Distinguishing the effects of epistasis and pleiotropy using a variant of the NK model

Jonathan M. Fisher[1] and Jason H. Moore[1]

[1]Dartmouth Medical School, One Medical Center Drive, HB 7937, Lebanon, NH 03766
jonathan.m.fisher@dartmouth.edu

Abstract

Pleiotropy and epistasis are central to understanding how genes are expressed. Kauffman's NK model is used ubiquitously to investigate gene expression in evolution and other contexts; it is widely understood to reflect the results of epistasis, but it is less often used to study pleiotropy. In this paper we introduce the NEP model, a variant of the NK model which allows epistasis and pleiotropy to be studied individually. We apply our methods to global and local optima and adaptive walks, and elucidate new insights into Kauffman's complexity catastrophe.

Introduction

Pleiotropy, which refers to a single locus affecting more than one trait, and epistasis, several loci collectively affecting a single trait, have long been recognized to be fundamental to our understanding of gene expression (Tyler et al., 2009). Epistasis is widely encountered in humans (Moore, 2003) and other organisms (Remold and Lenski, 2004; Bonhoeffer et al., 2004), as is pleiotropy (Ostrowski et al., 2005; Wagner et al., 2008; Scarcelli et al., 2007). Epistasis and pleiotropy are also seen to play a key role in evolution (Phillips, 2008; Fenster et al., 1997). Thus it is important to form a clear picture of the mechanisms of epistasis and pleiotropy and their effects on phenotypes.

Kauffman's NK model (Kauffman and Levin, 1987), a computational model of genomes in fitness landscapes, has been widely used to investigate properties of fitness spaces (Kauffman, 1993; Weinberger and Stadler, 1993; Macken and Perelson, 1989; Orr, 2005) and evolution thereon (Østman et al., 2010). A number of variants of the NK model have also been studied, such as the infinite-allele variant (Welch and Waxman, 2005) and the block model (Perelson and Macken, 1995); the NK model and its variants have been shown to be applicable to a variety of biological phenomena (Macken and Perelson, 1989; Perelson and Macken, 1995; Kauffman and Weinberger, 1989; Orr, 2006).

Epistasis and pleiotropy can be tuned in the NK model, but they always vary in tandem, which makes it difficult to study the two effects separately. In this paper we describe the NEP model, a variant of the NK model in which epistasis and pleiotropy can be tuned independently.

Methods

Models

The NK Model The NK model comprises a population of genomes, each of which consists of N loci, with A alleles at each locus. The model defines one trait for each locus; the locus interacts epistatically with K other loci in determining that trait. The fitness of a genome is calculated by averaging the fitness contributions of all of the traits. Each trait is represented by a $(K+1)$-dimensional table, with the length along each dimension equal to A, where the values in the table are stochastically chosen from a uniform distribution. The fitness contribution for each trait is selected from the table by choosing the row corresponding to the allele of the base locus, the column corresponding to the allele of the next locus, etc.

Choosing which other loci interact with a given locus can either be done deterministically, by having each locus interact with the succeeding K loci (where the genome is assumed to be circular), or stochastically, by choosing K other loci at random. Because the NK model contains a trait/table for each of the N loci, there are N traits/tables in total. In the rest of the paper we will refer to traits and tables interchangeably.

Fig. 1 gives an example genome with its associated tables; the top part of this figure refers to the NK model, and the bottom part contains tables that are added for the NEP model. In this example $N = 4$ and $A = 2$, so each genome contains 4 loci with 2 alleles each; the interaction degree, K, is 1. The horizontal bar in the middle of the diagram is the genome, and the 4 dots on the bar are the loci. The numbers $0, 1, 1, 0$ along the genome are the alleles at each locus.

In this example, each consecutive pair of loci interact in a trait, as do the outer two loci, for a total of four traits. Each trait is shown in the diagram as a two-dimensional lookup table above the genome, which is linked to its pair of loci by a pair of lines. The first pair of alleles is (0, 1), so the corresponding fitness contribution is the .71 shown in the

