J. theor. Biol. (1995) **173,** 427–440

Search Strategies for Applied Molecular Evolution

STUART A. KAUFFMAN AND WILLIAM G. MACREADY

The Sante Fe Institute, 1399 *Hyde Park Road*, *Santa Fe*, *NM* 87501, *U.S.A.*

(*Received and accepted on* 10 *November* 1994)

A new approach to drug discovery is based on the generation of high diversity libraries of DNA, RNA, peptides or small molecules. Search of such libraries for useful molecules is an optimization problem on high-dimensional molecular fitness landscapes. We utilize a spin-glass-like model, the *NK* model, to analyze search strategies based on pooling, mutation, recombination and selective hill-climbing. Our results suggest that pooling followed by recombination and/or hill-climbing finds better candidate molecules than pooling alone on most molecular landscapes. Our results point to new experiments to assess the structure of molecular fitness landscapes and improve current models.

1. Introduction

A new approach to drug discovery, variously called applied molecular evolution, direction evolution and molecular diversity, is beginning to transform the search for drug candidates (Joyce, 1992; Kauffman, 1992). Unlike traditional biotechnology, the new techniques are based on the generation of high diversity libraries of more or less random DNA, RNA, proteins or small organic molecules, coupled with screening or selection among the resulting libraries for molecular candidates of interest. Molecules of interest might be mimics of naturally occurring hormones, hence agonize, antagonize or modulate the action of such hormones, might mimic epitopes of pathogens and hence serve as candidate vaccines, might serve as biosensors, might act as catalysts for a variety of reactions, might function as *cis*-acting regulatory DNA sequences within host DNA, or as novel genes whose products impinge upon the genomic regulatory circuitry within cells. At present, maximum diversity libraries approach 10^{14} for DNA, RNA and peptides or polypeptides (Ellington & Szostak, 1990; Beaudry & Joyce, 1992; Bartel & Szostak, 1993), and in the millions for small organic molecule libraries. The medical and basic science potential of applied molecular evolution appears to be great.

In parallel with the emerging technologies to generate and screen or select from such libraries, a growing effect is underway to study the structure of molecular "fitness landscapes" in order to understand how to optimize search for useful molecules. At present, there are three broad approaches to such search: pooling strategies, mutation and selection for fitter variants, and the use of recombination among candidate molecules. In the present article, we discuss a specific model of molecular fitness landscapes, the *NK* model (Kauffman *et al*., 1989; Weinberger, 1991; Kauffman, 1993), itself based on spin-glasses (Stein, 1992). We utilize this family of landscapes to explore both the relative usefulness of the three strategies mentioned, as well as the information about the structure of molecular fitness landscapes that current experimental efforts can uncover. The *NK* model is a first statistical model of molecular fitness landscapes. While the *NK* model has been applied with some success to maturation of the immune response (Kauffman *et al*., 1988; Kauffman & Weinberger, 1989), it can best be viewed as offering a test-bed for examining optimization strategies in molecular search. More refined theory about optimizing search for useful molecules awaits improved data on the structure of real molecular fitness landscapes.

The article is organized as follows. In the Section 2, fundamental features of rugged fitness landscapes and the *NK* model are introduced. In the Section 3, we describe pooling strategies and describe predictions

0022–5193/95/080427+14 $08.00/0

about results of such pooling strategies as a function of fitness structure. We ask whether pooling strategies yield highly fit molecular candidates when used alone, or whether fitter candidates are likely to be found by following pooling strategies with mutation and selection. Section 4 explores the efficacy of recombination without or with mutation and selection among pool strategy candidates to yield even fitter candidates. We discuss uses of single and double mutant variants of pool-optimal sequences to find better sequences in Section 5 and summarize how the various features of smooth or rugged fitness landscapes change in coordinated ways as visualized by the experimental approaches.

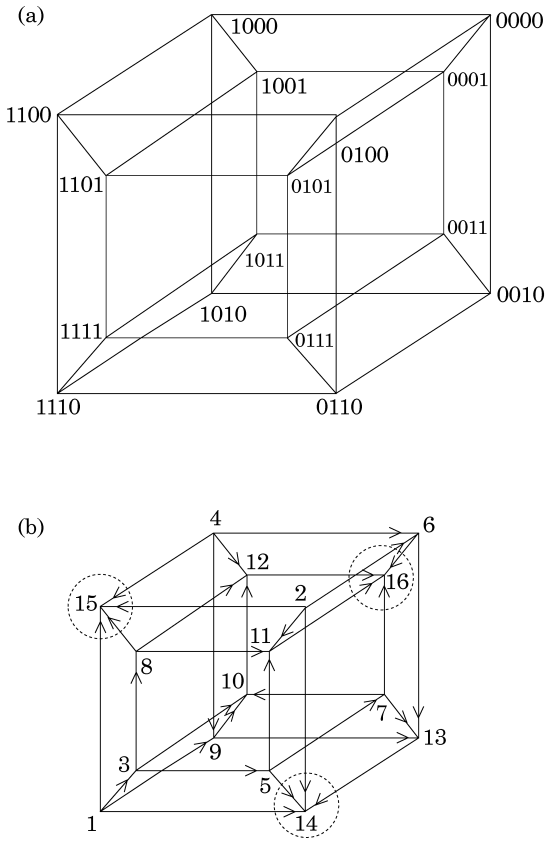


FIG. 1. (a) The four-dimensional Boolean hypercube; (b) with rank ordered fitnesses.

2. Fitness Landscapes and the *NK* Model

We define next the concepts of sequence spaces and molecular fitness landscapes. Consider sequences of letters drawn from an alphabet of A characters. If the sequence is of length N then there are A^N possible sequences of that length. For example there are 20^N possible proteins of N amino acids and 4^N possible polynucleotides of N nucleotides. To define a fitness landscape on a sequence space we need to add two features.

First, in order to define a sequence space we need a notion of neighboring sequences. The usual measure of distance in sequence spaces is Hamming distance. The Hamming distance counts the number of positions where two sequences differ. The $N(A-1)$ closest neighbors of a sequence are called its one-mutant neighbors. For each of the positions we can change the symbol to any of the $A-1$ differing symbols and we can do this at any of the N positions. (Similarly there are $\binom{N}{2}$ $(A-1)$ two-mutants whose Hamming distance is 2 from the wild-type sequence). In the case of $A=2$ with Hamming distance measuring the distance between sequences then each sequence of length N has N one-mutant neighbors. This allows definition of a sequence space in which the sequences of length N lie on the N-dimensional Boolean hypercube with the edges of the cube connecting one-mutant neighbors. The case of bit strings of length 4, hence the four-dimensional Boolean hypercube, is shown in Fig. 1(a).

The final component of a fitness landscape is a mapping from sequences to real numbers. This number may be the sequence's ability to perform some function. For example it may be a polymer's affinity for a specific ligand, say the estrogen receptor, in specified conditions. Then think of that affinity as a "height" at each point on the Boolean hypercube. The distribution of these heights forms a molecular fitness landscape for the specific function of binding the estrogen receptor. In Fig. 1(b), we have rank-ordered the 16 possible 4-bit tetramers from the worst, 1, to the best, 16. In this particular case, we have in fact assigned the rank orderings to the 16 vertices completely at random, creating a random fitness landscape.

The simplest adaptive walk on a molecular fitness landscape (Maynard Smith, 1970) starts with an arbitrary polymer, considers a randomly chosen one-mutant neighbor, and, if that neighbor is fitter, moves to the fitter neighbor. This allows the edges of the hypercube [Fig. 1(a)] to be labeled with arrows showing the directions "uphill" in fitness from any vertex [Fig. 1(b)]. An adaptive walk, in this simple sense, follows arrows uphill from an initial polymer until a polymer is reached which is fitter than all its one-mutant neighbors. Such a polymer is a local peak on the fitness landscape. Figure 1(b) shows a number of features of molecular fitness landscapes. These include the number of directions uphill at each point in the walk and how the number of directions uphill dwindles to 0 as peaks are approached, the expected number of random mutants tried on a walk to a peak, the fraction of local optima accessible to typical

initial polymers and the fraction of polymers able to climb to the global peak by adaptive walks. Beyond the simplest versions of adaptive walks, we can consider walks via two-mutant neighbors, three-mutant neighbors, or less fit neighbors, measures of the correlation structure of landscapes, and adaptive processes closer to population genetics which include mutation, recombination, selection and drift (Wright, 1932; Amitrano *et al*., 1989; Eigen *et al*., 1989; Fontana *et al*., 1989; Kauffman & Weinberger, 1989; Derrida & Peliti, 1991; Fontana *et al*., 1991; Stadler, 1992).

