A Computational Model of Gene Regulatory Networks and its Topological Properties

Ângela Gonçalves[2,3] and Ernesto Costa[1,2]

[1] Department of Informatics Engineering, University of Coimbra
[2] Centre of Informatics and Systems of the University of Coimbra
[3] European Space Agency, Rome
angela.goncalves@esa.int

Abstract

A new model for Gene Regulatory Networks (GRN) is proposed. The model is potentially more biologically sound than other approaches, and is based on the idea of an artificial genome from which several products like genes, mRNA, miRNA, non-coding RNA, and proteins are extracted. These products are connected giving rise to a heterogeneous directed graph. The topology of the obtained networks is studied using degree distributions. We make some considerations about the biological meaning of the outcomes of these simulations.

Introduction

Sequencing the human genome was a tremendous breakthrough, but today's great challenge is deciphering how genes determine the phenotypic traits of an organism and how the genome controls the development of organisms. Although biology's central dogma explains the basic process of gene expression into protein phenomena like cellular differentiation, the ability of cells with the same genetic information to behave differently according to their function in the organism, is not accounted for in the dogma. The answers to such questions lie in complex networks of interactions, known as regulatory networks, between genes and other molecules including proteins, the very products of gene expression. Regulatory networks are highly non-linear and have thousands of variables: finding a computational model for them is a difficult albeit important task. Various approaches for modeling gene regulatory networks (GRNs) appeared in the last decades focusing on regulation at transcription level, the best known form of regulation. However, recent studies revealed that regulation occurs at any stage of protein synthesis including transcription, RNA processing, mRNA decay, translation and post-translation. In this paper we propose a new model for gene regulatory networks, called HeRoN, that introduces a level of biological detail that not present in previous models, and study the topological properties of the networks using degree distributions.

The paper is organized as follows: in section 2, we will give a brief explanation of the biological concepts of gene regulation; a review of different approaches is given in section 3, and, in section 4, our own model is presented; section 5, will describe the experimental setup and the topological aspects of GRNs based on degree distributions; some concluding remarks are presented in the last section.

Gene Expression Regulation

Gene expression can be decomposed in three stages: transcription, processing and translation. In the transcription phase, a RNA molecule is created by complementing the DNA sequence of the gene starting in a place called the promoter of a gene. Transcription also ends when a particular signal is found. After transcription the RNA transcript is processed and certain non-coding sequences, called introns, are removed. The remaining sequences are joined and form a mature mRNA molecule. This mRNA molecule is then translated into a protein according to a known relation called the genetic code.

The central dogma posit gene expression as a one-way process where information flows from the genes to the proteins. What actually happens in organisms is that genes, proteins, mRNAs and other types of molecules in the cells, are able to interact with each other given rise to a regulatory process, which can occur at any stage of expression. One of the most well-known regulation mechanism acts at the transcription initiation stage. This type of regulation consists in the binding of certain proteins, called transcription factors, to particular sequences in the genome physically helping or making impossible the initiation of transcription. With this mechanism some genes are able to regulate the expression of other genes or even themselves. Other type of regulation occurs at the transcription termination and is influenced by many types of molecules in a cell. The regulation processes is highly non-linear and in Gonçalves and Costa (2007a) we studied the dynamics characteristics of GRN, namely the emergence of three types of behaviors: fixed, periodic and chaotic. One of the goals of this paper is to study the static topological properties of GRN, aiming at acquiring some insights about biological aspects of the process of genetic regulation.

State of the Art

In recent years several models for Gene Regulatory Networks have been proposed. The majority of models make the simplifying assumption that the control of gene expression resides only in the regulation of gene transcription. Due to lack of space we only briefly mention some of the known models. For an in-depth description (see Gonçalves and Costa, 2007b)

One early and influencing discrete approach adopted a complex system view of the genome (Kauffman, 1993). Using Random Boolean Networks, Kauffman represented the regulatory system as a network of logical components connected at random. Despite the interesting insights of Kauffmans model it was unable to give much explanation for the regulatory mechanisms and, to many, did not exhibit sufficient parallels with the real networks due to its abstract nature (Reil, 1999). Rather than using a network for a base level representation, another discrete and promising model, the Artificial Genome, originally proposed by Reil (1999) used a more biological framework being based on a DNA-like sequence representing the genome from which the network structure could be extracted. A similar model was proposed by Banzhaf (2003) and described by Hallinan and Wiles (2004), Watson et al. (2004) and Willadsen and Wiles (2003). In the original Artificial Genome model a random string of bases is generated to represent the genome of an organism. The string is searched for promoter sequences which are, by convention, '0101'. The six digits following the promoter will represent the gene sequence (see Figure 1A). The sequence between genes will be the regulatory region for the following gene. An operation is applied to the gene sequence to create a gene product. This operation represents the entire expression process and consists of incrementing each gene's digit by one, modulo 4 (the number of bases) (see Figure 1B). The resulting sequence is the gene product which will be used to search for matches in the regulatory regions of all genes (see Figure 1C). Each match represents a regulatory link between the gene that originated the gene product and the gene regulated by the region where the match occurred. Whether the regulation is inhibitory or excitatory depends on the value of the last digit of the gene product. After performing the matches a regulatory network can be extracted and displayed in the form of a graph. The Artificial Genome still comprise many simplifications, e.g., the merging of the entire process of gene expression, presenting no intermediate products and determining an arbitrary operation for the creation of a gene product.

