# Genome Profiling: A Realistic Solution for Genotype-Based Identification of Species[1]

Koichi Nishigaki,[2] Mohammed Naimuddin, and Keiichi Hamano[3]

Department of Functional Materials Science, Saitama University, 255 Shimo-Okubo Urawa, Saitama 338-8570

Received April 6, 2000; accepted April 28, 2000

**Species identification is the basis of Biology and has been carried out based on phenotype. Although some genes, such as that for 16S rRNA, have been used for species confirmation, identification of species based only on genotype has never been done before, although recent whole genome sequencing studies have demonstrated it to be possible in principle. However, it is evidently unrealistic for routine experiments of species identification. This paper clarifies that a very limited amount of information derived from a genome sequence is sufficient for identifying the species. It also proves that *Genome Profiling* [Nishigaki, K., Amano, N., and Takasawa, T. (1991) *Chem. Lett.* 1097–1100], TGGE analysis of random PCR products, can not only fulfill such requirements, but also serve as a universal method to analyze species. Thus, this compact technology can be used in many fields of biology, especially in microbe-related disciplines such as microbial ecology and epidemiology where exact knowledge about all members of a population is essential but previously difficult to obtain. This is the first demonstration that genotype-based identification of species is possible using a simple and uniform protocol for all organisms.**

**Key words: Genome Profiling, random PCR, species identification and information amount, TGGE.**

Species identification based on phenotype has been the only but yet sufficient method for most species. However, microbes, which are phenotypically far less prominent, cannot always be identified by such a method. Therefore, a method that enables the genotype-based identification of species has long been awaited. In addition, since a genotype-based approach can be made without a significant influence from environmental factors, it would be simpler and more reproducible. The well-known fact that genome DNA can be amplified for the sake of analysis by cloning or PCR is another advantage, which is particularly the case with species that are hard to culture. Besides whole genome sequencing, which has recently become realistic, there are other technologies that deal with genome DNAs such as RFLP (restriction fragment length polymorphism), RAPD (random amplified polymorphic DNA)-PCR, SSR (simple sequence repeat)-PCR, *etc.* (*1*, *2*), but none of these has been developed to identify species in general, mainly because of the insufficiency in the amount of information which they can provide. Whole genome sequencing or the sequencing of a substantial portion of the genome is surely determinative of species. However, it is too costly and laborious. From the viewpoint of the amount of information, it is too redundant only for identifying species. This can be easily understood if one considers that a trillion people can be uniquely identified only by a single number of 10 orders of magnitude called PIN (personal identification number), as introduced in some nations. Therefore, all the traits of a person are not necessary for this purpose. This is also true in identifying species. In order to clarify this fact, this paper first deals with the notion of the amount of information sufficient for species identification and then demonstrates that *Genome Profiling* (*3*, *4*) is fully suitable for this purpose. This paper also emphasizes the fact that such technology enables us to conveniently identify a species genotypically, and will play a key role in the field of genome-related sciences such as genome microbiology and genome ecology.

## MATERIALS AND METHODS

*Preparation of Genome DNAs*—The cells used were clinical strains of *Escherichia coli* (O157:H7) collected at Saitama Hygiene Institute (Saitama), baker's yeast commercially obtained from Nisshin Flour Milling, and rat (pc12) and mouse (p19) cells, gifts from Dr.Watanabe at The Tokyo Metropolitan Institute of Gerontology. All genome DNAs were conveniently prepared by the same alkaline method (*4*, *5*). Briefly, 10 mg of cells was placed in an Eppendorf tube and heated for 1 min at 100°C. The cells were mixed with 10 μl of 0.5 M NaOH and then stirred for 1 min (5 min for yeast cells) using a motor-driven microhomogenizer #226 (Cosmo Bio, Tokyo). Immediately, a 5 μl aliquot of the lysate solution was mixed with 495 μl of 100 mM Tris-HCl (pH 8.0). Usually, a 3 μl aliquot of the mixture thus obtained was used as a template for 100 μl scale PCR. There-

[1] This research was supported in part by Grant-in-Aid from the Ministry of Education, Science, Sports and Culture of Japan.

