The approach for bacterial phenotype representation by using bacterial whole genomes.

Ken Kurokawa[1], Tatsuya Takagi[2], and Teruo Yasunaga[1]

1. Genome Information Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka, JAPAN.

2. Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka, JAPAN.

1 Introduction

Since the first complete sequence determination of a bacterial whole genome[1], more than 25 bacterial genomes have been completely sequenced. Under this situation, these bacterial whole genomes have been extensively analyzed in precisely. Koonin[2] said about the emerging paradigm and problems in comparative genomics, and listed 10 major open problems in comparative genomics. In these, one of the most interesting problems is "How does the genome determine the phenotype?".

When we analyzed the relationship among bacteria, in generally, we focused on the specific single gene that is shared among all species (i.e. 16S rRNA, Ileu tRNA synthetase, ATPase, Elongation factor, Glutamine synthetase etc.)[3]. However, the species phylogenies based on comparisons of single genes are rarely consistent with each other, due to gene duplication, horizontal gene transfer or high rates of evolution[3]. Moreover, there are many cases that the classifications with genotype are inconsistent with the classifications with phenotype[4,5].

In bacterial species, the gene acquisition and gene loss by gene duplication or horizontal gene transfer are one of the most significant factors in bacterial evolution[6]. In *Thermotoga maritima*, at least 51 regions show the evidence of horizontal gene transfer[7]. Among them, 42 regions are similar to regions on the genomes of other thermophiles and the genes on the regions might represent significant thermophilic features[7]. Therefore, these transferred genes may play an important role in determining certain phenotype. Thus, it becomes important to analyze not only homologous genes shared among species but also gene set pattern that represents which gene exists or not (e.g. results of acquisition or loss of genes) in species.

In this paper, we performed the classification of bacteria based on gene set patterns of completely sequenced 25 bacterial whole genomes by using the metrical multidimensional scaling (MDS) method[8].

2 Methods

Bacterial protein sequences encoded by 25 completely sequenced genomes were gathered

from NCBI genome database[9].

The analysis described in this paper is considered to be *in silico* simulation of microarray and DNA chip strategies. First, we performed BLAST program[10] and obtained the "Score" (normalized score) as the homology score using each *Escherichia coli* gene (amino acid sequence) as the query sequence and the other bacterial genes (amino acid sequences) as the database to be searched. *E.coli* has the largest number of genes in the 25 completely sequenced bacterial genomes. Therefore, we chose *E.coli* genes as the reference gene set. Then, we listed up the highest homology score in each *E.coli* gene (homology list) for each bacterial genome. In the case that there was no homologous gene, the homology score was set to zero. Next, we made the matrix that consisted of the 24 bacterial homology lists, and constructed the Euclidean distance matrix. Finally, MDS method was applied to the distance matrix for clustering bacterial genomes and all the coordinate values were plotted versus the embedded two-dimensional space[11]. Furthermore, to examine the influence of the reference gene set, we used *Mycoplasma genitalium*, the smallest number of genes, and *Sacharomyces serevisiae*, Eucarya, as the reference gene set instead of *E.coli* gene set.

3 Results & Discussions

Fig. 1a shows the result of MDS clustering using *E.coli* genes as the reference gene set and we can find three clusters (with one exception: *T.maritima*) in the figure. Archaea and Bacteria were clearly separated with each other. Moreover, Bacteria were divided into the two separate clusters. When we used Yeast genes as the reference gene set, the result also shows three clearly separated clusters that was substantially identical to the clusters in the analysis by using *E.coli* gene set (Fig. 1b). When we used *M.genitalium* genes as the reference gene set we could separate Archaea and Bacteria. However, it was not able to separate Bacteria into two clusters (Fig. 2a). Interestingly, when we performed this analysis without Archaea, we could separate Bacteria into two clusters identical to the clusters in the analysis with *E.coli* and Yeast gene sets used (Fig. 2b). It seems that the reason of this drop of the resolution is caused by small number of the genes of *M.genitalium* used as the reference gene set. Therefore, while we could separate between distantly related species like Archaea and Bacteria because of the difference of gene set patterns between Archaea and Bacteria is large, we could not separate between neighboring species of Bacteria because of the difference of gene set pattern among Bacteria is small.

