Journal of Theoretical Biology 254 (2008) 252–263

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Hierarchical distribution of ascending slopes, nearly neutral networks, highlands, and local optima at the *d*th order in an NK fitness landscape

Takuyo Aita *

Graduate School of Science and Engineering, Saitama University, Saitama 338-8570 Japan

ARTICLE INFO

Article history:
Received 17 October 2007
Received in revised form
6 June 2008
Accepted 7 June 2008
Available online 14 June 2008

Keywords:
Fitness landscape
Sequence space
Adaptive walk
NK model
In vitro evolution
Evolvability

ABSTRACT

We obtained several structural features of an NK fitness landscape by analytical approach. Particularly, we focused on spatial distributions of "ascending slopes", "highlands", "nearly neutral networks", and "local optima" along the fitness coordinate W, from the viewpoint of adaptive walks with step-width d, where d is the number of mutated sites (Hamming distance) after a generation. The parameter k governs the degree of the ruggedness on the NK landscape, and we handled cases where k is moderate against the sequence length. From the foot up to the middle region on the landscape, many ascending slopes exist (high evolvability) and these slopes extend up near the "highland", which is mathematically defined as the specific region $W = W_d^*$ where the expectation of the fitness increment becomes zero. Denoting the standard deviation of the fitness change at $W = W_d^*$ by SD^*, we considered the existence of "nearly neutral networks", which percolate in the fitness band between $W - SD^*$ and $W + SD^*$. Our results suggest that the highland corresponds to a phase-transition threshold of the formation of the nearly neutral networks. Near or over the highland, "local optima at the dth order" appear drastically (low evolvability), where d means the radius of their basins. The value of W_d^* increases with d increasing. Then, as the fitness (= altitude) becomes higher, the basin size of the local optima increases. This leads to a conclusion that it is very hard or impossible for walkers with step-width d to reach near the global peak when d is a realistic large value: $d = 1$–6, and suggests that the region over the middle in real landscapes may be considerably smooth with small k-values to maintain high evolvability.

1. Introduction

Biological evolution process is comprehended as the "adaptive walk" or "hill-climbing" on a fitness landscape in genotype space (sequence space). The concept was first proposed by Wright (1932), and has been developed in evolutionary biology (Maynard-Smith, 1970; Eigen, 1992; Kauffman, 1993; Gavrilets, 2004). In the field of *in vitro* molecular evolution, which handles artificial evolution of protein or DNA sequences, the concept of "fitness" is extended to a molecular physicochemical property such as propagation rate (Eigen, 1985), enzymatic activity, binding affinity, or thermostability (Arnold, 2000; Matsuura and Yomo, 2006). Thus, the fitness landscape is regarded as the "evolutionary attribute" of biopolymers. Many theoretical studies have been done based on various mathematical models of the landscapes (Voigt et al., 2000; Gavrilets, 2004). One of the most familiar models is the NK model (Kauffman and Weinberger, 1989; Kauffman, 1993). The NK model is a mathematical model describing a complex system in which an arbitrary element is affected by other k elements. For a protein, an amino acid site corresponds to the element in the NK model. If the k sites that cooperatively affect the jth site are located near the jth site, this model is called the "adjacent neighbor (NK) model", whereas, if the k sites are located randomly through the sequence, this model is called the "random neighbor (NK) model" (Kauffman, 1993). The fitness landscape constructed by the NK model is called the "NK (fitness) landscape". The NK landscape with $k = 0$ has a single peak, whereas the landscape becomes more rugged and has more local peaks with k increasing.

