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 (Species identification dots) and PaSS (Pattern similarity score) were employed for measuring the closeness (or distance) between genomes. Based on the closeness (PaSS), we can build 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

Shamim Ahmed and Koichi Nishigaki

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@ems.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 on 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).](image-url)
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 from 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 (The whole procedure is shown in Fig. 2). 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 a temperature-gradient generator, and incubated at 65°C.

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:

\[ \text{PaSS} = 1 - \frac{1}{n} \sum_{i=1}^{n} \left| P_i - P'_i \right| \]  \hspace{1cm} (0 \leq \text{PaSS} \leq 1) \hspace{1cm} (1)

where \( P_i \) and \( P'_i \) 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:

\[ d' = 1 - \text{PaSS} \] \hspace{1cm} (2)

However, the distance, \( d' \), has been introduced here does not have a nature of the conventional distance that of Descartesian, like \( \sqrt{\left| x_i - y_i \right|^2 + \left| z_i - w_i \right|^2} \) (where \( x_i \) 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.
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:

\[
d_c = \alpha d_{xa} + \beta d_{xb} + \gamma |d_{sa} - d_{sb}|, \tag{3}
\]

where \(\alpha\), \(\beta\), \(\gamma\) are weighing parameters, \(d_x\), \(d_{xa}\), \(d_{xb}\) and \(d_{ab}\) represent distances between relevant clusters such as Cluster x and Cluster a for \(d_{xa}\).

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 usingRnd 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 °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 perturbed at random for both mobility and temperature coordinates within a range of 0.1 to 1.0 and 15°C to 65°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, $\gamma$, 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;

$$\overline{\gamma} = \frac{1}{s} \sum \gamma(s_j)$$

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

Fig. 3 Spiddos-shift. The shift (s) from $P_1 (T_{1,m1})$ to $P_2 (T_{2,m2})$ is shown, which may be caused by point mutation and/or insertion/deletion mutation in the corresponding DNA sequence.

The average robustness was found to be rather high (~0.03) and monotonous 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 weighted.

![Fig. 4 Scoring 1 or 0. If phylogenetic tree kept constant topologically after the spiddos-shift operation (a→b), then robustness score, $\gamma$, was put 1 while, if changed as in the case of a→c, $\gamma = 0$](image-url)
Average robustness (average s₅₀) 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 1). (e) A case of double spiddos-shift (spiddos 1 & 4). (f) The effect of spiddos-shift combination. The theoretical s₅₀ expected for each number of spiddos-shift are spotted and connected with a line.

APPENDIX

<table>
<thead>
<tr>
<th>No.</th>
<th>Species / Conventional name</th>
<th>Family</th>
<th>Order</th>
<th>Class</th>
<th>Phyllum</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Typha orientalis / Bulrush sp.</td>
<td>Typhaceae</td>
<td>Typhales</td>
<td>Mono*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>Arundinaria argenteosstrata / Bamboo sp.</td>
<td>Poaceae</td>
<td>Cyperales</td>
<td>Mono*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Tricyrtis hirta / Lily sp.</td>
<td>Liliaceae</td>
<td>Liliales</td>
<td>Mono*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>Cosmos bipinnatus / Cosmos sp.</td>
<td>Asteraceae</td>
<td>Asterales</td>
<td>Dico*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>Taraxacum officinale / Dandelion sp.</td>
<td>Asteraceae</td>
<td>Asterales</td>
<td>Dico*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>Callicarpa dichotoma / Beauty-berry sp.</td>
<td>Verbenaceae</td>
<td>Lamiales</td>
<td>Dico*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A7</td>
<td>Gardenia jasminoides / Gardenia sp.</td>
<td>Rubiaceae</td>
<td>Rubiales</td>
<td>Dico*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A8</td>
<td>Papaver nudicaule / Poppy sp.</td>
<td>Papaveraceae</td>
<td>Papaverales</td>
<td>Dico*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A9</td>
<td>Viola swinertickiana / Pansy sp.</td>
<td>Violaceae</td>
<td>Violales</td>
<td>Dico*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td>A10</td>
<td>Camellia sasanqua / Camellia sp.</td>
<td>Theaceae</td>
<td>Theales</td>
<td>Dico*</td>
<td>Anth*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


REFERENCES