[2] To whom correspondence should be addressed. Tel/Fax· +81-48-858-3533, E-mail: koichi@fms.saitama-u.ac.jp

[3] Present address. Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba, 292-0812.

Abbreviation: GP, *Genome Profiling*, *spiddo* (species identification dot).

fore, we define the genome DNA to be the whole DNAs thus extracted, including satellite DNAs and organelle (mitochondrion and chloroplast) DNAs. Considering the dynamic nature of genetic materials, this definition, we think, is not only convenient but also universal.

The empirically well-established fact that this technology does not require high quality DNA samples (allowing less intact and more contaminated samples) duc to the nature of PCR is another advantage of this method.

*Genome Profiling*—This comprised two methods· random PCR and TGGE (or DGGE). Random PCR was performed using a single primer of 12 nucleotides at a low annealing temperature such as 28°C. The PCR reaction buffer (100 μl) contained 200 μM dNTP (N = G,A,T,C), 0.5 μM primer, 10 mM Tris-HCl (pH 9.0), 50 mM KCl, 2.5 mM $MgCl_2$, 0.1% Triton X-100, 0.02 unit/μl Taq DNA polymerase (NEB, Massachusetts), and template DNA at around $10^{-17}$M. Random PCR was carried out in cycles of denaturation (94°C, 30 s), annealing (28°C, 2 min), and extension (47°C, 2 min) using PCR machines such as PTC-100TM (MJ Research, Massachusetts). Ninety microliters of random PCR products was mixed with 100 μl of a buffer solution containing 40 mM Tris-HCl (pH 8.0), 20 mM sodium acetate, 1 mM EDTA, 0.3% (w/v) xylene cyanol and 20% (w/v) sucrose. The mixture was then layered on the top of a slab gel [4% (w/v) acrylamide gel containing 8 M urea] and briefly subjected to electrophoresis in order to let the DNAs migrate into the gel. The gel was then placed on the stage of a TGGE apparatus, TG-180 (Taitec, Saitama) with one side (top) covered with a glass plate and the other side (bottom) touching the surface of the stage *via* a thin Gelbond film (FMC). The temperature gradient (30–70°C) was set perpendicular to the direction of DNA migration. Electrophoresis was usually performed for 75 min at 400 V.

*Computer Analysis*—Binding frequencies of oligonucleotides to genome DNAs were analyzed with a computer program, PCRAna-A1 [a derivative of PCRAna (*6*)]. The program module of PCRAna-A1 calculates the stabilities, in terms of Gibbs free energies, of binding structures, which usually contain bulges and mismatches, in addition to Watson-Crick base pairings (see Fig. 1a). The parameters used were essentially the same as used previously (*6*) (*i.e.*, allowed number of coil parts, 2 nts; Allowed length of coil part, 3 nts; minimum number of base pairs in helix part, 8; temperature, 30°C; minimum number of continuous base pairs at the 3′ end of a primer; and others). The cut-off energy, which was used to reduce the amount of computation and has an influence on the number of stable structures found, was empirically set at –6 kcal/mol, which is known to fit *in-vitro* experiments (*6*).

## RESULTS AND DISCUSSION

Although a complete match in genome sequence means an unequivocal identity in species, even a fractional match between two genome sequences is often sufficient for identification. Obviously, a decision on identity can be made based on comparison with the genome sequence of the "reference" species. Therefore, the less that is required for comparison, the more realistic it becomes. Then, what is a sufficient amount of information to identify species based on genome sequence?

