J. theor. Biol. (1993) **165**, 269–289

A Model for the Evolution of Networks of Genes

BERTRAND CLARKE†, JAY E. MITTENTHAL‡ AND MARK SENN§

†*Department of Statistics, 2021 West Mall, University of British Columbia, Vancouver, BC V6T 1Z2, Canada,*
‡*Department of Cell and Structural Biology, 505 S. Goodwin Street, (and Center for Complex Systems Research, Beckman Institute; and College of Medicine), University of Illinois, Urbana, IL 61801, and*
§*Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.*

(*Received on* 8 *February* 1992, *Accepted in revised form on* 10 *March* 1993)

An organism persists through the activity of structural genes, which is co-ordinated by clusters of coupled regulatory genes. During evolution, changes of coupling within a cluster can increase the reliability with which its structural genes perform a task. To study the evolution of coupling, we have simulated and analyzed a stochastic model for a simple problem. The assumptions of the model are these: A network of regulatory genes co-ordinates the synthesis of four structural proteins, which associate in distinct heterodimers that form a heterotetramer. Mutation in cis-regulatory regions produces transitions among 64 types of network. In a population, each network reproduces in proportion to its fitness, which depends on its probability (reliability) of synthesizing the tetramer. Fitness-dependent attrition keeps the size of the population constant. Regulatory genes occur in a sequence of levels; each level is associated with a different family of transcription factors. The following results emerge: Because different messengers within a family can give networks with the same connectivity, the 64 types of networks cluster into eight equivalence classes. During evolution with a low mutation rate, high-fitness classes can be approached through various paths on a fitness landscape. With a higher mutation rate, networks remain more uniformly distributed among the 64 types, and lower-fitness networks remain preponderant. An initially homogeneous population becomes more heterogeneous through mutation, but selection according to fitness later reduces its diversity. During this process the dispersion of the population over the possible networks increases, then decreases as the population approaches a unique steady state.

1. Introduction

Although the evolution of individual proteins and nucleic acids has been studied extensively, the evolution of networks of interacting genes is not well understood. In such a network, or net, each gene can bind gene products that control its activity. Nets of genes underlie the phenotype of an organism; they mediate the processes that carry it through its life cycle and enable its lineage to persist. Therefore, to understand the evolution of organisms, it is essential to understand how nets of genes evolve.

0022–5193/93/023269+21 $08.00/0

The evolution of gene nets has been studied by modeling the dynamics of a population of nets (e.g. Weisbuch, 1986, 1991; Langton *et al.*, 1992; Kauffman, 1993). Typically, the nets differ in the alleles that are present and in their connectivity—in the genes to which the product of each gene binds. Mutations of various kinds can change the connectivity of a net, or the input–output relations for its genes. A point mutation converts each net to a net which is its neighbor in a net space. If each net has a fitness in meeting constraints, which determines its rate of proliferation, the fitnesses define a fitness landscape on the net space. As a population of nets mutates and proliferates, points corresponding to the nets flow on the fitness landscape.

Random Boolean nets have been used to investigate the flow of a population on its fitness landscape (reviewed by Kauffman, 1989*a*, *b*, 1991, 1993). Populations of interacting genetic rules that undergo mutation and recombination have also been used in genetic algorithms and classifier systems to solve optimization problems (Goldberg, 1989). Typically, in such models all genes are equivalent, at least initially. A Boolean function, chosen at random from a class of admissible functions, characterizes the input–output relation for each gene. The contribution that each allele makes to fitness, and the consequences of mutation, are also assigned at random. However, a subset of such nets with special features is especially well suited to represent actual gene nets. Because the properties of this subset may differ from those of the entire set, it is necessary to study this subset specifically. For example, Kauffman (1971, 1989*b*, 1991) noted that many genes have few inputs and are governed by canalizing Boolean functions, in which at least one input has one value that can determine the output of the genes, regardless of the other inputs. He found that such model nets often display patterns of global order analogous to those in real gene nets.

In particular, model and real nets with few inputs per gene and with canalizing functions often have dynamic modules. A dynamic module is a dynamical system in which the activity in a cluster of genes changes with time. This dynamical system can be activated by a small perturbation, such as a change in a single input. Examples of modules include the fight-or-flight response elicited by adrenalin, and the activity of a morphogenetic field in generating an organ (see Clarke & Mittenthal, 1992; Mittenthal *et al.*, 1992). In real nets the genes are heterogeneous in ways that seem likely to affect the evolution of modules: There are distinct structural and regulatory genes, and genes occur in families. To understand the significance of differences and similarities among genes for the evolution of modules, we have explored a model of a small gene net. The model incorporates several features found in real nets, as discussed in the following section, but it is not a model of a specific real system; rather, it is designed for ease of analysis and comprehension.

BIOLOGICAL FEATURES OF THE MODEL

In organisms, a net of regulatory genes controls the activity of structural genes, the products of which are enzymes and structural proteins. Many previous models of evolution have not distinguished these classes of genes. However, several lines of investigation suggest that changes in the connectivity of regulatory genes influence

the performance of organisms in different ways than changes in structural genes. Wilson (1975) and King & Wilson (1975) found little difference in the structural proteins of chimpanzee and human; therefore, they proposed that these species have major differences in the connectivity of regulatory genes. Hedrick & MacDonald (1980) have argued on a theoretical basis that mutations in regulatory genes are more likely to underlie major rapid changes during evolution than are mutations in structural genes. Levinthal (1990) proposed that changes in capabilities of bacteria depend more on changes in connectivity of regulatory genes than on changes in the structural genes they control.

These arguments led us to formulate a model in which regulatory genes activate structural genes to produce subunits of an enzyme. In the nets of the model, as in many real nets, regulatory genes are coupled in series and in parallel to form a tree, or cascade, without feedback loops. Each regulatory gene synthesizes a transcription factor, or messenger, that can regulate the activity of a target gene by binding to its input region at a cis-regulatory site (cis element). The cascade consists of a sequence of stages, or levels. The genes at one level are activated synchronously and are all members of the same family; a different family occurs at each level.

A theoretical argument led us to this structure: In a sequence of coupled genes, if the same messenger appears at two stages, production of the messenger at the later-acting stage can feed back to act on the cis element at the earlier-acting stage, resulting in an infinite recursive loop. If the messengers at the earlier- and later-acting stages are not identical but are in the same family, a mutation may also produce a recursive loop. Such loops do occur; for instance, in flowers of *Arabidopsis*, a homeotic mutation that converts the carpels to a new flower can generate a flower with a stack of more than 70 organs (Yanofsky *et al.*, 1990). However, the possibility of such a loop may be a luxury allowed only in particular pathways of differentiation. In more universal sequential processes a different family of transcription factors may be used at each stage. This is the case in the cell cycle (Bodnar, J., personal communication).

Our model uses five regulatory genes in two families. Alternative connections among the genes produce an ensemble of 64 types of nets. A point mutation in a regulatory gene converts one type of net to another. Each type has a distinct connectivity, and is assigned a fitness that increases with its reliability in synthesizing a heterotetrameric structural protein. In the nets with high fitness, genes are activated in a hierarchy of dynamic modules: Modules make two heterodimers, and a higher-level module activates these two molecules to make the tetramer. We calculated the movement of a population of nets over the fitness landscape by evaluating the time dependence of the probabilities of occurrence of the 64 types in the population, with mutation among types.

IMPLICATIONS OF THE MODEL

Biological systems often embody a hierarchy of dynamic modules (Simon, 1962; Mittenthal *et al.*, 1992). Our previous study of the model used here (Clarke & Mittenthal, 1992) showed that such a hierarchy can increase the reliability with

which a small net of genes can synthesize the tetramer. We conjectured that hierarchical modular organization is common because it is likely to evolve. The present work supports this idea. It shows that if low-level modules evolve earlier, organisms embodying these modules are more fit, become a predominant part of the population, and are the ancestors of organisms with higher-level modules. However, in other scenarios, high-level modules can evolve without the intermediate predominance of organisms with low-level modules.

