

Evolutionary Consequences of Selected Locus-Specific Variations in Epistasis and Fitness Contribution in Kauffman's NK Model

Daniel Solow, Apostolos Burnetas, Theresa Roeder, and Neil S. Greenspan*

Department of Operations Research and Operations Management

*Institute of Pathology

Case Western Reserve University

Cleveland, OH 44106

e-mail: dxs8@po.cwru.edu

March 7, 2001

Abstract

Mathematical analysis and computer simulations are used to evaluate three modifications to Kauffman's NK model in an attempt to incorporate unexplored aspects of epistatic interaction between loci in genome evolution. Two modifications—one to the amount and the other to the distribution of epistatic interaction—further support Kauffman's conclusion that high levels of epistatic interaction lead to a decrease in overall fitness of the genome. The third model, however, provides a condition under which a high amount of epistatic interaction at certain loci results in higher genome fitness.

1 Introduction

Numerous mathematical models have been suggested for studying genome evolution [see, for example, the model in Macken & Perelson (1989), the “multiplicative” models of Franklin & Lewontin (1970); Lewontin (1974); and Ewens (1979)]. One class of such models is epitomized by the NK model of Kauffman and Levin (1987) and Kauffman (1993). In the

NK model, a haploid genome is represented as an ordered list of N loci, each of which can have any number of alleles. To control the amount of epistatic interaction at each locus i , Kauffman introduces an integer parameter, K , in which $K = 0$ indicates no interaction and $K = N - 1$ indicates interaction with all remaining $N - 1$ loci.

Kauffman's *NK* models of genome evolution rely on an intentionally-simplified representation of haploid genomes in which there are, in most cases, two alleles at each locus, additive contributions to fitness from different loci, and constant extents of epistatic interaction at different loci. Real genomes, however, contain loci with wide variation in the number of alleles, varying and non-additive net contributions to fitness, and disparate numbers and magnitudes of epistatic interactions. For example: (1) there is limited nucleotide sequence variation at the β 2-microglobulin locus [Hansen et al. (1993)], while at the major histocompatibility complex loci, such as HLA-B, there are over one hundred alleles [Bodmer et al. (1995)], (2) some loci when subjected to homozygous deletion, such as the epidermal growth factor receptor, are lethal [Threadgill et al. (1995); Sibia and Wagner (1995)], whereas others, such as the interleukin-2 locus, are not [Schorle et al. (1991)], and (3) the sickle and wild-type alleles at the β -globin locus would be expected to have different numbers of epistatic interactions involving different mechanisms of epistasis based on known properties of the respective gene products [Allison (1954); Stryer (1988); Greenspan (1998)].

Modifications to the basic *NK* model have been developed in an attempt to model epistatic interactions in more realistic ways. For example, in the "block" model presented in Perelson and Macken (1995), the N loci are partitioned into B blocks. It is assumed that the loci within each block interact epistatically with all other loci within that block but that there is no epistatic interaction between blocks. In this model, the amount of epistatic interaction at a locus i depends on the number of loci in the block to which locus i belongs.

In an effort to study other patterns of epistatic interaction, three modifications to the *NK* model are proposed in this paper. The ones described in Sections 2 and 3 support Kauffman's conclusions in that in these modified models, even few loci with high epistatic interaction are detrimental to fitness of the genome. In contrast, the model genomes in Section 4 benefit from high epistatic interaction at certain loci.

In the remainder of this section, various applications of the *NK* model are described

and a detailed description of the NK model is presented.

1.1 Applications of the NK Model

The NK model has been applied in many other biological settings involving the evolution of, for example: antibody variable domain amino acid sequences in the humoral immune response [Kauffman and Levin (1987); Kauffman et al. (1988); Macken and Perelson (1989); Macken et al. (1991); Kauffman (1993); Perelson and Macken (1995)], protein or RNA sequences or conformations [Weinberger (1988); Amitrano et al. (1989); Flyvbjerg and Lautrup (1992); Fontana et al. (1993); Schuster and Stadler (1994)] and molecular quasi-species [Eigen et al. (1989)].

The NK model has also found applications outside the field of biology. For example, Levinthal (1997) uses the NK paradigm in a business environment to model the process of organizational change. In that article, an organization is represented by a vector of N binary attributes. The contribution of each attribute to the overall fitness of the organization is influenced by the values of K other attributes. The author examines the effect of the value of K on the number of local peaks of the fitness landscape, each of which is associated with a dominant organizational form.

NK models borrowed ideas and notation from the field of spin glasses [Derrida (1981)] in physics. Spin glasses models can be viewed in the context of fitness landscapes [Weinberger (1991)]. In this setting, the spins (up or down) of N atoms are affected by the spins of K other surrounding atoms. Evolution in this setting involves variation in the spin of each atom so as to minimize the total energy of the collection of atoms.

1.2 A Description of the NK Model

The simplest version of the NK model has two alleles at each locus. In this case, a genome is represented mathematically as a binary N -vector, $\mathbf{x} = (x_1, \dots, x_N)$, in which $x_i = 1$ means that one of the two alleles is present at locus i and $x_i = 0$ means that the other allele is present at that locus. Geometrically, each of the 2^N binary N -vectors is a corner point of the N -dimensional unit cube.

