Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 4387–4391, May 1994
Evolution

Evolution of gene networks by gene duplications: A mathematical model and its implications on genome organization

(transcriptional regulation/molecular evolution/homeobox gene)

ANDREAS WAGNER

Department of Biology, Center for Computational Ecology, Osborn Memorial Laboratories, Yale University, P.O. Box 6666, New Haven, CT 06511

Communicated by Frank H. Ruddle, January 5, 1994

ABSTRACT **Networks of transcriptional regulators have key roles in metazoan development. Important forces in the evolution of these networks are gene duplications and gene deletions, events that may change the spatiotemporal expression pattern of network genes. A measure for the probability of such changes after gene-duplication events is proposed. This measure is based on a simple mathematical model that describes such networks as dynamical systems and on properties of ensembles of these dynamical systems. It is predicted that this probability depends only on the fraction of genes duplicated in a single event and that it is largest if ≈40% of the genes in a network are duplicated. This property is robust with respect to variations in model parameters. On these grounds, it is argued that (*i*) evolution of gene networks should preferentially occur either by duplication of single genes or by duplication of all genes involved in a network, and that (*ii*) tight linkage ("clustering") or strong dispersal are the two evolutionarily most favorable forms of genomic organization of genes forming such networks.**

The expression of most protein-coding genes in eukaryotes is regulated predominantly on the transcriptional level (1). The initiation rate of RNA polymerase II at these genes is mainly determined by the interaction of a "basal" transcription machinery with one or more general transcription factors bound to promoter sites in the vicinity of a gene (1, 2). The many complexities involved in transcriptional regulation, such as competition for binding sites on the DNA by transcription factors (3) and posttranslational regulation of the activity of transcription factors themselves—e.g., by protein phosphorylation (4), differential splicing (5), and heterotypic dimerization (6)—strongly suggest that a comprehensive quantitative theory of transcriptional regulation will not be available in the foreseeable future. The lack of any such theory is regrettable in light of the observation that transcription factors play crucial roles in the early development of metazoans. Two well-investigated examples, early *Drosophila* development (7) and axial patterning in vertebrates (8), indicate that sets ("networks") of genes encoding transcriptional regulators that mutually regulate each other's expression stand at the very basis of cell-fate determination and regional determination in metazoan embryos. Although early zygotic genes in *Drosophila* and Antennapedia-class homeobox genes in *Drosophila* and vertebrates are thus far the empirically best-understood examples, circumstantial evidence from different developmental processes (9–11) strongly suggests that similar networks will be found to act in those processes as well.

Although individual regulatory proteins in networks investigated thus far seem to display a remarkable degree of structural and functional conservation (8, 12), the structure of networks themselves seems to evolve slowly over time. Duplications and deletions of single genes, as well as duplications of all or a large fraction of the genes involved in a network, seem important agents of change in these systems (13). Due to the lack of quantitative models of transcriptional regulation, we poorly understand what kind of evolutionary phenomena are to be expected in these networks. Are duplications of genes interacting in a network-like fashion likely to perturb developmental processes? Do duplications of different numbers of genes cause different degrees of such perturbations? Finally and more specifically, is the fact that duplication of, for example, half a homeobox cluster never been observed coincidental or is it to be expected on theoretical grounds?

This contribution proposes a simplified model that may help to approach these and similar questions. No attempt is made to fully cover the biochemical phenomena underlying transcriptional regulation. Instead, the consequences of the network character of the regulatory system under consideration are emphasized, and it is this network character upon which conclusions are based. Possible consequences of the effects of gene duplications on the genomic organization of the genes involved are discussed.

THE MODEL

Different and only partially overlapping sets of transcription factors are expressed in different cells or different regions at any given stage of development of an organism. The model to be developed below refers to the expression of transcription factor genes only in one developmental stage and only in one set of cells (nuclei) that have an expression pattern in common—e.g., a set of nuclei in a part of a *Drosophila* blastoderm expressing a specific subset of gap genes and pair-rule genes. A subset of N such genes $\vec{G} := \{G_1, \ldots, G_N\}$, the products of which mutually regulate each other's expression on the transcriptional level, will henceforth be referred to as a "network." The size of known candidate networks that presumably rely mostly on transcriptional regulation among their member genes is small, probably being much less than 100 genes for any given network.