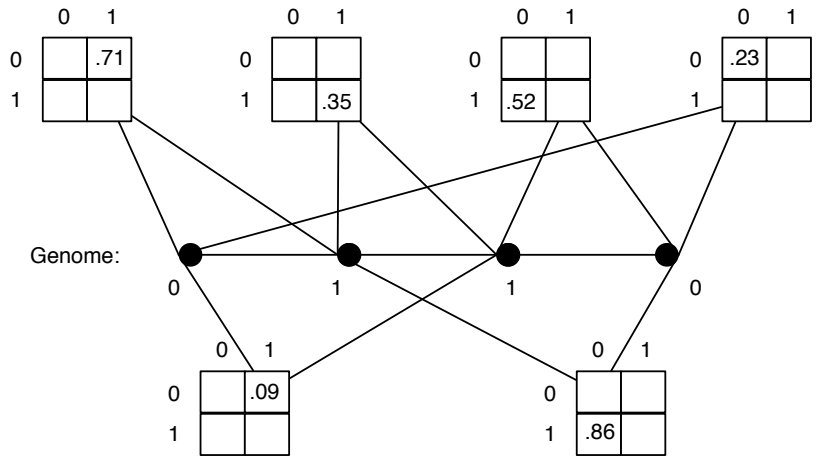

Figure 1: An example genome with its tables. The genome, with alleles, runs across the center of the figure; each pair of loci is linked to a table which represents the trait determined by that pair of loci. The single entry shown in each table is the fitness chosen by the allele values of the corresponding loci. The four tables along the top of the figure constitute an NK model, and all six tables together form an NEP model.

$0th$ row and the $1st$ column; the fitness contributions for the other pairs of loci are shown similarly. (In an actual instance of the model, of course, all of the values in each table would be filled in.) The overall fitness of the genome is then the average of the four lookup values.

The parameter K has traditionally been used to tune the degree of epistasis in the model: K determines the degree of epistasis, because each locus in the model interacts with K other loci. However, we note that K also determines the degree of pleiotropy, because each locus appears in $K + 1$ tables. The NK model has no way to separate epistasis from pleiotropy, however, which can lead to uncertainty about which of the two is causing any particular observed effect.

The NEP Model In order to separate epistasis from pleiotropy, we introduce the NEP model, adding the two new parameters E and P. To assist in describing E and P, we define T as the number of tables in the model. For i from 1 to T, E_i is the number of loci in the ith table — in other words, the number of loci that are used to look up the value in the ith table. The loci used in each table are chosen at random. We refer to E_i as the epistatic dimension of table i. For j from 1 to N, P_j is the number of tables in which the jth locus appears, referred to as the pleiotropic dimension of that locus. The NEP model is a generalization of the NK model, and the parameters E and P can be seen in the NK model: since the parameter K refers to how many *other* loci a given locus interacts with, while our parameter E refers to the *total* number of interacting loci, any given NK model has $K = E - 1$; similarly, $K = P - 1$. The difference between the NK model and the NEP model is that the NK model fixes $P \equiv E$ but the NEP model allows P to be different from E.

When two or more loci interact epistatically in the NEP model, the fitness contribution for that interaction is chosen from a stochastic lookup table. The stochasticity of the lookup table means that the interaction between the loci in determining the fitness is highly non-linear, consistent with the usual definition of epistasis.

Since a genome can have A possible values at each locus, there are a total of A^N genomes in the fitness space in either the NK model or the NEP model.

In the example of the NK model in Fig. 1 described above, $E = 2$ and $P = 2$; as always in the NK model, $P \equiv E$. However, in the NEP model we can modify the example to have P differ from E, by adding the two tables shown below the genome. The dimension E of each table is still 2, but now each locus is linked to 3 tables, so $P = 3$.

Referring to Fig. 1, there are two ways to count the links between the genome and the trait tables: If we count the links where they connect to the tables, we get $\sum_{i=1}^{T} E_i$ since the epistatic dimension of each table is equal to the number of links connecting to it. On the other hand, if we count the links where they connect to the loci on the genome, we get $\sum_{j=1}^{N} P_j$ since the pleiotropic dimension of each locus is equal to the number of links connecting to it. Since those two counts must be equal, we find that

$$\sum_{j=1}^{N} P_j = \sum_{i=1}^{T} E_i \, . \tag{1}$$

This equation can be rewritten as $N\bar{P} = \bar{E}T$, where $\bar{P}$ and $\bar{E}$ are the average values of P_j and E_i respectively. In practice we usually either choose to have P_j be the same for all loci, $P_j \equiv P$, or choose to have E_i be the same for all tables, $E_i \equiv E$. In this paper we take the former approach.

Note that the embedded-landscape model (Altenberg,

1994; Heckendorn et al., 1999) has previously introduced flexibility in the number of tables to the NK model, in the same way as the NEP model; however, the embedded-landscape model has generally been used in studies of computational complexity and has typically not delineated the effects of epistasis and pleiotropy.

Numerical Analysis

The current study employs an NEP model with $N = 20$ and $A = 2$. We simulated 400 different models: for P equal to each of 1 through 20 we chose, using the equation $NP = \bar{E}T$, the value of $\bar{E}$ closest to each of 1 through 20, for a total of 400 $(P, \bar{E})$ pairs. When $\bar{E}$ was an integer we chose all of the tables to have that value for E_i. When $\bar{E}$ was not an integer we chose some of the tables to have E_i equal to the next integer below $\bar{E}$, and some the next integer above, with the counts of each chosen to give the desired average $\bar{E}$. This ensured that the tables were as similar in size as possible. Each pair $(P, \bar{E})$ gave a model, and each model was run 100 times, populating the lookup tables with a different random seed each time; all of the statistics for each model were averaged across those runs.

