BioSystems 101 (2010) 194–199

Contents lists available at ScienceDirect

BioSystems

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

# A trade-off relationship between energetic cost and entropic cost for *in vitro* evolution

Takuyo Aita

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

ARTICLE INFO

*Article history:*
Received 3 June 2010
Received in revised form 8 July 2010
Accepted 9 July 2010

*Keywords:*
Artificial evolution
Fitness landscape
Optimization
Information
Search algorithm
Sequence space

ABSTRACT

In this paper, we consider two complementary cost functions for the landscape exploring processes to obtain the global optimum sequence through *in vitro* evolution protocol: one is the entropic cost $C_{etp}$, which is based on the deviation from the uniformity of a mutant distribution in sequence space, and the other is the energetic cost $C_{eng}$, which is based on the total number of sequences to be generated and evaluated. Based on a prior knowledge about the structure of a given fitness landscapes, the conductor of the experiment can think up the efficient search algorithm (ESA), which requires the minimum number of points (=sequences) to be searched up to the global optimum. For five typical fitness landscapes, we considered their respective (putative) ESA, $C^*_{etp}$ and $C^*_{eng}$ based on the ESA. As a result, we found a trade-off relationship between $C^*_{etp}$ and $C^*_{eng}$ for every case, that is, $C^*_{eng} + C^*_{etp}$ is approximately equal to the logarithm of the volume of the sequence space. $C^*_{etp}$ and $C^*_{eng}$ are interpreted in terms of the information-theoretic concepts.



## 1. Introduction

*In vitro* evolution is comprehended as an optimization process on a given fitness landscape, where "fitness landscape" is defined by the mapping from sequences (=genotype) to the corresponding fitness values (Wright, 1932; Maynard-Smith, 1970). Recent development of this field has been reviewed in some books (e.g. Arnold, 2000) and review papers (e.g. Romero and Arnold, 2009). Ideally, it is desiable to find the global optimum sequence at the summit. The efficient search algorithm on fitness landscapes is deeply dependent on our knowledge about statistical orders of the structures of fitness landscapes (Wolpert and Macready, 1997), where "the efficient search algorithm (which is abbreviated to ESA)" is defined as a particular strategy that requires the minimum number of points (=sequences) to be searched up to the global optimum. For example, we can say that the fitness landscapes with high order, such as Fujiyama landscape, have a lot of information about their structures. That is, we know the mutational effect is additive for Fujiyama landscape. Therefore, we should think up the efficient search algorithm based on the available information.

Generally, the search process is conducted by two processes: mutagenesis and selection. In the mutagenesis process, controlling the distribution of mutant sequences in sequence space requires a "cost" that originates from the negative entropy. Therefore, we introduced the entropic cost: $C_{etp}$. If a site-directed mutagenesis is applied and the mutant sequences are distributed in a local area in sequence space, $C_{etp}$ is large due to strong control of mutations. This control is done by our intelligence or replication fidelity of replication enzyme (e.g. DNA polymerase) (Ishii et al., 1989; Cady and Qian, 2009). If mutant sequences are distributed randomly in sequence space, $C_{etp}$ is small due to no control of mutations. On the other hand, the energy consumed through all the processes is proportional to the total number $N$ of sequences to be generated and evaluated. Therefore, we introduced the energetic cost defined by $C_{eng} = \ln N$.

As a whole, there seems a trade-off relationship between energetic cost $C_{eng}$ and entropic cost $C_{etp}$ based on the corresponding ESA. For example, it is easy to find the global optimum sequence of Fujiyama landscape, because one has only to identify the fittest letter for each site by positional scanning method. In this case, $C_{eng}$ is small due to small number of $N$, while $C_{etp}$ is large due to strong control of mutations. On the other hand, as for random rugged landscape, an exhaustive search over the sequence space is necessary by performing random synthesis of all possible sequences. In this case, $C_{eng}$ is large due to vast numbers of $N$, while $C_{etp}$ is small due to no control of mutations.

