IIASA
Science for
Global Insight

International Institute for Applied Systems Analysis
Schlossplatz 1 • A-2361 Laxenburg • Austria
Telephone: (+43 2236) 807 342 • Fax: (+43 2236) 71313
E-mail: publications@iiasa.ac.at • Internet: www.iiasa.ac.at

Interim Report **IR-99-052**

Correlation Analysis of Fitness Landscapes

Hannelore Brandt (brandt@iiasa.ac.at)
Ulf Dieckmann (dieckman@iiasa.ac.at)

Approved by

Gordon J. MacDonald (macdon@iiasa.ac.at)
Director, IIASA

December 1999

Contents

Abstract

Fitness landscapes underlie the dynamics of evolutionary processes and are a key concept of evolutionary theory. Recent research on molecular folding and on evolutionary algorithms has demonstrated that such landscapes are also important for understanding problems of chemistry and of combinatorial optimization. In these cases free energy or cost functions are used instead of biological fitness functions defined on genotypes.
However, the image of a three dimensional landscape with many peaks and valleys turns out to be misleading. Genotypes tend to differ in numerous characteristics, resulting in multidimensional fitness landscapes. Properties of these landscapes are very different from those of low dimensional ones. The main intention of this study is to investigate how these features affect the duration of adaptive walks on such landscapes. For this purpose we focus on the *Travelling Salesman Problem* (TSP), which amounts to finding the shortest tour visiting a given set of locations. By comparing theoretical predictions for the duration of adaptive walks to the actual waiting times observed for an evolutionary algorithm we demonstrate that a sufficiently fine-grained correlation matrix succeeds in capturing essential structural features of the TSP fitness landscape. To test the performance of correlation-based predictions for a class of fitness landscapes with varying degree of neutrality, we have analyzed evolutionary waiting times on *NKp fitness landscapes*. We show that for low degrees of neutrality, correlation statistics again prove to be an excellent basis for predicting waiting times, while for very high degrees of neutrality, a population's drift along neutral networks turns out to require incorporation of additional information on network topologies.

About the Authors

Hannelore Brandt
Institute of Mathematics
Strudlhofgasse 4
A-1090 Vienna, Austria

Ulf Dieckmann
Adaptive Dynamics Network
International Institute for Applied Systems Analysis
A-2361 Laxenburg, Austria

Acknowledgment

This article was written at the International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, where the author cooperated with the Adaptive Dynamics Network project and participated in the Young Scientists Summer Program 1999.

Correlation Analysis of Fitness Landscapes

Hannelore Brandt
Ulf Dieckmann

1 Introduction

The notion of a fitness landscape has permeated the analysis of evolutionary processes for more than 60 years (Gavrilets 1997). It has proved to be a powerful metaphor for investigating specific problems in biology, chemistry, computational optimization and even economy. Originally used to facilitate understanding of selection and mutation of biological genotypes, the concept has recently been transferred to the study of abstract genotypes such as sequences of RNA molecules (Fontana *et al.* 1993, Schuster *et al.* 1994, Schuster 1997), solutions of optimization tasks (Stadler and Schnabl 1992), or organizational structures of business firms. All these different entities can be envisaged as genotypes of an evolutionary algorithm and thus share some basic properties: they can be represented by a vector and, in order to measure their performance, they can be assigned a certain fitness value. Whereas biological fitness is defined as the expected number of offspring produced by an individual, in the case of RNA sequences fitness can be taken as the free energy of the folded molecular sequence. For many optimization problems cost functions can be thought of as determining fitness values. Mutation of a genotype then amounts to changing entries of an individual's genotype vector, and selective pressures assign fitter genotypes a higher probability of being taken over to the next generation.

Arranging genotypes in an abstract topological space with each genotype situated next to those which can be reached by a single mutation, and adding one dimension to include the fitness values of genotypes leads to the picture of a fitness landscape. A population of individuals can then be seen as a cloud of points on the surface, with the combined effect of mutation and selection forcing the population to perform a hill-climbing process towards fitness peaks.

For a long time, the notion of rugged fitness landscapes, involving many local peaks separated by fitness valleys, has dominated the discussion of adaptive processes (Kauffman and Levin 1987). However, for many evolutionary processes the intuitive image of a three-dimensional landscape with its emphasis on peaks and valleys (see Figure 1) may be inappropriate. Hill-climbing on such a rugged fitness landscape a population will soon end up at a local peak: selective pressures will prevent it from crossing the surrounding adaptive valleys to reach a higher fitness peak. Yet, biological evolution and evolutionary algorithms do not appear to become entrapped in local fitness peaks as often as this intuitive picture suggests (Schuster 1996).

The reason is that most genotypes differ in much more than two properties, result-

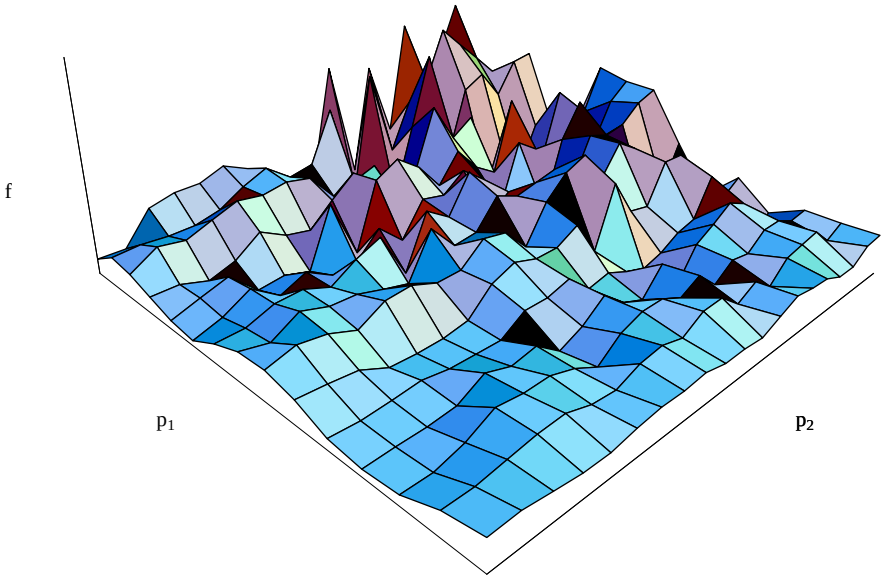

Figure 1: *If genotypes differ in only two properties* (p_1, p_2) *they can be arranged in a two-dimensional space, with each individual next to those which can be reached by a single mutation. Assigning a certain fitness value* f *to all genotypes leads to a fitness landscape. Evolution in areas where the surface is smooth typically results in steady evolutionary change, whereas in rugged regions a population can easily end up at a local peak.*

ing in fitness landscapes with dimensions much higher than three. It turns out that the structural features of these high-dimensional landscapes are very different from those of low-dimensional ones. For many high-dimensional landscapes the problem of being stuck in a local adaptive peak far away from the global optimum might even be non-existent.

From investigating special cases we know that the features of fitness landscapes can vary to a high degree. On landscapes which are rather smooth, the global fitness optimum is reached within a relatively short time, whereas on more rugged landscapes (not necessarily implicating the existence of many local peaks) a population needs to evolve over more generations to attain the highest peak.

It is therefore natural to ask which structural statistics of fitness landscapes determine the durations of adaptive walks. Having identified such statistics would allow for classifying fitness landscapes in such a way that the performance of evolutionary processes can be predicted. Even though some statistics have been suggested for this purpose and are already well-analyzed (Weinberger 1990, Stadler 1992, Stadler 1996, Barnett 1997), presently discussed statistics do not seem to be appropriate for obtaining sufficiently accurate predictions of evolutionary waiting times.

In this study we focus on specific fitness landscapes of well-known problems and investigate the durations (or waiting times) of evolutionary processes on these landscapes. We introduce a new type of correlation statistics, different from those used so far, and show by comparing observed and predicted waiting time distributions that

these statistics might be very useful for understanding, predicting, and classifying evolutionary processes on high-dimensional fitness landscapes.

In Section 2 we examine three different fitness landscapes of the *Travelling Salesman Problem*, obtained by using different mutation operators. Section 3 extends this analysis to the *NKp Model* of epistatic evolution. A summary of our findings and a sketch of open questions resulting from this study is provided in Section 4.

2 The Travelling Salesman Problem

2.1 General information

A salesman who has to visit each city on a given list, knowing the distance between all pairs of cities, will try to minimize the length of his tour. This optimization task is called *The Travelling Salesman Problem* (TSP). Although it has been studied in great detail, there exists no algorithm which finds the shortest tour faster than by complete testing of all possible tours. The time until the optimal solution is found thus grows more than exponentially with an increase in the number of cities, and the TSP therefore belongs to the class of hard integer programming problems with non-polynomial solution times.

In many cases of actual interest, however, the focus is not on detecting the shortest possible tour, but on finding a tour that is sufficiently close to the optimum within a feasible computation time. For this purpose it is convenient to let candidate solutions simply evolve towards better ones. Such algorithms, which make use of basic principles of evolution like mutation and selection (sometimes crossover and recombination are also considered), are called *evolutionary algorithms* and have been established as efficient tools for finding quasi-optimal solutions of many optimization problems (Beasley 1997, Pasemann *et al.* 1999).

In this study the landscape of the TSP is chosen as a benchmark problem because of its canonical genotype-to-fitness map and the attention that it has received in recent studies of fitness landscapes (Stadler and Schnabl 1992, Reidys and Stadler 1999).
For the following examples of the TSP, 25 cities have been distributed randomly according to a uniform distribution over the square $[0, 327]^2$. Every tour starts in City 1, visits each other city once and ends again in City 1.

The biological terms, taken to describe evolutionary processes, are used as follows in the context of the TSP:

- Genotypes

  Each possible tour corresponds to a genotype; its vector representation is given by the sequence of cities. For every genotype, the length of the resulting tour is calculated according to a two-dimensional distance table involving all pairs of cities. For 25 cities, the total number of different genotypes is 24!, which is of order 10^{23}. Because of this extremely large number of possibilities a strategy of testing all tours in order to find the best one is infeasible.

