World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:1, No:5, 2007

Error-Robust Nature of Genome Profiling Applied for Clustering of Species Demonstrated by Computer Simulation

Shamim Ahmed and Koichi Nishigaki

***Abstract*—**Genome profiling (GP), a genotype based technology, which exploits random PCR and temperature gradient gel electrophoresis, has been successful in identification/classification of organisms. In this technology, spiddos (S*pecies identification dots*) and *PaSS* (*Pattern similarity score*) were employed for measuring the closeness (or distance) between genomes. Based on the closeness (*PaSS*), we can buildup phylogenetic trees of the organisms. We noticed that the topology of the tree is rather robust against the experimental fluctuation conveyed by spiddos. This fact was confirmed quantitatively in this study by computer-simulation, providing the limit of the reliability of this highly powerful methodology. As a result, we could demonstrate the effectiveness of the GP approach for identification/classification of organisms.

***Keywords*—**Fluctuation, Genome profiling (GP), Pattern similarity score (*PaSS*), Robustness, Spiddos-shift.

I. Introduction

ADVANCES in methods and technologies have enabled us to know more and more in detail of biological systems. Now, we can, in principle, obtain the whole genome sequence of almost all organisms. However, biological systems are too complicated and sophisticated for us to know the whole of them even in this post-genome era. We can not freely experiment and utilize the genome information as a whole. Thus, this fact gave an impetus to the emergence of systems biology. We have too less tools to dig out significant information out of genomes. In this context, we have developed such a tool termed Genome profiling (GP) [1].

GP is a technology that enables the genotype-based identification of species. Traditionally, organisms have been identified and classified on the basis of their phenotypes. Conventional techniques, however, face difficulties in such cases as classifying characterless organisms like microbes [2] and analyzing communities composed of a huge number of various organisms [3] owing to both of the instability of phenotypes, which are easily affected by environmental factors [4], and the insufficiency in the number of experts [5]. Recently, genotype-based approach has become possible owing to the development of sequencing technology. However, it is still difficult to apply sequencing approaches to the analysis of a large number of species due to logistic reason. In most biological fields, the analysis of complex systems comprising various species has been an important theme, demanding an effective method for handling a vast number of species. A realistic solution to these problems has been to characterize organisms according to the sequence of their small subunit ribosomal RNA (16S/18S rRNA), an approach that has been applied to various organisms, initiated by Woese and his collaborators [6]~[8]. Similarly, cytochrome oxygenase subunit 1 (COX1), gyrase, and other genes have been used for this purpose [9]. The superiority of these approaches is that they are based on the popular and well-established sequencing technology and can provide the determinate result of the nucleotide sequence, which can be further computer-analyzed and can fuel the activity of Bioinformatics. Nevertheless, this approach cannot be said to be a readily usable method for classifying species because (i) it is rather costly and time-consuming for applying to a large number of species (e.g., >100), especially for scientists in general all over the world, and (ii) it often results in an insufficient amount of information for identifying and classifying species [9]. The latter problem can be overcome by sequencing additional genes [9]~[11]; however, this makes the approach more complicated and less accessible. In our former studies [12], we have presented a solution for the universal classification of species together with demonstrations of its effectiveness, which includes a test applying it to taxonomically well-established organisms such as plants, fish, and insects with obtaining a successful result [12]. Owing to its convenience and its highly informative nature, this technique of classification based on GP can be widely applied to biological researches in general.

Technologically, Genome profiling (GP) is based on a temperature gradient gel electrophoresis (TGGE) analysis of random PCR products [1]. For the sake of data refinement, a computer-aided technology such as introduction of species identification dots (spiddos), which correspond to structural

Koichi Nishigaki is with the Graduate School of Science and Engineering, Department of Functional Materials Science, Saitama University, Saitama 338-8570, Japan (phone: +81-48-858-3533; fax: +81-48-858-3533; e-mail: koichi@fms.saitama-u.ac.jp)

Shamim Ahmed is with the Graduate School of Science and Engineering, Department of Functional Materials Science, Saitama University, Saitama 338-8570, Japan (e-mail: shamim1174@yahoo.com)

transition points of DNAs and pattern similarity score (*PaSS*) was developed [13]. *PaSS* was shown to be usable for quantitatively measuring the closeness or distance between genomes [13]. To our surprise, the quantitative expression of *PaSS* was proved to be very effective, even though the accuracy of the measure given by *PaSS* is assumed to be limited, *a priori*, due to its stochastic nature [14]~[16]. In Genome profiling, there are some steps that are stochastic in nature and can influence determination of the *PaSS* value: for example, random PCR may or may not select a DNA fragment containing mutations, and the degree of displacement of spiddos caused by a point mutation depends on the type of mutation such as A to G or A to T substitution [17]-[18]. Especially, it was our great surprise that all kinds of organisms dealt (fish, plants, insects) were classified in a complete match with the traditional classifications, which were established based phenotypes, using only a single genome profile for each organism (Figure 1). Since the hierarchy gaps between taxa employed in the phenotype-based classification are arbitrary and are set to be equal, there is no quantitative meaning in the apparent distance of phenotype-based classification. On the other hand, the species-to-species distance expressed in genotype-based one has a semi-quantitative meaning, which is very intriguing. This fact immediately posed us with three questions:

1) 1. How generally applicable is this technology with such a limited amount of information (i.e., that obtained by a single primer) or how robust is this approach? In other words, how much data (how many genome profiles) is required to meet with a truly universal identification/classification of species?

