PHYSICA D

Effects of alternative connectivity on behavior of randomly constructed Boolean networks

Chikoo Oosawa, Michael A. Savageau*

Department of Microbiology and Immunology, The University of Michigan Medical School, 5641 Medical Science Building II, Ann Arbor, MI 48109-0620, USA

Abstract

Random Boolean networks (RBNs) were first introduced and used to model regulatory gene networks by Kauffman in 1969 and since then have been frequently used to model networks at various levels of biological organization. The standard model has a characteristic topology in which the network has N genes with an average of K input connections selected at random. We have used RBNs to examine regulatory gene networks with four different topologies, which are characterized by different rank distributions of output connections that vary from uniform to highly skewed. Among these is the topology of the standard model, which is included for purposes of comparison, and a topology with a power-law rank distribution, which is based on recent data for the regulatory gene network of the bacterium *Escherichia coli*. We also examine effects of bias in the distribution of Boolean functions for the network. The dynamical properties and mutual information of these networks depend not only on their size but also on their topology and Boolean functions. Networks with the more uniform rank distributions exhibit longer lengths of attractors, larger numbers of attractors, and less mutual information. Networks with the more skewed rank distributions have complementary properties. When viewed as biological decision-making networks, those with either the most uniform or the most skewed rank distributions have disadvantageous properties. The intermediate rank distribution exhibited by the regulatory gene network of *E. coli* avoids these disadvantages.
© 2002 Elsevier Science B.V. All rights reserved.

1. Introduction

A striking property of regulatory gene networks in cells is the ability to switch expression of specific genes ON or OFF according to circumstances such as developmental stage of the organism, stress response to environment change, and so on. The resulting patterns of gene expression vary temporally and spatially, and the individual pattern can be viewed as the outcome of a set of decisions executed by the network.

Microbes such as *Escherichia coli* and *Salmonella typhimurium* have been extensively used as model organisms to study the mechanisms of gene regulation [1]. These studies have determined molecular details of the regulatory mechanisms and also have revealed many global aspects of the regulatory gene network [2]. For example, both global and specific regulators have been found to activate or repress gene

* Corresponding author. Tel.: +1-734-764-1466;
fax: +1-734-763-7163.
E-mail address: savageau@umich.edu (M.A. Savageau).

transcription by binding to particular sites on DNA [3]. Global and specific regulators have many and few binding sites in the genome, respectively. Furthermore, expression of each regulator is controlled by the regulator itself and/or by other regulators. Hence, gene expression is regulated as part of a complex, interconnected network whose essential feature is the set of regulatory genes and their interaction via the encoded regulators and their binding sites on the DNA. This network manifests an abundance of simultaneous communications among the genes in the genome.

Other proteins (enzymes, transporters, structural elements, etc.) and RNAs (ribozymes, structural elements, etc.) are encoded in the genome and have their expression modulated by the regulatory gene network. The activities of these effector molecules in turn influence the level of numerous small molecular-weight signaling molecules, many of which also can affect gene expression at various levels directly or indirectly. Nevertheless, regulation of gene expression in bacteria appears to be controlled primarily at the level of transcription, and transcriptional regulators play an essential role in determining the state of gene expression.

We shall assume for simplicity that the topological structure of a regulatory gene network can be depicted as a directed graph. Each *node* corresponds to a transcriptional unit that includes a regulatory gene. Each *arc* represents an interaction between transcriptional units. The origin of an arc is the output of one transcriptional unit and its destination is a target binding site, which is the input of a second transcriptional unit. Recent applications of DNA microarray technology have shown directly the dynamic changes in levels of RNA expression under different physiological conditions [4,5]. The results make it clear that the regulatory gene network is a highly nonlinear dynamic system. Therefore, we are interested in regulatory gene networks that are depicted not only as directed graphs but also as nonlinear dynamical systems.

The mathematical investigation of large nonlinear dynamical systems is limited by numerical as well as analytical difficulties. Some form of simplification is required to even begin addressing this class of problems. One of the most fruitful simplifications has been random Boolean networks (RBNs) [6], which

provided a tractable model for such investigations. In addition to the abundant numerical studies using RBNs (e.g. [6–18]), there have been a number of analytical studies (e.g. [19–22]). However, all these have involved the standard model for the RBN with a given distribution of topologies. The implications of alternative distributions have yet to be examined. The work described in this paper uses numerical methods to address this issue.

There are two principal questions that we shall address in this paper. (1) What type of topology characterizes regulatory gene networks? (2) How does topology affect the dynamics of regulatory gene networks? In order to answer the first question, we surveyed recent publications that show relationships between regulators and their target binding sites in the *E. coli* genome and found a power-law rank distribution (see Section 2.2.3). The power-law rank distribution means that a small number of regulatory molecules have many target sites and a large number have only a few target sites. To address the second question, we modified the output connectivity of the standard model for RBNs. The standard model [6] is characterized as follows: (1) Transcriptional units are randomly connected with a fixed number of inputs. (2) Each unit is assigned a Boolean function selected at random with equal probability. (3) The pattern of expression for a given gene is represented by a sequence of binary states, each of which is determined by the Boolean function and the previous state of the inputs for the gene. We introduced four different rank distributions for the output connections in the model and investigated their effects on the dynamical properties and mutual information of the network. Our results reveal statistical tendencies associated with the different distributions.

2. Methods

The characterization of large and complex genetic networks can only be achieved with the systematic tools provided by mathematical modeling and computer analysis. However, there are different approaches that have been used for this purpose, each having its

Table 1

Binary output and internal homogeneity for all types of Boolean functions with two binary inputs[a]

Inputs	Type															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0 0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0 1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1 0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1 1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

[a] Internal homogeneity equals 0.5 (for types 3, 5, 6, 9, 10, and 12), 0.75 (for types 1, 2, 4, 7, 8, 11, 13, and 14), or 1.0 (for types 0 and 15). See text for discussion.

distinct advantages and disadvantages [23]. In this paper, we use RBNs to explore properties that are likely to be present with high probability regardless of mechanistic details. Simple discrete/deterministic elements, whose behavior is easily computed, permit the exploration of large populations of networks and the application of the requisite statistical techniques; with other more complex elements the computations quickly become prohibitive. No one description can efficiently and accurately represent phenomenon at all hierarchical levels of organization and all scales of time and space. We acknowledge that the element description is only an approximation to that of the real genetic elements, and instead we focus on gross systemic properties that emerge through the interaction of the elements. An analogy is the statistical mechanical view of the ideal gas laws. Previous analyses with RBNs have led to predictions that are supported by experimental evidence (Section 2.2). Further justification of this approach can only come through additional evidence of this sort.

In this section, we briefly review: (1) the elements of the standard model for RBNs; (2) the four different rank distributions of output connections that are used in our modified model; (3) the methods used to study network dynamics; (4) the statistical measures used to characterize the results.

2.1. Random Boolean networks

RBNs are an idealized discrete dynamical model that have been used to represent the regulatory gene network for a generic organism [6]. The standard model for RBNs has the following characteristics: (a) RBNs have N nodes and directed connections between these nodes. (b) The state of each node is either ON or OFF. (c) The input signals to each node come directly from the outputs of exactly K other nodes chosen at random, where K is called the *input connectivity*. (Self-inputs and duplicate inputs are usually prohibited.) (d) There is an equiprobable choice to fix the Boolean function by which the current input signals to each node are combined to determine its next state (Table 1). (e) The states of the nodes are updated in a synchronous fashion. Networks with different numbers of nodes are often used to investigate the dependence of dynamical properties on network size, and many networks are generated and analyzed under fixed K and N to identify the statistical regularities [7]. Although we appreciate the limitations of the RBN model as a representation of real gene networks, we have retained most properties of the standard model in order to facilitate comparisons with the abundant literature devoted to the standard model. The principal difference is our modification of the method for generating the random connections. We consider RBNs with four different topologies.

