Journal of Mathematical Chemistry 20 (1996) 1–45

Landscapes and their correlation functions

Peter F. Stadler [a,b]

[a] *Theoretische Biochemie, Institut für Theoretische Chemie, Universität Wien, A-1090 Vienna, Austria*
[b] *Santa Fe Institute, Santa Fe, NM, USA*

Received 10 August 1995; revised 30 April 1996

Fitness landscapes are an important concept in molecular evolution. Many important examples of landscapes in physics and combinatorial optimization, which are widely used as model landscapes in simulations of molecular evolution and adaptation, are "elementary", i.e., they are (up to an additive constant) eigenfunctions of a graph Laplacian. It is shown that elementary landscapes are characterized by their correlation functions. The correlation functions are in turn uniquely determined by the geometry of the underlying configuration space and the nearest neighbor correlation of the elementary landscape. Two types of correlation functions are investigated here: the correlation of a time series sampled along a random walk on the landscape and the correlation function with respect to a partition of the set of all vertex pairs.

1. Introduction

Since Sewall Wright's seminal paper [1] the notion of a *fitness landscape* underlying the dynamics of evolutionary optimization has proved to be one of the most powerful concepts in evolutionary theory. Implicit in this idea is a collection of genotypes arranged in an abstract metric space, with each genotype next to those other genotypes which can be reached by a single mutation, as well as a value assigned to each genotype. Such a construction is by no means restricted to biological evolution; Hamiltonians of disordered systems, such as spin glasses [2,3], and the cost functions of combinatorial optimization problems [4] have the same basic structure. It has been known since Eigen's [5] pioneering work on the molecular quasispecies that the dynamics of evolutionary adaptation (optimization) on a landscape depends crucially on detailed structure of the landscapes itself. Extensive computer simulations, see, e.g., [6,7] have made it very clear that a complete understanding of the dynamics is impossible without a thorough investigation of the underlying landscape [8,9].

The landscapes of a number of well known combinatorial optimization problems such as the Traveling Salesman Problem (TSP) [10], the Graph Bipartitioning Problem (GBP) [11], or the Graph Matching Problem (GMP) have been investigated in some detail, see [12–14]. A detailed survey of a variety of model landscapes

derived from folding RNA molecules into their secondary structures has been performed recently [6,7,15–26].

Most of the knowledge about landscapes has so far been derived using statistical methods, considering random models of landscapes rather than a single landscape. The distribution of local optima and the statistical characteristics of down-hill walks have been computed for the uncorrelated landscape of the random energy model [27–29]. Furthermore, two one-parameter families of tunably rugged landscapes have been studied extensively: the Nk model and its variants [17,30–32] and the p-spin models [33–36]. Local optima of 2-spin models are considered in [37–41]. While the statistical approach is the natural one, e.g., in the physics of spin glasses, it seems to be rather contrived in evolutionary biology because it is by no means clear what a reasonable statistical model should look like, even if there is a computational procedure to model *the* landscapes of, say, RNA free energies.

A theory of landscape is based on three ingredients: we are given a *finite*, but very large set V of "configurations" and a "fitness function" $f : V \to \mathbb{R}$. The third ingredient is a notion of neighborhood between the configurations, which allows us to interpret V as the vertex set of a graph Γ. We will refer to Γ as the *configuration space* of the landscape f. Let us briefly discuss two examples here:

Consider the set of RNA molecules of given chain length n. A particular molecule x can be represented as a string of length n taken from the alphabet $\{\mathbf{G}, \mathbf{C}, \mathbf{A}, \mathbf{U}\}$; molecular biologists call this string the *sequence* of the RNA. The "fitness" function f is, for instance, the free energy of folding x into its secondary structure [15]. *In silico* the folding is done by an algorithm containing a large number of experimentally determined parameters [42]. In nature as well as in *in vitro* experiments variation is introduced by mutations, predominantly point-mutations. Neighboring sequences are thus those that differ in only a single position. The resulting graph is known as the sequence space [43,44].

A very different example is the travelling salesman problem. A salesman starts from his home city and visits exactly once each of the n cities on a given list, then he returns home. The configurations are the possible tours, i.e., all permutations of the cities on the salesman's list. The numerical value assigned to a particular tour τ is its total length $f(\tau)$. The notion of neighborhood between different tours is much less obvious here than in the biological example above. Usually one says that two tours are neighbors if they can be interconverted by a simple operation on the list of cities, such as swapping two cities (transpositions), or inverting the order of a contiguous part of the list. It turns out that the performance of an optimization heuristic depends crucially on the choice of the neighborhood relation. We will return to this topic later in this contribution.

Conceptually, there is a close connection between the (biological) *landscapes* and the *Potential Energy Surfaces* (PES) that constitute one of the most important issues of theoretical chemistry [45,46]. As a consequence of the validity of the Born-Oppenheimer approximation, the PES provides the potential energy as a function of the nuclear geometry of the system, $U(R)$. PES are therefore defined on a high-