2) Why did these quite different approaches, one is phenotype based and another is genotype based, provide the same classification result?

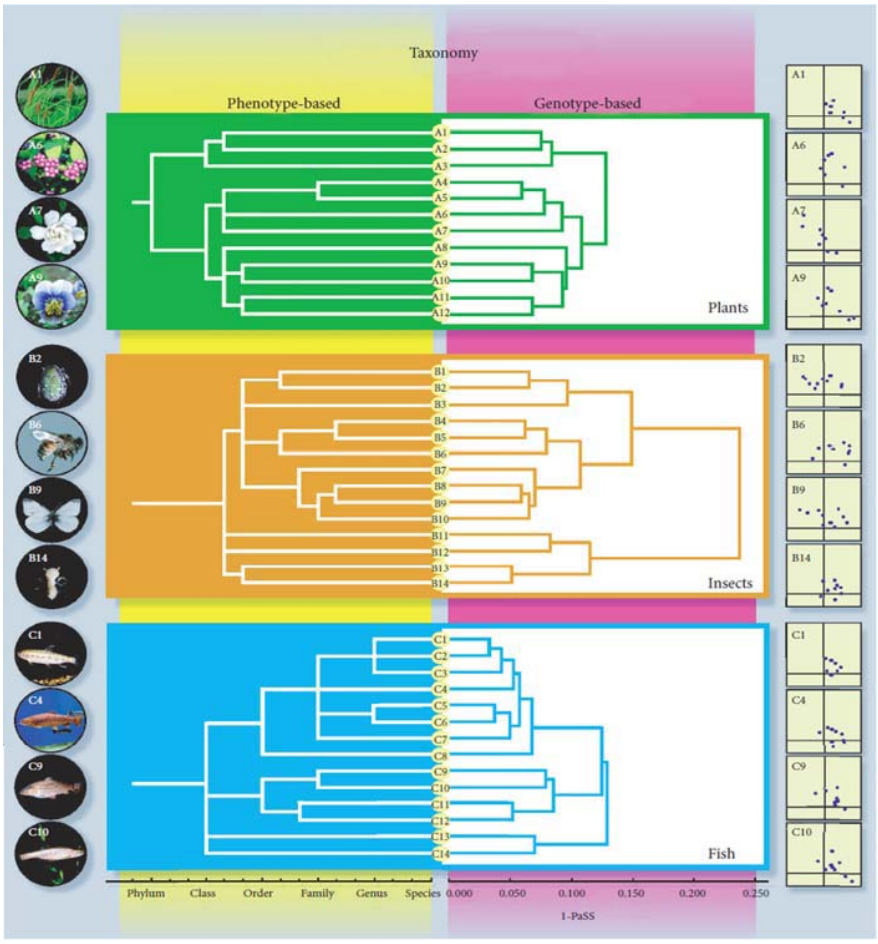


Fig. 1 Phylodendrons of plants (A1 ☐ A12), insects (B1 ☐ B14), and fish (C1 ☐ C14). Phenotypic (left) and genotypic (right) trees are drawn on the basis of taxonomic hierarchy or *PaSS* value, respectively. The same nomenclatures of these organisms are appearing in appendix as supplementary Table 1. Photographs (leftmost) and *spiddos* (rightmost) are included to illustrate the technique. Trees were drawn by the *group average method* (plants) or the *median method* (insects and fish). Figure was taken from ref. 12, (*International Journal of Plant Genomics* has a policy of free distribution).

3. How much is the distance given by the GP approach accurate? Namely, what is the nature of the genome distance defined here?

All of these problems have a profound meaning and are very challenging.

In this study, we tackled the first problem, i.e. "robustness", by building an *in-silico* model experiment. That is based on the simulation experiments of GP; we generated five genomes, each containing eight spiddos generated at random. The effect of *spiddos-shift* on the score of *PaSS* (or genome distance) and then, the consequent clustering result was analyzed, representing the degree of robustness of GP from the matching ratio between the results obtained by the phenotype-based and genotype-based approaches.

II. Methodology

A. Genome Profiling (GP)

Preparation of DNA is carried out by the alkaline extraction method in general except a few special cases [19]. Briefly, the procedures adopted are as follows: 1) An aliquot containing cells is transferred into an Eppendorf tube; 2) After adding 3 µl of 0.5 M NaOH, the sample solution is incubated at 94°C for 5 min and then at 64°C for 60 min; 3) the sample solution is neutralized with 5 µl of 200 mM Tris-HCl (pH 8.0) buffer, and incubated at 65°C.

