Rapid Communication Genomics & Informatics Vol. 2(1) 53-56, March 2004

DNA Microarrays for Comparative Genomics: Identification of Conserved and Variable Sequences in Prokaryotic Genomes

Jae-Chang Cho*

Institute of Environmental Sciences and Department of Environmental Sciences, Hankuk University of Foreign Studies, Yongin, Korea

Keywords: DNA microarray, comparative genomics

Introduction

Bergey's manual, the authority on described microorganisms, lists several thousands bacterial species (http://www.cme. msu.edu/bergeys/), and analyses on the DNAs extracted from natural environments indicate there may be at least 10,000 species per one gram of soil (Torsvik *et al.*, 1990) and most of them are not characterized yet. Only a few bacterial genome sequences are currently available, and it is practically impossible to sequence whole genomes of all bacterial species on this planet. One of the key issues in microbiology is how to characterize and compare those extremely diverse prokaryotic genomes. This is particularly important for finding and characterizing new microorganisms for diverse purposes, from environmental remediation to pharmaceutical discovery.

Comparative genomics of bacteria starts from measuring genetic or taxonomic distances between genomes of interest. The dissimilarities between the genomes are the consequences of evolutionary diversification, and the prokaryotic evolutionary relationships have been mostly derived from sequence comparisons among 16S ribosomal RNA genes (Woese, 1987; Olsen *et al.*, 1994), although other conserved genes such as RNA polymerase gene (*rpo*), and DNA gyrase gene (*gyr*) are frequently used. However, the phylogenetic reconstruction employing similarity assessments of the aligned homologous genes or regions are likely to be biased. Different genes are under different types and intensities of selection pressures, hence the phylogenies based on different genes frequently shows incongruence. Difficulties intrinsic to the conventional approaches summarized by Karlin and Mrazek (1999) include: alignments of distantly related sequences and very long sequences (e.g., genome sequences) are generally not feasible; different phylogenetic reconstruction may result for the same set of organisms based on analysis of different protein and gene, although attempts are made to overcome this by averaging over many genes or proteins; resultant trees may be dependent on details of the alignment algorithm employed, and often assume the constant rates of evolution on the various branches, which may be violated.

In contrast to single gene phylogeny-based approaches, whole genomic DNA-DNA hybridization method evaluates overall similarity between test genomes, and is adopted as the current official method for bacterial species determination (Wayne *et al.*, 1987). While the sequence analyses of one or several genes sample only < 1/1000 of the genome, the whole genomic DNA-DNA similarity reflects the overall differences between genomes tested, and provides more robust, at least not-biased, estimates for the taxonomic distances as well as genetic similarities between the genomes. However, in spite of the straightforward nature of the whole genomic DNA-DNA hybridization method, the method is not popularly used, since this method needs laborious cross-hybridizations to find the similarity relationships between test genomes (organisms). Moreover, the method cannot work when applied to analyze previously uncharacterized genomes (e.g., genomes of environmental isolates), because we cannot select appropriate reference genomes to be used for the cross-hybridizations. Among others, a fatal drawback of the method when applied to comparative genomics is that it gives us only single numerical value (e.g., 70% similarity). The value reflects the average similarity, but all other information is concealed in this single number.

Microarrays for comparative genomics

DNA microarrays have been widely used for functional genomics to evaluate gene expression by competitive hybridizations between different populations of mRNA expressed under different culture conditions. The relative extents of hybridizations of target mRNAs to probes on the microarray provide information on the degree of expression. Recently several research groups applied

* Corresponding author: E-mail: choje@hufs.ac.kr, Tel +82-31-330-4350, Fax +82-31-330-4529
Accepted 8 March 2004

the DNA microarray-based approach to other fields, such as SNP (single nucleotide polymorphism) and mutation detection (Hacia *et al.*, 1999; Gerry *et al.*, 1999), genetic linkage analysis and population genetics (Chakravarti, 1999), gene titration (Cho and Tiedje, 2002), and comparative genomics (Behr *et al.*, 1999; Cho and Tiedje, 2001; Murray *et al.*, 2001).