Associated with each of the 2^N genomes, \mathbf{x} , is a fitness in the form of a real number,

$f(\mathbf{x})$, between 0 and 1. A value close to 0 indicates poor fitness and a value close to 1 indicates good fitness. In Kauffman’s model, the fitness, $f(\mathbf{x})$, is the average of the fitness contributions, $f_i(\mathbf{x})$, from each locus i , that is:

$$f(\mathbf{x}) = \frac{\sum_{i=1}^N f_i(\mathbf{x})}{N} \quad (1)$$

The specific way the contribution to fitness of locus i is calculated is based on the epistatic interactions affecting that locus, as indicated by the value of the integer parameter K . Kauffman (1993) assumes that the contribution of each locus i to the overall fitness of the genome depends on the allele at locus i as well as on the alleles at K other loci (for example, the $K/2$ loci on either side of locus i , wrapping around, if necessary). There are 2^{K+1} possible combinations for the alleles at these $K + 1$ loci, so there are 2^{K+1} possible fitness contributions for each locus. For simulation purposes, 2^{K+1} uniform 0 – 1 random numbers are generated for each locus. Then, the value of $f_i(\mathbf{x})$ is the random number that corresponds to the combination of alleles at locus i and the K loci that affect locus i .

Given values for N, K , and the N tables of 2^{K+1} uniform 0 – 1 random numbers, the collection of all 2^N binary vectors, together with their fitnesses, as defined by (1), constitute the *NK model*. Evolution in this model involves moving from one genome to another in search of a genome having the best possible fitness. There are many ways to do so (for example, gradient descent, recombination, genetic algorithms, and so on). In this paper, evolution is modeled by considering a *one-mutant neighbor*, hereafter called a “neighbor”, of a genome. A *neighbor* of a genome \mathbf{x} is a genome \mathbf{y} in which the allele at exactly one locus i of \mathbf{y} is different from the allele at locus i of \mathbf{x} , all other alleles being the same. Note that the fitness of a neighbor \mathbf{y} may or may not be better than the fitness of \mathbf{x} . The *NK* model together with this one-mutant neighborhood structure of a genome constitutes the *NK fitness landscape*.

Evolution is now assumed to proceed by an *adaptive walk*, as follows. Starting with an arbitrary genome, a sequence of neighboring genomes with successively better fitnesses are visited until obtaining a genome whose fitness is better than the fitness of all its neighbors. This final genome is referred to as a *local maximum*. Modeling evolution in this way affords the ability to obtain analytical results in certain cases that can then be supported by

simulations. For example, one question Kauffman set out to answer in this framework was how the values of N and K affect the average fitness of the final genome obtained from these adaptive walks.

Kauffman (1993) provides a mathematical analysis for the case $K = 0$, referred to as a highly *correlated* landscape. In this case, starting with any initial genome, an adaptive walk reaches the best possible genome, whose expected fitness is shown mathematically to be $2/3$. When $K = 0$, the contribution of each locus to the overall fitness depends only on the allele at that locus. A genome with maximum fitness is obtained by an adaptive walk that successively sets the allele at each locus to its best value, 0 or 1, as determined by the random-number table for that locus. The expected number of steps needed to do so is $N/2$.

Kauffman (1993) also provides an analysis for the case $K = N - 1$, referred to as a highly *uncorrelated* landscape. In this case, changing the allele at one locus changes the fitness contributions of all loci. Kauffman argues analytically that when N tends toward infinity, an adaptive walk results in a local maximum whose expected fitness approaches $1/2$. It is also shown that the expected number of local maxima is $2^N/(N + 1)$ and that the expected number of steps needed to reach a local maximum is $\log(N - 1)$.

From the analysis for these two extreme values of K , one can identify what is referred to in this paper as the *path-length phenomenon*, namely, that the expected number of steps needed to reach a local maxima is inversely correlated with the number of local maxima—the more the expected number of steps, the fewer the expected number of local maxima.

One might also infer that as the expected number of steps needed to reach a local maximum decreases, the expected fitness decreases. This, however, turns out to be true only for a certain range of values for K . Specifically, Kauffman uses computer simulations to determine the average fitness of a local maximum for different values of K . Those results indicate that, as N gets large, for small positive values of K , the expected fitness exceeds the fitness of $2/3$ associated with $K = 0$. But then, as K increases, the expected fitness of the local maximum decreases toward $1/2$. Kauffman refers to this phenomenon—of decreasing fitness associated with increasing epistatic interaction—as the *complexity catastrophe*.

Intuitively, the foregoing results are due to a trade-off that arises as K increases. The larger the value of K , the greater the number of possible values for the fitness contributions

of each locus, thus resulting in more allelic combinations that could result in fitness contributions close to 1. However, as K increases, there are more conflicts between the loci in that whatever allele is chosen at locus i , that allele is likely to benefit the fitness contributions of some loci while being detrimental to the fitness contributions of other loci. The simulation results obtained from the NK model indicate that for small positive values of K , the benefits of having more choices for the fitness contributions of the individual loci outweigh the few conflicts. However, as K increases, the negative effects of the increasing number of conflicts dominate the benefits of having more choices for the fitness contributions of the individual loci, thus resulting in the complexity catastrophe.