GP is composed of two major experimental steps: random PCR and temperature gradient gel electrophoresis (TGGE) (The whole procedure is shown in Fig. 2). Random PCR is a process in which DNA fragments are sampled at random from genomic DNA through a mismatch containing hybridization of a primer to a template DNA during PCR [20]. Random PCR can be performed using a single primer of dodeca-nucleotides (pfM12, dAGAACGCGCCTG) with the 5'-end Cy3-labeled. This primer sequence has been recommended for general use including the application to animal cells [21]. The PCR reaction (50 µl) usually contains 200 µM dNTPs (N= G,A,T,C), 0.5 µM primer, 10 mM Tris-HCl (pH 9.0), 50 mM KCl, 2.5 mM $MgCl_2$, 0.02 unit/µl Taq DNA polymerase (Takara Bio, Shiga, Japan) and a particular amount of template DNA. Random PCR is carried out with 30 cycles of denaturation (94°C, 30 s), annealing (26°C, 2 min) and extension (47°C, 2 min) using e.g., a PTC-100TM PCR machine (MJ Research, Inc., Massachusetts, USA). The DNA samples are subjected to µ-TGGE [3], which adopts a tiny slab gel of 24 × 16 × 1 mm^3 for electrophoresis using a temperature-gradient generator, µ-TG (Taitec, Saitama, Japan). In each run of electrophoresis, an internal reference DNA is co-migrated. The 200-bp reference DNA (the 191-bp bacteriophage fd gene VIII, sites 1350~1540 attached to a 9-bp sequence, CTACGTCTC, at the 3'-end) is experimentally determined to have a melting temperature of 60°C under standard conditions. The gel used is composed of 6% acrylamide (acrylamide:bis = 19:1) containing 90 mM Tris-HCl (pH 8.0), 90 mM boric acid, 2 mM EDTA and 8 M urea. The linear temperature gradient is run from 15°C to 65°C. After electrophoresis, DNA bands are detected with a fluorescence imager (e.g., Molecular Imager FX, Biorad, Hercules, CA) or by silver staining [22].

B. Data processing employing spiddos and PaSS

Genome profiles obtained by GP technology are highly informative but difficult to manage due to their complexity. However, this inconvenience could be overcome by introducing the featuring points, designated as spiddos (species identification dots), which can represent genome profiles compactly [13]. The featuring points, or spiddos, correspond to the points where structural transitions of DNA occur, such as double-stranded to single-stranded DNA [23]. Spiddos can be used to provide a sufficient amount of information for identifying species [13]. Using spiddos, we can define the pattern similarity score (*PaSS*) between two genomes as follows:

$$PaSS = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\vec{P}_i - \vec{P}_i'\right|}{\left|\vec{P}_i\right| + \left|\vec{P}_i'\right|} \qquad (0 \le PaSS \le 1) \qquad (1)$$

where $\vec{P}_i$ and $\vec{P}_{i'}$ correspond to the normalized positional vectors (composed of two elements, mobility and temperature) for spiddos P_i and $P_{i'}$ collected from two genome profiles (discriminated with or without a prime), respectively, and *i* denotes the serial number of spiddos. A database site has been constructed (*On-web GP* [28]) in order to provide semi-automatic data processing [24]. The *PaSS* value thus introduced is empirically known to be a good measure to quantify the closeness or the distance between two species (or cells) [13].

C. Genome Distance

Genome distance as a practical form (*d'*) [16] can be defined by deriving from *PaSS* as follows:

$$d' = 1 - PaSS \qquad (2)$$

However, the distance, *d',* has been introduced here does not have a nature of the conventional distance that of Descartesian, like $\left|\vec{d}_{1,3}\right| = \left|\vec{d}_{1,2} + \vec{d}_{2,3}\right|$ (where $\vec{d}_{i,j}$ means position vector), but still have a, though non-linear, additive nature. If *d'* is sufficiently small (*d'*<< 1), it means that the two genomes of interest belong to the same species [3]. Genome distance, *d'*, which is experimentally obtainable, can serve as a convenient substitution for the true genome distance which needs to be discussed in relation to genetic distance [25], although it leaves a lot to be theoretically refined.

D. Cluster Analysis

Sato and others in our laboratory have developed a clustering/displaying program termed FreeLighter on the basis of Ward's method [26]-[27], which is a type of nearest neighbor method with an objective function of minimizing the "error sum of squares". These methods are based on the distance defined in Eq.3 which implies that Clusters *a* and *b* are to be merged into *c*, and *x* is an arbitrary cluster:

$$d_c = \alpha_a d_{xa} + \alpha_a d_{xb} + \beta d_{xb} + \gamma |d_{xa} d_{xb}|, \quad (3)$$

where α_a, α_b, β_a, and γ are weighing parameters, d_c, d_{xa}, d_{xb}, and d_{ab} represent distances between relevant clusters such as Cluster *x* and Cluster *a* for d_{xa}.

