Evolving Specific Network Statistical Properties using a Gene Regulatory Network Model

Miguel Nicolau, Marc Schoenauer

Evolving Specific Network Statistical Properties using 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. The generation of network topologies with specific, user-specified statistical properties is the aim of this paper. This is achieved through the use of an artifical Gene Regulatory Network Model, shown previously to be able to correctly seed the population of an evolutionary algorithm, with the aim of steering the evolution towards the desired topologies. This method had previously been shown to be able to evolve scale-free topologies; the results obtained in this paper reinforce the applicability of the method, showing that the evolution of small-world topologies is also possible.

1 Introduction

Network generative procedures targeted toward specific topology properties are generally either iterative processes that sequentially add nodes using problem-dependent rules [1, 2], or ad hoc stochastic procedures modifying existing networks [3]. However, such procedures need to be designed anew whenever a new target property is wanted. Goal-directed procedure, on the other hand, only require a measure of the desired properties, and some (generally stochastic) optimization method able to search the space of network topologies. Unfortunately, such approaches generally suffer from two main drawbacks: the curse of dimensionality, making it untractable to optimize large networks; the bootstrap problem, in that random topologies barely indicate any meaningful path toward good solutions of the problem at hand.

An alternative is to use specific initialisation procedures that are able to produce diverse enough, though not random, topologies as a starting point for further optimization based topology design. Artificial Genetic Regulatory Networks offer such a procedure: previous work has demonstrated that simple models of GRNs were able to boost the evolutionary optimization of topologies in order to create scale-free topologies.

Building on this work, the present paper uses similar ideas to extend the generality of the approach, and introduces an original generative procedure to design Small World topologies [4]. This is a specific kind of topology, where most nodes are not directly connected, yet the average connection distance between any two nodes is very low. This kind of topology has been shown to exist in

many real-life networks, such as biological transcriptional networks [5], computer networks [6], or social networks [7].

The results obtained in the current study show that the presented method is applicable to the evolution of small-world networks, therefore reinforcing the usability of such approaches to the design of network topologies with user-specified statistical properties.

This paper is structured as follows: Section 2 presents the gene regulatory model used; Section 3 introduces the definition of small-world topologies, and the method used to extract topologies for the regulatory model. Then Section 4 presents and analyses the results obtained, and finally section 5 draws conclusions and future work directions.

2 The Gene Regulatory Network Model

The expression of genes in a genome is regulated by *Transcription Factors*. These are special proteins, produced by other genes, which can enhance or inhibit the production of their target genes; the networks of interactions between genes and the proteins they produce are termed *Gene Regulatory Networks* (GRN).

The model used in this work was first proposed by Wolfgang Banzhaf [8]. It represents a genome as a bit string, and uses specific bit sequences as promoter sites, identifying the location of a gene. If a gene is found, the 5×32 bits following it represent the protein it produces, and the the 2×32 bits upstream from the promoter site represent the enhancer and inhibitor sites for this gene, respectively. Fig. 1 illustrates the model.

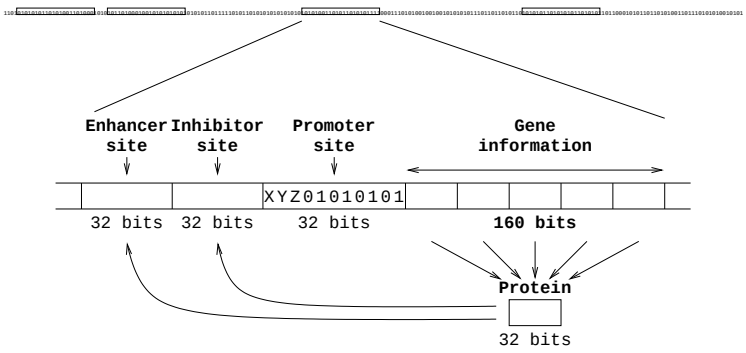

Fig. 1. Close-up view of the representation of a gene within the model.

The promoter site can be any sequence of bits; in this case, it is the sequence `XYZ01010101`, where X, Y and Z are any 8 bit sequence. The protein produced by the gene is a 32 bit binary sequence, extracted by a majority rule between all 5 sequences of 32 bits that compose it.

