Random model for RNA interference yields scale free network

D. Balcan[1] and A. Erzan[1,2,a]

[1] Department of Physics, Faculty of Sciences and Letters, Istanbul Technical University, Maslak 34469, Istanbul, Turkey
[2] Gürsey Institute, POB 6, Çengelköy, 34680 Istanbul, Turkey

Abstract. We introduce a random bit-string model of post-transcriptional genetic regulation based on sequence matching. The model spontaneously yields a scale free network with power law scaling with $\gamma = -1$ and also exhibits log-periodic behaviour. The in-degree distribution is much narrower, and exhibits a pronounced peak followed by a Gaussian distribution. The network is of the smallest world type, with the average minimum path length independent of the size of the network, as long as the network consists of one giant cluster. The percolation threshold depends on the system size.

PACS. 87.10.+e General theory and mathematical aspects – 89.75.Hc Networks and genealogical trees

1 Introduction

Although biology on the whole is a "knowledge based" discipline, with a strong traditional bias towards a reverse engineering "form follows function" approach to evolution, prominent workers in the field have stressed that natural selection could very well have operated on complex structures already present in the pre-biotic world. Eigen [1] pointed out that non-linear out of equilibrium systems were capable of amplifying random fluctuations participating in feed-back loops, while Kauffman [2,3] introduced random Boolean networks as null models of the genetic regulatory mechanism, and showed that they could spontaneously give rise to a great degree of complexity. Meanwhile, within the last two decades, we have learned a great deal about Self-Organized Criticality [4], namely the spontaneous emergence of scale free structures in open systems far from equilibrium, driven by conserved fluxes. Thus, within a statistical mechanics context, it is much more natural to consider an ensemble of different states among which complex structures arise spontaneously. Evolutionary processes can be regarded as inducing dynamics on the distribution of states in this high dimensional phase space.

In this paper we introduce a null model for gene interactions resulting in the regulation of gene expression. This model is based on sequence matching on a random bit-string representation of the chromosome and is therefore radically different from the random Boolean networks which have been considered before [2,3,5–8]. It can be considered as an out of equilibrium system subject to a constant mutation rate, achieving a steady state invariant under further mutations.

[a] e-mail: `erzan@gursey.gov.tr`

We present simulation results which display many qualitative features of gene regulation networks found in nature [9,10]. We find that the in- and out-degree distributions are qualitatively different from each other, the out-degree distribution exhibiting power law decay $n(k_{\text{out}}) \sim k_{\text{out}}^{\gamma}$, with $\gamma = -1$ and log-periodic oscillations for relatively small k. The in-degree distribution, on the other hand, is much more localized.

The network has smallest world characteristics, with the cluster diameter, or the average minimum path length being essentially independent of the cluster size as long as the network consists of one cluster.

The average clustering coefficients for the in and out bonds have been calculated as 0.648 and 0.034 respectively. It is interesting to note that the clustering coefficients for the in-bonds behave much like for classical random networks [11], while the out-bonds have a clustering coefficient typical of scale free networks. [12,13]

In Section 2, we will first motivate and then define the model. In Section 3, we will present our simulation results. In Section 4, we consider a toy model which mimics some of the features we observe in our simulations, and indicate ways in which this toy model may be improved in order to provide insights into how our basic model works. Section 5 contains our conclusions and a discussion.

2 Modelling RNA interference

2.1 Genomic regulatory networks

Protein networks, which are an important component of transcriptional gene regulation networks [14,15], display a scale free structure, with the out-degree distribution

characterised by a power law $n(k) \sim k^{\gamma}$, with the exponent $\gamma \simeq -2.5$ [9,10]. Gene regulation networks actually operate at many different levels [16]. Post-transcriptional gene regulation may involve many different mechanisms, among which are the formation of secondary structures on mRNA, the action of RNA binding proteins, the control of RNA lifetimes, or RNA interference [17]. The latter is a mechanism where RNA strips may go and directly bind upon complementary segments on messenger RNA destined to be translated into some protein, thereby suppressing the production of this protein. Although we are not aware of a scaling analysis of post-transriptional gene interaction, it would be a fair guess to assume that interaction complexes of various sizes may arise in this type of interaction as well.

In many protein-protein or intra-genomic interactions, as well as the translation mechanism itself, essential lock-and key mechanisms are in operation. For normal translation to take place, rRNA in the ribosomes must be able to recognize and match the different amino acids and the corresponding three-letter anti-codon on the mRNA. In this, the rRNA is aided by the intermediary tRNA, which assumes a very specific three dimensional structure depending on the amino acid to which it binds. In RNA interference, the short interfering RNA (siRNA) strips bind onto complementary sequences on the mRNA, via Watson-Crick base pairing [17–19].

It should be realised that many of these lock-and-key mechanisms eventually match linear codes, even though this matching may take a few intermediary steps. The three dimensional structures (so called secondary or tertiary structures) which come into play in tRNA can also be said to be determined by the sequence of ribo-nucleic acids on the tRNA. Clearly the simplest is direct Watson-Crick base pairing between complementary nucleic acid sequences, and it is this latter, as it appears in RNA interference, which we will take to be the paradigm for our model. In fact we will further simplify the matching condition to consist of the identity relation rather than complementarity, since both are one-to-one, we believe that this should not change our results.