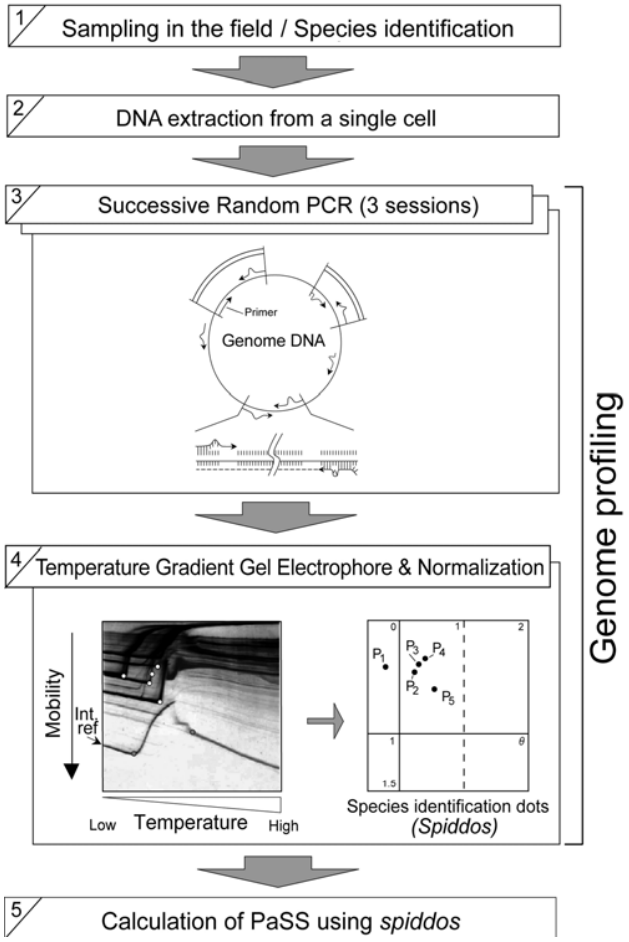


Fig. 2 The procedure used to identify species by GP: Random PCR was carried out. Note that primer binding occurs in a mismatch-containing structure in this relaxed mode of PCR (random PCR), thus enabling us to sample DNA fragments from various sites of the genome. In TGGE, DNA fragments layered on the top of a slab gel migrate downward in a horizontal line with a characteristic curvature caused by the temperature gradient. Featuring point(s) of each DNA fragment are assigned and processed to generate species identification dots (spiddos) with a computer. The *PaSS* (pattern similarity score) calculation is performed as described in the text. This figure was taken from *BMC Genomics* [16].

E. Evaluation of robustness in GP-based clustering

The following steps were adopted for the evaluation of robustness.

Step 1: Five set of spiddos representing genome profiles, A, B, C, D and E, were generated at random using Rnd function of Visual Basic 6 with eight spiddos contained for each. The ranges for the coordinates for spiddos, mobility and temperature, were set to be 0.1 to 1.0 and 15 to 65 ^{o}C, respectively (Table 1).

Step 2: Using Table 1, genome distances between a pair of genomes were calculated (Table 2a).

Step 3: Random numbers were generated between the range of -0.2 to 0.2 for the mobility and -5 to 5 for the temperature and were added to the corresponding coordinate of a particular spiddos of genome A (see Table 1). Then shifted genome distance between a pair of genomes was calculated (Table 2b). Trials were done by 10000 times for each random shift of a spiddos, thus generating 10000 of similar tables. The degree of spiddos shift was evaluated as shift (s) and recorded in each time step.

Step 4: Step 3 was repeated with changing the spiddos to be shifted into spiddos 2, 3, 4, 5, 6, 7 and 8 of genome A. the effects of double and quadruple spiddos shifting were also measured by the same way.

Step 5: using Tables 2a (genome distance) and 2b (shifted genome distance), clustering analyses were performed adopting *FreeLighter* program to generate phylogenetic trees.

Step 6: If the phylogenetic trees obtained for Tables 2a and 2b were topologically the same, then it scored 1, else 0 (Figure 4).

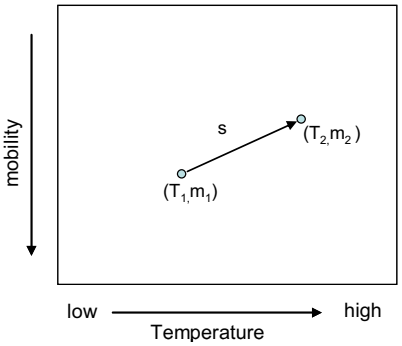
Step 7: Statistics was taken for all the results thus obtained: Scores (1 or 0) were collected and plotted against their corresponding delta shift (Figure 5).

III. Results and Discussion

Genome profiling has been shown to be applicable to a variety of purposes. To improve the performance of GP technology, various factors that affect GP sensitivity and reproducibility should be elucidated. In this sense, the robustness of GP results is of great interest.