In the three analyses (Fig.1a, Fig.1b, Fig.2b), the one cluster of Bacteria consists of *Synechosystis* sp., *E.coli*, *Mycobacterium tuberculosis*, *Aquifex aeolicus*, *Neisseria meningitidis*, *Haemophilus influenzae*, *Helicobacter pylori*, *Campylobacter jejuni* (group A), and the other cluster consists of *Chlamydia trachomatis*, *Chlamydia pneumoniae*, *Rickettsia prowazekii*, *Borrelia burgdorferi*, *Treponema pallidum*, and *Deinococcus radiodurans* (group B). This classification is

inconsistent with the phylogenies based on 16S rRNA. For example, *R.prowazekii*, alpha-proteobacteria, is separated from the other proteobacteria (*N.meningitidis*, *E.coli*, *H.influenzae*, *C.jejuni*, *H.pylori*) that are classified into the group A cluster. Moreover, *Synchosystis* sp., which is positioned near Chlamydia in 16S rRNA-based phylogenies, is classified into the group A cluster in the present analysis. The bacteria classified into the group B cluster are all pathogenic bacteria without *D.radiodurans*. Among of these pathogenic bacteria, Rickettsia (*R.prowazekii*) and Chlamydia (*C.trachomatis*, *C.pneumoniae*) has a special life cycle as the obligate intracellular parasite bacteria. Spirochete (*B.burgdorferi*, *T.pallidu*) is fastidious and difficult to culture in vitro, requiring a specially enriched media and cultured cell[12]. Rickettsia, Chlamydia and Spirochete both have host parasitic life cycle. On the contrary, the bacteria classified into the group A are facultative intracellular bacteria. Thus, it seems that the clusters in the present analyses represent the phenotypic features of life cycle, especially concerned with metabolizing and host parasitic life styles.

In the 16S rRNA-based phylogeny, *D.radiodurans* clusters with Chlamydia. In our analysis, it is interesting that *D.radiodurans* is classified into the group A when using *E.coli* gene set as the reference while *D.radiodurans* is classified into the group B when using *M.genitalium* gene set as the reference. Because *E.coli* and *M.genitalium* are members of the group A and the group B bacteria respectively, these results indicate that *D.radiodurans* have the genes similar to both group A and B bacteria. It suggests that *D.radiodurans* have some phenotypical features similar to not only group A bacteria but also group B bacteria.

T.maritima was classified into the group A with Yeast gene set. However, when *E.coli* gene set was used, *T.maritima* was isolated from both Bacteria and Archaea group. These results suggests that *T.maritima* has the specific gene set not shared in both *E.coli* and Yeast, and these genes are similar to those of Archaea. The fact that *T.maritima* is placed between Bacteria and Archaea is consistent with the result of 16S rRNA based phylogenetic analysis[13]. When *M.genitalium* gene set was used as the reference, *T.maritima* was also isolated from both group A and B, however, *T.maritima* is much close to *Bacillus subtilis*. It is consistent with the fact that more than 21% of genes in *T.maritima* are homologous to *B.subtilis* genes, and these are not similar to the other bacterial genes[7]. Interestingly, these homologous genes between *T.maritima* and *B.subtilis* are also similar to genes in Mycoplasma, and this implies that these genes may represent some phenotypic features common for *E.coli*, Yeast and *M.genitalium*.