Among the phenomena we study here are local optima and adaptive walks, both of which are important in studying the dynamics of evolution (Orr, 2005). If the fitness of a given genome is greater than the fitness of any genome at a Hamming distance of 1 from the given genome, that fitness is said to be a local optimum. An adaptive walk is a sequence of genomes that starts at a randomly-selected genome and proceeds by single fitness-improving mutations until it reaches a local optimum.

We collected the following statistics: variance in fitness within each fitness space; global optimum fitness, defined as the largest fitness value in the fitness space; mean local optimum fitness; number of local optima in the fitness space; average length of adaptive walks, defined as the number of fitness-improving mutations traversed in the walk; fitnesses attained in adaptive walks, defined as the fitness of the last genome in the adaptive walk (by definition, a local optimum); first step up in adaptive walks, defined as the difference in fitness between the first genome and the second genome in an adaptive walk; last step up in adaptive walks, defined as the difference in fitness between the second-to-last genome and the last genome in an adaptive walk; and the maximum step up, defined as the largest difference in fitness between any two adjacent genomes in the fitness space. These are some of the most commonly-measured statistics in studies of the NK model (Kauffman, 1993; Macken and Perelson, 1989) and its variants (Perelson and Macken, 1995).

In each run, the variance, the global optimum, the mean local optimum, and the number of local optima were calculated by iterating across all 2^{20} genomes in the fitness space. The maximum step up was calculated by iterating across all pairs of genomes. To calculate statistics on adaptive walks, 1,000 adaptive walks were simulated in each run; each reported statistic is then the average value of that statistic across the 1,000 adaptive walks simulated.

Representative results from the 400 models are described below.

Results

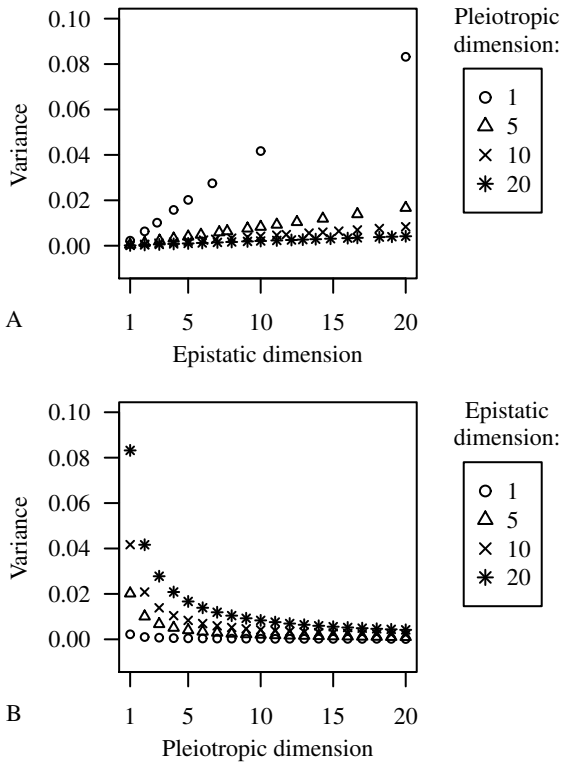

Figure 2: Fitness variance by epistatic dimension and pleiotropic dimension.

Fitnesses and Optima

Variance in fitness increased approximately linearly with epistatic dimension (Fig. 2A) and decreased with pleiotropic dimension (Fig. 2B). This can be understood by considering what happens when we increment either the epistatic dimension or the pleiotropic dimension. First, if we increase $\bar{P}$ by 1 while holding $\bar{E}$ fixed, the equation $N\bar{P} = \bar{E}T$ tells us that we need to multiply T by $(\bar{P} + 1)/\bar{P}$. Since the fitness of any given genome is calculated by averaging one entry from each of the T tables, multiplying T by $(\bar{P} + 1)/\bar{P}$ means that each calculated fitness is averaged across more table-entries, which decreases the fitness variance.

Secondly, consider incrementing $\bar{E}$ while holding $\bar{P}$ fixed. This requires us to multiply T by $\bar{E}/(\bar{E} + 1)$, decreasing the value of T, which has the opposite effect to

incrementing $\bar{P}$: it increases the variance of fitnesses. Incrementing $\bar{E}$ also increases the variance by increasing the degrees of freedom of the model, which is equal to the number of separately-generated stochastic table-entries. Since each table contains $A^{\bar{E}}$ entries, for any given values of A, $\bar{E}$, and T, the number of separately-generated stochastic table-entries is

$$TA^{\bar{E}}. \tag{2}$$

If we increment $\bar{E}$ while holding $\bar{P}$ fixed then the new total number of entries is given by

$$T\frac{\bar{E}}{\bar{E}+1}A^{\bar{E}+1}. \tag{3}$$

Thus the net effect of incrementing $\bar{E}$ is to multiply the number of separately-generated stochastic table-entries, and thus the degrees of freedom, by $A\bar{E}/(\bar{E}+1)$. For $A = 2$, this is greater than 1 as long as $E > 1$.