2.2. Alternative topologies

We define the topology of the network in terms of the rank distribution of input and output connections. The number of input connections per node is considered to be small. The original work of Kauffman [6] suggested that a value of $K = 2$ for the input connectivity is relevant for the regulatory gene network of an

organism. Average values greater than $K = 2$ or 3 result in networks with complex erratic behaviors (the term "chaotic" has been used to describe behavior in which nearly all of the system's variables are changing and there are many patterns that are easily disturbed by slight perturbations), whereas average values less than $K = 2$ result in networks with a very limited behavioral repertoire (the term "frozen" has been used to describe behavior in which nearly all of the system's variables are unchanging and this pattern is not easily changed by perturbations) [7]. Thus, the most biologically meaningful behaviors were predicted for values of $K = 2$ or 3, and further studies have suggested that the evolutionary potential of these networks is greatest when they are operating in this range (which is described as the "edge of chaos") [8,9]. This predicted input connectivity based on the behavior of RBNs is supported by the more direct biological estimates of input connectivity that come from the work of Thieffry et al. [24,25]. They examined the number, type and location of regulatory binding sites in the promoter regions of *E. coli* and concluded that the average input connectivity is low, between 1 and 2. On the basis of these studies we have fixed the input connectivity for all our models at $K = 2$.

The 16 possible Boolean functions with $K = 2$ are shown in Table 1. For instance, expression of the maltose operon requires the presence of both the global regulator CRP and the specific regulator MalT. This condition can be considered as the logical AND function. However, in most cases the Boolean function is unknown, and so we have followed the procedure used in the standard model and randomly assigned Boolean functions to each node.

We have considered four different network topologies that, given the fixed input connectivity described above, are characterized by their output connectivity. Once the input connectivity K is determined, the total number of connections in a network is also fixed automatically. If a network contains N nodes and each node has exactly K input connections, the total number of connections in the network becomes KN. These connections are assigned by different methods to the output of each node and the resulting topologies are termed Uniform, Exponential, *E. coli* and Extreme.

2.2.1. Uniform topology

For this topology all nodes have exactly two output connections. Hence, the rank distribution for the number of output connections is uniform. The network in this case is constructed by selecting an unoccupied input at random and connecting it to an unoccupied output selected at random. This procedure is followed until all $2N$ connections have been assigned. In this topology, there is no central regulator; each individual regulator has equal potential to influence the network (Fig. 1).

2.2.2. Exponential topology

This is the topology of the original RBN, as described in the previous section. The rank distribution of output connections is exponential-like. The network in this case is constructed by selecting an unoccupied input at random and connecting it to an output selected at random. This procedure is followed until all $2N$ connections have been assigned. In some cases there will be nodes with no output connections. Also the rank distribution for the number of output connections can be different for each network. The average rank distribution among 1000 networks (see Section 2.4) appears to be exponential-like and dependent upon network size (Fig. 2A). The initial slopes of these rank distributions decrease with increasing network size. Thus, the rank distribution characteristic of the Exponential topology approaches that of the Uniform topology for large networks (Fig. 1).

2.2.3. E. coli topology

We have determined the *E. coli* topology from data in three recent publications [26–28]. We tabulated the number of regulated operons with known binding sites for each known regulator and rank ordered the regulators according to the number of operons with the corresponding target site. The results when plotted give a power-law rank distribution. Fig. 2B shows the normalized power-law rank distribution, with global regulators such as CRP and FNR having a large number of binding sites and specific regulators such as LacI and MalT having a small number. Although the three curves in Fig. 2B are based upon data that were available at different times, and hence the total number
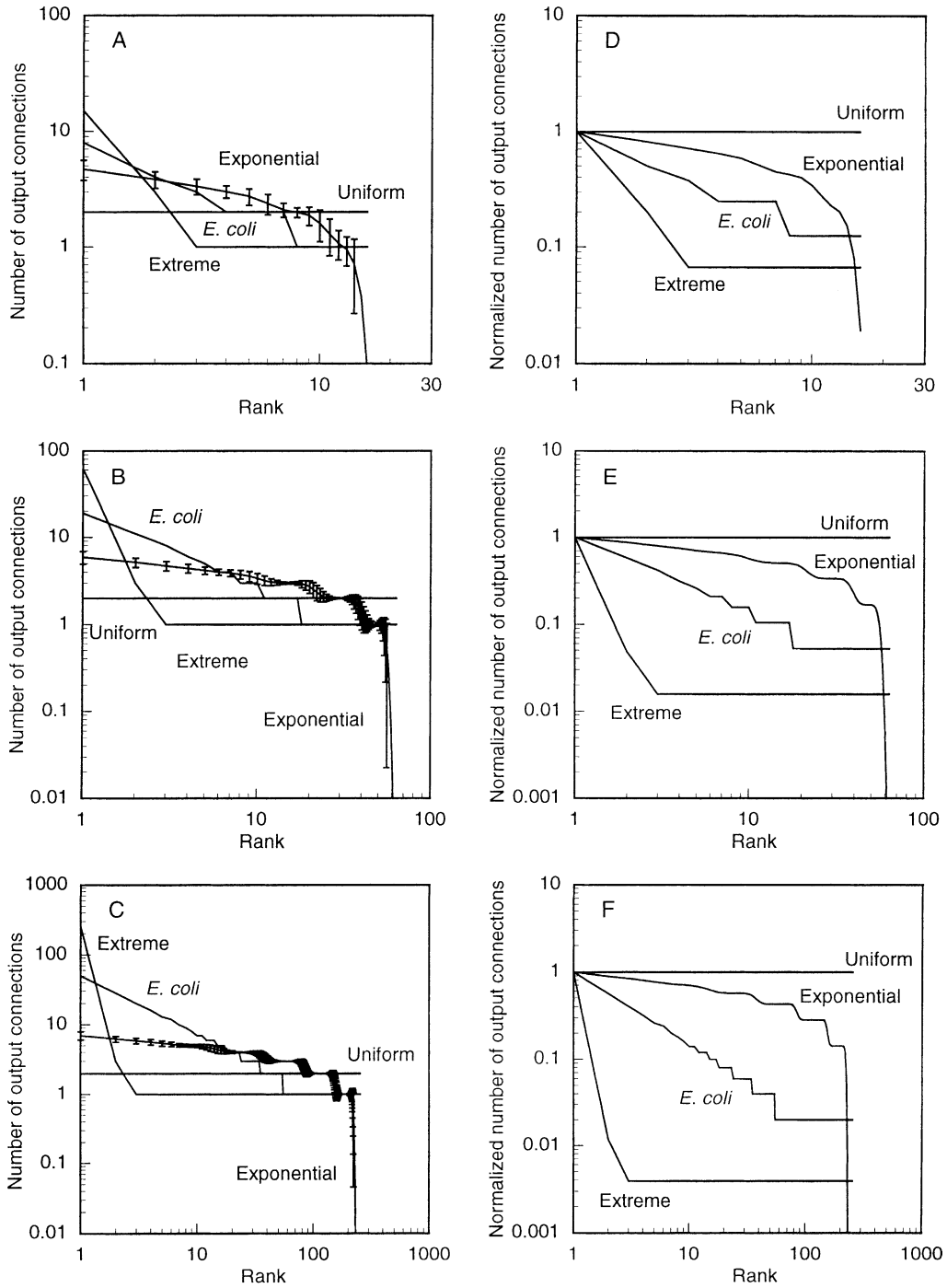
Fig. 1. Rank distribution of output connections in RBNs of increasing size with four different topologies. The numbers of output connections per node are represented in the left-hand panels: (A) $N = 16$; (B) $N = 64$; (C) $N = 256$. These distributions are normalized with respect to their maximum, rank 1 number, and represented in the right-hand panels: (D) $N = 16$; (E) $N = 64$; (F) $N = 256$. The four different topologies are chosen to span a full range of distributions from uniform to highly skewed; this is most easily seen in panel F. Error bars are shown for the Exponential topology based on 1000 randomly constructed networks. See text for discussion.
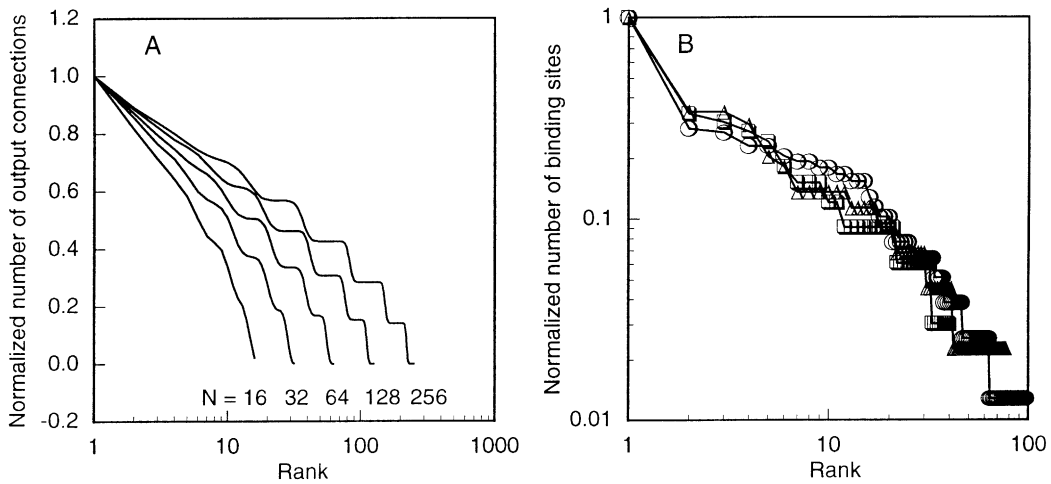
Fig. 2. Rank distribution of output connections based on the standard model for RBNs and on data from *Escherichia coli*. (A) Normalized rank distributions for networks of increasing size from $N = 16$ to 256 according to the standard model [6]. Note that the data are represented in a semi-logarithmic plot, and that they approach an exponential-like distribution with increasing network size. (B) Normalized rank distributions according to three data sets taken from *E. coli*: (□) Collado-Vides et al. [26]; (△) Otsuka et al. [27]; (○) Huerta et al. [28]. Note that these data are represented in a log–log plot, and that they approximate a power-law distribution with an exponent of -0.8.