- Fitness

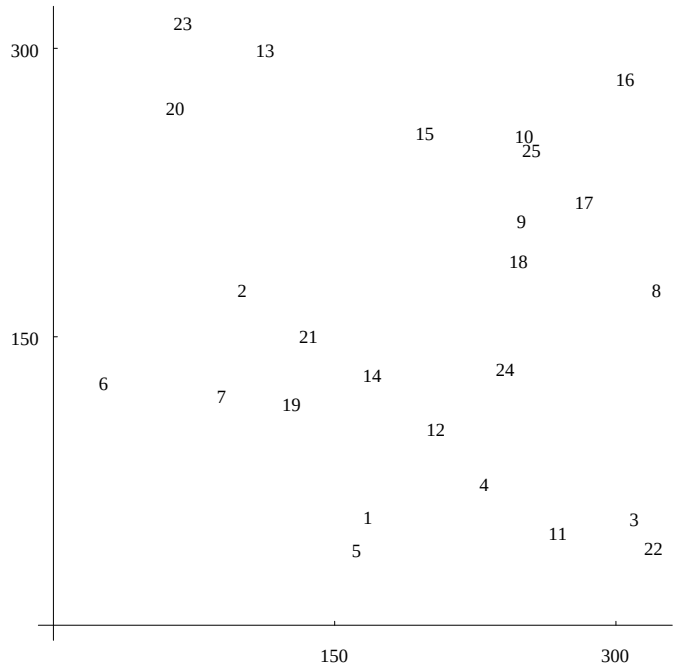

Figure 2: *The 25 cities of a sample TSP.*

The fitness f of a specific genotype g is given by

$$f(g) = \frac{1}{l(g)}$$

where $l(g)$ is the length of the genotype g.

- Mutation

  To mutate a genotype, every operator that changes the vector in a way that the mutated genotype is still a possible tour can be considered. Here, three frequently used mutation operators are chosen (Manderick 1997). For each of these mutation operators, two positions within the tour, corresponding to two indices of the genotype vector, are chosen at random.

  - point mutation: the cities at two indices of the vector are swapped. If the fourth and the eight index are chosen, a mutant of the tour $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...]$ is the vector $[1, 2, 3, \mathbf{8}, 5, 6, 7, \mathbf{4}, 9, 10, ...]$.
  - reverse mutation: the order of cities between two indices is reversed. A mutant of the tour $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...]$ is the vector $[1, 2, 3, \mathbf{8,7,6,5,4}, 9, 10, ...]$.
  - remove and reinsert: the city at the first index is taken out and reinserted at the second index. A mutant of the tour $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...]$ is either the vector $[1, 2, 3, \mathbf{5,6,7,8,4}, 9, 10, ...]$ or $[1, 2, 3, \mathbf{8,4,5,6,7}, 9, 10, ...]$, depending on the order in which the two indices were chosen.

- Selection

In every generation, each genotype produces two offspring individuals which are both once mutated. The best third of the union of the old population and the offspring is taken over to the next generation. This process thus keeps the population size (chosen at 15 individuals below) constant. Two individuals that can be transformed into each other by a single mutation are called neighbors.

- Evolutionary waiting times

  Evolutionary waiting times are stochastic variables defined as the number of generations necessary for a population to evolve between two given fitness values. In this work we concentrate on waiting times as these provide crucial statistics of an evolutionary process.

The fitness landscape of the TSP (Stadler 1992, Stadler and Schnabl 1992, Fontana *et al.* 1993, Happel and Stadler 1996) as well as those derived from other well known problems (Schuster *et al.* 1994, Schuster 1997, Barnett 1998) have already been studied in detail. Most of those analyses focus on a specific type of auto-correlation function to describe the structure of fitness landscapes. This function can be defined based on time series obtained from unbiased random walks on a given landscape. During such walks, the genotype of a single individual is repeatedly mutated without considering the fitness of resulting mutants. These auto-correlation functions are utilized to measure the ruggedness of a given fitness landscape. However, as pointed out by Barnett (1998), these auto-correlation functions alone may be inadequate to characterize evolutionary dynamics. Another property of fitness landscapes, their degree of neutrality, also seems to play an important role. Two genotypes are said to be *neutral*, if they have the same fitness. If large networks of such neutral genotypes are embedded in a fitness landscape, a population is to drift along them. Hill-climbing is then confined to the occasional transfer of the population to neutral networks of higher fitness. As shown by Barnett (1998), the degree of neutrality does not necessarily influence the auto-correlation function of a landscape and thus can be tuned independently of a landscape's ruggedness. Understanding the interplay of ruggedness and neutrality is crucial for a characterization of a wide variety of fitness landscapes (Huynen *et al.* 1996).

2.2 Point Mutation

Point mutations of a TSP genotype are realized by choosing at random two different indices of the genotype vector. At these two positions, the entries of the vector are then swapped. This generates an offspring genotype with a tour length different from that of its parent. In the next subsections we investigate, starting with results on the real evolutionary process, the performance of different reduced descriptions of the resulting fitness landscape. The salient question here is which structural statistics of the landscape are essential for understanding and predicting the duration of a population's adaptive walk on this landscape.

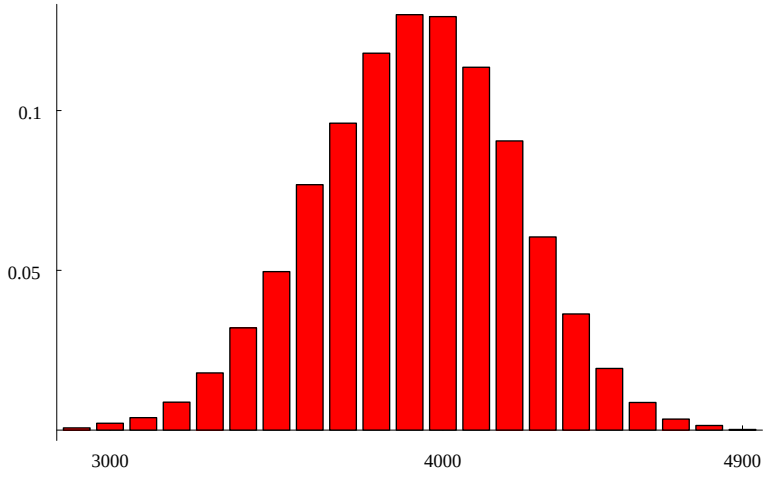

Figure 3: *Length distribution of random TSP tours.*

2.2.1 The Evolutionary Algorithm

Before we focus on the evolutionary waiting times on our specific TSP landscape, we give some basic information concerning the behavior of the utilized evolutionary algorithm.

Randomly produced tours normally have a length between 2900 and 4900. The distribution of these length values is shown in Figure 3.

A population of 15 individuals, each producing two mutated offspring per generation, rapidly tends to climb the fitness landscape, see Figure 4, where the shortest tour length decreases from about 3000 to about 1500 in 200 generations. The best fitness in the population normally remains constant for a number of generations and then suddenly jumps to a higher level. This is a common property of evolutionary algorithms and is referred to as epochal evolution.

The best tour found by all different mutation operators discussed in this paper has a length of 1369 and typically is found within 100-200 generations. This solution certainly is very close to the global optimum of this TSP.

To test the different reduced descriptions of fitness landscapes studied in this paper, it is necessary to obtain statistics of evolutionary waiting times for different fitness intervals. For this purpose, we have chosen initial and final fitness values from the interval 1/5000 to 1/2900; producing random tours with fitness values in this range is relatively easy, and this is a prerequisite for obtaining valid statistical distributions of waiting times. In particular, initial fitness values are chosen $1/5000, 1/4300$ and $1/3600$, and final fitness values $1/4300, 1/3600$ and $1/2900$. To construct the distribution of evolutionary waiting times from a certain initial fitness f_i to a final value f_f we proceed as follows. In generation 0, the population is initialized with a random genotype that has a fitness of approximately f_i. The number of generations necessary until one individual of the population reaches fitness f_f is stored as the waiting time of a run. For a given pair of initial and final values, results of 2000 such runs are combined to obtain the distribution of waiting times; an example is shown in Figure 6.

In the following subsections we consider landscape statistics of increasing com-

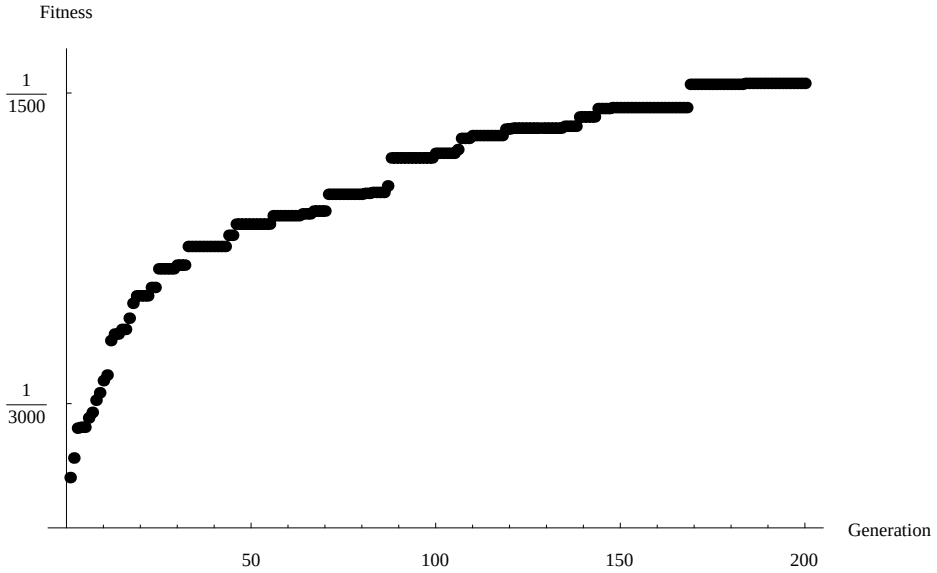

Figure 4: *The evolution of the highest fitness in a population. Periods of constant fitness are interspersed with sporadic jumps, a characteristic property of evolutionary algorithms.*