In this model, all proteins produced are transcription factors, and therefore they all regulate the expression of all genes, including their own producing gene. Regulation works by matching a protein and the regulating sites of a gene with the `XOR` operation: the result is the regulating strength. The enhancing and inhibiting signals regulating the production of protein p_i are then calculated as:

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

where N is the number of proteins, c_j is the concentration of protein j, $u_{i,j}$ is the `XOR` result 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 $\sum_i c_i = 1$ at all times, which results in competition between binding sites for proteins.

2.1 Initialisation Method

Although the binary genomes used within the model can be randomly created, an initialisation method has been proposed [8], based on a *Duplication and Mutation* (DM) process. It involves creating a random 32 bit sequence, followed by a series of length duplications associated with a low mutation rate. It has been shown [9] that this process of growing genomes can also occur in nature.

3 Network Topologies

Even though the model used is overly simplified compared to what is known of biological GRNs, an interesting issue is to find out whether or not the resulting interaction network exhibits particular properties resembling those found in natural networks. Previous work [10] has shown that evolution of linear genomes to achieve Scale-Free topologies [11–13, 5, 14] is possible; the current work concentrates on Small-World topologies [4, 5].

3.1 Small-World Networks

Small-World networks are characterised by a short average distance between any two nodes. Watts [4] describes them in terms of their characteristic path length L, and clustering coefficient C, with C_i being the percentage of all nodes connected to node i that are also connected to each other[1]. Given these definitions, a graph with n nodes and average vertex degree k is a small-world network, if it satisfies $L \geq L_{rand} \sim \frac{ln(n)}{ln(k)}$ and $C \gg C_{rand} \sim \frac{k}{n}$, with $n \gg k \gg ln(n) \gg 1$.

[1] The phenomenon $C_i = 1.0$ is also known as a *clique*.

3.2 Artificial Regulatory Networks

Once a genome has been constructed with the model described (see Section 2), its regulatory network topology can be analysed: genes will be represented as nodes, and the proteins they produce will be directed edges towards the genes they regulate; the weight of those edges will be the number of complementary bits between the regulating protein and the target gene (as seen in Eq. 1).

As all produced proteins regulate all genes, the resulting graph is complete. However, a *threshold* can be set on the minimum weight a connection must have, before being considered as an edge on the resulting regulating network.

Using different thresholds will therefore result in different networks. For example, Fig. 2 and 3 show regulating networks of the same random genome for two different thresholds (23 and 24). While almost all nodes are connected on Fig. 2, increasing the threshold by one removes many connections, and the graph on Fig. 3 is only a small sub-graph of the previous one (nodes which become isolated are not shown, which explains the smaller number of genes).

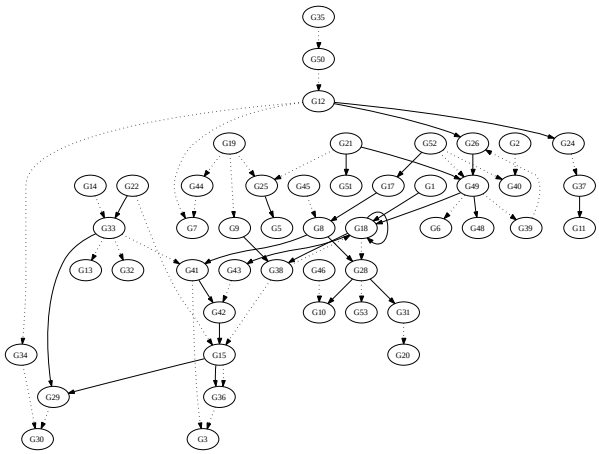

Fig. 2. Regulatory network for a random genome of length $L_G = 32768$, at a threshold of 23 bits. Solid edges are enhancing interactions, dotted edges are inhibiting ones.

Networks extracted from genomes initialised with the DM technique have remarkably different topologies, as seen in Fig. 4. The use of DM steps with low mutation result in a much shallower hierarchy of genes, with a few master genes, and most other genes poorly connected. Increasing the threshold removes connections, but the same master genes are still present, as can be seen in Fig. 5.