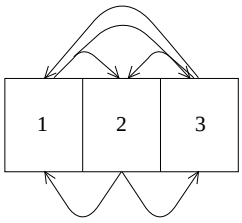
The *NK* model generates a large family of fitness landscapes. It consists of a set of N sites where each site can be in A alternative states. In the *NK* model, we posit that each site makes a contribution to the fitness of the overall polymer which depends upon its own state, s_i, and the state of K other sites in the polymer, $\{s_i, \ldots, s_{iK}\}$. Thus, K reflects "epistatic" interactions among the sites. The K sites which influence each site may be chosen in any way. For example, the K sites may be its left and right flanking neighbors, may be chosen at random among the N, or may be chosen in any other way. For each of the A^{K+1} combinations of states of the $K+1$ sites, the fitness contribution of the site in question, $F_i^{(K)}$, is assigned, once and for all, from the uniform distribution between [0, 1]. The fitness of any polymer is defined as the mean of the fitness contribution of its N sites as in eqn (1).

$$F\{s\} = \frac{1}{N}\sum_{i=1}^{N} F_i^{(K)}(s_i;\, S_i, \ldots, s_{iK}). \qquad (1)$$

A specific example of this construction of an *NK* landscape for $N=3$, $K=2$ and $A=2$ is found in Fig 2(a)–(c). Figure 2(a) first specifies the epistatic connections between sites, then Fig. 2(b) specifies the site fitness for all possible combinations of neighbors. Finally, Fig. 2(c) shows the polymer fitnesses on the Boolean hypercube with arrows pointing in the uphill directions.

The *NK* model yields a family of fitness landscapes as the major parameters, N and K, are altered. For $K=0$, each site is independent of all other sites. The resulting fitness landscape is Fujiyama-like with a single peak and smooth sides. When K takes its maximum value, $K=N-1$, each site is affected by every site, as exemplified by Fig. 2(a)–(c). A change in state at any site therefore affects all sites, yielding a new random fitness contribution by each site. Thus, one-mutant neighbors have fitnesses which are fully random with respect to one another. The $K=N-1$ limit corresponds to the random landscape shown in Fig. 1(b), and to Derrida's "random energy" model of spin glasses (Derrida, 1981). Random landscapes

(a)



(b)

s_1	s_2	s_3	$F_1^{(2)}$	$F_2^{(2)}$	$F_3^{(2)}$	$F = \frac{1}{N}\sum_{i=1}^{N} F_i^{(2)}$
0	0	0	0.6	0.3	0.5	0.47
0	0	1	0.1	0.5	0.9	0.50
0	1	0	0.4	0.8	0.1	0.43
0	1	1	0.3	0.5	0.8	0.53
1	0	0	0.9	0.9	0.7	0.83
1	0	1	0.7	0.2	0.3	0.40
1	1	0	0.6	0.7	0.6	0.63
1	1	1	0.7	0.9	0.5	0.70

(c)

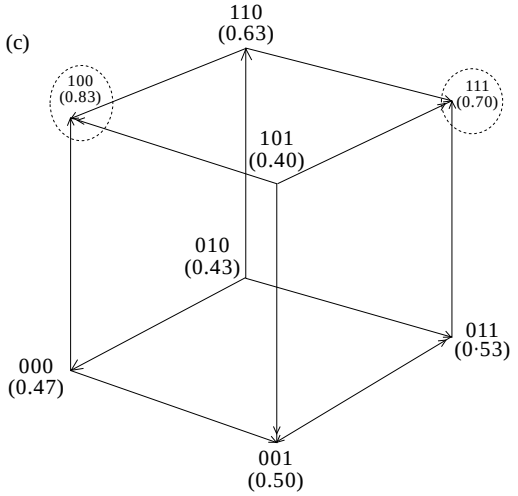


FIG. 2. (a) Assignment of the $K=2$ epistatic interactions for each of the $N=3$ sites, (b) Random assignment of fitness values, (c) Fitness landscape on the Boolean cube indicating uphill directions.

have on the order of $2^N/(N+1)$ local optima, walk lengths to optima scale as $\ln N$, every step "uphill" the expected number of directions uphill is halved, the expected number of mutants tried on a random adaptive walk is N, only a tiny fraction of local optima are accessible from any initial point, and the global optimum is accessible from only a tiny fraction of sequence space (Kauffman & Levin, 1987; Macken & Perelson, 1989; Kauffman, 1993).

As K increases from $K=0$ to $K=N-1$, landscapes become increasingly rugged and multipeaked. This reflects the increasing levels of conflicting constraints among the sites as the richness of epistatic interactions, K, increases. Because of these increasing conflicting constraints, as K increases, local peaks dwindle in height. Figures 3(a)–(d) summarize these features. Other features of *NK* landscapes are described elsewhere (Kauffman, 1993).

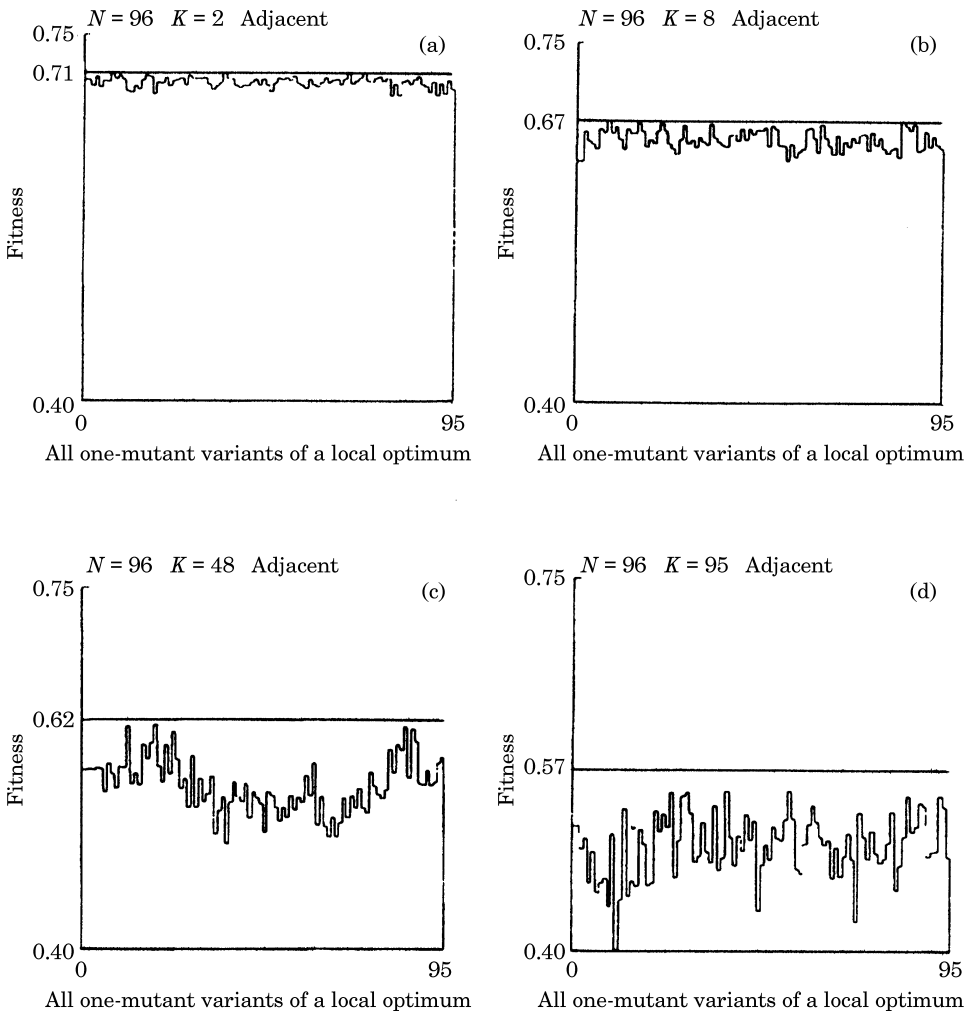


FIG. 3. The ruggedness of an NK landscape for $N=96$, showing the fitness of all 96 one-mutant variants of a random local optimum for (a) $K=2$, (b) $K=8$, (c) $K=48$, and (d) $K=95$.

