

# Measuring Neutrality of Fitness Landscapes Based on the Nei's Standard Genetic Distance

Yoshiaki Katada<sup>1</sup>, Kazuhiro Ohkura<sup>2</sup>, and Kanji Ueda<sup>3</sup>

<sup>1</sup> Graduate School of Science and Technology, Kobe University, Kobe, JAPAN

<sup>2</sup> Faculty of Engineering, Kobe University, Kobe, JAPAN

Phone/Fax: +81-78-803-6135

Email: {katada, ohkura}@rci.scitec.kobe-u.ac.jp

<sup>3</sup> RACE (Research into Artifacts, Center for Engineering),

The University of Tokyo, Meguro, JAPAN

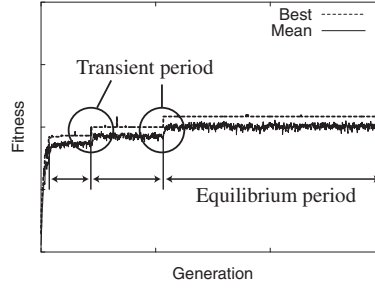
Phone: +81-3-5453-5887 Fax: +81-3-3467-0648

Email: ueda@race.u-tokyo.ac.jp

**Abstract.** In recent years, not only ruggedness but also neutrality has been recognized as an important feature of a fitness landscape for genetic search. Since the traditional theories based on the schema theory are not useful for problems including neutrality, new theoretical guidelines should be developed for effective genetic search. This research direction has been introduced almost for ten years, but the progress is not very fast. For instance, only a few methods have been proposed or tested for measuring neutrality on the problem to be solved. In this paper, we propose the use of the *Nei's standard genetic distance*, which originates from population genetics, for measuring the neutrality of a landscape. Several computer simulations are conducted by using a standard genetic algorithm in order to investigate the validity of the proposed approach. The terraced NK landscape, which is a popular test function in this field, is adopted for a test function. The results suggest to us that the Nei's genetic distance in natural evolution is a reliable method for measuring neutrality in artificial evolution after minor modifications.

## 1 Introduction

Neutrality have attracted much research interest in recent years [1][2]. This feature, caused by highly redundant mappings from genotype to phenotype, is also found in natural systems, and has been of particular interest to evolutionary theorists and molecular biologists [3]. Neutrality has been found in many real-world applications of artificial evolution, such as evolution of neural network controllers in robotics [4], on-chip electronic circuit evolution [5]. Landscapes which include neutrality have been conceptualized as containing *neutral networks*. Harvey [6] first introduced the concept of *neutral networks* into the GA community. He defined it as follows: "A neutral network of a fitness landscape is defined as a set of connected points of equivalent fitness, each representing a separate genotype: here connected means that there exists a path of single (neutral) mutations



**Fig. 1.** Typical evolutionary dynamics on a fitness landscape featuring neutral networks: Instead of being stuck in a local optimum, a population is exploring in the genotype space during the equilibrium period.

which can traverse the network between any two points on it without affecting fitness.”

The evolutionary dynamics on neutral networks can be classified into two kinds of period, i.e., a transient period and an equilibrium period (Fig. 1) [7][8]. During an equilibrium period, a population is located around the *dominant phenotype* in the genotype space maintaining in a cluster, analogously to *quasi-species* [9], and moves along the neutral pathway until it finds a portal to a next higher neutral network (Fig. 1). The discovery of the portal leads to a transient period which is expected to be very short compared to the equilibrium period. This evolutionary dynamics is clearly different from the traditional explanations based on the schema theory.

In the GA community, the majority of fitness landscape descriptions have based on ruggedness and modality [10–12]. This is because GA practitioners customarily avoid redundancy in the coding of problems for genetic search. However, Nimwegen *et al.* [7] suggested that the building block hypothesis[10][13] does not hold for systems that mainly show equilibrium periods. This means that the traditional theories based on the schema theory[10][13] are not useful for problems including neutrality. If equilibrium periods seem dominative, ruggedness is not enough to measure the search difficulty. Another measure, i.e., neutrality is required.