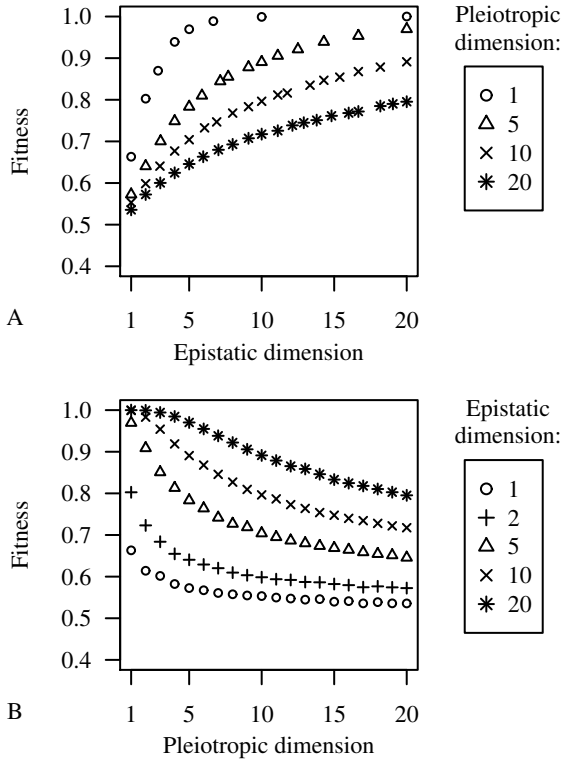

Figure 3: Global optimum fitness by epistatic dimension and pleiotropic dimension.

Fitnesses of global optima increased with epistatic dimension; for smaller pleiotropic dimension (up through about 5) the global optimum approached 1.0 (Fig. 3A). The global optima decreased with pleiotropic dimension (Fig. 3B). The global optimum is an extreme value of fitness, so as the variance increases with epistatic dimension, the global optimum also increases. The opposite argument shows why the global optimum decreases with pleiotropic dimension.

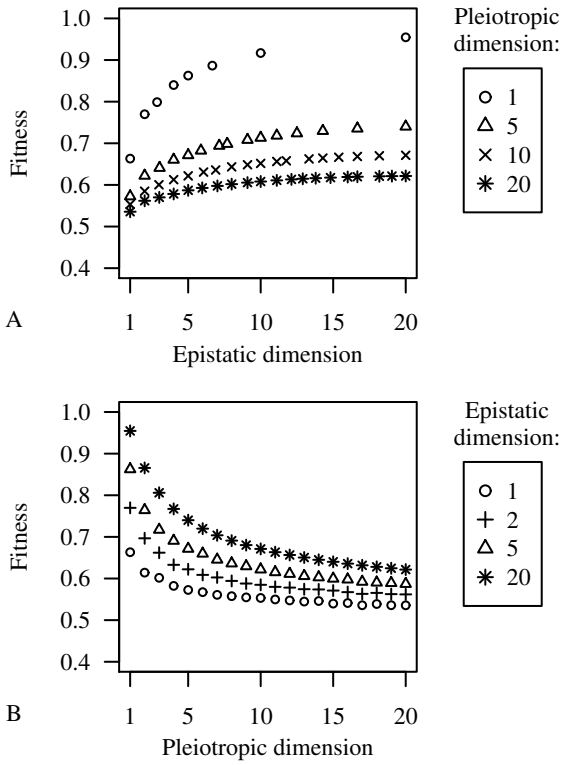

Figure 4: Local optimum fitness by epistatic dimension and pleiotropic dimension.

As with global optima, fitnesses of local optima increased with epistatic dimension (Fig. 4A) and decreased with pleiotropic dimension (Fig. 4B). Since local optima are local extreme values, the same argument as with global optima indicates why local optima increase with epistatic dimension and decrease with pleiotropic dimension.

The number of local optima increased dramatically with epistatic dimension (Fig. 5A); for high and low epistatic dimensions, the number of local optima did not vary with pleiotropic dimension, but for intermediate epistatic dimensions, lower pleiotropic dimensions gave slightly higher numbers of local optima (Fig. 5B). Note also that for an epistatic dimension of 2, the graph of the number of local optima by pleiotropic dimension is a little noisy; later we will notice the same phenomenon in a different way with the lengths of adaptive walks.

Kauffman proves (Kauffman, 1993) that for $K = N - 1$ the number of local optima is very large; the proof uses only the epistasis effects of K, not the pleiotropy effects. The results we get here are consistent with that, in that the number of local optima increases with epistatic dimension but varies little with pleiotropic dimension.