3. Pooling Strategies

The first approach to the generation of high-diversity libraries of peptides and practical search among such libraries for useful polymers is the pooling strategy introduced by Geysen (Geysen *et al.*, 1987), and subsequently utilized by numerous other workers (Houghten *et al.*, 1991). Here one wishes to find a peptide, in a concrete case, a hexamer, which is able to bind to a specific monoclonal antibody or to a specific receptor. An example would be the estrogen receptor. The central idea is to create specific pools that partition the 20^6 ($=64$ million) possible hexapeptides into non-overlapping pools such that members of any pool share specific subsequences. These pools are tested for binding capacity and the best pool is picked. This pool is then used over cycles to create successive subpools each of which share additional subsequences. At each cycle the sub-pools are tested for binding and the best sub-pool is used for subsequent cycles until a single specified optimal sequence is identified. For example, one pool might have hexamers all of which have glycine at position 3 and alanine at position 4, but be random at the remaining four positions.

In Geysen's original approach, each pool is localized on one of 400 pins formed by a 20×20 array. Thus, each pin carries a mixture of hexamers which are identical in two of their amino acids, and random at the remaining four positions. The pins are tested with labeled estrogen receptor. The pin with the most bound counts is chosen for further use. For example, if the pin with phenylalanine and tryptophane at positions 3 and 4 binds the most counts, in a subsequent iteration, these two amino acids are held fixed at positions 3 and 4 and 400 partitioning sub-pools of the possible simultaneous choices at positions 2 and 5 but random in positions 1 and 6 are generated and localized on the 400 pins. Testing with labeled estrogen receptor picks

out the pin with the most counts, hence fixes the amino acids at positions 2, 3, 4 and 5. A final round of 400 partitioning sub-pools fixes the remaining 1st and 6th position of the hexamer.

Pooling strategies are remarkably effective in producing hexamers with relatively high affinity for a variety of antibody and receptor ligands. Such strategies are now in rather wide use. Thus, it is germane to characterize the features of pooling as a search strategy and compare it to alternatives based on mutation, recombination and selection, or mixtures of all these approaches.

The most obvious strength of pooling is that it allows a large *in vitro* library to be screened simultaneously via partitioning into subsets which share specific subsequences. Conversely, in the first two cycles of screening, the counts bound on any one pin reflect the average behavior of the mixture of sequences at that pin, and hence may reflect high affinity of one member of the mixture, or modest affinity of many members of the mixture. Therefore, an iterative process which sequentially picks the best pin-pool at each step is not guaranteed to find the true optimal sequence among the 64 million hexapeptides. Indeed, iterative pooling to achieve a ''pool optimal'' sequence does not guarantee that the pool-optimal hexamer is itself even a local peak on the molecular fitness landscape.

In order to explore the implications of fitness landscape structure for pooling strategies, we utilized the NK model, fixing $N=6$ to represent hexamers, $A=20$ to reflect the 20 possible amino acids, and allowed K to vary from 0 to its maximum value, 5. We considered the case where the K sites affecting any site are drawn at random among the N, and the case where the K sites are the left and right flanking neighbors. In this latter case we modeled peptides as closed ring structures so that each amino acid was affected by K other sites.

We modeled two alternative pooling strategies. In the first, method 1, the first cycle of 400 pools were fixed in positions 3 and 4 and random at the remaining positions, and the pool with the highest average fitness—modeling the best pin—was selected. On the second cycle, 400 pools were fixed in positions 2 and 5 and the fittest pool chosen, thus fixing positions 2, 3, 4 and 5. On a final third cycle, the remaining 400 possibilities, each a unique sequence fixed at all six positions, were tested and the fittest chosen. This process yields a ''pool optimum''. In a second pooling strategy, method 2, mimicking that used by Geysen, the initial cycle treated 400 pools fixed at positions 3 and 4, and picked the fittest pool. The second cycle, however, considered the 20 possible trimers fixed at positions 2, 3 and 4 plus the 20 possible trimers fixed at positions 3, 4 and 5. The best of these 40 pools were selected, and via further iterations, the remaining three positions in the model hexamer were fixed. At each pooling cycle, the maximum diversity of model hexamers in any single pool —hence pin—was limited to 1000 for computational convenience.

Figures 4(a)–(f) and 5(a)–(f) show the model results for the distribution of fitnesses over the 400 sub-pools at each of the three cycles in method 1 and over the 400 or 40 sub-pools at each of the five cycles in method 2. These distributions model the expected distribution of labeled estrogen receptor bound to the set of pins at each pooling cycle.

Figures 4 and 5 reveal interesting general features. When $K=0$ landscapes are smooth and single peaked, successive pooling cycles yield fitness distribution which shift successively to higher fitness, modeling successively higher average bound counts to the pins. As K increases, this pattern changes. For $K=3$ for example, the fitness distributions of the first and second cycles nearly overlap. When K increases to its maximal value, 5, fitness landscapes are fully random, and the peaks of all early pool cycles nearly coincide. Similarly, the fitness of the best candidate at each cycle of pooling and screening improves more or less uniformly on $K=0$ landscapes, but as K increases and landscapes become more rugged, the best candidate at each cycle improves slowly over early cycles and more rapidly over later cycles.

Published data are not available to compare the predictions of Figs 4 and 5 with real molecular data. However, unpublished data from Geysen suggests that, using method 2, successive peaks of the pool-pin distribution shift to the right over the five cycles of pin screening. If correct, Geysen's data, and further unpublished data from Houghten showing successive shifting to higher mean affinities over pooling cycles (Geysen *et al*., 1987; Houghten *et al*., 1991), would be inconsistent with fully random landscapes among hexamer peptides, and suggest that landscapes are at least modestly correlated.

Figures 6(a) and (b) compare the fitness of pool optima under methods 1 and 2 for linear and random K connections. In general, method 1 outperforms method 2 more profoundly as landscape ruggedness increases.

In order to test whether pool optimal sequences were even local peaks on the corresponding model fitness landscapes, we carried out adaptive walks from each pool optimum. To do so the 19×6 one-mutant variants of the pool optimum were generated, and adaptive walks uphill via all fitter variants were carried

