Research

A database for the provisional identification of species using only genotypes: web-based genome profiling

Takehiro Watanabe, Ayumu Saito, Yusuke Takeuchi, Mohammed Naimuddin and Koichi Nishigaki

Address: Department of Functional Materials Science, Saitama University, 255 Shimo-Okubo, Saitama, Saitama 338-8570, Japan.

Correspondence: Koichi Nishigaki. E-mail: koichi@fms.saitama-u.ac.jp

Published: 28 January 2002

Genome ***Biology*** 2002, **3(2)**:research0010.1–0010.8

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/2/research/0010

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 28 August 2001
Revised: 22 October 2001
Accepted: 5 December 2001

Abstract

Background: For a long time one could not imagine being able to identify species on the basis of genotype only as there were no technological means to do so. But conventional phenotype-based identification requires much effort and a high level of skill, making it almost impossible to analyze a huge number of organisms, as, for example, in microbe-related biological disciplines. Comparative analysis of 16S rRNA has been changing the situation, however. We report here an approach that will allow rapid and accurate phylogenetic comparison of any unknown strain to all known type strains, enabling tentative assignments of strains to species. The approach is based on two main technologies: genome profiling and Internet-based databases.

Results: A complete procedure for provisional identification of species using only their genomes is presented, using random polymerase chain reaction, temperature-gradient gel electrophoresis, image processing to generate 'species-identification dots' (spiddos) and data processing. A database website for this purpose was also constructed and operated successfully. The protocol was standardized to make the system reproducible and reliable. The overall methodology thus established has remarkable aspects in that it enables non-experts to obtain an initial species identification without a lot of effort and is self-developing; that is, species can be determined more definitively as the database is used more and accumulates more genome profiles.

Conclusions: We have devised a methodology that enables provisional identification of species on the basis of their genotypes only. It is most useful for microbe-related disciplines as they face the most serious difficulties in species identification.

Background

A biological species is usually defined in principle as a set of actually or potentially interbreeding organisms, but as interbreeding is very difficult to measure, species have in practice been identified by their phenotypic traits. Until recently, progress in most microbe-related disciplines has been hampered by the enormous effort needed to identify less prominent traits. We are now in an age when we can identify species based on the genome (genotype) [1], although this does not change the principle that taxonomy is defined by phenotypes [2]: according to the generally accepted rules of taxonomy, a strain belongs to a species if it falls within the range of phenotypes that define that species. This situation has been brought about by the success of the ribosomal RNA

approach to phylogenetics [3-5]. Well-conserved molecules, such as 16S rRNA in particular, have been used to give a species a molecular identifier and to draw phylogenetic relationships. The 16S rRNA-based approach has been widely accepted and has proved successful in phylogenetic tree-making and even in identifying species. In this context, the Ribosomal Database Project has been established [6]. There are other similar approaches, such as one based on the gyrase gene [7] and multilocus sequence typing [8]. Nonetheless, it has been impossible in practice to analyze all the constituents of a microbial population, not only because of the huge size of such populations (more than 10^8 cells per ml) but also because of lack of suitable methodology. Although there are methods other than gene and genome sequencing for analyzing genomes, such as restriction-fragment length polymorphism (RFLP), amplified fragment-length polymorphism (AFLP), Octamer-based genome scanning (OBGS), random polymerase chain reaction (PCR) and others [9-12], most cannot be used to identify species without a knowledge of phenotypic traits. In reality, there is no general methodology that enables us to identify species by genotype only, although many approaches use genotypic information (DNA sequences) to complement phenotypic information.

We have recently demonstrated the possibility of species identification by genotype using genome profiling [13], which is a temperature-gradient gel electrophoresis (TGGE) analysis of random PCR products [14]. In particular, the use of 'species-identification dots' (spiddos), which are feature points in genome profiles, is very useful for objective and reproducible data processing [15,16]. We present here a universal method for provisional genotype-based species identification based on these technological advances and using the Internet environment, which enables us to identify species in general. This paper also presents the important concepts of genome distance and genome sequence space, which are essential for species identification based on genotype.