In this study, an *in-silico* model experiment was performed to investigate the robustness in the clustering result. For this, five genome profiles (A, B, C, D and E), each containing eight spiddos, were generated (Table 1), and one or more spiddos were perturbed at random for both mobility and temperature coordinates within a range of 0.1 to 1.0 and 15^{o}C to 65 ^{o}C, respectively, mimicking the real GP experiments (Figure 3). Table 1 was used for calculating genome distance, *d'*, between the genomes using Eq. 2 (Table 2a). Since the *PaSS* value is governed by stochastic events (for example, random PCR may or may not select a DNA fragment containing mutations, and the degree of displacement of spiddos depends on the type of point mutation (A to G or A to T or else)), spiddos-shift was arbitrarily generated in its degree, selected point, and the

number of spiddos shifted. Table 2b represents shifted genome distance thus obtained ten thousands of similar tables generated. Obviously, in Table 2b, those cells which have no relation to genome A kept constant, and the degree of change in the relevant cells differs from cell to cell. Both Tables 2a and 2b were subjected to clustering analysis using *FreeLighter* program. In the same way clustering results were obtained for the shifting of spiddos 1, 2, 4, 5, 6, 7 and 8 separately and represented the same figure like Fig. 5a. As shown in Fig. 4, if the phylogenetic tree kept constant topologically after the operation of spiddos-shift, then the robustness score, γ, was set unity, otherwise 0. Statistical representations are provided for these results in Fig. 5. In Figure 5, the average robustness, $\overline{\gamma}$, is defined to be;



Fig, 3 Spiddos-shift. The shift (s) from P_1 (T_1,m_1) to P_2 (T_2,m_2) is shown, which may be caused by point mutation and/or insertion/deletion mutation in the corresponding DNA sequence.

TABLE I GENOME PROFILES ARBITRARILY GENERATED AND USED FOR THIS SIMULATION EXPERIMENT

	Genome A		Genome B		Genome C		Genome D		Genome E	
	Mobility	Temp	Mobility	Temp	Mobility	Temp	Mobility	Temp	Mobility	Temp
Sipddos 1	0.77	54	0.34	51	0.52	60	0.94	64	0.94	50
Spiddos 2	0.5	52	0.4	39	0.25	26	0.99	57	0.67	16
Spiddos 3	0.69	43	0.92	56	0.44	48	0.44	46	0.62	55
Spiddos 4	0.77	63	0.67	45	0.52	16	0.19	36	0.94	22
Spiddos 5	0.65	51	0.88	63	0.4	56	0.41	54	0.58	62
Spiddos 6	0.48	24	0.32	55	0.17	41	0.75	61	0.59	32
Spiddos 7	0.71	21	0.23	30	0.41	23	0.65	20	0.83	29
Spiddos 8	0.42	52	0.26	39	0.77	70	0.45	29	0.29	60

TABLE II GENOME DISTANCES (A) AND ONE EXAMPLE OF SHIFTED GENOME DISTANCES (B), IN WHICH A PERTURBATION EFFECT OF SPIDDOS SHIFT WAS INCORPORATED

(a)

	A	B	C	D	E
A	0	0.026	0.017	0.012	0.027
B	0.026	0	0.026	0.025	0.036
C	0.017	0.026	0	0.017	0.029
D	0.012	0.025	0.017	0	0.032
E	0.027	0.036	0.029	0.032	0

(b)*

	A	B	C	D	E
A	0	0.025	0.018	0.016	0.032
B	0.025	0	0.026	0.025	0.036
C	0.018	0.026	0	0.017	0.029
D	0.016	0.025	0.017	0	0.032
E	0.032	0.036	0.029	0.032	0

*One of 10,000 different tables

$$\overline{\gamma(s_i)}=\left(\sum_e \gamma(s_j)\,|\,s_i \le s_j < s_i+\Delta s\right)\Big/\left(\sum_e 1\right) \qquad (4)$$

where $\sum_e$ stands for taking the summation of the flanking term over all of the relevant events and Δx is the interval of sections.

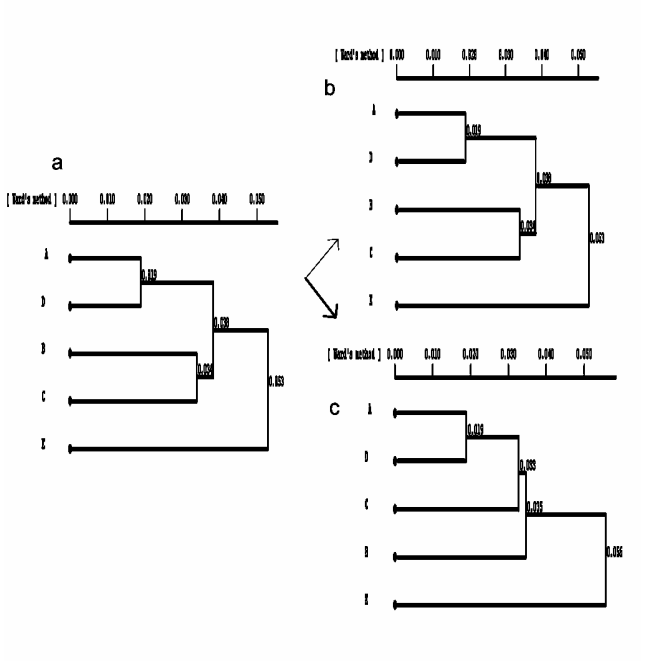


Fig. 4 Scoring 1 or 0. If phylogenetic tree kept constant topologically after the spiddos-shift operation (a→b), then robustness score, γ, was put 1 while, if changed as in the case of a→c, γ = 0