In this paper, focusing on five typical fitness landscapes, we considered their respective (putative) ESA, $C_{etp}$ and $C_{eng}$ based on the ESA. Our aim is to describe quantitatively a trade-off relationship between $C_{eng}$ and $C_{etp}$, and to interpret them in terms of the information-theoretic concepts.

*E-mail address:* taita@mail.saitama-u.ac.jp.

0303-2647/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2010.07.003

## 2. Definition

### 2.1. Deviation from the uniformity of a mutant distribution in sequence space

We consider all conceivable sequences of $\lambda^\nu$, where $\nu$ is the number of all sites and $\lambda$ is the number of available letters for each site. The $\lambda$ is much larger than one ($\lambda \gg 1$). Then, an arbitrary sequence $s$ ($s = 1, 2, \cdots, \lambda^\nu$) is mapped into the corresponding point in the $\lambda$-valued $\nu$-dimensional sequence space.

Consider that sequences with an infinite population are distributed in sequence space according to the underlying probability distribution $Q(s)$ $\left(\sum_{s=1}^{\lambda^\nu} Q(s) = 1\right)$, where $Q(s)$ is a probability of being occupied by a sequence $s$ in sequence space. Let $d$ be the Hamming distance between two arbitrary sequences among the population, and let $P(d)$ $\left(\sum_{d=0}^{\nu} P(d) = 1\right)$ be the probability distribution of $d$ over all pairs of them. That is, the probability distribution $Q(s)$ is converted to the probability distribution $P(d)$ by

$$P(d) = \sum_{s=1}^{\lambda^\nu} \sum_{s'=1}^{\lambda^\nu} \delta(d(s, s'), d)\, Q(s) Q(s'), \tag{1}$$

where $d(s, s')$ is the Hamming distance between sequences $s$ and $s'$, and $\delta(x, x_0)$ is the Kronecker's delta defined by

$$\delta(x, x_0) \equiv \begin{cases} 1, & \text{if } x = x_0 \\ 0, & \text{if } x \neq x_0. \end{cases}$$

Let $B(d)$ be a back ground distribution where sequences are distributed randomly over the sequence space. As the deviation from the uniformity of a mutant distribution in the sequence space, we introduce the following relative entropy $D$:

$$D \equiv \sum_{d=0}^{\nu} P(d) \ln \frac{P(d)}{B(d)}, \tag{2}$$

$$B(d) = \binom{\nu}{d} \left(1 - \frac{1}{\lambda}\right)^d \left(\frac{1}{\lambda}\right)^{(\nu - d)}. \tag{3}$$

Eq. (3) is derived in Appendix A, based on the "profile" of given sequences (Gribskov et al., 1987).

### 2.2. Energetic cost and entropic cost for in vitro evolution

In this paper, we consider landscape exploring processes to obtain the global optimum sequence through *in vitro* evolution protocol. *In vitro* evolution is conducted by iterating the evolution cycle, which consists of generating mutant sequences of $N_t$, measuring fitness values of them and selecting the fittest one from among them. The $N_t$ sequences consist of heterogeneous sequences. The step number of the iteration process is denoted by $t$ ($t = 1, 2, 3, \cdots, t_e$). Suppose that the following two different costs are generated in the *in vitro* evolution: one is the "energetic cost" $C_{eng}$, which is based on the total number of sequences to be generated and evaluated, the other is the "entropic cost" $C_{etp}$, which is based on the deviation from the uniformity of a mutant distribution in sequence space.