2.2 The model

The model is defined as follows. We postulate a "chromosome" to consist of a sequence of fixed length L, of independently and identically distributed random numbers with the probability distribution

$$P(x) = p\delta(x-2) + (1-p)/2[\delta(x-1) + \delta(x)]. \quad (1)$$

We define a "gene" to consist of a sequence of 0's and 1's situated between the ith and $i + 1$st occurance of the symbol "2," and we will denote the ith gene by

$$G_i = \{x_{i,1}, x_{i,2}, \ldots x_{i,\ell_i}\} \quad i = 1, \ldots, s \quad (2)$$

where $x_{i,\mu} \neq 2$, $\mu = 1, \ldots \ell_i$ and ℓ_i is the length of the ith gene. We have used periodic boundary conditions, but
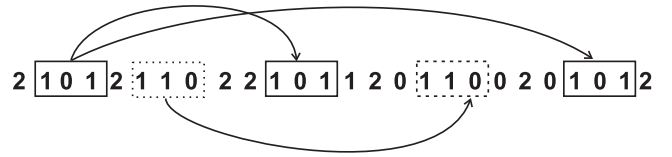

Fig. 1. The random sequence of symbols representing a chromosome. The "2" represents a start or stop sign for a gene, and the 0's and 1's code the genes. The arrows indicate that one gene sequence is embedded in another one.

one could just as well agree to end the chromosome always with a "2" at the $L + 1$st site. Then

$$\sum_{i}^{s} \ell_i = L - s, \quad (3)$$

where s is the number of genes (the number of times the symbol 2 appears) on the chromosome. Let n_ℓ be the number of genes of length ℓ. It obeys the sum rule

$$\sum_{\ell=0}^{L-s} n_\ell = s. \quad (4)$$

For a given number s of genes, the number of possible realisations of a given set $\{n_\ell\}$ is

$$K[\{n_\ell\}] = \frac{s!}{\prod_{\ell=0}^{L-s} n_\ell!}, \quad (5)$$

and the most probable distribution $\overline{n}(\ell)$ can easily be found by using Lagrange multipliers, to be $\overline{n}(\ell) = Lp^2(1-p)^{\ell}$, in the limit of large L.

With these definitions we obtain a sequence of genes, seperated from each other by the symbol 2. (See Fig. 1.) In case there are more than one consecutive 2's, they will be considered to bracket null genes. Clearly, for large L, the number of non-null genes, N, will fluctuate around $\overline{N} = Lp - Lp^2$.

Each of the non-null genes constitutes a node in our gene regulation network. The interactions do not depend on the proximity of the genes along the chromosome. We define the adjacency matrix w_{ij} by the matching condition such that

$$w_{ij} = \begin{cases} 1 & G_i \subset G_j \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

By $G_i \subset G_j$ we mean $x_{i,\mu} = x_{j,\mu+\nu}$ for $\mu = 1, \ldots \ell_i$ for at least one integer ν such that $0 \leq \nu \leq \ell_j - \ell_i$. Note that $w_{ij} = 1$ implies that $\ell_i \leq \ell_j$; in the case of the equality, $w_{ij} = 1$ if and only if the two sequences G_i and G_j are identical, i.e., congruent. Thus, two genes are said to interact if the sequence G_i occurs at least once as an unbroken subsequence of G_j, i.e., if one can be embedded in the other at least once.

Clearly this adjacency (or connectivity) matrix is directed. Moreover connectivity is "transitive" in the sense that $w_{ij} = w_{jk} = 1$ implies that $w_{ik} = 1$. The latter condition gives rise to a preferential attachment of incoming

bonds to large genes, while small genes have an enhanced distribution of out-bonds. However, we will see in the next section that the degree distribution is scale free for the out-bounds, but not for the in-bonds.

A simple argument tells us that our network is of the smallest world [12] type. If we take the out-bonds, we see that due to the transitivity, any two successive edges linking say vertex ij and jk, necessarily imply the existence of another directed edge ik. Thus $l_{\min}^{(o)} \equiv 1$, as long as the network consists of a single cluster. The in-bonds follow the same argument, giving $l_{\min}^{(i)} \equiv 1$ for a network with one giant cluster.

The question of whether a giant cluster always exists or whether we can identify the analog of a "percolation threshold" we will address in the next section, where we will also report numerical results for the minimum path length for undirected bonds (i.e., ignoring the directionality of the edges). We expect that the qualitative behaviour of the network should not depend on p as long as p is bounded away from the percolation threshold. Requiring the number of vertices, N, to be larger than unity, i.e., $\overline{N} = Lp - Lp^2 = Lp(1-p) > 1$ gives $p(1-p) > 1/L$ (however this lower limit turns out not to be tight enough, i.e., $p_c > 1/L$). The average gene size, again in the limit of large L is $\langle \ell \rangle = (1-p)/p$. The obvious requirement for a non-trivial network, that $\langle \ell \rangle > 1$ yields $p < 1/2$. We therefore expect to find scaling behaviour, if any, for $p_c < p < 1/2$.