DNA microarrays can be applied to the comparative genomics with experimental designs similar to those used in functional genomics, and eliminate the above disadvantages of conventional methods. Whole genomic DNA-DNA similarity can be estimated from the similarity coefficients calculated from DNA microarray hybridization patterns. In the study of DNA microarrays fabricated with genome fragments from fluorescent *Pseudomonas* spp. (Cho and Tiedje, 2001), regression analysis showed a good agreement between DNA-DNA reassociation values and the microarray hybridization pattern similarities. The coefficient of determination (r^2) was ca. 0.7, and order 1 of linear relationship with the regression coefficient of 0.7 (slope) indicated that the microarray method is similar in resolution to the whole genomic DNA-DNA hybridization method. The microarray method showed the linearity over a broader span of DNA similarity values (50 to 100%) but provided slightly less resolution at >70% DNA similarity values than for REP-PCR fingerprinting method (Rademaker, 2000). The microarray method, however, could distinguish closely related genomes and, more importantly, provided resolution over the gap between REP-PCR fingerprinting and 16S rRNA gene analysis (Cho and Tiedje, 2000). Cluster analysis can be performed on the hybridization patterns of spotted DNA probes across all test genomes. In a gene expression data analysis, clusters indicate that the genes belonging to each cluster tend to turn on and off simultaneously, but the grouping for the comparative genomics indicates only that the hybridization patterns of the cluster members are similar to a certain degree. If the spotted DNAs on the array form such a cluster, it suggests but does not confirm conserved or variable sequences.

Evenness index

Identification of conserved and variable sequences can be achieved by characterizing the shape of hybridization signal distribution with evenness index (Cho and Tiedje, 2001). The evenness (E) value of each spotted DNA sequence, standardized entropy, is calculated based on information theory (Pielou, 1966; Legendre and Legendre, 1998) using $E = (-\sum p \log p)/\log q$, where p is the relative proportion of log-normalized hybridization signal ratio (R) and q is the total number of hybridizations performed (the number of test genomes). Since the distribution of the E values can be highly skewed (skewness = -0.86 for *Pseudomonas* spp.), the E values should be normalized. The arc cosine-transformed evenness value, θ_E, is used to represent the degree of conservation of each probe sequence. Fig. 1 describes the inherent characteristics of the evenness angle. If a spotted DNA fragment is extremely conserved in all test genomes (e.g., rRNA genes), the angle (θ_E) would show its minimum value (0°). It is noteworthy that the variable and conserved sequences cannot be reliably identified by cluster analysis, but are easily revealed by θ_E values. DNA fragments showing a small angle (high evenness) tend to show high hybridization signal ratio with low standard deviation, indicating that they show as high a hybridization signal as many genomes tested and hence can be considered conserved sequences. In contrast, DNA sequences showing a large angle (low evenness) tend to show low average signal ratio with high standard deviation, indicating that they show appreciable hybridization signal only to the very closely related genomes and hence can be considered variable sequences (Fig. 1 and 2). The average angle (θ_E) of the genome fragments sampled from fluorescent *Pseudomonas* spp. was 35° (Cho and Tiedje, 2001). DNA fragments with θ_E values lower than 1 standard deviation (SD) below the mean showed appreciable hybridization signal (R > 1) for the genomes from strains of wide taxonomic range, and DNA fragments with θ_E values of 1 SD above the mean showed appreciable hybridization only when hybridized to the reference strain. Recent studies on the genomes of world-wide collection of *Pseudomonas fluorescens* strains using the evenness index identified conserved sequences involved in information storage and processing, cellular process and metabolism. The calculated similarities between those sequences and their corresponding GenBank matches were about 85 ~ 90% with nucleotide diversity (Π) ranging from 15 ~ 22. The study also revealed that 15% of *Pseudomonas fluorescent* genomes were highly variable sequences (data not shown).

$D_{1/\tan(\theta)}$ as genome-wide genetic distance

Using the calculated θ_E values, we can also construct a relationship between θ_E value and taxonomic distance of genome (Fig. 2), where valley-shaped regions could be caused by selection pressure, resulting in subsequent speciation events. The genome fragments with low θ_E values have almost identical sequences and are distributed over a wide taxonomic range, while the fragments with high θ_E values are distributed over a narrow taxonomic range. When empirical results by Cho

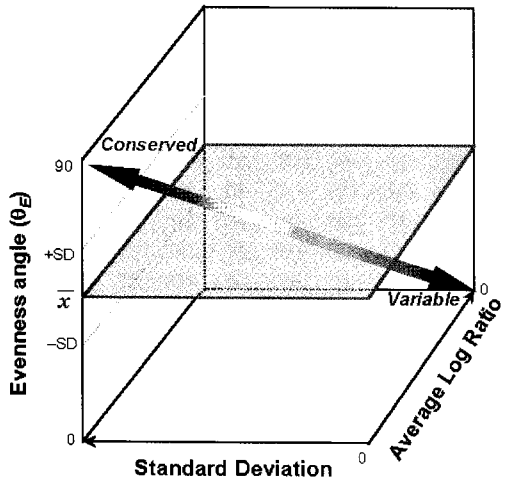

Fig. 1. Evenness value (θ E) scatter diagram, with average and SD of log hybridization signal ratio.