Several models have been proposed that treat variables as continuous values and calculate them through differential equations. The Additive Regulation Model, and models based on the S-System power law are some examples. In the Additive Regulation Model (D'Haeseleer, 2000) variables are continuous and updated synchronously. This model can be represented as a matrix of positive, negative or zero connections (see Figure 2).

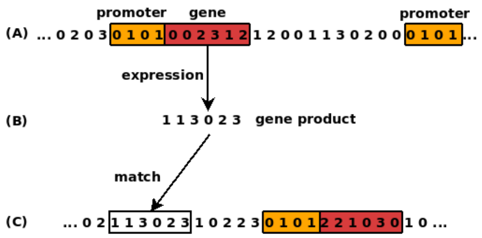
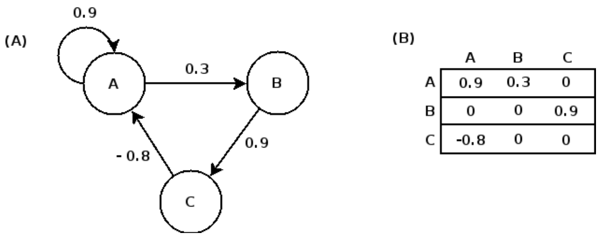

promoter
gene
promoter
(A) ... 0 2 0 3 0 1 0 1 0 0 2 3 1 2 1 2 0 0 1 1 3 0 2 0 0 0 1 0 1 ...
expression
(B) 1 1 3 0 2 3 gene product
match
(C) ... 0 2 1 1 3 0 2 3 1 0 2 2 3 0 1 0 1 2 2 1 0 3 0 1 0 ...

Figure 1: Artificial Genome model.



	A	B	C
A	0.9	0.3	0
B	0	0	0.9
C	-0.8	0	0

Figure 2: A Graph of a Regulatory Network

When a matrix entry is nonzero, there is a regulatory connection from gene product i to gene product j. If the entry is positive the regulation is enhancing and if it is negative the regulation is repressive. The expression level of each gene x_i could be given by the weighted sum of all variables:

$$\frac{dx_i}{dt} = S(\sum_j w_{ji} x_j + b_i) - D_i x_i \tag{1}$$

with x_i the expression level of the ith variable, b_i a bias term that indicates if the gene is expressed in the absence of regulatory inputs, w_{ji} the weight in the matrix from gene j to gene i, $S()$ a sigmoidal function and D_i the decay rate of gene i.

Finally, an S-system is a parameterized set of nonlinear differential equations:

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{n} x_j(t)^{g_{ij}} - \beta_i \prod_{j=1}^{n} x_j(t)^{h_{ij}} \tag{2}$$

where x_i is the expression level of gene i, n is the number of network components, $\alpha_i \geq 0$, $\beta_i \geq 0$ are rate constants and g_{ij}, h_{ij} represent the interactive affectivity of x_j to x_i. The first product describes all influences that are excitatory (increase x_i) and the second product all influences that are inhibitory (decrease x_j). These systems have a rich structure but number of parameters that have to be estimated is large (Noman and Iba, 2005).

With the continuous models just described biologically plausible features such as decay rates of molecular products

(D'Haeseleer, 2000) can be included and reverse engineering/learning algorithms can be used to determine their parameters from real data (Ando and Iba, 2001; Sakamoto and Iba, 2001; Noman and Iba, 2005), however, as with the discrete models discussed, the blackbox approach of the process they use makes understanding the mechanisms of gene regulation at the various levels more difficult.

The above considerations prompted the creation of a new model called HeRoN with a string based framework similar to the Artificial Genome breaking the process down to its important steps and overcoming some of its simplifications. The networks derived by the HeRoN model can be represented by a graph where the nodes represent the different products involved in the process of gene expression, thus heterogeneous, and the arcs establish the interactions between the products.

HeRoN: a Model for a Heterogeneous Gene Network

The proposed model HeRoN takes a string from a four symbol alphabet representing the genome and derives from it various products such as genes, proteins and some more intermediate products. The expression algorithm is a six-step process that will now be described with some detail.

1. Generate the genome The genome, implemented as a string of integers, is randomly generated given a size parameter. Each integer corresponds to a base: 0 - T(U), 1 - A, 2 - G, 3 - C.