The average robustness was found to be rather high (~0.03) and monotaneous as shown in Fig. 5a & 5e provided that a single spiddos-shift was applied (except for the case of spiddos 3). This has a profound meaning since the error range of GP experiments is established to be less than 0.01 (that is 1%) [13]-[14] and it is known to be further diminished to ~0.5% by introducing a double internal reference method (which adopts two independent internal reference molecules instead of the current single one to raise the accuracy (to be published elsewhere)). As the number of spiddos-shift increases from 1 to 4 (Fig. 5a, b and c), s_{50} is gradually increasing than the theoretically expected value (Fig. 5f), indicating the random canceling effect of accumulated spiddos-shifts (each of them occurs independently without any bias or orientation). Among these simulations, Fig. 5d offers a noteworthy result though it can not be completely rationalized nor confirmed yet. The apparent phenomenon of oscillation may be due to the crowdedness of the area with spiddos where spiddos 3 is located in the GP plane. This may be the reason why a small fluctuation of the coordinate, i.e. spiddos-shift, can result in a large difference in the clustering result in which the neighboring effect is weighed.

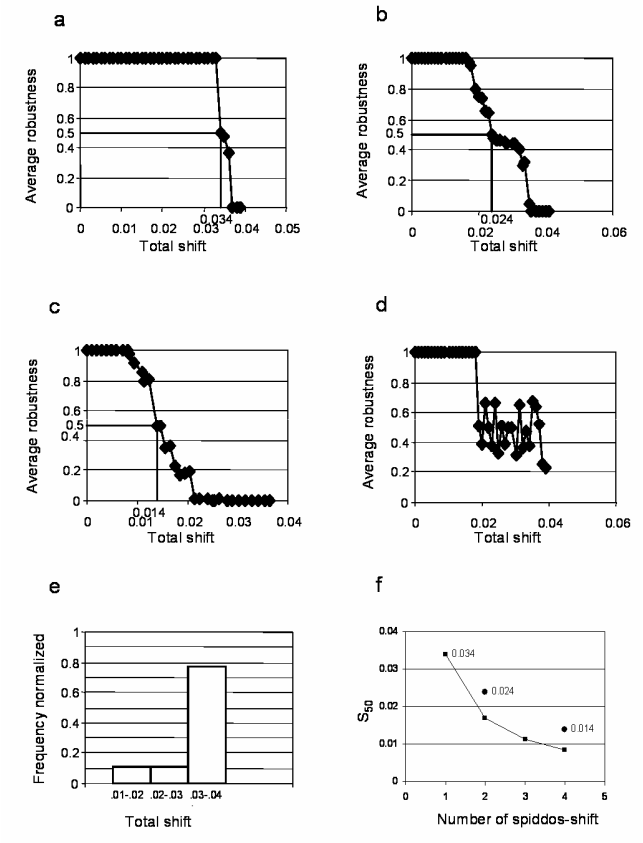


Fig. 5 Average robustness of the phylogenetic tree topology against perturbations (spiddos-shift in GP) (a) single spiddos-shift (case of spiddos 1). (b) A case of double spiddos-shift (spiddos 1 & 4) occurring at a time. (c) A case of quadruple spiddos-shift (spiddos 1, 4, 5 and 7 were shifted at a time). (d) An abnormal case of a single spiddos-shift (spiddos 3). (e) Distribution of $\overline{\gamma}$ for the cases of the single spiddos-shift. (f) The effect of spiddos-shift combination. The ordinate expresses s_{50} (the value of total shift which gives 50% average robustness). The theoretical s_{50} expected for each number of spiddos-shift are spotted and connected with a line.