The modeling shows the importance of level-specific families of transcription factors in several respects. First, the families imply the existence of equivalence classes. All nets in an equivalence class have the same connectivity among levels (though they may use different messengers), the same fitness and equivalent transitions. An equivalence class is a many-to-one mapping; it is degenerate, as the genetic code is degenerate in that several triplets code for the same amino acid. Dealing with equivalence classes allows one to focus on the significant structural differences between nets of different classes. The use of equivalence classes reduces the complexity of the fitness landscape and of the population's movement on it, if it is sufficient to trace the relative positions and dynamics of the classes rather than of types of nets. In our model, the fitness landscape for classes has only two peaks separated by a valley. Second, because a different family of messengers is used at each level, mutations can change connectivity independently at different levels. This dissociation among levels may accelerate the accumulation of nets with higher fitness.

The results show that the prevalence of classes—the number of members of each class present in the population at a given time—depends on the pattern of allowed transitions, on the mutation rate, on the fitness of each class and on its cardinalty—the number of nets it contains. The fittest classes become the most prevalent only when the mutation rate is sufficiently low. As the mutation rate increases, all nets tend to become equally prevalent. Consequently, equivalence classes with high cardinality are more prevalent than the fittest classes. Thus, in our model, mutation rate behaves somewhat as does temperature in equilibrum statistical mechanics.

The appearance of an analog of temperature in evolution suggests the possibility that its conjugate variable in thermodynamics, a measure of dispersion or entropy, may be of interest for evolution. Indeed, Brooks & Wiley (1988) suggested that dispersion could be used to characterize an evolving population. We find that an initially homogeneous population becomes more heterogeneous through mutation, but that selection according to fitness subsequently reduces the diversity of the population. This pattern of early experimentation followed by later standardization is common in evolution at all taxonomic levels (Gould, 1989). Our calculations show that correspondingly, for biologically reasonable mutation rates, the dispersion increases, then decreases.

In section 2 we present the qualitative assumptions of the model and then state it formally, presenting the dynamical equations and demonstrating the stability of the steady state. In section 3 we present the time courses of prevalence and dispersion inferred from the model, showing how these depend on the pattern of transitions, cardinality, fitness and mutation rate. Section 4 discusses implications of the model.

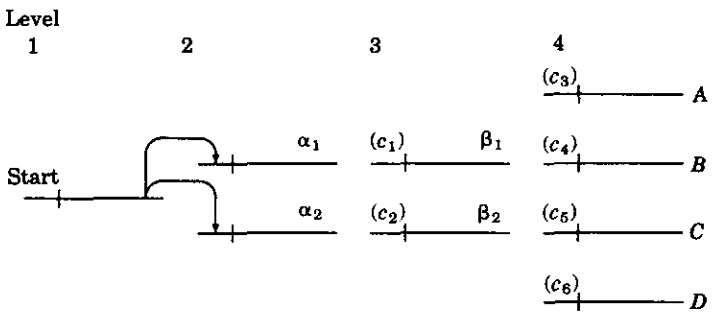

FIG. 1. Genes and levels of the nets in the model. We assume that the net is activated when a messenger binds to the cis-element Start of the level 1 gene. The product of this gene binds to both genes at level 2. A Greek letter denotes the messenger made by the output region of the regulatory genes at levels 2 and 3. The indices $c_1 \ldots c_6$ denote cis elements capable of binding one of the two messengers made by the previous level, but not both. *A*, *B*, *C* and *D* are monomeric proteins synthesized by the four structural genes at level 4.

2. Model

QUALITATIVE ASSUMPTIONS OF THE MODEL

The model nets have four levels of genes—three levels of regulatory genes and one level of structural genes. As Fig. 1 shows, the first level has one regulatory gene, the second and third levels have two regulatory genes each. The fourth level has four structural genes. Each gene has an input region containing a single cis-regulatory element that can bind a transcription factor (messenger). When the cis element binds a messenger, the output region of the gene synthesizes a protein, a messenger (for regulatory genes) or a structural protein (for structural genes). Transcription and translation are subsumed in this process.

The messengers made by genes at each level belong to a different family, and only bind to cis elements of genes at the next level. Thus, as Fig. 1 shows, each gene at level 2 makes a messenger in the alpha family of transcription factors. Each type of alpha messenger can bind only to a specific cis element, which may be present in the input region of either gene (or both, or neither) in level 3. Similarly, the genes of level 3 make transcription factors in the beta family, and each of these can bind to a specific cis element that may occur in the input region of any structural gene. The products of the four structural genes are structural proteins *A*, *B*, *C* and *D*; these monomers associate to form dimers *AB* and *CD*, which form the tetramer *ABCD*. Since each of the six cis elements can bind one of two kinds of inputs, these assumptions generate the set of $2^6 = 64$ possible nets of interest here.

Because the nets are trees, their internal (physiological) dynamics can be neglected; in each type of net the same set of genes will be activated whenever the gene at the first level is activated. Only the evolutionary dynamics associated with mutation is of interest. Each time step corresponds to a generation. During a time step, a cis element may be unchanged or may mutate to another element in the same family.

We assume that all such mutations are equally probable. The probability of a mutation is assumed to be sufficiently small that one may neglect the possibility of two mutations in one time step. So, during a time step, a given net can remain unchanged with probability p, or can change one of its six cis elements, producing a different net. We call $1-p$ the mutation rate; each mutation has probability $(1/6)(1-p)$. Using a computer, we generated the 64 possible nets and the 64×64 matrix for transitions resulting from mutation. The elements on the main diagonal of the transition matrix are p. Off the main diagonal, an element is $(1/6)(1-p)$ if the transition is possible; otherwise the element is 0. The matrix is symmetric, since each transition is reversible.

During a time step, each type of net reproduces in proportion to its prevalence in the population and to its fitness. As in our previous study of the *ABCD* model (Clarke & Mittenthal, 1992), we assigned fitnesses by assuming that, in its physiological operation, each regulatory gene is slightly unreliable in producing a messenger that can activate genes at the next level. Then the reliability of a net—the probability that it makes a tetramer—can be calculated from the reliabilities of the regulatory genes, although the difficulty of the calculation depends on the lifetimes of the monomers. Here, we did not make a formal calculation, but assumed qualitatively that the reliability, and hence the fitness, of a type of net is greater, the fewer regulatory genes are required to activate the structural genes. The numerical values of fitnesses were assigned *ad hoc*, because small variations in them did not affect the results. The fitnesses are fractions near unity, whereas the fitnesses conventional in population genetics have values slightly greater than unity, and represent the average number of progeny per organism. The fitnesses that we use could be converted to conventional fitnesses by multiplying by a positive constant; however, this factor would cancel out in our equations.

The 64 nets segregate into eight equivalence classes under graph isomorphism. Two nets are isomorphic if interchanging labels of cis elements or monomers converts one net to the other, given the constraint that the tetramer forms through two heterodimers. Within each equivalence class the nets are microscopically distinct but macroscopically equivalent, in that they have the same structure and therefore the same fitness, and they make isomorphic transitions to other classes. Thus, the equivalence classes define a small set of macroscopically distinct types of nets, making the analysis less cumbersome. The eight equivalence classes show four qualitative patterns of net connectivity characterized by Clarke & Mittenthal (1992), and shown in Fig. 2. The nets with highest reliability have a shotgun pattern of connectivity, in which all four structural genes are activated by the same messenger. In a modular pattern, one messenger activates synthesis of A and B, while another activates synthesis of C and D. This pattern of activation matches the association of monomers in the dimers AB and CD, and so defines dynamic modules that make AB and CD; it has a relatively high reliability. The remaining nets perform with relatively low reliability; in them, messengers activate synthesis of monomers in patterns unrelated to the subsequent association of monomers. A pseudomodular net activates synthesis of three monomers with one messenger; these monomers form only one of the two dimers. A non-modular net synthesizes two pairs of monomers, but the monomers of a pair do not associate to form a dimer.