of cases is greater in the more recent publications, the results are very similar. This suggests that the experimental data in each case is an unbiased sampling from the actual distribution. The exponential parameter determined for these rank distributions is approximately equal to -0.8. This value was used for all the network sizes we examined (see Fig. 1). The network in this case is constructed in two stages. First, the *E. coli* rank distribution of output connections is determined with the minimum number of outputs being one. Each node is allotted a number of outputs at random according to this distribution. Second, the connections are made by selecting an unoccupied input at random and connecting it to an unoccupied output selected at random. This procedure is followed until all $2N$ connections have been assigned.

2.2.4. Extreme topology

The opposite of the Uniform rank distribution, for which no node has more connections than another, is the Extreme rank distribution, for which one node has connections to all other nodes in the network and the remaining nodes have only one or a few connections. More specifically, one node has $(N - 1)$ output con-

nections, another has three and the rest have only one (Fig. 1). Among our four topologies, this one is called Extreme because its rank distribution exhibits the maximum skewness. In this case, the skewness increases with increasing network size (Fig. 1). The network in this case also is constructed in two stages. First, the Extreme rank distribution of output connections is determined with the minimum number of outputs being one. Each node is allotted a number of outputs at random according to this distribution. Second, the connections are made by selecting an unoccupied input at random and connecting it to an unoccupied output selected at random. This procedure is followed until all $2N$ connections have been assigned.

2.2.5. Effects of network size

Fig. 1D–F show normalized rank distributions for N equal to 16, 64 and 256, respectively, demonstrate clear differences among the four topologies. The Uniform and *E. coli* topologies retain their same normalized rank distributions independent of network size, whereas the Exponential and Extreme topologies have rank distributions that change with network size. The differences between the *E. coli* and Exponential

topologies increase with network size, as the normalized rank distribution of the Exponential topology approaches that of the Uniform topology. The differences between the *E. coli* and Extreme topologies also increase with network size, because of changes in the skewness of the normalized rank distribution for the Extreme topology as mentioned above (see Fig. 1).

2.3. Dynamics of Boolean networks

Dynamics of the RBN described above are determined by the following equation:

$$X_i(t + 1) = B_i(X(t)), \quad i = 1, 2, \ldots, N, \qquad (1)$$

where $X_i(t)$ is the binary state, either 0 or 1, of node i at time t, $B_i(\cdot)$ is the Boolean function used to update the state of node i, and $X(t)$ is a binary vector that gives the states of the N nodes in the network. An initial vector $X(0)$ is assigned and successive states of the nodes are updated synchronously.

The dynamical behavior of these random networks is represented by the time series of states, which corresponds to the time course of gene expression. The time course of gene expression follows a transient phase from an initial state until a periodic pattern, called an attractor, is eventually established. The length of the transient phase and the length of the attractor are determined by the number of time steps before and after entering the attractor respectively. Different initial states for the same network may lead to the same or different attractors. A large number of different initial states are explored in an effort to estimate the number of distinct attractors for a given network. By comparing the dynamics of networks with different topologies, we can expect to identify distinctive properties associated with each topology.

2.4. Statistical considerations

2.4.1. Ensembles of networks

We examined 20 different ensembles of networks corresponding to five different sizes ($N = 16, 32, 64, 128$ and 256) and four different topologies. Each ensemble consisted of 1000 networks randomly constructed as described above. The dynamic behavior of

each network was followed from 2000 different initial states.

2.4.2. Sampling of Boolean functions and initial states

Internal homogeneity [9,10], which is a measure that reflects the tendency for a node with a given Boolean function to give the same output (see Table 1), can be used to characterize the distribution of Boolean functions in a RBN. Because we selected Boolean functions at random from an uniform distribution in our studies, the average internal homogeneity for networks of different sizes and topologies was 0.6875, as expected. This was also confirmed by direct examination of the frequencies for all Boolean functions in the networks (data not shown).

Initial states were sampled uniformly from among the 2^N possible states in a network with N nodes. The number of different initial states, 2000, was selected so that we might easily compare our results with those in the literature. As far as we are aware, this is the largest number that has been used in previous numerical experiments, e.g., see [11]. Given the large number of possible states, the number 2000 can only be considered a sampling used to estimate relative numbers of attractors.

2.4.3. Types of averages

Dynamical properties such as number of attractors in a network, length of attractor and transient length, have been used to characterize networks under different conditions. The values of these properties in theory can be astronomical (up to 2^N). Because of this fact, median values have been used to provide representative results. For instance, Kauffman [6] conjectured that both the median length of attractors and the median number of attractors increase as the square-root of N. Bhattacharjya and Liang [11] demonstrated that the distribution of attractor lengths is noisy with a very long tail that exhibits a power-law distribution. Bagley and Glass [12] showed slow convergence of mean values in attractor lengths. They also showed long tails in distributions for the number of attractors. Values of the arithmetic mean are usually sensitive to the large values found in distributions with very long tails.

For the above reasons, we have used medians as representative values to characterize the distributions for numbers of attractors, attractor lengths, and transient lengths. In order to characterize the effects of network topology on the distributions of dynamical behaviors, we also have used the mode and quartiles of these distributions.

2.4.4. Overall activity

RBNs with average $K = 2$ exhibit behaviors in which a large fraction of nodes, typically 60–80%, fall into fixed 1 or 0 states and maintain the same state in all attractors [13,9]. The large fraction of nodes with fixed state implies a high degree of similarity in the expression patterns (as represented by the temporal sequence of the binary state vector) for the different attractors. As the average K increases, the RBNs show increasingly disordered behaviors and eventually most nodes change their state in all attractors [7].

We define the *active ratio* for each attractor as

$$1 - \frac{\text{Number of nodes with fixed state in the attractor}}{N}.$$
(2)

This provides a rough measure for the overall activity of each attractor.

Networks may have multiple attractors that are revealed by following the dynamics from 2000 different initial states, and some nodes of these networks may remain unchanged with state 0 or 1 in all of the attractors. Accordingly, we define the *active ratio* for each network as

$$1 - \frac{\text{Number of nodes with fixed state in } \textit{all} \text{ attractors}}{N}.$$
(3)

This represents the overall activity of each network.

2.4.5. Mutual information

Mutual information [29], which can be considered a measure of dependence similar to a correlation function [30], is widely used to characterize discrete dynamical systems [14]. Mutual information can be calculated directly from the expression patterns for any attractor of finite length. The mutual information function, MI, between two binary sequences, is given by

$$\text{MI}(\text{IN}; \text{OUT}) = H(\text{IN}) + H(\text{OUT}) - H(\text{IN}; \text{OUT}),$$
(4)

where $H(\text{IN})$ and $H(\text{OUT})$ are the individual Shannon entropies of the IN and OUT sequences and $H(\text{IN}; \text{OUT})$ is the joint entropy of the IN and OUT sequences (see Appendix A for further details). The Shannon entropy is a measure of randomness in the expression patterns. If a binary sequence has equal numbers of ones and zeroes, then the entropy of the sequence has a maximum value of one. If a binary sequence has only ones or only zeros, then the entropy has a minimum value of zero. The mutual information between two sequences has a value of zero if they are uncorrelated, e.g., because one or both sequences is unchanged throughout the length of an attractor; it has a positive value (up to one) if they are correlated. (Note that an input sequence IN_i for the node i is the same as the output sequence of an up-stream node OUT_j and that the output sequence OUT_i for the node i is the same as the input sequence of a down-stream node IN_k.)