Each such network is visualized as a dynamical system, in which an initial gene expression state or activation state of the network—i.e., an array of concentrations of proteins at time $t = 0$, $\vec{P}(0) := \{P_1(0), \ldots, P_N(0)\}$ encoded by the genes $\{G_1, \ldots, G_N\}$ changes in time due to cross-regulation and auto-regulation of the expression of the member genes by their gene products. This initial state may be a response to an extracellular signal, such as a growth factor or a specific composition of nutrients in the medium. It will be imposed onto the network by the products of one or more "upstream" genes that are not themselves part of the network, insofar as their activity is not regulated by members of the network. Note that the borders of such a network are somewhat arbitrary: the upstream genes may well encode transcriptional regulators, too—e.g., a retinoic acid receptor acting on homeobox genes in a developing vertebrate limb, as long as

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

their activity is independent of the rest of the network. Also, the members of a network will regulate the transcription of genes downstream of the network—e.g., structural genes or genes involved in cell–cell signaling (8, 14).

To arrive at an analytically and computationally tractable mathematical model and to minimize the number of parameters involved in the model, a number of simplifying assumptions are necessary. (*i*) It is assumed that expression of the genes in the network is regulated exclusively on the transcriptional level. (*ii*) Each gene of the network is assumed to produce one and only one species of an active transcriptional regulator. (*iii*) In lack of a comprehensive kinetic theory of transcriptional regulation, it is simply assumed that enhancer elements mediating each regulator's effect on the expression of a target gene act independently from enhancer elements for other regulators of the same gene. (*iv*) In line with empirical evidence (15–18) it is assumed that strong cooperative effects of transcriptional activation by individual transcription factors are mainly responsible for strong transcriptional activation (repression) of a target gene. For reasons of computational convenience, the admissible concentration range for each P_i in $\{P_1, \ldots, P_N\}$ will be normalized and restricted to the interval [0,1], where $P_i = 1$ corresponds to the maximal possible concentration—i.e., the corresponding gene G_i is in a state of maximal transcriptional activation. On the basis of the above simplifications and observations, the following set of equations describing the dynamics of the expression states of the genes in a network is proposed:

$$P_i(t + \tau) = g_c(f_i(t) + \epsilon_i)$$

$$f_i(t) := \sum_{j=1}^{N} w_{ij} P_j(t) \qquad i \epsilon \{1, \ldots, N\}. \qquad [1]$$

τ is a time constant characteristic for the process under consideration and will depend on biochemical parameters, such as the rate of transcription or the time necessary to export mRNA into the cytoplasm for translation. The constants $w_{ij} \epsilon \Re$ in f_i describe the "strength" of interaction of the product of gene j with gene i—i.e., the degree of transcriptional activation ($w_{ij} > 0$) or repression ($w_{ij} < 0$) that the transcriptional regulator produced by gene j has on gene i. These constants define a matrix of "connectivities" $w = (w_{ij})$ of the network. It is the relative size of the connectivities that is relevant to the dynamics. $g_c(x)$ is some sigmoidal activation or "gain" function—e.g., $1/[1 + \exp(-cx)]$. The system's dynamics will be very sensitive to stochastic perturbations if any of its arguments is close to zero, unless the slope c of $g_c(x)$ is steep near $x = 0$. Therefore, only the limiting case of $c \to \infty$ will be considered. ϵ_i is a constant that reflects either a basal transcription rate ($\epsilon_i > 0$) of gene G_i or influences of upstream genes on gene G_i. Because the relevant properties of the network are independent of any such constant and because a large value of $|\epsilon|$ may override and, thus, obscure the effect of regulators inside the network, it will be set to zero. Further, a change of variables $\vec{S} := 2\vec{P} - (1, \ldots, 1)$ will facilitate analytical treatment. Eq. **1** thus becomes

$$S_i(t + \tau) = \sigma\left[\sum_{j=1}^{N} w_{ij} S_j(t)\right] = \sigma[h_i(t)], \qquad [2]$$

where $S_i = 1$ corresponds to gene G_i being ON and ($S_i = -1$) corresponds to G_i being OFF. $\sigma(x)$ is the sign function [$\sigma(x) = -1$ for $x < 0$, $\sigma(x) = +1$ for $x > 0$ and $\sigma(0) = 0$]. Clearly, many simplifications are involved—for example, about the absence of time delays, negligence of spatial and diffusion effects, and a small half-life of the proteins compared to τ. Although Eq. **2** is similar to a formalism used in the theory of neural computation (19) and in "spin glass" models of gene networks (20), the model is conceptually different from the latter class of models in that it is concerned with a specific type of genes.