APPENDIX

No.	Species / Conventional name	Family	Order	Class	Phylum
A1	*Typha orientalis* / Bulrush sp.	Typhaceae	Typhales	Mono*	Anth*
A2	*Arundinaria argenteostriata* / Bamboo sp.	Poaceae	Cyperales	Mono*	Anth*
A3	*Tricyrtis hirta* / Lily sp.	Liliaceae	Liliales	Mono*	Anth*
A4	*Cosmos bipinnatus* / Cosmos sp.	Asteraceae	Asterales	Dico*	Anth*
A5	*Taraxacum officinale* / Dandelion sp.	Asteraceae	Asterales	Dico*	Anth*
A6	*Callicarpa dichotoma* / Beauty-berry sp.	Verbenaceae	Lamiales	Dico*	Anth*
A7	*Gardenia jasminoides* / Gardenia sp.	Rubiaceae	Rubiales	Dico*	Anth*
A8	*Papaver nudicaule* / Poppy sp.	Papaveraceae	Papaverales	Dico*	Anth*
A9	*Viola xwittrockiana* / Pansy sp.	Violaceae	Violales	Dico*	Anth*
A10	*Camellia sasanqua* / Camellia sp.	Theaceae	Theales	Dico*	Anth*
A11	*Davidia involucrata* / Dove tree sp.	Davidiaceae	Cornales	Dico*	Anth*
A12	*Hydrangea macrophylla* / Hydrangea sp.	Hydrangeaceae	Rosales	Dico*	Anth*
B1	*Chilocorus rubidus* / Beetle sp. 1	Coccinellidae	Coleoptera	Inse*	Arth*
B2	*Oxycetonia jucunda* / Beetle sp. 2	Scarabaeidae	Coleoptera	Inse*	Arth*
B3	*Bombylius major* / Horse fly sp.	Bombyliidae	Diptera	Inse*	Arth*
B4	*Camponotus japonicus* / Ant sp. 1	Formicidae	Hymenoptera	Inse*	Arth*
B5	*Formica japonica* / Ant sp. 2	Formicidae	Hymenoptera	Inse*	Arth*
B6	*Apis mellifera* / Bee sp.	Apidae	Hymenoptera	Inse*	Arth*
B7	*Limenitis camilla* / Butterfly sp. 1	Nymphalidae	Lepidoptera	Inse*	Arth*
B8	*Anthocharis scolymus* / Butterfly sp. 2	Pieridae	Lepidoptera	Inse*	Arth*
B9	*Pieris rapae crucivora* / Butterfly sp. 3	Pieridae	Lepidoptera	Inse*	Arth*
B10	*Eurema laeta* / Butterfly sp. 4	Pieridae	Lepidoptera	Inse*	Arth*
B11	*Gonolabis marginalis* / Earwig sp.	Anisolabididae	Dermaptera	Inse*	Arth*
B12	*Bothrogonia ferruginea* / Stinkbug sp.	Cicadellidae	Hemiptera	Inse*	Arth*
B13	*Blattella germanica* / Cockroach sp.	Blattellidae	Blattaria	Inse*	Arth*
B14	*Reticulitermes speratus* / Termite sp.	Rhinotermitidae	Isoptera	Inse*	Arth*
C1	*Oncorhynchus masou* / Salmon sp. 1	Salmonidae	Salmoniformes	Acti*	Chor*
C2	*Oncorhynchus tshawytscha* / Salmon sp. 2	Salmonidae	Salmoniformes	Acti*	Chor*
C3	*Oncorhynchus mykiss* / Rainbow trout	Salmonidae	Salmoniformes	Acti*	Chor*
C4	*Salmo trutta* / Brown trout	Salmonidae	Salmoniformes	Acti*	Chor*
C5	*Salvelinus malma malma*/ Dolly Varden	Salmonidae	Salmoniformes	Acti*	Chor*
C6	*Salvelinus leucomaenis* / Whitespotted char	Salmonidae	Salmoniformes	Acti*	Chor*
C7	*Hucho perryi* / Japanese huchen	Salmonidae	Salmoniformes	Acti*	Chor*
C8	*Osmerus eperlanus mordax* / Rainbow smelt	Osmeridae	Salmoniformes	Acti*	Chor*
C9	*Cyprinus carpio* / Carp sp. 1	Cyprinidae	Cypriniformes	Acti*	Chor*
C10	*Phoxinus percnurus* / Carp sp. 2	Cyprinidae	Cypriniformes	Acti*	Chor*
C11	*Misgurnus anguillicaudatus* / loach sp. 1	Cobitidae	Cypriniformes	Acti*	Chor*
C12	*Barbatula barbatula* / loach sp. 2	Balitoridae	Cypriniformes	Acti*	Chor*
C13	*Silurus asotus* / Amur cat fish sp.	Siluridae	Siluriformes	Acti*	Chor*
C14	*Cottus nozawae* / Bullhead sp.	Cottidae	Scorpaeniformes	Acti*	Chor*

†This table is built based on NCBI's Taxonomy (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=taxonom) and Iwanami Biology Encyclopedia, 4th edition [30]. * Mono: Monocotyledonopsida, Dico: Dicotyledonopsida, Anth: Anthophyta, Inse :Insecta, Acti : Actinopterygii, Arth: Arthropoda, Chor : Chordata.

REFERENCES

[1] Nishigaki K., Naimuddin M., Hamano K., "Genome profiling: a realistic solution for genotype based identification of species," *J Biochem*, 128:107-112, 2000.