In our case, each network node i responds to two binary input sequences or patterns, IN1_i and IN2_i, and produces a single binary output sequence or pattern, OUT_i. We define the mutual information, MI_i, between the input and output of each node i as

$$\text{MI}_i = \left[\text{MI}(\text{IN1}_i; \text{OUT}_i)^* \, \text{MI}(\text{IN2}_i; \text{OUT}_i) \right]^{1/2},$$
$$i = 1, 2, \ldots, N.$$
(5)

The node functions as an integrative device when the output reflects mutual information from both inputs. The geometric mean yields positive mutual information for the node only when the mutual information contributed by each of the inputs is positive. The arithmetic mean provides a higher estimate for the mutual information of the node, but there is no change in the relative order of the mutual information for the different topologies (data not shown). From this perspective, the alternative definition for the means does not affect our conclusions.

We define the entropy and mutual information for each attractor as sums of the entropy and mutual information for the individual nodes of the network:

$$H(attractor) = \sum_{i=1}^{N} H(OUT_i), \tag{6}$$

$$MI(attractor) = \sum_{i=1}^{N} MI_i. \tag{7}$$

We also define the entropy and mutual information for each network as arithmetic averages of the entropy and mutual information for the individual attractors of the network. Every network has at least one attractor, and we expect to identify a reasonable sampling of all attractors, since we uniformly sample the state space with 2000 different initial states. Thus, these network averages are representative values.

3. Results

3.1. Number of attractors vs. topology

3.1.1. Distribution

The number of attractors in networks with 64 nodes and the Exponential topology has a distribution with a broad dispersion (Fig. 3A). The distributions for networks with other sizes and topologies are similar in that their mode is always one attractor; they differ with regard to the maximum probability and the shape of decay. Fig. 3B shows the maximum probability (the probability of having a single attractor) as a function of network size and topology. Approximately 50% of networks with the Extreme topology have only one attractor, regardless of network size. Networks with the other topologies manifest a lower probability of having a single attractor, and this probability decreases as network size increases. In addition, this figure illustrates the typical tendencies found in many of our results. (1) The differences between pairs of networks with alternative topologies tend to increase with network size. The Exponential topology is an exception. (2) The differences between networks with Exponential and Uniform topologies tend to decrease with network size. This occurs because the rank distribution of output connections in networks with the

Exponential topology tends to converge towards that in networks with the Uniform topology as network size increases, as noted earlier (Fig. 1).

3.1.2. Averages

As mentioned in Section 2, it is difficult to characterize the average number of attractors in RBNs with a single measure because of the long dispersed tails in the distributions. In Fig. 3C and D we show two measures for the average number of attractors: the 2nd quartile (median) and the 3rd quartile. The square-root relationship between network size and median number of attractors in networks with the Exponential topology [6,12] is confirmed in Fig. 3C. This relationship also appears to hold for the Uniform topology and for the *E. coli* topology, at least with the larger network sizes. The square-root relationship is evident in Fig. 3D for networks with the *E. coli* topology, but not for the others. In the case of larger network sizes, the 3rd quartile for the number of attractors (Fig. 3D) tends to be less than the arithmetic mean (data not shown). The steeper relationship exhibited by networks with the Uniform and Exponential topologies is a manifestation of the long highly dispersed tails in the distributions associated with these topologies; conversely, the shallower relationship exhibited by networks with the Extreme topology is a consequence of the very narrow, highly skewed distribution characteristic of this topology.

A sample size of 2000 initial states is the largest that has been used in previous studies to estimate the number of attractors for a given network. We also have used this sample size so that we might compare our results with those of other investigators. Nevertheless, in an effort to determine the effectiveness of this sample size, we constructed 1000 RBNs of each topology and size $N = 256$, varied the number of different initial states, and determined the average number of different attractors. The results are shown in Fig. 4. The number of attractors appears to reach a limit in the case of the Extreme distribution. However, for the other distributions, the numbers show a diminishing tendency to limit as the distributions become more uniform. Indeed, the percentage difference between any two curves increases monotonically with the number
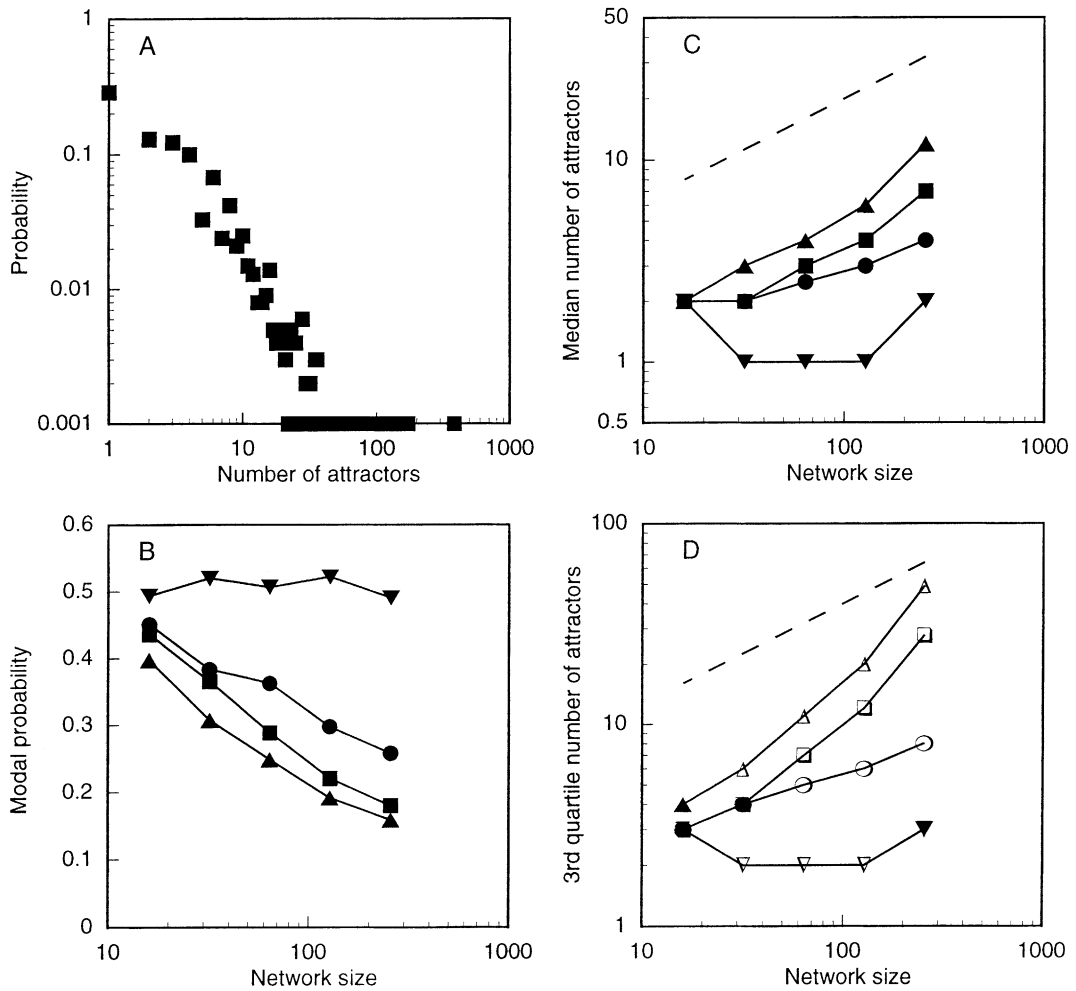
Fig. 3. Statistics for the number of attractors in RBNs with four different topologies ((▲) Uniform; (■) Exponential; (●) *E. coli*; (▼) Extreme) and various network sizes. (A) Probability distribution for the number of attractors in a network of size $N = 64$ and an Exponential topology. (B) Modal probability of the distribution for each topology. (C) Median number of attractors. (D) 3rd quartile number of attractors. An open symbol in this panel indicates that the arithmetic mean is larger than the corresponding 3rd quartile value. The dashed line shows a square-root relationship.

of initial states. Thus, it appears that the differences between topologies observed in Fig. 3 are actually underestimates.