2. Search the genome for genes and create them The genome is searched for given sequences that represent the gene promoters. In real biological systems there are some promoter sequences that appear in most genes of many organisms, called consensus sequences, and the more a sequence in a genome resembles them, the more efficient the transcription. To achieve this, a threshold symbolizing the binding strength between a RNA polymerase and the genome, was set as a parameter. A sequence in the genome, with the same size as the given promoter sequence, is considered to be a valid promoter in the genome when its percentage of match with the given sequence is equal or above the threshold. Each time a valid promoter is found the genome is searched for a termination sequence. When such termination sequence, chosen to be a poly-A sequence of adjustable size, has been found a gene is created. Each gene consists of a promoter sequence, the coding sequence and the regulatory region. The coding sequence is the region located between the promoter and the termination sequence. The regulatory region is the region located between the end of the previous gene (after its termination sequence) and the promoter (see Figure 3A).

3. Generate RNA transcript from the genes The RNA transcript is generated by complementing the bases on the coding sequence of the gene according to the pairing A-T and C-G. In the four integer alphabet 1 and 0 are the complement of each other as are 3 and 2 (Figure 3A-B).

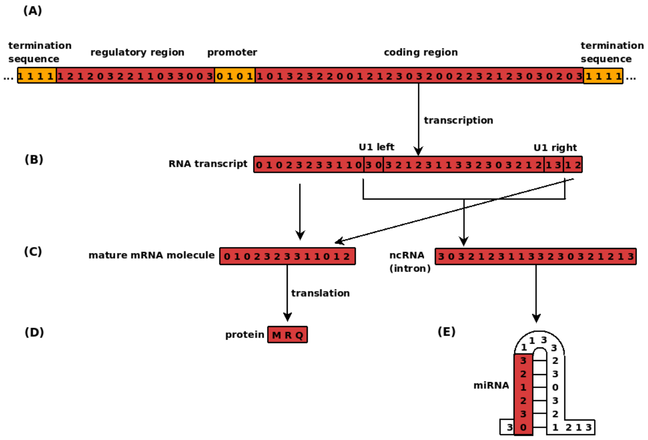


Figure 3: The HeRoN model. See text for an explanation.

4. Splice the RNA transcript which generates the mature mRNA and introns Splicing the RNA transcripts means that each RNA transcript is searched for introns that are removed from the sequence and stored into a list of components called ncRNAs. Introns are detected by means of two sequences, U1 left and U1 right, that simulate the role of U1 srRNA molecule that has two highly conserved consensus sequences complementary to the 5 and 3 ends of essentially all mRNA introns (Zhang and Rosbash, 1999). The new sequences created from the RNA transcripts with the introns removed are called mRNA (Figure 3B-C).

5. Translate the mRNA into proteins Each mRNA molecule is scanned for the start codon sequence (AUG). When this sequence is found the mRNA is read three bases at a time until a stop codon is found (UAA, UGA, or UAG). Each three bases are translated into one amino acid according to the genetic code table. The stop codon is not considered as part of the protein (Figure 3C-D).

6. Search the ncRNAs from miRNAs and create them The model incorporates a mechanism of RNA interference that regulates the stability of mRNA by triggering its degradation. This mechanism was added to the model when it was noticed that a large number of RNA transcripts did not produce proteins because they missed the start codon. Searches in biology literature for similar phenomena led to the subject of non-coding/junk DNA. Junk DNA has been a name given by researchers to large portions of DNA for which no function has yet been identified, including introns and large portions of intergenic sequences. Having found evidence that genes considered to be junk DNA have a regulatory influence (Martens et al., 2004) and that this kind of DNA makes up to 95% of chromosomes, researchers reversed their opinions on the usefulness of junk DNA, changing its name to non-coding DNA. In particular, the regulatory role of noncoding genes relates to the RNAi mechanism. This mechanism of transcriptional gene silencing is induced by the association between proteins and RNA. The resulting molecules are called small interfering RNA (siRNA), when they derive from exogenous sources (outside the cell), or are called microRNA (miRNA), when they are produced from non-coding genes in the cells own genome. miRNAs are short single-stranded RNA stretches of 21 to 23 nucleotides that are processed from primary transcripts known as pre-miRNA to short stem-loop structures called pre-miRNA and finally to functional miRNA (Gregory et al., 2006). The effect of this regulation mechanism is that while some genes are transcripted at a normal rate they are not expressed because they are degraded before they leave the nucleus. To incorporate this influence in the model it was determined that if the resulting protein has no sequence, because the mRNA misses the start codon, that mRNA molecule is considered to be non-coding and therefore is added to the ncRNA list where the introns were already stored. All ncRNAs are then scanned for hairpin loops with a minimum length. This indicates the presence of miRNAs that are then considered as another product in the model (Figure 3C-E).