Results and discussion

Figure 1 shows one of the results obtained using the protocols described in the Materials and methods. For the query species, the closest species as judged by spiddos and a list of genome profiles within the tolerance (τ), together with the annotation attached to it, is given (Figure 1). If there is a genome profile among the list annotated with species, then it means that the query species is identified with the confidence defined by the pattern similarity score (PaSS; see Materials and methods). If the value of PaSS is very high (that is, close to unity), then it is highly probable that it is indeed an exact match. In contrast, if the value is not sufficiently close to unity, then it may be only a related species (not the exact species), belonging to the same genus or family or any of the higher taxonomical categories, depending on the value of PaSS. Although we do not yet have enough data to determine the PaSS value at which it is safe to identify a species, we have a preliminary idea, based on experience, that 0.95 (Z score $\approx$ 4) may be a critical value [15]. The important challenge of how to reconcile the difference between identification of species by phenotype, which conventional taxonomy has adopted, with that based on genotype, is discussed later. An important aspect of this system is that one does not need to be a specialist in the relevant biological field to obtain an initial identification of an unknown organism. All that is required is to register the genome profile of the unknown species on the database. Therefore, an incomplete set of phenotypic data, which do not reach the criteria for species identification (say, peculiar behaviors or unusual properties), can also be registered and later used without having to undertake further laborious phenotypic identification (Figure 2). All the information regarding a given species (in other words, all the entries within a certain PaSS value) will be connected automatically, generating a volume of data on a particular species. Scientists can work cooperatively to identify species and collect their phenotypic traits (Figure 2). In conventional approaches to identification, most of which have been phenotype-based, those data that failed to meet the required criteria for identification were left unconnected, and could not be used later because there was no convenient way of correlating them with a given species without knowing the species name (Figure 2). Thus, our approach of genotype-based species identification, utilizing genome profile and the Internet, will be of great help to the field of taxonomy.

Key concepts of the on-web genome profiling

In evaluating the effectiveness of this methodology, the nature of PaSS must first be considered, as it plays the most important part in the method. As PaSS is calculated on the basis of the coordinates of spiddos (see Equation 1 in Materials and methods), the nature of spiddos must be thoroughly investigated. If two genomic DNAs contain common sequence regions that can be amplified by random PCR using the same primer, the resultant DNAs will usually generate similar spiddos (by definition, the spiddos obtained by TGGE represent the crucial points of a genome profile, points at which the temperature corresponds to the beginning of a prominent structural transition in DNA [15]). As shown schematically in Figure 3, two corresponding spiddos derived from two closely related species can be connected by a displacement vector, which consists of two independent elements of mobility (μ) and temperature (θ). The differences in each element ($\Delta\mu$ and $\Delta\theta$) can be related to the differences between two sequences as shown in Figure 3. The displacement in the ordinate is caused by the difference in length between the two DNAs and is caused by deletion or insertion, whereas that in the abscissa is mainly caused by point mutation (although insertion/deletion can also contribute).

As the extent of these changes is roughly proportional to the evolutionary time since the species diverged, we can expect that the summation of the displacement of each spiddo is

