Article

Diffusion limited aggregation and the fractal evolution of gene promoter networks

Preston R. Aldrich
Department of Biological Sciences, Benedictine University, Lisle, IL 60532, USA
E-mail: paldrich@ben.edu

Received 27 June 2011; *Accepted 31 July 2011*; *Published online 1 September 2011*
IAEES

Abstract
Gene promoter networks (GPNs) are systems-level representations of the base pair-sharing relationships (graph edges) among promoters (graph nodes). It has been shown in the bacterium *E. coli* that these networks can contain a fractal nucleus of strong associations suggesting a self-organizing complexity. Here I report results of twenty seven in silico simulations for a diffusion limited aggregation model which accounts for much of the fractal structure previously observed in GPNs. Parameters varied in the model included (*a*) the frequency of gene duplication events, and the extent of (*b*) attraction and (*c*) repulsion presented by the DNA-protein binding chemistry. Both duplication and attraction had significant effects on fractal topology of the GPN nucleus, whereas repulsion due to DNA-protein binding chemistry did not, at least for the levels explored in these simulations. Since repulsion is thought to be a key feature of fractal networks, it is likely that the repulsion in GPNs arises from the sparseness of the promoter space. The generation of a finite random set of promoters leads to sparse occupancy of promoter space which itself presents a considerable repulsion away from the consensus motif, working against the DNA-binding protein's efforts to organize the system of promoters over evolutionary time. This interplay between attractive and repulsive forces in a GPN is sufficient to generate a fractal topology.

Keywords regulon; network; transcription factor.

1 Introduction

The binding of RNA polymerase is an important stage in gene expression. RNA polymerase will readily bind to DNA non-specifically although transcription is generally inefficient in such cases, and helper proteins such as σ-factors in bacteria and transcription factors more generally serve to specify and optimize the binding to particular places in the genome (Weaver, 2012; Hinckle and Chamberlin, 1972). These binding sites are referred to as promoters. Some transcription factors bind to numerous promoters in the genome defining a regulon of genes. Such global regulators are influential in committing the genome to broad-scale gene expression events such as in response to heat shock or nitrogen depletion.

GRNs and GPNs are two different systems-level views of transcriptional control structure (Fig. 1). The GRN, or Gene Regulatory Network (e.g., Davidson and Levin, 2005), represents genes as nodes and regulatory interactions as directed edges or arcs. Here, a transcription factor points to the genes that it regulates within the network. The GPN, or Gene Promoter Network (Aldrich et al., 2010), represents transcription factor binding

sites (i.e., promoters) as nodes and the extent of base pair sharing between sites as weighted edges in the network. GPN edges are undirected and any promoter can connect to any other promoter in the network provided they share at least one position-specific base pair in common. Complexity is reduced by thresholding, removing the weak edges representing low-bp sharing.

Prior work (Aldrich et al., 2010) with σ-factor GPNs in *E. coli* revealed a fractal nucleus of strong associations among the promoters of several regulons (e.g., Fig. 2). Promoter sets identified by RegulonDB (Gama-Castro et al., 2008) did not include a consensus motif for the regulons examined, instead the promoters exhibited considerable sequence variation and clustered around the non-existent consensus motif. Self-similar structure became evident after the removal of weak edges, particularly at the upper phase transition. The phase transition is the point at which the largest connected component (the largest set of nodes that remain interconnected) abruptly declines in size (number of nodes) as more edges are removed.

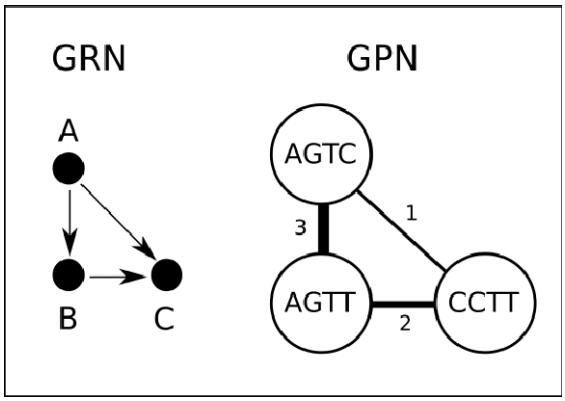

Fig. 1 Basic structure of a GRN (gene regulatory network) and a GPN (gene promoter network). The GRN is a directed network in which nodes represent genes or gene products wherein some gene products (proteins) can act as regulators of other genes (in which case they are termed transcription factors). Regulatory relationships are denoted by arrows. In the figure, gene A regulates both genes B and C, while gene C regulates neither A nor B. In a GPN, nodes represent gene promoters, the cis-regulatory regions that bind to regulatory proteins such as transcription factors or σ-factors in bacteria. The network is undirected with weighted edges denoting the extent of base pair-sharing between promoters. The emphasis of the present study is the GPN.