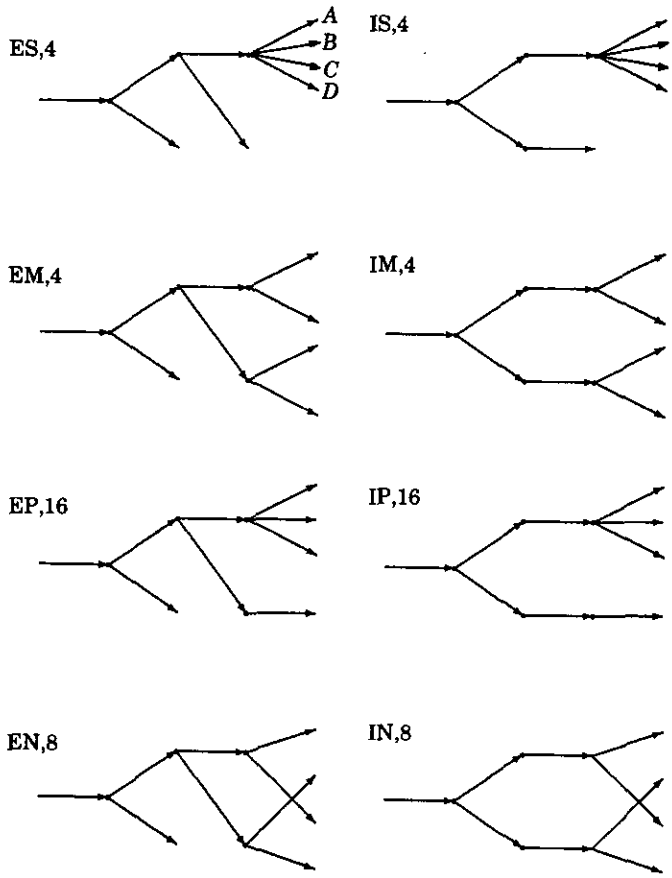

FIG. 2. Canonical nets from each of the eight equivalence classes. The classes are listed in order of decreasing fitness, although classes ES and IS have the same fitness. The cardinality of each class—the number of microscopically distinct nets in each class—is next to its label. E, efficient; I, inefficient; S, shotgun; M, modular; P, pseudomodular; N, non-modular. For the representative of the ES class, the synthesis of *A*, *B*, *C* and *D* is indicated; these monomers combine to form the dimers *AB* and *CD*, which form the tetramer *ABCD*.

Each of these four patterns is represented in two equivalence classes. In nets of the four classes called "efficient", any diversity of messengers occurs at level 4; both level 3 genes bind the same alpha messenger. This reduction in the total diversity of messengers increases the reliability of efficient nets relative to nets called "inefficient", in which level 3 genes bind different alpha messengers. Note that there are relatively few high-fitness nets—eight shotgun and eight modular—compared to 48 lower-fitness nets.

Figure 3 shows the fitness landscape for transitions among the equivalence classes, the class landscape. The restriction to transitions within families of messengers appears as a dissociation of alpha transitions (between efficient and inefficient nets) from beta transitions (among shotgun, modular, pseudomodular and non-modular nets). That is, efficient and inefficient forms of a net type can interconvert without a

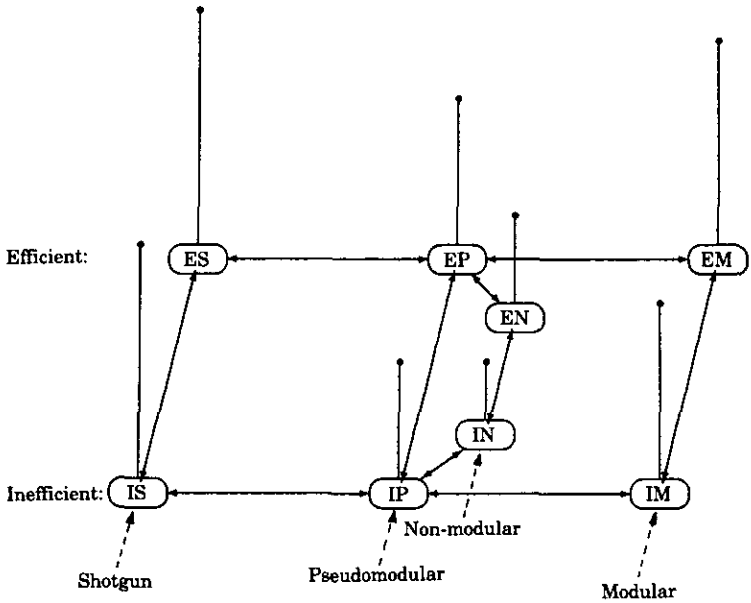

FIG. 3. The transition diagram; solid arrows indicate the allowed transitions between equivalence classes. Dashed arrows denote the four patterns of net connectivity. Relative fitnesses of the eight classes are indicated by the heights of the vertical lines. The black dots atop the lines define the fitness landscape, with a valley separating two peaks.

change of modularity; and transitions of modularity can occur without a transition of efficiency. Transitions between the high-fitness modularity classes, shotgun and modular, can only occur through lower-fitness pseudomodular and non-modular nets. In terms of the fitness landscape, transitions between two peaks can only occur by traversal of a valley.

The landscape in Fig. 3, for transitions among the eight classes, is a simplified form of the fitness landscape for transitions among the 64 nets, the net landscape. A point mutation (one-step transition) converts each net into one of six neighboring nets. Thus, the net landscape should be displayed on a six-dimensional hypercube. To explore the topography of the net landscape, note that transitions between two nets in the same equivalence class require more than one step. One-step transitions occur between nets of the shotgun classes, ES and IS, which have the same fitness. So, in the net landscape, a ridge is associated with the shotgun nets, as in the class landscape. However, all transitions among nets in the modular classes, EM and IM, proceed through nets of the lower-fitness pseudomodular and non-modular classes. Thus, the net landscape has a spike corresponding to each modular net, surrounded by lower fitness neighbors; it is quite rugged near the modular nets.

FORMAL REPRESENTATION OF THE MODEL

We considered two models that differ in the sequence in which mutation and reproduction occur. A net may mutate, with a corresponding change in fitness, before

it reproduces asexually, as in a unicellular organism. Alternatively, a net may have a fixed fitness that determines its rate of asexual reproduction, but may produce mutant offspring. This model represents a multicellular organism with a soma (body) that has a constant phenotype and fitness throughout its life; the organism reproduces according to its fitness. The soma contains a germ line that can produce mutant offspring. For the limited range of fitness values that we used, the two models gave nearly equivalent results. Hence, we only present the unicell model, in which mutation precedes asexual reproduction.

Imagine a population of nets of distinct types not yet grouped into equivalence classes. Let the vector $v(t) = (v_1(t), \ldots, v_{64}(t))'$ represent the proportions of each of the 64 types in the population at time t, where each $v_i(t) \geq 0$, $\sum_{i=1}^{64} v_i(t) = 1$, and the prime denotes transposition. To each type we associate a fitness f_i which is positive but not otherwise constrained. We assemble the f_i's into a fitness matrix **F**: **F** has the ordered f_i's on its main diagonal and all its off-diagonal entries are zero. The transition matrix **T** depends on the mutation rate $1-p$. Its (i,j)-th entry $t_{i,j}$ is the probability of a transition from type j to type i.