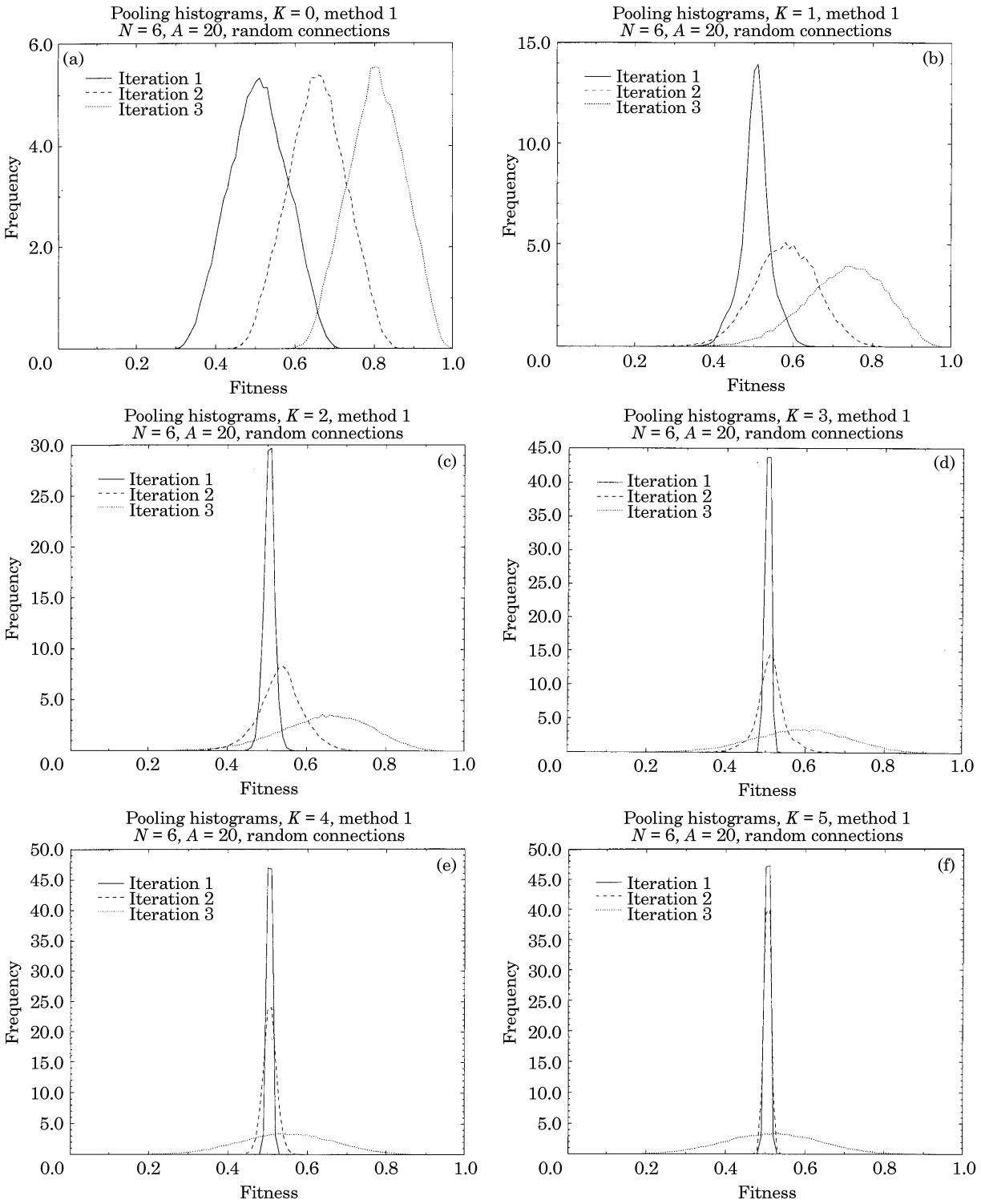


FIG. 4. Fitness histograms under pooling iterations for method 1 with (a) $K=0$, (b) $K=1$, (c) $K=2$, (d) $K=3$, (e) $K=4$, and (f) $K=5$, with random connections.

out until each lineage terminated on a local peak. Thus, each pool optimum might itself be a local peak, or might yield a branching walk to one or many alternative local peaks. We characterized several features of these walks: the probability that a pool optimum was a local peak, mean and maximum number of steps from the pool optimum to local peaks, the mean and maximum number of peaks found and

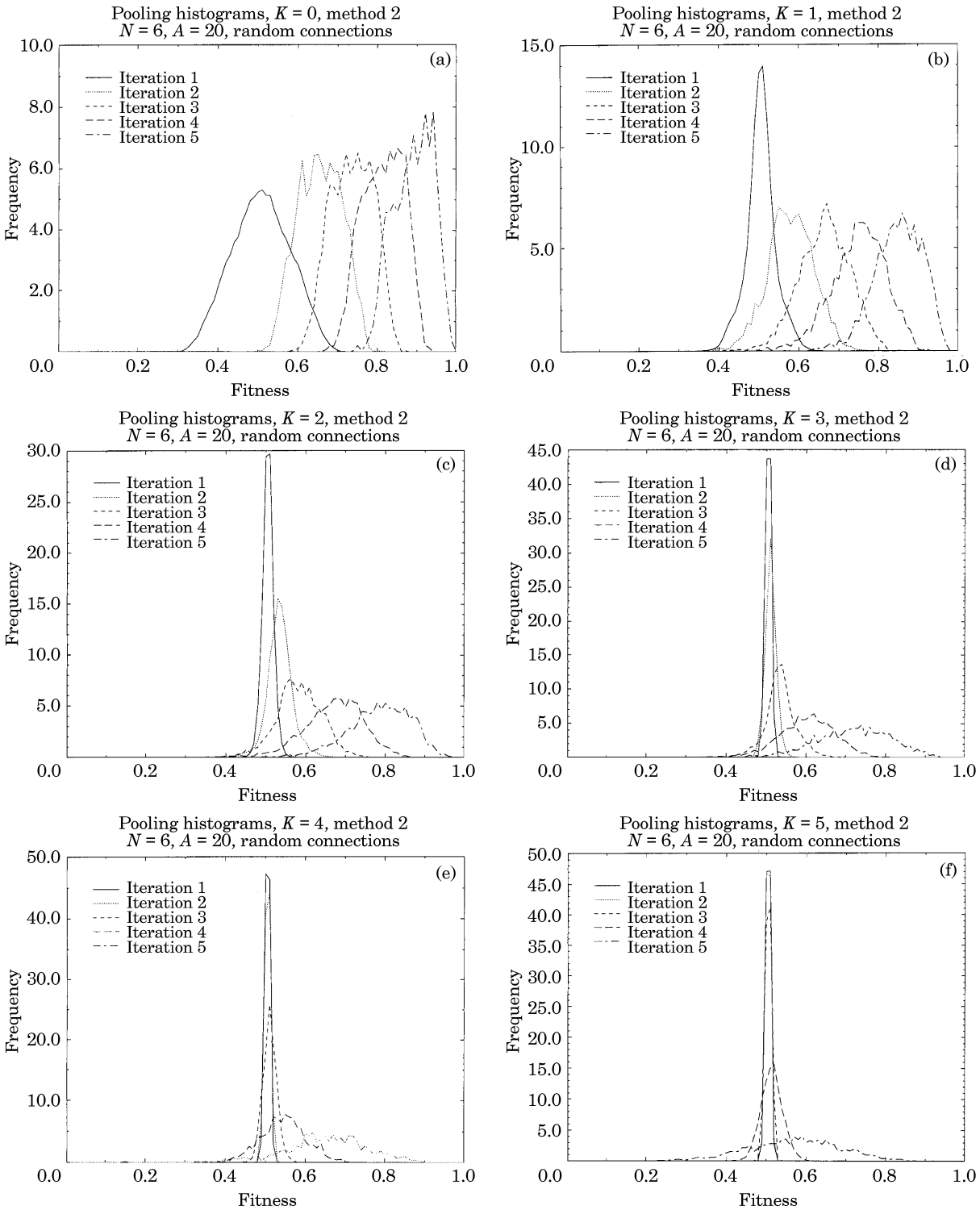


FIG. 5. Fitness histograms under pooling iterations for method 2 with (a) $K=0$, (b) $K=1$, (c) $K=2$, (d) $K=3$, (e) $K=4$, and (f) $K=5$, with random connections.

the number of directions uphill at each step. The results are shown for method 2 with random connections in Figs 7, 8, 9 and 10. Results are similar for linear connections and method 1.