Aldrich et al. (2010) showed that the GPN nucleus can have several interesting features. (*a*) It can exhibit strong visual symmetry. (*b*) The central region of the GPN tends to be vacant suggesting some source of repulsion is at work. (*c*) A symmetric nucleus generally has a significantly fractal topology as measured by the box covering/network coloring method (Song et al., 2005, 2007). (*d*) The fractal dimension of the set of nuclei examined was on average $d = 1.731$. (*e*) The position of the upper phase transition occurred at the place expected for a random graph according to percolation theory (Erdös and Rényi, 1960). (*f*) Promoter abundances scaled as a power-law in the genome. Given these points of evidence, and since the development of fractal structure in networks is thought to arise in response to repulsive forces (Song et al., 2006), Aldrich et al. (2010) proposed the diffusion limited aggregation (DLA) model as a mechanism for the evolution of GPNs. In the generic 2-dimensional DLA model proposed by Witten and Sander (1981), particles diffuse randomly as a Brownian motion, occasionally sticking to a growing cluster. Aldrich et al. (2010) recognized this as growth through preferential attachment, but not to the oldest particles as in a scale-free model of network growth (Barabasi and Albert, 1999). Instead, particles attach preferentially to the growing arms of the cluster since the

arms physically obstruct access to the central region. It appears as though the center repulses any new additions. A fractal dimension of $d = 1.7$ is typical of systems arising by DLA (Liebovitch, 1998).

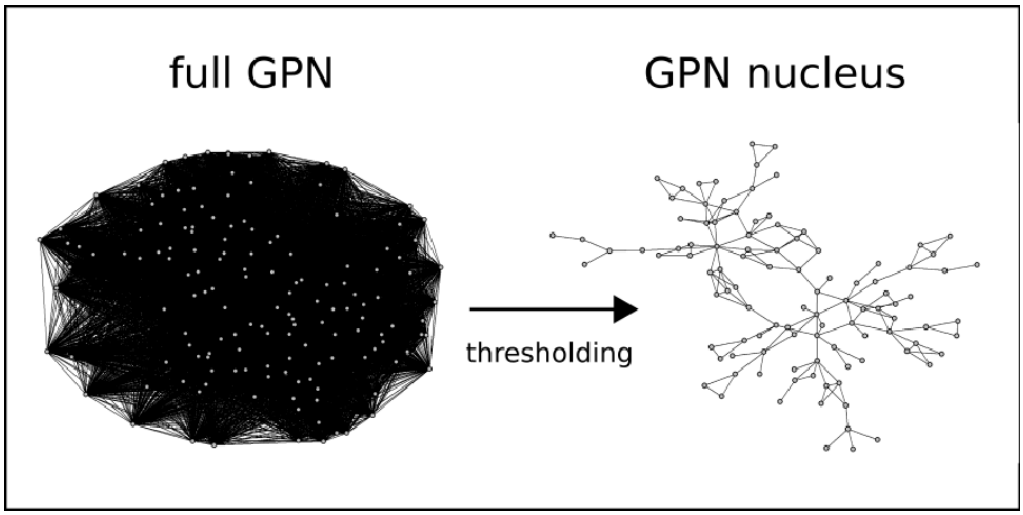

Fig. 2 Full GPN and its nucleus for the σ-54 regulon built from predicted promoters obtained from the RegulonDB database (Gama-Castro et al., 2008). The original GPN (left) contained 154 nodes and the largest connected component evaluated at the phase transition gave a nucleus of 105 nodes (right) after removal of all edges with weights < 10 bp (i.e., $m = 10$). Networks were visualized using the program Pajek (Batagelj and Mrvar, 1998) and the Kamada-Kawai projection (Kamada and Kawai, 1989).

Aldrich et al. (2010) conjectured that a similar DLA process might apply to GPN growth. Here promoters arise randomly and add preferentially to the periphery of the network (Fig. 3, DLA model), not to the central hub (SF model). In addition to the numerical bias that there generally are more peripheral nodes to which to attach, there also are biochemical aspects of the system that could contribute to the development of a fractal nucleus. A GPN growing by DLA could be regulated by both repulsive and attractive forces, mediated on the micro-scale through DNA-protein binding chemistry, and on the macro-scale by population-level fitness, organized around a consensus promoter. The consensus would form an attractor in promoter space because it represents the optimal binding chemistry for the DNA-binding protein. It is known that promoters whose sequence departs too far from the consensus would be weak and ineffective in its binding capacity (Hawley and McClure, 1983). Yet it has been observed that consensus and canonical motifs rarely participate directly in transcription perhaps because they bind the transcription factor or σ too firmly, preventing promoter clearance and elongation (Hawley and McClure, 1983; Huerta and Collado-Vides, 2003; Ellinger et al., 1994). The resulting lowered population-level fitness would repulse additions from the GPN center. These dynamics are analogous to the inter-atomic attractive and repulsive forces that include the van der Waals interactions.

Here I test the DLA model of preferential attachment through in silico simulations.