I will focus on a subset of networks described by Eq. **2**—namely, those that, given an initial state $\vec{S}(0)$, converge ultimately to a stable equilibrium state $\vec{S}^{eq}$—i.e., a state such that $S_i^{eq} = \lim_{t \to \infty} S_i(t)$ holds for all i. Empirically speaking, $\vec{S}^{eq}$ is interpreted as a stable expression pattern of network genes attained through cross-regulation and autoregulation within the network. Any such pattern will induce expression of specific downstream genes affecting the phenotype of the organism. Although the model thus introduced is a haploid one, all results derived from it will hold in the diploid case as long as there is little allelic variation in the magnitude of connectivities.

Duplication of one or more genes in a network creates another network in a higher dimensional state space. Assume, without loss of generality, that G_1 through G_k are duplicated. The activation state of a network is then expanded, according to a function π:

$$\pi: \{-1, 1\}^N \to \{-1, +1\}^{N+k}$$

$$(S_1, \ldots, S_k, S_{k+1}, \ldots, S_N) \to$$

$$(S_1, S_1, \ldots, S_k, S_k, S_{k+1}, \ldots, S_N).$$

The matrix w is transformed into a $(N + k) \times (N + k)$ matrix w^d, in which columns and rows 1 through k are duplicated. The interaction strengths within the duplicated part of the network are kept identical to those in the corresponding original part. The state at time t of a network with some genes duplicated and the corresponding equilibrium state, if any such state is attained, will be denoted as $\vec{S}^d(t)$ and $\vec{S}^{d,eq}$, respectively. It is important to note that even if a network and a network derived from it by duplication of a number of genes will have the "same" initial states, $\vec{S}(0)$ and $\vec{S}^d(0) = \pi[\vec{S}(0)]$, subsequent states will, in general, not be equal, in the sense that $\vec{S}(t) \neq \pi^{-1}[\vec{S}^d(t)]$. Here π^{-1} denotes the inverse of π—i.e., a projection of the higher-dimensional state space in the original state space. Note that π^{-1} is invertible because the values of each pair of state variables corresponding to original and duplicated gene are the same at any time t if the networks had corresponding original states.

To compare equilibrium states attained before and after duplications, given the corresponding initial states $\vec{S}(0)$ and $\pi[\vec{S}(0)]$, the Hamming distance

$$d_h[\vec{S}^{eq}, \pi^{-1}(\vec{S}^{eq,d})] := \frac{1}{2} - \frac{1}{2N} \sum_{i=1}^{N} S_i^{eq}[\pi^{-1}(S_i^{eq,d})] \qquad [3]$$

will be used.

In lack of comprehensive evidence that allows conclusions regarding properties of "typical" initial states, equilibrium states, or connectivity matrices, a statistical characterization of networks having simple properties in common will be attempted. (*i*) It is assumed that the matrix w is drawn from a probability distribution with density $\rho(w) = \Pi_{i,j=1}^{N} \rho(w_{ij})$, where $\rho(w_{ij})$ is the density of entry w_{ij}. $\rho(w_{ij})$ is chosen to be the same for all entries and symmetrical around zero. These assumptions imply two network properties—namely, that each transcriptional regulator can activate or repress transcription depending on the promoter it acts upon (21) and that the numbers of transcriptional repressors and activators are approximately equal. (*ii*) Initial state and equilibrium state are characterized by the mean number of genes being ON—i.e., they are specified probabilistically as $\text{Prob}[S_i(0) = 1] = p_0$ and $\text{Prob}[S_i^{eq} = 1] = p_{eq}$ ($p_0, p_{eq} \epsilon [0, 1]$). Only properties of the whole ensemble of dynamical systems defined by the set of triplets:

$$E := (w, \vec{S}(0), \vec{S}^{eq}) \quad [4]$$

will be considered. Because $S_i^{eq} = \lim_{t\to\infty} S_i(t)$ is a constraint that any matrix w chosen for a given pair $[\vec{S}(0), \vec{S}^{eq}]$ has to satisfy, the ensemble is characterized by patterns of correlations between connectivities as well as between connectivities and the states of the network, interfering with efforts to analyze Eq. 4 analytically. Duplication of the same k genes in all elements of E induces a new ensemble $E^d := [w^d, \pi(\vec{S}(0)), \vec{S}^{eq,d}]$. Only elements of E that reach an equilibrium state after duplication have counterparts in E^d.