3.2. Length of attractors vs. topology

3.2.1. Distribution

The length of attractors has a distribution similar to that in Fig. 3A in that the mode is always length one

(data not shown). Fig. 5A shows the maximum probability (the probability of having an attractor of length one) as a function of network size and topology. Approximately 50% of networks with the Extreme topology have attractors of length one, regardless of network size. Networks with the other topologies manifest a lower probability of having attractors of length one, and this probability decreases as network size increases.
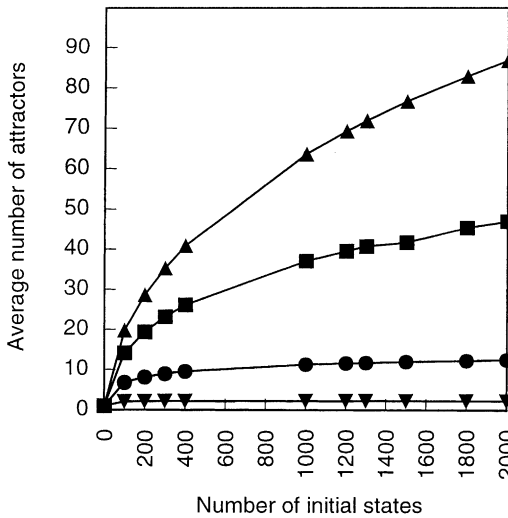
Fig. 4. Median number of attractors in RBNs with four different topologies ((▲) Uniform; (■) Exponential; (●) *E. coli*; (▼) Extreme) determined from various sample sizes of initial states. See text for discussion.

3.2.2. Averages

In Fig. 5B and C we show values for the 2nd quartile (median) and the 3rd quartile of attractor lengths. The median values (Fig. 5B) for networks with the Exponential topology follow approximately the square-root relationship as previously observed [7,12,15]. The same is true of networks with the Uniform topology. The shallower relationship exhibited by networks with the *E. coli* topology is a consequence of the narrower and more skewed rank distribution characteristic of this topology. The same is true of networks with the Extreme topology, but in this case the tendency is more exaggerated because the skewness of the rank distribution increases with network size unlike the case with the other topologies (see Fig. 1). Similar tendencies are apparent in the values for the 3rd quartile (Fig. 5C), except for the increase at larger network sizes exhibited by networks with the *E. coli* topology.

3.3. Length of transients vs. topology

3.3.1. Distribution

The length of transients has a very different distribution from that in Fig. 3A. The modal value is
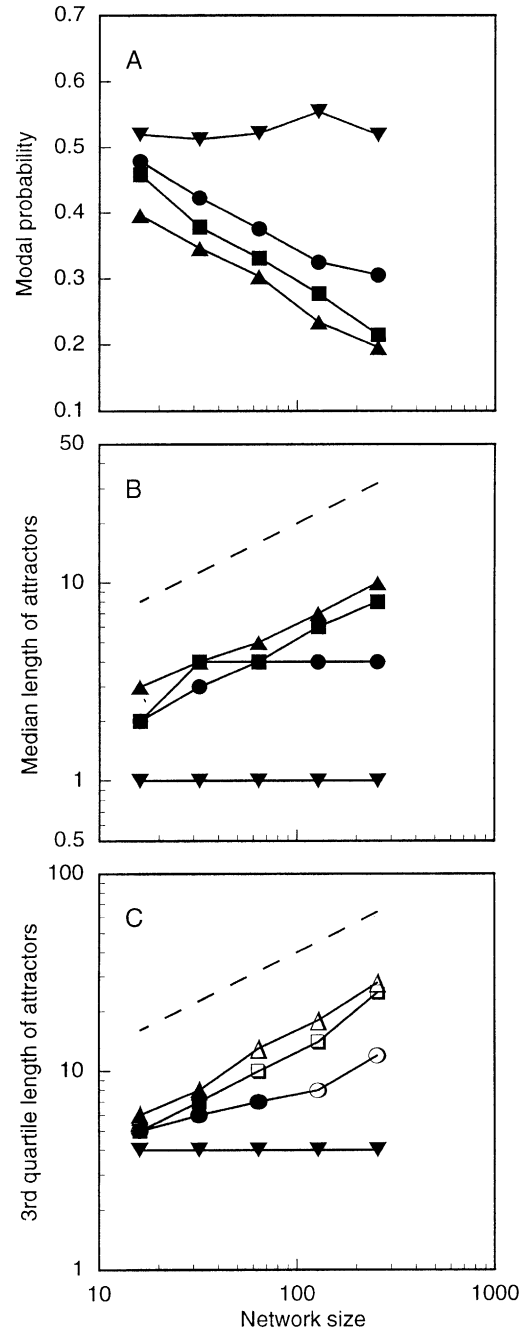

Fig. 5. Statistics for attractor length in RBNs with four different topologies ((▲) Uniform; (■) Exponential; (●) *E. coli*; (▼) Extreme) and various network sizes. (A) Modal probability of the distribution for each topology. (B) Median attractor length. (C) 3rd quartile of attractor length. An open symbol in this panel indicates that the arithmetic mean is larger than the corresponding 3rd quartile value. The dashed line shows a square-root relationship.

not at an extreme but has a more central location in the distribution (Fig. 6A), and both the modal value and the probability of the modal value vary with network size. The modal value of transient length increases with network size for all topologies (Fig. 6B), whereas the probability of the modal value decreases (Fig. 6C).

3.3.2. Averages

Fig. 6D and E show the 2nd quartile (median) and 3rd quartile values for transient length, respectively. Transient length increases with network size in a similar fashion for both quartile values and all topologies, unlike the case for attractor number and the case for attractor length. Thus, differences in transient length between topologies are less pronounced.

3.4. Other intermediate topologies and lower internal homogeneity

Figs. 3, 5 and 6 show large differences between networks with the *E. coli* and Extreme topologies. In order to investigate these tendencies further, we have examined networks with topologies that are intermediate between these two. Networks that have a rank distribution of output connections governed by a power-law with exponent equal to -2.0 exhibit behavior that is intermediate between that of networks with the *E. coli* and Extreme topologies (data not shown). Thus, topology has a graded effect on network behavior as the skewness of the rank distribution varies from Uniform to Extreme.

RBNs with Exponential topology, Boolean functions chosen at random, and input connectivity $K = 2$ show ordered behavior, whereas the same networks with higher input connectivity show disordered behavior [9]. Disordered behavior also is manifested in RBNs with Exponential topology and input connectivity $K = 2$, but with a frequency of XOR functions greater than expected if all Boolean functions were chosen at random [16]. In order to investigate the effects of alternative topologies on networks with more disordered dynamics we have examined RBNs with only XOR functions rather than higher input connectivity. (It is physically more difficult to generate highly

Table 2
Dynamical properties of RBNs with different topologies and only XOR functions[a]

Property	Statistic	Topology	
		Uniform	Extreme
Number of attractors	Median	8	4
	3rd quartile	29[b]	6[b]
Attractor lengths	Median	105	7
	3rd quartile	511[b]	24[b]
Transient lengths	Median	3	4
	3rd quartile	4	5

[a] Data for networks of size $N = 16$.
[b] The arithmetic mean is larger than the 3rd quartile.

skewed distributions in networks with high input connectivity.) We find that RBNs with only XOR functions show disordered behavior with a larger number of attractors and attractors of longer length, and that the alternative topologies influence these properties in the manner described above for networks with all Boolean functions chosen at random (Table 2). Thus, the effect of topology on the dynamical properties of the network is independent of internal homogeneity as defined by Walker and Ashby [10] and Kauffman [9] and shown in Table 1. These results suggest, but do not prove, that the influence of the alternative topologies might also be significant in networks with higher input connectivity, although the biological significance of such higher input connectivity remains to be established.

3.5. Analysis of expression patterns

In this section, we examine various measures of information to ascertain the degree of correlation or randomness within the expression patterns of RBNs and to see if these measures have any relationship to the topology of the network.