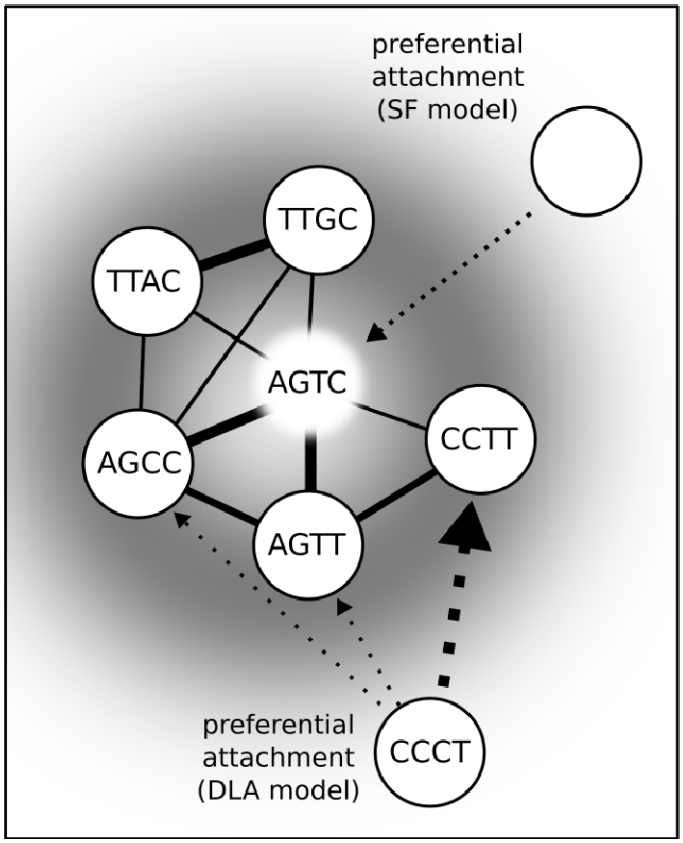

Fig. 3 Growth of a GPN through preferential attachment. Under the classic scale-free (SF) model of network growth, a new node would attach preferentially to the node with highest degree. In the diffusion limited aggregation (DLA) model, a new node attaches to the set of promoters that overlap in base composition. Even without consideration of DNA-binding chemistry, there is a greater probability of random matches to the numerous peripheral nodes in the network compared to the one central hub, which may even be absent. In any case, the binding chemistry of the transcription factor serves to organize promoter space around this motif forming an attractor in promoter space. In the figure, the dark torus marks the region of optimal promoter similarity assuming both attraction from the central consensus and repulsion from too close a match (strong version of DLA).

2 Methods

2.1 Random promoters and GPN formation

Random promoter networks, or random grammar networks, were introduced in the first study of fractal GPNs (Aldrich et al., 2010).

Nodes are formed in a random GPN by generating a set of n promoters, each through F (footprint size) random draws from a uniform base distribution (A, C, G, T). In the present study I emulated the footprint of the σ^{54} system (data as provided by RegulonDB (Gama-Castro et al., 2008)), fractal structure reported by Aldrich et al. (2010). This σ-factor has F = 11-bp footprint with a –10 (5 bps) box and –35 (6 bps) box separated by a short spacer (5-6 bps). Spacer sizes were drawn from the observed distribution of sizes in the σ^{54} system. Each full random GPN contained 500 random promoters.

Edges are formed as pairwise non-zero measures of similarity between promoter pair sequences i and j (A_{ij}) evaluated simply as the number of base pairs shared. These weighted edge values are used to form the adjacency matrix, $\boldsymbol{A}$. A network or graph G is then generated based on the matrix $\boldsymbol{A}$. Networks were produced and analyzed using Python and networkx (Hagberg et al., 2008; http://networkx.lanl.gov/), though visualized using Pajek (Batagelj and Mrvar, 1998; http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

2.2 Thresholding of GPNs

Thresholding was achieved through serial *m*-slices (de Nooy et al. 2005). In an *m*-slice one removes all edges from a weighted graph *G* below a critical threshold *m* (where $1 < m < F$ = footprint size in bp). In serial extractions, edges were removed based on a sliding *m*-threshold value. At each step as *m* increased to *F*, the largest connected component was extracted from *G* and evaluated for the number of nodes and edges. The largest connected component is the maximal subgraph *G'* containing all nodes still interconnected after removal of edges not meeting the threshold criterion. A largest connected component is termed a giant component when it contains at least half the nodes present in the full graph *G*. The giant component of a random graph fractures (or emerges) near the phase transition (Erdös and Rényi, 1960).

2.3 Assessment of fractal structure

I wrote Python/networkx script that implemented the renormalization procedure of Song et al. (2007) to evaluate the fractal dimension of the largest collected components extracted from the GPNs at the phase transitions. The technique is a graph coloring exercise founded on the traditional box-covering method of fractal measurement. In brief, for a given box length (l_B), or shortest path length between nodes, each node is colored in a fashion such that neighbors of like color are no further away than the current box length. Then the network is renormalized by collapsing adjacent nodes into a single node if they share the same color. This enforces the graph coloring rule that no two adjacent nodes can share the same color. The value N_B then gives the minimum number of boxes of length l_B required to cover the graph of N_B nodes, and is equal to the graph size (node count) following renormalization. Considering a range of box lengths, a plot of l_B versus N_B on a log-log scale will be linear for networks with a fractal topology. On a normalized series of graphs with minimum size *N*, the fractal dimension d_B is obtained from linear regression of the log-log transformation of the general scaling relation:

$$\frac{N_B}{N} = l_B^{-d_B}.$$

Fractal structure of each GPN nucleus was judged relative to that expected of a random network. Box length (l_B) and graph size (N_B) scale as a power-law when the network has a fractal structure, and so the relationship should be well-described by a linear function under a log-log transformation of both box length and graph size (Song et al., 2005). By contrast, a random network is more likely to follow an exponential distribution which is best modeled as a linear function under a simple log transformation of one variable. Each simulation was comprised of 100 replicate random GPNs, each appraised using the above method for its fit to power and exponential models. The better model was chosen based on a comparison of the coefficients of determination (R^2) for linear least-squares regressions. The end of a simulation yielded the fraction out of 100 random GPNs giving a superior fit to a power-law (fractal) model.