From the expression algorithm to the network

The expression algorithm described above creates a list of products and stores their corresponding sequences and references to the products from which they derived. To extract the interaction network between these products it is necessary to determine the bindings between them, namely between proteins and genes and between miRNAs and genes. Finding the interactions between miRNAs and genes is simple since the two products are made of the same components, nucleotides, and their binding is a simple match between complementary sequences. The other type of binding involves elements that do not interact in a linear manner and are made up of different components, amino-acids and nucleotides. In biological systems the proteins ability to locate and bind with certain DNA sequences depends not only on the involved amino-acid and nucleotide sequences but also on the protein's three-dimensional structure and on the DNA double stranded structure. Many solutions exist that try to predict DNA-protein binding sites (Baker and Sali, 2001) and this is still an open topic in Bioinformatics. In addition to these approaches some authors find it important to examine the individual interactions between the amino-acids and the nucleotides since underlying the bindings are the discrete interactions between them (Hoffman et al., 2004). Databases such as the Amino Acid-Nucleotide Interaction Database (AANT) categorize amino-acid-nucleotide interactions from experimentally determined protein-nucleic acid structures. In our model a protein and a DNA sequence are perfectly aligned and the statistical table of the entire AANT database along with a binding threshold is used to determine if they bind. For each amino-acid in the protein its binding probability with the corresponding nucleotide in the DNA is given by the AANT statistic table. Given the interactions, one of four methods, called average, maximum, minimum and random, is used to compare them with the threshold (see Figure 4).

Aminoacid	A(%)	C(%)	G(%)	T(%)
Alanine (Ala, A)	24.2	17.3	24.0	24.6
Arginine (Arg, R)	19.6	24.1	35.7	12.2
Asparagine (Asn, N)	25.5	20.0	23.9	17.7
Aspartate (Asp, D)	13.3	34.2	37.0	1.5
Cysteine (Cys, C)	29.1	18.8	24.8	23.1
Glutamine (Gln, Q)	28.0	17.7	29.4	13.7
Glutamate (Glu, E)	19.1	34.8	33.0	4.8
Glycine (Gly, G)	20.1	22.9	32.1	17.0
Histidine (His, H)	25.3	16.2	37.7	14.2
Isoleucine (Ile, I)	21.4	26.4	30.8	11.4
Leucine (Leu, L)	9.5	31.1	30.2	19.4
Lysine (Lys, K)	23.7	22.8	30.7	16.3
Methionine (Met, M)	22.1	27.9	22.1	9.8
Phenylalanine (Phe, F)	17.7	24.1	40.5	17.7
Proline (Pro, P)	37.0	11.0	21.0	2.0
Serine (Ser, S)	28.2	20.9	27.2	19.7
Threonine (Thr, T)	24.6	20.2	27.8	23.1
Tryptophan (Trp, W)	14.4	30.2	24.8	21.8
Tyrosine (Tyr, Y)	28.4	27.4	23.6	15.0
Valine (Val, V)	25.0	35.3	20.0	1

Table 1: Statistical table of the entire AANT database. Along with the name of the amino-acids are the conventional three-letter and one-letter abbreviations.

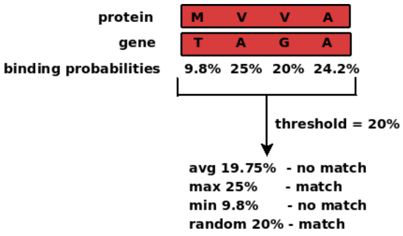


Figure 4: Protein binding example. Each amino-acid-nucleotide pair is searched for in the AANT statistic table (see Table 1 for a complete description).

If using the average method an average of all the probabilities is calculated. For the maximum and minimum methods, the respective maximum or minimum probability is chosen. For the random method the probability of a random amino-acid-nucleotide pair, from the sequence, is chosen. In the example of Figure 4 it was the V-G pair. The components are said to bind if the resulting probability is above or equal to the threshold.

The information gathered about the interaction between the components is then used to create a graph representation of the network where each gene creates many products, most of them ncRNAs and a single mRNA molecule. Each mRNA either creates a protein or a miRNA molecule, miRNA molecules can also derive from ncRNAs. All connections starting at a miRNA molecule end at an mRNA molecule and are repressive, while connections between proteins and genes can be either activating or repressive (see Figure 5).

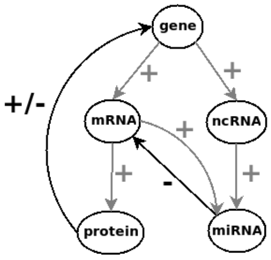


Figure 5: Activation/deactivation relations between the different products. The positive and negative signs near the edges represent, respectively, activation or deactivation of a product. The black colored edges represent the regulatory connections while the grey edges represent the "creation" of a product.

Experimental Setup and Results

Now that the model has been described it is time to present the experimental study that was carried out. Here we will be concerned only with the topological properties of GRNs.