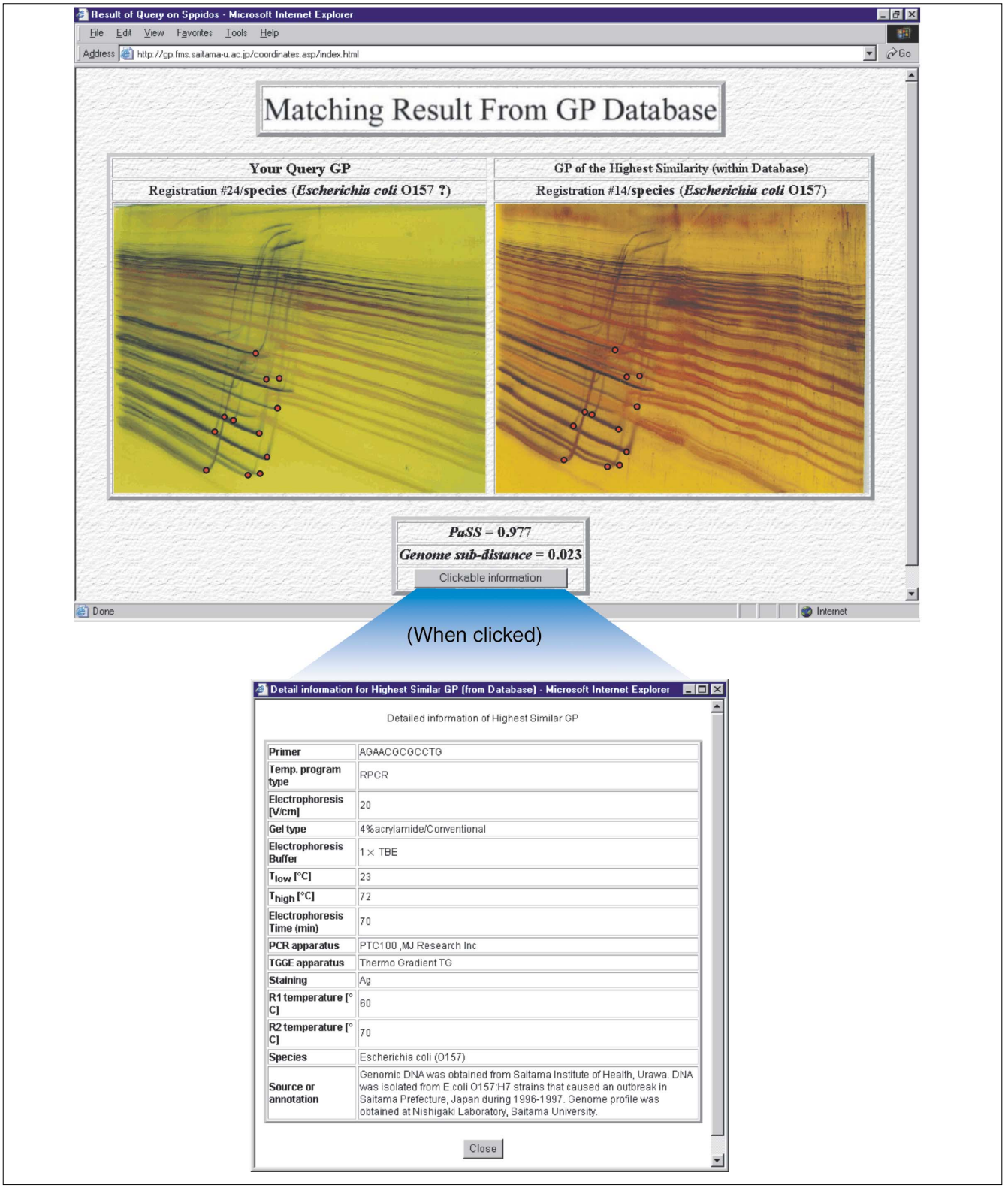

Figure 1
Result for a trial of on-web genome profiling. After uploading a genome-profile image and assigning spiddos and then subjecting it to a database search, a result will be displayed as shown, with the values of PaSS and genome distance to the closest species in the database. Note that a PaSS value close to unity infers that the query species is close to (or even the same as) the one retrieved from the database. The information on the selected species (right) already registered in the database can be viewed by clicking the button.

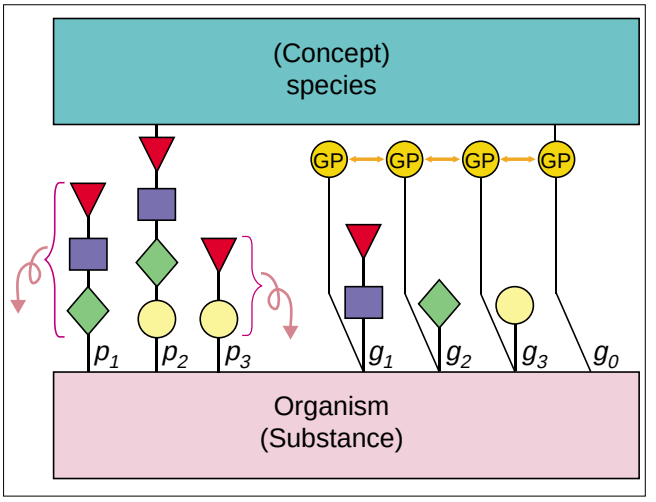

Figure 2
How to assign species in phenotype-based and genotype-based approaches. Phenotype-based approaches (indicated by p_1, p_2 and p_3) are heavily dependent on the traits (phenotypic or behavioral, appearing as different shapes) to identify species. In order to clarify such traits, sophisticated instruments and expert skills are often required. p_2 represents a successful identification attempt, where all the required traits for identifying the species have been obtained, whereas p_1 and p_3 are not successful because of insufficient information. Identity confirmed by genome profile in the genotype-based approaches makes it easy to compare and link unknown species (g_1-g_3) to known ones (g_0) without the requirement for extensive knowledge of phenotypic traits. Thus, the traits of each organism can be attributed to a particular species.

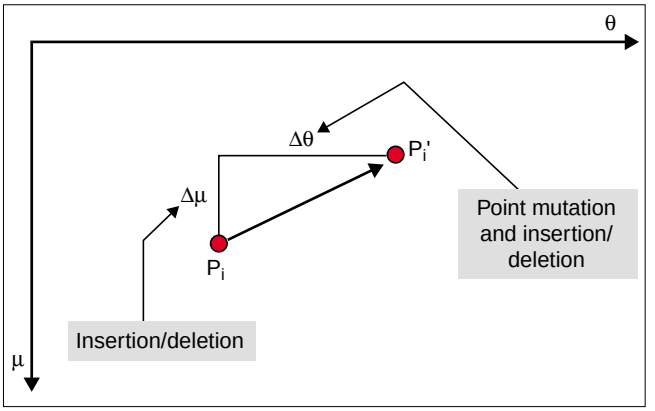

Figure 3
Causes of displacement in spiddos. The displacement between two spiddos (P_i and P_i') from two genome profiles can be decomposed into two elements, Δμ and Δθ. Δθ, which results from the shift in melting temperature, must have been caused mainly by point mutation and sometimes by deletion/insertion. On the other hand, Δμ, which is a measure of length, must be a result of insertion/deletion events occurring in the DNAs.