The original NK model was proposed by Kauffman and Levin (1987) and many slightly modified models have been studied (Barnett, 1998; Newman and Engelhardt, 1998; Iguchi et al., 2005). In almost all cases, they have adopted a binary sequence space, where the number of available letters is two (0 or 1). We denote that by $\lambda = 2$. Many theorists have been interested in such issues as the height of the global peak and local peaks, the number of local peaks, and the walk length from a random point. These issues were investigated numerically by computer simulation in most studies (e.g. Kauffman, 1993). The first analytical study was conducted by Weinberger (1991). Recently, several rigorous analytical studies were conducted by Evans and Steinsaltz (2002), Durrett and Limic (2003), and Limic and Pemantle

* Tel.: +81 48 858 3756.
E-mail address: taita@mail.saitama-u.ac.jp

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.06.002

(2004), in which they adopted the adjacent neighbor model with $\lambda = 2$. Iguchi et al. (2005) investigated the effect of a scale-free network of inter-sites interaction on the properties of the NK landscape. The NK model has also been adopted to study the neutrality on the evolution or neutral evolution (Barnett, 1998; Ohta, 1998; Newman and Engelhardt, 1998).

Meanwhile, several studies tried to estimate the k-value in real landscapes by fitting the NK model to experimental data. The k-value seems inherent in the physicochemical property of individual biopolymers. Kauffman and Weinberger (1989) applied the NK model to affinity maturation of the V region in immunoglobulin and estimated that k is about 40, from the viewpoint of the number of steps of adaptive walk up to the local optima (Kauffman, 1993). Fontana et al. (1993) examined the RNA free energy landscape by computer experiments, and estimated $k = 7$–8 for this landscape in terms of autocorrelation on the landscape. In our previous paper (Hayashi et al., 2006; Aita et al., 2007), we demonstrated that the experimental data in an *in vitro* molecular evolution can be explained quantitatively by our adaptive-walk theory on an NK landscape, which was defined by the random neighbor (NK) model. Hayashi and coworkers carried out *in vitro* molecular evolution beginning with a defective fd phage carrying a random polypeptide (139 a.a.) in place of the g3p minor coat protein D2 domain, which is essential for phage infection. Through 20 mutation-selection cycles, the random polypeptide evolved gradually up to the middle region of the fd-phage infectivity landscape in the free energy scale (Hayashi et al., 2006). Our analysis of the experimental data suggested $k \approx 27$ for this landscape (Aita et al., 2007).

Therefore, in light of the great importance of the NK landscapes, we examined analytically structural features of the NK landscape ($\lambda = 20$, random neighbor model) from the viewpoint of an adaptive walker with step-width d. Particularly, we were concerned with spatial distributions of the local optima at the dth order, where d means the radius of their basins. Almost all previous studies considered cases where $d = 1$, because they handled natural evolution in which mutation occurs infrequently. We want to make a note that our interest is not in natural evolution but in the *in vitro* molecular evolution, which can control the number of mutated sites (Hamming distance) d after a generation and population size. In addition to this, our originality lies in that we analytically obtained the local fitness distribution over all possible d-fold point mutants generated from a reference sequence with fitness W (Aita et al., 2007). Many statistical properties of the landscape were derived from the local fitness distribution. In this paper, we will refer to the spatial distributions of "ascending slopes", "highlands", "nearly neutral networks", and "local optima" along the fitness coordinate.

2. Model of the NK fitness landscape

We consider all conceivable amino acid sequences with a chain length of ν, and λ letters are available at every site, where λ is large enough to satisfy $(\lambda - 1)/\lambda \approx 1$. Then, each sequence is mapped into the corresponding point in the λ-valued ν-dimensional sequence space. The fitness W for a given sequence "$\mathrm{A}_1\mathrm{A}_2,\ldots,\mathrm{A}_\nu$" is defined by

$$W = \sum_{j=1}^{\nu} w_j(\mathrm{A}_j|\mathrm{A}_{j_1},\mathrm{A}_{j_2},\ldots,\mathrm{A}_{j_k}), \tag{1}$$