The amount of information required for identification, $I_s$, can be estimated as follows. If the genome of a species can be uniquely specified by a sequence stretch of $s$ nucleotides that occurs only once throughout all the genomes of all species, then $I_s$ can be related to the minimum value of $s$. If the number of all species and their average genome size are $m$ and $\langle n \rangle$, respectively, then, the following formula holds true based on the theory of expectation value:

$$m \cdot \langle n \rangle \cdot p^s \leqq 1 \qquad (1)$$

where $p$ is the probability of finding a particular nucleotide (G, A, T, or C) at a particular site. It is safe to estimate that $m \leq 10^{30}$. This is because the number was derived based on a supposition that a radial column space standing on a 1 $cm^2$ surface area of the Earth (assumed to be $5 \times 10^{18}$ $cm^2$), contains $2 \times 10^{11}$ kinds of species. This means that even bacterial flora with a population of $2 \times 10^9$ cells/ml (equivalent to the saturation population of *E. coli*) must be piled up to a height of 1 m, and what, is more, letting them be *mutually independent species*. Obviously, this is an overestimate. No sea water (even if the depth is considered) is so heavily populated with different species. In reality, the number is estimated around $10^8$ species by ecologists (*7*), which is deceptively smaller than that estimated above. On the other hand, the average genome size, $\langle n \rangle$, can be given as $10^{12}$ bp since one of the largest genomes known is $1.3 \times 10^{11}$ bp for a plant (*Psilotum nudum*) (*8*).

The value of $p$ cannot be always 1/4 (the value for a random sequence) since genome sequences are by no means random. Here, we can conservatively place the value at 1/2, although this might be argued. Thus, the minimum value of $s$ that satisfies Eq. 1 is obtained by substituting $m = 10^{30}$, $\langle n \rangle = 10^{12}$, and $p = 1/2$: $s = 140$. This means that a sequence of only 140 nucleotides or 140 bits of information [because each letter is considered to contain 1 bit ($= -\log_2 p = -\log_2 1/2$) in average] is sufficient to specify a genome, discriminating the organism from all the other species. The risk probability, $P_R$, that two species will have the same sequence of s nucleotides is:

$$\begin{aligned} P_R &= 1-\{{}_{\upsilon}C_0\, q^0(1-q)^{\upsilon} + {}_{\upsilon}C_1\, q^1(1-q)^{\upsilon-1}\} \\ &= \{(\upsilon-1)q\}^2 = (\upsilon\, q)^2 \qquad (2) \end{aligned}$$

where $q = p^s$ and $\upsilon = m \cdot \langle n \rangle$. This implies that if $s = 200$ (thus $q = p^s = (1/2)^{200} = 10^{-60}$ while $\upsilon = 10^{42}$), probability $P_R$ is as small as $10^{-36}$. Considering this risk, the amount of information sufficient for identifying species, $I_s$, is provided by a sequence of only 200 letters (200 bits). Regardless of the exact value of $I_s$, the fact that each species can be identified by fewer than 200 bits of information is very important for genome science. This means *each genome can be identified by a very short sequence contained within its genome*. Besides, any sequence within a genome, whether it is continuous or fragmented, is usable for this purpose so far as it can be uniquely specified due to the additive nature of information.

Therefore, it is of technological importance is to establish a method by which a sequence can be uniquely identified under a set of common criteria. Here, we demonstrate that *Genome Profiling* (GP) is fully competent for meeting these requirements. GP is composed of two elementary techniques as described in Fig. 1 and "MATERIALS AND METHODS," random PCR and T/DGGE [temperature and denaturing gradient gel electrophoreses are interconvertible (*9*, *10*)]. Random PCR (*3*, *6*), which adopts one or more arbitrary primer(s), is performed at a lower temperature than conventional PCR, and generates a set of DNA fragments