approximately proportional to the time since divergence, and thus to the genome-to-genome distance. By using a sufficient number of spiddos, we can obtain statistically reliable results. Empirically, we know that 8-10 spiddos, which can be obtained from a single genome profile, can be significant. However, since the more spiddos the better the result, we tentatively made it a rule to adopt four genome profiles (≈32-40 spiddos) - that is, four random-PCR products - as a current standard of initial species identification. Therefore, PaSS has the theoretical and empirical basis to be used as a measure of similarity between genomes, although the extent of its effectiveness remains to be shown experimentally as data accumulates. We have introduced a measure of distance, d', obtained from PaSS as formulated in Equation 2 (see Materials and methods), for the sake of convenience [15].

We call d' a genome sub-distance because it is based not on the whole but a part of the genome sequence. Thus, we introduce (true) genome distance, d, as in Equation 3 (see Materials and methods). Genome distance must have a close relationship with genetic distance, as defined by Nei and others [17-19], although there is a difference in the definition. The genetic distance based on sequences is basically the Hamming distance (the number of different letters at each corresponding position of two sequences of letters that are optimally aligned) between two nucleotide (or amino-acid) sequences. In aligning sequences arbitrariness is introduced, depending on the algorithm and parameters used [20]. Another constraint on genetic distance is that it is usually obtained from a limited number of genes, although that is also the case for genome distance. As genome distance is easier to obtain in practice using our method, it should be easier to obtain a lot of data on it compared with genetic distance. On the basis of d (in practice d'), we can construct phylogenetic trees and genome sequence space (an imaginary spherical space in which all the genomes (individuals) can be uniquely located in a finite manner based on the distance between genomes, providing clusters of species (K.N., unpublished observations)).

Although the applicability and effectiveness of genome distance for such purposes needs to be further investigated, it is obvious that an organism that has near-zero genome distance from a certain standard species, as an average over four or more genome sub-distances obtained from as many genome profiles, can be easily assigned to that same species with a high level of confidence. We are not claiming, however, to be able to give the correct taxonomical name to any species using this method. The greater the number of strains registered in the database, the more easily will a species be assigned. Basically, no special efforts, except expanding the database and using sophisticated algorithms, are necessary to raise the proportion of correct assignments. This is the self-developing nature of the database. Therefore, this methodology has two potential great advantages for tentative species identification: first, expertise is not always necessary; and second, database building can be carried out in a self-developing manner (that is, by acquiring more and more accurate data on species) with no waste of information.

Materials and methods

The principle that a species can be identified on the basis of its similarity to a standard species remains unchanged in the shift from phenotype-based to genotype-based methodology. Therefore, the essence of our methodology resides in finding a sufficiently closely related species by way of a measure of similarity - a pattern similarity score (PaSS). Note that this genotype-based methodology cannot define species under the current taxonomy regime, in which phenotype is used as the defining characteristic of species [2].

General protocol for on-web genome profiling

Although genome profiling is the basic technology for our current purpose, provisional species identification based on genotype can be fulfilled only by using computer-aided database technology, which is most effectively constructed in the Internet environment. As this methodology is based on use by a large number of scientists, the protocol must be designed to be reproducible and easy to carry out. The processes have been deliberately designed with this in view, and are presented on our website [21].

Genome profiling consists of two basic technologies: random PCR and TGGE, which have been well established [22-24]. However, if it is to be used for the purpose of general and universal applications, well-defined standardization is absolutely required to obtain significant results. We have carried out such standardization for genome profiling. The main topics included in the protocol are: preparation of genome DNAs; the set of primers used for random PCR and the internal reference DNAs used for TGGE; experimental conditions for random PCR; and the experimental conditions for TGGE. The protocol also includes the related procedures (extraction of spiddos, calculation of PaSS, and others).