Two complementary approaches were pursued to characterize properties of Eq. 4 with respect to gene duplications. In the first approach, samples of the space E were obtained numerically. A pair of state vectors $[\vec{S}(0), \vec{S}^{eq}]$ was chosen with a pseudo-random number generator, according to the rules outlined above, and a stochastic search in the space of connectivity matrices was done until a matrix w was found that satisfied (Eq. 4) for the given pair of states. This stochastic search, a simulated annealing procedure in the matrix space, utilized a large (>100) set of matrices generated randomly according to the probability distribution described above. Exploration of the space was done by adding pseudo-random numbers distributed with density $\rho(w_{ij})$ to a fraction of the individual entries of the matrices. Once a matrix with the desired property had been found, a new pair of state vectors was chosen, and a new search was carried out. By these means, a sample E of size 100 was obtained. Several independent duplications of randomly chosen k-tupels of genes were done on each of the members of this sample, and sample statistics on Eq. 3 were calculated.

The second approach is an analytical approximation that assumes that the correlations occurring in the ensemble are weak. As will be seen below, both approaches yield qualitatively identical results.

RESULTS

Deviations from an optimal expression state of network genes will most likely cause deleterious phenotypic effects. The probability of any such deviation after a gene-duplication event—i.e., the probability that $d_h(\vec{S}^{eq}, \pi^{-1}(\vec{S}^{eq,d})) \neq 0$ in the ensemble of Eq. 4—will therefore be used as a probabilistic measure of the effect of gene duplication. From Eq. 2 it is clear that duplication of a whole network will have no effect. Intuitively, one might therefore assume that the effect of duplicating k out of N genes increases monotonically from zero as (*i*) k is increased from zero or (*ii*) as k is decreased from N, thus leading to a maximal effect for some intermediate value of k. Also, the increase in effect should be larger in case (*i*), such that the maximum effect occurs for $k < N/2$. The following considerations suggest that this intuition is correct.

Initially, it will be assumed that $p_0 = p_{eq} = 0.5$—i.e., a class of networks is considered in which 50% of the genes are "ON" in the initial and in the equilibrium state. Assuming that correlations between states and connectivities and correlations among connectivities in the ensemble of Eq. 4 are weak and that the marginal distribution of individual connectivities in E is well-approximated by the original distribution, $\rho(w_{ij})$, one can write the quantity $h_i(0)$ of gene G_i at time $t = 0$ as the sum of N stochastically independent, identically distributed random variables $X_j := w_{ij}S_j(0)$—*i.e.*, $h_i(0) = X_1 + \ldots + X_N$. It is easy to see that the probability distribution of X_j is identical to the distribution of w_{ij}. The quantity corresponding to $h_i(0)$ after duplication of k genes, denoted as $h_i^d(0)$, is given by $h_i^d(0) = 2X_1 + \ldots + 2X_k + X_{k+1} + \ldots + X_N$, if $\pi[\vec{S}(0)]$ was used as an initial state for the network with duplications. Next I assume that

$$\text{Prob}\{d_h(\vec{S}(\tau), \pi^{-1}[\vec{S}^d(\tau)]) \neq 0\} \propto \text{Prob}\{d_h(\vec{S}^{eq}, \pi^{-1}[\vec{S}^{d,eq}(\tau)]) \neq 0\} \quad [5]$$

holds, implying that information about the states after the first-time step is sufficient to make qualitative assertions regarding the probability of displacement of equilibrium states after gene duplication. $\text{Prob}\{d_h(\vec{S}(\tau), \pi^{-1}[\vec{S}^d(\tau)]) \neq 0\}$ depends only on the number of sign changes that occur in the quantities $h_i^d(0)$ with respect to $h_i(0)$. The probability of such a sign change if k genes are duplicated, p_k^N, is given by

$$p_k^N := \text{Prob}[h_i(0)h_i^d(0) < 0] = 2\int_0^\infty \left(\int_{x/2}^x \rho_k(y)dy\right)\rho_{N-k}(x)dx. \quad [6]$$

Here $\rho_l(x)$ denotes the density of a sum of l of the random variables X_i—i.e., the l-fold convolution of the density $\rho(w_{ij})$.