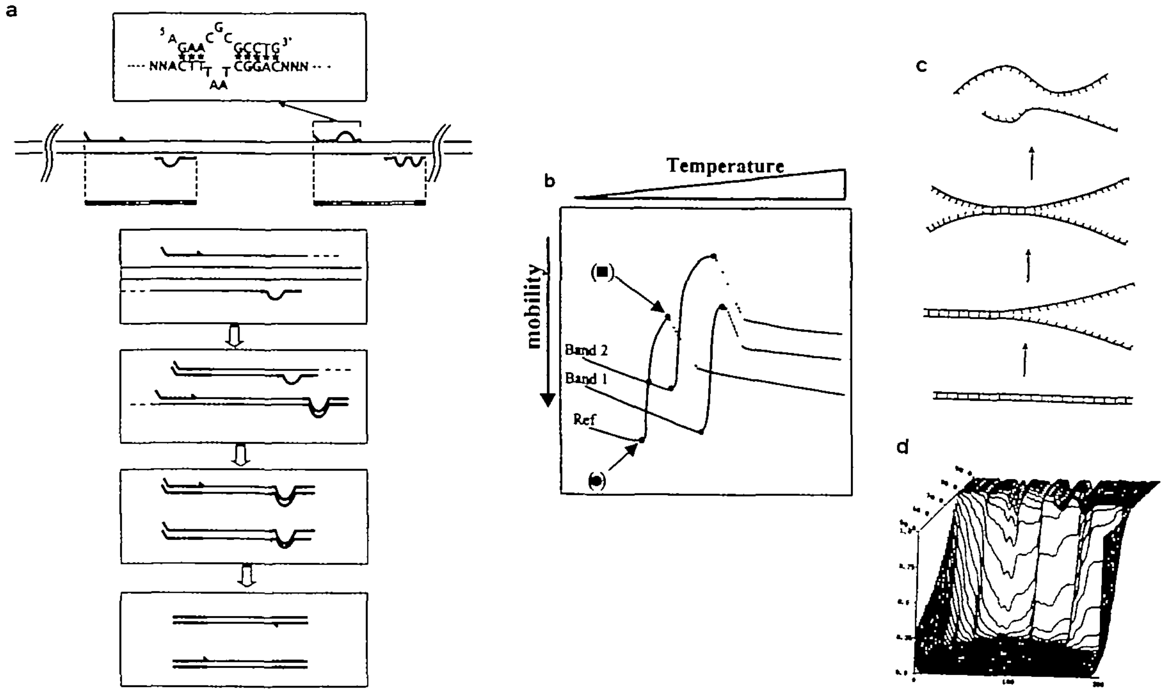

Fig 1 **Schematic representation of *Genome Profiling*.** a· Random PCR A kind of PCR that employs arbitrary primers and is operated at a lower annealing temperature than usual begins with the binding of primers containing mismatches and/or bulges. b TGGE analysis Sample DNAs, charged in a line on the top, migrated towards the bottom under a temperature gradient rising from left to right, and were separated depending on their size and structure. A result thus obtained, which corresponds to the genome profile of *E coli* (Fig. $2a_1$), is shown Featuring points [called *species identification dots* (*spiddos*)] are marked (● and ■) The filled circles (●) represents the position at which temperature ($T_i$) DNA melting initiates, while the filled squares (■), the position for strand dissociation (*9*) A more determinative and better-defined approach is in progress using these featuring points (soon appear elsewhere) c. The secondary structures appearing in the melting are shown compared with the reference DNA (204 bp) marked with an arrow head, which is an interpretation of the melt map theoretically obtained (shown in Fig 1d). d Melt map of a DNA. The lateral, longitudinal and depth-directional coordinates are for nucleotide positions in sequence, (1–$\theta$) (where $\theta$ is helix probability), and temperature, respectively. This figure was obtained using the computer program *Poland* provided by Dr. Steger (*12*) and modified in our laboratory.

that are, in principle, predictable from knowledge about the sequences of the template (genome) and primer(s) (*6*). This prediction was made based on the stabilities of primer binding structures and was experimentally shown to be reliable (*6*) [recently shown to be applicable to *E. coli* (*11*)]. When the DNA fragments thus obtained are subjected to TGGE, sequence-specific melting profiles appear on the gel plane after staining (Fig. 1c). These profiles contain the features indicated in Fig.1-c. These points are known to represent any transition in DNA melting, which is also theoretically predictable (*9*, *12*, *13*), as shown partially in Fig. 1d. Figure 2 shows typical GP results obtained with various species, bacteria to animal, in experiments performed under identical conditions. Evidently, species-specific genome profilings are observable for all categories of species including (i) mammals and fungi (newly reported here) and (ii) bacteria, plants and invertebrates [previously reported for more than 20 species (*3*, *4*)]. Close inspection of the DNA bands makes it clear that each DNA band has a reproducible intrinsic pattern (*10*). The degree of closeness between species is also evident by visual inspection of rat and mouse, within the same family (c and d in Fig. 2), and for two strains of *E. coli* and budding yeast ($a_1/a_2$ and $b_1/b_2$), within the same species. Since the method used here for preparing DNA samples is common, a universal stage for comparing species based on genotype is given here.