In the unicell model we apply the matrix **FT** to $v(0)$ to obtain $v(1)$ by

$$v(1) = \frac{\mathbf{FT}v(0)}{I'\mathbf{FT}v(0)},$$

where I is the vector of length 64 in which all entries are unity. Similarly, $v(2)$ is obtained from $v(1)$ by applying **FT** and normalizing. In general, we have

$$v(t+1) = \frac{\mathbf{FT}v(t)}{I'\mathbf{FT}v(t)}. \tag{1}$$

(In the analogous equation for the soma-germ model, the order of **F** and **T** is reversed.) We can represent (1) in terms of co-ordinates in general by

$$v_i(t+1) = \frac{f_i \sum_j t_{ij} v_j(t)}{\sum_{ij} f_i t_{ij} v_j(t)}. \tag{2}$$

The character of this non-linear dynamic is more evident if it is written as a difference equation:

$$v_i(t+1) - v_i(t) = \frac{\left[f_i \sum_j t_{ij} v_j(t) - v_i(t) \sum_{ij} f_i t_{ij} v_j(t) \right]}{\sum_{ij} f_i t_{ij} v_j(t)}. \tag{3}$$

Expression (3) resembles the classical competition equations for population dynamics (May, 1981), in that the numerator contains "birth" terms that are linear in the v_i's and "death" terms that are quadratic in $v_i v_j$. It shows the competition among nets to survive the attrition that keeps the total size of the population constant. Note that if all the f_i's are 1, then the model coincides with the Markov chain defined by T. The competition expressed by the normalization destroys the Markov property by introducing non-linearity.

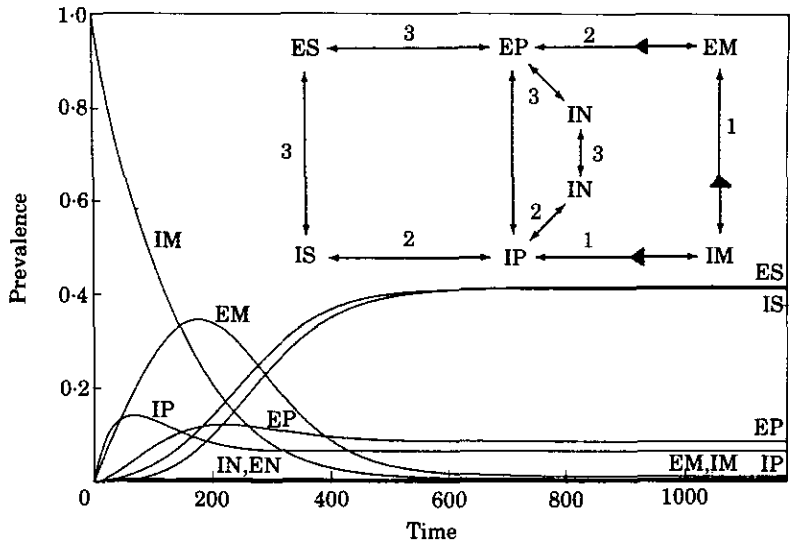

FIG. 4. The time course of class prevalences when the population is initially distributed uniformly over net types within the inefficient modular (IM) class. $p = 0{\cdot}99$. The fitness values used to obtain the results in Figs 4–8 were: 0·9 (the fitness of the ES and IS nets), 0·85 (EM), 0·8 (IM), 0·75 (EP), 0·7 (IP), 0·65 (EN), 0·6 (IN). In the inset transition diagram, large arrowheads denote directions of major consistent transitions.

To pass from the dynamical equations for v to the prevalences of the eight equivalence classes we define a new vector $z(t) = (z_1(t), \ldots, z_8(t))'$, where $z_i(t)$ is the proportion of nets of class i in the population. Clearly,

$$z_i(t) = \sum_{j \in [i]} v_j(t), \tag{4}$$

for $i = 1, \ldots, 8$ and $[i]$ denotes the equivalence class of type i. The curves plotted in Figs 4–8 use the time courses of the $z_i(t)$'s calculated as in (4).

For the dynamical system defined by (1) it can be shown that there exists a unique asymptotically stable equilibrium point. A formal statement and proof of this result is given in the Appendix. This ensures that the convergences obtained computationally in the next section are valid.

3. Results

For the unicell model we consider the time course of class prevalences with various initial conditions and mutation rates.

TIME COURSE OF CLASS PREVALENCES WITH VARIOUS INITIAL CONDITIONS

First, we contrast the flow of the population of nets over the class landscape, if all nets are initially in a high-fitness vs. a low-fitness class. Nets in the inefficient modular

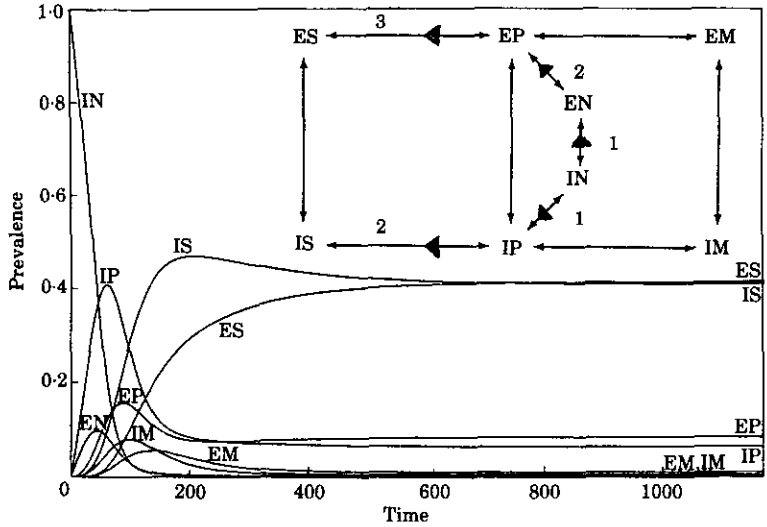

FIG. 5. As Fig. 4, for the inefficient non-modular (IN) class.

(IM) class have a high fitness; they are in the modular peak on the class landscape. An IM net has two clusters of genes, each of which makes a dimer (Fig. 2). These clusters might have evolved initially if it was selectively advantageous to make the dimers *AB* or *CD* separately. However, the fitness of IM nets is lowered because these clusters are coupled to the first (start) gene indirectly, through pathways involving distinct genes at levels 2 and 3. Alternatively, all nets might initially be in the inefficient non-modular (IN) class, at the bottom of the valley in the class landscape. In IN, as in IM, two clusters of two structural genes are coupled through separate pathways to the start gene. However, IN is the class with the lowest fitness because the clusters do not match the constraint of forming the dimers *AB* and *CD*.

Figure 4 shows the flow from an initial population of IM nets. The two dimer-making clusters tend to become more tightly coupled with time, first through transitions to IP and EM, then through EP to the shotgun classes ES and IS. Thus, with dimer-making clusters a hierarchy of modules is transiently predominant, in EM (and, of course, in IM), before the shotgun classes become more abundant. By contrast, as Fig. 5 shows, in the less plausible initial state with all nets in the IN class, there is a transient peak in other low-fitness nets, IP and EN, and scarcely any peak of modular nets before shotgun nets predominate. Thus, the organization with highest fitness can evolve through a hierarchy of modules with an evolutionarily plausible initial condition, or may bypass this hierarchy in a less plausible scenario.