Figure 7 shows that pool optima are local optima about 50% of the time on smooth landscapes, and are increasingly likely to be local optima as landscape ruggedness increases. Figure 8 shows that typical walks

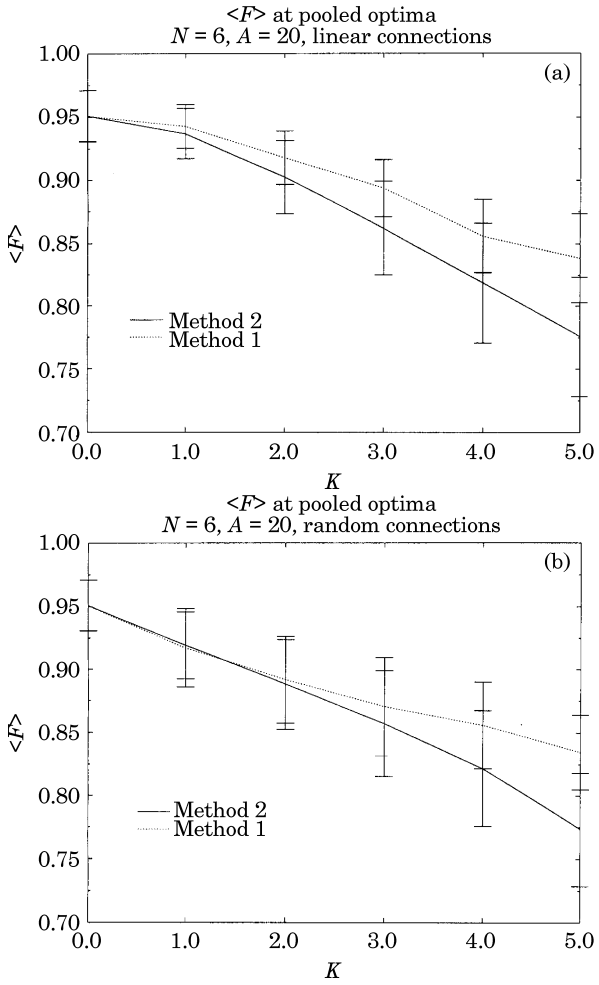


FIG. 6. Expected fitnesses obtained by pooling as a function of K under (a) linear connections and (b) random connections.

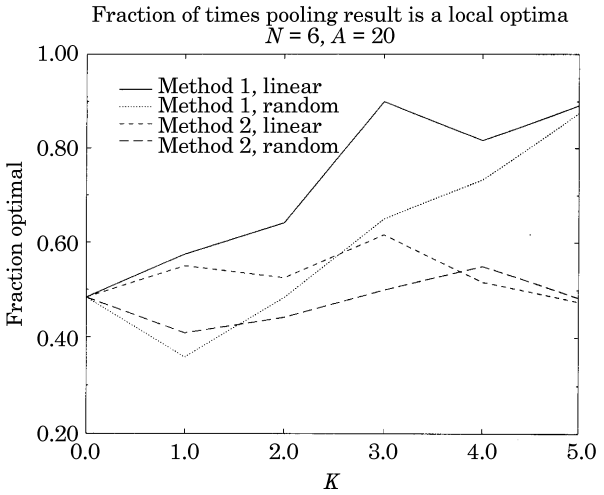


FIG. 7. Fraction of times the pool result is a local optimum vs. K under methods 1 and 2 for linear and random connections.

to local peaks are short, averaging less than a single step due to cases where the pool optimum is a local optimum. Occasionally, however, long walks of up to 14 steps are encountered. Figure 9 shows that, typically, only one or two neighboring local peaks are accessible from the pool optimum. Figure 10 characterizes the dwindling number of directions uphill along walks from pool optima.

It appears to be experimentally established that pool optima are not always local peaks on the fitness landscape. Houghten (Houghten *et al.*, 1991; Houghten 1994, personal communication) has found at least one case where the pool optimum is a local optimum on the hexamer landscape and another case where the pool optimum is not a local optimum. Assessing the frequency with which pool optima are local peaks and the other features predicted in Figs 8–10 is experimentally straightforward. It requires trying all $19 \times 6 = 114$ one-mutant variants of the pooled optima and carrying out adaptive walks via one-mutant variants to the nearby peaks. Our simulations strongly suggest that better peptides may often be found by one-mutant adaptive walks from the pooled optimum rather than ceasing search with the pool optimum alone.

4. Recombination

Recombination between two biopolymers of equal length consists in cleaving each and ligating the "left" end of one with the "right" end of the second. In the case of polypeptides, left and right correspond to 3′ and 5′ ends of single stranded sequences. Molecular recombination has the property that wherever two polymers are identical to one another, they remain identical under the operation of recombination. For example, recombination between the binary strings

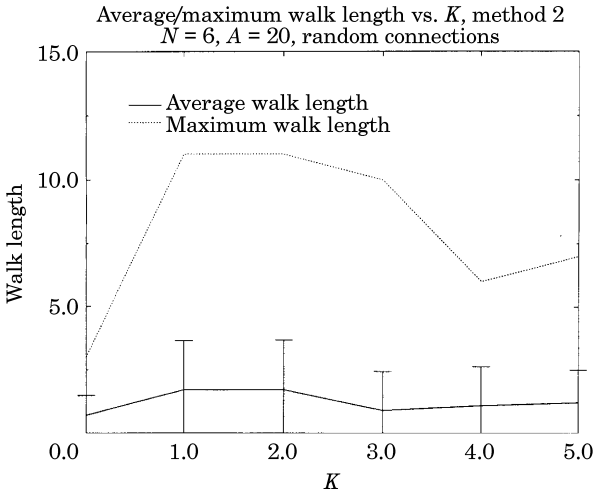


FIG. 8. Expected walk lengths to optima for $K = 0$–5 for method 2 with random connections.

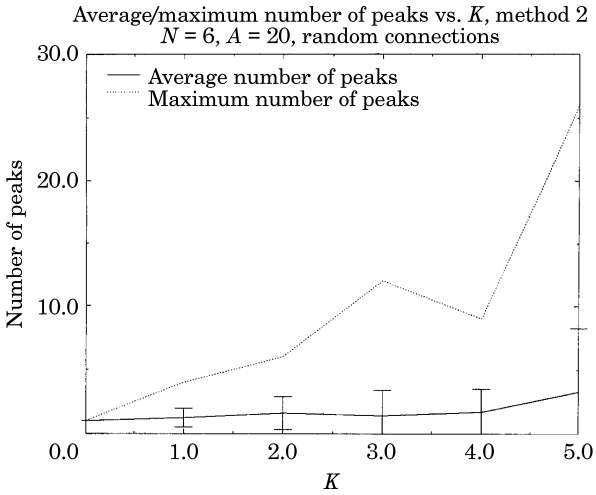


FIG. 9. Expected number of peaks for $K=0$–5 for method 2 with random connections.

(0000) and (0011) cannot alter the 0 values in the first two positions where the two sequences are identical, but can, in principle, yield novelty in the final two positions, such as sequences (0001) and (0010). In a high-dimensional sequence space, recombination is a search process restricted to the subspace where the parental sequences differ, but lying in the hyperplanes where parental sequences are identical.

Previous work using the *NK* model suggests that cycles of recombination followed by hill climbing via one-mutant neighbors to local peaks was useful in locating subregions of rugged landscapes where very high fitness peaks are concentrated (Kauffman, 1993). In contrast, if such subregions do not exist, recombination did not appear to be a useful search strategy. The existence of subregions where high peaks congregate in the *NK* model only occurs for relatively low values of *K*. Here the correlation length of the fitness landscape can span the entire space, implying that the space as a whole is non-isotropic. When this is true, high peaks carry mutual information about the locations of other higher peaks with ever larger "drainage" basins from which those peaks are attainable via adaptive walks. Conversely, when *K* is a large fraction of *N* landscapes are rugged and isotropic, no special subregions exist, and recombination between isotropic subregions appears not to help search.