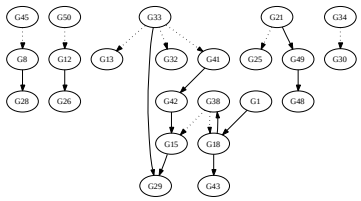

Fig. 3. Gene regulatory network for the genome from Fig. 2, at a threshold of 24 bits.

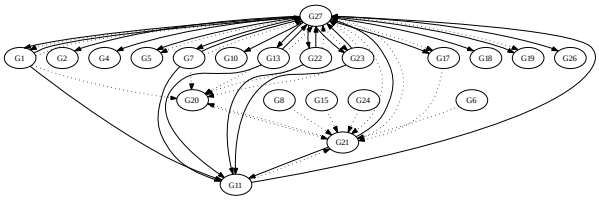

Fig. 4. Gene regulatory network for a genome of length $L_G = 32768$, created using 10 duplication events and a mutation rate of 1%, at a threshold of 16 bits.

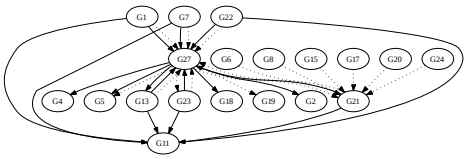

Fig. 5. Gene regulatory network for the genome from Fig. 4, at a threshold of 17 bits.

3.3 Connectivity variance

In order to analyse the effect of the threshold value on the network connectivity, an approach similar to that of Kuo et al. [15] has been used: 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. 6 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 DM 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 to no connectivity. 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 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.

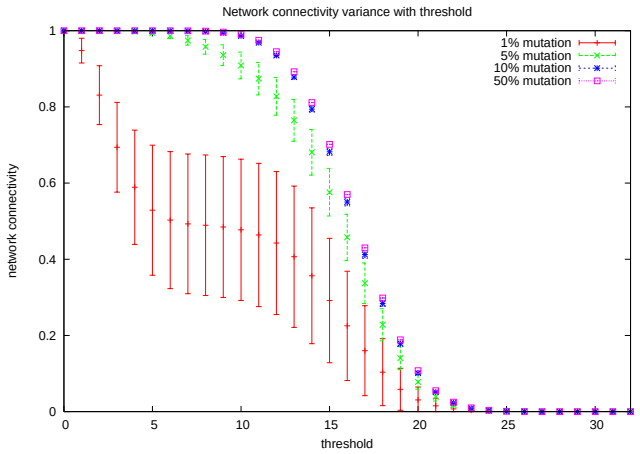

Fig. 6. Fraction of edges in a graph as compared to a fully connected network (and standard deviations), versus threshold parameter, based on samples of 100 genomes, created using 14 duplication events, and mutation rates of 1%,5%, 10%, and 50%.

4 Evolution of Statistical Properties

The objective of this section is to test whether genomes created with the model can be evolved, such that their regulating networks exhibit precise statistical properties, consistent with the definition of small-world topologies.

4.1 The Evolutionary Algorithm

A population of bit-string genomes is evolved, using the simple bit-flip mutation as the only variation operator. The evolution is a straightforward $(25+25)-ES$: 25 parents give birth to 25 offspring, and the best 25 of the 50 parents+offspring become the parents of next generation. The mutation rate is adapted following the 1/5 rule of Evolution Strategies [16]: its rate is initially set to 1% (per bit),

and 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[2].

In order to compare the evolvability of DM initialised genomes and completely random populations, 50 independent runs of 50 generations were performed with each type of genome, for each of the following fitness functions.

4.2 Fitness Function

The objective is to make "as true as possible" the relationship $n \gg k \gg ln(n) \gg 1$, as seen in Section 3.1. Candidate objective functions thus should try to maximise the three (normalized) following terms:

$$\frac{n-k}{n} \; ; \; \frac{k-\ln(n)}{k} \; ; \; \frac{\ln(n)-1}{\ln(n)}$$

Furthermore, in order to be considered small-world, a network must satisfy $L \geq L_{rand}$, and $C \gg C_{rand}$, so networks with $C < C_{rand}$ should be discarded, and relevant objective functions should maximise the following terms:

$$(L_{rand} - L)^+ \; ; \; \frac{C - C_{rand}}{C_{rand}}$$

A possible fitness function for an Evolutionary Algorithm is thus given by a weighted sum of those terms, whose weights should be tailored to the problem at hand:

$$F(x) = \alpha_1 \frac{n-k}{n} + \alpha_2 \frac{k-\ln(n)}{k} + \alpha_3 \frac{\ln(n)-1}{\ln(n)} + \alpha_4 (L_{rand} - L)^+ + \alpha_5 \frac{C - C_{rand}}{C_{rand}} \tag{4}$$

Note that, in the current work, all these weights were set to 1.