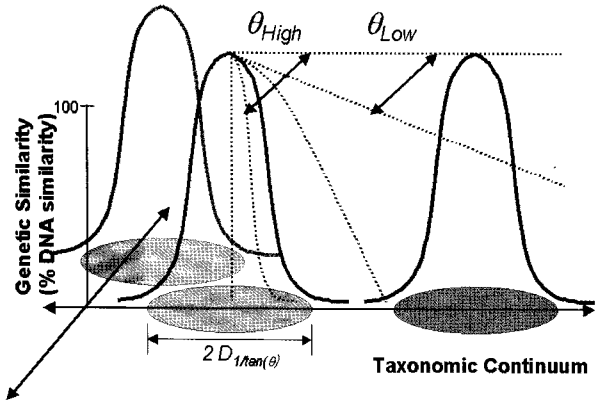

Fig. 2. Proposed relationship between θ E value and genetic distance in taxonomic continuum. Taxonomic is multidimensional, and hence, genetic similarity peaks could also be a multidimensional structure, but diagram is drawn as shown (three-dimensional) for convenience. Dotted lines indicate the degree of conservation of sequences with different θ E values.

and Tiedje (2001) were applied to this diagram, the degree of conservation within strain level, species level, closely related species level, and genus level correspond roughly to θ_E values of $>50°$, 50° to 20°, 20° to 10°, and $<10°$, respectively. Additionally, an alternative genetic (or taxonomic) distance [$D_{1/\tan(\theta)}$] can be calculated {$D_{1/\tan(\theta)} = 1/[\tan(\theta_E)]$}. The range of θ_E values for *Pseudomonas* species resulted in a $D_{1/\tan(\theta)}$ of ca. 2.7, indicating a radius of taxonomic range for a species. This alternative to calculating genetic distance by using genome-wide analyses may be useful for delineating species, although the values would be expected to vary with microbial groups.

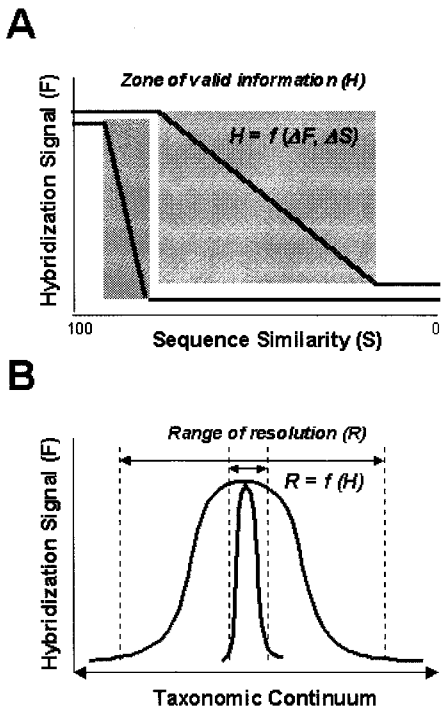

Fig. 3. Zone of valid information (A) and resolution range (B) of oligonucleotide probe-based (green line) and DNA fragment probe-based (red line) microarrays.

Oligonucleotide-based arrays vs PCR product-based arrays

The genetic content of microarray resides in the immobilized nucleic acid sequences on the microarray, and the identity of these sequences determines what information can be obtained from microarray experiments. These nucleic acids can be synthesized directly (or deposited after syntheses) on the microarray (e.g., oligonucleotides) or they can be other DNA fragments (e.g., PCR-products and purified cDNA clones), which are mechanically deposited on the array substrata. The length of oligonucleotides varies between 20 to 70 nucleotides (Li *et al.*, 2001) depending on the source of oligonucleotide content, and the oligonucleotide probes have benefits over DNA fragment probes: different parts of the same gene can be represented on the array, which enables a more robust design of array experiments; oligonucleotide probes offer precise control over the genetic composition on the array. DNA fragments probes include libraries of cDNA clones, expressed sequence tags (ESTs), and PCR-amplified fragments corresponding to open reading frames (ORFs) in genomic DNA, and the optimal length for DNA fragment probes

is between 500 to 1,000 nucleotides. Examination of potential cross hybridization between related sequences, such as those derived from a gene family, has revealed that ORF-type probes could not distinguish target DNAs with > 80% sequence similarities. Despite the above differences between oligonucleotide probes and DNA fragments probes, both give reliable results and are popularly used in functional genomics. However, when the microarrays are applied to comparative genomics, there is a noteworthy differences between oligonucleotide probes and DNA fragments probes in terms of the amount of underlying information.