Parameter	Used values
genome size	20000, 100000 and 500000
miRNA binding site size	4, 5, 6 and 7
inhibition rate	0, 0.25, 0.50 and 0.75
binding threshold	29, 32, 33 and 34
binding choice	avg, max, min and rand

Table 2: Variable parametrization

Table 2 shows the variable parameterization used throughout the experiments performed. The fixed parameters are: sequence 0101 for the promoter, promoter match of at least 75%, sequence 1111 for the termination sequence, and a binding site size of 6 for the proteins. Experiments were run for all possible combinations of the 'used values' mentioned in Table 2 . Each combination of the variable parameters was run 10 times. The initial set of active genes for each of the runs was randomly determined from a uniform distribution.

Topology of the obtained networks

The different topology classes of networks, i.e., regular lattice, small-world and random networks, arise from the different ways large sets of elements connect. A network where each node is connected to its nearest spatial neighbors is the so called regular lattice. Starting with a regular lattice and randomly rewiring a portion of the links creates small-world networks. At the extreme random networks are formed, where every pair of elements is connected at random. Like most social and biological networks, such as the World Wide Web, the immune system, the brain and ant colonies, to name just a few examples, genetic regulation networks possess certain non-trivial topological features. For instance, while nodes on regular lattices have constant degree and ordinary random networks have Poisson degree distributions, it is found that many real-world networks have degree distributions measurably different from these. This strongly suggests that there are features of such networks that would be missed if they were to be approximated by an ordinary random graph or lattice (Newman et al., 2001), thus many recent works on real-world complex systems focus on the subject of small-world and scale-free networks. While there are several statistical properties of graphs that may be used to characterize their topology (e.g., average path length, clustering coefficient), the work done on HeRoN concentrates on the degree distributions of the obtained networks. Most frameworks, with very few exceptions (Newman et al., 2001), for the study of graph statistical properties have been developed for unipartite, undirected graphs. It is, however, an important aspect for us to consider directed and heterogeneous graphs (graphs with nodes of different types), since this is the case of the network graphs obtained with the HeRoN model. One consequence of the graph being directed is that nodes have two different kinds of edges, the ones arriving at the node and the ones leaving the node - these will be referred to, respectively, as input and output connections. This is particularly important in analyzing the degree distribution of the nodes and therefore nodes of different kinds will be analyzed separately in relation to input and output connectivity.

Figure 6 and Figure 7 show the input and output degree distributions for each kind of node for a 20,000 base long genome on the left column, and for a 500,000 base long genome on the right column. Each column refers to the same network, obtained with a binding threshold of 29, 'avg' binding choice, miRNA binding size of 6 and an inhibition rate of 0. Table 3 gives the number of products of each kind in networks of different sizes with this parameterization.

#genome	#genes	#mRNA/prot	#ncRNA	#miRNA
	(8%)	(6%)	(80%)	(1%)
20000	57	46	556	9
100000	253	194	2993	50
500000	1361	1043	14531	259

Table 3: Number of each kind of product for different genome sizes, binding threshold = 29 and binding choice = avg. The number of proteins is the same as the number of mRNAs.

Whilst the 'genome size' parameter does not seem to

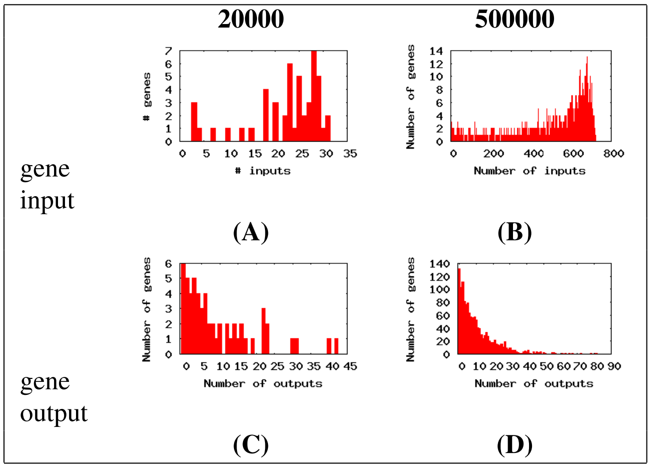


Figure 6: Histograms giving the degree distribution of gene input and output connectivities for a 20,000 base long genome, on the left column, and a 500,000 base long genome, on the right column. **(A)** and **(B)** Gene input connectivity distribution. **(C)** and **(D)** Gene output connectivity distribution.