4.3 Results

Fig. 7 shows the best fitness in the population averaged over the 50 runs, for both initialisations procedures, using both fitness cases.

The results obtained show that DM initialised genomes get off their starting blocks with topologies which are much closer to being small-world, as defined in equation 4. Not only that, but their structure actually allows them to continuosly improve their fitness score through evolution.

Random genomes, on the other hand, have starting topologies which score badly. Furthermore, their regular topologies have very similar scores across different runs, and show little signs of evolution over time.

[2] 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.

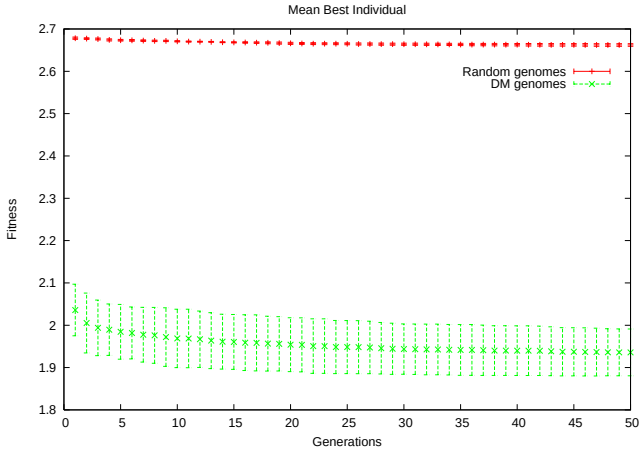

Fig. 7. Average best fitness per generation across 50 independent runs, for random and 1% DM initialised genomes. Error bars plot the standard deviation across runs.

Table 1 shows the relevant statistical measures of random and DM initialised networks, both without or with evolution. Of relevant interest is the fact that, after evolution, DM initialised networks exhibit a high clustering coefficient, while keeping a characteristic path length similar to that of randomly initialised genomes.

Table 1. Initial and evolved results, for random and DM initialised genomes.

	Generation	Threshold	N	k	ln(N)	L	C	Fitness
Random	0	22.199	1072.400	129.300	6.978	2.016	0.121	**2.812**
Genomes	50	23.000	1093.099	42.899	6.996	2.140	0.040	**2.662**
Initialised	0	12.900	1910.799	90.700	7.554	1.962	0.404	**2.274**
Genomes	50	11.699	1782.400	44.399	7.479	1.976	0.810	**1.936**

4.4 Analysis

The shallow hierarchies observed in initialised genomes exhibit characteristics similar to those of Small-World topologies, leading to the results observed. To analyse the reasons leading to such a different in the extracted network topologies, a sequence of DM steps was analysed, 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 drawn (Fig. 8 and 9) using a threshold of 13[3].

Fig. 8. Gene network after 6 duplication events.

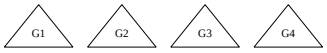

Fig. 9. Gene network after 7 duplication events.

The shape of the genes determines their starting location, when mapped to the original 32 bit sequence. For example, for the genome in Fig. 9, the starting locations for its genes were 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 drawn using the same shape.

This also explains why there are no connections between genes in Fig. 9. 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. 10). Most genes are still duplications of the 9^{th} bit of the original sequence; however, **G7** starts at a different location, and is thus drawn using a rectangular shape.

As the connectivity between genes is established by the difference between regulation sites and proteins, genes originating from different locations are more likely to be connected, even using lower threshold values. This can be seen in Fig. 10: genes labelled with different 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 outwards inhibiting connection to **G7**, whereas **G1** also has an inwards inhibiting connection originating from the same gene.