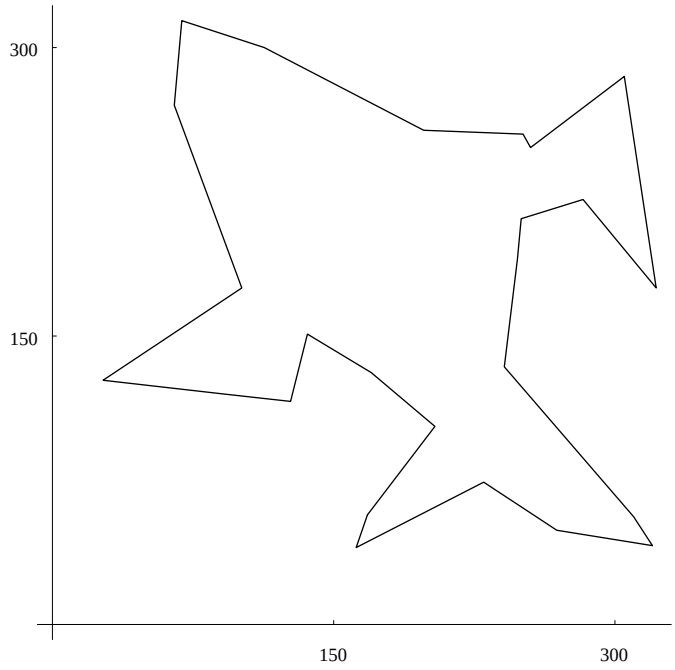

Figure 5: *The shortest tour found for our sample TSP.*

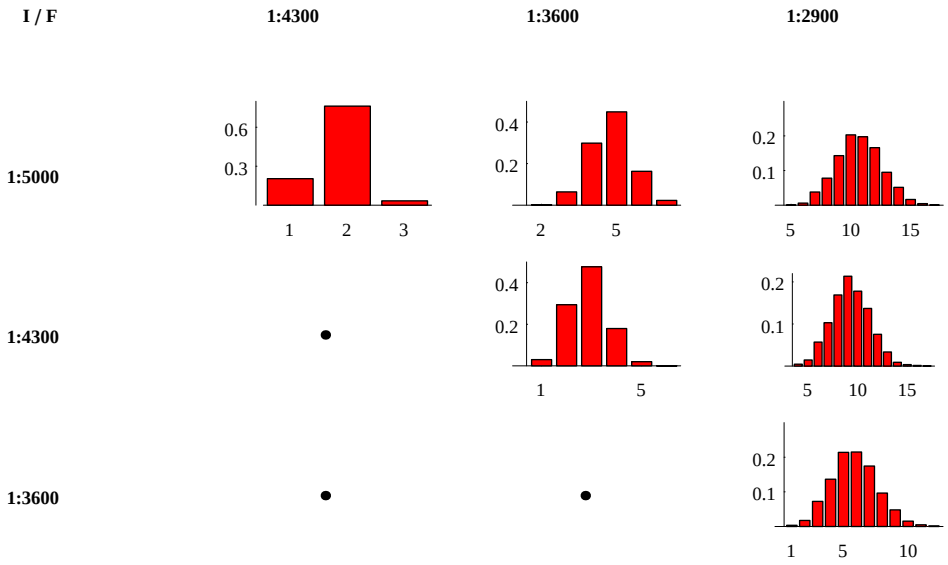

Figure 6: *A matrix of waiting time distributions for 3 initial and 3 final fitness values, based on an evolutionary algorithm that uses point mutations. This graph serves as a target for results derived from different reduced descriptions of fitness landscapes, and allows to assess their absolute and relative performance. Since decreases in population fitness cannot arise in the evolutionary algorithm considered here, waiting times for three pairs of fitness values are infinite; the corresponding distributions are replaced with filled circles.*

plexity and compare their suitability for predicting the actual distributions of evolutionary waiting times.

2.2.2 Percolation

The basic concept of percolation theory is a grid in a multidimensional space with each lattice site being independently filled with probability p (Kesten 1982, Grimmett 1989). If p exceeds a certain threshold, a subset of the filled sites forms a connected infinite cluster that percolates through the entire grid. Cluster statistics have been used to study a wide variety of problems (Sahimi 1994, Stauffer and Aharony 1995).

One of the simplest ways of analyzing a given fitness landscape is to divide all elements of the genotype space into two classes by introducing a certain fitness threshold. Those genotypes with fitness beyond the threshold are in class 1, all others in class 0. All elements of the multidimensional genotype space of a TSP can then be considered as belonging to either of these two clusters with probability p and $1 - p$, respectively. A percolation approximation of a fitness landscape then amounts to (i) considering only the labels 0 or 1, while ignoring actual fitness values, and (ii) assuming that genotypes independently belong to either cluster according to a probability p. The probability p is estimated from a large number of random TSP tours.

The evolutionary algorithm is then imitated as follows: The entire population is

in class 0 initially. In each generation, 300 new individuals (for 25 cities each TSP genotype has $25 * 24/2 = 300$ neighbors under point mutation) are assigned label 0 or 1 according to p. Out of these, 30 offspring (15 individuals produce two offspring each) are chosen. If an offspring in the population belongs to the higher fitness class (class 1), the process stops. Otherwise it continues with the next generation.

To compute the matrix of waiting times, percolation probabilities for the three different final fitness values are calculated. Initial fitness are ignored in this approximation.

In order to calculate the probability $p_{stop}(g)$ for the described process to end in a certain generation g, we first have to define some variables:

A ... total number of different genotypes

N ... number of neighbors for each genotype

m ... number of offspring per generation

p ... probability for one individual to be in the higher fitness class (class 1)

A_1 ... number of all genotypes in class 1 ($\approx Ap$)

The probability p_k to have k neighbors in class 1 out of N possible is then given by

$$\begin{aligned} p_k &= \frac{\binom{A_1}{k}\binom{A-A_1}{N-k}}{\binom{A}{N}} \\ &= \frac{A_1!\,(A-A_1)!\,N!\,(A-N)!}{k!\,(N-k)!\,(A_1-k)!\,(A-A_1-N+k)!\,A!} \\ &= \binom{N}{k}\frac{A_1!\,(A-A_1)!\,(A-N)!}{(A_1-k)!\,(A-A_1-N+k)!\,A!}. \end{aligned}$$

If N and k are relatively small compared to A (which is true for our TSP landscape), this equation is well approximated by

$$p_k \approx \binom{N}{k} p^k (1-p)^{N-k}.$$

Now, m offspring are chosen out of the N neighbors; these are not necessarily different. The probability p^* to have at least one individual of higher fitness in the next generation can now be calculated,

$$p^* = 1 - \sum_{i=0}^{N-1} p_i \left(\frac{N-i}{N}\right)^m.$$

The probability $p_{stop}(g)$ is then given by

$$p_{stop}(g) = (1-p^*)^{g-1} p^*.$$

For our TSP landscape we have $A = 24!\,, N = 300\,,$ and $m = 30$;

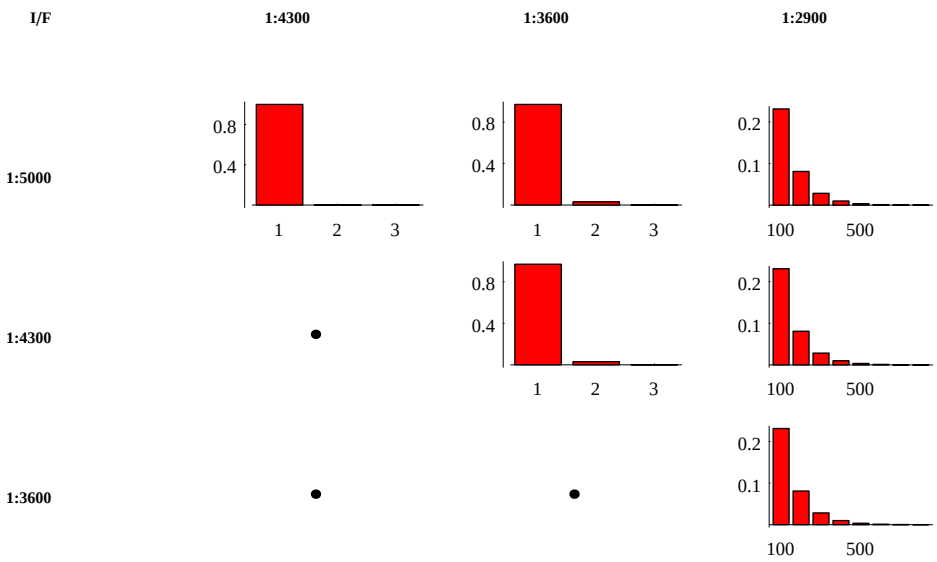

Figure 7: *Waiting time distributions for reaching certain final fitness thresholds as predicted by the percolation approximation. Initial fitness values are ignored in this approximation. Compare the depicted distributions to those in Figure 6.*

It is not surprising that this drastic simplification of the fitness landscape is too coarse. The waiting time statistics in Figure 7 show that the actual evolutionary process is much faster than this reduced description suggests. In the actual process, the probability for choosing a neighbor with fitness above the final value increases over the generations as the population successively attains higher fitness values. The percolation approximation cannot capture this critical effect.

As a next step we thus incorporate a critical landscape feature: the neighborhood of a genotype strongly depends on its own fitness.

2.2.3 Monomorphic correlation

The results obtained for the percolation approximation suggest dividing all genotypes into more than just two fitness classes, with each class having a different distribution of neighbor fitness. The correlation c_{ij} between classes i and j is determined by the probability for a random neighbor of an individual of class i to be in class j. The correlation matrix $C = (c_{ij})$ is then used to define the transition matrix T of a Markov chain that approximates the evolutionary algorithm (Rudolph 1997). In this approximation, the whole population is still considered to reside in the same fitness class, and is thus assumed to be monomorphic. A transition from class i to a higher fitness class j occurs if at least one offspring belongs to fitness class j, but no offspring is in one of the classes higher than j. Transition to lower fitness classes are not possible; the population will therefore remain in the same class if no offspring possesses a higher fitness. To calculate the probability t_{ij} for a transition from i to j we need the following variables, assuming that i and j are fixed:

c_+ ... union of all fitness classes higher than j

c_0 ... class j

c_- ... union of all fitness classes lower than j

p_l ... probability for a random neighbor of an individual of class i to belong to $c_l, l = +, 0, -$

N ... number of neighbors for each genotype

m ... number of offspring per generation

The probability w_{lk} to have k neighbors in c_l is determined by a binomial distribution, $w_{lk} = \binom{N}{k} p_l^k (1-p_l)^{N-k}$. Again, m individuals are chosen out of N neighbors. The probability t_{ij} for a transition from class i to j is the probability to have no offspring in c_+ but at least one in c_0. Thus,

$$t_{ij} = \sum_{n=0}^{N-1} w_{0n} \left(\frac{N-n}{N}\right)^m \quad - \quad \sum_{n=1}^{N} w_{2n} \left(\frac{n}{N}\right)^m .$$

The resulting transition matrix $T = (t_{ij})$ defines a Markov chain and allows us to derive expectations for the moments of waiting time distributions. In particular, the mean number of generations needed for attaining the absorbing state (final fitness) from different starting classes (initial fitness) can be computed. These results are presented below.