Consider that, in the $t$ th step, mutant sequences of $N_t$ are generated according to the underlying probability distribution $Q_t(s)$ $\left(\sum_{s=1}^{\lambda^\nu} Q_t(s) = 1\right)$, where $Q_t(s)$ is a probability of being occupied by a sequence $s$ in sequence space. Note that the shape of $Q_t(s)$ is set up by the "conductor", who sets up the experimental algorithm and implements it. The probability distribution $Q_t(s)$ is converted to the probability distribution of the Hamming distance $d$ between arbitrary two sequences, $P_t(d)$ $\left(\sum_{d=0}^{\nu} P_t(d) = 1\right)$, by using Eq. (1). Therefore, we apply the measure $D$ defined in Eq. (2) to the probability distribution for the $t$ th step, $P_t(d)$:

$$D_t \equiv \sum_{d=0}^{\nu} P_t(d) \ln \frac{P_t(d)}{B(d)}. \tag{4}$$

We note the following two similar but different cases: one is the case where a single particular sequence $s^*$ is generated ($N_t = 1$), and the other is the case where a single arbitrary sequence $s$ is generated ($N_t = 1$). For the former case, $D_t = \nu \ln \lambda$, because $Q_t(s)$ is given by the Kronecker's delta, $Q_t(s) = \delta(s, s^*)$, and then $P_t(d) = \delta(d, 0)$. For the latter case, $D_t = 0$, because $Q_t(s)$ is the uniform distribution, $Q_t(s) = 1/\lambda^\nu$, and then $P_t(d) = B(d)$. Then, we define the "entropic cost" through the whole evolution process by

$$C_{etp} \equiv \langle D \rangle \equiv \frac{1}{t_e} \sum_{t=1}^{t_e} D_t. \tag{5}$$

In this paper, we use $\langle X \rangle$ as the arithmetic mean of a $t$-dependent quantity $X_t$ over the whole process from $t = 1$ to $t = t_e$.

The energy consumed through all the processes is proportional to the total number of sequences to be generated and evaluated. Let $N_{tot}$ be the number of points (=sequences) to be searched up to reaching the global optimum sequence: $N_{tot} = \sum_{t=1}^{t_e} N_t$. We define the "energetic cost" by

$$C_{eng} \equiv \ln N_{tot}. \tag{6}$$

The total cost is defined by

$$C_{tot} \equiv C_{eng} + C_{etp}. \tag{7}$$

Both $C_{etp}$ and $C_{eng}$ range from 0 to $\nu \ln \lambda$.

### 2.3. Typical fitness landscape and efficient search algorithm

Here, we define a "typical fitness landscape" as a set of all possible landscapes described by a given mathematical model with several parameters (e.g. Fujiyama landscape, NK landscape). A typical fitness landscape has its characteristic order and then the conductor knows it as a prior knowledge about the landscape. The conductor's job is to obtain the global optimum sequence of the given typical landscape by implementing an "efficient search algorithm", where the efficient search algorithm (which is abbreviated to ESA) is, in this paper, defined as a particular algorithm that requires the minimum number of points (=sequences) to be searched up to reaching the global optimum. This minimum number is denoted by $N^*_{tot}$. Based on the prior knowledge about the given landscape, the conductor can think up the corresponding ESA. $C_{etp}$, $C_{eng}$, and $C_{tot}$ based on the corresponding ESA are denoted by $C^*_{etp}$, $C^*_{eng}$ ($= \ln N^*_{tot}$), and $C^*_{tot}$, respectively.

The ESA gives the minimum value of $C_{eng}$. We can define another search algorithm that gives the minimum value of $C_{tot}$. However, determining this algorithm for a given landscape is so difficult. Therefore, we leave it and focus on the ESA in this paper.

## 3. Application to typical fitness landscapes

In this section, as for five typical fitness landscapes, we consider their respective ESA, $C^*_{etp}$, $C^*_{eng}$ and $C^*_{tot}$ based on the ESA. In some cases, although the conductor can think up a putative ESA for a given landscape, the putative ESA is however difficult to be proven true rigorously. Therefore, the ESAs we described below are putative ones. The results are complied in Table 1.

### 3.1. Additive fitness landscape (Fujiyama landscape)

In this case, each letter at every site in a given sequence contributes to the fitness independently and additively, that is,