The next problem concerns the amount of information conveyed by *Genome Profiling*, $I_{GP}$ Random PCR begins with the primer binding to a template DNA, which is usually not completely complementary but yet defines the positions of both ends of the products along the template DNA. This process, selecting binding sites from the whole template DNA, can be interpreted to bear some amount of information. In order to evaluate this, a computational approach was made (Fig. 3). Figure 3 shows how often a primer binds to a template (of unit length) as a function of their sequences. As expected, there is a positive correlation between the stability of completely complementary structures (as a representative of primer-binding structures) and the frequency of binding to a template. It also shows that there are binding sites, roughly, in every 20–50,000 nucleotides for all cases tested (various types of primers against bacteria to yeast) under physiological conditions. Therefore, [considering that a particular sequence of $s$ nucleotides is expected to appear in each $4^s$ in a random sequence ($p$ = 1/ 4)], the binding frequencies can be estimated to be equivalent to those of 3–9 nucleotides in perfect match (*i.e.*, 6–18 bits per binding in this case). Thus, the generation of each random PCR product specified by two binding sites conveys more than 12 (= 2 × 6) bits.

Secondly, the position of each featuring point is known to be related to the sequence of the corresponding DNA (*9*).

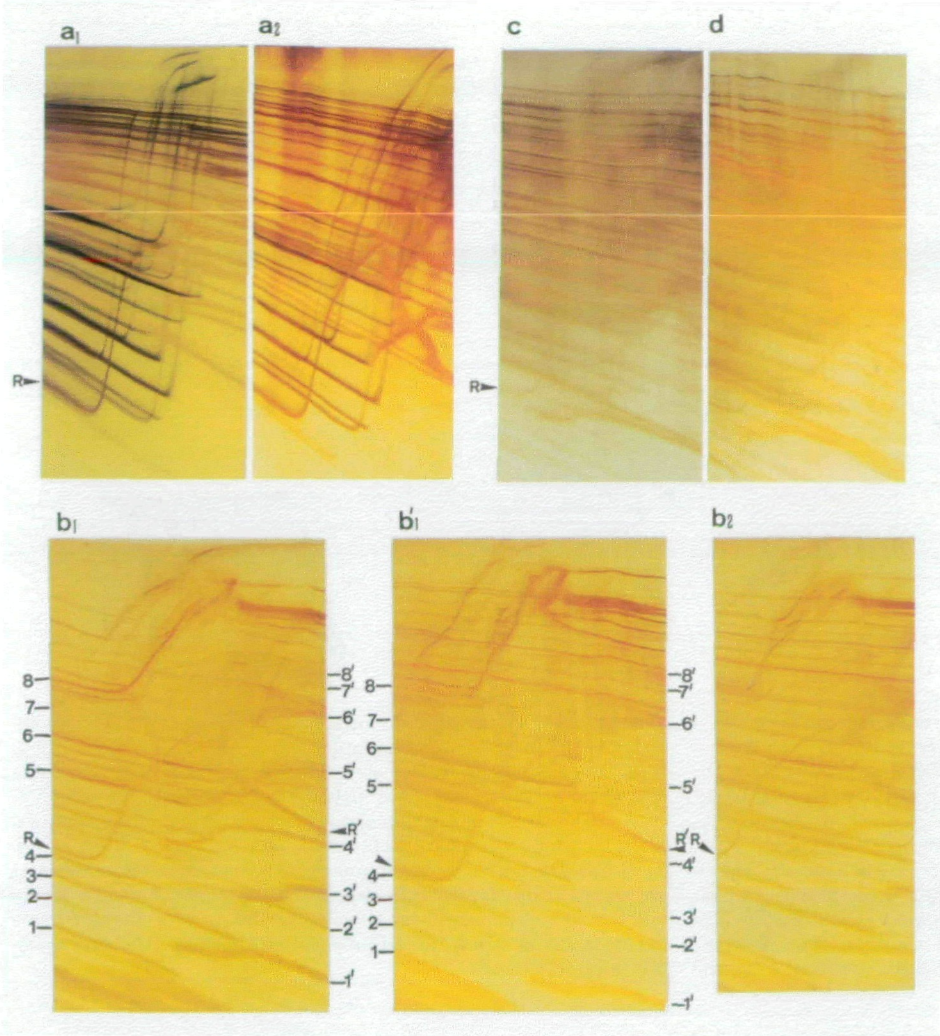