In order to test the efficacy of recombination after pooling, we modified methods 1 and 2. At each step, rather than utilizing only the fittest pool-pin, we retained the two fittest pool-pins. Each was used over the remaining pool screening cycles, thereby generating eight terminal candidate pool optima for method 1 and 32 for method 2. Interestingly, many of the eight or 32 pool candidates are often fitter than the pool optimum which results from choosing the best pin at each iteration. This suggests that such an expanded

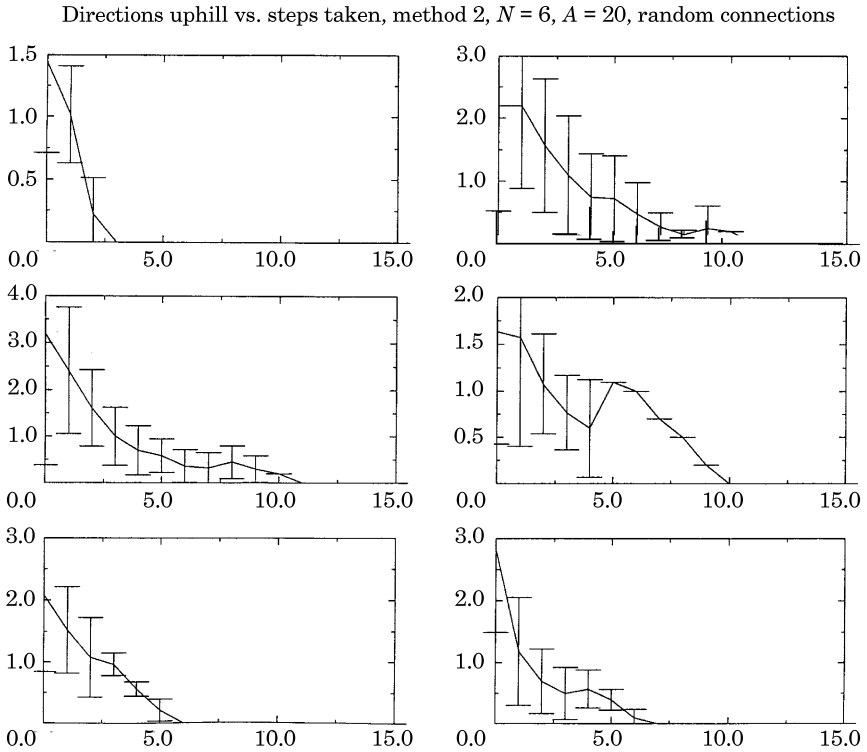


FIG. 10. Expected number of directions uphill as a function of steps uphill for $K=0$–5 for method 2 with random connections.

pooling protocol may be useful. To test the efficacy of recombination alone as a search strategy, we created all possible recombinants at all five possible internal cleavage sites in our model hexamers. The results consistently yield polymers which were less fit than the parental candidate pool optima. Recombination alone does not help find fitter peptides.

In order to test whether recombinant hexamer polymers lie in subregions of sequence space from which they might climb by adaptive walks to higher peaks than the parental pool optimal sequences, we then carried out adaptive walks via fitter one-mutant variants, following a single pathway upwards from each polymer at each step choosing the fittest mutant. In addition, we carried out similar adaptive walks from pool optimal hexamer polymers. The results for random connections and method 2 [Fig 11(a)–(c)] show that, when followed by adaptive walks, recombination does aid the search for higher peaks for all values of K among such hexamers.

Experimental tests of the efficacy of recombination alone or recombination followed by adaptive hillclimbing are again straightforward. Our numerical results suggest that recombination alone is unlikely to yield fitter peptides but recombination followed by hillclimbing should be useful.

5. Single and Double Mutants

Landscape ruggedness makes predictions about the probability of finding fitter variants at any given mutation distance from an initial polymer. For the $K=0$ Fujiyama landscape and an initial polymer located at the single, hence global, peak, no polymer at any distance is fitter, and typically, those at a one-mutant distance are fitter than those at a two-mutant distance. Conversely, for fully random landscapes, a polymer on a local peak with respect to one-mutant variants may still be less fit than some two mutant variants. Beyond the one-mutant range from a local peak, the density distribution of fitter mutants is independent of search distance. On rugged, multi-peaked, but correlated landscapes, the probability of finding fitter variants of a highly fit polymer decreases with search distance.

In order to test the implications of the NK family of landscapes for these properties, we located single pool optimal candidates by methods 1 or 2, then generated all 1 mutant and 2 mutant variants of the pool optima. In Fig. 12(a)–(c) we show the fitness distribution of one- and two-mutant variants for $K=1$, $K=3$ and $K=5$ landscapes. On correlated landscapes $K<5$, the mean of the one-mutant distribution is higher than the two-mutant distribution. This reflects the correlation

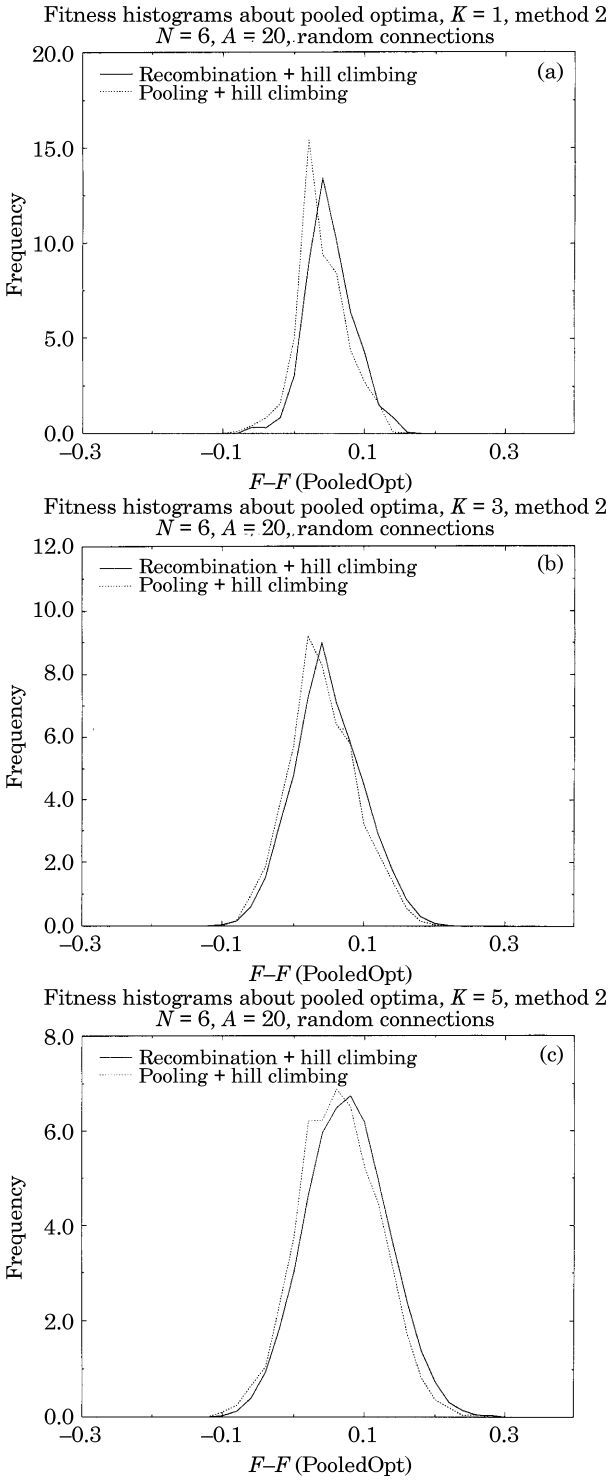


FIG. 11. Comparison of fitness distributions about the pool optimum under pooling with hill climbing and recombination and pooling with hill climbing alone under method 2 for random connections for (a) $K=1$, (b) $K=3$, (c) $K=5$.

structure of the landscape, points further away are less likely to be correlated with the high pooled fitness. In the extreme, $K=5$, landscapes are nearly fully random

and fitness is independent of search distance except if the pool optimal polymer happens to be a local peak; hence the one-mutant and two-mutant distributions nearly coincide.