2 Modifications to the Amount of Epistatic Interaction

The first modification of the NK model is designed to test the sensitivity to changes in the *amount* of epistatic interaction at each locus. In particular, in the NK model, it is assumed that the contribution to fitness of each locus i [namely, $f_i(\mathbf{x})$] depends on the alleles at locus i and at K other loci. What would happen if $f_i(\mathbf{x})$ depends on the allele at locus i and on the alleles at K_i other loci? In other words, what happens if the amount of epistatic interaction varies from one locus to the next?

To study this question in such a way as to compare the results with the NK model, consider fixed values of N and K and define the total amount of epistatic interaction as $N * K$. This total is then allocated across the N loci in a systematic way, as follows. To test the extreme case, all the epistatic interaction is concentrated on as few loci as possible. (Those loci for which $K_i > 0$ are called *affected loci* and those for which $K_i = 0$ are called *unaffected loci*.) In subsequent simulations, these $N * K$ epistatic interactions are distributed systematically to more and more of the N loci until eventually, each of the N loci is affected by exactly K other loci, thus resulting in the original NK model. (In this sense, the proposed NK_i model is a generalization of the NK model.)

To be specific, consider the case when $N = 16$ and $K = 4$, so the total amount of epistatic interaction is $N * K = 64$. To concentrate these 64 epistatic interactions on as few affected loci as possible, observe that the epistatic interaction at any one locus can be at most $N - 1 = 15$. Thus, $K = 4$ affected loci can each have $K_i = 15$, so the contributions to

fitness of these loci depend on their own allele and those at all other loci. To distribute the 4 leftover epistatic interactions, two additional affected loci, each with $K_i = 2$, are needed. All remaining $N - K - 2 = 10$ loci each have $K_i = 0$, so their contributions to fitness depend only on their own alleles. In subsequent simulations for $N = 16$ and $K = 4$, the 64 epistatic interactions are distributed to more and more loci, as shown, for example, in the following table:

Simulation	Amount of Epistatic Interaction at Locus															
Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	0	0	0	0	2	15	15	15	15	2	0	0	0	0	0
2	0	0	0	0	8	8	8	8	8	8	8	8	0	0	0	0
3	0	0	5	5	5	5	5	7	7	5	5	5	5	5	0	0
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

The loci that affect the fitness contributions of those loci where $K_i > 0$ are chosen at random from the N loci, so the specific locations of the loci where $K_i > 0$ are irrelevant. In particular, the fact that these loci are centrally located does not affect the results of the subsequent simulations. (A partial simulation indicated that similar results are obtained when the loci that affect locus i are chosen as the $K_i/2$ loci on either side.)

These groups of simulations were performed for NK models with $N = 8, 16, 24, 48, 96$ and $K = 2, 4, 8, 16, 24, 48$, so long as $K < N$. Note that when $K = N - 1$, each of the N loci must have $K_i = N - 1$. Because this is the only way to distribute the total epistatic interaction in this case, these simulations are omitted. For each remaining combination of N and K , 500 different landscapes with one adaptive walk on each were simulated, except for the case $N = 96$, in which 200 trials were performed.

The average fitness and average number of steps needed to reach a local maximum for $N = 48$ and $N = 96$ are presented in Figures 1 and 2, respectively. The last point on each curve coincides with the original NK model. The results are similar for the other values of N and K and confirm Kauffman's observations on the complexity catastrophe. Furthermore, they indicate that for virtually all values of N and K , the expected fitness of a local maximum increases (almost linearly) with the number of loci for which $K_i > 0$. That is, the more uniform the distribution of epistatic interaction across loci, the better

the expected fitness of a local maximum.

Although no analytical explanation was found for this outcome, further statistics for the case when $N = 48$ were collected. For each value of K , the simulations revealed that the average contribution to fitness of each affected locus exceeded substantially the average contribution of each unaffected locus, regardless of how many affected loci there are. This means that, in the NK_i model, high epistatic interaction at the affected loci is more beneficial to the fitness contribution of those loci than the lack of epistatic interaction at the unaffected loci. One plausible explanation for this phenomenon is that the contribution to fitness of each unaffected locus depends only on its own allele. However, an adaptive walk evidently chooses the alleles at most loci to benefit the fitness contributions of the affected loci. In any event, because the average contribution to fitness of affected loci is greater than that of unaffected loci, it follows that as the number of affected loci increases—that is, as the distribution of epistatic interaction becomes more uniform—the fitness of the genome increases. This explanation is consistent with the trade-off between the number of choices and the number of conflicts associated with increasing K , as discussed in Section 1.

It was also found that, for a fixed number of affected loci, as K increases, the average contributions of each affected locus and each unaffected locus both decrease, thus confirming that higher epistatic interaction leads to a decrease in average fitness of a local maximum.

Turning to the number of steps needed to reach a local maximum, one might anticipate that, for a fixed value of K , this number increases as the distribution of epistatic interaction across loci increases. This is because, in general, better fitness occurs when there are fewer local maxima. Assuming that the path-length phenomenon is applicable, when there are fewer local maxima, more steps are needed to reach a local maximum. This correlation of an increase in the number of steps needed to reach a local maximum with an increase in the distribution of epistatic interaction across loci is evident from the simulation results in Figures 1(b) and 2(b).