Fig 2. **Genome Profiling of various species.** GPs of *E coli* ($a_1$, $a_2$), yeast ($b_1$, $b_1'$, $b_2$), rat (pc12) (c) and mouse (p19) (d) are shown, where the same primer (pfM12; dAGAACGCGCCTG) was used for all. Two strains of *E coli* ($a_1$, $a_2$) and yeast ($b_1$, $b_2$) were used as templates for GP To show the range of reproducibility, two GPs, obtained with the same strain in separate experiments beginning from the DNA preparation, are shown ($b_1$, $b_1'$) The internal reference (204 bp, $T_i$ = 60°C) is shown by an arrow with R (double-strand form) and *R'* (single-strand form). Some of the DNA bands in $b_1$ and $b_1'$ are assigned as 1–8, with or without a prime meaning single- or double-strand form, respectively

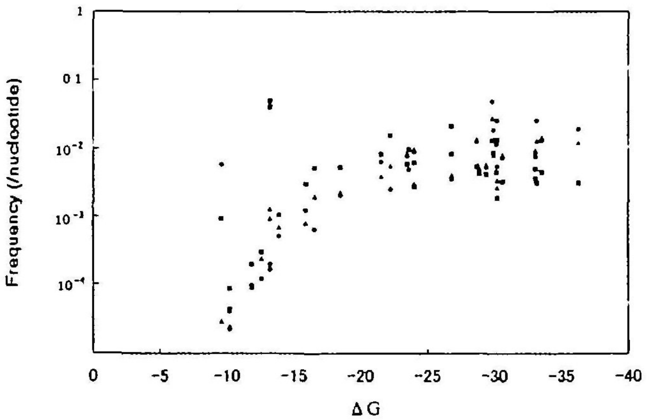

Fig. 3. **How often do primers bind to genome DNAs?** The frequency of primer binding in the random PCR model was examined using genome DNAs from *S. cerevisiae* (□) and *E. coli* (○), together with *B. subtilis* (△), for the sake of generality, using 33 primers of 12 nucleotides with different sequences. The primers used have binding stability of *ΔG* ranging from −5 to −40 kcal/mol (calculated as Watson-Crick base-paired structure) The frequencies were normalized by dividing by the number of the total nucleotides in each genome

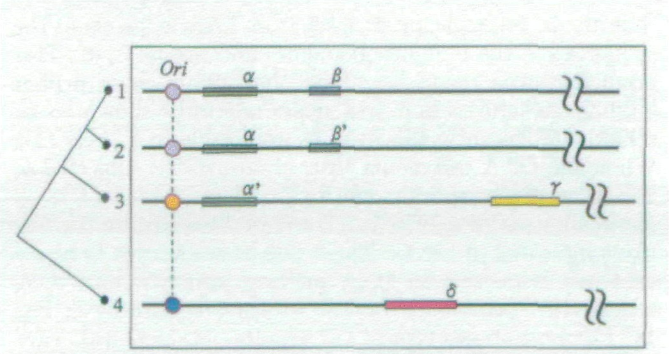

Fig. 4 **Interpretation of random PCR-amplified regions.** The boxed regions were amplified by random PCR. Relationship among species is shown by the phylogenetic tree to the left. The regions, α and α' (also, β and β'), are commonly conserved genetic fragments (ccgf). The closer the species, the greater the number of ccgf between them. *Ori* is depicted for convenience to show the relative position in the genome. This interpretation could be taken for Fig. 2 by identifying genomes 1–4 as those of mouse, rat, yeast and *E. coli*, respectively.