In order to test the influence of the model parameters on GPN nucleus fractality, analysis of variance (ANOVA) was conducted on the full set of 27 simulations using Matlab (The MathWorks Inc., Natick, MA). The frequency of power-law GPNs was used as the dependent variable and the three-level parameters DUP, ATT, and REP as the main factors, testing both main effects and two-way interactions.

2.4 In silico model

Using the Python/networkx program, simulated GPNs were formed as random grammar networks wherein a 'concensus sequence' was randomly generated and used to seed the growth of the GPN – though it was not allowed to formally enter the GPN. Additional promoters were added to the GPN over time by duplicating

existing promoters or as a purely random process. Parameters of the model were varied to explore evolutionary mechanisms that might yield a fractal topology including DNA binding chemistry.

Of the standard evolutionary forces (mutation, selection, drift, and migration), all but migration were considered explicitly (Table 1; see also Suppl. Information). (*1*) Mutation was modeled as either point mutations or duplications of existing promoters in the growing GPN. (*2*) Selection was modeled as (*2a*) the attractive forces of the DNA-protein binding, and (*2b*) as repulsive forces potentially arising from overly tight binding and failed promoter clearance. Composite population fitness coefficients, *w*, were generated across these modes simply as the cross-product of the respective fitness under each of the models. Values were standardized to set the highest fitness within a simulation to $w = 1.0$ (see Suppl. Table 4). (*3*) Random genetic drift was handled implicitly through the finite size of the simulations (100 replicates of 500 promoters).

2.4.1 Mutation: duplication rate

A new promoter was produced by duplicating an extant promoter with probability *D*. Alternatively a new promoter was generated *de novo* from a uniform random base composition with probability 1-*D*.

2.4.2 Attraction from DNA-protein binding

In the model it was assumed that the σ factor had a binding chemistry optimized to a single promoter sequence, which was generated randomly each replicate. This optimal promoter (or more loosely, the 'consensus promoter') was not allowed to enter the GPN but was allowed to influence the composition of the GPN (except under conditions 'ATT-none', see Suppl. Table 1).

Attractive forces influencing the formation of random promoters were mediated through population-level fitness coefficients (Suppl. Table 1) that represented the selective advantage of a genome containing a given promoter based on the condition that it shares *x* bases with the consensus promoter motif (optimal binding chemistry). The no attraction model (ATT-none) assumed an absence of any chemical specificity in DNA-protein binding and served as a control whereby the extent of bp-sharing with the consensus did not affect fitness. Under the weak attraction (ATT-weak) model, the fitness function $w(x)$ was at its maximum when the random promoter was an exact copy of the consensus motif, and $w(x)$ declined gradually as *x* declined. The same held for the strong attraction (ATT-strong) model except $w(x)$ declined more rapidly as *x* declined, modeling a tighter binding chemistry. For a specific example, under ATT-weak a random promoter that shared 8 out of 11 bases with the consensus promoter had a fitness coefficient of $w = 0.727$. In practice this meant that the random promoter had a chance $P = 0.727$ of entering the GPN, subject to the drawing of a pseudorandom number (r); in the event of $r > 0.727$ the promoter was rejected and another random promoter generated and considered for entry.

Table 1 Evolutionary and cellular forces modeled through the 27 in silico simulations. Each of the main model parameters (DUP(*D*), ATT(w_{att}), and REP($w_{rep\text{-}fpc}$)) were varied over three levels: DUP (none=0, weak=0.15, strong=0.30); w_{att} (see Suppl. Table 1 for values); $w_{rep\text{-}fpc}$ (see Suppl. Table 2 for values).

Evolutionary force	Cellular force	Model	Parameter
mutation	promoter **duplication**	DUP	*D*
	point mutation		1-*D*
natural selection	**attraction** of DNA-protein binding permitting efficient transcription initiation	ATT	w_{att}
	repulsion from failed promoter clearance due to overly tight DNA-protein binding	REP	$w_{rep\text{-}fpc}$
random genetic drift	**intrinsic repulsion** from consensus arising from random occupancy of diffuse promoter space		$w_{rep\text{-}int}$

Table 2 Simulation results for the 27 models of GPN evolution. Model, simulation number; DUP, duplication rate (none=0, weak=0.15, strong=0.30); ATT, attraction (see Suppl. Table 1 for values); REP, repulsion due to failed promoter clearance (see Suppl. Table 3 for values); Exponential, fraction of replicate GPN nuclei that gave a better fit to the exponential (random) model; Power, fraction of replicate GPN nuclei that gave a better fit to the power-law (fractal) model.