3 Modifications to the Distribution of Epistatic Interaction

The next modification to the NK model arises by asking the following question. Given that the contribution to fitness of locus i depends on its allele and the alleles at K other

loci, does it matter *which* K loci affect locus i ? Kauffman (1993) studied three forms of this epistatic *distribution*, one of which is where the contribution to fitness of each locus is affected by its own allele and those at the $K/2$ loci on either side of locus i (wrapping around, if necessary). The second is a model in which each locus is affected by its own allele and those at K other randomly-chosen loci. (Kauffman found no significant difference in the final fitness of a local maximum obtained by an adaptive walk with these two variations.) The third is a model in which the epistatic distribution is “falling off exponentially from the first site to the N th site, such that the weighted proportion of epistatic outputs from each site i is $e^{-0.1i}$ ” [c.f. Kauffman (1993)].

In this section, a more systematic study of the effects of the distribution of epistatic interaction is conducted. To compare the results with the NK model, consider fixed values of N and K , so the contribution to fitness of each locus i is affected by its own allele and the alleles at K other loci. To control which loci affect locus i , choose $A \geq K$ *effector loci*. Then, the K loci that affect any locus i are chosen among these A effector loci, for example, $K/2$ of these effector loci on either side of locus i .

In the extreme case, when $K = A = 1$, there is one effector locus—say, locus 1. In this case, the contribution to fitness of each locus i depends on the alleles at locus i and at locus 1. Thus, the allele at locus 1 affects the contribution to fitness of all other loci and hence, locus 1 has a high level of epistatic interaction. At the other extreme, when $A = N$, a version of the NK model results. In this sense, the NK/A model is a generalization of the NK model.

One interesting observation is that there are at most 2^A local maxima in the NK/A model. To see why this is so, consider again the case when $K = A = 1$, with locus 1 being the effector locus. Temporarily fixing the allele at locus 1, the fitness contribution of any other locus $i \geq 2$ now depends only on its own allele. Thus, a necessary condition for a local maximum is that the allele at each locus $i \geq 2$ have its best value, 0 or 1, depending on the random-number table for that locus and on the fixed allele of the effector locus 1. The resulting genome is a local maximum if changing the allele at the effector locus 1 does not lead to further improvement. In any case, for each of the two alleles in the effector locus 1, there is at most one local maximum. Thus, in the case $A = 1$, there are at most $2^A = 2$

local maxima.

More generally, if the fitnesses of all genomes are distinct—which is the case when the random number generator for the fitnesses is unbiased—then, for an arbitrary value of A , there are at most 2^A different local maxima in the NK/A model. To see that this is so, suppose there were more than 2^A different local maxima. Because there are only 2^A combinations of alleles at the A affector loci, at least two different local maxima, say, $\mathbf{x} \neq \mathbf{y}$, must have the same combination of alleles at the A affector loci. It is now shown that these assumptions lead to a contradiction. In particular, it has already been stated that the alleles at the A affector loci of \mathbf{x} and \mathbf{y} are the same. It will now be shown that the alleles of the remaining $N - A$ loci of \mathbf{x} and \mathbf{y} are also the same, which contradicts the fact that $\mathbf{x} \neq \mathbf{y}$. To see that the alleles at the remaining loci are the same, as noted in the case $A = 1$, for fixed values of the A affector loci, the contribution of each of the remaining $N - A$ loci depend only on their own alleles. Thus, at a local maximum, there is a unique setting of alleles at the $N - A$ affector loci that achieves the maximum contribution to fitness of these loci (assuming distinct fitness values for each genome). Hence, the alleles at the remaining $N - A$ loci of both local maxima \mathbf{x} and \mathbf{y} are also the same. This contradiction establishes that there are at most 2^A local maxima in the NK/A model.

The simulations were first performed for $N = 48$ and $K = 2, 4, 8, 16, 24$. For each value of K , the number of affector loci, A , is set initially to $A = K$. During each successive run, the value of A is increased to the next multiple of 5, so long as $A < 48$, with the final run having $A = 48$. For each set of K and A values, 500 simulations were performed in which loci $1, \lfloor N/A \rfloor + 1, \lfloor 2N/A \rfloor + 1, \dots, \lfloor (A - 1)N/A \rfloor + 1$ are the A affector loci. Each locus is then affected by the $K/2$ affector loci on either side of locus i . A similar approach, involving 200 simulations each, was used for $N = 96$ and $K = 2, 4, 8, 16, 24, 48$, with the only difference being that the value of A is increased to the next multiple of 10, so long as $A < 96$ and the final run having $A = 96$.

To illustrate, for $N = 48$ and $K = 4$, simulations were run for which there were $A = 4, 5, 10, 15, 20, 25, 30, 35, 40, 45$, and 48 affector loci. When $A = 10$, loci $1, 5, 9, \dots, 37$ are the 10 affector loci. Thus, locus 2, for example, is affected by loci 1 and 37 on the left and by loci 5 and 9 on the right.

The average fitness and average number of steps needed to reach a local maximum are reported in Figures 3 and 4. The results from these simulations indicate that for these values of N , higher average fitnesses of local maxima are associated with higher values of A , when K is small. That is, when K is small, the greater the number of affector loci, the better the average fitness obtained by an adaptive walk. Alternatively stated, high epistatic interaction, in the form of few affector loci that affect many loci, results in decreased average fitness. However, as K increases, the increase in average fitness associated with increasing the number of affector loci starts to level off and even decreases slightly for large K .