[2] Amann R. I., Ludwig W., and Schleifer K. H., "Phylogenetic identification and in situ detection of individual microbial cells without cultivation," *Microbiological Reviews*, vol. 59, no. 1, pp. 143–169, 1995.
[3] Sebat J. L., Colwell F. S., and Crawford R. L., "Metagenomic profiling: microarray analysis of an environmental genomic library," *Applied and Environmental Microbiology*, vol. 69, no. 8, pp. 4927–4934, 2003.
[4] Miner B. G., Sultan S. E., Morgan S. G., Padilla D. K., and Relyea R. A., "Ecological consequences of phenotypic plasticity," *Trends in Ecology and Evolution*, vol. 20, no. 12, pp. 685–692, 2005.
[5] American Museum of Natural History, "The Global Taxonomy Initiative: using systematic inventories to meet country and regional needs," in *DIVERSITAS/Systematics Agenda 2000 International Workshop*, New York, NY, USA, 1999.
[6] Cole J. R., Chai B., Farris R. J., et al., "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis," *Nucleic Acids Research*, vol. 33, pp. D294–D296, 2005.
[7] Maidak B. L., Cole J. R., Parker Jr. C. T., et al., "A new version of the RDP (Ribosomal Database Project)," *Nucleic Acids Research*, vol. 27, no. 1, pp. 171–173, 1999.
[8] Woese C.R., Kandler O., and Wheelis M. L., "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.
[9] Urwin R., and Maiden M. C. J., "Multi-locus sequence typing: a tool for global epidemiology," *Trends inMicrobiology*, vol. 11, no. 10, pp. 479–487, 2003.
[10] Sorokin A., Candelon B., Guilloux K., et al., "Multiplelocus sequence typing analysis of *Bacillus cereus* and *Bacillus thuringiensis* reveals separate clustering and a distinct population structure of psychrotrophic strains," *Applied and Environmental Microbiology*, vol. 72, no. 2, pp. 1569–1578, 2006.
11] Maiden M. C. J., Bygraves J. A., Feil E., et al., "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 6, pp. 3140–3145, 1998.
12] Kouduka M., Sato D., Komori M., Kikuchi M., Miyamoto K., Kosaku A., Naimuddin M., Matsuoka A., and Nishigaki K., "A Solution for Universal Classification of Species Based on Genomic DNA," *International Journal of Plant Genomics,* Vol. 2007, Article ID 27894, 8 pages, 2007.
13] Naimuddin M., Kurazono T., Zhangc Y., Watanabe T., Yamaguchi M., Nishigaki K., "Species-identification dots: a potent tool for developing genome microbiology," *Gene*, 261:243-250. 2000.
14] Futakami M., and Nishigaki K., "Measurement of DNA mutations caused by seconds-period UV-irradiation," *Chemistry Letters*, vol. 36, no. 3, p. 358-359, 2007.
15] Futakami M., Salimullah M., Miura T., Tokita S., and Nishigaki K., "Novel mutation assay with high sensitivity based on direct measurement of genomic DNA alteration: comparable results to the Ames test," *J. Biochem.* 141, 675-686, 2007.
16] Kouduka M., Matsuoka A., and Nishigaki K., "Acquisition of genome information from single-celled unculturable organisms (radiolaria) by exploiting genome profiling (GP)," *BMC Genomics*, vol. 7, p. 135, 2006.
17] Myers R. M., Fischer S. G., Lerman L. S., and Maniatis T., "Nearly all single base substitutions in DNA fragments joined to a GC-clamp can be detected by denaturing gradient gel electrophoresis," *Nucleic Acids Research*, vol. 13, no. 9, pp. 3131– 3145, 1985.
18] Salimullah M., Mori M., Nishigaki K., "High throughput three-dimensional gel electrophoresis for versatile utilities: a stacked slice-gel system for separation and reactions (4SR)," *Genomics Proteomics Bioinformatics*, 4(1): 26-33, Feb 2006.
[19] Wang H., Qin M., Cutler AJ., "A simple method of preparing plant samples for PCR," *Nucleic Acids Res* 21:4153-4154, 1993.
[20] Sakuma Y., and Nishigaki K., "Computer prediction of general PCR products based on dynamical solution structures of DNA," *Journal of Biochemistry*, vol. 116, no. 4, pp. 736–741, 1994.
[21] Hamano K., Takasawa T., Kurazono T., Okuyama Y., Nishigaki K., "Genome Profiling-Establishment and practical evaluation of its methodology," *Nikkashi* 1996:54-61, 1996.
[22] Biyani M., Nishigaki K., "Hundredfold productivity of genome analysis by introduction of microtemperature-gradient gel electrophoresis," *Electrophoresis*, 22:23-28, 2001.
[23] Nishigaki K., Husimi Y., Masuda M., Kaneko K., Tanaka T., "Strand dissociation and cooperative melting of double-stranded DNAs detected by denaturant gradient gel electrophoresis," *J Biochem,* 95:627-35, 1984.
[24] Watanabe T., Saito A., Takeuchi Y., Naimuddin M., Nishigaki K., "A database for the provisional identification of species using only genotypes: web-based genome profiling," *Genome Biol*, 3:00101.1, 2002.
[25] Nei M., Takezaki N., "Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA," *Genetics*, 144, 389-399, 1996.
[26] Jobson J. D., "*Applied Multivariate Data Analysis*, *Categorical and Multivariate Methods*," Springer, New York, NY, USA, vol. 2, 1992.
[27] Ward Jr. J. H., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.
[28] On-web GP [http://gp.fms.saitama-u.ac.jp/]