We now consider the other factors on which the population flow depends—the pattern of transitions among types of nets, the cardinality and fitness of the classes, and the mutation rate. The conclusions given here apply only to mutation rates near $p = 0{\cdot}99$, as in Figs 4 and 5; other mutation rates will be considered later.

As regards transitions among types of nets, one might expect mutation to generate sequentially the classes that are more transitions distant from the initial class; this is

generally so. For example, starting from IM, the prevalence of classes one step away, IP and EM, rises and peaks most rapidly (Fig. 4). Then nets leave IP and EM for IN, EP and IS, which are at the distance $d = 2$ transitions from IM, and for EN and ES, which can be reached by three transitions. However, pairs of classes that are equivalent distances from IM can peak at different times. For example, IN peaks earlier than IS, though both have $d = 2$; EN peaks earlier than ES, though both have $d = 3$. Here, the class with the greater cardinality peaks earlier: IN and EN, with eight nets, peak earlier than IS and ES, which have four. Classes with larger cardinality might be expected to peak earlier because there are more ways to enter and leave them.

The cardinality of a class affects its asymptotic prevalence, as well as its transient prevalence. The low-fitness classes EP and IP are more prevalent asymptotically than the fitter classes EM and IM, because the former have the greatest cardinality. The stability analysis shows that the asymptotic prevalences are independent of the initial condition (cf. Figs 4 and 5); they depend only on the mutation rate, the fitnesses and the transitions among nets. If two classes have the same pattern of transitions and cardinality but have different fitness, the fitter class has a higher asymptote. This is evident on comparing the efficient (fitter) and inefficient classes of each type, such as EP and IP or EM and IM. (ES and IS have the same fitness.)

In summary, our results suggest several conclusions for p near 0·99. (i) Mutation tends to generate sequentially the classes that differ from the initial class by a greater number of transitions. (ii) Classes with larger cardinality tend to display more rapid transients. (iii) The asymptotic prevalence of a class tends to increase with its cardinality and its fitness. (iv) The asymptotic prevalences are independent of the initial condition; they depend only on the mutation rate, fitnesses and transitions. These conclusions apply to all of the cases we have examined—the unicell and somagerm models, starting with the whole population in each of the eight classes, or starting with the population equally distributed among the 64 types of nets.

THE DEPENDENCE OF THE TIME COURSE OF PREVALENCES AND OF ENTROPY ON THE MUTATION RATE

The mutation rate, $1-p$, affects the preceding conclusions. Figure 6 shows that, starting from IM, if the mutation rate is lower, the approach to the asymptotic values is slower. However, the sequence of peaks does not change. At a lower mutation rate, fitness becomes more important relative to cardinality in determining prevalence. For example, note that the transient peak in EM is larger, relative to peaks in the four lower-fitness classes, at higher p. Figure 7 shows the dependence of the asymptotic prevalences on p. As the mutation rate approaches zero, the asymptotic prevalence of all classes but the fittest approaches zero, and the population becomes equally partitioned between the two fittest classes, ES and IS.

At high mutation rate the asymptotic prevalence of each class tends to become proportional to its cardinality. That is, the nets of the population become equally partitioned among the 64 types of nets. However, at high mutation rate the formal model does not correspond to the biological situation. In the formal model we

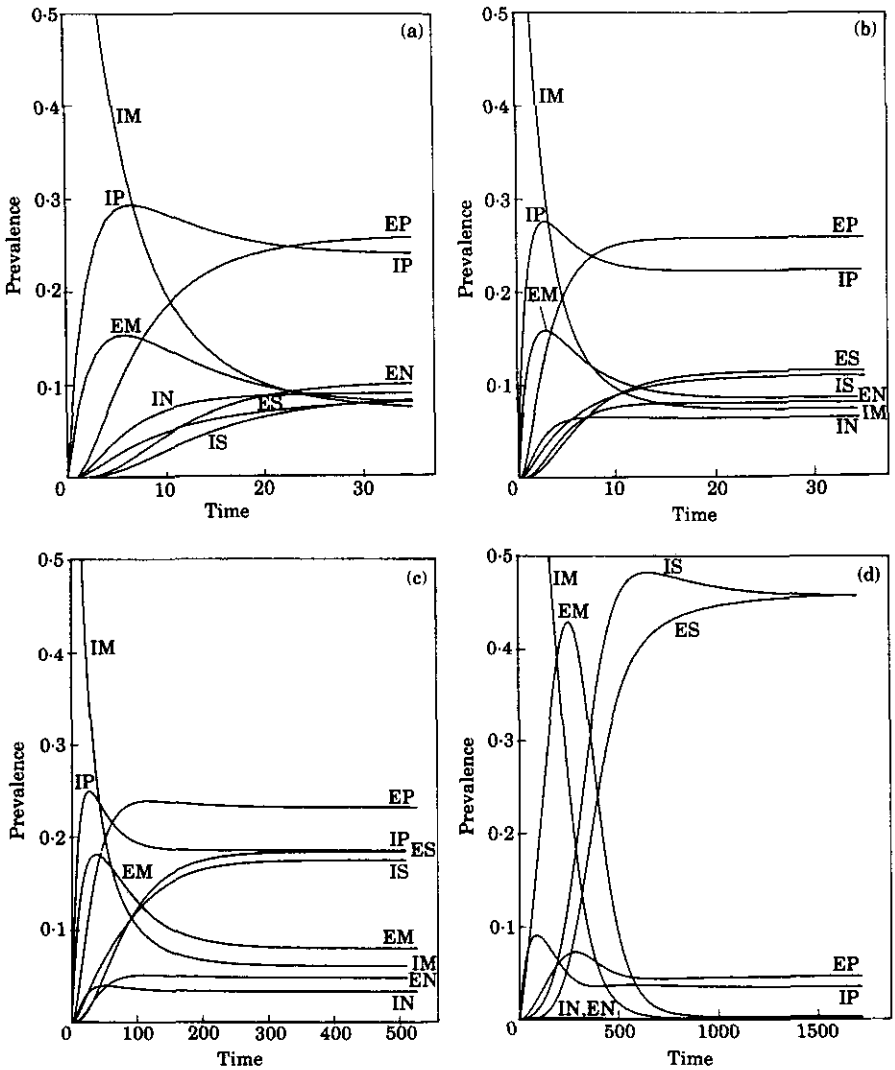

FIG. 6. As Fig. 4, starting from the IM class, for different mutation rates.

assumed that the number of mutations per generation was likely to be only 0 or 1. This assumption is reasonable at biologically appropriate low mutation rates. As the mutation rate increases there is likely to be more than one mutation per generation. We have not modeled this situation.

We obtained the time dependence of dispersion in two senses. The net dispersion, for the population of 64 net types, is defined by $\sum_{i=1}^{64} v_i(t) \log(1/v_i(t))$. The class dispersion for the set of the eight classes, is defined by $\sum_{i=1}^{8} z_i(t) \log(1/z_i(t))$. Figure 8 shows the time course of these measures of dispersion for two mutation rates. Initially, the nets are uniformly distributed among the four net types in the IM class.