From the results in Figures 3(b) and 4(b), for a fixed value of K , the number of steps needed to reach a local maximum decreases as the number of affector loci increases. Furthermore, the rate of decrease in the number of steps appears to be steeper for larger values of K .

4 Modifications to the Contributions of Individual Loci

The models described in Sections 2 and 3 are based on the likelihood that, in real biological settings, the amount of epistatic interaction varies as a function of locus. In this section, another modification to the NK model is proposed in an attempt to differentiate the impact of each locus and its epistatic interactions on the overall fitness of the genome. This modification is obtained by writing the fitness, $f(\mathbf{x})$, as defined in (1), in the following form:

$$f(\mathbf{x}) = \frac{f_1(\mathbf{x}) + \cdots + f_N(\mathbf{x})}{N} = \frac{1}{N}f_1(\mathbf{x}) + \cdots + \frac{1}{N}f_N(\mathbf{x}) \quad (2)$$

In (2), the contribution to fitness, $f_i(\mathbf{x})$, of each locus i is weighted by the same amount, namely, $1/N$. The model suggested now is one in which each $f_i(\mathbf{x})$ is weighted by a number w_i , where $0 < w_i < 1$ and $w_1 + \cdots + w_N = 1$. That is,

$$f(\mathbf{x}) = w_1f_1(\mathbf{x}) + \cdots + w_Nf_N(\mathbf{x}) \quad (3)$$

Observe that when each $w_i = 1/N$, (3) reduces to (2). In this sense, the NK/W model is a generalization of the NK model.

The analysis in Kauffman (1993) for the NK model when $K = 0$ applies to the NK/W model also. Specifically, when $K = 0$, the value of $f_i(\mathbf{x})$ depends only on the allele at locus

i. A genome \mathbf{x}^* with maximum fitness is obtained by an adaptive walk that successively sets the allele at each locus to its best value, 0 or 1, as determined by the random-number table for that locus. Thus, the expected value of $f_i(\mathbf{x}^*)$ is the maximum of two independent, identically distributed (i.i.d.) uniform 0 – 1 random variables, which is $2/3$. It then follows from (3) that the expected value of $f(\mathbf{x}^*)$ is

$$E[f(\mathbf{x}^*)] = E\left[\sum_{i=1}^N w_i f_i(\mathbf{x}^*)\right] = \sum_{i=1}^N w_i E[f_i(\mathbf{x}^*)] = \sum_{i=1}^N w_i \left(\frac{2}{3}\right) = \frac{2}{3} \sum_{i=1}^N w_i = \frac{2}{3}$$

In contrast, the analysis in Kauffman (1993) for the case $K = N - 1$ that leads to the complexity catastrophe is not applicable in the NK/W model. This is because, in the NK model with $K = N - 1$, a local maximum is obtained by an adaptive walk on an N -dimensional unit cube in which the fitness of each genome \mathbf{x} is the sum of the following N i.i.d. random variables, each of which has a uniform distribution between 0 and $1/N$:

$$\frac{f_1(\mathbf{x})}{N}, \frac{f_2(\mathbf{x})}{N}, \dots, \frac{f_N(\mathbf{x})}{N}$$

For large N , the central limit theorem implies that the distribution of $f(\mathbf{x})$ is approximately normal with mean $1/2$ and variance $1/(12N)$, which converges to 0 as N approaches infinity. Thus, when N is large, the fitness of any genome is close to $1/2$ with high enough probability to justify the fact that even the fitness of a local maximum is approximately $1/2$.

For the NK/W model with $K = N - 1$, although the expected value of $f(\mathbf{x})$ is still equal to $1/2$, the variance of $f(\mathbf{x})$ is

$$\text{Var}(f(\mathbf{x})) = \sum_{i=1}^N \frac{w_i^2}{12}$$

This variance might not approach 0 as N approaches infinity, in which case the expected fitness of a local maximum is no longer $1/2$. In other words, in the NK/W model, high epistatic interaction may no longer lead to the complexity catastrophe. One such model that actually benefits from high levels of epistasis at certain loci is presented next.

4.1 The Case of One Heavily-Weighted Locus

Consider now a special case of the NK/W model in which, say, locus 1 is heavily weighted relative to the weights of the other loci. For example, in the extreme case when $w_1 = 1$

and $w_2 = \dots = w_N = 0$, the fitness of the genome, as defined by (3), depends only on the fitness contribution of locus 1, that is,

$$f(\mathbf{x}) = f_1(\mathbf{x})$$

In this case, at a local maximum, the alleles of the loci that affect locus 1 will be set to achieve a high fitness contribution for locus 1, without concern for how those settings affect the contribution of other loci. Thus, when K is large, the contribution of locus 1, and of the whole genome, will be close to 1. In other words, in this special case, high epistatic interaction at locus 1 is beneficial.