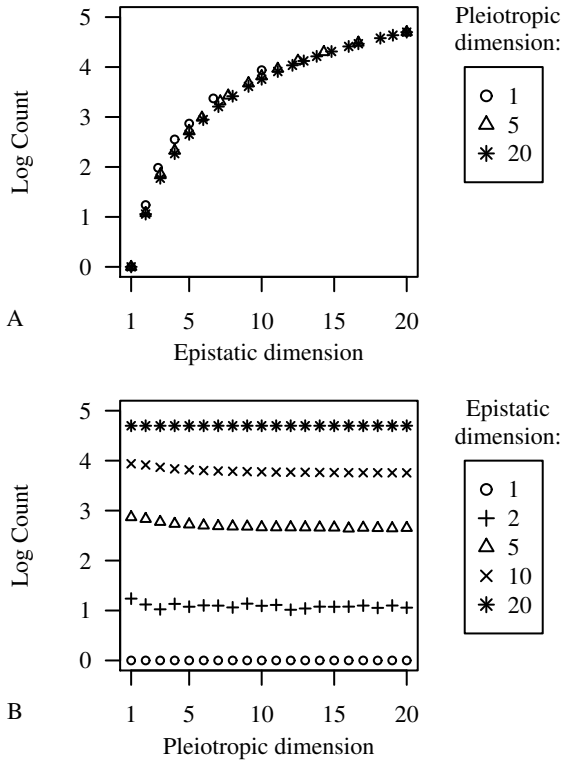

Figure 5: Log number of local optima by epistatic dimension and pleiotropic dimension.

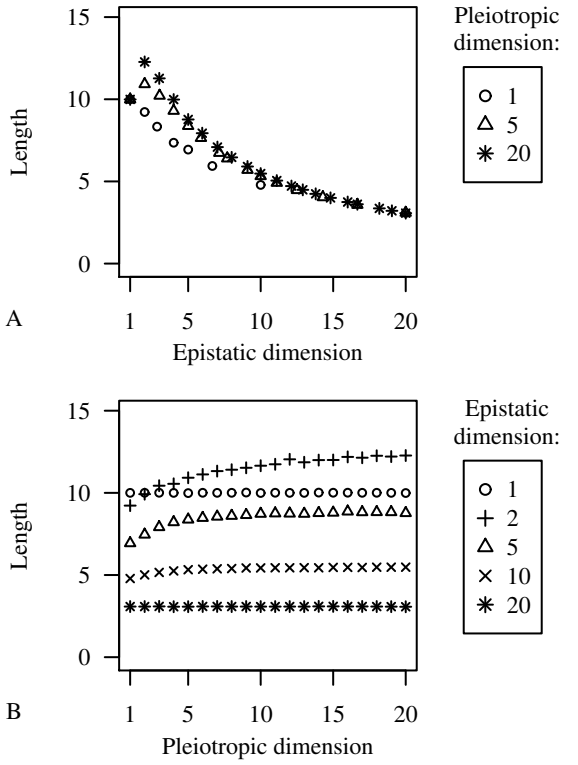

Figure 6: Length of adaptive walks by epistatic dimension and pleiotropic dimension.

Adaptive Walks

In general the length of adaptive walks decreased with epistatic dimension (Fig. 6A) and increased or held steady with pleiotropic dimension (Fig. 6B). For epistatic dimension of 1 the average length was 10, regardless of pleiotropic dimension; otherwise, for each pleiotropic dimension, the average length peaked at an epistatic dimension of 2 and decreased from there. For epistatic dimension between 2 and about 10, lengths of adaptive walks increased as a function of pleiotropic dimension.

Kauffman proves (Kauffman, 1993) that in an NK model where $A = 2$ and $K = 0$ the average length of an adaptive walk is $N/2$. Again, the proof uses only the epistasis effects of K, and thus holds for the NEP model; this explains why the average length is 10 in our models when the epistatic dimension is 1 (which is equivalent to $K = 0$). For epistatic dimensions greater than 1, the average adaptive-walk length is inversely related to the number of local optima, and thus it decreases with epistatic dimension and increases with pleiotropic dimension for low epistatic dimensions. The inverse relationship with number of local optima is also why we see the same noisiness in the graph of adaptive-walk length by pleiotropic dimension for epistatic dimension of 2 as we did with the number of local optima.

Attained fitness in adaptive walks increased with epistatic dimension, with the effect being most pronounced for low pleiotropic dimension (Fig. 7A); attained fitness decreased with pleiotropic dimension (Fig. 7B).

Adaptive-walk lengths decrease with epistatic dimension but attained fitnesses increase; the opposite is true with respect to pleiotropic dimension. This apparent contradiction is resolved by Figures 8 and 9, which show the typical steps in adaptive walks.