3.5.1. Overall activity

Fig. 7A and B show the relationship between network size and the average active ratio for attractors and networks, respectively. The differences between networks with the different topologies increase with network size. Networks with the more Uniform topologies exhibit the least activity for large networks. The average active ratios for networks with Uniform
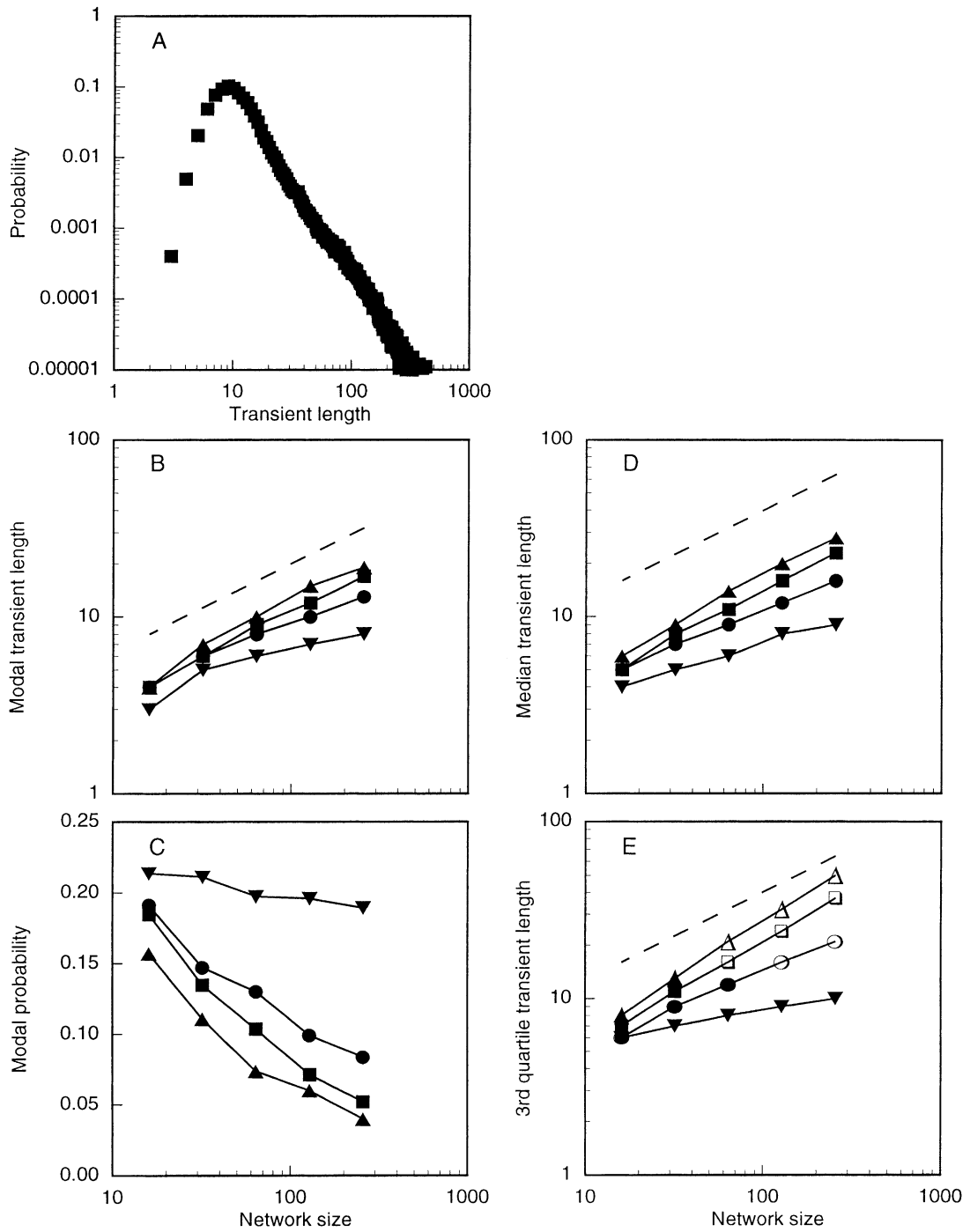
Fig. 6. Statistics for transient length in RBNs with four different topologies ((▲) Uniform; (■) Exponential; (●) *E. coli*; (▼) Extreme) and various network sizes. (A) Probability distribution for transient lengths in a network of size $N = 64$ and an Exponential topology. (B) Modal transient length. (C) Modal probability of the distribution for each topology. (D) Median transient length. (E) 3rd quartile transient length. An open symbol in this panel indicates that the arithmetic mean is larger than the corresponding 3rd quartile value. The dashed line shows a square-root relationship.
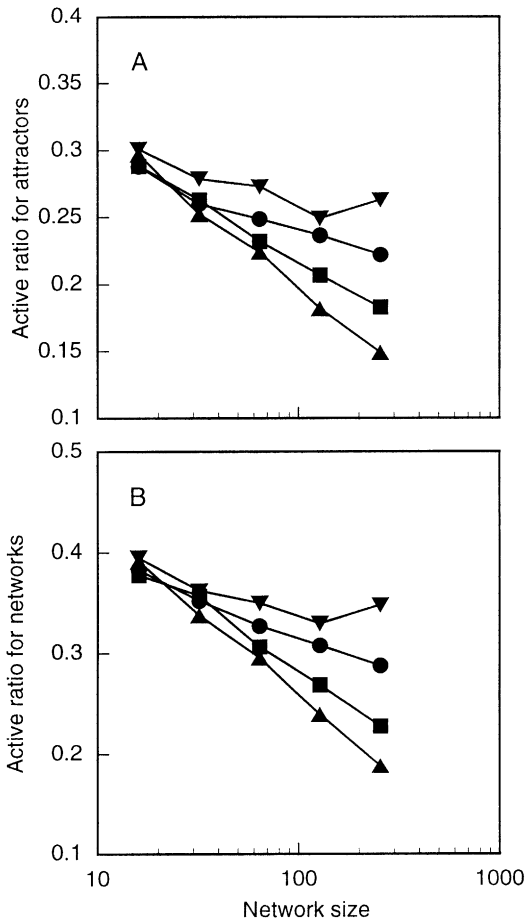
Fig. 7. Active ratio for each topology ((▲) Uniform; (■) Exponential; (●) *E. coli*; (▼) Extreme) as a function of network size. Each point represents the average for 1000 networks. (A) Average for all attractors in 1000 networks. (B) Average for 1000 networks. Note that the active ratio for a given network is always greater than or equal to the average active ratio for the attractors of that network. Note that the data are represented in a semi-logarithmic plot. See text for further discussion.

and Extreme topologies that utilize all Boolean functions (Fig. 7A) are smaller than those that utilize only XOR functions (Table 3). This shows that the dynamical properties of XOR networks (with either Extreme or Uniform topologies) are more disordered than those of the corresponding networks that utilize all of the Boolean functions. The average active ratios for networks that utilize only XOR functions are smaller for the Extreme topology than for the Uniform topology (Table 3).

Table 3

Arithmetic mean for measures of overall activity in RBNs with different topologies and only XOR functions[a]

Property	Topology	
	Uniform	Extreme
Active ratio for attractors[b]	0.929	0.818
Active ratio for networks[c]	0.962	0.907
Entropy	14.65	12.81
Mutual information	0.440	1.317

[a] Averages for 1000 randomly constructed networks of size $N = 16$ and 2000 initial binary states for each network.

[b] Average active ratio for all attractors in all networks. See Eq. (2) in the text.

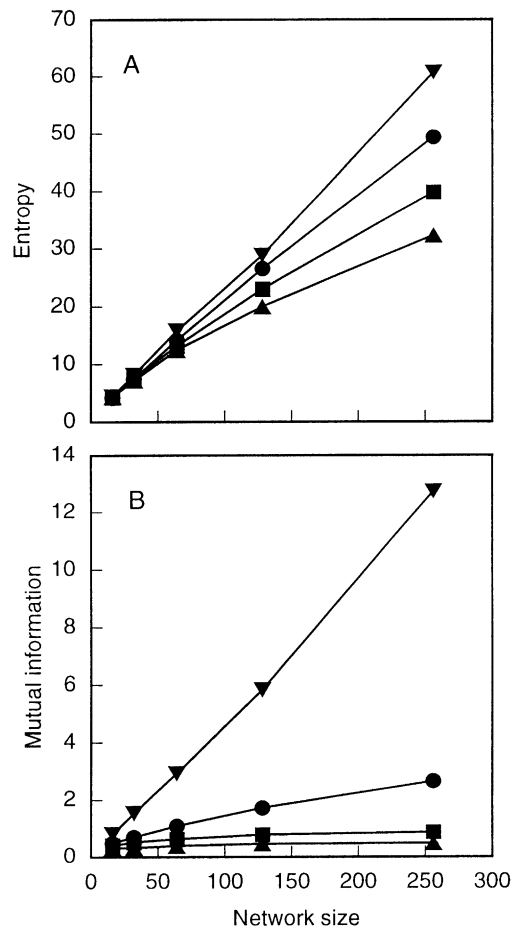[c] Average active ratio for all networks. See Eq. (3) in the text.