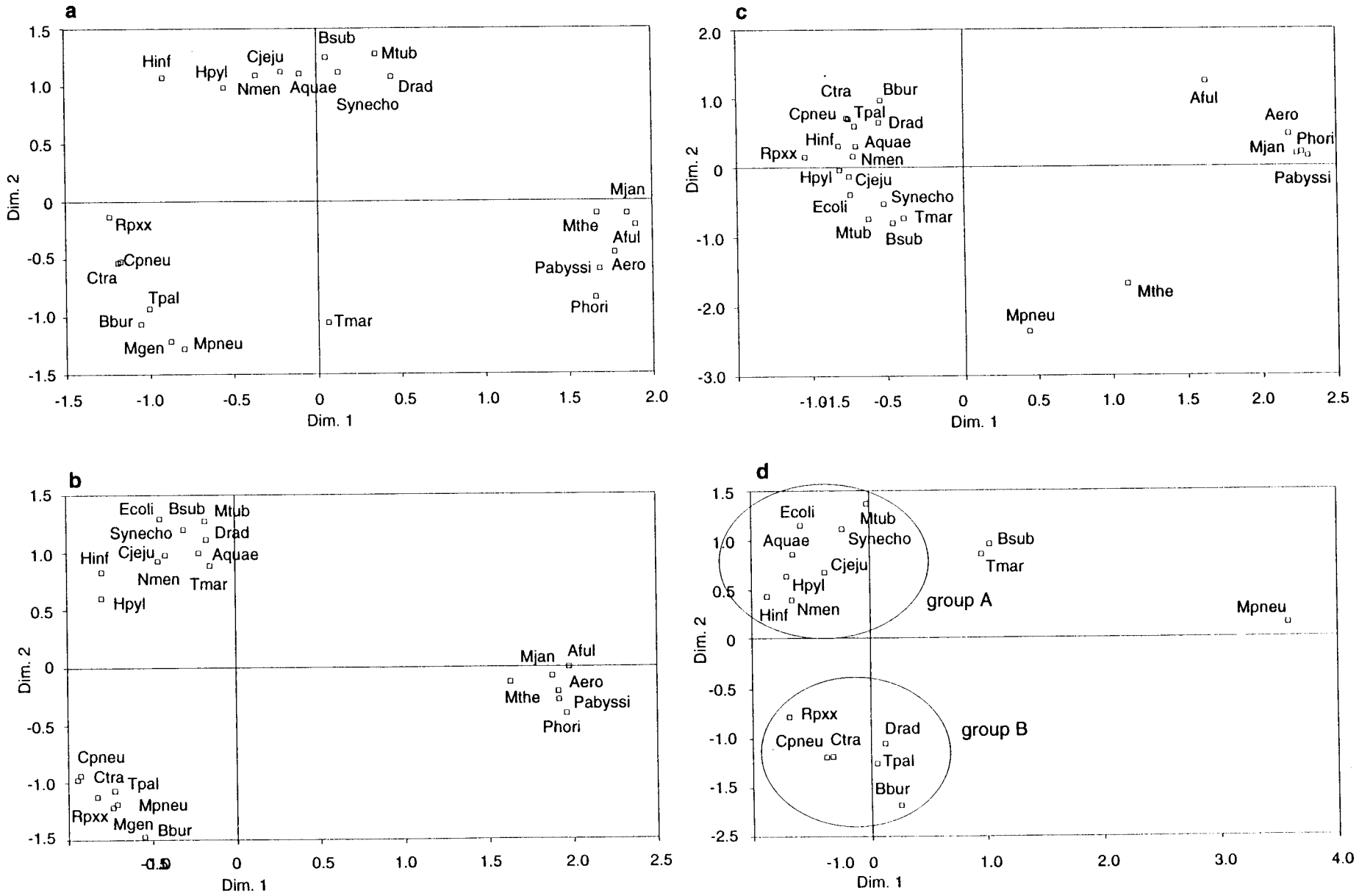
A.aeolicus has been considered as the most deeply branching species within Bacteria based on the phylogenetic analysis of 16S rRNA[14]. On the contrary, according to the phylogeny based on orthologous gene content[4], *A.aeolicus* was clustered with *Synechosystis* sp. with high bootstrap value, and this implied that *A.aeolicus* is not the deepest branching eubacteria but clearly Bacteria[4]. Furthermore, according to the phylogenetic analysis with 33 different genes for which homologues

were conserved all sequenced species, there are significant differences in the topologies of different genes, and *A.aeolicus* position is different from rRNA-based position[7]. In our analysis, *A.aeolicus* was also classified into Bacterial group B in every reference gene set. Therefore, the 16S rRNA based phylogenetic position of *A.aeolicus* should be considered to be ambiguous.