Coarse-grained correlation. We begin by introducing four fitness classes, seperated by the three final fitness values used, e.g., in Figure 6. The correlation matrix $C = (c_{ij})$, where c_{ij} is the probability for a random neighbor of an individual of class i to belong to class j, is estimated by randomly mutating random genotypes of class i. The resulting 4×4 transition matrix defines a Markov chain for which waiting time distributions are computed. These turn out to be closer to the actual ones; yet, systematic differences of mean values and variances illustrate the need for further refinement of this correlation-based approach.

Figure 9 shows that even if the population already is in the class next to the final one, producing an offspring the fitness of which exceeds the final fitness threshold takes too much time. The many transitions within classes, leading from the lower bound of a class' fitness range to the upper bound are neglected by only allowing for a small number of classes. This observation suggests to introduce a fine-grained classification of fitness values.

Fine-grained correlation. To improve the predictive accuracy of the correlation approximation we consider a 31×31 correlation matrix. 30 equally spaced fitness thresholds between tour lengths 2900 and 5000 serve as the boundaries of a fine-grained classification. As before, the correlation matrix is obtained by randomly generating neighbors of random genotypes of a given class. In this manner, small changes in fitness values, which can be decisive for the dynamics of an evolutionary algorithm, are no longer neglected.

The fact that most of the probability mass of a correlation matrix is concentrated around its diagonal indicates that neighboring genotypes tend to possess similar fitness values. Yet, the possibility that they belong to distant fitness classes is given. Using the same principles for constructing a Markov chain as described above, Figure 11 shows the resulting transition matrix $T = (t_{ij})$ of the process.

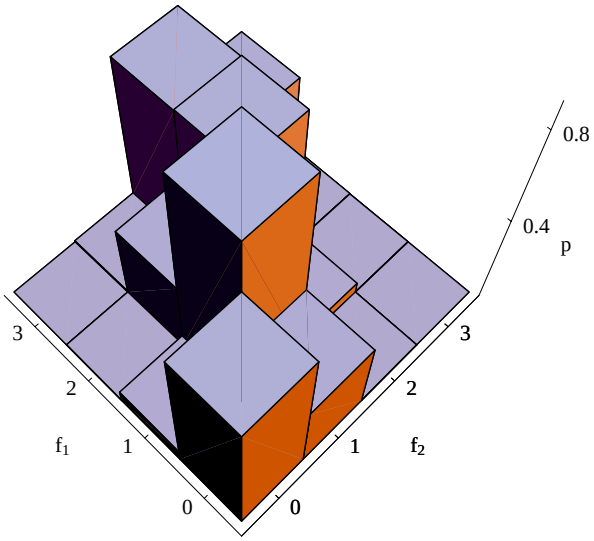

Figure 8: *The correlation matrix for 4 fitness classes. $p(f_1, f_2)$ denotes the probability for a random offspring of an individual of fitness class f_1 to belong to class f_2. The classes range from 0 (lowest fitness) to 3 (highest fitness). These statistics seem to require refinement.*

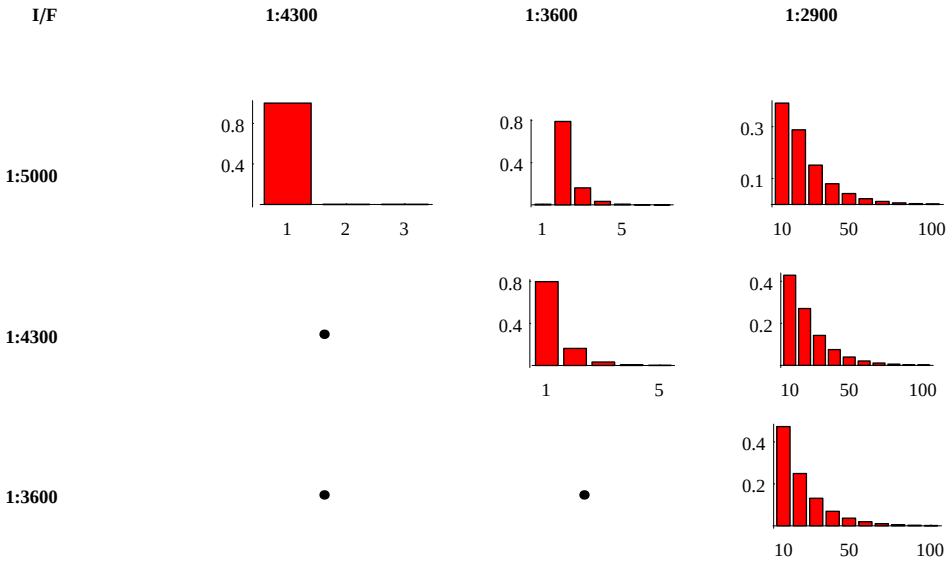

Figure 9: *Waiting time distributions for transitions between given initial and final fitness values as predicted by monomorphic evolution based on a coarse-grained correlation approximation. Comparison of depicted distributions to those in Figure 6 shows that waiting times are overestimated by this approach.*

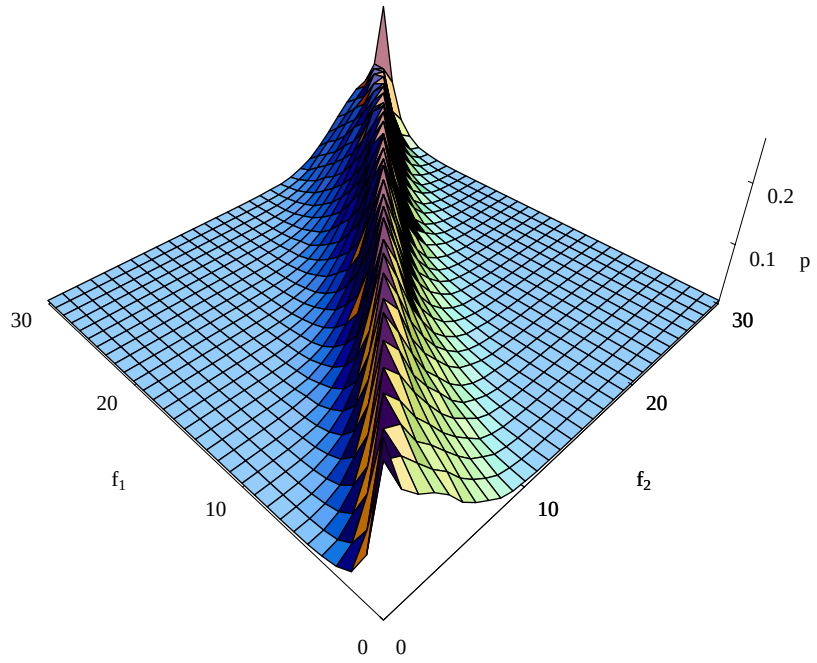

Figure 10: *The correlation matrix for 31 fitness classes. $p(f_1, f_2)$ denotes the probability for a random offspring of an individual of fitness class f_1 to belong to class f_2. The classes range from 0 (lowest fitness) to 31 (highest fitness). The figure shows that for all classes the neighbors of a given genotype tend to have the same or a similar fitness value as that genotype.*

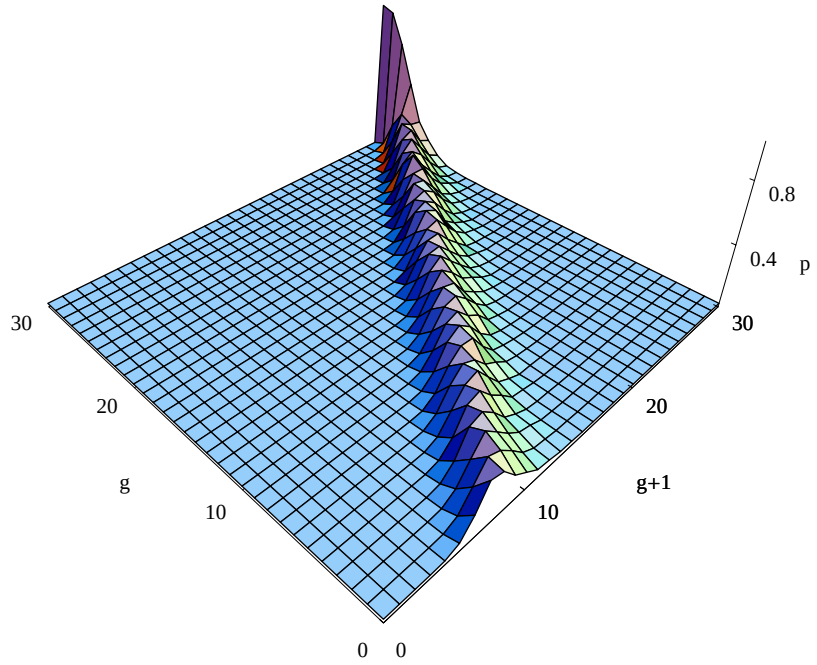

Figure 11: *The transition matrix of the Markov chain that provides a correlation-based approximation of the evolutionary algorithm. A population at generation g jumps to higher fitness classes in the next generation with probability p. Class 31 is called an absorbing state; if the population reaches it, the process stops.*

The fine-grained correlation matrix provides a detailed summary of the adjacency relations between the different fitness classes; the neighborhood structure for different TSP genotypes should therefore be described with sufficient accuracy. We thus might expect that the waiting time distributions derived from this simplified process are a close match to the actual ones. And, indeed, the fine-grained monomorphic correlation approximation is the first approach presented here that succeeds in capturing many of the qualitative and quantitative features of the evolutionary algorithm as summarized in Figure 6. The results presented in Figure 12 therefore underline that a fine-grained correlation matrix as defined above carries salient information about a fitness landscape's structure.