The elements w_{ij} of the connectivity matrix are equal to unity with probabilities

$$P_{ij} = P(\ell_i, \ell_j), \tag{7}$$

which depend on the lengths (ℓ_i, ℓ_j) through ℓ_i and $\nu = \ell_j - \ell_i$. It is trivial to see that

$$P(\ell, \ell) = \left(\frac{1}{2}\right)^\ell. \tag{8}$$

We may make a mean field theory type of approximation to $P(\ell, \ell + \nu)$ by neglecting the correlations between overlapping subsequences of G_j, and obtain,

$$P(\ell, \ell + \nu) \simeq \left(\frac{1}{2}\right)^\ell (1 + \nu). \tag{9}$$

In the Appendix we have computed $P(\ell, \ell + \nu)$ explicitly for $\nu = 1, 2$. However, so far it has not been possible to extract the form of the degree distribution analytically.

The directed graph gives rise to different kinds of vertices, shown in Figure 2, which allow different classes of clustering coefficients C. Let $k_{\text{out}}(i)$, $k_{\text{in}}(i)$, $k(i)$, be, respectively, the out- degree, in-degree and the total degree of the vertex i. The clustering coefficient at a given vertex i is defined as

$$C_i = \frac{2E(i)}{k(i)[k(i) - 1]} \tag{10}$$

where $E(i)$ is the number of edges connecting the nearest neighbors of i. Thus it is the number of pairs of nearest
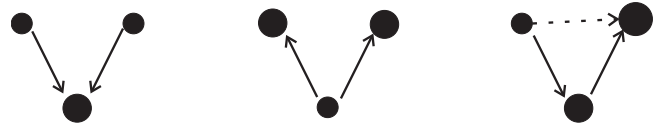

Fig. 2. Different kinds of vertices allowed on our directed random network. While the in or out-neighbors of a vertex may or may not be connected to each other, in the case of the third configuration, with a pair of mixed bonds, the neighboring nodes are necessarily connected as shown, due to transitivity.

neighbors directly connected to each other, normalized by the largest number of such connections possible. We may extend this concept to directed graphs and define,

$$C_{\text{out}}(i) = \frac{2E_{\text{out}}(i)}{k_{\text{out}}(i)[k_{\text{out}}(i) - 1]} \tag{11}$$

and similarly for C_{in}, where $E_{\text{out}}(i)$ (respectively $E_{\text{in}}(i)$) are the number of edges connecting out (in) nearest neighbors of i. Note that any pair of incoming and outgoing bonds passing through i necessarily defines a triangle, due to transitivity, as shown in Figure 2. Thus, the clustering coefficient may be conveniently decomposed as

$$C_i = \frac{k_{\text{in}}(i) \, k_{\text{out}}(i) + E_{\text{in}}(i) + E_{\text{out}}(i)}{k(i)[k(i) - 1]/2}. \tag{12}$$

This is a null model in the sense that no assumptions have been made as to the fitness of any particular type of interaction; the resulting interactions depend only on the random sequences coded in the genes and on the distribution of gene lengths. This random network provides the "tabula rasa" on which we assume natural selection will subsequently act. As such, it is of great interest to determine the properties of the null network, which turn out to be highly non-trivial. This is the task to which we turn in the next section.

3 Simulation results

To characterize the network defined by our model, we generated random chromosomes as defined above. The statistical properties of the network obtained from the totally random chromosome were checked to be invariant under a constant mutational load, with a mutation probability of 0.01. A mutation is affected as follows. If the symbol (x) occupying the site to be mutated happens to be a 0 or 1, then it is flipped, i.e., we set $x = \text{mod}_2(x + 1)$. If $x = 2$, then it exchanges places with either its right or left nearest neighbor, with equal probability. It should be noted that the first case corresponds to a substitutional mutation, whereas the second to a shifting of the position of the start sign which shifts the reading frame of the gene. The latter gives rise to far-reaching modifications, since the three-letter codes corresponding to the different amino acids will be completely modified along the whole gene if the reading frame is shifted.

The most remarkable property of the random network generated in this way is that it has a scale free out-degree
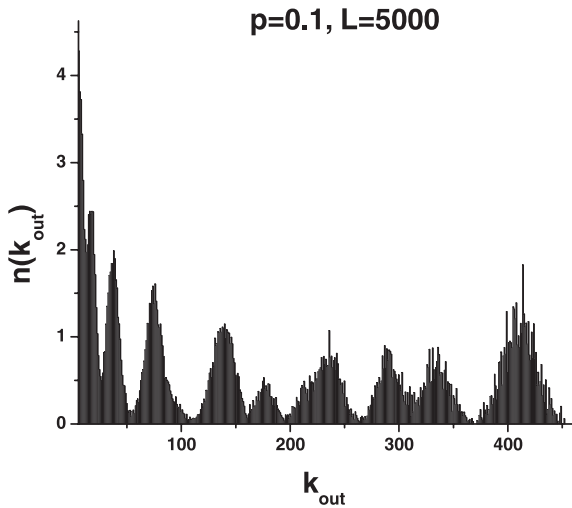
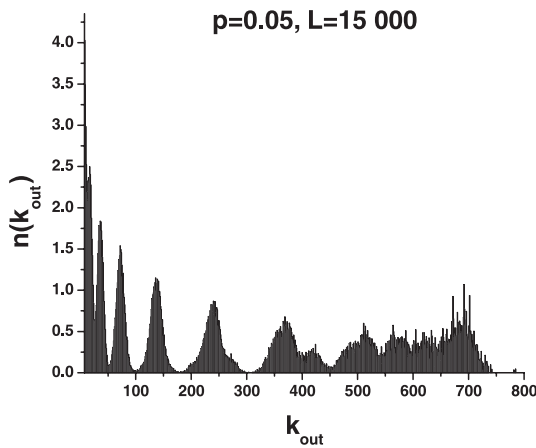**Fig. 3.** The out degree distribution, for $p = 0.1$ and $L = 5000$.

