Biol. Cybern. 63, 325–336 (1990)

Biological Cybernetics

Correlated and Uncorrelated Fitness Landscapes and How to Tell the Difference

E. Weinberger

Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104, USA

Received September 30, 1989/Accepted in revised form March 7, 1990

Abstract. The properties of multi-peaked "fitness landscapes" have attracted attention in a wide variety of fields, including evolutionary biology. However, relatively little attention has been paid to the properties of the landscapes themselves. Herein, we suggest a framework for the mathematical treatment of such landscapes, including an explicit mathematical model. A central role in this discussion is played by the autocorrelation of fitnesses obtained from a random walk on the landscape. Our ideas about average autocorrelations allow us to formulate a condition (satisfied by a wide class of landscapes we call AR(1) landscapes) under which the average autocorrelation approximates a decaying exponential. We then show how our mathematical model can be used to estimate both the globally optimal fitnesses of AR(1) landscapes and their local structure. We illustrate some aspects of our method with computer experiments based on a single family of landscapes (Kauffman's "N-k model"), that is shown to be a generic AR(1) landscape. We close by discussing how these ideas might be useful in the "tuning" of combinatorial optimization algorithms, and in modelling in the experimental sciences.

Introduction

There has been considerable recent interest in considering evolution as a combinatorial optimization problem, that is, a problem in finding the best of a large, but finite number of possibilities. Biologists, such as Kauffman and Levin (1987), have embraced this paradigm in the hope that they might learn something new about evolution, and computer scientists, such as Holland (1981) and Brady (1985), hope to use evolutionary strategies in developing new methods of solving combinatorial optimization problems. Similar issues have also attracted the attention of physicists, such as Palmer (1989) and Stein (1989). They speculate that the thermodynamics of glassy systems, such as polymers and other more or less random covalent networks, is intimately related to the complex structure of the barrier heights present in the potential surface. The common denominator in all of this work is a notion that Kauffman and Levin have called a "rugged landscape.". If one is a biologist, such a landscape can be interpreted as a fitness landscape; if one is a computer scientist, the landscape is the set of allowable configurations in some optimization problem; and if one is a physicist/chemist, it is a glass and/or spin glass.

To the biologist, the notion of an adaptive landscape is not new, having been proposed in 1932 by Sewell Wright (Wright 1932). Subsequent developments in biological thinking have only reinforced the power of this idea. Molecular biology has made clear the essential discreteness of the genome, and thus the finitude of the number of possible organisms. It is, perhaps, easier to make sense of the concept of an adaptive landscape if one considers the evolution of individual molecules, rather than entire organisms; hence, Smith's notion of a "peptide space" (Smith 1970). In such a space, one arranges all peptides of a specified length in such a way that nearest neighbors differ by a single amino acid substitution at a single site. One might then define the fitness of the peptide as its ability to bind to a particular substrate, to catalyze a specific reaction, etc. The lethal nature of the sickle cell anemia mutation in humans emphasizes the fact that even a single point mutation can result in a dramatic change in fitness. Furthermore, the selection of an optimal enzyme to catalyze a particular reaction involves a host of complex tradeoffs: the enzyme must bind the substrate tightly enough so that the reaction will proceed, but not so tightly that it will not be released when the reaction is completed, the enzyme must not also catalyze competing reactions, and, of course, the enzyme must not interfere with the action of other enzymes.

Present address: Max-Planck-Institut für biophysikalische Chemie, Postfach 2841, Am Fassberg, D-3400 Göttingen-Nikolausberg, Federal Republic of Germany