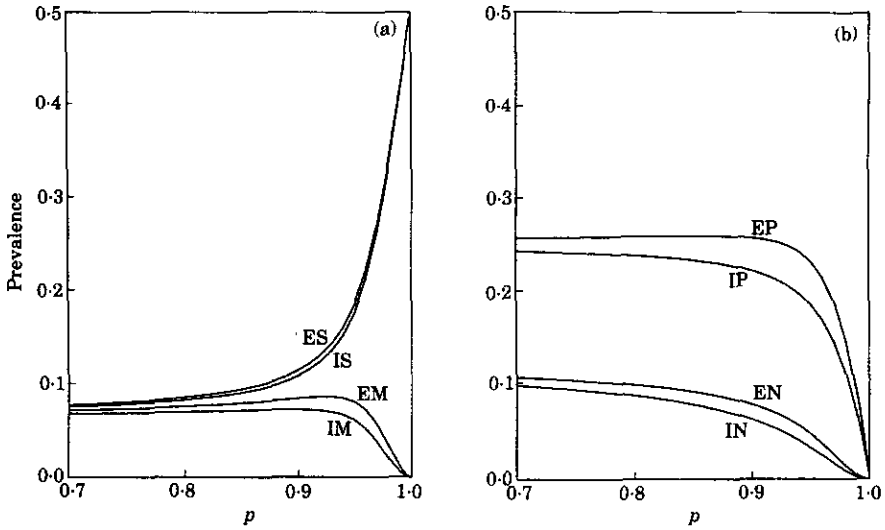

FIG. 7. Graphs of asymptotic prevalence vs p. (a) Fitter classes; (b) less fit classes.

Therefore the class dispersion is initially zero. At a high mutation rate both dispersions increase monotonically to an asymptotic value. However, at a biologically reasonable low mutation rate, both dispersions increase to a maximum and then decrease to an asymptotic value. This is the behavior expected for the pattern of evolution suggested by Gould (1989) in which early experimentation precedes later standardization.

4. Discussion

We have explored a stochastic model for the flow of genetic nets on a fitness landscape. Here, we relate our model to previous models and to biological observations.

RELATION TO PREVIOUS MODELS

The nets in our model are random Boolean nets in which N genes interact. K genes affect the state of each gene through a Boolean function; these patterns of dependency specify the connectivity of a net. Such NK nets have been extensively studied (see reviews by Kauffman, 1989*a*, *b*, *c*, 1991). In static NK models the states of genes do not vary with time; the interactions among genes represent epistatic effects, as in classical population genetics (Ewens, 1979; Feldman, 1989). In dynamic NK models the Boolean functions govern transitions in the states of genes during successive time steps.

We used a dynamic NK model to investigate the evolution of modular organization in Boolean nets with families of genes. Each net has the same $N = 9$ genes, and each gene has $K = 1$ inputs; a gene is on when its input is present. Because the nets

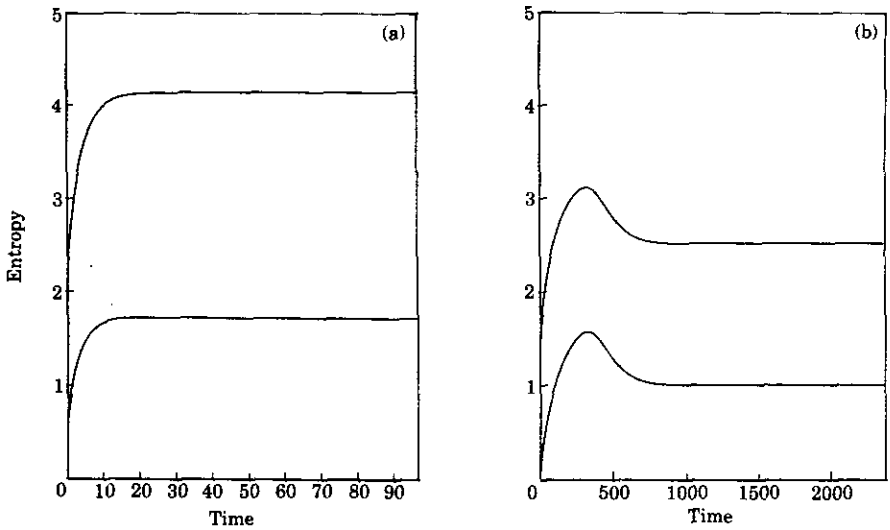

FIG. 8. Time course of dispersion. In each panel the upper curve shows the net dispersion and the lower curve shows the class dispersion.

have $K = 1$ and no loops, activity propagates unidirectionally through a net (see Weisbuch, 1991, Chapter 2). Thus, these nets provide simple models for dynamic modules of genes with a cascade of gene activations, such as Britten & Davidson (1969) proposed. The genes are not all equivalent, but are allocated to four levels—three families of regulatory genes, and four structural genes. The connectivity of possible nets is limited, because all genes at a level encode messengers in the same family, which can only bind to cis elements of genes at the following level.

Our results show features generic to $K = 1$ nets and features distinctively associated with the partitioning of genes into families and levels, as we now discuss.

FITNESS, FITNESS LANDSCAPES, AND FLOWS ON THEM

In our model the fitness is the reliability with which unreliable regulatory genes activate the structural genes co-ordinately. Thus, the fitness depends directly only on the activation of structural genes. Similarly, Kauffman & Smith (1986) defined fitness only in terms of the states of a subset of genes, regarding the states of the remaining genes as hidden variables. They found little effect of the size of the subset on the evolution of high-fitness nets. Our conclusions and theirs suggest that hidden (regulatory) genes are under as rigorous selection as the directly selected (structural) genes, even though selection on hidden genes is indirect.

As Kauffman (1971, 1989*b*) remarked, $K = 1$ nets have modules, but the connectivity that gives a particular pattern of modules has low stability against point mutations. This low stability was evident in the ruggedness of the net landscape for our model: Each modular net had high fitness but was surrounded by low-fitness neighbors, as was a plateau of high fitness for shotgun nets.

As Kauffman (1989*b*) discussed, a model with more than one input per gene—that is, with redundancy—can give more stable modules. In this regard it is interesting to consider the effect on the net landscape of increasing the number of inputs per gene from one to two, in the set of genes in Fig. 1. Specifically, suppose each gene in levels 3 and 4 has two cis-regulatory elements. As before, each cis element can bind a messenger from the preceding level, and a point mutation replaces one cis element by another in the same family. Suppose also that the Boolean function relating the input of a gene to its output is "or"; that is, if either cis element binds its messenger, or if both do, the gene is active.

In this $K = 2$ model, because genes have redundant inputs, a point mutation tends to reduce the fitness of these nets less than in nets with one input per gene. For example, in the fittest of the 4^6 nets with two inputs, each gene has non-identical cis elements. This is so because such a net can make the tetramer if either messenger from level 3 is present. Any other net can only make the tetramer if both messengers are present, or in response to one of the two messengers. Each gene with non-identical cis elements has two equivalent forms—the cis element which binds one type of messenger followed by the cis element which binds the other type, and the reverse. Consequently, the fittest nets form an equivalence class of cardinality 2^6. A point mutation in any of the fittest nets gives a net with fitness lower than the maximum, but higher than the fitness of a shotgun net. (This assumes that fitnesses are calculated as in Clarke & Mittenthal, 1992.) By contrast, if each gene has only one cis element, mutation of an optimal, shotgun, net (or of a near-optimal, modular, net) gives a pseudomodular net, with a much larger reduction in fitness.

Furthermore, in the $K = 2$ nets, transitions between modular nets can occur directly or through shotgun nets as well as through less fit nets. These additional paths of transition, and the effect of redundancy on mutation, may make the net landscape for $K = 2$ nets less rugged than the landscape for one-input nets.

Our model combines mutation as in a Markov chain with reproduction and attrition. Attrition produces a non-linear, non-Markovian stochastic process. More fit nets reproduce at a greater rate and undergo less attrition. We assume that the population of nets is so large that some nets of every type survive attrition. This assumption preserves the microscopic reversibility of transitions among net types and thus prevents the population from becoming trapped at a local maximum of fitness. Correspondingly, in genetic algorithms, each type of schema in a population of schemata undergoes stochastic reproduction or attrition in proportion to the ratio of its fitness to the average fitness of the population (Goldberg, 1989: 30). By contrast, the alternative procedure of discarding all of the least fit organisms allows a population to be trapped at a local maximum (e.g. Fontana & Schuster, 1987).