Preparation of genomic DNA

Briefly, the alkaline extraction method was selected for simplicity as follows: 10 mg of cells or tissue are placed in an eppendorf tube and heated for 1 min at 100°C. The cells are mixed with 10 μg 0.5 M NaOH and stirred for 1 min (or 5 min or so for stiffer cells such as yeast) using a micro-homogenizer, if necessary, with added quartz sand. Immediately, a 5 μl aliquot of the lysate is mixed with 495 μl 100 mM Tris-HCl (pH 8.0). Usually, a 3 μl aliquot of the mixture thus obtained is used as a template for 100-μl-scale PCR. In some cases, such as *Escherichia coli*, which does not have a strong cell envelope, these cell-breakdown processes can even be omitted and the cells can be directly used in PCR. In other cases, such as fungi, thorough mechanical treatment (grinding with quartz sand) is needed. Thus, minimal and common procedures are preferred as much as possible for simplicity and generality in so far as they are consistent with the purity and integrity of the DNA samples. DNA samples thus prepared were shown to be identical with those DNAs prepared by the more elaborate conventional method of Thomas [25] as a PCR template [26]. This seems quite natural, as PCR can be carried out successfully in the presence of contaminating proteins or polysaccharides, irrespective of the DNA cleavages introduced, unless the regions of DNA to be amplified are completely cleaved. Nonspecific binding of proteins, which gives footprint effects, will change the yield but not the molecular ratio of random PCR products as long as the binding is totally stochastic. We also adopt a universal, convenient definition for genome DNA - that it is composed of all DNAs thus prepared, including dynamic elements such as satellite and organelle DNAs, and is irrespective of haploid or diploid status of the cells. Therefore, the DNA samples for genome profiling can be prepared in a common, technically well-defined method for all organisms.

Set of primers for random PCR

Technically important restrictions are introduced by selecting a standard set of primers for random PCR (T.W., A.S., M.N. and K.N., unpublished observations). It is important to carry out random PCR with all kinds of organisms using the same primers so that all species can be compared on the same platform. We have initially selected four oligonucleotides (pfM12: dAGAACGCGCCTG; pfM19: dCAGGGCGCGTAC; $d(TGC)_3$; $d(T_3G_3)_2$) as a standard set of random PCR primers. The primers pfM12 and pfM19 were selected on the basis of the abundant experimental background on them, whereas $d(TGC)_3$ and $d(T_3G_3)_2$ were rather theoretically favored (K.N. and A.S., unpublished observations). 'Oligonucleotide-stickiness analysis', which monitors oligonucleotide-binding sites along the template DNA (K.N. and A.S., unpublished observations), was exploited to determine the universal primers and moderately sticky oligonucleotides were selected. These four primers can be fluorescently labeled for convenience. More primers can be used to obtain more detailed information or to supplement insufficient information provided by the four primers about particular pairs of organisms. The information provided by such extra primers can explore in a more detailed manner the local landscape in genome sequence space. In contrast, the standard primers give us rough relationship between any pair of organisms.

Internal reference DNAs

Internal reference bands, which are provided by DNAs of a known melting pattern, are used to calibrate each genome profile, giving highly reproducible results [22].

Conditions for random PCR

Random PCR is usually carried out under standard conditions: 10 ng template DNA, 50 pmol primer DNA, 250 μM of each dNTP, 50 mM Tris-HCl (pH 8.8), 15 mM $(NH_4)_2SO_4$, 10 mM $MgCl_2$, 0.45% Triton X-100, 200 μg/ml bovine serum albumin and 2 units of *Taq* DNA polymerase (Biotech International). PCR was carried out in 30 cycles of 30 sec at 94°C, 2 min at 28°C and 2 min at 47°C, using a thermal cycler PTC-100TM (MJ Research, MA). Annealing temperature can be attenuated depending on the size of the template

DNA (in general, the larger the template, the greater the number of DNA fragments generated by random PCR).

Experimental conditions for TGGE

TGGE analysis of random PCR products is carried out with co-migrating internal reference DNAs. TGGE can be either the conventional type or a micronized type [16]. At least two feature points are extracted from the band pattern of the internal reference DNA(s), and then used for calibration of genome profiles or species identification dots (spiddos) [15] as described below. After calibration, sufficiently high reproducibility of the pattern of spiddos is guaranteed [16].

Extraction of spiddos

Although the genome profile is a kind of reduction of information contained in the whole genome sequence, it is still too complicated to deal with as it is. Thus, a second reduction is carried out by extracting feature points (spiddos) from the genome profiles. Double-stranded DNAs are known to melt in an intrinsically determined manner, depending on their sequence, when heated gradually [27]. All the intermediate states of DNA have their own structure and mobility in gel. Spiddos correspond to the structural transition points appearing in band patterns (Figure 4). Currently, there are four kinds of spiddos: initial melting point (P_{ini}); minimum mobility point (P_{min}); isomobility point (P_{iso}); and the end melting point (P_{end}). Empirically, P_{ini} is the most reproducible. Therefore, P_{ini} is recommended for working spiddos wherever possible. Further details are given in the standard protocol on our website [21].