Model	DUP	ATT	REP	Exponential (random)	Power (fractal)
1	none	none	none	0.97	0.03
2	none	none	weak	0.97	0.03
3	none	none	strong	0.91	0.09
4	none	weak	none	0.98	0.02
5	none	weak	weak	0.97	0.03
6	none	weak	strong	0.89	0.11
7	none	strong	none	0.64	0.36
8	none	strong	weak	0.63	0.37
9	none	strong	strong	0.74	0.26
10	weak	none	none	0.61	0.39
11	weak	none	weak	0.61	0.39
12	weak	none	strong	0.64	0.36
13	weak	weak	none	0.66	0.34
14	weak	weak	weak	0.70	0.30
15	weak	weak	strong	0.62	0.38
16	weak	strong	none	0.60	0.40
17	weak	strong	weak	0.50	0.50
18	weak	strong	strong	0.53	0.47
19	strong	none	none	0.34	0.66
20	strong	none	weak	0.34	0.66
21	strong	none	strong	0.31	0.69
22	strong	weak	none	0.24	0.76
23	strong	weak	weak	0.33	0.67
24	strong	weak	strong	0.40	0.60
25	strong	strong	none	0.50	0.50
26	strong	strong	weak	0.37	0.63
27	strong	strong	strong	0.33	0.67

2.4.3 Repulsion from failed promoter clearance

This portion of the model assumed that the optimal binding chemistry might not be optimal for transcription. A promoter with the consensus motif could bind σ efficiently yet prevent RNA polymerase from clearing the promoter if binding were too tight (e.g., Ellinger et al., 1994), in which case transcriptional elongation would be difficult to achieve. The null model REP-none (Suppl. Table 2) assumed no such interactions. Under the weak repulsion (REP-weak) model, the fitness function $w(x)$ was at its minimum ($w = 0$) when the random promoter was a perfect match to the consensus ($x = 11$), and $w(x)$ increased rapidly as x declined, reaching a maximum ($w = 1.0$) for $x \leq 9$ (arbitrary choice). Similar dynamics held for the strong repulsion (REP-strong) model except $w(x)$ increased more slowly as x declined ($w = 1.0$ for $x \leq 7$), modeling a broader region of base similarity in which promoter clearance was impeded.

2.4.4 Intrinsic repulsion from the random drift of GPN formation

Preliminary investigations of the model showed that most randomly generated promoters shared few bases with the ‘consensus’ promoter that had been used to seed the GPN. As the network grew and promoter space

filled in, new promoters could link to existing promoters – though not necessarily directly to the original 'consensus'. Thus the frequency distribution of bp-sharing in a random GPN is inherently skewed toward low values of sequence similarity. This native or default distribution of bp-sharing was utilized in the fitness function and the method of generating random sequences (see next two sections).

2.4.5 Composite measures of the fitness functions

Random promoters were allowed to enter the GPN provided a pseudorandom number was no greater than the promoter's composite fitness function, w_{comp}. A promoter entering with no constraints entered under w_{comp} = 1.0. The composite fitness coefficient was used to simulate the GPNs consisting of selective pressures due to differences in attractive forces of DNA-protein binding (w_{att}), repulsive forces due to failed promoter clearance ($w_{rep\text{-}fpc}$), and intrinsic repulsion ($w_{rep\text{-}int}$). These probabilities were multiplied to obtain a composite measure of fitness.

$$w_{comp} = w_{att} \times w_{rep-fpc} \times w_{rep-intr}$$

Pseudorandom numbers were used to select from this composite probability distribution x values (bp's shared with the consensus motif of footprint size F). For each random promoter, the number of bases to mutate from the consensus was then given by $m = F - x.$, and pseudorandom numbers were used to select the m bases and replace them with one of the four possible bases. This approach reduced the run time considerably compared to randomly forming promoters and then rejecting them if they failed to meet the acceptance threshold.

For any given simulation, each fitness value was standardized by the largest value in the distribution to set that value to 1.0 and allow the simulation to run faster (see Supplementary Tables).

2.4.6 Overall workflow in the simulations

I) Specify state conditions for the model parameters

II) Generate random optimal consensus promoter (though don't allow consensus to enter GPN)

III) For n = 1 to 500

A) Generate a new promoter

a) By duplication at rate D

b) Or by uniform random process at rate 1-D

1) Randomly determine how many bp's the new promoter will differ from the consensus

This is calculated from the composite (model components 1-3) probability density associated with each state of bp-sharing (see Suppl. Table 4)

2) Randomly mutate a consensus sequence to meet this criterion

B) Extract largest connected component from upper phase transition of the random GPN via m-slice

C) Conduct fractal analysis as specified by Song et al. (2007)

D) Evaluate least squares R^2 fit to linear transformations assuming each of the two models:

a) Fractal topology, expecting a power-law with best fit from a log-log transformation

b) Random topology, expecting an exponential relationship with best fit from a simple log transformation of one variable

E) The count is kept as to the number of replicates out of 100 that give a best fit to one or the other models

IV) Return to #1 and change model conditions

3 Results

Each of the 27 models yielded a combination of GPNs some of whose nuclei were best described as exponential (random) and some as power-law (fractal) (Table 2). On average the frequency of random GPNs within a simulation was greater (mean, 0.61; range, 0.24-0.98) compared to the frequency of fractal nuclei (mean, 0.42; range, 0.02-0.76). All of the simulations in which duplications were frequent (DUP = strong) produced at least as many fractal GPN nuclei as exponential nuclei (frequency of fractal nuclei $\geq$ 0.50).

Visual examples of simulated GPN nuclei are represented (Fig. 4) for four of the 27 models (across rows). Networks are shown at and near the phase transitions (across columns). Note the following: (*a*) Duplication has the effect of increasing the modularity or heterogeneity of linkages in the network; Fig. 4E has several dense modules of nodes whereas such substructure is less evident in the purely random Fig. 4B. (*b*) Attraction has the effect of organizing the promoter space around the consensus sequence as seen in the topological differences between Fig. 4H (with attraction) versus Fig. 4B and E (without attraction). (c) Repulsion has the effect of vacating the central portion of the GPN as seen in Fig. 4K (with repulsion) compared to the other figures above it (without repulsion).