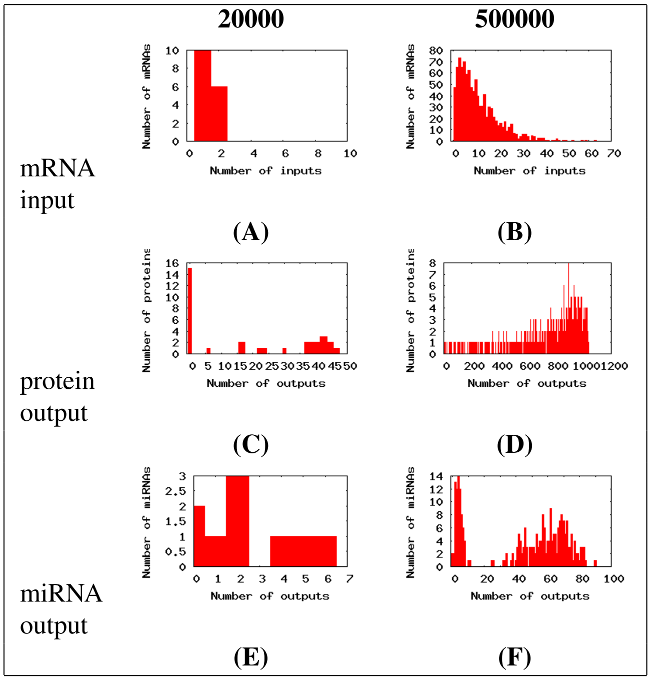


Figure 7: Histograms giving the degree distributions of the different species for a 20,000 base long genome, on the left column, and a 500,000 base long genome, on the right column. **(A)** and **(B)** mRNA input connectivity distribution. **(C)** and **(D)** Protein output connectivity distribution. **(E)** and **(F)** miRNA output connectivity distribution.

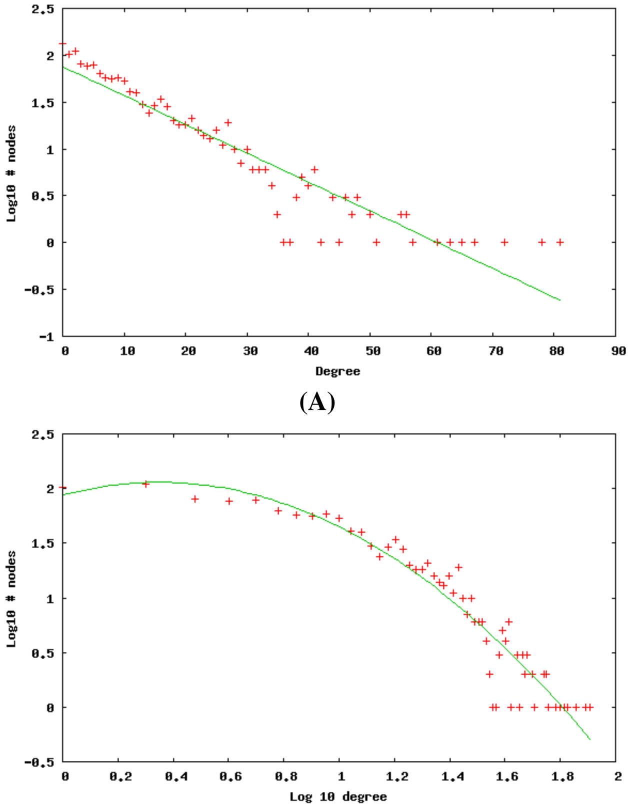
qualitatively alter the connectivity distributions, as can be seen by comparing the columns within Figure 6 and Figure 7, the two parameters 'binding threshold' and 'binding choice' determine the input connectivity distribution of genes and the output connectivity distribution of proteins. With a binding threshold of 29 and a 'max' or 'avg' binding choice, a heavily left skewed distribution with a fat tail is found for both the gene input and protein output connectivity (Figure 6B and 7D). These types of distributions are consistent with studies on other complex systems with directed graphs (Newman et al., 2001). The output distribution for the miRNA (Figure 7F) has two peaks. The left most one is the output distribution for the miRNA sequences that contain the subsequence 30 (U1 left), while the rightmost peak is the output distribution for the rest of the miRNA sequences. The reason for this is that the miRNA binds with mature mRNA sequences that have a low probability of containing the U1 left sequence because it is usually sliced with the introns. mRNA molecules have the U1 left sequence when this sequence is not followed by an U1 right sequence in the RNA transcript and therefore is not removed.

Regarding the gene output connectivity the shape of the distribution (Figure 6D) is always maintained, because the parameters that could alter it (promoter, promoter match, termination sequence, left and right u1 and u1 match) were kept fixed throughout the experiments. Figure 8 shows the linear-log and the log-log plots for the gene output. On the linear-log plot the distribution falls on a straight line, indicating an exponential decay of the distribution of connectivity. On the log-log plot the distribution decays faster than a power law would, since if the distribution had a power law tail it would fall on a straight line on this plot.