Calculation of PaSS and genome distance

A set of spiddos (around ten), assigned to a genome profile on a computer display, is processed to calculate the normalized mobility and temperature of each point. A measure of similarity of two genomes - the PaSS - is introduced as follows.

$$\mathrm{PaSS} = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\vec{P}_i^{(1)} - \vec{P}_i^{(2)}\right|}{\left|\vec{P}_i^{(1)}\right| + \left|\vec{P}_i^{(2)}\right|} \qquad (1)$$

$\vec{P}$ of each spiddo (1 to n) is its position vector and is a function of temperature and mobility (that is, $|\vec{P}| = P(T, m)$). The superscripts 1 and 2 in parentheses in Equation (1) represent genomes 1 and 2, respectively. PaSS will be unity for a complete match in two sets of spiddos. In general, $0 \le \mathrm{PaSS} \le 1$. Genome distance and genome sub-distance (d') are derived from PaSS as follows:

$$d' = (1 - \mathrm{PaSS})/\mathrm{PaSS} \qquad (2)$$

$$d = \lim_{n\to\alpha}\sum_{i=1}^{n} d'_{(i)}/n \qquad (3)$$

Where $d'(i)$ is the ith genome sub-distance obtained with the ith primer used for random PCR.

Computer-aided data acquisition

The overall process of obtaining an on-web genome profile is shown in Figure 5. There are two steps in this methodology: the local phase and the database phase. In the local phase, genome profiling is carried out for the organism of interest, following the standard protocol presented on our website [21] and outlined in the previous sections. After obtaining a genome profile, the database is accessed and the database phase is begun as a client. The database site requires the client to input an image of the genome profile, to assign

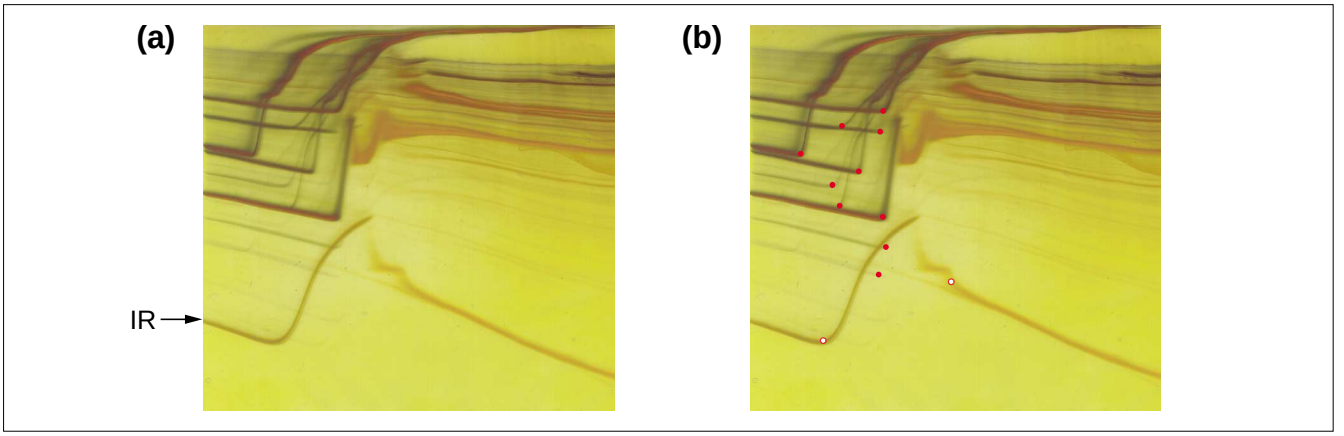

Figure 4

Spiddo assignment. **(a)** A genome profile before processing. The temperature gradient is set from left (low) to right (high) and the direction of migration is top to bottom; IR, internal reference band used for normalization. **(b)** The spiddos of the genome profile are marked with red filled circles; those of the IR are indicated with red open circles. All the spiddos except for the rightmost one are at the first transition of DNA melting (P_{ini}). Although there are four kinds of spiddos (dots), as described in Materials and methods, P_{ini} is used for simplicity as these points are clearly visible.