The oligonucleotide microarrays could be problematic in comparative genomics because of the high specificity, which distinguishes perfect matches from even one-base mismatches. Positive matches between genomes of under comparative genomics study (i.e., not identical, but closely related genomes) require that many of the oligonucleotides are identical within each gene's set of oligonucleotide probes. Since it is unlike that several stretches of 20-base perfect matches exist for most genes in the different genomes, the oligonucleotide microarray may give many false negatives compared to the microarrays fabricated with longer DNA fragments such as ORF arrays that do not depend on such perfect matches (Dong *et al.*, 2001). When calculating similarity and correlation coefficients, comparisons of matrices with full of negatives or near zero values cause the calculated statistics to be insignificant, and following analyses may be invalid. While hybridization signals from oligonucleotide probes tend to be all-or-none like, the extents of hybridization signals from ORF-type probes tend to be progressive according to the similarities between probe and target sequences (Fig. 3). Assuming Markov chain, the information contents (H) of oligonucleotide (20-mer) and DNA fragment (1.0kb) are 3.01 and 150.51, respectively. The information content is proportional to the length of the probe, and the DNA fragment of 1.0kb provides 50 times more information on sequence differences.

Acknowledgments

This work was supported by the Korea Research Foundation Grant (KRF-2002-015-CS0057).

References

Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S., and Small, P. M. (1999). Comparative genomics of BCG vaccines by whole genomic DNA microarray. *Science* 284, 1520-1523.

Chakravarti, A. (1999). Population genetics-making sense out of sequence. *Nat. Genet.* 21, 56-60.

Cho, J.-C. and Tiedje, J. M. (2000). Biogeography and degree of endemicity of fluorescent *Pseudomonas* in soil. *Appl. Environ. Microbiol.* 66, 5448-5456.

Cho, J.-C. and Tiedje, J. M. (2001). Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Appl. Environ. Microbiol.* 67, 3677-3882.

Cho, J.-C. and Tiedje, J. M. (2002). Quantitative detection of microbial genes using DNA microarrays. *Appl. Environ. Microbiol.* 68, 1425-1430.

Dong, Y., Glasner, J. D., Blasttner, F. R., and Triplett, E. W. (2001). Genomic interspecies microarray hybridization: rapid discovery of three thousand genes in the maize endophyte, Klebsiella pneumoniae 342, by microarray hybridization with *Escherichia coli* K-12 open reading frames. *Appl. Environ. Microbiol.* 67, 1911-1921.

Gerry, N. P., Witowski, N. E., Day, J., Hammer, R. P., and Barany, G. (1999). Universal DNA microarray method for multiplex detection of low abundance point mutations. *J. Mol. Biol.* 292, 251-262.

Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., L. C. Brody, Wang, D., Lander, E. S., Lipshutz, R., Fodor, S. P., and Collins, F. S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* 22, 164-167.

Karlin, S. and Mrazek, J. (1999). Prokaryotic genome-wide comparisons and evolutionary implicatrions. In *Bacterial Genomes: Physical Structure and Analysis*, F. J. deBruijn, J. R. Lupski, G. M. Weinstock, eds. (Norwell, MA: Kluwer Academic Publishers), pp.196-212.

Legendre, P. and Legendre, L. (1998). *Numerical Ecology* (Elsevier Science, Amsterdam, The Netherlands).

Li, F. and Stormo, G. D. (2001). Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* 17, 1067-1076.

Murray, A. E., Lies, D., Li, G., Nealson, K., Zhou, J., and Tiedje, J. M. (2001). DNA/DNA hybridization to microarray reveals gene-specific differences between closely related microbial genomes. *Proc. Natl. Acad. Sci. USA* 98, 9853-9858.

Olsen, G. J., Woese, C. R., and Overbeek, R. (1994). The winds of evolutionary change: Breathing new life into microbiology. *J. Bacteriol.* 176, 1-6.

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *J. Theor. Biol.* 13, 131-144.

Rademaker, J. L., Hoste, L. W., Louws, F. J., Kersters, K., Swings, J., Vauterin, L., Vauterin, P., and deBruijn, F. J. (2000). Comparison of AFLP and rep-PCR genomic fingerprinting with DNA-DNA homology studies: *Xanthomonas* as a model system. *Int. J. Syst. Evol. Microbiol.* 50, 665-677.

Torsvik, V., Gokoyr, J., and Daae, F. L. (1990). High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* 56, 782-787.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., and Truper, H. G. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* 37, 463-464.

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* 51, 221-271.