Although correlation approximation of the fitness landscape results in predictions that are not far away from observations on the actual evolutionary algorithm, it is interesting to ask why the simplified process is always a bit faster than the evolutionary algorithm itself. A detailed investigation of fitness distributions for all generations of the evolving population reveals that in each generation fitness values of genotypes in the population are spread over a wide range of classes. For this reason, the assumption of monomorphism (implying that all individuals of the population belong to a single fitness class) is inaccurate. Instead of assuming transitions of the whole population's fitness from one class to another, individuals in the population need to be envisaged as belonging to different fitness classes.

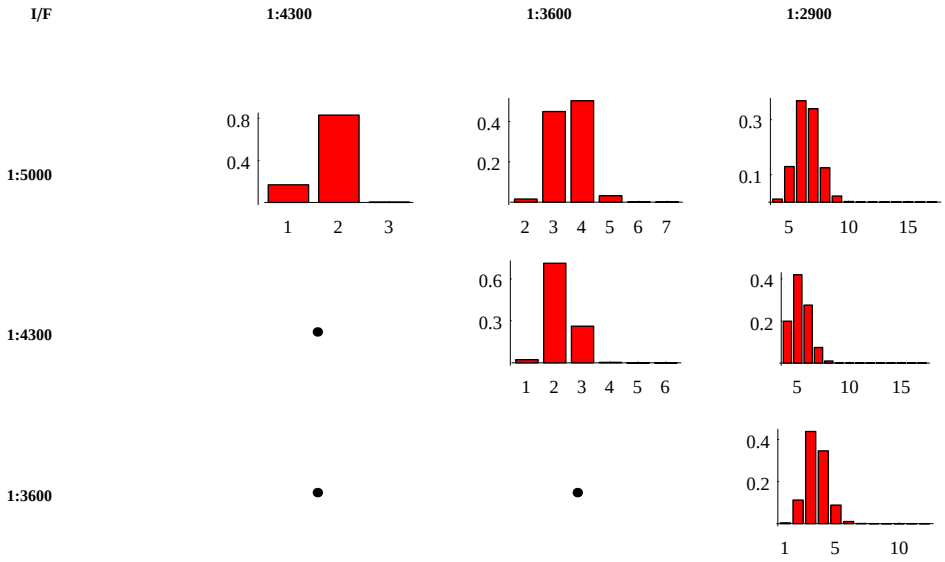

Figure 12: *Waiting time distributions as predicted by monomorphic evolution based on a fine-grained correlation approximation. Comparing these distributions with those in Figure 6 shows that a relatively good approximation of the actual evolutionary algorithm has been achieved.*

2.2.4 Polymorphic correlation

In this subsection we further improve our approximation of the evolutionary algorithm of the TSP. We use the same statistics as before, namely the 31×31 matrix of correlation probabilities for each pair of fitness classes. However, instead of considering only one class that represents the fitness of the whole population, different individuals of the population can now belong to different fitness classes in each generation. The population's state in one generation is no longer a specific fitness class, but is given by a frequency distribution over all 31 possible fitness classes. In other words, we allow the population to be polymorphic. As before, offspring from a given fitness class are produced according to the probabilities provided by the correlation matrix. The transition matrix of this Markov process describes the probability for a population with a certain frequency distribution of fitness values to jump to another composition of fitness classes in the next generation. As there are $\binom{45}{15}$ population states, the transition matrix was not calculated. Instead, we have directly implemented the stochastic process and have combined the outcome of 2000 trials to construct the distribution of waiting times, shown in Figure 13.

It turns out that, at least for this special fitness landscape, the polymorphic correlation approximation is an excellent way to predict the time scales of evolution. Comparing these results with those derived by applying the monomorphic correlation approximation shows that allowing for the specific composition of fitness values within a population is crucial for obtaining accurate predictions. These findings raise the question whether this approach will also perform for other fitness landscapes in a similarly accurate and equally successful way. A first test is to investigate the fitness landscapes of other TSPs, resulting from considering new mutation operators.

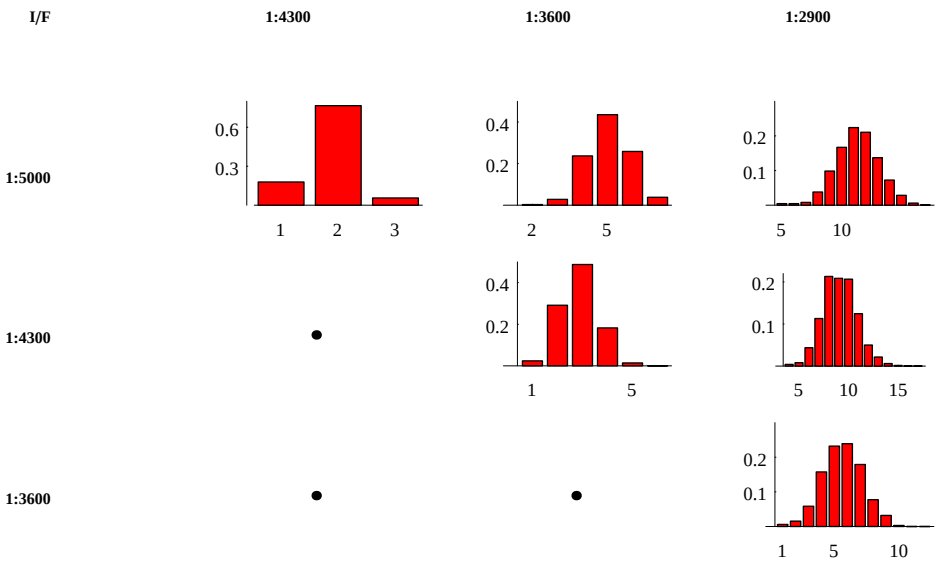

Figure 13: *Waiting time distributions as predicted by polymorphic evolution based on a fine-grained correlation information. Comparison with Figure 6 shows that this approach yields almost exact predictions of waiting times. Although the full TSP is approximated by a* 31×31 *matrix, the match with the actual process is remarkably good.*

2.3 Reverse Mutation

In this section, we use the same configuration of 25 cities and maintain all other parameters, only the mutation operator is changed. For producing an individual's offspring by reverse mutation, we choose two indices. The cities between the smaller and the larger index are now rearranged in reverse order. As two genotypes can be neighbors under reverse mutation while being seperated by a large distance in genotype space under other mutation operators, resulting fitness landscapes may have very different features. The importance of the mutation operator for determining the structural features of fitness landscapes is underscored by the fact that under reverse mutation the evolutionary algorithm needs many more generations for reaching the three final fitness classes, than were required under point mutation. On the other hand, tours with length under 1400 are hardly ever found within 100 generations under point mutation, whereas this task seems to be achieved much more rapidly under reverse mutation. Even the tour of length 1369, see Figure 5, was found within this time limit.

As the polymorphic correlation approach provided the best approximation of the actual waiting times under point mutation, here we focus on an evaluation of this way of imitating the evolutionary process under reverse mutation. As in the case of point mutation, the correlation matrix is estimated by generating random neighbors of random genotypes. The structure of neighborhoods is similar to the one obtained for point mutation and yet possesses some different properties.

The waiting time distributions obtained from 2000 simulation runs of the correlation-based stochastic process exhibit a very close match with the waiting times

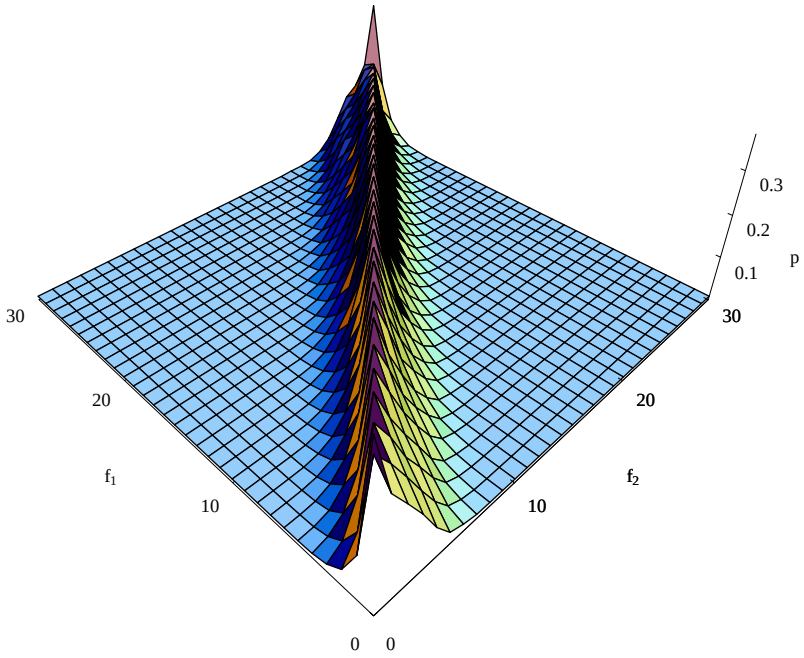

Figure 14: *The correlation matrix for reverse mutation. Producing an offspring with significantly higher fitness appears to be easier under point mutation, at least in the depicted range of fitness values. Under reverse mutation, individuals in class 0 (lowest fitness) hardly ever have neighbors which are at least five fitness classes higher.*

of the evolutionary algorithm, see Figure 15. Again, the polymorphic correlation approximation is successful in describing the evolutionary algorithm.

We conclude our investigations of the TSP by investigating a third mutation operator, resulting in yet another fitness landscape.

2.4 Remove-and-Reinsert Mutation

To apply the remove-and-reinsert mutation operator, we choose two indices successively. The city at the first index will now be taken out and inserted at the second index. The cities in between move backwards or forwards by one index. Notice that for such remove-and-reinsert mutations the order of the two indices is important, whereas point mutations and reverse mutations are symmetric in this respect. Using the same configuration of the TSP, each genotype now has 600 neighbors. Nonetheless, the correlation matrix has the same basic features as before, characterized by a strong emphasis of correlations along the diagonal, see Figure 16.

A comparison of actual and predicted waiting times for remove-and-reinsert mutation, see Figure 17, demonstrates again that the correlation matrix captures all the information necessary to describe the behavior of the much more complex evolutionary algorithm.

Statistics describing the structure of neighborhoods for a fine-grained fitness classification of genotypes have proved to carry the appropriate information for predicting evolutionary waiting times on TSP landscapes. We now have to leave the realm of TSP landscapes and demonstrate successful applications of this method to other landscapes with widely different structural features. In certain respects, the TSP fitness landscape is exceptional; in particular, neighbors that have exactly the same fitness, so-called neutral neighbors, are occurring very rarely. The impact of neutral networks on TSP landscapes may thus be negligible (Huynen *et al.* 1996).