Fig. 4. The out degree distribution for $p = 0.05$ and $L = 15000$.
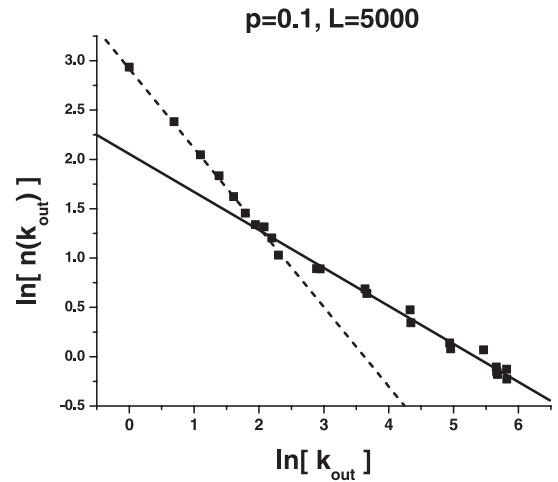

Fig. 5. The log-log fit to the envelope of the out-degree distribution shown in Figure 3. The slopes of the dashed and continuous lines are -0.9 and -0.39, respectively.
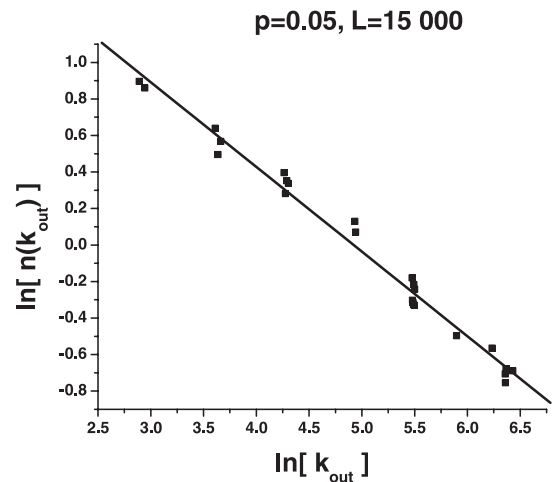

Fig. 6. The log-log fit to the envelope of the out-degree distribution for $p = 0.05$, $L = 15000$ (Fig. 4). The fit is to a slope of -0.46.

distribution, and a qualitatively different in-degree distribution which is much less broad, with a pronounced narrow peak followed by a Gaussian peak. These qualitative behaviours match the results found on protein and genomic networks to a surprising extent [9,10].

We display in Figures 3, 4 the out-degree distribution $n(k_{\mathrm{out}})$ for two sets of parameters, namely $L = 5000$, $p = 0.1$ and for $L = 15000$ $p = 0.05$, with, the number of genes fluctuating around $\overline{N} = 450$ and $\overline{N} = 712$, respectively. The data have been averaged over 500 independent realizations. The distributions, which are strongly log periodic, with a power law envelope $n_m(k_{\mathrm{out}}) \sim k_{\mathrm{out}}^{\gamma'}$, have the same characteristic behaviour for the two cases. The fits to the envelope of the peaks are shown in Figures 5, 6. We see a marked break in the log-log fit for larger p, with a crossover from $\gamma' \simeq -1$ to $\gamma' = -0.45 \pm 0.06$ at about $\ln(k_{\mathrm{out}}) = 2.5$. This crossover behavior is just off-scale in Figure 6, as can be seen in Figure 7, with the incipient scaling with a power of -1 again extending out to $\ln(k_{\mathrm{out}}) \simeq 2$.

It is also worthwhile mentioning that the integral over the peaks to the rightmost of the out-degree distribution simply gives the network size N, as it should, because these peaks come from the smallest genes of unit length which are embedded in all the other genes. Introducing a larger lower cutoff to the gene size (i.e., a cutoff bigger than unity) would somewhat restrict the range of k_{out}, but also yield cleaner scaling behaviour for large values of k_{out} by effectively eliminating these peaks.

It should be noticed that for $k_{\mathrm{out}} \leq 100$, the dips in the degree distribution are about evenly spaced on the logarithmic scale, with a scale factor of $1/2$. We conjecture that each individual peak corresponds to a particular gene-length ℓ. (This conjecture is borne out by comparing the integrals under the peaks with n_ℓ for $p = 0.1$, $L = 5000$ but is much more noisy for $p = 0.05$, $L = 15000$.) Increasing ℓ by unity, exactly halves the leading contribution to the to k_{out}, from genes of length ℓ.