The typical first step up in an adaptive walk increased with epistatic dimension (Fig. 8A) and decreased with pleiotropic dimension (Fig. 8B). The same is true for the typical last step (Fig. 9), with one exception: for epistatic dimensions greater than about 10, the typical last step up was lower for a pleiotropic dimension of 1 than for the next few pleiotropic dimensions. Again we see the effects of variance: increasing the variance with epistatic dimension allows for larger steps up, while decreasing it with pleiotropic dimension decreases the possible steps up. The phenomenon whereby average steps up increase with increasing K in the NK model, allowing higher fitness to be achieved with fewer steps, has been observed by Østman et al. (2010), who ascribed the phenomenon to a combination of pleiotropy and epistasis. In contrast, we find that the increasing amplitude of steps up is due solely to increasing epistasis, not increasing pleiotropy,

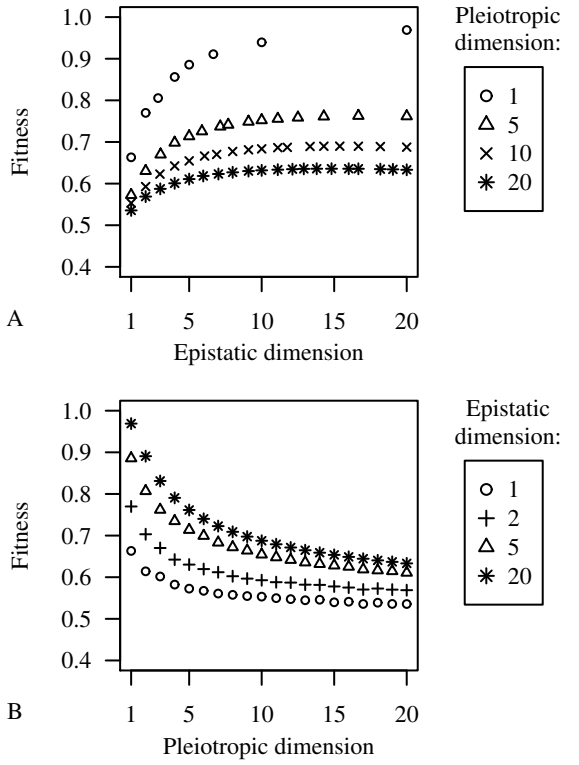

Figure 7: Attained fitness of adaptive walks by epistatic dimension and pleiotropic dimension.

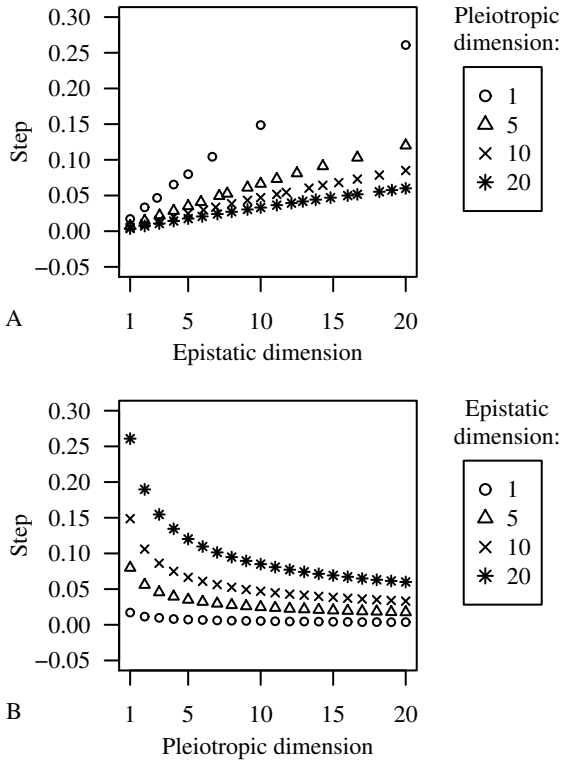

Figure 8: First step of adaptive walks by epistatic dimension and pleiotropic dimension.

and, in fact, that amplitudes of steps up decrease with increasing pleiotropy.

Except for an epistatic dimension of 1 (where they were the same), the typical last step was less than the typical first step for all epistatic dimensions and pleiotropic dimensions; the effect was more pronounced for higher epistatic dimension and lower pleiotropic dimension. The reason for this is that on the last step of an adaptive walk the fitness of the genome is already quite high, limiting the remaining steps available.

The maximum step up, defined as the largest difference in fitness between two adjacent genomes, increased with epistatic dimension (Fig. 10A) and decreased with pleiotropic dimension (Fig. 10B). Again we see the result of the fact that fitness variance increases with epistatic dimension and decreases with pleiotropic dimension.