In the next chapter we therefore turn our attention to fitness landscapes that allow for tuning the degree of neutrality.

3 NKp Fitness Landscapes

The family of NK landscapes was first introduced by Kauffman (1993) and later extended to the family of NKp landscapes (Weinberger 1990, Fontana *et al.* 1993). The original idea was to envisage a bitstring of length N as a genotype. All loci (positions on the genotype vector) therefore carry one of the alleles (entries) 0 or 1. For each locus i of the genotype, we choose at random K different other loci, which we then call epistatically linked to locus i. The fitness contribution of locus i thus depends on the entries of K other loci and on the locus i itself. To each of the resulting 2^{K+1} combinations we assign a random number from the interval $[0, 1]$, which determines the contribution to fitness by locus i. In order to calculate the fitness of the whole genotype, we need a fitness table for each locus and for each combination of linked loci. The final fitness value is the sum of the fitness contributions of all loci, divided by N.

Adding the parameter p is a proximate way of incorporating neutrality into the model: The contribution of a specific combination of loci is 0 with probability p,

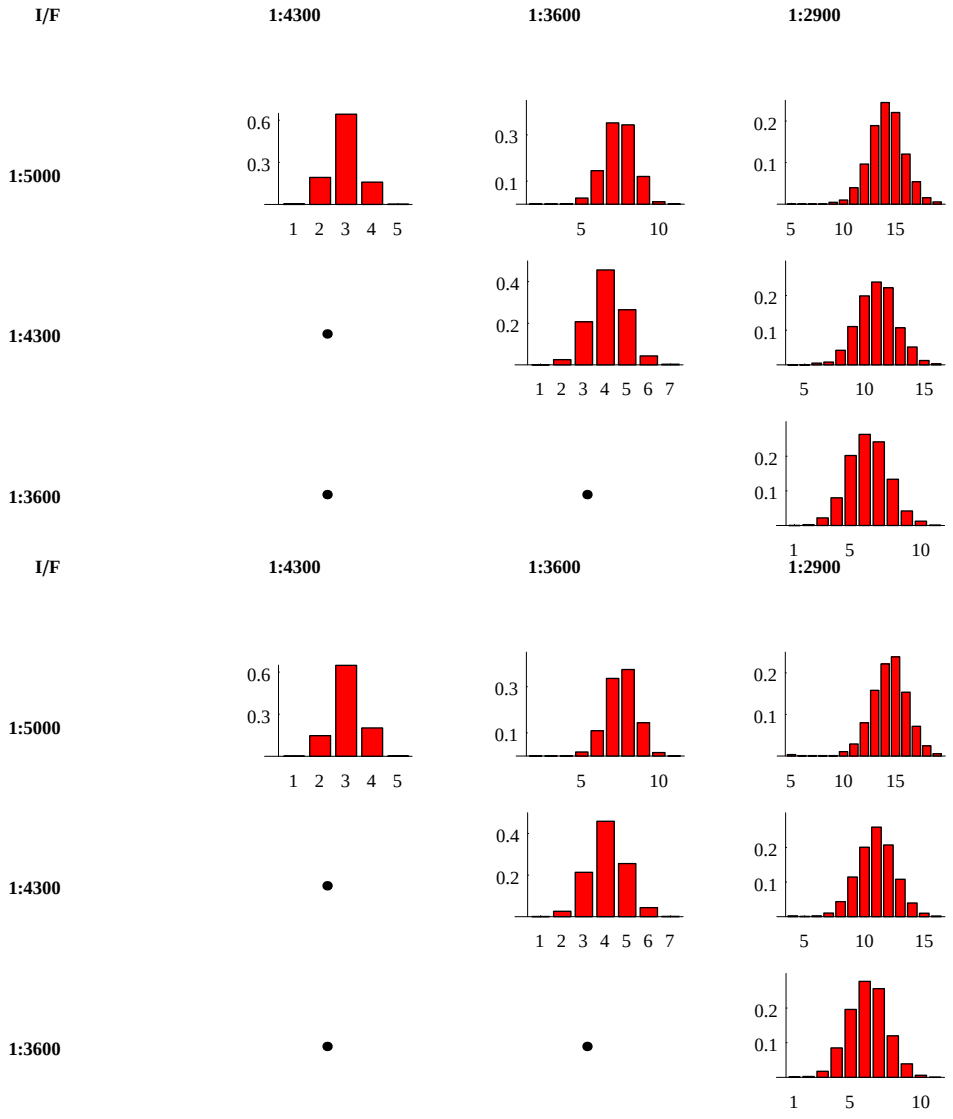

Figure 15: *Waiting times of the evolutionary algorithm with reverse mutation (top) and distributions predicted by the polymorphic correlation approximation (bottom). The predictive accuracy is remarkable.*

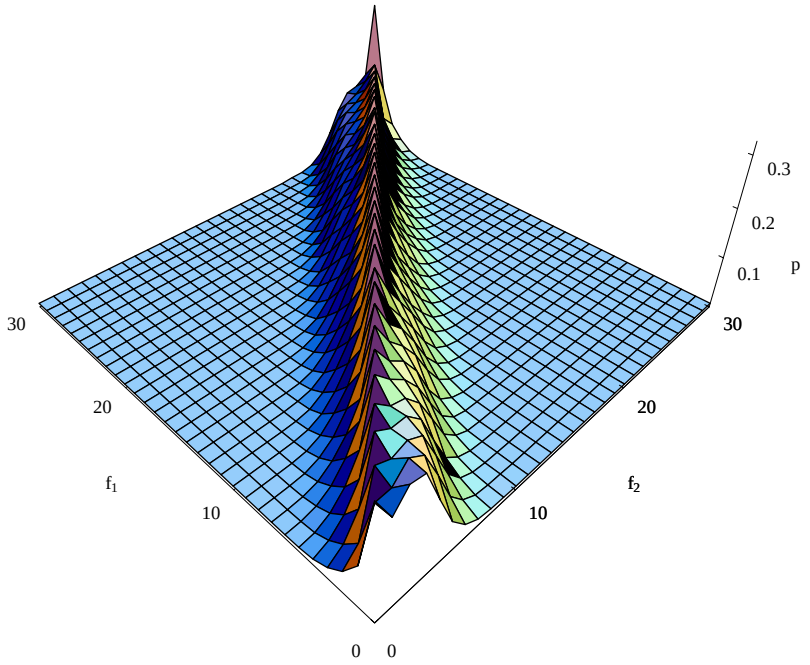

Figure 16: *The correlation matrix for remove-and-reinsert mutation.*

and is assigned randomly from the range $[0, 1]$ with probability $1 - p$. This model is even easier to justify biologically than the original NK model. Many combinations of alleles do not influence a genotype's fitness. The probability for neutral neighbors on the resulting fitness landscape can now be adjusted by simply changing the parameter p.

Mutating a genotype of the NK or NKp model is changing a random vector entry from 1 to 0 and vice versa. In this study we used the same definitions of selection and waiting times as for the TSP, only the population size was reduced to five.

In the subsections below, we present two types of NKp landscapes with different degree of neutrality.

3.1 Low neutrality

To examine a NKp landscape with low probability of neutral neighbors, we chose the following parameters: $N = 15, K = 3, p = 0.3$. We proceed as we did for the TSP and estimate the correlation matrix, see Figure 18.

Comparing the actual waiting time distribution with those predicted by a polymorphic correlation approximation, shows encouraging results. For low degrees of neutrality the correlation matrix again seems to provide enough information to characterize evolution on NKp fitness landscapes.

3.2 High neutrality

For landscapes with a very high degree of neutrality, a simple correlation approximation may be insufficient. Differences of neighborhood structures within one fitness

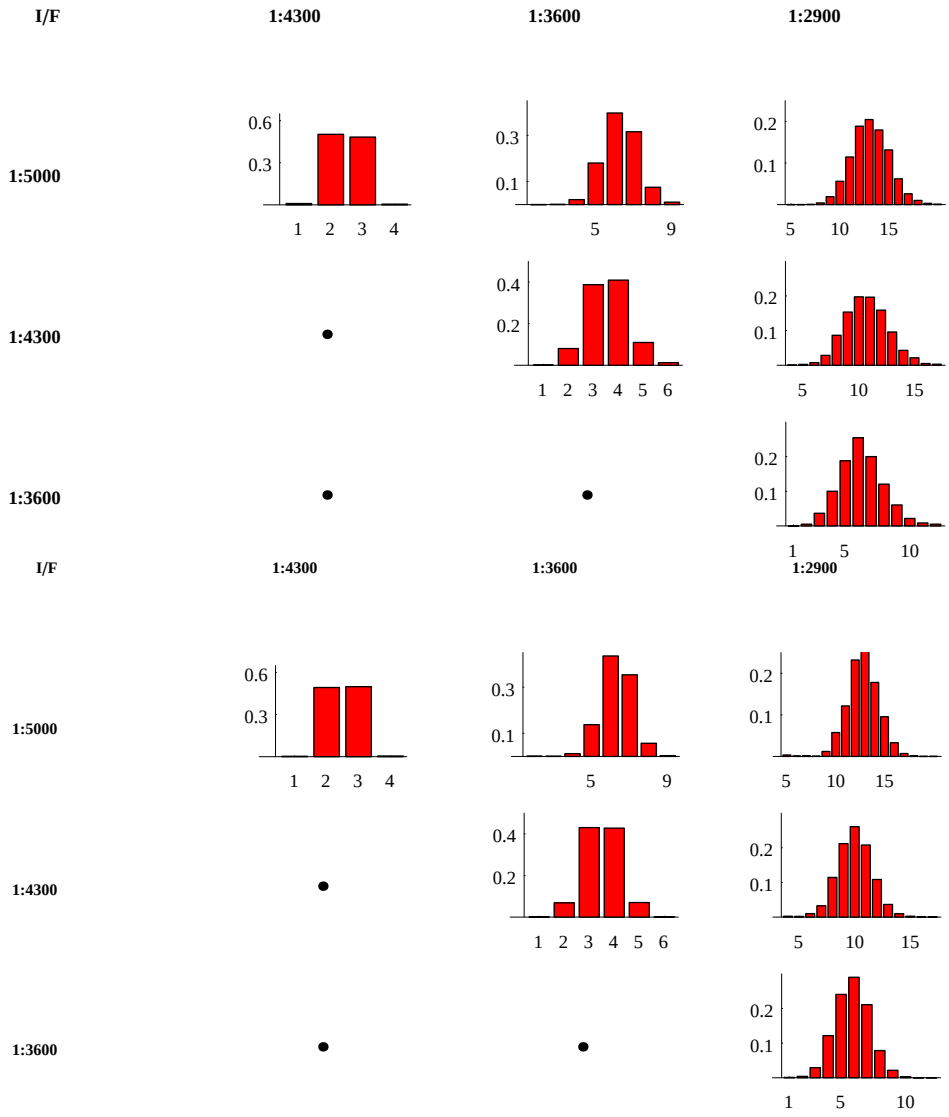

Figure 17: *Actual (top) and predicted (bottom) waiting times for remove-and-reinsert mutation.*