To the best of our knowledge, the **Smith’s fitness evolvability portraits [2] are only the method for measuring neutrality.** This is defined as the probability that the offspring fitness is greater than or equal to the parent fitness. In their measurement, a certain distinctive characteristic defined to be neutral between two fitness values must be set by GA practitioners. This value has a great influence on this measure, especially in the case that the fitness is evaluated as a real value or in a noisy environment. However, it has not been shown which distinctive characteristic should be considered to be neutral.

According to Kimura’s neutral theory [3] and Ohta’s nearly neutral theory [14][15], the number of gene substitutions of each genotype increases with the increase of neutrality. In population genetics, there are several statistical methods

for estimating the number of gene substitutions. From the viewpoint of measuring neutrality, the number of gene substitutions could be an index of neutrality. This idea would be beneficial in artificial evolution because there is no need to treat any fitness values.

This paper examines whether the Nei's standard genetic distance is adequate for measuring neutrality of fitness landscapes. The next section describes the Nei's standard genetic distance. Section 3 applies the Nei's genetic distance to describe one of tunably neutral landscapes called the Terraced NK landscape and shows the validity of the Nei's genetic distance for measuring neutrality. Section 4 discusses the condition that the Nei's genetic distance can be used. Conclusions are given in the last section.

## 2 The Nei's Standard Genetic Distance

Genetic distance is a term of population genetics used for estimating gene differences per locus between populations. Although there are several definitions for this, the Nei's standard genetic distance [16] is adopted in this paper.

The Nei's standard genetic distance is defined as follows. Consider two populations,  $X$  and  $Y$ . Let  $x_{ik}$  and  $y_{ik}$  be the frequencies of the  $k$ -th alleles ( $i = 1, \dots, N, k \in \{1, 2\}$  in the standard GA) in  $X$  and  $Y$ , respectively. The probability of identity of two randomly chosen genes is  $j_{xi} = x_{i1}^2 + x_{i2}^2$  in the population  $X$ , while it is  $j_{yi} = y_{i1}^2 + y_{i2}^2$  in the population  $Y$ . The probability of identity of a gene from  $X$  and a gene from  $Y$  is  $j_{xyi} = x_{i1}y_{i1} + x_{i2}y_{i2}$ . The normalized identity of genes between  $X$  and  $Y$  with respect to a locus is defined as

$$I_i = \frac{j_{xyi}}{\sqrt{j_{xi}}\sqrt{j_{yi}}}. \quad (1)$$

It is equal to 1.0 if the two populations have the same alleles in identical frequencies, and 0.0 if they have no common alleles. The normalized identity of genes between  $X$  and  $Y$  with respect to the average in all loci is defined as

$$I = \frac{J_{XY}}{\sqrt{J_X}\sqrt{J_Y}}, \quad (2)$$

where,  $J_X = \sum_{i=1}^N j_{xi}/N$ ,  $J_Y = \sum_{i=1}^N j_{yi}/N$  and  $J_{XY} = \sum_{i=1}^N j_{xyi}/N$ . The genetic distance between  $X$  and  $Y$  is defined as

$$D = -\log_e I, \quad (3)$$

under the assumption that the mutation rate per locus is sufficiently small. However, the above definition cannot be applied to the standard GA directly, because it is assumed that a new allele always appears on a locus when a mutation occurs, while "back mutations [14]" frequently occur in the standard GA, due to the binary coding scheme. Therefore, the genetic distance between the population at the initial generation and the one at the last generation is calculated

as:

$$D_{final} = \sum_1^{T-1} D_{t,t+1} \quad (4)$$

where  $T$  is the number of the last generation and  $D_{t,t+1}$  is the genetic distance between the population in the  $t$ -th and the  $(t + 1)$ -th generation. The rate of gene substitution is defined as the genetic distance per generation.

Based on the studies of population genetics, the following hypotheses are needed:

- (1) The number of substitutions increases with the increase of neutrality.
- (2) The number of substitutions is largest when random sampling is applied as a selection method.
- (3) The mutation rate per locus is small so that the above two expectations hold.