To get a better grip on the degree distribution, we numerically integrated the curves in Figures 3, 4. The result
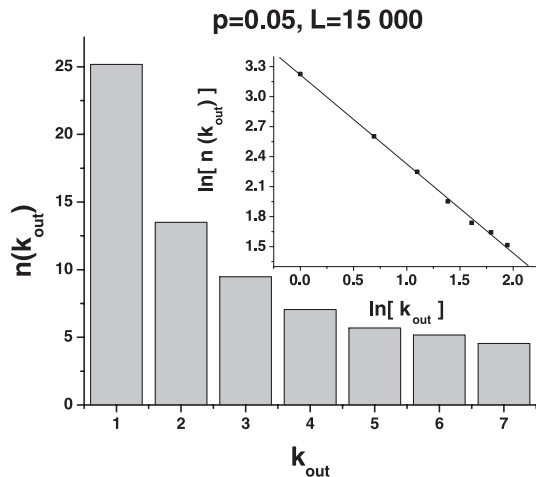
Fig. 7. The initial peak of the out-degree distribution for $p = 0.05$ and $L = 15000$, and the double logarithmic graph of the same, showing incipient power law behaviour with the power -0.89.
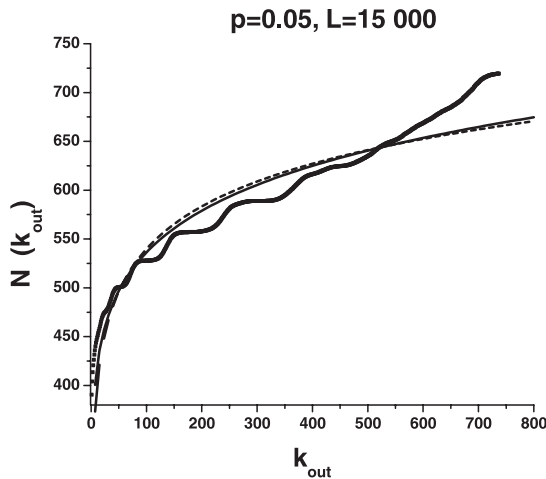

Fig. 8. The wavy line is the cumulative out degree distribution for $p = 0.05$, $L = 15000$. The dashed line is a logarithmic fit, while the continuous one $\sim k_{\mathrm{out}}^{0.11}$.

is shown for $p = 0.05$ in Figure 8. We find that

$$\mathcal{N}(k_{\mathrm{out}}) \equiv \int^{k_{\mathrm{out}}} n(z)dz \sim \ln(k_{\mathrm{out}}), \qquad (13)$$

or a very small power, $\sim k_{\mathrm{out}}^{\gamma+1}$ where numerically $\gamma + 1 < 0.1$, at least for k_{out} that are not too large. Thus we are led to believe that the overall out degree distribution scales like

$$n(k_{\mathrm{out}}) \sim k_{\mathrm{out}}^{\gamma} f\left(k_{\mathrm{out}}/\overline{N}\right), \qquad (14)$$

with $\gamma = -1.0 \pm 0.1$. The scaling function $f(x) \sim$ const. for $x \ll 1$, and is log-periodic for intermediate values of its argument. Scaling breaks down for $x \sim O(1)$. Note that γ is smaller in absolute value than that reported for real gene regulation networks [9,10,15], thus evolution seems to have narrowed the distribution somewhat, restricting the number of genes that may be affected by any one gene.
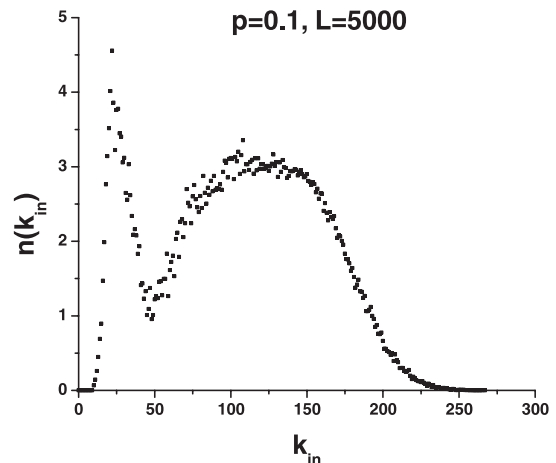

Fig. 9. The in-degree distribution for $p = 0.1$ and $L = 5000$. The in-degree distribution is much narrower that the out-degree distribution and displays two peaks.
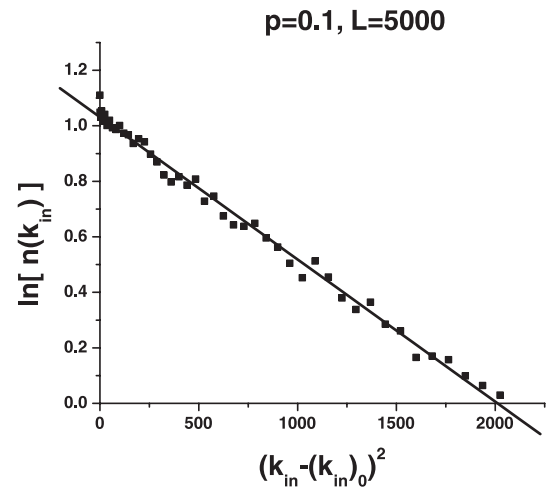

Fig. 10. The second peak of the in-degree distribution for $p = 0.1$ and $L = 5000$ can be fit to a Gaussian.