All three parameters (DUP, ATT, REP) considered independently appeared to influence the fractality of the GPN nuclei (Fig. 5). The influence of attraction chemistry was much stronger than that of repulsion chemistry, and in both cases the effect was evident mainly at the strongest, most extreme parameter level. Duplication rate had the largest effect on topology, even on a weak level. These results should be interpreted with the caveat that the range of values chosen for the parameters may have influenced the extent of the response in each case. Regardless, the results show the trend that increases in each parameter associate with greater fractal topology.

Quantitative assessment of the pattern using ANOVA (Table 3) showed that duplication and attraction had significant effects on GPN nucleus fractality, as did their interaction. However, repulsion did not have a significant effect on fractal topology.

Table 3 ANOVA results for simulations showing the influence of duplication rate and attraction and repulsion of DNA-binding chemistry on the extent of fractality in the nuclei of GPNs.

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
duplication	1.14521	2	0.57260	115.46	**0.0000**
attraction	0.06112	2	0.03056	6.16	**0.0240**
repulsion	0.00170	2	0.00085	0.16	0.8458
dupl. x attr.	0.12575	4	0.03144	6.34	**0.0134**
dupl. x repul.	0.00024	4	0.00006	0.01	0.9997
attr. x repul.	0.01139	4	0.00285	0.57	0.6894
Error	0.03967	5	0.00496		
Total	1.38507	26			