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

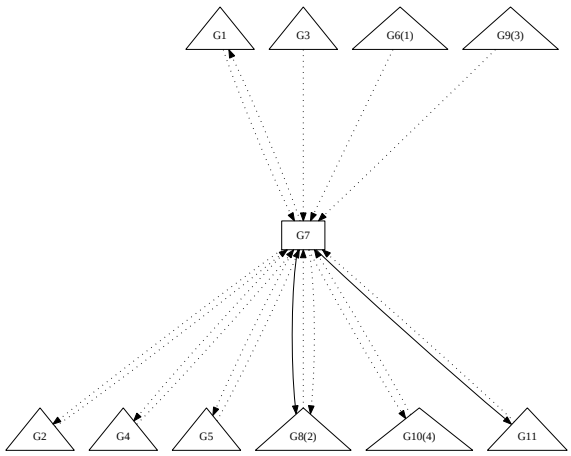

Fig. 10. Gene network after 8 duplication events.

With 9 DM steps, the network becomes a lot more complex (Fig. 11). 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 do not mix 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. 12), creating a network with 50 genes. Although hard to analyse at the naked eye, one can clearly see its shallow hierarchy, with a few highly connected nodes, to which most other nodes connect to. One can also see the appearance of a third type of gene, labelled with a pentagonal shape, which becomes the most connected gene. Table 2 shows a list of the gene families, along with their count, initial location, corresponding initial bit, and average number of inwards, outwards, and total connections.

Table 2. 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

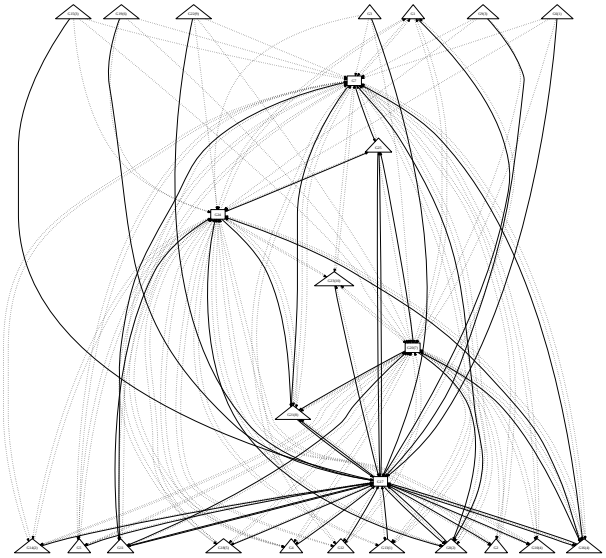

Fig. 11. Gene network after 9 duplication events.

Although this network is just an example, many networks were found to follow the same mechanics while being extracted from genomes grown with DM steps. It shows that the tendency of initialised 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.

5 Conclusions and Future Work

The experimental results obtained with this work show that the use of a gene regulatory network model allows for the construction of network topologies with specific statistical properties, in this case small-world topologies. Once again the initialisation procedure plays a vital role in the seeding of potential models, not only giving them a head-start when compared to random models, but also leading to better final statistical measures, with specific emphasis on clustering coefficient values (as seen in Table 1).

This way of evolving topologies also allows one to fine-tune the objective function, such that specific relationships between n, k and $ln(n)$ (see Section 3) are possible.