Under the above assumptions, the covariance $\langle (w_{ij}S_j)(w_{kj}S_j)\rangle$ over E vanishes for any $i \neq k$, implying that the number of quantities h_i that change signs after duplication is binomially distributed, yielding

$$\text{Prob}\{d_h(\vec{S}(\tau), \pi^{-1}[\vec{S}^d(\tau)]) \neq 0\} = 1 - (1 - p_k^N)^N. \quad [7]$$

A class of distributions of w_{ij} will now be considered that allows numerical evaluation of Eq. 6 and, ultimately, generalizations to a wider class of distributions. Assume that the density $\rho(w_{ij})$ is given by $(1 - c)\delta(w_{ij}) + cp(w_{ij})$ for all i,j, where δ denotes the Dirac delta function. p denotes a Gaussian density, $p(w_{ij}) = 1/(\sqrt{2\pi}\sigma)\exp[-w_{ij}^2/(2\sigma^2)]$, with variance σ^2 and mean zero. c is the probability of a connectivity being different from zero. Nc is the mean number of genes that influence the transcriptional state of any given gene. Therefore, c is a measure for the "density" of regulatory interactions in the network. Henceforth p_k^N from Eq. 6 will be denoted as $p_k^N(c)$, indicating that it may depend on c. In the case of $c = 1$, corresponding to a "fully connected" network with Gaussian distribution of connectivities, it can easily be seen that Eq. 6 is invariant with respect to changes in σ. This is a very important property, because it follows that p_k^N depends only on the ratio $r := k/N$—i.e., no matter what number of genes are involved in a network and no matter what the variance of the distribution of individual connectivities is, the mean effect of a gene duplication depends only on the fraction of genes duplicated. Results of numerical integration of Eq. 6 are shown in Fig. 2*A*. Most importantly, the effect of gene duplications, as defined in Eq. 7, is a unimodal function of r, with the maximum located at $r_{max} \approx 0.41$—i.e., on average, the effect of gene duplications on a network is largest if roughly 40% of the genes are duplicated. It is only this qualitative feature of p_k^N that will be relevant here. A comparison of Eq. 6 and an equivalent form, $p_k^N = 2\int_0^\infty \rho_k(x)\int_x^{2x}\rho_{N-k}(y)dydx$, shows that under the above assumptions $p_{N-k}^N = p_{k/4}^{N-k/4}$, and therefore that $p_{N-k}^N < p_k^N$ for $k/N < r_{max}$—i.e., p_k^N is not mirror symmetric around r_{max}, as exemplified by Fig. 2*A*. Fig. 1 shows that these theoretical predictions agree qualitatively with the statistics of duplication effects obtained from a computer-generated sample of E. Although differences in effects observed in the simulations are smaller, unimodality as well as a slight displacement of the peak to the left of $k/N = 0.5$ can be observed. This agreement is quite remarkable because the assumption of Eq. 5 implies that

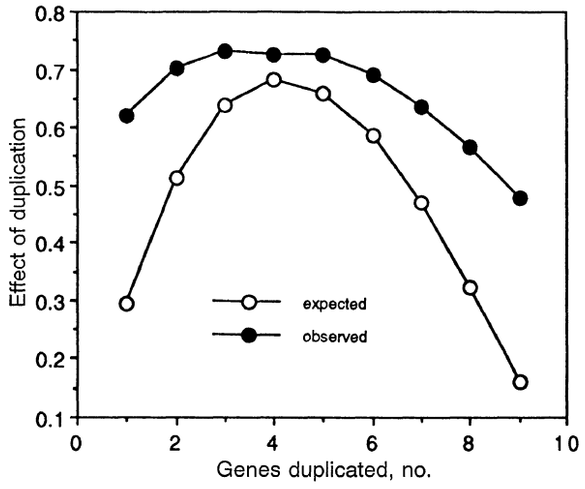

FIG. 1. Effects of gene duplications on a network of $N = 10$ genes with "dense" ($c = 1$) regulatory interactions. $p_0 = p_{eq} = 0.5$, $k \in \{1, \ldots, 9\}$, $p(w_{ij})$ Gaussian. "Expected" values are given by $1 - (1 - p_k^N)^N$ as in Eq. 7, using numerical integration of Eq. 6. "Observed" values represent numerical results obtained by simulating the effect of duplication of k genes out of N genes for each network in a sample of size 100 of the ensemble E. Numbers presented are based on at least 990 duplications for each k. Depicted is the fraction of networks that attain, after duplication, a stable equilibrium state that is different from the original equilibrium state.

all "dynamical" aspects of the dynamical systems in E are neglected by the analytical approach.