where $w_j(\mathrm{A}_j|\mathrm{A}_{j_1},\mathrm{A}_{j_2},\ldots,\mathrm{A}_{j_k})$ is the "site-fitness", i.e., a fitness contribution from a particular letter A_j at the jth site when the k sites $\{j_1,j_2,\ldots,j_k\}$ are occupied by the particular letters $\{\mathrm{A}_{j_1},\mathrm{A}_{j_2},\ldots,\mathrm{A}_{j_k}\}$. The k sites $\{j_1,j_2,\ldots,j_k\}$ are randomly chosen from all $\nu - 1$ sites except the jth site ("random neighbor model"). The assignment of site-fitness values is modeled as follows: with a set of letters $\{\mathrm{A}_{j_1},\mathrm{A}_{j_2},\ldots,\mathrm{A}_{j_k}\}$ given, a site-fitness value of an arbitrary letter a (e.g., $a = \mathrm{Ala},\mathrm{Cys},\ldots,\mathrm{Tyr}$) for each site is randomly once assigned from the following set of λ values ("quenched model"):

$$w_j(a|\mathrm{A}_{j_1},\mathrm{A}_{j_2},\ldots,\mathrm{A}_{j_k}) \in \left\{\varepsilon\left(1-\frac{2i}{\lambda-1}\right)\middle| i = 0,1,2,\ldots,\lambda-1\right\}, \tag{2}$$

where ε is a positive constant: $\varepsilon > 0$. We do not allow the degeneracy of assignment, that is, $w_j(a|\cdots) \neq w_j(a'|\cdots)$ for $a \neq a'$. Therefore, the underlying density function of site-fitness w at each site is given by the comb function:

$$\frac{1}{\lambda}\sum_{i=0}^{\lambda-1}\delta\left(w-\varepsilon\left(1-\frac{2i}{\lambda-1}\right)\right), \tag{3}$$

where $\delta(x)$ is Dirac's delta function. The reason why we fix the site-fitness distribution as shown in Eq. (3) lies in analytical tractability demonstrated in Appendix C.[1] We note that our theoretical conclusion is robust to the shape of the site-fitness distribution (Limic and Pemantle, 2004) and also robust to the site-dependence of ε-values (Aita et al., 2004). From Eq. (3), we can see that the mean of the site-fitness values over λ letters is equal to zero, while the variance, denoted by σ^2, is given by

$$\sigma^2 \approx \frac{\varepsilon^2}{3}, \tag{4}$$

which is a well-known property of the discrete uniform distribution.

The fitness landscape resulting from this model is called the "NK landscape", although there are several differences from the original NK landscape (Kauffman and Weinberger, 1989; Weinberger, 1991; Kauffman, 1993). In the case of $k = 0$, the fitness landscape has a single peak. As the k-value increases, the fitness landscape becomes more rugged. In spite of the ruggedness, there are so many slopes from the foot to the middle region, whereas local peaks with a large basin size are likely to appear at high altitudes on the landscape. In this paper, we focus on the cases where k is moderate or large, and describe these features quantitatively from an analytical approach.

In this paper, we describe several landscape properties, such as the occurrence probability of local optima, along one-dimensional fitness coordinate W. The probability density of the fitness W over all possible sequences is given approximately as the following normal distribution with the mean 0 and variance $\mathscr{V}$:

$$\mathscr{N}(W|0,\mathscr{V}) \quad \text{for} \quad -H \leqq W \leqq H, \tag{5}$$

where

$$H \equiv \varepsilon\nu \quad \text{and} \quad \mathscr{V} = \sigma^2\nu \approx \frac{\varepsilon^2\nu}{3} \tag{6}$$

(derivation is shown in Appendix A). The mean of fitness over the whole sequence space corresponds to the "foot" of the landscape, while regions where $W < 0$ corresponds to an "undersea" and is negligible for the adaptive walks that start from random points, which are likely to be at the foot. Since the fitness at the global peak takes about $H(=\varepsilon\nu)$,[2] then H corresponds to the height of the landscape from the foot to the global peak. In this paper, we focus on the regions from the foot to the global peak: $0 \leqq W \leqq H$.

In the NK model mentioned above, an arbitrary single-point mutation causes the changes in site-fitness at about $1 + k$ sites,

[1] In more familiar variants of the NK model, site-fitness values are assigned from a continuous (uniform or normal) distribution.

[2] This is not necessarily guaranteed for $k \geqq 1$.