It was these considerations that inspired Kauffman and Levin (1987) to draw the analogy between biological constraints and those imposed by the combinatorial optimization problems that have received much recent attention in computer science (see, for example, Garey and Johnson 1979). Perhaps the most famous example of this class of problems is the travelling salesman problem (TSP), which is to find the shortest tour through N cities from an initial city, visiting each city once, and returning at the end to the initial city. For as few as 11 cities, the number of possible tours is in the millions, growing as $((N-1)/2)!$. This is typical of combinatorial optimization: out of a finite, but extremely large set of entities, one would like to find one that is, in some well defined sense, the "fittest". For a large class of such problems (e.g. the class of NP-complete problems), it is thought that even the most efficient algorithms that solve them must require a computational effort that grows at least exponentially with some measure of the "size" of the problem (such as the number of cities that the traveling salesman must visit). Indeed, the only known method of finding the optimal traveling salesman tour is to exhaustively search through the list of possible tours, checking each to see if each is shorter (and therefore "fitter") than all of those previously encountered. Because of the extreme amounts of time required for such a search, which can easily exceed the age of the universe for even moderate N on the largest of computers, algorithms that settle for sub-optimal tours are used in practice. The effective intractability of other combinatorial optimization problems places a similar constraint on algorithms to solve these problems, as well. Typically, these algorithms use a generalization of the fact that tours that visit the cities in more or less the same order have more or less the same length. It is therefore profitable to search the vicinity of a good solution in the hopes of finding a better solution. This heuristic is incorporated into the algorithms by explicitly defining "neighborhoods", and searching from neighbor to neighbor.

Essential features of the above discussion of the travelling salesman problem are also present in our evolutionary paradigm. In the case of a 20 amino acid peptide space, there are 20^N possible proteins of length N which would need to be searched to find the "fittest" protein for a particular function. How can an evolving biological system search through such a vast number of possibilities? Obviously, it can't. Instead, it would seem that evolution, like practical combinatorial optimization algorithms, makes do with sub-optimal solutions. Indeed, it has been shown previously (Weinberger 1987a, b) that there is a detailed analogy between evolutionary optimization and the combinatorial optimization technique known as simulated annealing (Kirkpatrick et al. 1983). The "local hill climbing" heuristic described above, which, in the context of combinatorial optimization, involves transitions to randomly selected fitter neighbors, is one of the simplest of combinatorial optimization schemes.

It is useful to consider a rugged landscape as an abstract mathematical object, both because some important considerations emerge immediately, and because a framework will be provided for subsequent discussion. This we do in the next two sections, motivating our remarks by considering the two concrete examples of combinatorial optimization problems mentioned previously, the travelling salesman problem and optimization on peptide space. The second of these secions rigorously defines the notion of a "correlation length", and, more generally, an autocorrelation function for such combinatorial landscapes. We will see that this autocorrelation function is indeed a decaying exponential, as implied by the existence of a single correlation length, for the wide class of landscapes that we call AR(1) landscapes. We then use these theoretical tools to explain why the two most obvious methods of optimization on combinatorial landscapes, random search and local hill climbing, do remarkably badly in finding the point on the landscape with the globally optimal fitnesses. The next section shows how our theoretical framework can be used to predict aspects of the local structure of a landscape, illustrating our method with computer experiments based on a simple family of two amino acid peptide landscapes (the "N-k model") proposed in Kauffman et al. In this section, we also demonstrate the intriguing fact that N-k landscapes are generic members of the class of AR(1) landscapes. The final section of the paper discusses the significance of our results.

Some Thoughts about the Abstract Structure of Fitness Landscapes

In all of the applications mentioned above, one would like to know the details of specific landscapes. Unfortunately, this "best of all possible worlds" scenario is unrealistic, due to the enormous amount of experimental data required. A more realistic goal would be to gather a moderate amount of data about the landscape and infer likely statistical properties of the ensemble of landscapes that might fit the data. It is therefore appropriate to begin a mathematical description of rugged fitness landscapes by assuming that the fitnesses of its points are random variables, and inferring properties of their moments. Although this program can be carried out if the random variables have essentially any distribution, our results will be most useful if the random fitness values have a joint Gaussian distribution. In that case, the distribution is completely specified by a vector of mean fitness values and a matrix specifying the covariance of the fitnesses. It is also useful to note that any linear combination of jointly Gaussian variables is, itself, Gaussian.

There are both theoretical and simulation results showing that the marginal fitness distributions are, in fact, Gaussian for a wide class of landscapes of practical interest. (Of course, the fact that all of the marginal distributions are Gaussian is necessary, but not sufficient to conclude that a collection of random variables is jointly Gaussian. See Feller (1972) for counter-examples. However, Karlin and Taylor (1975) note that the