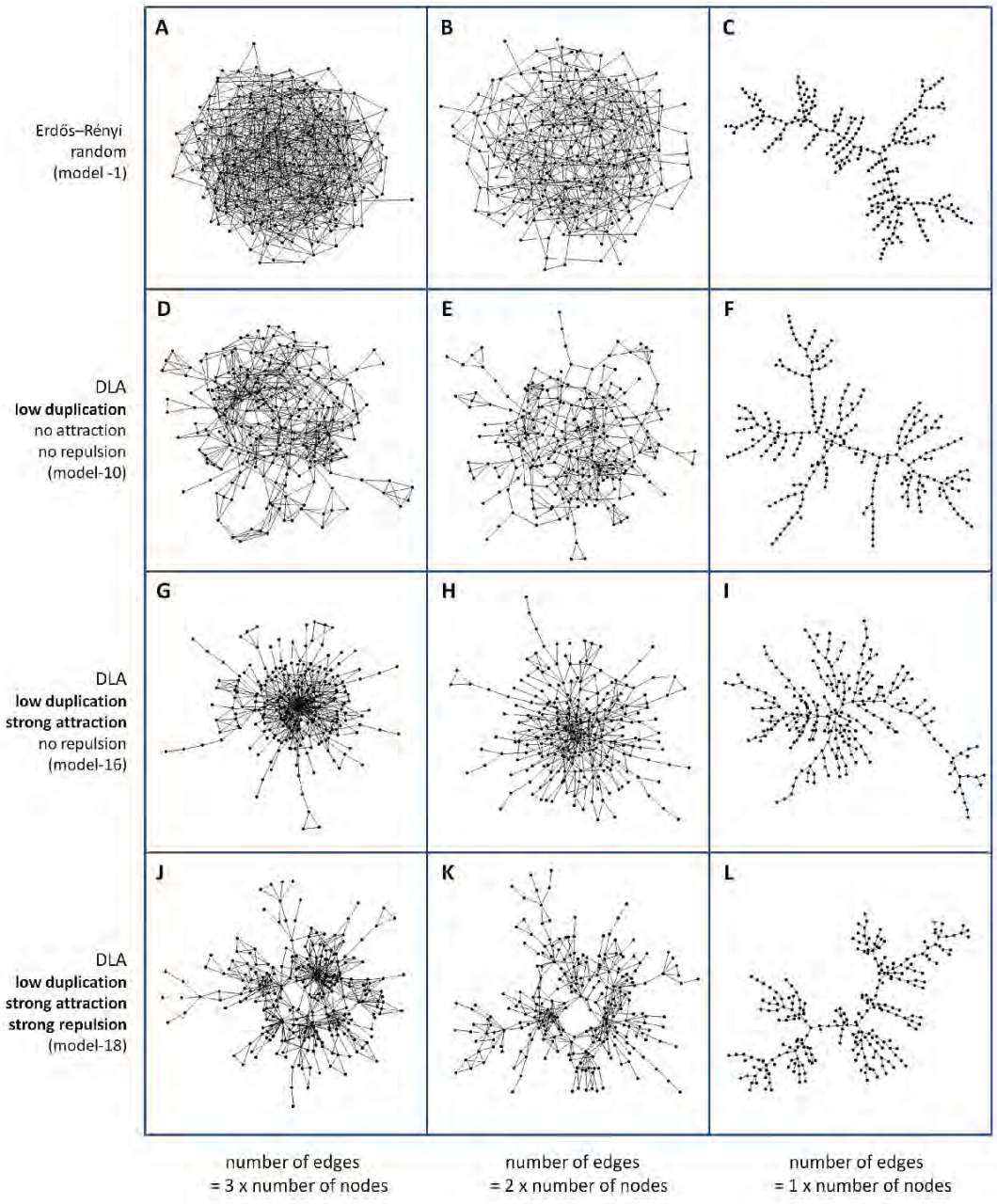

Fig. 4 Examples of simulated GPN nuclei captured from phase transitions. Four of the 27 models are shown here in the four rows (models 1, 10, 16, and 18), while columns represent the GPN nucleus just before the phase transition (left column), roughly at the phase transition (center column), and after the phase transition (right column). Each graph contains n=250 nodes and e=3n, 2n, or n edges (left to right) extracted as an m-slice. (**A-C**) Purely random graphs consistent with the Erdös-Rényi model in which there was no promoter duplication (DUP=none), no attractive force based on DNA-binding chemistry (ATT=none), and no repulsive force based on DNA binding chemistry (REP=none). (**D-F**) Low duplication rate but no DNA-binding chemistry specified. (**G-I**) Low duplication rate and strong DNA-binding chemical specificity, but no repulsion based on DNA-binding chemistry. (**J-L**) Low duplication rate combined with high attraction and repulsion in DNA-binding chemistry. Fractal dimensions and coefficients of determination for fit to the fractal model follow: **A**, d_B=2.706 (R^2=0.844); **B**, 2.316 (0.841); **C**, 1.358 (0.975); **D**, 2.148 (0.920); **E**, 1.837 (0.916); **F**, 1.439 (0.944); **G**, 1.971 (0.918); **H**, 1.910 (0.895); **I**, 1.523 (0.956); **J**, 2.001 (0.934); **K**, 1.872 (0.950); **L**, 1.495 (0.967)

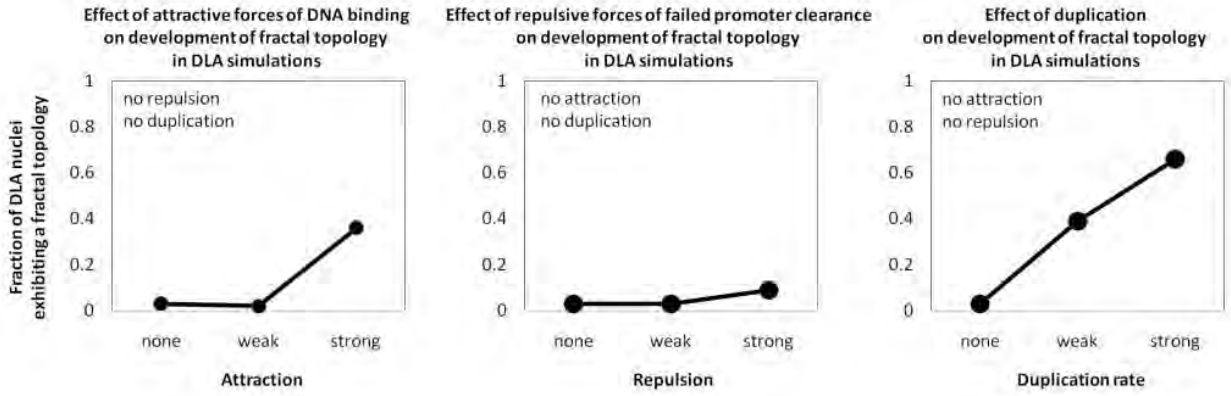

Fig. 5 Graphical representations of the results of DLA simulations with each parameter effect considered in isolation. Each figure shows the effects of one of the three evolutionary forces on the development of GPN nucleus topology, measured as the number out of 100 replicates giving a better fit to the power-law model.

4 Discussion

These results reject the strong version of the DLA model that includes a repulsive force deriving from DNA-protein binding chemistry. The results are consistent with the weaker form of the model that includes duplication, attraction, and intrinsic repulsion arising from a sparse promoter space, which pushes against the attractive forces of the DNA-binding protein yielding fractal topologies.

Gene duplication was shown to have the strongest effect on the fractal structure of the GPN nucleus. It is possible, though, that this is influenced by the levels selected for duplication versus the levels chosen for the fitness-based parameters (ATT and REP). Nevertheless, increasing the rate of duplication increased the extent of fractality. Gene duplication has been recognized for some time as an important factor in the evolution of genomes (Ohno, 1970; Zhang, 2003). Moreover, the duplication process has been implicated as the primary causal factor in the development of self-similarity (fractal structure) in the genome which includes a variety of aspects of genome organization that scale as a power-law (Luscombe et al., 2002; Koonin et al., 2006).

Attraction also proved important in the development of a fractal GPN nucleus, even when the consensus (optimal chemistry) promoter was excluded from the growing GPN; in such a case the consensus-GPN system behaves much like an attractor in a chaotic system (Kauffman, 1993). Not only is the attractive DNA-protein binding chemistry critical in the process of transcription initiation (on a local scale), it is critical to the development of organization in promoter space (on a broader scale). Without such a force, promoter space is inherently random and un-ordered as represented in the top row of Fig. 4 which shows GPNs that are essentially equivalent to the classic Erdös-Rényi (1960) random graphs. It takes the organizing influence of the DNA-binding protein, the transcription factor or σ-factor, to organize the promoters over evolutionary time, drawing them closer to the optimal binding chemistry. Natural selection likely mediates these outcomes in that a gene with a promoter that is too different from the optimal chemistry will not be transcribed and this may come at a fitness cost to the organism carrying this mutant variety of promoter.