The in-degree distribution, which we display in Figure 9 shows two peaks. The second can be fit to a Gaussian, as shown in Figure 10. The first peak is more skewed than a Poissonian, but may be fitted reasonably well by a distribution of the form $f(x) \sim (x - x_0)^{\delta} \exp[-\xi(x - x_0)]$, where $\delta = 2.2$ and $\xi = 0.15$. (Fig. 11).

The total degree distribution again displays a modulated structure, as can be seen in Figure 12.

We have already discussed that the network will behave trivially either for very small N (all very large genes with very small probabilities for interaction) or for large p (i.e., $p > 1/2$) where most of the chromosome will be occupied by non-coding partitions (the symbol 2), and many null genes. We have determined the threshold value, p_c, below which the the genes are too few and too long, such that the single giant cluster breaks up into more than one component. The results are reported in Table 1, for L ranging from 15000 to 1000, and undirected edges. Although we have only a few points, the dependence (see Fig. 13) fits a power law, with $p_c \sim L^{-\alpha}$, with $\alpha \simeq 3/4$,
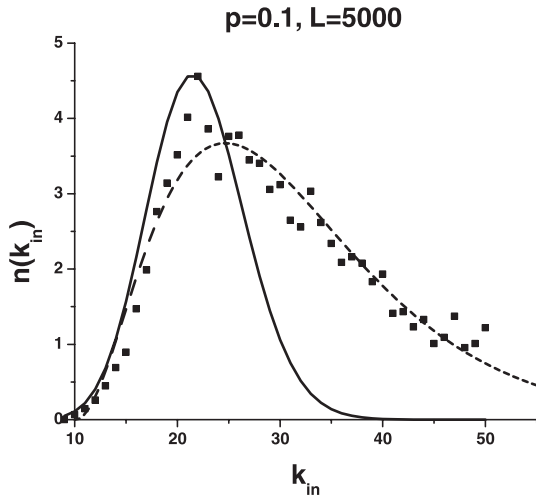
Fig. 11. The first peak of the in-degree distribution, with the tail of the Gaussian distribution subtracted. The fits are to a Poisson distribution (continuous line) $\exp(-z)z^k/k!$ with $z = \overline{k}$ computed from the data points, and the function $f(k) = 0.09(k-10)^{2.2}\exp[-0.15(x-10)]$ (dashed line).
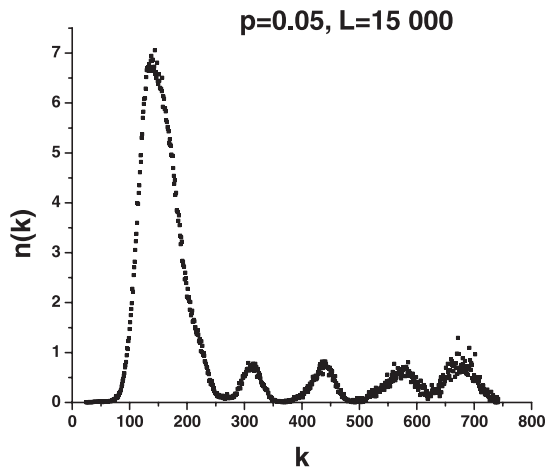

Fig. 12. The total degree distribution for $p = 0.05$, $L = 15000$.

Fig. 13. Dependence of the "percolation threshold" on L. The power law fit gives $p_c \sim L^{-3/4}$.

Table 1. The "percolation threshold" for strings of different length L. The average number of non-null genes at the percolation threshold is also reported.

L	$\langle N \rangle$	p_c
15000	177	0.012
5000	112	0.023
1500	83	0.059
1000	54	0.086

Table 2. The average minimum path length for undirected edges, with $p = 0.05$, for different L.

L	$\langle N \rangle$	$\langle l_{\min} \rangle$
15000	717	1.669
10000	479	1.670
5000	239	1.670
2500	120	1.670
2000	95	1.669
1750	83	1.672
1650	79	1.673

within the range explored. In the limit of $L \to \infty$, $p_c \to 0$, but less fast than $1/L$.

On the other hand our expectations that there is a reasonably large interval of p's ($p_c < p < 1/2$) where the results do not depend quantitatively on the precise value of p are borne out, as can be seen from our calculations for the minimum average path length for undirected edges.

We argued in Section 2 that as long as there is a single giant cluster, the directed minimum path length is identically 1. For undirected paths this is no longer true and we must determine $\langle l_{\min} \rangle$ numerically. Nevertheless, we can show that $l_{\min} \leq 4$. Note that for $p > p_c$, there will be an abundance of short genes of unit length, which will contain the symbol 0 or 1 with equal probability. These will have outgoing edges to all the typical genes which have an admixture of 0s and 1s, so that most such genes will be certainly linked by paths of length at most equal to 2. The "worst case" is that of atypical cases of all 1's all
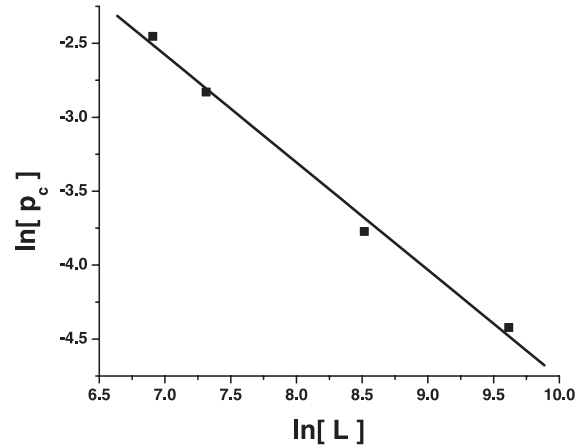
or all 0's which require at least one intermediate typical gene to link up ($11111 - 1 - 011100 - 0 - 00000$), which gives $\max(l_{\min}) = 4$. We have calculated $\langle l_{\min} \rangle$ for fixed p and different L. We consider undirected edges and find that it at most depends very weakly on N, the number of vertices. We find, e.g., for L ranging from 1650 to 15000, and $p = 0.05$ that $\langle l_{\min} \rangle$ ranges only from 1.673 to 1.699. (Tab. 2)