A similar conclusion holds when the weight w of locus 1 is relatively high, with the remaining weight, $1 - w$, distributed, say, evenly among the rest of the loci, so $w_1 = w$ and $w_j = (1 - w)/(N - 1)$, for $j = 2, \dots, N$. To see that the fitness of the final genome obtained by an adaptive walk in this model is largely determined by the contribution of locus 1, consider a genome \mathbf{x} and one neighbor of \mathbf{x} , say, \mathbf{y} , for which $f_1(\mathbf{y}) < f_1(\mathbf{x})$. If w is sufficiently close to 1, an adaptive walk starting from \mathbf{x} will not move to \mathbf{y} , even if the contributions to fitness of the remaining loci of \mathbf{y} are all 1 because this is not sufficient to compensate for the fitness decrease of the heavily-weighted locus 1. Specifically, the difference in total fitness between \mathbf{x} and \mathbf{y} is:

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= \sum_{i=1}^N w_i (f_i(\mathbf{y}) - f_i(\mathbf{x})) \\ &= w(f_1(\mathbf{y}) - f_1(\mathbf{x})) + \frac{1-w}{N-1} \sum_{i=2}^N (f_i(\mathbf{y}) - f_i(\mathbf{x})) \\ &\cdot w(f_1(\mathbf{y}) - f_1(\mathbf{x})) + \frac{1-w}{N-1} \sum_{i=2}^N (1 - 0) \\ &= w(f_1(\mathbf{y}) - f_1(\mathbf{x})) + 1 - w \end{aligned}$$

The expression on the right side of the final inequality above is negative when w is sufficiently close to 1.

This discussion suggests that it is reasonable to approximate an NK/W model in which only locus 1 is heavily weighted by a model where the $K+1$ loci that affect locus 1 (including locus 1 itself) follow an adaptive walk to maximize solely $f_1(\mathbf{x})$ (without concern for how the alleles in those $K+1$ loci affect the contributions of other loci) and the alleles at the

remaining $N - K - 1$ loci are set randomly. Let \tilde{f}_1 be the expected value of the contribution to fitness of locus 1 at the local maximum \mathbf{x} reached by an adaptive walk that maximizes the contribution of locus 1. Then the expected value of the local maximum of the NK/W model with one heavily-weighted locus is approximately

$$\begin{aligned} E[f(\mathbf{x})] &= w_1 E[f_1(\mathbf{x})] + w_2 E[f_2(\mathbf{x})] + \cdots + w_n E[f_n(\mathbf{x})] \\ &\approx w \tilde{f}_1 + \frac{1}{2}(w_2 + \cdots + w_n) \\ &= w \tilde{f}_1 + \frac{1-w}{2} \end{aligned} \quad (4)$$

The value \tilde{f}_1 is the expected value of an adaptive walk on a $(K + 1)$ -dimensional unit cube in which the fitnesses of the 2^{K+1} corner points are i.i.d. uniform random variables. A system of this type has been studied by Macken and Perelson (1989). Using recursive analysis with an approximating Markov process, they show that the expected value of the local maximum reached by such an adaptive walk approaches the ideal value of 1. In particular, for K large, their results imply that

$$E[\tilde{f}_1] \approx 1 - \frac{0.63}{K}. \quad (5)$$

Substituting (5) in (4), an approximation for the expected value of a local maximum \mathbf{x} in the current NK/W model, when w is close to 1 and K is large, is

$$E[f(\mathbf{x})] \approx w \left(1 - \frac{0.63}{K}\right) + \frac{1-w}{2} \quad (6)$$

The expression on the right side of (6) is increasing in K , which means that *higher epistatic interaction leads to higher expected fitness, provided that the weight of locus 1 is sufficiently close to 1*. In other words, this model achieves the advantages of having a large number of choices for the fitness contribution of the heavily-weighted locus without suffering from the disadvantage of epistatic conflicts.

A simulation is presented to show the effects of heavily weighting one locus. Specifically, the weight of one chosen locus is controlled and the effects of varying N and K are studied. Equivalently stated, for fixed values of N and K , the effects on overall fitness of using various weights for one chosen locus are obtained. The simulations for this model were performed for $N = 8, 16, 24, 48, 96$ and $K = 2, 4, 8, 16, 24, 48$, with $K < N$ and also for

$K = N - 1$. 500 adaptive walks were generated for all N, K values except for $N = 96$, for which 200 were generated. Each locus is affected by $K/2$ loci on either side, wrapping around the end of the genome, if necessary.

For each combination of N and K , the simulations were performed first for $w_1 = 1/N$, which correspond to the original NK model, and then for $w_1 = 0.1, 0.2, \dots, 0.9$. In each case, the remaining $N - 1$ loci have equal weights $w_i = (1 - w_1)/(N - 1)$.

The average fitness of a local maximum and the average number of steps needed to reach a local maximum are presented in Figure 5 for $N = 48$ and in Figure 6 for $N = 96$. For small values of w_1 , the average fitness of a local maximum is decreasing as K increases, which is consistent with the discussion in the previous sections. However, when w_1 is large, this behavior is reversed in that the average fitness of a local maximum increases with K . The average number of steps needed to reach a local maximum, however, appears to be independent of the weight. The following tables provide approximate values of the weight of the heavily-weighted locus at which the expected fitness of a local maximum for two different values of K cross over:

Cross-over Weights for $N = 48$							Cross-over Weights for $N = 96$							
	K							K						
K	2	4	8	16	24	47	K	2	4	8	16	24	48	95
2	–	0.1	0.2	0.3	0.35	0.5	2	–	0.1	0.2	0.3	0.35	0.4	0.5
4		–	0.4	0.45	0.5	0.6	4		–	0.3	0.4	0.45	0.5	0.65
8			–	0.45	0.6	0.7	8			–	0.4	0.5	0.6	0.7
16				–	0.6	0.85	16				–	0.5	0.6	0.8
24					–	0.9	24					–	0.7	0.85
							48						–	0.85

These results verify the prediction that, with sufficiently high differentiation in the individual loci weights, the higher number of choices for the fitness contribution of a heavily-weighted locus—resulting from high epistatic interaction at that locus—leads to better fitness.