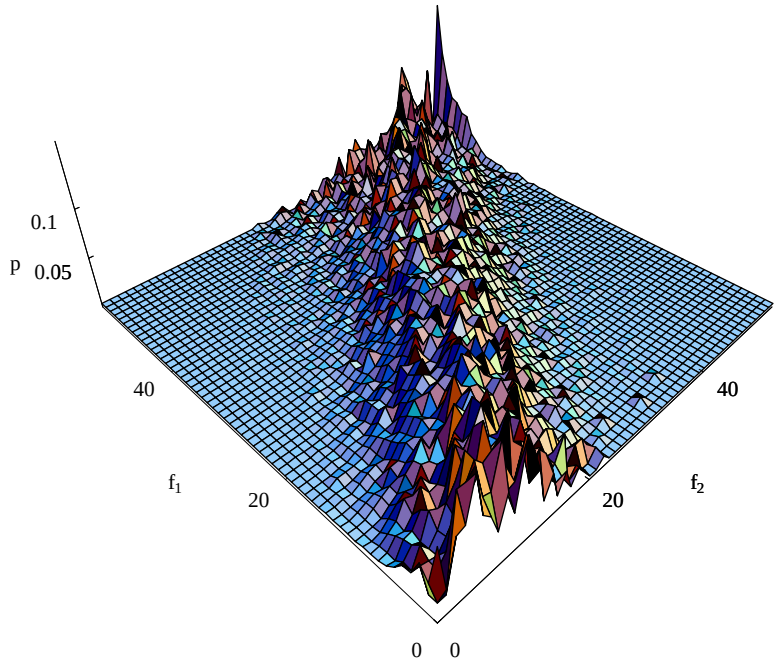

Figure 18: *The correlation matrix for NKp genotypes belonging to 52 different fitness classes. The matrix shows the same property of strong emphasis on the diagonal as did the various TSP landscapes.*

class are neglected. Such differences are expected to occur as a population drifts from the entrance point of a neutral network to an exit, from where it can reach a higher fitness class. Even if the probability for finding a better neighbor is very high at a neutral network's exit, the population probably still needs to evolve for several generations to reach such 'portals'. A simple correlation approximation does not account for this subtlety and can thus only provide mediocre estimates of evolutionary waiting times, see Figure 21.

For $p = 0.99$, a division into more than 17 fitness classes was not possible, valid statistics are only obtained for this relatively coarse classification, see Figure 20.

An improvement of these predictions probably might require incorporation of new statistics of the landscape, allowing to describe population drifts along neutral networks.

4 Conclusions

Although the metaphor of a fitness landscape on which a population is evolving towards adaptive peaks has served as an important basis for understanding evolutionary processes in different areas of science, the question which landscape statistics are critical for predicting evolutionary change on these landscapes so far has not been resolved. In this paper, we have suggested a potential answer to this question, based on polymorphic correlation approximations of evolutionary algorithms.

One of the essential features of an evolutionary algorithm concerns the probability for a population of a certain initial fitness to reach a final fitness threshold within

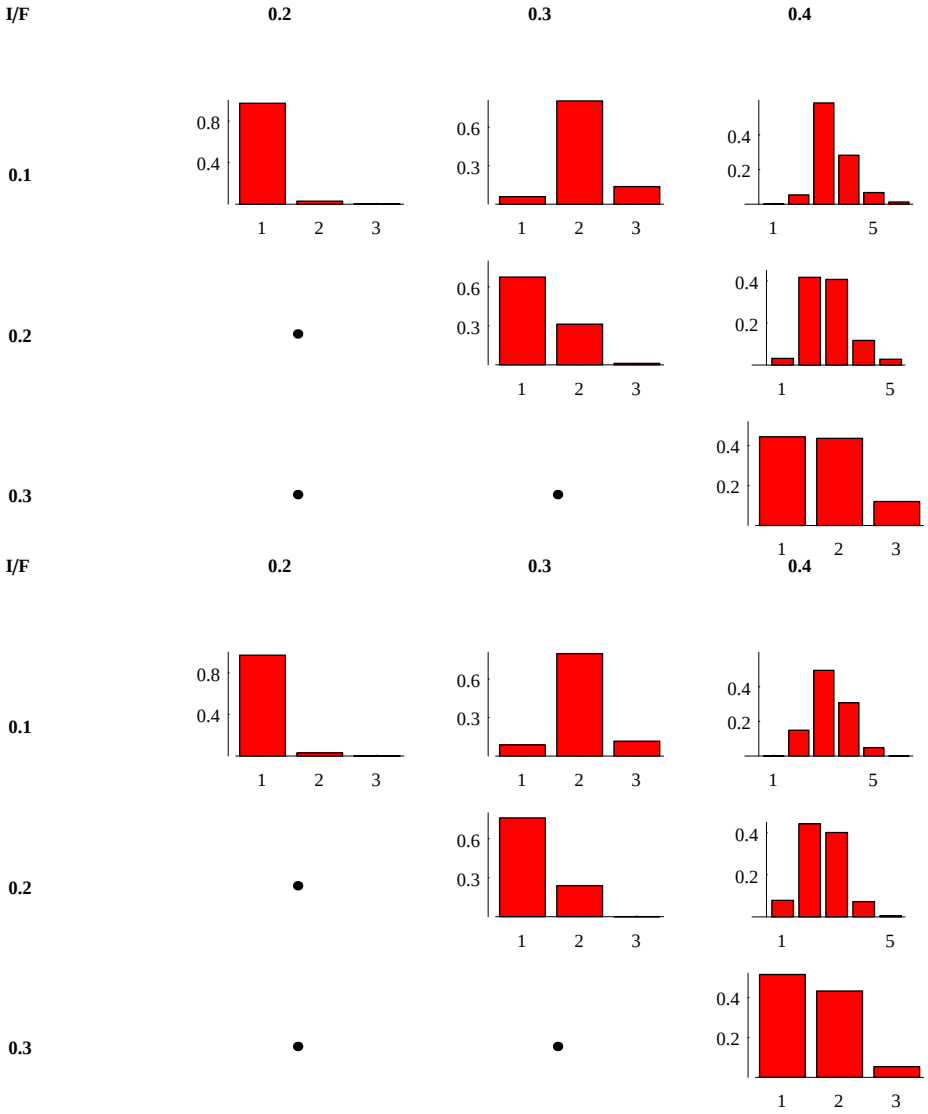

Figure 19: *Actual (top) and predicted (bottom) waiting times for evolution on a NKp landscape with low degree of neutrality* ($p = 0.3$). *The similarity of both results is again very satisfactory.*

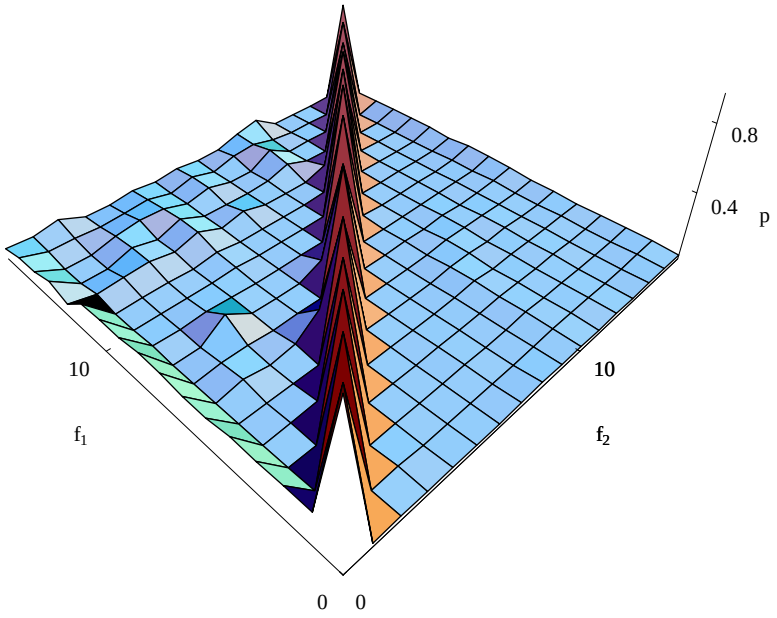

Figure 20: *The correlation matrix for a NKp landscape with p=0.99 shows the high probability for neutral neighbors.*

g generations. This information is summarized in the distribution of waiting times. Comparing the actual waiting times of evolutionary algorithms for several specific problems to predictions based on various candidate statistics served to assess the relative merits of those simplified landscape descriptions.

In Section 2 we have focused on the Travelling Salesman Problem, combined with the point mutation operator. Within this setting, the complexity of landscape statistics has been increased in a sequence of several steps:

1. *Percolation.* All individuals on the landscape are considered to have fitness values above or below a given threshold with probability p and $1 - p$, respectively (Gavrilets and Gravner 1997). The results of an approximate evolutionary process based on this simplification showed that introducing more fitness classes and accounting for their specific distribution of neighboring fitness values are vital steps for overcoming the poor predictive accuracy of the percolation approximation.

2: *Coarse correlation with monomorphic population.* In a next step we have used correlation statistics of the landscape. This has enabled us to take into account that genotypes in different fitness classes are surrounded by different neighborhood structures. Although a coarse classification of fitness values into just a small number of fitness classes, while treating the population as being monomorphic, resulted in improved predictions relative to the percolation approximation. Yet, the simplified processes still resulted in too long waiting times.