The Nei's genetic distance can be applied as the measure of neutrality in artificial evolution as long as the above hypotheses hold.

### 3 Neutrality in a Tunably Neutral NK Landscape

#### 3.1 A Terraced NK Landscape

A terraced NK landscape was employed as the test function in our computer simulations. This is the tunably neutral landscape proposed by Newman *et al.* [17]. A terraced NK landscape has three parameters:  $N$ , the length of the genotype;  $K$ , the number of epistatic linkages between genes; and  $w$ , the contribution of a locus to the fitness of the entire genotype.

The fitness value is calculated as follows: The fitness contribution of the  $i$ -th locus,  $w_i$ , is an integer generated randomly in the range  $0 \leq w_i < F$ ,  $i = 1, \dots, N$ . To calculate the fitness,  $W$ , of a genotype, the fitness contribution of each locus is averaged, and then divided by  $F - 1$ , normalizing  $W$  to the range 0.0 to 1.0. More formally:

$$W = \frac{1}{N(F-1)} \sum_{i=1}^N w_i. \quad (5)$$

The neutrality of the landscape can be tuned by changing the value of  $F$ . The neutrality of the landscape is maximized when  $F = 2$ , and is effectively non-existent as  $F \rightarrow \infty$ .

#### 3.2 Simulation Conditions

We applied two genetic algorithms: the standard GA (SGA) and the (*random-sampling*,  $q$ )-algorithm. The (*random-sampling*,  $q$ )-algorithm employs standard bit mutation at the rate of  $q$  as the genetic operation and random sampling

as a selection method where  $M$  offsprings are sampled from  $M$  ancestors with replacements. This model was used to investigate the effect of random sampling and mutation on the genetic distance and the rate of gene substitution of a population. This model is approximately equivalent with Kimura's stochastic genetic models to study random genetic drift and the expected time of fixation of a mutant gene [3]. Kimura has suggested that the rate of gene substitution is largest when the selective advantage of a new mutation over the original allele is zero except that the new mutation is deleterious in a small population.

Computer simulations were conducted using a population size of 50. The SGA used standard bit mutation as the genetic operation. For both algorithms, the per-bit mutation rate,  $q$ , was set at 0.008, based on the hypothesis (3) in Section 2. Crossover was not employed. Tournament selection was adopted for the SGA. The tournament size was set at 2. Each run lasted 2,000 generations. We conducted 50 independent runs for each problem under the landscape parameters,  $N = 20$ ,  $K \in \{0, 2, 6, 12, 19\}$ ,  $F \in \{2, 3, 4, 6\}$ . The results were averaged over 50 runs.

### 3.3 Simulation Results

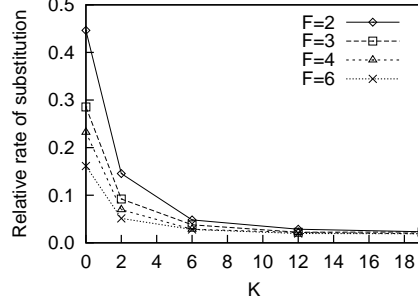
In this experiment, the number of substitutions increased approximately linearly over generations in all runs. Therefore, the rate of substitution is shown by using the method of least squares on the results of all runs because the rate of substitution is equivalent to the gradient of the number of substitutions over generations.

Table 1 shows the rate of substitution for the SGA. Notice first that the rate of substitution increased with the decrease of  $F$  with all  $K$ s. This means that the rate of substitution increases with the increase of neutrality. This tendency is consistent with the neutral theory. Second, the rate of substitution decreased with the increase of  $K$  with all  $F$ s. This means that not only neutrality but also ruggedness have an influence on the rate of substitution. This is consistent with Ohta's results for NK landscapes with weak selection based on the nearly neutral theory, where the number of substitutions decreases with the increase of  $K$  [15, 18].

The rate of substitution for (*random-sampling*,  $q$ )-algorithm was 0.004576. Thus, the rate of substitution for (*random-sampling*,  $q$ )-algorithm was always larger than any others for the SGA with  $K$  and  $F$ .

