome, by varying the threshold parameter (as per Fig. 8).

Another reason for the greater efficiency of duplication/divergence initialised genomes as initial population for evolution is their ability to improve fitness by varying the size of genomes. This is illustrated on Fig. 13, that shows that random genomes keep roughly the same size for all genomes in the population across evolution, with very small variance across runs; initialised genomes, on the other hand, vary their size much more, with a much higher variance across runs. Even though the mutation 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 improved the fitness 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.

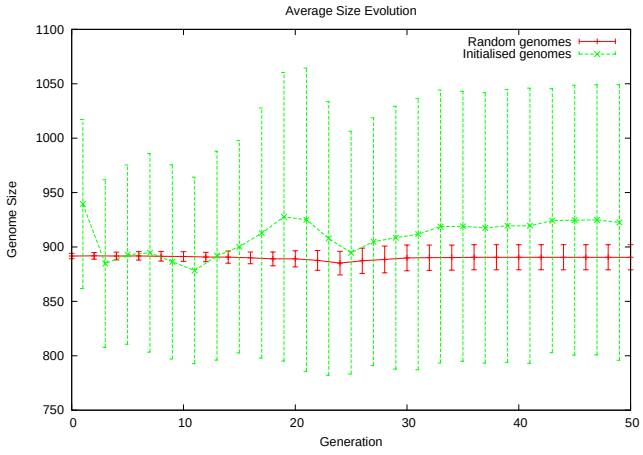

Fig. 13. Average population genome size across 50 independent runs, for random genomes and 1% duplication/divergence initialised genomes. Error bars plot the standard deviation.

The difference in terms of evolution potential with regard to scale-freeness can further be seen in Fig. 14, that displays the effectiveness of mutations during typical runs: when starting from randomly created genomes, most mutation events are harmful after just a few generations; when using an initial

population of genomes built through duplication/divergence, however, evolution lasts much longer, with neutral mutations starting to appear only after 24 generations, and with some beneficial mutations appearing as late as generation 46.

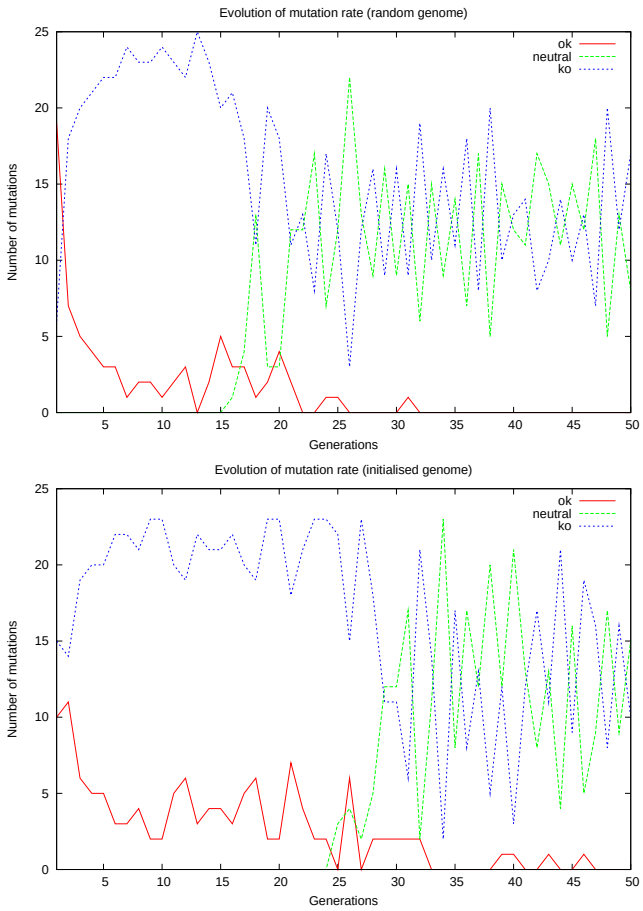

Fig. 14. Evolution of mutation effectiveness for random genomes (top) and 1% duplication/divergence initialised genomes (bottom), taken from a typical run. OK (resp. KO) mutations improved (resp. degraded) the fitness, while neutral mutations did not modify it.

Fig. 15 shows two examples of typical evolved networks extracted from initialised 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.

However, this figure also highlights some of the drawbacks of the experimental setup used here:

- By using logarithmic binning, the resulting vertex degree distribution consists of only a few points;
- The use of logarithmic binning in the fitness function also hinders evolution somewhat, as small changes to a node created by bit-level mutation are "diluted" across the bin to which it belongs;
- The incorporation of the error measure in the fitness function makes networks with a small number of (well aligned) points in the vertex degree distribution act as local minima, from which it is very hard to escape, in spite of the penalisation of small networks in the fitness.