Recent models for the flow of populations on fitness landscapes (Kauffman & Levin, 1987; Kauffman, 1989*a*, *b*; Kauffman & Weinberger, 1989) have investigated the characteristics of adaptive walks on fitness landscapes. In an adaptive walk the fitness of the population increases monotonically. The concept of an adaptive walk derives from an analysis by Gillespie (1983) of the increase in fitness associated with mutations at a single locus with multiple alleles. He argues that if selection is strong and the mutation rate is low, a fitter allele rapidly spreads through the population,

but fitter alleles only arise infrequently. From any allele a one-step mutation to a fitter allele can occur, until the fittest allele dominates the population. Thus, a simplified analysis of this process in terms of two time scales is possible in the asymptotic case of strong selection and weak mutation.

We did not assume that the population makes an adaptive walk, because that assumption seemed inappropriate for our model, in which the spectrum of fitnesses is not broad. If the entire population is initially placed at a local maximum, mutations will convert some nets to lower-fitness types. These will proliferate at an appreciable rate if their fitnesses are not much less than the fitness at the nearby maximum. Because there is always a non-zero probability of finding a new peak (if it exists) in finite time, lower-fitness nets may undergo transitions to another local maximum. Thus, the population will flow over the landscape, remaining partially but not wholly concentrated at local maxima, as Figs 4–6 show. Only as the mutation rate approaches zero does the flow approximate an adaptive walk; Fig. 6(d) most nearly represents this case.

FAMILIES AND LEVELS OF GENES; EQUIVALENCE CLASSES

We assumed that if a cis element mutates it still binds a messenger in the same family; one mutation can change its specificity. Empirical evidence supports these assumptions. Cis elements for different families of messengers often differ in the number and identity of bases in the consensus sequence, and in the presence or absence of a palindromic repeat (Harrison, 1991). However, within a family such as the homeodomain family the base sequences of different cis elements often differ in only one or two bases (Laughon, 1992). Thus, differences between families are much greater than differences within families.

Where two or more different mutations can switch the binding specificity of a cis element, the single mutation that we have invoked represents a macromutation subsuming the alternative ways to make the transition. The model is thereby oversimplified in that the macromutation represents micromutations that are not necessarily equivalent for subsequent evolution. One could make a model at the level of micromutations, but it would have many more types of nets and might be computationally intractable.

Level-specific families of messengers suggest two processes that are probably important for evolution. First, transitions within different families are dissociable. This dissociability is likely to stratify the search for maximum fitness and so to speed the population's approach to optimum fitness, as in a search with stratified sampling. A net with dissociable level-specific transitions resembles a simple lock with a sequence of wheels bearing numbers, in which the correct position for each wheel can be determined independently of the positions of the other wheels (Simon, 1962; see also Rasmussen, 1987). (We have not shown that stratification speeds attainment of maximum fitness; this would have required comparing our results to those from a model without level-specific families. It is unclear whether level-specific families of messengers affect the population dynamics of a network in other ways.)

Second, our model assumed that level-specific families of messengers imply the existence of equivalence classes. That is, because alternative cis elements in the same family can occur at each level, there can be alternative nets with equivalent connectivity. Thus, equivalence classes allow one to relate myriad microscopic variables, which describe molecular changes in genes, to a small number of macroscopic variables that can describe transitions among phenotypes during evolution.

We computed the time courses of ensemble-averaged prevalences for the equivalence classes, rather than simulating individual realizations. We examined the effects of four factors on the time courses—the pattern of allowed transitions, the cardinality of the class, its fitness and the mutation rate in the population. In general, the transition from one equivalence class to another could occur in diverse microscopic ways and through diverse intermediate equivalence classes. That is, a polyphyletic origin of a particular equivalence class was likely.

Polyphyletic transitions among taxa of organisms may be common in nature (e.g. Willmer, 1990). Furthermore, because several microscopically distinct nets constitute each equivalence class, descendants of a net might mutate to other equivalence classes and then return to the first class. Such paths of mutation presumably underlie the phenomenon of homoplasy, which is common in some taxonomic groups (Wake, 1991). Thus, we suggest that because equivalence classes exist, polyphyletic origins and homoplasy have been common in evolution.

As a model for macroevolution, our model is unrealistically limited in that we fixed the set of levels of genes and the set of genes at each level. Therefore, the relative proportions of equivalence classes reached a steady state. However, this is unlikely to occur in biological systems for two kinds of reasons. First, the physical and biological environment of a population is unlikely to remain constant over a sufficiently long time. Second, the transitions within families that we modeled occur on the most rapid of three time scales for transitions. On a longer time scale, new messengers will be generated within existing families of genes, by duplication and divergence of existing genes. On a still longer time scale, these processes can be expected to generate new families of messengers. The continuing evolution of novel messengers will prevent convergence to steady-state proportions of equivalence classes. The population will undergo a random walk on a sufficiently high-dimensional lattice that it will not attain a stable pattern of class prevalences in biological time. The occurrence of a quasi-stable state under these circumstances depends on the relative rates of processes at the three time scales, and on the rate at which higher levels of complexity such as multicellularity and social organization evolve. The methods we have used to examine a simple model for evolution in an artificially constrained case can be used to investigate these more complex situations.

APPENDIX

To establish stability for the two models we use Birkhoff's projective metric (see Seneta, 1981: 81) which we denote by d_B. Let $S:\Delta^{64}\to\Delta^{64}$ be defined by either (1) or (4), where Δ^{64} denotes the collection of all probability vectors in 64-dimensional real space.

PROPOSITION

The dynamical system defined by S is asymptotically stable with a unique equilibrium point independent of the initial vector.

PROOF

First note that the set $\{\omega | S\omega = \omega\}$ contains all the limit points of any sequence $\langle S^n v(0) > |_{n=0}^{\infty}$. Indeed, by the Bolzano–Weierstrass theorem, at least one limit point exists, and all limit points are in Δ^{64}. Let $\bar{v}$ be the limit point of the subsequence $\langle S^{n_i} v(0) > |_{i=0}^{\infty}$. Then, for any metric d

$$d(S\bar{v}, \bar{v}) = d\left(\lim_{i\to\infty} S^{n_i} v(0), \bar{v}\right) = \lim_{i\to\infty} d(S^{n_i} v(0), \bar{v}) = 0. \tag{A.1}$$

So $\bar{v} \in \{\omega | S\omega = \omega\}$.

Now it is enough to show that there is exactly one limit point. Birkhoff's projective metric is defined for vectors x, y with strictly positive entries x_i and y_i to be

$$d_B(x, y) = \max_{i,j} \log \frac{x_i y_j}{x_j y_i}. \tag{A.2}$$

On the interior of Δ^{64}, d_B has all the properties of a metric and is contractive (see Seneta, 1981: 90, ex. 3.1 and p. 83, Lemma 3.2) for matrices with all entries strictly positive.

First, suppose S is as in (1). It is straightforward to see that for $p \in (0, 1)$, T^k has all entries strictly positive for $k \geq 6$. Consequently, $(\mathbf{FT})^k$ has all entries strictly positive for $k \geq 6$. As a result, all members of $\{\omega | S\omega = \omega\}$ are interior points of Δ^{64}.