4.2 The Case of Two Heavily-Weighted Loci

The second model considered in this section is one with two heavily-weighted loci, each with weight $w/2$. Specifically, fix loci i and j and assume that $w_i = w_j = w/2$ and $w_k = (1 - w)/(N - 2)$ for $k \neq i, j$. In the event that none of the K loci that affect the fitness contribution of locus i also affect the contribution of locus j , locus i and the loci that affect locus i behave independently of locus j and the loci that affect locus j . Provided that w is sufficiently close to 1, the arguments in Section 4.1 apply for loci i and j separately. In particular, this model can be approximated by another in which each of the two sets of i and j affectors follow an adaptive walk to maximize f_i and f_j , respectively, and the alleles at the remaining loci are set at random. For large K , the same approximation as in (5) holds for each of the two local optima of f_i and f_j . Therefore, (6) also provides a valid approximation for the fitness of a local maximum for this model.

When K is greater than or equal to $N/4$, some loci that affect the fitness contribution of locus i will also affect the fitness contribution of locus j . A computer simulation shows the impact of having these overlapping effector loci.

The simulations for this model were performed by weighting 2 loci, (loci $N/4$ and $3N/4$) each with weight $1/3$. The remaining weight was distributed evenly among the other $N - 2$ loci, that is, $w_i = 1/(3(N - 2))$. 500 adaptive walks were completed for $N = 8, 16, 24$, and 48, and 200 trials were completed for $N = 96$. For each value of N , the simulations start with $K = 2$, and the value of K is doubled as long as $2K < N/2$ (which implies that the two heavily-weighted loci have disjoint effector sets). When $2K \geq N/2$, these two loci are no longer affected by entirely different effectors. In this case, K is set to $N/2$ and successively increased by 2, 4, 8, and 16 so long as $K < N$ and also set to $K = N - 1$, in order to study the effect of increasing overlap in the effector sets.

The average fitness and average number of steps needed to reach a local maximum are reported in Figure 7(a) and (b). It can be observed that the average fitness increases with K , as long as K is such that the sets of effectors of the two heavily-weighted loci are disjoint. However, as the number of common effectors increases, the fitness starts decreasing and continues to decrease due to the increased number of epistatic conflicts. Only in the case $N = 8$ does the fitness fall below the initial fitness associated with $K = 2$. The

simulation results indicate that when there are two heavily-weighted loci, i and j , high epistatic interaction at these loci is desirable provided that the loci that affect the fitness contribution of locus i do not affect the fitness contribution of locus j . Even when there are overlapping affector loci, the decrease in overall fitness is not as substantial as in the NK model.

Turning to the number of steps needed to reach a local maximum, for each value of N , this number decreases as the value of K increases. This decrease is relatively large until there are overlapping affector loci, at which point the decrease in the number of steps needed to reach a local maximum is more gradual.

A comparison of 200 simulations of this two heavily-weighted NK/W model and the NK model for $N = 96$ is given in Figure 8. This comparison indicates that for all values of K , the average fitness of a local maximum in this NK/W model is greater than that of the original NK model. Furthermore, the decrease due to the complexity catastrophe is much slower in the NK/W model.

Conclusions and Directions for Future Research

Three variations of Kauffman's NK model for studying the effects of epistatic interaction on genome evolution are presented. In the NK_i model described in Section 2, the amount of epistatic interaction, K_i , varies with each locus i . In this case, computer simulations confirm Kauffman's complexity catastrophe in that whenever one or more loci have high levels of epistatic interaction (indicated by a large value of K_i), the overall fitness of the final genome obtained by an adaptive walk decreases.

In the NK/A model presented in Section 3, the distribution of epistatic interaction is controlled by choosing A affector loci. For fixed values of N and K , the K loci that affect any locus i are chosen among these A affector loci. Even in this model, high epistatic interaction, in the form of few affector loci that affect many loci, decreases fitness of a local maximum obtained by an adaptive walk.

The complexity catastrophe can be attenuated, to some degree, by differentiating the weights of the fitness contributions from each locus, as done in the NK/W model presented in Section 4. Specifically, the fitness of a genome is now the weighted sum of the fitness

contributions from each locus. In the case when there is one heavily-weighted locus, high epistatic interaction at this one locus is shown, both analytically and by simulation, to lead to better fitness obtained from an adaptive walk. When there are two heavily-weighted loci, better fitness is obtained when there is high, non-overlapping epistatic interaction at these two loci. Even when there are overlapping affector loci, the decrease in fitness associated with increasing K in this NK/W model is not as great as in the NK model.

In this paper, the expected value of a local maximum and the number of steps needed to obtain it were investigated. An area for future research is to explore, for each variation of the NK model, the expected number of local maxima, the number of neighbors examined, the number of local maxima that can be reached from a given initial genome, and the number of genomes that result in the same local maximum.

Perhaps the most important result of this work is that it appears worthwhile to seek ways to differentiate the contribution of loci and their allelic settings to the overall fitness of a genome. As shown in this paper, one way to do so, based on biological considerations, is to weight the fitness contributions of each locus differently. The consequences of doing so in the NK model using an adaptive walk to model evolution are clear. It would be interesting to explore how this type of modification affects other models of genomic evolution on fitness landscapes, such as population-based search and recombination. Other approaches to differentiating the contribution of the loci in these models are also worth investigating.