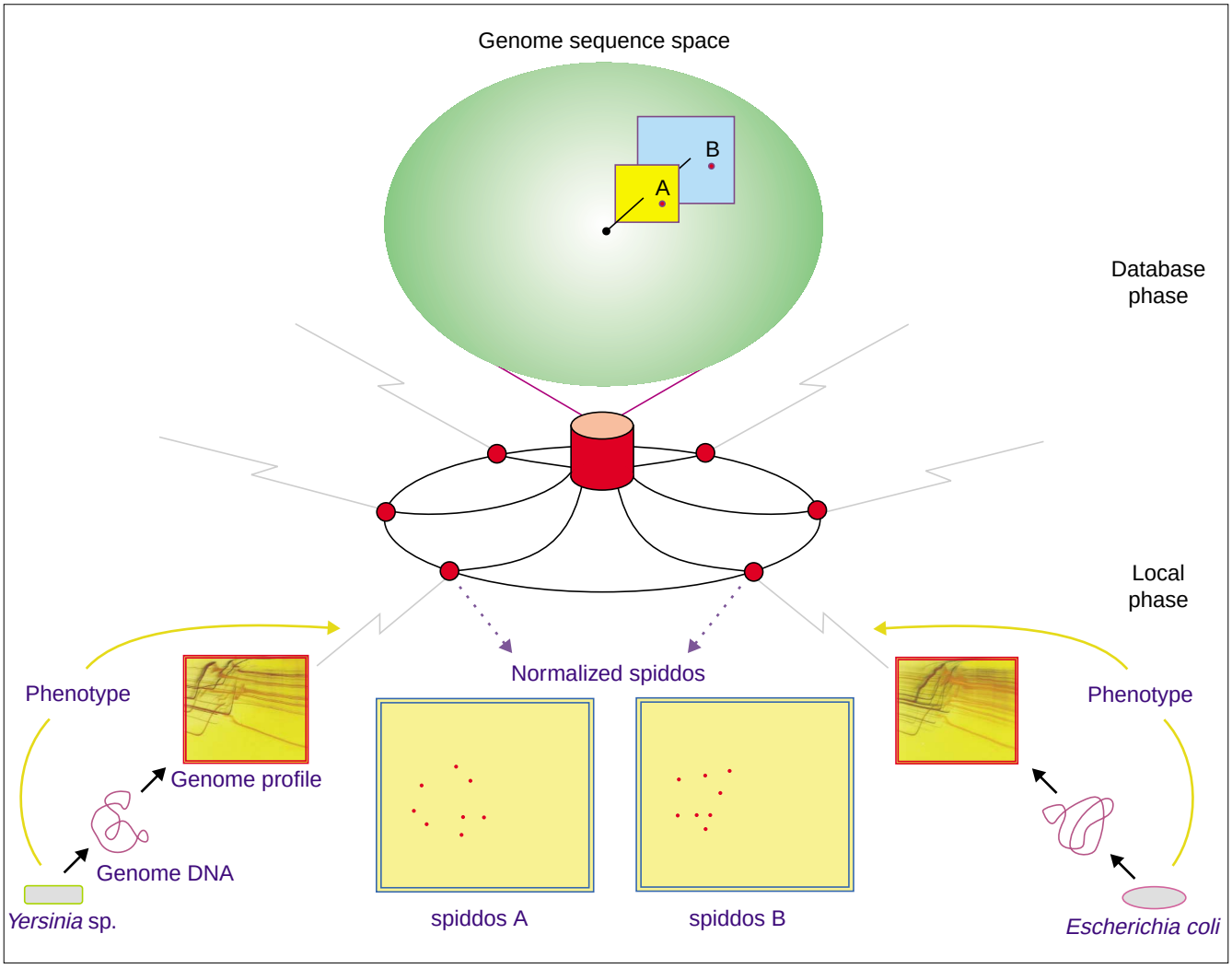

Figure 5
On-web genome profiling. The overall procedures to tentatively identify species by genotype only (genome profiling) are shown. Genome profiles are prepared by TGGE of random PCR products obtained from the genome DNA of a particular organism at the client site (the local phase). After accessing the database (represented by the red cylinder), a client (red circle) has spiddos assigned to each genome profile, which are used to calculate the measure of similarity, PaSS, and will finally get an output of the nearest species registered in the database (this phase of the process is called the database phase). Genome sequence space, with the location of the genomes A and B, is shown in green above the database of genomes.

spiddos on the genome profile (Figure 4), and to fill in relevant data on the online form. The site will search the database for species with the most similar pattern of spiddos by calculating the PaSS [15].

Acknowledgements

This study was supported in part by a Grant-in-Aid (09272203) from the Ministry of Education, Science, Sports and Culture of Japan. M.N. was supported by the Japan Society for Promotion of Science (13001147).