**Table 1.** The rate of substitution for the SGA with  $q = 0.008$

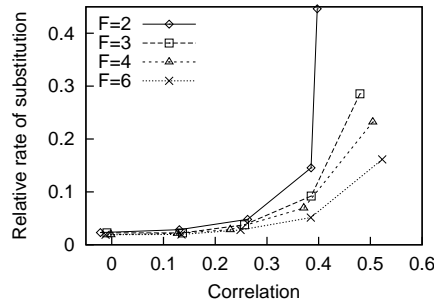
$\begin{smallmatrix} & K \\ F & \end{smallmatrix}$	0	2	6	12	19
2	0.002042	0.000666	0.000220	0.000131	0.000106
3	0.001307	0.000421	0.000173	0.000102	0.000104
4	0.001063	0.000318	0.000130	0.000099	0.000087
6	0.000738	0.000235	0.000128	0.000090	0.000087



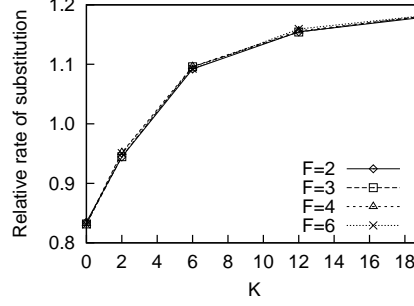
**Fig. 2.** Relative rate of substitution for the SGA and the  $(random\text{-}sampling, q)$ -algorithm with  $q = 0.008$

In order to summarize the results obtained in this experiment, the relative rate of substitution,  $\gamma$ , for the SGA and the  $(random\text{-}sampling, q)$ -algorithm was calculated. This is formulated as follows:  $\gamma = \alpha / \alpha_{random}$ , where  $\alpha$  is the rate of substitution for the SGA and  $\alpha_{random}$  is the one for the  $(random\text{-}sampling, q)$ -algorithm. When  $\gamma < 1.0$ , the rate of substitution for the SGA is less than the one for the  $(random\text{-}sampling, q)$ -algorithm. When  $\gamma > 1.0$ , the rate of substitution for the SGA is more than that for the  $(random\text{-}sampling, q)$ -algorithm. The results are shown in Fig. 2. We can recognize the tendencies for  $K$  and  $F$  more clearly. Thus, the hypothesis (1) and (2) hold when the landscape includes neutrality.

In real-world problems, the increase of ruggedness of a fitness landscape is predicted by the decrease in the correlation length [2, 11]. Fig. 3 shows the relative rate of substitution at each correlation coefficient corresponding to  $K$  with all  $F$ s in our results, where the correlation coefficient was calculated as a fitness correlation between genotypes at Hamming distance 1 [12]. Thus the relative rate of substitution predicts the increase of neutrality for  $F$ s around each correlation coefficient.



**Fig. 3.** Relative rate of substitution as a function of the correlation coefficient



**Fig. 4.** Relative rate of substitution for the SGA and the  $(\text{random-sampling}, q)$ -algorithm with  $q = 0.1$

## 4 Discussion

Additional computer simulations were conducted to observe the transition of the Nei's genetic distance in the condition of a large mutation rate, which induces the inconsistency with the hypotheses at Section 2. The simulation settings were the same as Section 3.2 except that  $q$  was set at 0.1 for the two algorithms. The results are shown in Fig. 4. Surprisingly,  $\gamma$  increased with the increase of  $K$  with all  $F$ s. In addition to this, no significant differences were found between the graphs of different  $F$ s.

The rate of substitution for the SGA was higher than that for the  $(\text{random-sampling}, q)$ -algorithm when  $K > 6$ , because  $\gamma > 1.0$  with all  $F$ s. This would be caused by the effect of the larger mutation rate than the *error threshold*[7] which changes artificial evolution into a random search.

From the above, we confirmed that the Nei's genetic distance can be used as long as the mutation rate is sufficiently small compared with the error threshold.