Although the clustering resolutions are different among the results for reference gene sets, the bacterial genomes are classified into the three clusters. Snel *et al.*[4] constructed the phylogenetic tree based on gene content. Their phylogenetic tree reflected the standard 16S rRNA-based tree, however, it did not correlate with phenotype[4]. In this paper, we described the bacterial genome clustering analysis based on both gene homology and gene set patterns. We can separate the Bacterial genomes into two clusters; one consists of the host parasitic bacteria and the other consists of the facultative intracellular bacteria, and our analysis can represent the specific phenotypic features especially concerned with metabolizing and host parasitic life styles. In near future, with the availability of more genomes, we will be able to give the answer to the question, "How does the genome determine the phenotype?".

Figure Legends

Figure 1. The two-dimensional plot of coordinate values of each bacterial genome (a) using *E.coli* gene set as the reference, (b) using Yeast gene set as the reference, and (c) using *M.genitalium* gene set as the reference with Archaea and (d) without Archaea. **Aero:** *Aeropyrum pernix*, **Aful:** *Archaeoglobus fulgidus*, **Aquae:** *A.aeolicus*, **Bbur:** *B.burgdorferi*, **Bsub:** *B.subtilis*, **Cjeju:** *C.jejuni*, **Cpneu:** *C.pneumoniae*, **Ctra:** *C.trachomatis*, **Drad:** *D.radiodurans*, **Ecoli:** *E.coli*, **Hinf:** *H.influenzae*, **Hpyl:** *H.pylori*, **Mgen:** *M.genitalium*, **Mjan:** *Methanococcus jannaschii*, **Mpneu:** *M.pneumoniae*, **Mthe:** *Methanobacterium thermoautotrophicum*, **Mtub:** *M.tuberculosis*, **Nmen:** *N.meningitidis*, **Synecho:** *Synechocystis* sp., **Pabyssi:** *Pyrococcus abyssi*, **Phori:** *Pyrococcus horikoshii*, **Tmar:** *T.maritima*. For a detail explanation of both group A and B, see the text.



Kurokawa, K. *et al.* Figure 1.

References

1. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).
2. Koonin, E. V. The emerging paradigm and open problems in comparative genomics. *Bioinformatics* **15**, 265-266 (1999).
3. Forterre, P. Protein versus rRNA: Problems in rooting the universal tree of life. *ASM News* **63**, 89-95 (1997).
4. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nature Gent.* **21**, 108-110 (1999).
5. Wolf, Y. I., Aravind, L, & Koonin, E. V. Rickettsiae and Chlamydiae evidence of horizontal gene transfer and gene exchange. *Trend in Genet.* **15**, 173-175 (1999).
6. Jain, R., Rivera, M. C., & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. USA* **96**, 3801-3806 (1999).
7. Nelson, K. E. *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima. Nature* **399**, 323-329 (1999).
8. Torgerson, W. Multidimensional scaling : I. Theory and method. *Psychometrika* **17**, 401-419 (1952).
9. http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html
10. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc. Acids Res.* **25**, 3389-3402 (1997).
11. The MDS analysis determined by SPSS (SPSS, Chicago, 1999), Trends 9.01 software.
12. Fraser, C. M. *et al.* Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi. Nature* **390**, 580-586 (1997).
13. Olsen, G. J., Woese, C. R. & Overbeek, R. The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **176**, 1-6 (1994).
14. Deckert, G. *et al.* The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus. Nature* **392**, 353-358 (1998).