In the more realistic case of $c < 1$—i.e., a case where not all other genes influence the expression of any given gene, gene duplications can affect connectivities that are zero. Taking this into consideration, $p_k^N(c)$ becomes

$$p_k^N = \binom{N}{k}^{-1} \sum_{h=1}^{N} \sum_{i=max(h+k-N,0)}^{min(k,h)} \binom{h}{i}\binom{N-k}{h-i} c^h (1-c)^{N-h} [p_i^h(c)|_{c=1}]. \quad [8]$$

Fig. 2*A* shows results of numerical integration of Eq. **8** for several values of c, indicating that for any given ratio k/N effects of gene duplications decrease but that the overall qualitative pattern is conserved in that $p_k^N(c)$ is unimodal as a function of k. Also, note that the amount of this decrease is only minor between $c = 1.0$ and $c \approx 0.5$, whereas below $c < 0.5$ it becomes fairly large. Fig. 2*B* shows the corresponding results for some computer-generated samples of E, again showing good agreement between theory and simulation results. The small discrepancies from the expected patterns in Fig. 2*B* are attributable to the fact that, due to the vast amounts of computing time required, only small samples could be obtained. Theory predicts that in this range changes in effects are small, such that the sample sizes available do not provide the amount of statistical significance required.

Changing the mean number of transcriptionally active genes in the initial state and in the equilibrium state in the sense that $p_0 = p^{eq} \neq 0.5$ is permitted should, according to the assumptions of the model, have no effect on equilibrium states, because the covariance $\langle (w_{ij}S_j)(w_{ik}S_k) \rangle$ is still zero. This is where the agreement between theory and simulation breaks down because computer simulations show a slight reduction in the effect of a duplication of any given number of genes if these values are decreased or increased (results not shown). However, unimodality of duplication effect as a function of the fraction of genes duplicated still holds, leaving the qualitative relationships observed thus far unaffected.

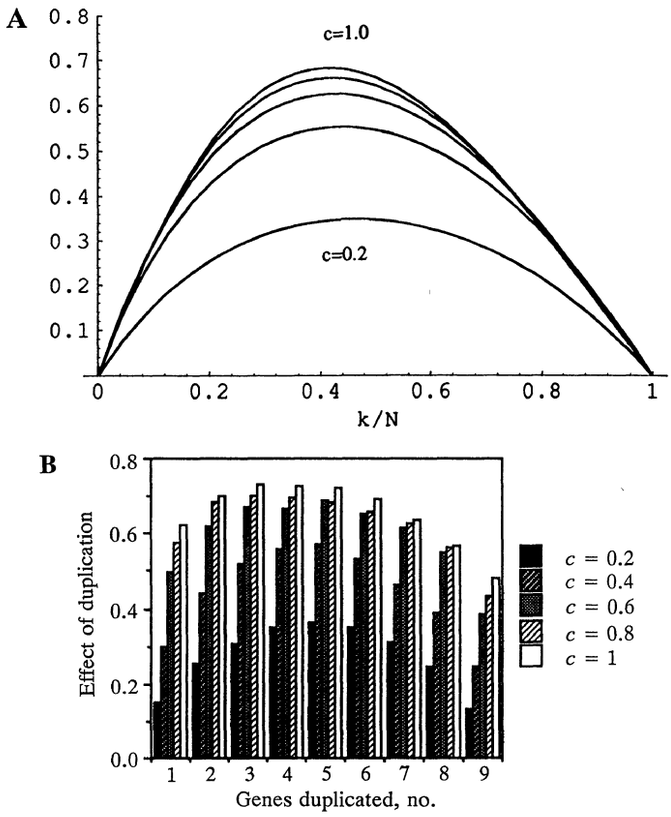

FIG. 2. Dependence of the effect of gene duplications on the "density" of regulatory connections, c. $N = 10$, $p_0 = p_{eq} = 0.5$, $k \in \{1, \ldots, 9\}$, $p(w_{ij})$ Gaussian. (*A*) Analytically expected values, given by $1 - [1 - p_k^N(c)]^N$, where $p_k^N(c)$ was obtained by numerical integration of Eq. **8**. For any given value of k/N, effects decrease monotonically as c is decreased. (*B*) Results obtained by simulating the effect of duplication of $k = 1$ through $k = 9$ out of 10 genes for each network in a sample of size 100 of the ensemble E. Figures presented are based on at least 990 duplications for each $k \in \{1, \ldots, 9\}$. Depicted is the fraction of networks that attain, after duplication, a stable equilibrium state that is different from the original equilibrium state. Calculation of G-statistics (22) indicated that, for each k, differences in the observed frequencies are significant at $P \ll 10^{-4}$.