Acknowledgment

We thank Bennett Levitan, whose extensive and substantive comments on all aspects of the paper led to this improved version. We also are grateful to Ming-Chi Tsai for his help in tracking down many of the references.

References

- Allison, A. C. (1954.), 'Protection afforded by sickle-cell trait against subtertian malarial infection', *Br. Med. J.* **i**, p. 290.
- Amitrano, C., Peliti, L., & Saber, M. (1989), 'Population dynamics in a spin-glass model of chemical evolution', *J. Mol. Evol.* **29**, p. 513.

- Bodmer, J.G., Marsh, S.G.E., Albert, E.D., Bodmer, W.F., Bontrop, R.E., Charron, D., Dupont, B., Erlich, H.A., Mach, B., Mayr, W.R., Parham, P., Sasazuki, T., Schreuder, G.M.T., Strominger, J.L., Svejgaard, A., and Terasaki, P.I. (1995), ‘Nomenclature for factors of the HLA system’, *Tiss. Antigens* **46**, p. 1.
- Derrida, B. (1981), ‘Random-energy model: An exactly solvable model of disordered systems’, *Phys. Rev. B Condens. Matter* **24**, p. 2613.
- Eigen, M., McCaskill, J. & Schuster, P. (1989), ‘The molecular quasispecies’, *Adv. Chem. Phys.* **75**, p. 149.
- Ewens, W. (1979), *Mathematical Population Genetics*, Springer Verlag, New York.
- Flyvbjerg, H. & Lautrup, B. (1992), ‘Evolution in a rugged fitness landscape’, *Phys. Rev. A At. Mol. Opt. Phys.* **46**, p. 6714.
- Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D., & Schuster, P. (1993), ‘RNA folding and combinatorial landscapes’, *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **47**, p. 2083.
- Franklin, I. & Lewontin, R. (1970), ‘Is the gene the unit of selection?’, *Genetics* **65**, p. 707.
- Greenspan, N.S. (1998), “Genomic logic, allelic inference, and the functional classification of genes”, *Perspectives Biol. Med.*, **41**, p. 409.
- Hansen, T.H., C. B. & Sachs, D. (1993), The Major Histocompatibility Complex, in *Fundamental Immunology*, 3rd. ed., W. E. Paul (ed.), Raven Press, chapter 16, p. 595.
- Kauffman, S. A. (1993), *The Origins of Order*, Oxford University Press, Oxford.
- Kauffman, S. A., & Levin, S. (1987), ‘Towards a general theory of adaptive walks on rugged landscapes’ *J. Theor. Biol.* **128**, p. 11.
- Kauffman, S. A., Weinberger, E. D., & Perelson, A.S. (1988), ‘Maturation of the immune response via adaptive walks on affinity landscapes’, in *Theoretical Immunology, Part One, SFI Studies in the Sciences of Complexity*, Ed. A. S. Perelson, Addison-Wesley Publishing Company.

- Levinthal, D.A. (1997), 'Adaptation on rugged landscapes', *Management Science* **43**, p. 934.
- Lewontin, R. (1974), *The Genetic Basis of Evolutionary Change*, Columbia University Press.
- Macken, C. A., Hagan, P. S., & Perelson, A. S. (1991), 'Evolutionary walks on rugged landscapes' *SIAM J. Appl. Math.* **51**, p. 799.
- Macken, C. A. & Perelson, A. S. (1989), 'Protein evolution on rugged landscapes', *Proc. Natl. Acad. Sci. USA* **86**, p. 6191.
- Perelson, A. S. & Macken, C. A. (1995), 'Protein evolution on partially correlated landscapes', *Proc. Natl. Acad. Sci. USA* **92**, p. 9657.
- Schorle, H., & Horak, I. (1991), 'Development and function of T-cells in mice rendered interleukin-2 deficient by gene targeting', *Nature* **352**, p. 621.
- Schuster, P. & Stadler, P. F. (1994), 'Landscapes: complex optimization problems and biopolymer structures', *Comput. Chem.* **18**, p. 295.
- Sibilia, M. & Wagner, E. (1995), 'Strain-dependent epithelial defects in mice lacking the EGF receptor', *Science* **269**, p. 234.
- Stryer, L. (1988), *Biochemistry*, W. H. Freeman and Company, New York.
- Threadgill, D.W., Dlugosz, A.A., Hansen, L.A., Tennenbaum, T., Lichti, U., Yee, D., LaMantia, C., Mourton, T., Herrup, K., Harris, R.C., Barnard, J.A., Yuspa, S.H., Coffey, R.J., and Magnuson, T. (1995), 'Targeted disruption of mouse EGF receptor: Effect of genetic background on mutant phenotype', *Science* **269**, p. 230.
- Weinberger, E. D. (1988), 'A more rigorous derivation of some properties of uncorrelated fitness landscapes', *J. Theor. Biol.* **134**, p. 125.
- Weinberger, E. D. (1991), 'Local properties of Kauffman's N-K model: A tunably rugged energy landscape', *Phys. Rev. A At. Mol. Opt. Phys.* **44**, p. 6399.