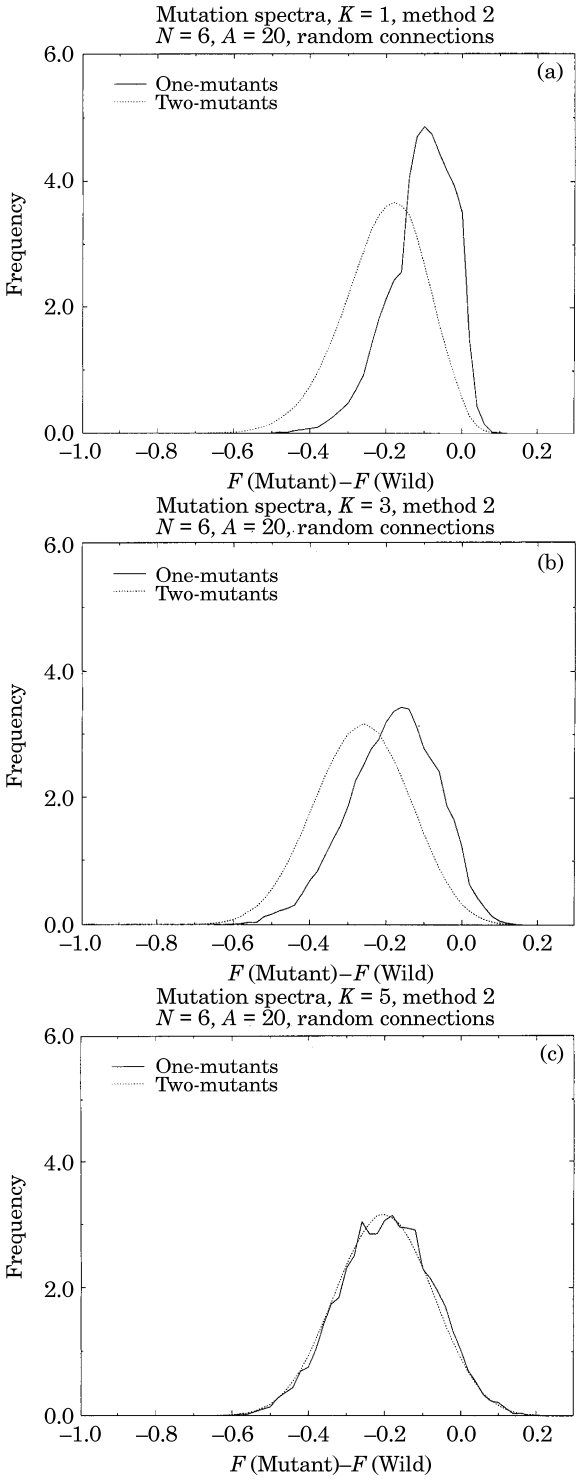


FIG. 12. Fitness distributions of one- and two-mutants about pooled optimum under method 2 with random connections for (a) $K=1$, (b) $K=3$, and (c) $K=5$.

In Fig. 13 we summarize typical one-mutant fitness distributions about the pooled optima for various K. The histograms have a marked K dependence. The mean decreases with K while the variance increases. As K increases the histograms become more Gaussian reflecting the decreasing correlation structure. For all values of K, one-mutant variants fitter than the pool optimum are found. This simple experiment performed on real molecular landscapes would yield information on the ruggedness of the landscape.

It is often the case with evolved proteins that two independent and fitter one-mutant variants of an initial protein can be combined to form the double mutant which is even fitter than the additive effects of the two single mutations alone. Figure 14 shows the number of cases of such super-additive double mutations among all double mutants formed from all single mutants of pool optimal polymers based on method 2. The results show that, on smooth landscapes, $K=0$, super-additive double mutants are not found, while for more rugged landscapes due to larger K, super-additive mutants arise with modest frequency.

The fact that super-additive mutants arise in evolved polymers, and have also been found among the double mutants of pool optimal hexamers generated by method 2, suggests that hexamer landscapes cannot be too smooth. In terms of the *NK* model, K must be 1 or greater. Testing the frequency distribution of super-additive mutants experimentally again should give clues to the structure of molecular fitness landscapes.

6. Discussion

Effective development of applied molecular evolution to generate and identify useful polymers will require sophisticated understanding of the structure of molecular fitness landscapes and means to optimize search upon them. At present, neither the structure of such landscapes, nor optimal search strategies, are understood. The two problems are interrelated, for optimal search methods depend upon landscape structure. To take a trivial example, were landscapes Fujiyama like, with a single peak, greedy hill climbing from any point would suffice. On fully random landscapes, no search strategy helps.

The *NK* model has been useful in fitting features of maturation of the immune response, suggesting that the *NK* family of landscapes are at least a useful first model of the statistical features of molecular fitness landscapes, (Kauffman *et al*., 1988; Kauffman, 1993). While better data and theory will yield more refined

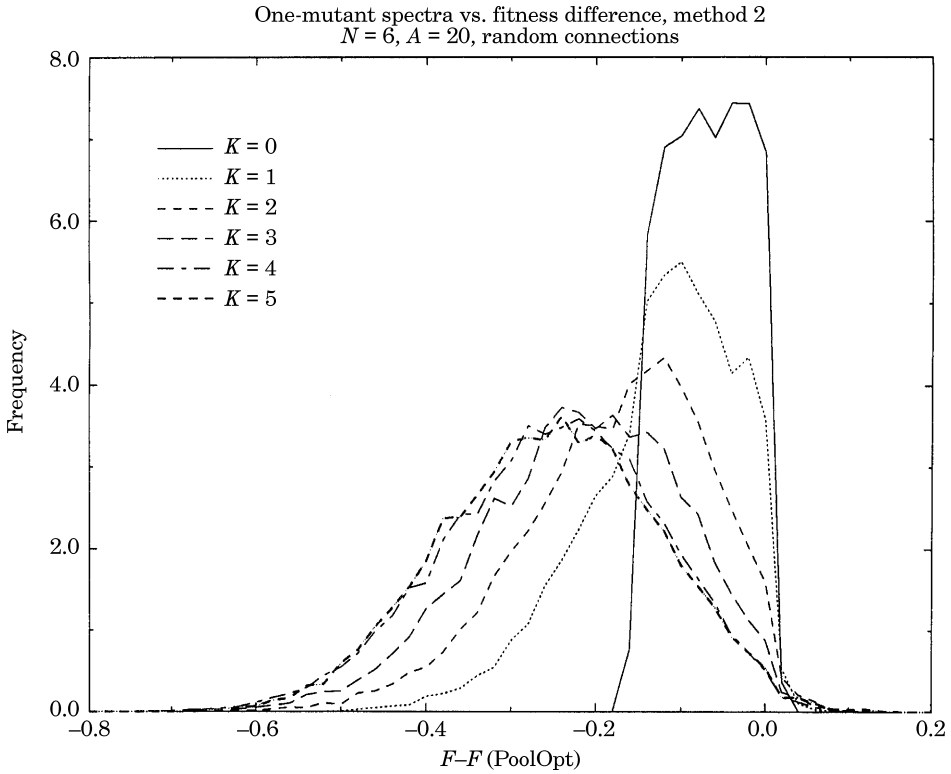


FIG. 13. Fitness distribution vs energy difference from pooled optima for all one-mutants under method 2 with random connections.