The fact that networks of transcriptional regulators have different initial states and different equilibrium states in different cell types has thus far been neglected. However, a line of reasoning similar to the one above suggests that Eq. **5** still holds for any of the pairs of states involved, as long as pairs of states are stochastically independent. Also, computer simulations in which two pairs of states are assigned to each network show that there is no change in the qualitative relationships outlined above (results not shown).

The distributions involved in calculating p_k^N are assumed to be distributions of sums of independent random variables. Because such sums are asymptotically normally distributed, one might suppose that the results obtained above hold also for different types of distributions for w_{ij}, especially when N is large. Using double gamma distributions for individual connectivities [$p(x) = [2\Gamma(a)]^{-1} e^{-|x|} |x|^{a-1}$ for some parameter $a > 0$] in simulations analogous to the ones described above yields patterns qualitatively identical to those reported above even for network sizes as small as $N = 10$ (results not shown). This suggests that unimodality of effects as a function of k/N may be a generic property of Eq. **6**, conserved over a range of distribution types.

DISCUSSION

Little is known about specific properties of networks of transcriptional regulators, such as the distribution of the

number of genes that are transcriptionally active in different cell types or the average number of genes that regulate the expression of any given gene in the network. It is therefore reassuring that the model introduced above makes qualitatively identical predictions over a wide range of model parameters. Results obtained are robust with respect to changes in the "density" of regulatory interactions within a network and with respect to biases in initial state and equilibrium state. They are insensitive to distribution parameters and, presumably, to the specific distribution type used for individual interaction strengths. The main results can be briefly summarized as follows: The probability that a gene-duplication event alters the equilibrium expression pattern of network genes is a unimodal function of the fraction of network genes that are duplicated in a single duplication event. It is highest when ≈40% of the genes are duplicated. Also, according to the model, duplication of all network genes does not affect their expression pattern.

The measure for the effect of gene duplications used above refers to equilibrium states as being "identical" if the corresponding set of genes is expressed in the original network and in the network after duplications. Because of the high selective burden superimposed onto such equilibrium states (21), effects of changing the expression state of as few as only one gene will, in most cases, be highly deleterious if not lethal. Thus, even the small differences in effects evident from the simulation results in Fig. 1 will affect the kind of duplication events that are likely to be tolerated and that will therefore be predominant in the evolutionary record. Also, regulatory networks will, in most cases, have different initial states and equilibrium states for different tissues and/or developmental stages. Those states may be independently affected by gene duplications, which will massively enhance differences in effects of duplications of different numbers of genes.

Probably partly due to their ancient evolutionary origins (23) clusters of Antennapedia class homeobox genes have developed manifold and peculiar interdependencies in their expression patterns (6). Although they may, therefore, not present ideal examples for illustration of the principles proposed above, some data regarding their evolution is available, and the observed patterns coincide well with the predictions made above. Duplications of whole clusters as well as duplications of individual genes have left traces in the evolutionary record. However, there is as yet no reported incidence of a duplication of, for example, half of the genes of a cluster in a single event, in agreement with the prediction that such events are less likely to be tolerated by the developmental system.