Some authors have shown evidence for the occurrence of three classes of small-world networks in real world networks: scale-free networks, characterized by a vertex connectivity distribution that decays as a power law; broad-scale networks, characterized by a connectivity distribution that has a power law regime followed by a sharp cutoff like an exponential or Gaussian decay of the tail; and single-scale networks, characterized by a connectivity distribution with a fast decaying tail, such as exponential or Gaussian. The question of why this range of possible structures for small-world networks exists is explained by the preferential attachment of new nodes that gives rise to the power law distributions. In the broad-scale and single-scale networks there are constraints limiting the addition of new links (Amaral et al., 2000). One constraint exists for the connection of new nodes to genes that could account for the faster decay of the tail of the gene output distribution. Genes have outputs to two different types of nodes: mRNA nodes and ncRNA nodes. While a gene only produces one mRNA, it can produce several ncRNAs and, as such, those connections are the most significant in terms of the overall degree distribution. Bigger genes have higher probability of producing several ncRNAs but their ability to produce them decays each time a ncRNA is produced because it shortens the sequence being searched (search for ncRNAs continues after the last found ncRNA). As with the output of the genes, mRNAs receive two kinds of inputs: each mRNA receives one single input from a gene and possibly several inputs from miRNAs, so the shape of the distribution (Figure 7B) depends, mainly, on the miRNAs. Parameters that influence the input distribution connectivity of the mRNAs are the genome size and the miRNA binding site size. Figure 9 shows the linear-log and the log-log plots of the mRNA input frequency. Similar to the output connectivity of genes, the linear-log plot falls on a straight line, with an exponential decay and the log-log decays faster than a power law would, therefore indicating that there may be constraints limiting the addition of new links between the miRNAs and the mRNAs. Since bigger mRNAs have higher probability of having more inputs, the shape of the input distribution may be greatly influenced by the size distribution of the mRNAs. The scale-free nature of these networks is thus arguable.

Conclusions

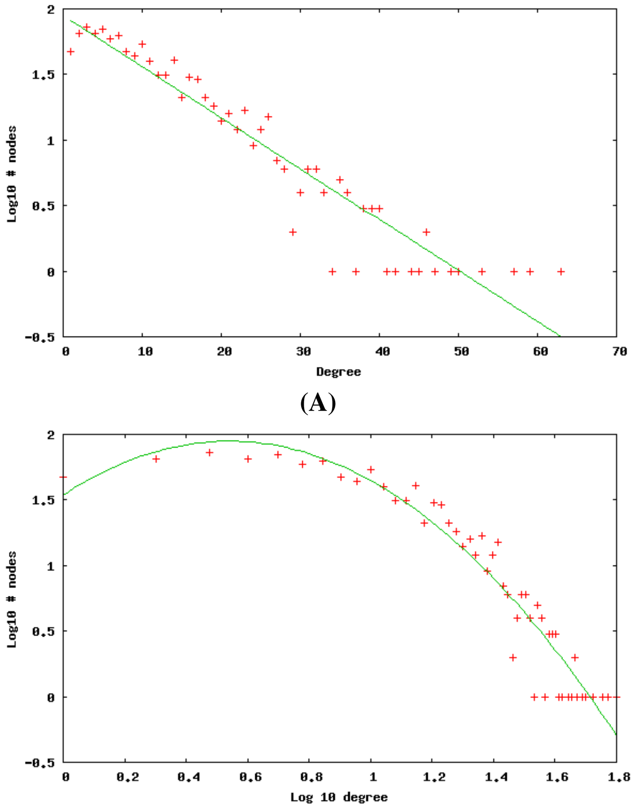
The proposed model, HeRoN, introduces a new level of biological detail. The separation of the several processes and the representation of all the products involved in heterogeneous networks allowed, in particular, to extend the model to incorporate a RNA interference mechanism. From the networks obtained some interesting observations about their topology and dynamics can be made. From the static point of view, although many authors claim that the genetic regulatory networks have scale-free topologies, most of them (Geard, 2004; Hallinan and Wiles, 2004; Watson et al., 2004; Willadsen and Wiles, 2003) are not based on experimental results for this concrete type of network but rather on other biological networks, such as the protein-protein interaction networks and metabolic networks. Others (Liebovitch et al., 2006) use experimental mRNA concentration data to extract the networks thus ignoring all regulation other than regulation of transcription initiation. This could lead to misleading results since the presence of an mRNA does not mean that the protein it produces, which is a potential transcription factor, is actually synthesized, as other regulation mechanisms, such as the miRNA negative regulation, may be acting on the mRNA. A model that does not account for these mechanisms may incorrectly assume regulations between genes that are actually regulated by other products. Another question of



(A)

(B)

Figure 8: **(A)** Linear-log plot of the gene output connectivity. **(B)** Log-log plot of the gene output connectivity.



(A)

(B)

Figure 9: **A** Linear-log plot of the mRNA input connectivity. **B** Log-log plot of the mRNA input connectivity.

importance is that most models make a one-mode projection of an intrinsically heterogeneous network, i.e., they assume a network where all nodes are genes and the edges between them represent regulation relations. When such one-mode projection is made some information is obviously discarded (Newman et al., 2001). As was observed in real-world statistical data of other problems, real complex systems do not always have power law distributions because they are subject to constraints. In the HeRoN model we could not only find degree distributions that are constrained but we also introduced the study of degree distributions for some intermediate products.