References

1. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eukarya.** *Proc Natl Acad Sci USA* 1990, **87:**3140-3145.
2. Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, *et al.*: **Report of the ad hoc committee on reconciliation of approaches to bacterial systematics.** *Int J Syst Bacteriol* 1987, **37:**463-464.
3. Olsen GJ, Overbeek R, Larsen N, Marsh TL, McCaughey MJ, Maciukenas MA, Kuan WM, Macke TJ, Xing Y, Woese CR: **The Ribosomal Database Project.** *Nucleic Acids Res* 1992, **20:**199-200.
4. Preparata RM, Meyer EB, Preparata FP, Simon EM, Vossbrinck CR, Nanney DL: **Ciliate evolution: the ribosomal phylogenies of the tetrahymenine ciliates.** *J Mol Evol* 1989, **28:**427-441.
5. Martinez JG, Bescos I, Sala JJR, Valera FR: **RISSC: a novel database for ribosomal 16S–23S RNA genes spacer regions.** *Nucleic Acids Res* 2001, **29:**178-180.
6. Maidak BL, Cole JR, Lilburn TG, Parker CT Jr, Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM: **The RDP-II (Ribosomal Database Project).** *Nucleic Acids Res* 2001, **29:**173-174.
7. Yamamoto S, Harayama S: **PCR amplification and direct sequencing of *gyrB* genes with universal primers and their application to the detection and taxonomic analysis of**

Pseudomonas putida **strains.** *Appl Environ Microbiol* 1995, **61:**1104-1109.

8. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, *et al.*: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci USA* 1998, **95:**3140-3145.
9. Tsipouras P: **Restriction fragment length polymorphisms.** *Methods Enzymol* 1987, **145:**205-213.
10. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M: **AFLP: a new technique for DNA fingerprinting.** *Nucleic Acids Res* 1995, **23:**4407-4414.
11. Kim J, Nietfeldt J, Benson AK: **Octamer-based genome scanning distinguishes a unique subpopulation of *Escherichia coli* O157:H7 strains in cattle.** *Proc Natl Acad Sci USA* 1999, **96:**13288-13293.
12. Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV: **DNA polymorphisms amplified by arbitrary primers are useful as genetic markers.** *Nucleic Acids Res* 1990, **18:**6531-6535.
13. Nishigaki K, Naimuddin M, Hamano K: **Genome profiling: a realistic solution for genotype-based identification of species.** *J Biochem* 2000, **128:**107-112.
14. Nishigaki K, Amano N, Takasawa T: **DNA profiling: an approach of systematic characterization, classification, and comparison of genomic DNAs.** *Chem Lett* 1991, **1991:** 1097-1100.
15. Naimuddin M, Kurazono T, Zhang Y, Watanabe T, Yamaguchi M, Nishigaki K: **Species-identification dots: a potent tool for developing genome microbiology.** *Gene* 2000, **261:**243-250.
16. Biyani M, Nishigaki K: **Hundredfold productivity of genome analysis by introduction of microtemperature-gradient gel electrophoresis.** *Electrophoresis* 2000, **22:**23-28.
17. Barnabas J, Goodman M, Moore GW: **Descent of mammalian alpha globin chain sequences investigated by the maximum parsimony method.** *J Mol Biol* 1972, **69:**249-278.
18. Nei M, Chakraborty R: **Genetic distance and electrophoretic identity of proteins between taxa.** *J Mol Evol* 1973, **2:**323-328.
19. Tateno Y, Nei M, Tajima F: **Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species.** *J Mol Evol* 1982, **18:**387-404.
20. Argos P: **Sensitive methods for determining the relatedness of proteins with limited sequence homology.** *Curr Opin Biotechnol* 1994, **5:**361-371.
21. **On-Web GP** [http://gp.fms.saitama-u.ac.jp]
22. Nishigaki K, Tsubota M, Miura T, Chonan Y, Husimi Y: **Structural analysis of nucleic acids by precise denaturing gradient gel electrophoresis: I. Methodology.** *J Biochem* 1992, **111:**144-150.
23. Henco K, Harders J, Wiese U, Riesner D: **Temperature gradient gel electrophoresis (TGGE) for the detection of polymorphic DNA and RNA.** *Methods Mol Biol* 1994, **31:**211-228.
24. Wartell RM, Hosseini S, Powell S, Zhu J: **Detecting single base substitutions, mismatches and bulges in DNA by temperature gradient gel electrophoresis and related methods.** *J Chromatogr A* 1998, **806:**169-185.
25. Berns KI, Thomas CA Jr: **Isolation of higher molecular weight DNA from *Hemophilus influenzae*.** *J Mol Biol* 1965, **11:**476-490.
26. Hamano K, Takasawa T, Kurazono T, Okuyama Y, Nishigaki K: **Genome profiling- establishment and practical evaluation of its methodology.** *Nikkashi* 1996, **1996:**54-61.
27. Wada A, Yabuki S, Husimi Y: **Fine structure in the thermal denaturation of DNA: high temperature-resolution spectrophotometric studies.** *CRC Crit Rev Biochem* 1980, **9:**87-144.