3: *Fine correlation with monomorphic population.* A fined-grained correlation matrix served as the basis for the next step and was supposed to better represent the possibility of small changes in fitness values, which can be critical for describing the evolutionary process. Predicted time scales of evolution lay quite close to those of the actual evolutionary algorithm. However, the fact that the waiting times were systematically underestimated led us to conjecture that the simultaneous presence

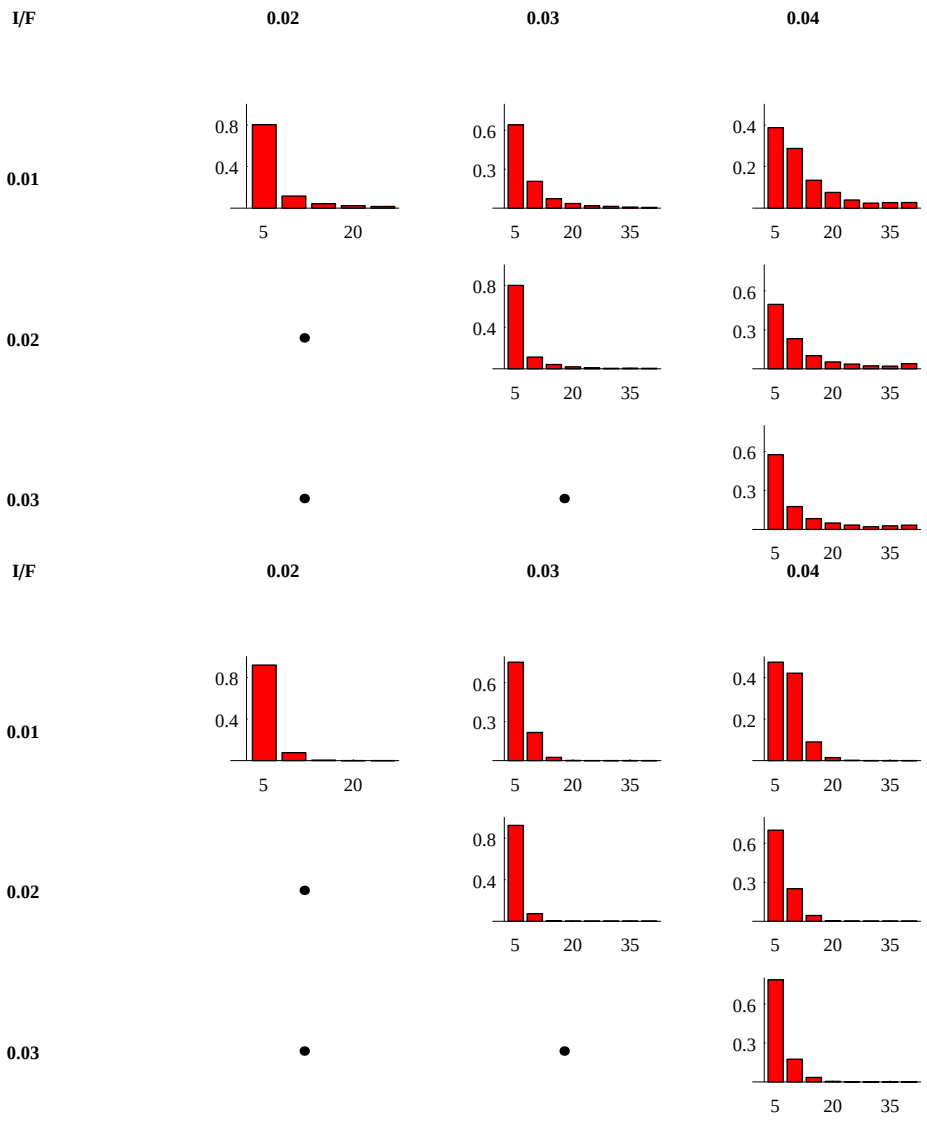

Figure 21: *Actual (top) and predicted (bottom) waiting times for evolution on a NKp landscape with a very high degree of neutrality ($p = 0.99$). The comparison shows that actual evolution takes many more generations to attain a final fitness class; the many steps within one fitness class that probably result from diffusion along neutral networks are not considered by the correlation approximation.*

of different fitness values within a population is important for predicting the mode of evolutionary change.

4: *Fine correlation with polymorphic population.* Allowing the population to be polymorphic, so that individuals in one generation can belong to different fitness classes resulted in remarkably accurate predictions of evolutionary waiting times. All the information needed for this very close approximation of the evolutionary algorithm is provided by a fine-grained correlation matrix.

Until now, mainly two different correlation functions have been used to describe fitness landscapes; these are often referred to as *auto-correlation functions*. Based on these functions, however, rough approximations of the behavior of the evolutionary process were possible (Stadler 1995, Manderick 1997). One type of these functions is based on the auto-correlation of fitness values in time series that result from random walks on a fitness landscape. The dynamics of such random walks are only determined by mutation and fitness values or selective pressures are not considered. Auto-correlations are then averaged over all possible initial conditions for the random walk. The second type of auto-correlation function is based on considering the correlation of fitness values between pairs of genotypes at varying mutational distances. It can be shown that these two types of auto-correlation functions carry equivalent information (Stadler 1995).

The correlation measure presented in this study at the same time reduces and enhances the statistical information provided: the focus of this new measure is on single mutational steps (reduced information) but the initial fitness of considered individuals is maintained in the measure (enhanced information). The success of these landscape statistics, allowing for almost perfect prediction of evolutionary waiting times, shows that we open up a new pathway for improving our understanding of fitness landscapes.

Our study of NKp fitness landscapes with varying degree of neutrality confirmed that the correlation matrix of a landscape provides all the information necessary for describing evolution in systems with a rather low probability for neutral neighbors. For landscapes with a higher degree of neutrality, it seems that additional considerations are required. For example, a population that drifts along a neutral network can probably be represented by a diffusion process, which either takes additional landscape statistics into account or utilizes the information provided by the correlation matrix in an innovative way.

In summary, we have shown that for many fitness landscapes a sufficiently fine-grained correlation matrix provides an excellent basis for predicting evolutionary change on these landscapes. Evolutionary processes on landscapes with very high degrees of neutrality require additional study.

References

Barnett L. (1997). Tangled Webs - Evolutionary Dynamics on Fitness Landscapes with Neutrality MSc. diss., School of Cognitive and Computing Sciences, Sussex Univ. UK.

Barnett L. (1998). Ruggedness and neutrality - the NKp family of fitness landscapes, *Artificial Life VI, Proceedings of the Sixth International Conference on Artificial Life, MIT press*

Beasley D. (1997). Possible applications of evolutionary computation, In: "Handbook of Evolutionary Computation" (A1.2) Bắck T., Fogel D.B. and Michalewicz Z. (eds.) *Institute of Physics Publishing and Oxford University Press*

Fontana W., Stadler P.F., Bornberg-Bauer E.G, Griesmacher T., Hofhacker I.L., Tacker M., Tarazona P., Weinberger E.D., and Schuster P. (1993). RNA folding and combinatory landscapes, *Phys. Rev. E*, **47**(3): 2083-2099.

Gavrilets S. (1997). Evolution and Speciation on Holey Adaptive Landscapes, *Trends Ecol. Evol.*, **12**(8): 307-312.

Gavrilets S. and Gravner J. (1997). Percolation on the fitness hypercube and the evolution of reproductive isolation, *J. Theor. Biol.* **184**(1): 51-64.

Grimmett G. (1989). Percolation, *Springer*

Happel R. and Stadler P.F. (1996). Canonical Approximation of Fitness Landscapes, *Complexity* **2**: 53-58.

Huynen M.A., Stadler P.F. and Fontana W. (1996). Smoothness within ruggedness: The role of neutrality in adaptation, *Proc. Natl. Acad. Sci. (USA)* **93**(1): 397-401.

Kauffman S.A. and Levin S. (1987). Towards a general theory of adaptive walks on rugged landscapes, *J. Theor. Biol.* **128**: 11-45.

Kauffman S.A. (1993). The Origins of Order - Self-Organization and Selection in Evolution, *Oxford University Press.*

Kesten H. (1982). Percolation theory for mathematicians, *Birkhaeuser.*

Manderick B. (1997). Correlation analysis, In: "Handbook of Evolutionary Computation" (B2.7.3) Bắck T., Fogel D.B. and Michalewicz Z. (eds.) *Institute of Physics Publishing and Oxford University Press.*

Pasemann F., Steinmetz U. and Dieckmann U. (1999). Evolving structure and function of neurocontrollers, *Proceedings of the 1999 Congress on Evolutionary Computation* July 6th to 9th 1999, Madison, Washington DC, USA (1999).

Reidys C.M. and Stadler P.F. (1999). Neutrality in Fitness Landscapes, *Appl. Math. & Comput.* in press 1999.

Rudolph G. (1997). Stochastic processes, In: "Handbook of Evolutionary Computation" (B2.2) Bắck T., Fogel D.B. and Michalewicz Z. (eds.) *Institute of Physics Publishing and Oxford University Press.*

Sahimi M. (1994). Applications of percolation theory, *Taylor&Francis.*

Schuster P., Fontana W., Stadler P.F. and Hofacker I.L. (1994). From sequences to shapes and back: A case study in RNA secondary structures, *Proc. Roy. Soc. Lond. b Bio.* **255**(1344): 279-284.

Schuster P. (1996). How does complexity arise in evolution?, *Complexity* **2**(1): 22-30.

Schuster P. (1997). Landscapes and Molecular Evolution, *Physica D* **107**(2-4): 351-365.

Stadler P.F. (1992). Correlation in landscapes of combinatorial optimization problems, *Europhys. Lett.* **20**(6): 479-482.

Stadler P.F. and Schnabl W. (1992). The landscape of the travelling salesman problem, *Phys. Lett. A* **161**(4): 337-344 .

Stadler P.F. (1995). Towards a Theory of Landscapes, In: "Complex Systems and Binary Networks" (Proc. of the Guanajuato Lectures 1995) Lopéz-Peña R., Capovilla R., García-Pelayo R., Waelbroeck H. and Zertuche F. (eds.), *Springer*.

Stadler, P.F. (1996). Landscapes and their Correlation Functions, *J.Math.Chem.* **20**(1-2): 1-45.

Stauffer D. and Aharony A. (1994). Introduction to percolation theory, *Taylor&Francis*.

Weinberger E. (1990). Correlated and Uncorrelated Fitness Landscapes and How to Tell the Difference, *Biol.Cybern.* **63**(5): 325-336.