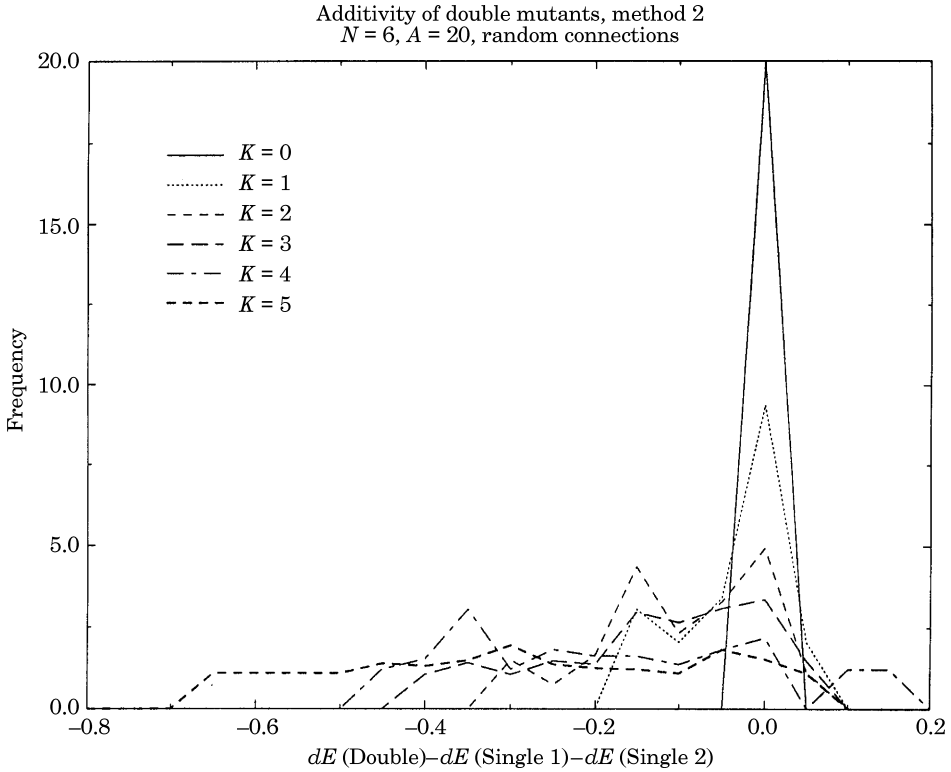


FIG. 14. Probability of additivity of double mutants about a pooled optima under method 2 with random connections. Density where $dE(\text{Double}) - dE(\text{Single1}) - dE(\text{Single2}) > 0$ is super-additive while density where $dE(\text{Double}) - dE(\text{Single1}) - dE(\text{Single2}) < 0$ is sub-additive.

pictures of molecular fitness landscapes, the results reported above are useful in at least three respects:

- The results make suggestions about optimization among hexamer peptides. Thus, the results suggest that often pooling does not even locate local peaks. Hill climbing from pool optima to local peaks should be simple and effective. In addition, expanding standard pooling to include the top few pins at each cycle appears useful. Furthermore, the results suggest that recombination among several pool optima candidates, followed by hill climbing via mutation and selection from those recombinants, yields peptides residing on higher peaks that hill climbing from pool optima candidates without recombination. Thus, use of recombination to find good subregions of hexamer space seems useful and is likely to extend to search in the larger sequences spaces corresponding to longer polymers.
- These predictions of the *NK* model indicate simple experiments against which the landscape model may be tested. If pool optima are often not even local peaks, that is directly testable by carrying out mutation and selection from such pool optima. The statistics of walk lengths to local peaks, number of peaks accessible in the vicinity, and dwindling numbers of directions uphill along such walks are directly testable. If recombination alone among several pool optima candidates typically yields polymers less fit than the parental polymers, that too is directly testable. Further, if hill climbing from such recombinant polymers typically locates better peaks than hill climbing from pool optimal candidates, that prediction is testable. Finally, the distribution of super-additive mutants can be assessed.
- These predictions, and the capacity to test each, make it clear that all these features reflect the structure of the underlying molecular fitness landscape and the search strategies used to find polymers of interest. The *NK* model is merely a first statistical model of molecular fitness landscapes. Improved data and theory will lead to improved pictures of such landscapes. The kinds of predictions made above would all be verified were they based on an adequate model of molecular landscapes and search upon them. Thus, the feature and predictions specified above are among the *desiderata* of an adequate theory. Additional features of importance include the presence or absence of one or more "consensus" subsequences, typically found experimentally among small peptides, and the distribution of extended ridges or isolated islands of sequences with a desired property in sequence space.

It is interesting that, even with the sparse data available, we can already reach conclusions. From the fact that, over pooling cycles, the observed distribution of bound counts over the pins shifts progressively to higher counts, we can conclude that hexamer peptide landscapes are not fully random, and must be moderately correlated. From the existence of super-additive double mutants, we can conclude with moderate confidence that landscapes are unlikely to be extremely smooth and single peaked. If one had to make an informed bet based on the *NK* model of hexamer peptides, peptide landscapes correspond roughly to $K=3$ landscapes, rugged but correlated. In turn, this conclusion, were it valid, would make predictions about structure function relations, for it implies that each amino acid in a hexamer makes a contribution to binding which is influenced by about three other amino acids in that peptide. Better models of fitness landscapes should yield better insight into such structure function relationships.

REFERENCES

Amitrano, C., Peliti, L. & Saber, M. (1989). Population Dynamics in a spin glass model of chemical evolution. *J. molec. Evol.* **29,** 513–525.

Bartel, D. P. & Szostak, J. W. (1993). Isolation of new ribosomes from a large pool of random sequences. *Science* **261**, 1411–1418.

Beaudry, A. A. & Joyce, G. F. (1992). Directed evolution of an RNA enzyme. *Science* **257,** 635–641.

Derrida, B. (1981) Random energy model: an exactly solvable model of disordered systems. *Phys. Rev. B* **24,** 2613–262.

Derrida, B. & Peliti, L. (1991). Evolution in a flat fitness landscapes. *Bull, math. Biol.* **53,** 355–382.

Eigen, M., McCaskill, M. & Schuster, P. (1989). The molecular quasispecies. *Adv. Chem. Phys.* **75,** 149–263.

Ellington, A. D. & Szostak, J. W. (1990). *In vitro* selection of RNA molecules that bind to specific ligands. *Nature, Lond.* **346,** 818–822.

Fontana, W. & Schuster, P. (1989). Physical aspects of evolutionary optimization and adaption. *Phys. Rev. A* **40,** 3301–3321.

Fontana, W., Griesmacher, T., Schanbl, W., Stadler, P. F. & Schuster, P. (1991). Statistics of landscapes based on free energies, replication, and degradation rate constants of RNA secondary structures. *Methods Chem.* **122,** 795–819.

Geysen, H. M., Rodda, S. J., Mason T. J., Tribbick, G. & Schoofs, P. G. (1987). Strategies for epitope analysis using peptide synthesis. *J. Immun. Methods* **102,** 259–274.

Houghten, R. A., Pinilla, C., Blondelle, S. E., Appel, J. R., Dooley, C. T. & Ceurvo, J. H. (1991). Generation and use of synthetic peptide libraries for basic research and drug discovery. *Nature, Lond.* **354,** 84–86.

Joyce, G. F. (1992). Directed molecular evolution. *Sci. Am* **267 (6),** 48–55.

Kauffman, S. A. (1992). Applied molecular evolution. *J. theor. Biol.* **157,** 1–7.

Kauffman, S. A. (1993). *Origins of Order*. Oxford: Oxford University Press.

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. & WEINBERGER, E. D. (1989). *J. theor. Biol.* **141,** 211–245.

KAUFFMAN, S. A., WEINBERGER, E. D. & PERELSON, A. S. (1988). Maturation of the immune response via adaptive walks on affinity landscapes. In: *Theoretical Immunology I: Sante Fe Institute Studies in the Sciences of Complexity* (Perelson, A. S., ed.) pp. 349–382. New York: Addison Wesley.

MACKEN, C. A. & PERELSON, A. S. (1989). Protein evolution of rugged landscapes. *Proc. natn. Acad. Sci. U.S.A.* **86,** 6191–6195.

MAYNARD SMITH, J. (1970). Natural selection and the concept of a protein space. *Nature, Lond.* **225,** 563–566.

STADLER, P. F. (1992). Correlation in landscapes of combinatorial optimization problems. *Europhys. Lett.* **20,** 479–482.

STEIN, D. L. (ed.) (1992). *Spin Glasses and Biology*. Singapore: World Scientific.

WEINBERGER, E. D. (1991). Local properties of Kauffman's NK-model, a tunably rugged energy landscape. *Phys. Rev. A* **44,** 6399–6413.

WRIGHT, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics* **1,** 356–366.