Therefore, featuring points contain some amount of sequence information. Although it is difficult to determine the total amount, it can be approximated as follows: Assume that each featuring point spificies a spicific one out of 100 (*e.g.*, 10 × 10) *blocks* spread on the 2-dimensional gel plane (the size of each block need not be same so far as each block can be resolved), and that the probability for a featuring point to be located on one of these blocks is equal; then, ~7 bits (= $-\log_2 P = -\log_2 1/100$) can be assigned to each featuring point. These featuring points are analyzed to correspond to one of the following: initiation, propagation, or termination of DNA melting. Usually, there are more than two featuring points, melting initiation and strand dissociation (*9*, *13*), observed for each band (shown as *spiddos* in Fig. 1). Actually, each band pattern is highly characteristic and can present far more information as shown in Fig 2. Hence, the information collected amounts to more than 14 bits. Considering that these are usually more than 5 bands observable per gel (*4*), the total amount of information provided by a single GP, $I_{GP}$, is·

$$I_{GP} \geqq \{(6 \times 2) + (7 \times 2)\} \times 5 = 130 \qquad (3)$$

$I_{GP}$ is already more than half of the upper limit of $I_s$ (= 200 bits) discussed above. Since this estination is very conservative, the results show that a single round of GP is actually sufficient for genome identification. Even if not sufficient, performing another round of GP using a different primer will suffice. Consequently, GP is a *purpose-sufficient* method that can easily conform to the requirements for species identification.

*Significance of Genome Profiling*—As shown in Fig. 4, random PCR can be taken as making copies of some regions of a genomic DNA. The products can be identical (α in Fig. 4) or very similar (α/α' and β/β', pairwisely) for closely related species as seen in Fig. 2. As a result, the overall GB pattern looks similar to such species (Fig. 2, c and d). [Although the overall similarity of the patterns of two kinds of GPs can easily be evaluated semi-quantitatively by visual inspection, especially when the two organisms are sufficiently close (see Fig. 2), such an approach is less objective and less convenient for dealing with data. Therefore, we are advancing a technology in which only the featuring points on GP are utilized for point-by-point comparison, making things much simpler and clearer, although there is reduction in the amount of usable information (still sufficient to identify species). (Details to be reported elsewhere)]. However, as two species become phylogenetically distant, their random PCR products become less or unrelated as in γ and δ in Fig 4. Nevertheless, the products are uniquely defined so far as the combination of species and primer(s) is defined, which is the reason these products can be used for species identification. Owing to the rather robust nature of random PCR against point mutations, deletions/insertions inside and genome-scale shufflings (that is, random PCR amplifies the same region of DNA, regardless of these mutations), we can expect to trace from one species to a distant species following after such *commonly conserved genetic fragments* (ccgf). Here, it should be emphasized that if a gene is more conserved in evolution, it is less informative for comparative analysis. Therefore, sequencing such a highly conserved gene will not convey sufficient information as a single gene. On the other hand, less conserved regions, although informative, are difficult to amplify for all species using a simple protocol due to their high rate of mutation. Therefore, a very limited number of genes, such as those for 16S rRNA and gyrases, are currently used. Thus, for comparative analysis of a wide range of species, random PCR performed with a common primer must be far more powerful than any specific PCR approaches.

GP conveys not only size information but also sequence-derived information. Both types of information are absolutely essential not only in terms of the amount of information, but also for accurate band-assignment in gel electrophoresis. Based only on mobility information, as in in the case of RFLP (*2*, *15*) and RAPD-PCR (*16*, *17*), there is often ambiguity (*15*) in assignment. The amount of information should be sufficiently large to confirm band identity. This is the point in which GP is advantageous over other related technologies.