Fixing the length L and varying p (Tab. 3) shows again that as long as $\langle l_{\min} \rangle$ is defined, i.e., just above the "percolation threshold" it is already very close to the value it will have for larger p.

The clustering coefficients are markedly different for the in- and out-degree distributions, as well as the total connectivity. We find, for $p = 0.05$ and $L = 15000$, $\langle C_{\text{out}} \rangle = 0.034$, $\langle C_{\text{in}} \rangle = 0.648$ and $\langle C \rangle = 0.534$. We have verified in the last case that this clustering coefficient is reproduced by computing the average total degree

Table 3. The average minimum path length $\langle l_{\min} \rangle$ for undirected edges, for different p and L. Note that $\langle l_{\min} \rangle$ varies very little over the entire range from $p \simeq p_c$ to $p \simeq 1/2$.

p	L	$\langle N \rangle$	$\langle l_{\min} \rangle$
0.013	15000	194	1.858
0.014	15000	209	1.851
0.016	15000	237	1.84
0.018	15000	267	1.82
0.02	15000	295	1.81
0.05	15000	717	1.669
0.1	10000	452	1.543
0.48	1500	374	1.360

per node and dividing by the total number of nodes, i.e., $\langle C \rangle \simeq \langle k \rangle / N$. Thus, the total connectivity behaves very much like in classical random graphs.

4 Analytical results on toy model

In this section we would like to present a number of the results which may be obtained from a hierarchical model. In particular, we would like to exploit the transitivity property, to compute certain properties of a toy version of our original model.

Let us consider a tree network, with a branching ratio b. Take all connections emanating from nodes at the mth generation to the ones at the $(m+1)$st generation, to be directed. If one applies the transitivity rule, then this automatically generates further connections that link directly a node at the mth generation to all the nodes below it on the same branch of the tree. Then, the out-degree (we will drop the index "out" from now on) of a node at the mth level will be

$$k(m) = \sum_{m'=0}^{M-m-1} b^{m'} = \frac{b^{M-m} - 1}{b - 1}, \qquad (15)$$

where M is the total number of levels and $k(m) \sim b^{M-m}$ for large $M - m$, which we may safely assume. The frequency of the nodes at the mth level is b^m. This gives, for the out-degree distribution,

$$n(k) \sim \text{const.} \left(\frac{1}{k} \right). \qquad (16)$$

We now introduce redundancies within the same level. Let a fraction r of the b downstream neighbors emanating from any node be identical to each other, with $r = a^m$, $b^{-1} < a < 1$ being some constant. The nodes that are identical within any given level are connected to each other by two-way bonds within the sub-branch where they are located. I will call this subset of nodes the clones at any given level. This means that the out-degree k of any one of these nodes is now $k' = (a^m b) b^{M-m}$, since each interconnected node inherits all the downstream neighbors of the $a^m b$ clones to which it is connected. The interconnections between nodes further downstream do not introduce

any further change in $k'(m)$, since all the nodes that they can connect are already connected to the nodes at level m. The number of clone-nodes (not all identical to each other, but only in groups of $a^m b$) is clearly $a^m b^m$. Then, the out-degree distribution of the nodes belonging to the clone-sets becomes,

$$n(k') = b^{\gamma-1} \left(\frac{k'}{N} \right)^{\gamma} \qquad (17)$$

where $N = b^M$ and with

$$\gamma = \frac{\ln a + \ln b}{\ln a - \ln b}. \qquad (18)$$

Note that under our assumptions above, $-1 < \gamma < 0$. For large enough k this then becomes the *leading* contribution to the out degree distribution, rather than $\gamma = -1$.

5 Conclusions and discussion

The model we have presented for the scaling behaviour of gene interaction, more specifically RNA interference, turns out to be a rich model which is worthwhile to investigate in its own right, as well as providing a highly structured starting point on which further evolutionary pressures could act.

We have checked that the statistical properties of the network are robust under random point mutations. Thus, the scale free random network may be seen as the attrator for a process starting from, say, a sequence of uniform genes, subjected to random point mutations over a very long time period. The scale free steady state can be considered as the outcome of a process of self-organization under random perturbations.

In conclusion it is worthwhile to note that in RNA interference, longer RNA chains are cut up by the "Dicer" enzyme into smaller segments, siRNA [17], which can be conveniently matched with complementary sequences in a larger number of mRNAs. This trick would change the power of the out-degree distribution from that found here.