Discussion

In the NEP model, referring to the equation $N\bar{P} = \bar{E}T$ we see that one way to increase the pleiotropic dimension P is by increasing T by adding one or more traits. As the results here have shown, increasing pleiotropy in this way tends to decrease the local and global optima. On the other hand, when a biological species adds a trait its fitness generally goes up; it may be that when a biological species adds a trait it also increases the overall number of loci, thus avoiding an increase in pleiotropy. Further investigation will be required to resolve this question.

Effects of Epistasis and Pleiotropy

The overall effect of epistasis is to increase the variance in fitness, and of pleiotropy to decrease it. We see this first in the direct measurements of variance of fitness; we also see it in measurements of global and local optima: less variance reduces the heights of the available maxima. Attained fitnesses in adaptive walks match closely with mean local optima, so they, too, decrease with decreasing variance. First and last steps in adaptive walks, and maximum single steps, also trend in the same direction as variance. The situation here is more complicated, however: because a single step means mutating a single allele, we are very far from the extreme-value considerations that apply to global and local optima.

A key difference between the NEP model and the embedded-landscape model is that the latter does not fully separate the effects of epistasis and pleiotropy. For example, (Smith and Smith, 1999) find that "the epistasis parameter K has little effect on the global fitness statistics, and does not affect the mean fitness values of the local optima at all." However, when we separated epistasis out from pleiotropy,

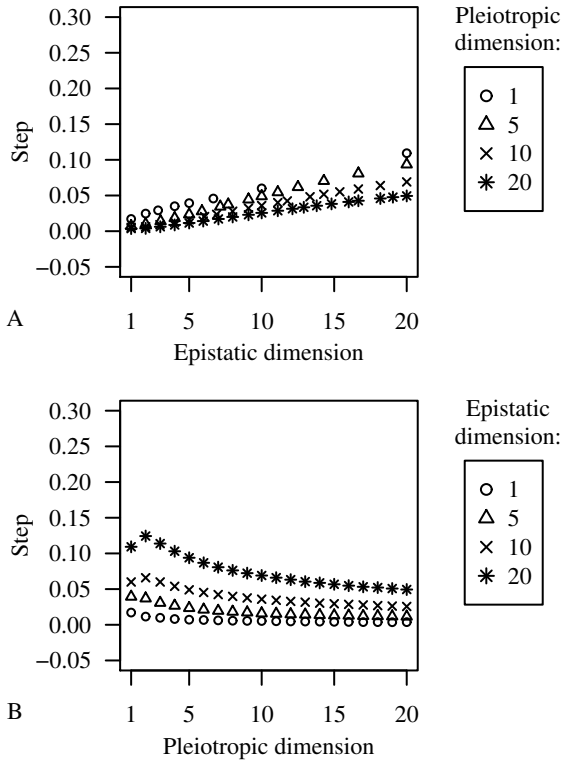

Figure 9: Last step of adaptive walks by epistatic dimension and pleiotropic dimension.

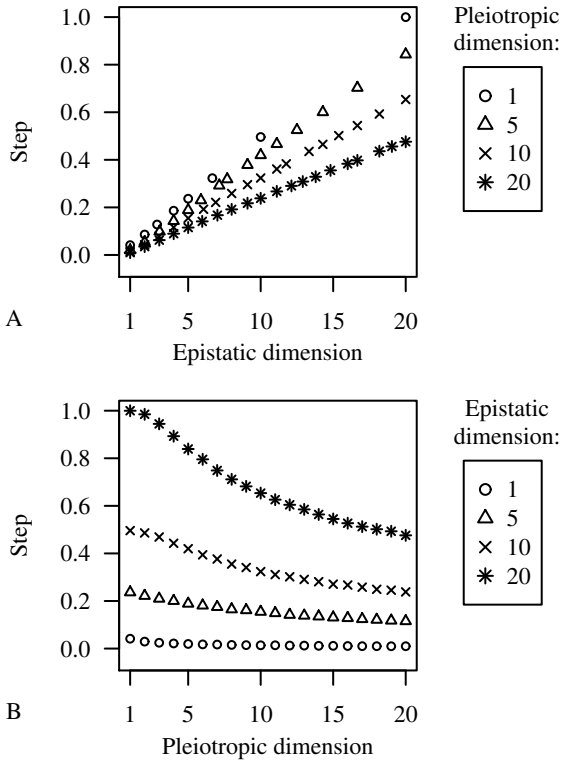

Figure 10: Maximum single step by epistatic dimension and pleiotropic dimension.

we found a clear dependency of global optimum fitness and mean local optimum fitness on epistasis.

The Complexity Catastrophe

In working with the NK model, Kauffman observed a tendency for fitnesses of local optima to decrease with increasing K, and coined the term "complexity catastrophe" to describe this (Kauffman, 1993). Previous authors have ascribed the decrease in local optima to an increase in epistasis (e.g., Kauffman (1993); Solow et al. (1999)); in contrast, in the context of the embedded-landscape model (Smith and Smith, 1999) stated that the decrease in local optima is due to an increase in the number of traits.