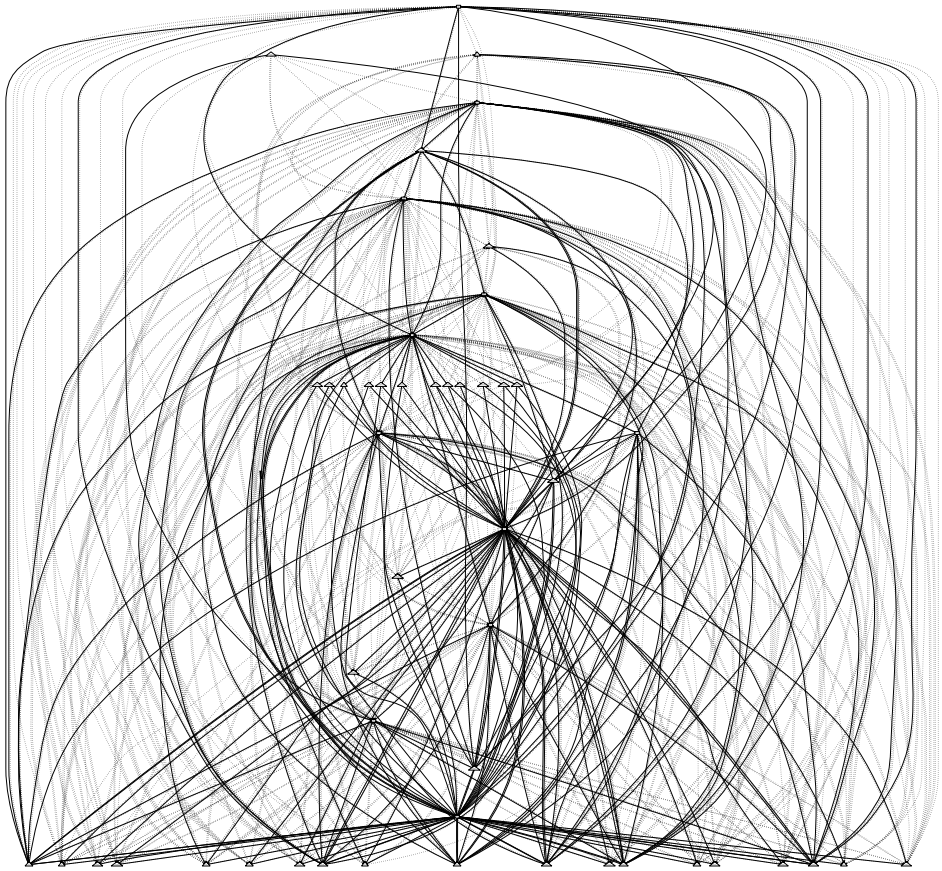

Fig. 12. Gene network after 10 duplication events.

A few problems remain with the current approach. The set of weights used in the fitness function (see 4) gives equal importance to the differences between the terms n, k, $ln(n)$ and 1, regardless of them having a few orders of magnitude of difference. Furthermore, the relationship $C \gg C_{rand}$, which fundamentally defines a small-world topology, is given the same fitness weight as the differences between the previous four terms. Although there is evolution with this setup, a better fitness function could be designed in the future.

Another interesting future work direction could be a multi-objective approach to this problem. Even more promising could be a multi-objective approach to join the current work and the previous work on scale-free networks [10], evolving network topologies which are both scale-free and small-world.

Acknowledgment

This work was supported by the Sixth European Research Framework (proposal number 034952, GENNETEC project).

References

1. Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.
2. Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
3. Romualdo Pastor-Satorras, Eric Smith, and Ricard V. Solé. Evolving protein interaction networks through gene duplication. *Theoretical Biology*, 222:199–210, 2003.
4. Duncan J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, 1999.
5. Vera van Noort, Berend Snel, and Martijn A. Huynen. 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, 2004.
6. Serguei N. Dorogovtsev and José F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.
7. John Guare. *Six Degrees of Separation*. Vintage, November 1990.
8. Wolfgang Banzhaf. Artificial regulatory networks and genetic programming. In Rick Riolo and Bill Worzel, editors, *Genetic Programming Theory and Practice*, chapter 4, pages 43–62. Kluwer Publishers, Boston, MA, USA, November 2003.
9. K. Wolfe and D. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.
10. Miguel Nicolau and Marc Schoenauer. Evolving scale-free topologies using a gene regulatory network model. In *IEEE Congress on Evolutionary Computation, CEC 2008, Hong-Kong, June 1-6, 2008, Proceedings*, pages 3748–3755. IEEE Press, 2008.
11. Stefan Wuchty. Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, 18:1694–1702, 2001.
12. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.
13. Nabil Guelzim, Samuele Bottani, Paul Bourgine, and Fran cois Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31:60–63, 2002.
14. M. Madan Babu, Nicholas M. Luscombe, L. Aravind, Mark Gerstein, and Sarah A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14:283–292, 2004.
15. P. Dwight Kuo, W. Banzhaf, and André Leier. 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, 2006.
16. Ingo Rechengerg. *Evolutionsstrategie '94*. Frommann-Holzboog, Stuttgart, 1994.