GP itself has an added property in that the origin of the genetic fragments does not matter (any unknown genes and genomes are acceptable) and uses a tiny portion of the whole genome (intact genomic DNA is less required). It also can be performed rapidly and is compatible with a high-throughput system due to its simplicity. The most promising application for GP needs to be pursued in such disciplines as microbial ecology, microbial epidemiology, and microbial environmental chemistry. In these disciplines, vast numbers of species need to be identified in order to study the interactions and distributions of all microbes. Such identification is impossible or too difficult to be performed by phenotype-based approaches. As a result of the application of GP, we can expect to identify things such as all constituent microbes in soil, intestine, and mouth; distribution pathways of endotoxin-generating bacteria such as O157; and the effects of chemical substances on the population of microflora in fresh water. GP is also applicable to the discrimination of, for example, edible mushrooms from poisonous ones.

The authors are grateful to Takashi Hasegawa and Yuusuke Takeuchi for technical assistance, and the anonimous English-corrector to J B for his/her excellent work

## REFERENCES

1. Caetano-Anolles, G (1996) Scanning of nucleic acids by *in vitro* amplification· new developments and applications. *Nat Biotechnol.* **14**, 1668–1674

2 Tsipouras, P (1987) Restriction fragment length polymorphisms *Methods Enzymol* **145**, 205–213

3. Nishigaki, K., Amano, N., and Takasawa, T (1991) DNA profiling—An approach of systemic characterization, classification and comparison of genomic DNAs. *Chem Lett* **1991**, 1097–1100

4. Hamano, K., Takasawa, T, Kurazono, T., Okuyama, Y, and Nishigaki, K. (1996) Genome profiling—Establishment and practical evaluation of its methodology *Nikkashi* **1996**, 54–61

5 Wang, H , Qin, M , and Cutler, A.J (1993) A simple method of preparing plant samples for PCR *Nucleic Acids Res.* **21**, 4153–4154

6 Sakuma, Y and Nishigaki, K. (1994) Computer prediction of general PCR products based on dynamical solution structures of DNA. *J. Biochem.* **116**, 736–741

7. Rosenzweig, M.L (1995) *Species Diversity in Space and Time*, Cambridge University Press, Cambridge

8. Cavalier-Smith, T (1985) *The Evolution of Genome Size*, Wiley, New York

9. Nishigaki, K., Husimi, Y, Masuda, M , Kaneko, K., and Tanaka, T (1984) Strand dissociation and cooperative melting of double-stranded DNAs detected by denaturant gradient gel electrophoresis. *J. Biochem.* **95**, 627–635

10 Nishigaki, K., Tsubota, M., Miura, T., Chonan, Y, and Husimi, Y. (1992) Structural analysis of nucleic acids by precise dena-

turing gradient gel electrophoresis: II. Applications to the analysis of subtle and drastic mobility changes of oligo- and polynucleotides. *J. Biochem.* **111**, 151–156

11. Nishigaki, K., Saito, A., Hasegawa, T., and Naimuddin, M , (2000) Whole genome sequence-enabled prediction of sequences performed for random PCR products of *Escherichia coli*. *Nucleic Acids Res.* **28**, 1879–1884
12. Fischer, S G. and Lerman, L.S (1983) DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. *Proc. Natl. Acad. Sci. USA* **80**, 1579–1583
13. Steger, G. (1994) Thermal denaturation of double-stranded nucleic acids: prediction of temperatures critical for gradient gel electrophoresis and polymerase chain reaction. *Nucleic Acids Res.* **22**, 2760–2768
14. Abrams, E.S. and Stanton, V.P., Jr. (1992) Use of denaturing gradient gel electrophoresis to study conformational transitions in nucleic acids. *Methods Enzymol* **212**, 71–104
15. Papadopoulos, D., Schneider, D., Meier-Eiss, J., Arber, W., Lenski, R.E., and Blot, M (1999) Genomic evolution during a 10,000-generation experiment with bacteria. *Proc. Natl Acad. Sci. USA* **96**, 3807–3812
16 Williams, J.G.K., Kubelik, A.R., Livak, K.J , Rafalski, J.A., and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* **18**, 6531–6535
17. Welsh, J. and McClelland, M. (1990) Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Res.* **18**, 7213–7218