The results described here allow us to clarify which part of the complexity causes the catastrophe. As pleiotropic dimension increases, both attained fitnesses and the fitnesses of local optima decrease. On the other hand, as epistatic dimension increases, both attained fitnesses and the fitnesses of local optima *increase*. The culprit in the complexity catastrophe is simply that decreasing variance lowers the local optima.

Kauffman further observed that as K increases, the mean value of local optima initially increases, and then starts to decrease. Looking at parts A and B of Fig. 4, we see the explanation of the observed trends: Increasing K corresponds to increasing E and P simultaneously, which results in simultaneous tendencies to increase and decrease local optima, respectively; at first the former tendency predominates, and then the latter.

Conclusion

Given the ubiquity of epistasis and pleiotropy in gene expression, and given the prevalence of the NK model in studying genetic phenomena, it is important to ensure that we fully understand those two forms of gene linkage in the context of the NK model. As described here, the NEP model can be used to distinguish the effects of epistasis and pleiotropy, allowing assumptions made in the literature to be re-examined, and allowing new insights to be gained going forward.

Acknowledgements

We thank Davnah Urbach and Jeff Kiralis and our other colleagues in the lab for their assistance with preparing this paper. This work was supported by NIH grants LM009012, LM010098, and AI59694.

References

Altenberg, L. (1994). Evolving better representations through selective genome growth. In Schaffer, J. D., Schwefel, H. P.,

and Kitano, H., editors, *Proceedings of the IEEE World Congress on Computational Intelligence*, pages 182–187. IEEE, Piscataway, NJ.

Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M., and Petropoulos, C. J. (2004). Evidence for positive epistasis in HIV-1. *Science*, 306:1547–1550.

Fenster, C. B., Galloway, L. F., and Chao, L. (1997). Epistasis and its consequences for the evolution of natural populations. *Trends in Ecology & Evolution*, 12:282–286.

Heckendorn, R. B., Rana, S., and Whitley, D. (1999). Test function generators as embedded landscapes. In Bäck, T. and Banzhaf, W., editors, *Foundations of Genetic Algorithms - 5*, pages 183–198. Morgan Kaufmann, San Francisco, CA.

Kauffman, S. A. (1993). *The Origins of Order*. Oxford University Press, New York.

Kauffman, S. A. and Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128:11–45.

Kauffman, S. A. and Weinberger, E. D. (1989). The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141:211–245.

Macken, C. A. and Perelson, A. S. (1989). Protein evolution on rugged landscapes. *Proceedings of the National Academy of Sciences*, 86:6191–6195.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, 56:73–82.

Orr, H. A. (2005). The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, 6:119–127.

Orr, H. A. (2006). The population genetics of adaptation on correlated fitness landscapes: The block model. *Evolution*, 60:1113–1124.

Østman, B., Hintze, A., and Adami, C. (2010). Impact of epistasis and pleiotropy on evolutionary adaptation. *arXiv:0909.3506v2*.

Ostrowski, E. A., Rozen, D. E., and Lenski, R. E. (2005). Pleiotropic effects of beneficial mutations in escherichia coli. evolution. *Evolution*, 59:2343–2352.

Perelson, A. S. and Macken, C. A. (1995). Protein evolution on partially correlated landscapes. *Proceedings of the National Academy of Sciences*, 92:9657–9661.

Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9:855–867.

Remold, S. K. and Lenski, R. E. (2004). Pervasive joint influence of epistasis and plasticity on mutational effects in escherichia coli. *Nature Genetics*, 36:423–426.

Scarcelli, N., Cheverud, J. M., Schaal, B. A., and Kover, P. X. (2007). Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. *Proceedings of the National Academy of Sciences*, 104:16986–16991.

Smith, R. E. and Smith, J. E. (1999). An examination of tunable, random search landscapes. In Bäck, T. and Banzhaf, W., editors, *Foundations of Genetic Algorithms - 5*, pages 165–182. Morgan Kaufmann, San Francisco, CA.

Solow, D., Burnetas, A., Roeder, T., and Greenspan, N. S. (1999). Evolutionary consequences of selected locus-specific variations in epistasis and fitness contribution in Kauffman's NK model. *Journal of Theoretical Biology*, 196:181–196.

Tyler, A. L., Asselbergs, F. W., Williams, S. M., and Moore, J. H. (2009). Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *BioEssays*, 31:220–227.

Wagner, G. P., Kenney-Hunt, J. P., Pavlicev, M., Peck, J. R., Waxman, D., and Cheverud, J. M. (2008). Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature Letters*, 452:470–473.

Weinberger, E. D. and Stadler, P. F. (1993). Why *some* fitness landscapes are fractal. *Journal of Theoretical Biology*, 163:255–275.

Welch, J. J. and Waxman, D. (2005). The *nk* model and population genetics. *Journal of Theoretical Biology*, 234:329–340.