The results indicate that repulsion arising from DNA-binding chemistry was not important in the development of fractal structure. It is conceivable that in some regulons a promoter motif might bind a transcription factor too tightly (e.g., Ellinger et al., 1994), such that optimal chemistry promoters might be

selected against and not appear in a realized GPN. However, the simulations run here suggest that, for the parameters considered, this is not as important a factor as duplication and attraction. It remains the case, though, that repulsion is thought to be a general feature of fractal networks (Song et al., 2006), so it is likely that the intrinsic repulsion discovered during these simulations is likely to play this role.

Intrinsic repulsion is a function of random genetic drift. The random generation of promoters yields a sparse promoter space that, when subject to an attractive organizing force, presents a repulsive force pushing back against the organizing principle. If all possible promoters existed in a GPN (saturated network) with a promoter footprint F, these 4^F promoters would yield a dense GPN with steps of only one base between adjacent promoters in the network. But this saturated GPN with $F = 11$ would consist of over four million promoters; real regulons and GPNs do not include all possible promoters. Instead, a random GPN of size $n = 500$ promoters forms a diffuse footprint in promoter space where most promoters share only a few bases with one another. Simulations of one million random promoters were performed to quantify this effect using the 11-bp footprint and spacer sizes from the RegulonDB σ^{54} predicted promoter set. This showed that over 97% of the random promoters shared no more than five bases (0-5 out of 11) with the consensus motif (Suppl. Table 3). The result is a frequency distribution of random promoter base sharing with the consensus that is heavily skewed away from the consensus motif, presenting a sort of intrinsic repulsion to the GPN. Set against the attractive chemical forces of the DNA-binding protein, a fractal GPN emerges.

Acknowledgements

Thanks to RegulonDB for use of their promoter database, A Hagberg and colleagues for networkx, Benedictine University for computer support, and anonymous reviewers for comments.

References

Aldrich PR, Horsley RK, Ahmed YA, et al. 2010. Fractal topology of gene promoter networks at phase transitions. Gene Regulation and Systems Biology, 4: 75-82

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

Batagelj V, Mrvar A. 1998. Pajek – program for large network analysis. Connections, 21(2): 47-57

Davidson E, Levin M. 2005. Gene regulatory networks. Proceedings of the National Academy of Sciences USA, 102(14): 4935

de Nooy W, Mrvar A, Batagelj V. 2005. Exploratory Social Network Analysis with Pajek. Cambridge University Press, New York, USA

Ellinger T, Behnke D, Bujard H, et al. 1994. Stalling of *Escherichia coli* RNA polymerase in the +6 to +12 region in vivo is associated with tight binding to consensus promoter elements. Journal of Molecular Biology, 239(4): 455-465

Erdös P, Rényi A. 1960. On the evolution of random graphs. Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 5: 17-61

Gama-Castro S, et al. 2008. RegulonDB (version 6.0): Gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Research, 36(Database issue): D120-124

Hagberg AA, Schult DA, Swart P. 2008. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008) ((Varoquaux G, Vaught T, Millman J eds.). Pasadena, CA, USA, 11-15

Hawley DK, McClure WR. 1983. Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Research, 11(8): 2237-2255

Hinckle DC, Chamberlin MJ. 1972. Studies of the binding of *Escherichia coli* RNA polymerase to DNA. I. The role of sigma subunit in site selection. Journal of Molecular Biology, 70(2): 157-185

Huerta AM, Collado-Vides J. 2003. Sigma70 promoters in *Escherichia coli*: Specific transcription in dense regions of overlapping promoter-like signals. Journal of Molecular Biology, 333 (2): 261-278

Kamada T, Kawai S. 1989. An algorithm for drawing general undirected graphs. Information Processing Letters, 31(1): 7-15

Kauffman SA. 1993. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York, USA

Koonin EV, Wolf YI, Karev GP. 2006. Power-Laws, Scale-Free Networks and Genome Biology. Springer, New York, USA

Liebovitch LS. 1998. Fractals and Chaos. Oxford University Press, New York, USA

Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M. 2002. The dominance of the population by a selected few: Power-law behavior applies to a wide variety of genomic properties. Genome Biology, 3(8): research0040.1–research0040.7

Ohno S. 1970. Evolution by Gene Duplication. Springer, New York, USA

Song C, Havlin S, Makse HA. 2006. Origins of fractality in the growth of complex networks. Nature Physics, 2(4): 275-281

Song C, Gallos LK, Havlin S, et al. 2007. How to calculate the fractal dimension of a complex network: The box covering algorithm. Journal of Statistical Mechanics, P03006

Song C, Havlin S, Makse HA. 2005. Self-similarity of complex networks. Nature, 433(7024): 392-395

van der Waals JD. Over de Continuiteit van den Gas- en Vloeistoftoestand (on the continuity of the gas and liquid state). Ph.D. thesis. Leiden, The Netherlands, 1873

Weaver RF. 2012. Molecular Biology (5^{th} ed.). McGraw-Hill, New York, USA

Witten TA, Sander LM. 1981. Diffusion-limited aggregation, a kinetic critical phenomenon. Physical Review Letters, 47(19): 1400-1403

Zhang J. 2003. Evolution by gene duplication: an update. Trends in Ecology and Evolution, 18(6): 292-298