Fig. 8. Entropy and mutual information calculated from the expression patterns for each topology ((▲) Uniform; (■) Exponential; (●) *E. coli*; (▼) Extreme) as a function of network size. Each point represents the average for 1000 randomly constructed networks. (A) Entropy; (B) Mutual information.

3.5.2. Entropy and mutual information in expression patterns

The average entropy of networks with the full range of Boolean functions increases with network size regardless of topology (Fig. 8A). Those topologies with a more uniform rank distribution of output connections exhibit the lowest entropy, which means less randomness in their expression patterns. The opposite relationship can be seen in networks that have only XOR Boolean functions. The XOR networks with the Extreme topology have lower entropy than those with the Uniform topology. The dynamical properties of XOR networks (with either Extreme or Uniform topologies) are more disordered than those of the corresponding networks with the full range of Boolean functions (compare results in Table 3 with those in Fig. 8A).

The mutual information for networks increases with network size regardless of topology (Fig. 8B). Those topologies with a more uniform rank distribution of output connections exhibit the lowest mutual information (Fig. 8B and Table 3), which means less correlation in their expression patterns.

4. Discussion

A number of large complex networks such as the World Wide Web, electrical power grids, and actors networks have been treated as abstract graphs to characterize properties associated with their connectivity [31]. Barabási and Albert [32] have demonstrated that real networks of large size have nodes with a power-law distribution of connections. They showed that networks with the power-law distribution display a high degree of tolerance against the random removal of a node [33], however, these same networks are vulnerable to the removal of a highly connected node. These studies have concentrated on the relationship between static structural properties of networks and their tolerance to removal of a node. The dynamic properties of these networks have not been addressed specifically.

Previous studies, which pioneered the use of RBNs to represent the dynamics of regulatory gene networks [9], have shown how numbers of attractors and attractor lengths are affected by the input connectivity, the types of Boolean functions used, and the size of the network. The topology of the networks used in these studies is characterized by a limited number of input connections ($K \sim 2$) and an exponential-like rank distribution of output connections.

In this paper, we have examined how dynamics are influenced by connectivity in RBNs that model regulatory gene networks. We have explored a wide range of network topologies, as manifested in the rank distribution of output connections, and related our results to the topology of the regulatory gene network for the bacterium *E. coli*. Our approach, which takes this class of models one step closer to an accurate representation of the real regulatory gene network, may provide information not only for regulatory gene networks but also for other complex networks with similar topologies. The significance of our results and areas for further development are discussed below.

4.1. Regulatory genetic networks as decision-making networks

The regulatory gene network of an organism must be capable of making "decisions" that produce appropriate patterns of gene expression in response to different inputs from the external environment. The regulatory gene network of an organism can be viewed as making decisions that correspond to different attractors. For example, the regulatory gene network of bacteriophage lambda makes decisions corresponding to at least two attractors, one for lytic growth and one for lysogenic growth. The decision-making process is based on interactions among regulatory elements of the phage and the external environment provided by its host [34,35]. If an organism is to survive in many different environmental conditions, then its regulatory gene network must be capable of making a large number of decisions that correspond to a large number of appropriate attractors. Moreover, these decisions must be made quickly in order to successfully establish the organism's response to stresses such as heat, radiation and starvation. Thus, we might expect organisms to evolve toward large numbers of attractors with short attractor lengths. This appears to be the case

for the *E. coli* topology, which has more patterns of decisions (more different attractors) than the Extreme topology and shorter times to establish the patterns (shorter attractor lengths) than the Uniform topology. It would be instructive to make a direct comparison of the numbers of attractors in our RBN and in real *E. coli*, but this is not possible because the actual number of attractors for the organism is currently unknown.

It should be emphasized that there is no exact correspondence between the number of distinct behavioral patterns in Boolean and in continuous models of real networks. For example, it is well known that a continuous model of an unbranched pathway with three dependent variables and end-product inhibition can in principle exhibit a wide range of dynamic behavior. This includes stable and unstable fixed points, foci, and limit cycles. These models also exhibit a wide spectrum of relaxation times, depending on the specific parameter values. The same system modeled as a Boolean network can exhibit two distinct attractors with lengths 2 and 6, and the transient lengths in these cases are zero. The longer attractor exhibits a wave-like pattern of activity around the negative feedback loop that is similar to the limit cycle behavior of the continuous model. The shorter attractor, while still showing an oscillatory pattern, is more like the fixed point of the continuous model. (Indeed, if one projects the unit cube onto a plane such that the two vertices of the shorter attractor have overlapping images, then the shorter attractor appears as a fixed point and the longer attractor appears as a limit cycle surrounding the fixed point.) This simple example also shows that a perturbation can switch the Boolean network from one attractor to the other without a positive feedback loop in the network. Thus, the "decisions" corresponding to such a switch in a RBN model are different from the "decisions" corresponding to a switch in a continuous model that involves positive feedback and a cusp catastrophe [36].

4.2. RBNs as discrete dynamical systems

Previous work has demonstrated two critical factors that influence the dynamical behavior of RBNs,

input connectivity and bias in the choice of Boolean functions. Here we have confirmed these findings and provide evidence for another critical factor; namely, network topology as manifested in the rank distribution of output connections. We have examined the standard topology of RBNs (Exponential) and three alternatives (Uniform, *E. coli*, and Extreme) designed to sample the full range of rank distributions from uniform to extremely skewed. Our results for the standard (Exponential) topology are consistent with those of Bagley and Glass [12] and Bhattacharjya and Liang [15]. Our results for the alternative topologies show that smaller numbers of attractors and shorter attractor lengths are characteristic of the more skewed topologies. The numbers of attractors and the attractor lengths, two properties that increase with network size, have very similar distributions that differ markedly from the distribution of transient lengths. The typical tendencies between topologies, as explained above, are found in many figures (Figs. 3, 5 and 6). They clearly demonstrate that topology is another critical factor influencing the dynamical behavior of RBNs.

The connections between nodes determine the routes for signal transmission. In networks with highly skewed topologies a few nodes account for a large portion of all the output connections. Thus, most nodes in such networks have a high degree of dependence upon the dictatorial nodes. This dependence is evident in Fig. 8B and Table 3, which show that mutual information is greater in networks with more skewed topologies. From the results in Figs. 3 and 5, which show that networks with more skewed topologies have fewer attractors and shorter lengths, we also would expect that an increase in mutual information would be accompanied by a decrease in the number of attractors and in attractor length.

The differences in both entropy (Table 3 and Fig. 8A) and active ratio (Fig. 7) between networks with all Boolean functions and those with only XOR Boolean functions show that the properties of XOR networks are more disordered. These two types of networks are able to increase their mutual information in diametrically opposed ways through the use of alternative topologies. XOR networks, which are

more disordered to begin with, increase mutual information with more skewed topologies, which decrease entropy (Table 3), whereas full Boolean networks, which are more rigidly ordered to begin with, increase mutual information with more skewed topologies, which increase entropy (Fig. 8). This is again an indication that high mutual information is achieved in an intermediate zone between frozen and chaotic behavior.

The active ratio in networks with more Uniform topology is smaller, which means a larger number of genes that are unable to change state both within an attractor and between attractors. However, the number of attractors in those networks is larger (Fig. 3), which suggests that networks with more Uniform topology will tend to have a higher degree of similarity in their patterns of gene expression (as represented by the temporal sequence of the binary state vector).

The limited number of input connections and the skewed distribution of output connections exhibited by the bacterium *E. coli* are consistent with a set of well-established experimental observations. This topology implies that most genes will have a small number of input connections from, and output connections to, other genes. This suggests that many gene systems can be studied as relatively isolated subsystems via the experimental manipulation of the manageable number of input and output variables. This is consistent with the wealth of experience in molecular genetics that has accumulated over the past several decades. On the other hand, this topology implies that a few genes will have a large number of output connections to other genes. These are the global regulators whose true sphere of influence has been revealed only with the advent of microarray technologies for monitoring the expression profile of the entire genome [4,5].

4.3. Extension of the results

The *E. coli* topology is the first biologically relevant distribution of connections to be incorporated into a RBN. This distribution is based on experimental data [26,28] and theoretical predictions of regulator–regulon relationships for *E. coli* [27]. The resulting *E. coli* rank distribution is based mainly on data for σ^{70} promoters. However, the complete genome sequence has been determined for *E. coli* [37] and dozens of other microbial species [38], so we can expect that data on the full connectivity of regulatory gene networks for several organisms will soon become available. These data will permit a more detailed exploration of the generality of the results presented here.