## 5 Conclusions

In this paper, we have proposed the use of the Nei's standard genetic distance, which treats only genotype data, for measuring neutrality of a landscape, and shown the validity of this approach by applying it to the Terraced NK landscapes. The characteristic of the rate of substitution estimated by the Nei's genetic distance can be summarized as follows:

When the mutation rate per locus is small,

- The number of substitutions is largest when random sampling is applied with mutation.
- The number of substitutions increases with the increase of neutrality.
- The number of substitutions decreases with the increase of ruggedness where the landscape includes neutrality.

Our results suggest that the genetic distance can provide a guideline for measuring neutrality and ruggedness. As for the next step, we will investigate how well our method applies to real-world problems whose fitness landscape includes

neutral networks, such as the evolution of artificial neural networks for robot control.

## References

1. Ebner, M., Langguth, P., Albert, J., Shackleton, M., Shipman, R. (2001) "On Neutral Networks and Evolvability," In *Proceedings of the 2001 IEEE Congress on Evolutionary Computation: CEC2001*, IEEE Press pp. 1-8.
2. Smith, T. Husbands, P. Layzell, P. and O'Shea, M. (2002) "Fitness Landscapes and Evolvability," *Evolutionary Computation*, 10(1):1-34.
3. Kimura, M. (1983) "The Neutral Theory of Molecular Evolution," Cambridge University, Press, New York.
4. Harvey, I. (1997) "Artificial Evolution for Real Problems," In Gomi, T., editor, *Evolutionary Robotics: From Intelligent Robots to Artificial Life (ER'97)*, AAI Books.
5. Thompson, A. (1996) "An Evolved Circuit, Intrinsic in Silicon, Entwined with Physics," In *Proceedings of the first International Conference on Evolvable Systems: From Biology to Hardware*, pp. 390-405.
6. Harvey, I., Thompson, A. (1996) "Through the Labyrinth Evolution Finds a Way: A Silicon Ridge," In *Proceedings of the first International Conference on Evolvable Systems: From Biology to Hardware*, pp. 406-422.
7. Nimwegen, E., Crutchfield, J., and Mitchell, M. (1999) "Statistical dynamics of the royal road genetic algorithm," In *Theoretical Computer Science*, Vol. 229, No. 1, pp. 41-102.
8. Barnett, L. (2001) "Netcrawling - Optimal Evolutionary Search with Neutral Networks," In *Proceedings of the 2001 IEEE Congress on Evolutionary Computation*, pp. 30-37.
9. Eigen, M., McCaskill, J. and Schuster, P. (1989) "The Molecular Quasi-species," In *Advances in Chemical Physics*, 75, pp. 149-263.
10. Goldberg, D. (1989) "Genetic Algorithms in Search, Optimization and Machine Learning," Addison-Wesley.
11. Weinberger, E. D. (1990) "Correlated and uncorrelated fitness landscapes and how to tell the difference," *Biological Cybernetics*, 63:325-336.
12. Manderick, B., Weger, M. and Spiessens, P. (1991) "The Genetic Algorithm and the Structure of the Fitness Landscape," In Belew, R. and Booker, B, editors, *Proceedings of the Fourth International Conference on Genetic Algorithms*, pp. 143-150, Morgan Kaufmann, San Mateo, California.
13. Holland, J. H. (1975) "Adaptation in Natural and Artificial Systems," University of Michigan Press.
14. Ohta, T. (1992) "The nearly neutral theory of molecular evolution," In *Annu. Rev. Ecol. Syst.*, 23:263-286.
15. Ohta, T. (1998) "Evolution by nearly-neutral mutations," In *Genetica*, 102/103, pp. 83-90.
16. Nei, M. (1972) "Genetic Distance between Populations," In *The American Naturalist*, Vol. 106, pp. 283-292.
17. Newman, M. and Engelhardt, R. (1998) "Effect of neutral selection on the evolution of molecular species," In *Proceedings of the Royal Society of London B*, Morgan Kaufmann, 256, pp. 1333-1338.
18. Ohta, T. (1997) "Role of random genetic drift in the evolution of interactive systems," In *Journal of Molecular Evolution*, Vol. 44, pp. S9-S14.