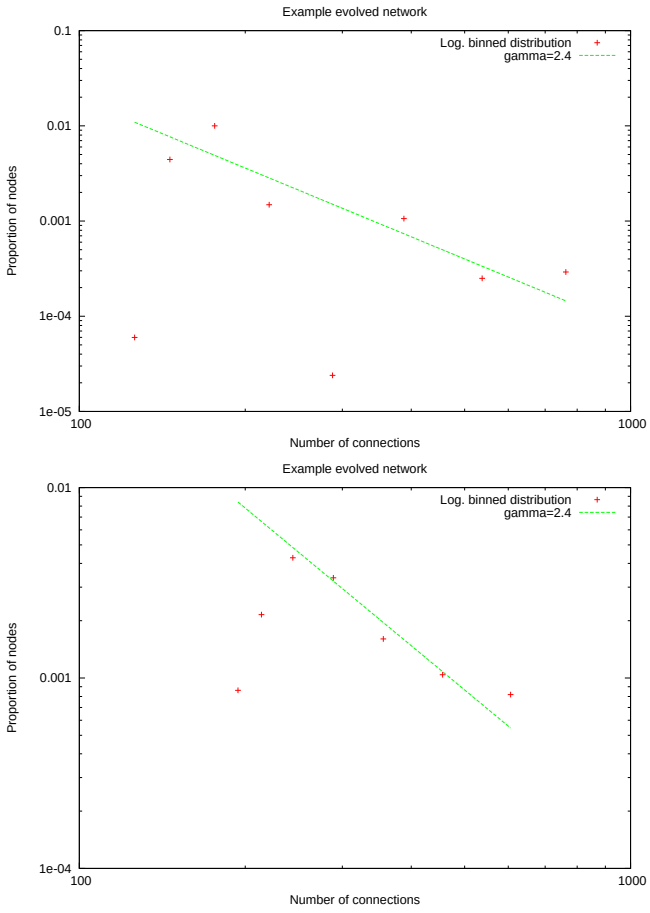

Fig. 15. Example networks extracted from best genomes, after evolutionary process based on initialised genomes. Vertex degree distribution was logarithmically binned, and plotted on a log/log scale. The value of γ was calculated to be 2.4 on both cases: a power-law with the same exponent (straight line) is also plotted for comparison.

The first two points can be addressed by using larger genomes, thus generating larger networks (typical network sizes were around $800 - 1000$ nodes here). The last point is a trickier one; a potential solution could be to raise even more the influence of the number of points n in the fitness, or setting a minimum size for it.

V. 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 optimizing a fitness measure that is directly connected to the topological property of the network, as opposed to the more classical generative methods where the scale-free property 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 W. Banzhaf [18] 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 of course several issues that still need to be addressed with the current approach. First of all, it remains to demonstrate that different values for the exponent can be reached that way, and not only the popular value of around 2.5 that was used here. Also, the use of logarithmic binning 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. 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: at the moment, a single evolution takes 4-6 hours of a recent Pentium computer (3.6GHz) for random networks, and 4-10 hours 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 IV-C), the influence of fixing the number of genes remains to be studied in more detail.

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

[1] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[2] S. Wuchty, "Scale-free behavior in protein domain networks," *Molecular Biology and Evolution*, vol. 18, pp. 1694–1702, 2001.

[3] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, pp. 651–654, 2000.

[4] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès, "Topological and causal structure of the yeast transcriptional regulatory network," *Nature Genetics*, vol. 31, pp. 60–63, 2002.

[5] V. van Noort, B. Snel, and M. A. Huynen, "The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model," *EMBO Reports*, vol. 5, no. 3, pp. 280–284, 2004.

[6] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, "Structure and evolution of transcriptional regulatory networks," *Current Opinion in Structural Biology*, vol. 14, pp. 283–292, 2004.

[7] S. Valverde, R. F. Cancho, and R. V. Solé, "Scale-free networks from optimal design," *Europhysics Letters*, vol. 60, no. 4, pp. 512–517, 2002.

[8] R. Pastor-Satorras, E. Smith, and R. V. Sole; "Evolving protein interaction networks through gene duplication," *Theoretical Biology*, vol. 222, pp. 199–210, 2003.

[9] P. D. Kuo and W. Banzhaf, "Small world and scale-free network topologies in an artificial regulatory network model," in *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, J. Pollack, M. Bedau, P. Husbands, T. Ikegami, and R. Watson, Eds. Bradford Books, USA, 2004, pp. 404–409.

[10] H. de Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *Computational Biology*, vol. 9, no. 1, pp. 67–103, 2002.

[11] K. Wolfe and D. Shields, "Molecular evidence for an ancient duplication of the entire yeast genome," *Nature*, vol. 387, pp. 708–713, 1997.

[12] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae," *Nature*, vol. 428, pp. 617–624, 2004.

[13] A. Bhan, D. J. Galas, and T. G. Dewey, "A duplication growth model of gene expression networks," *Bioinformatics*, vol. 18, no. 11, pp. 1486–1493, 2002.

[14] P. D. Kuo, W. Banzhaf, and A. Leier, "Network topology and the evolution of dynamics in an artificial regulatory network model created by whole genome duplication and divergence," *Biosystems*, vol. 85, no. 3, pp. 177–200, 2006.

[15] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási, "Evolutionary conservation of motif constituents in the yeast protein interaction network," *Nature Genetics*, vol. 35, pp. 176–179, 2003.

[16] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–1542, March 2004.

[17] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, pp. 308–312, 2004.

[18] W. Banzhaf, "Artificial regulatory networks and genetic programming," in *Genetic Programming Theory and Practice*, R. Riolo and B. Worzel, Eds. Boston, MA, USA: Kluwer Publishers, November 2003, ch. 4, pp. 43–62.

[19] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," *European Physical Journal B*, vol. 41, no. 2, pp. 255–258, 2004.

[20] K. Levenberg, "A method for the solution of certain problems in least squares," *Quart. Appl. Math.*, vol. 2, pp. 164–168, 1944.

[21] D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM J. Appl. Math.*, vol. 11, pp. 431–441, 1963.

[22] M. E. J. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary Physics*, vol. 46, pp. 323–351, 2005.

[23] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Physical Review E*, vol. 73, p. 016102, 2006.

[24] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378–382, 2000.

[25] I. Rechengerg, *Evolutionsstrategie '94*. Stuttgart: Frommann-Holzboog, 1994.

[26] R. Cohen and S. Havlin, "Scale-free networks are ultrasmall," *Physical Review Letters*, vol. 90, no. 5, pp. 058 701.1–058 701.4, 2004.