Appendix

We would like to compute the probabilities $P(\ell, \ell')$ in equation (7). For successive $\nu = \ell' - \ell > 0$, it is convenient to define subsequences of G_j of length ℓ_i with a shift λ, $0 \leq \lambda \leq \nu$, such that $g_{ij}^{(\lambda)} \equiv \{x_{j,1+\lambda}, x_{j,2+\lambda}, \ldots x_{j,\ell_i+\lambda}\}$. Then, we can obtain $P(\ell, \ell')$ in terms of joint probabilities

like $\mathcal{P}(G_i \neq g_{ij}^{(0)}, G_i = g_{ij}^{(1)})$, etc. Thus,

$$P(\ell, \ell) = \mathcal{P}\left(G_i = g_{ij}^{(0)}\right)$$

$$P(\ell, \ell+1) = \mathcal{P}\left(G_i = g_{ij}^{(0)}\right) + \mathcal{P}\left(G_i \neq g_{ij}^{(0)}, G_i = g_{ij}^{(1)}\right)$$

$$\cdots$$

$$P(\ell, \ell+\nu) = \mathcal{P}\left(G_i = g_{ij}^{(0)}\right) + \ldots$$
$$+ \mathcal{P}\left(G_i \neq g_{ij}^{(0)}, G_i \neq g_{ij}^{(1)}, \ldots, G_i = G_{ij}^{\lambda}\right)$$
$$+ \ldots$$
$$+ \mathcal{P}\left(G_i \neq g_{ij}^{(0)}, G_i \neq g_{ij}^{(1)}, \ldots, G_i = g_{ij}^{(\nu)}\right).$$

Let us display $\mathcal{P}(G_i \neq g_{ij}^{(0)}, G_i = g_{ij}^{(1)})$ for greater clarity. We have

$$\mathcal{P}\left(G_i \neq g_{ij}^{(0)}, G_i = g_{ij}^{(1)}\right) =$$
$$\left\langle \left(\sum_{\mu=1}^{\ell} \prod_{m=1}^{\mu-1} \delta_{x_{i,m} x_{j,m}}\right) \left(1 - \delta_{x_{i,\mu}, x_{j,\mu}}\right) \prod_{n=1}^{\ell_i} \delta_{x_{i,n} x_{j,n+1}} \right\rangle$$
$$= 2^{-\ell_i}\left[1 - 2^{-\ell_i}\right].$$

In particular we find

$$P(\ell, \ell+1) = 2^{-\ell}\left[1 + 1 - 2^{-\ell}\right]$$
$$= 2^{-\ell}\left[2 - 2^{-\ell}\right]$$
$$P(\ell, \ell+2) = 2^{-\ell}\left[3 - 2^{-\ell+1} - \frac{1}{3}\left(1 - 2^{-2\ell}\right)\right].$$

Note that the probabilities $\mathcal{P}$ do not depend upon the particular gene, except through the lengths of the sequences, as the genes do not overlap, and each of the symbols are independently and identically distributed, with equal probability, over 0 and 1.

References

1. M. Eigen, Naturwissenschaften **10**, 58 (1971)
2. S.A. Kauffman, J. Theor. Biol. **22**, 437 (1969)
3. S. Kauffman, *The Origins of Order: Self-organization and Selection in Evolution* (Oxford University Press, N.Y. 1993)
4. P. Bak, C. Tang, K. Wiesenfeld, Phys. Rev. Lett. **59**, 381 (1987); Phys. Rev. A **38**, 364 (1988); K. Sneppen, P. Bak, H. Flyvberg, M.H. Jensen, Proc. Natl. Acad. Sci. **92**, 5209 (1995)
5. H. Atlan, F. Fogelman-Soulie, J. Salomon, et al., Random Boolean Networks Cybernet. Syst. **12**, 103 (1981)
6. B. Derrida, G. Weisbuch, J. Phys. France **47**, 1297 (1986)
7. B. Derrida, Y. Pomeau, Europhys. Lett. **1**, 45 (1986)
8. B. Derrida, H. Flyvberg, J. Phys. A **20**, L1107 (1987)
9. S. Maslov, K. Sneppen, Science **296**, 910 (2002)
10. S. Maslov, *Midterm Conference COSIN on "Growing Networks and Graphs in Statistical Physics, Finance, Biology and Social Systems", Rome, 1–5 September 2003* (see this issue)
11. P. Erdös, Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960)
12. S.N. Dorogovstsev, J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002)
13. R. Albert, A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002)
14. N.A. Campbell, J.B. Reece, *Biology* (Addison Wesley, San Francisco, 2002), pp. 363–368
15. R.V. Sole, R. Pastor-Satorras, *Complex Networks in Genomics and Proteomics*, edited by S. Bornholdt, H.G. Schuster, *Handbook of Graphs and Networks* (Wiley-VCH Verlag, Berlin, 2002)
16. T. Maniatis, R. Reed, Nature **416**, 499 (2002)
17. G.J. Hannon, Nature **418**, 244 (2002)
18. B. Alberts et al., *Molecular Biology of the Cell* (Garland Science, N.Y., 2002), p. 375 ff
19. B. Alberts et al., Op. cit., pp. 451–452