This work, and the corresponding model, can be extended in several directions. An important issue that must be addressed is its scalability. Experiments with genomes of realistic dimensions should be performed. The genome of the E.Coli would be a good starting point since it is one of the smallest (4,600,000 bases long) and the most studied genome available. Then, the model could be improved and made more biological sound, by taking into account aspects such as the concentration of products (a continuous variable) and the time delays involved.

Finally, it would be interesting to observe how the alternative splicing of genes could alter the output degree distribution of genes, proteins and miRNAs. The HeRoN model would have to be extended to include this feature. Although several interesting observations were made by analyzing the degree distribution of the nodes, there are several other statistical properties that could be used to better understand them. Future work should include a study on the clustering coefficient, the average path length between nodes, the distribution of component (subgraph) sizes and the existence and size of a giant-component.

References

Amaral, L., Scala, A., Barthélémy, M., and Stanley, H. (2000). Classes of small-world networks. In *Proceedings of the National Academy of Sciences (USA)*, volume 97, pages 11149–52.

Ando, S. and Iba, H. (2001). Inference of gene regulatory model by genetic algorithms. In Kim, J.-H., editor, *Proceedings of Congress on Evolutionary Computation*, volume 1, pages 712–719.

Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540):93–96.

Banzhaf, W. (2003). *Artificial Regulatory Networks and Genetic Programming*, pages 43–62. Springer Verlag.

D'Haeseleer, P. (2000). *Reconstructing Gene Networks fro Large Scale Gene Expressin Data*. PhD thesis, University of New Mexico (USA).

Geard, N. (2004). Modeling gene regulatory networks: fromn systems biology to complex systems. Accs draft technical report, University of Queensland (Australia).

Gonçalves, Â. and Costa, E. (2007a). A computational model for gene regulatory networks. Technical Report TR 2007/06, CISUC - University of Coimbra.

Gonçalves, Â. and Costa, E. (2007b). Heron: a computational model of gene regulatory networks. In Neves, J., Santos, M., and Machado, J., editors, *New Trends in Artificial Intelligence*, pages 288–299.

Gregory, R., Chendrimada, T., and Shiekhattar, R. (2006). Microrna biogenesis: isolation and characterization of the microprocessor complex. *Methods in Molecular Biology*, 342:33 – 47.

Hallinan, J. and Wiles, J. (2004). Evolving genetic regulatory networks usinf an artificial genome. In *Proceedings of the Seconf Conference on Asia-Pacific Bioinformatics*, volume 29, pages 291–296. Australian Computer Society.

Hoffman, M., Khrapov, M., Cox, J., Yao, J., Tong, L., and Ellington, A. (2004). Aant: the aminoacid-nucleotide interaction database. *Nucleic Acid Research*, 32:174–181.

Kauffman, S. (1993). *The Origins of Order: self-organization and selection in evolution*. Oxford University Press.

Liebovitch, L., Jirsa, V., and Shehadeh, L. (2006). *Complexus Mundi: emergent patterns in nature*, chapter Structure of genetic regulatory networks: evidence for scale-free networks. World Scientific.

Martens, J., Laprade, L., and Winston, F. (2004). Intergenic transcription is required to repress the saccharomyces cerevisiae ser3 gene. *Nature*, 429:571–574.

Newman, M., Strogatz, S., and Watts, D. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64.

Noman, N. and Iba, H. (2005). Inference in gene regulatory networks using s-system and differential evolution. In Beyer, H., editor, *Proceedings of the 2005 Conference on Genetic and Evolutinoary Computation*, pages 439–446. ACM Press.

Reil, T. (1999). Dynamics of gene expression in an artificial genome: implications for biolocal and artificial ontogeny. In Floreano, D., Nicoud, J.-D., and Mondada, F., editors, *Advances in Artificial Life: 5th European Conference on Artificial Life*, pages 457–466. Springer Verlag.

Sakamoto, E. and Iba, H. (2001). Inferring a system of differential equations for a gene regulatory network by using genetic programming. In Kim, J.-H., editor, *Proceedings of the 2001 congress on Evolutionary Computation*, volume 1, pages 720–726.

Watson, J., Geard, N., and Wiles, J. (2004). Towards more biological mutation operators in gene regulation studies. *BioSystems*, 76:239–248.

Willadsen, K. and Wiles, J. (2003). Dynamics of gene expression in an artificial genome. In Sarker, R., editor, *The 2003 Congress on Evolutionary Computation*, volume 1, pages 185–190.

Zhang, D. and Rosbash, M. (1999). Identification of height proteins that cross-link to pre-mrna in the yeast commitmem complex. *Genes Dev*, 13:581–592.