Harris et al. [17] and Kauffman [18] have suggested that eukaryotic cells, which have higher input connectivity, are able to avoid chaotic behavior by biasing the choice of Boolean functions toward higher internal homogeneity. Their results demonstrate that ordered dynamics can be achieved even with high input connectivities ($K = 3$–5) when networks have this biased selection of Boolean functions. They found support for this prediction in their analysis of papers characterizing transcriptional regulation of eukaryotic genes. Our results suggest another strategy by which eukaryotic cells with higher input connectivity are able to avoid chaotic behavior, namely, evolving a more skewed distribution of output connections. This prediction could be tested when more information becomes available for the complete genome sequence of higher eukaryotes.

Synchronous updating of the states is a prominent feature of the standard model that we have retained for reasons of computational efficiency and for purposes of comparison with the existing literature. However, we have done limited preliminary experiments with a nonsynchronous updating scheme and found roughly similar results (data not shown). A more sophisticated algorithm is required for the determination of the transients and attractors, and the lengths of the transients are much longer, both of these factors contribute to excessively long computation times. Thus, we have postponed further analysis with this approach. A statistical approach developed for networks described by continuous variables has been applied to cases with a few different connections but many different values for the parameters characterizing the strengths of the connections [39]. Again, the computation time proved impractical for the large number of randomly connected networks examined here.

These alternative approaches have the potential to better represent real genetic networks, and further efforts to improve their computational efficiency are in progress.

The molecular mode of gene control is predicted to be a function of the demand for use of a regulator in the expression of the regulated genes [40], the positive mode is associated with high demand and the negative mode with low demand. Numerous experimental results support these predictions. This demand theory is based on natural selection for robustness in response to mutations in the regulatory mechanisms. The RBNs used for the studies in this paper could be used to model this evolutionary process and thereby provide an additional means to explore the implications of this demand theory.

Appendix A. Entropy

The entropy of a given pattern of expression (binary sequence) for a finite length attractor is defined as:

$$H = -[p_0 \log(p_0) + (1 - p_0) \log(1 - p_0)], \quad (A.1)$$

where the logarithms have base 2 and p_0 represents the proportion of zeros in the expression pattern.

A.1. Joint entropy

The joint entropy of two binary sequences, IN and OUT, is defined as:

$$H(\text{IN}; \text{OUT}) = \sum_{i,j=0,1} p_{ij}^{\text{IN OUT}} \log(p_{ij}^{\text{IN OUT}}), \quad (A.2)$$

where $p_{ij}^{\text{IN OUT}}$ is the joint probability that state i in the sequence IN and state j in the sequence OUT appear at the same time.

References

[1] F.C. Neidhardt (Ed.), *Escherichia coli* and *Salmonella cellular* and Molecular Biology, Vol. 2, second ed., American Society for Microbiology, Washington, DC, 1996.

[2] F.C. Neidhardt, M.A. Savageau, Regulation beyond the operon, in: *Escherichia coli* and *Salmonella cellular* and Molecular Biology, Vol. 2, second ed., American Society for Microbiology, Washington, DC, 1996, pp. 1310–1324.

[3] S. Gottesman, Bacterial regulation: global regulatory networks, Ann. Rev. Genet. 18 (1984) 415–441.

[4] J.L. DeRisi, V.R. Iyer, P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[5] H. Tao, C. Bausch, C. Richmond, F.R. Blattner, T. Conway, Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media, J. Bacteriol. 181 (1999) 6425–6440.

[6] S.A. Kauffman, Metabolic stability and epistasis in randomly constructed genetic nets, J. Theoret. Biol. 22 (1969) 437–467.

[7] S.A. Kauffman, Requirements for evolvability in complex systems: orderly dynamics and frozen components, Physica D 42 (1990) 135–152.

[8] S.A. Kauffman, Antichaos and adaptation, Sci. Am. 265 (1991) 78–84.

[9] S.A. Kauffman, Origin of Order, Oxford University Press, Oxford, 1993.

[10] C.C. Walker, W.R. Ashby, On temporal characteristics of behavior in certain complex systems, Kybernetik 3 (1966) 100–108.

[11] A. Bhattacharjya, S. Liang, Power-law distribution in some random Boolean networks, Phys. Rev. Lett. 77 (1996) 1644–1647.

[12] R.J. Bagley, L. Glass, Counting and classifying attractors in high dimensional dynamical system, J. Theoret. Biol. 183 (1996) 269–284.

[13] S.A. Kauffman, Emergent properties in random complex automata, Physica D 10 (1984) 145–156.

[14] C.G. Langton, Computation at the edge of chaos: phase transitions and emergent computation, Physica D 42 (1990) 12–37.

[15] A. Bhattacharjya, S. Liang, Median attractor and transients in random Boolean nets, Physica D 95 (1996) 29–34.

[16] K.E. Kürten, H. Beer, Inhomogeneous Kauffman models at the borderline between order and chaos, J. Stat. Phys. 87 (1997) 929–935.

[17] S.E. Harris, B.K. Sawhill, A. Wuensche, S.A. Kauffman, Biased eukaryotic gene regulation rules suggest genome behavior is near edge of chaos, Santa Fe Institute 97-05-039, 1997.

[18] S.A. Kauffman, Investigations, Oxford University Press, Oxford, 2000.

[19] B. Derrida, D. Pomeau, Random network of automata: a simple annealed approximation, Europhys. Lett. 1 (1986) 45–49.

[20] B. Derrida, D. Stauffer, Phase transition in two-dimensional Kauffman cellular automata, Europhys. Lett. 2 (1986) 739–745.

[21] B. Derrida, G. Weisbuch, Evolution of overlaps between configurations in random Boolean networks, J. Phys. (Paris) 47 (1986) 1297–1303.

[22] H. Flyvbjerg, An order parameter for networks of automata, J. Phys. Math. Gen. 21 (1988) L955–L960.

[23] R. Alves, M.A. Savageau, Systemic properties of ensembles of metabolic networks: application of graphical and statistical methods to simple unbranched pathways, Bioinformatics 16 (2000) 534–547.

[24] D. Thieffry, A.M. Huerta, E. Pérez-Rueda, J. Collado-Vides, From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*, BioEssays 20 (1998) 433–440.

[25] D. Thieffry, H. Salgado, A.M. Huerta, J. Collado-Vides, Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12, Bioinformatics 14 (1998) 391–400.

[26] J. Collado-Vides, B. Magasanik, J.D. Gralla, Control site location and transcriptional regulation in *Escherichia coli*, Microbiol. Rev. 55 (1991) 371–394.

[27] J. Otsuka, H. Watanabe, K.T. Mori, Evolution of transcription regulation system through promiscuous coupling of regulatory proteins with operons: structure from protein sequence similarities, J. Theoret. Biol. 178 (1996) 183–204.

[28] A.M. Huerta, H. Salgado, D. Thieffry, J. Collado-Vides, RegulonDB: a database on transcriptional regulon in *Escherichia coli*, Nucl. Acids Res. 26 (1998) 55–59.

[29] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Champaign, IL, 1963.

[30] W. Li, Mutual information functions versus correlation functions, J. Stat. Phys. 60 (1990) 823–837.

[31] D.J. Watts, S.H. Strogatz, Collective dynamics of small world networks, Science 393 (1998) 440–442.

[32] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.

[33] R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks, Science 406 (2000) 378–382.

[34] D. Thieffry, R. Thomas, Dynamical behaviors of biological regulatory networks II. Immunity control in bacteriophage lambda, Bull. Math. Biol. 57 (1995) 277–297.

[35] A. Arkin, J. Ross, H.H. McAdams, Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells, Genetics 149 (1998) 1633–1648.

[36] M.A. Savageau, Design principles for elementary gene circuits: elements, methods, and examples, Chaos 11 (2001) 142–159.

[37] F.R. Blattner, et al., The complete genome sequence of *Escherichia coli* K-12, Science 227 (1997) 1453–1462.

[38] http://www.tigr.org/tdb/mdb/mdbcomplete.html.

[39] R. Alves, M.A. Savageau, Effect of overall feedback inhibition in unbranched biosynthetic pathways, Biophys. J. 79 (2000) 2290–2304.

[40] M.A. Savageau, Demand theory of gene regulation: quantitative application to the lactose and maltose operons of *Escherichia coli*, Genetics 149 (1998) 1677–1691.