Let $v(0) \in \Delta^{64}$ be any initial vector. Then the Banach fixed point theorem implies that the sequence

$$\left\langle \frac{(\mathbf{FT}^{k(i+1)} v(0)}{I'(\mathbf{FT})^{k(i+1)} v(0)} \right\rangle \Bigg|_{i=0}^{\infty}$$

has a unique fixed point $\bar{v}$ independent of $v(0)$ in d_B. Note that normalizing constants do not matter, since they cancel out in the argument of the logarithm in (A.2). Similarly, the sequence

$$\left\langle \frac{(\mathbf{FT}^{k(i+1)+1} v(0)}{I'(\mathbf{FT})^{k(i+1)+1} v(0)} \right\rangle \Bigg|_{i=0}^{\infty}$$

has a unique fixed point $\bar{u}$ in Δ^{64} independent of $v(0)$. By the same reasoning as in (A.1), we see that

$$\begin{aligned} d_B(\bar{u}, \bar{v}) &= \lim_{i\to\infty} d_B(S^{k(i+1)+1} v(0), S^{k(i+1)} v(0)) \\ &\leq \lim_{i\to\infty} \alpha^i d_B(S^{k+1} v(0), 2^k v(0)), \end{aligned}$$

where $\alpha \in (0, 1)$. So $d_B(\bar{u}, \bar{v}) = 0$, i.e. $\bar{u} = \bar{v}$.

Now we can conclude that $\{\omega | S\omega = \omega\}$ is a singleton set, and that point is the unique stable equilibrium for S. If S is as in (4), the argument is similar. □

The above proposition obtains the uniqueness of the limit by using a fixed point theorem. While the Perron–Frobenius theorem applies and gives the existence of limiting distributions (see Seneta, 1981, Chapter 1), they need not be unique in general.

We thank Nigel Goldenfeld, who pointed out the correspondence between temperature and mutation rate, and Yoshi Oono, who suggested the soma-germ model and commented on the manuscript. Thanks also to Steve Lalley for directing us to Birkhoff's projective metric, and to Kim Brady for help with the graphics. We appreciate extensive comments by a referee.

REFERENCES

BRITTEN, R. J. & DAVIDSON, E. H. (1969). Gene regulation for higher cells: a theory. *Science* **165,** 349–357.

BROOKS, D. R. & WILEY, E. D. (1988). *Evolution as Entropy Toward a Unified Theory of Biology,* 2nd edn. Chicago, IL: University of Chicago Press.

CLARKE, B. & MITTENTHAL, J. E. (1992). Modularity and reliability in the organization of organisms. *Bull. math. Biol.* **54,** 1–20.

EWENS, W. J. (1979). *Mathematical Population Genetics.* Berlin: Springer-Verlag.

FELDMAN, M. W. (1989). Dynamical systems from evolutionary population genetics. In: *Lectures in the Sciences of Complexity. SFI Studies in the Sciences of Complexity* (Stein, D., ed.) pp. 501–526. Redwood City, CA: Addison-Wesley Longman.

FONTANA, W. & SCHUSTER, P. (1987). A computer model of evolutionary optimization. *Biophys. Chem.* **26,** 123–147.

GILLESPIE, J. H. (1983). A simple stochastic gene substitution model. *Theor. Pop. Biol.* **23,** 202–215.

GOLDBERG, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Reading, MA: Addison-Wesley.

GOULD, S. J. (1989). *Wonderful Life.* New York: W. W. Norton & Co.

HARRISON, S. C. (1991). A structural taxonomy of DNA-binding domains. *Nature, Lond.* **353,** 715–719.

HEDRICK, P. W. & MACDONALD, J. F. (1980). Regulatory gene adaptation: an evolutionary model. *Heredity* **45,** 83–97.

KAUFFMAN, S. A. (1971). Gene regulation networks: a theory for their global structure and behavior. *Curr. Top. develop. Biol.* **6,** 145–182.

KAUFFMAN, S. A. (1989*a*). Adaptation on rugged fitness landscapes. In: *Lectures in the Sciences of Complexity. SFI Studies in the Sciences of Complexity* (Stein, D., ed.) pp. 527–618. Addison-Wesley Longman.

KAUFFMAN, S. A. (1989*b*). Principles of adaptation in complex systems. In: *Lectures in the Sciences of Complexity. SFI Studies in the Sciences of Complexity* (Stein, D., ed.) pp. 619–712. Addison-Wesley Longman.

KAUFFMAN, S. A. (1989*c*). Origins of order in evolution: self-organization and selection. In: *Theoretical Biology. Epigenetic and Evolutionary Order from Complex Systems* (Goodwin, B. & Saunders, P., eds) pp. 67–88. Edinburgh: Edinburgh University Press.

KAUFFMAN, S. A. (1991). Antichaos and adaptation. *Sci. Am.* **265(2),** 78–84.

KAUFFMAN, S. A. (1993). Requirements for evolvability in complex systems: orderly dynamics and frozen components. In: *Thinking About Biology. SFI Studies in the Sciences of Complexity, Lecture Notes Volume III* (Stein, W. & Varela, F. J., eds). Reading, MA: Addison-Wesley.

KAUFFMAN, S. A. & LEVIN, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. theor. Biol.* **128,** 11–45.

KAUFFMAN, S. A. & SMITH, R. G. (1986). Adaptive automata based on Darwinian selection. *Physica D* **22,** 68–82.

KAUFFMAN, S. A. & WEINBERGER, E. D. (1989). The *NK* model of rugged fitness landscapes and its application to maturation of the immune response. *J. theor. Biol.* **141,** 211–245.

KING, M. C. & WILSON, A. C. (1975). Evolution at two levels: molecular similarities and biological differences between humans and chimpanzees. *Science* **188,** 107–116.

LANGTON, C. G., TAYLOR, C., FARMER, J. D. & RASMUSSEN, S. (eds). (1992). *Artificial Life II.* Redwood City, CA: Addison-Wesley.

LAUGHON, A. (1992). The DNA binding specificity of homeodomains. *Biochemistry*, **30,** 11357–11367.

LEVINTHAL, M. (1990). The evolution of complexity in metabolism. In: *Biosynthesis of Branched Chained Amino Acids* (Barak, Z., Chipman, D. M. & Schloss, J. V., eds) pp. 163–178. New York: Balaban Publishers.

MAY, R. M. (ed.). (1981). *Theoretical Ecology*, 2nd edn. Oxford: Blackwell Scientific Publications.

MITTENTHAL, J. E., BASKIN, A. B. & REINKE, R. (1992). Patterns of structure and their evolution in the organization of organisms: molecules, matching, and compaction. In: *Principles of Organization in Organisms. SFI Studies in the Sciences of Complexity, Proc. Vol. XIII* (Mittenthal, J. E. & Baskin, A. B., eds). Reading, MA: Addison-Wesley.

RASMUSSEN, N. (1987). A new model of developmental constraints as applied to the Drosophila system. *J. theor. Biol.* **127,** 271–301.

SENETA, E. (1981). *Non-negative Matrices and Markov Chains*. New York: Springer-Verlag.

SIMON, H. (1962). The architecture of complexity. *Proc. Am. Phil. Soc.* **106,** 467–482.

WAKE, D. B. (1991). Homoplasy: the result of natural selection, or evidence of design limitations? *Am. Natur.* **138(3),** 543–567.

WEISBUCH, G. (1986). Networks of automata and biological organization. *J. theor. Biol.* **121,** 255–267.

WEISBUCH, G. (1991). *Complex Systems Dynamics*. Redwood City, CA: Addison-Wesley.

WILLMER, P. (1990). *Invertebrate Relationships*. Cambridge: Cambridge University Press.

WILSON, A. C. (1975). Evolutionary importance of gene regulation. *Stadler Symp.* **7,** 117–133.

YANOFSKY, M. F., MA, H., BOWMAN, J. L., DREWS, G. N., FELDMANN, K. A. & MEYEROWITZ, E. M. (1990). The protein encoded by the *Arabidopsis* homeotic gene agamous resembles transcription factors. *Nature, Lond.* **346,** 35–39.