Evolutionary forces shuffling long stretches of DNA, such as duplications of parts of chromosomes are abundant and act on a fairly short time scale. It has been shown that such forces were involved in the evolution of homeobox gene clusters (24). There is also evidence for a positive correlation between the rate of anatomical evolution and the rate of chromosomal evolution, suggesting that chromosomal mutations may be an important factor in the evolution of developmental systems (25). It seems therefore likely that chromosomal mutations contribute in a major way to the evolution of developmentally relevant gene networks. The question then arises as to whether there is an optimal organization of the genes involved in a network, so that gene duplications caused by these events are unlikely to have immediate deleterious effects. For example, a scenario in which half of the genes of a network are tightly linked on one chromosome, whereas the other half is also tightly linked but located on a different chromosome is clearly unfavorable. In most cases, duplications by means of chromosomal rearrangements will affect 50% of the genes in a network causing perturbations in development. Two most favorable forms of organization are predicted by the model. Either all genes of a network should be very closely linked in one chromosome or individual genes should be overdispersed in the genome—i.e., they should be as "far apart" from each other as possible. Duplication events are then likely to affect the whole cluster of genes or only single genes, respectively. Note that, once established, a transition between these forms of organization is unlikely because intermediate stages would occur that are more vulnerable to duplications of only parts of the genes of the network. Moreover, in the case of tight linkage, various events may cause additional cohesive forces, as in the case of homeobox gene clusters where regulatory sequences for individual genes may be spread throughout the cluster (26). Which type of organization is more likely to occur will depend on the type of events responsible for most duplications. If there is, in general, a positive correlation between the location of the gene and its duplicate in the genome, clusters will be more abundant than dispersed patterns. Note that the approach used here is very much a statistical one, in the sense that no particular events are "prohibited" by the model. To test the predictions made by the model, it will be necessary to compare multiple related taxa or, if possible, several networks of transcriptional regulators within a single taxon. Although insufficient empirical data is as yet available to rigorously test these predictions, data from various "genome projects" under way, as well as ongoing characterization of candidate networks, will most likely make such tests feasible soon.

I thank L. Buss, R. Bürger, W. Fontana, H. Hadrys, B. Misof, E. Mjølsness, F. Ruddle, D. Stein, R. Vaisnys, and G. P. Wagner for numerous helpful discussions and for comments on earlier drafts of the manuscript. This paper is communication no. 7 from the Center for Computational Ecology of the Yale Institute for Biospherics Studies.

1. Johnson, P. F. & McKnight, S. L. (1989) *Annu. Rev. Biochem.* **58,** 799–839.
2. Mermelstein, F. H., Flores, O. & Reinberg, D. (1989) *Biochim. Biophys. Acta* **1009,** 1–10.
3. Goodrich, J. A. & McClure, W. R. (1991) *Trends Biochem. Sci.* **16,** 394–397.
4. Gogos, J. A., Hsu, T., Bolton, J. & Kafatos, F. C. (1992) *Science* **257,** 1951–1954.
5. Sorger, K. S. & Pelham, H. R. B. (1988) *Cell* **54,** 855–864.
6. Sassone-Corsi, P., Ransone, L. J., Lamph, W. W. & Verma, I. M. (1988) *Nature (London)* **336,** 692–695.
7. Ingham, P. W. (1988) *Nature (London)* **335,** 25–32.
8. McGinnis, W. & Krumlauf, R. (1992) *Cell* **68,** 283–302.
9. Olson, E. (1990) *Genes Dev.* **4,** 1454–1461.
10. Rosenfeld, M. G. (1991) *Genes Dev.* **5,** 897–907.
11. Schöler, H. (1991) *Trends Genet.* **7,** 323–328.
12. McGinnis, N., Kuziora, M. A. & McGinnis, W. (1990) *Cell* **63,** 969–976.
13. Kappen, C., Schughart, K. & Ruddle, F. H. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5459–5463.
14. Davidson, E. H. (1991) *Development (Cambridge, U.K.)* **113,** 1–26.
15. Carey, M., Lin, Y.-S., Green, M. R. & Ptashne, M. (1990) *Nature (London)* **345,** 361–364.
16. LeBowitz, J. H., Clerc, R. G., Brenowitz, M. & Sharp, P. A. (1989) *Genes Dev.* **3,** 1625–1638.
17. Struhl, K. & Oliviero, S. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 224–228.
18. Wright, A. P. H. & Gustafsson, J.-A. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 8283–8287.
19. McCulloch, W. S., Pitts, W. A. (1943) *Bull. Math. Biophys.* **5,** 115–133.
20. Kauffman, S. A. (1993) *The Origins of Order* (Oxford Univ. Press, New York).
21. Lufkin, T., Dierich, A., LeMeur, M., Mark, M. & Chambon, P. (1991) *Cell* **66,** 1105–1119.
22. Sokal, R. R. & Rohlf, F. J. (1981) *Biometry* (Freeman, New York).
23. Kenyon, C. & Wang, B. (1991) *Science* **253,** 516–517.
24. Schughart, K., Kappen, C. & Ruddle, F. H. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 7067–7071.
25. Wilson, A. C., Sarich, V. M. & Maxson, L. R. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 3028–3030.
26. Whiting, J., Marshall, H., Cook, M., Krumlauf, R., Rigby, P. W. J., Stott, D. & Allemann, R. K. (1991) *Genes Dev.* **5,** 2048–2059.