Classification of eubacteria based on their complete genome: where does Mycoplasmataceae belong?

Kenji Sorimachi^{1*} and Teiji Okayasu²

¹Department of Microbiology, and ²Center of Medical Informatics, Dokkyo University School of Medicine, Mibu, Tochigi 321-0293, Japan

*Author for correspondence (kenjis@dokkyomed.ac.jp).

Recd 20.08.03; Accptd 13.11.03; Published online 19.01.04

The amino acid compositions of 11 Gram-positive and 12 Gram-negative eubacteria were determined from their complete genomes. They were classified into two groups, 'S-type' represented by *Staphylococcus aureus* and 'E-type' represented by *Escherichia coli*, based on their patterns of amino acid compositions determined from the complete genome. These two groups were characterized by their concentrations of Arg, Ala and Lys. Mycoplasmas, which lack a cell wall, belonged to the 'S-type', while Gram-positive mycobacteria belonged to the 'E-type'. *Rickettsia prowazekii*, *Borrelia burgdorferi*, *Campylobacter jejuni* and *Helicobacter pylori*, which are Gram-negative, belong to the 'S-type'. The classification into two groups based on their amino acid compositions determined from the complete genome was independent of Gram staining. In addition, the amino acid composition based on the plasmid resembled that based on the parent complete genome.

Keywords: amino acid composition; genome; eubacteria; Gram staining; mycoplasmas; plasmid

1. INTRODUCTION

We recently showed that the basic pattern of cellular amino acid composition is conserved from bacteria to mammalian cells, and that their differences in cellular amino acid compositions seem to reflect biological evolution (Sorimachi 1999; Sorimachi *et al.* 2000, 2001). During these studies, it was found that the cellular amino acid composition of archaea resembled that calculated from the complete genome, assuming just for calculation that all genes are expressed equally in a cell (Sorimachi *et al.* 2001). However, this coincidence puzzled us for a long time because the assumption is obviously incorrect. Also, the amino acid composition of several proteins represented that of the whole cells (Sorimachi 1999). As a protein is a gene product, amino acid composition based on several genes may represent the amino acid composition determined from the complete genomes. Indeed, the amino acid composition based on not only a gene assembly but also on a single gene encoding more than 3000 amino acid residues represented that based on the

complete genome of *Saccharomyces cerevisiae* (Sorimachi & Okayasu 2003). The consistent result was obtained from *Encephalitozoon cuniculi* consisting of 11 chromosomes (Okayasu *et al.* 2004). These results indicate that the genome is homogeneously constructed with putative gene units with similar amino acid composition and without definite restricted boundaries. The amino acid compositions based on the genomes must reflect certain cellular characteristics. Therefore, we have applied genomic data to the classification of eubacteria in the present study.

Gram staining, giving Gram-positive and Gram-negative bacteria, is an established method for classifying eubacteria, and is apparently based on structural differences in cell walls that affect the bacterial staining. Eventually, the Mycoplasmataceae, which lack a cell wall, stains Gram-negative (Krieg 2001). However, the Mycoplasmataceae evolved from Gram-positive ancestors (certain clostridia) by reduction of genome size (Brooks *et al.* 2001). In the present study, we compare a new classification based on the complete genome with that created by Gram staining, and we conclude that, based on their genomic structures, mycoplasmas belong to the *Staphylococcus aureus* 'S-type', but not to the *Escherichia coli* 'E-type'.

2. MATERIAL AND METHODS

The genome data were obtained from GenomeNet (<http://www.genome.ad.jp>) and data analysed on a personal computer. The following genomic data were used: *S. aureus* (Kuroda *et al.* 2001), *Streptococcus pyogenes* (Ferretti *et al.* 2001), *Bacillus subtilis* (Kunst *et al.* 1997), *Clostridium perfringens* (Shimizu *et al.* 2002), *Listeria monocytogenes* (Glaser *et al.* 2001), *Mycoplasma plumonis* (Chambaud *et al.* 2001), *M. genitalium* (Fraser *et al.* 1995), *M. pneumoniae* (Himmelreich *et al.* 1996), *Ureaplasma urealyticum* (Glass *et al.* 2000), *Mycobacterium tuberculosis* (Fleischmann *et al.* 2002), *Mycob. leprae* (Cole *et al.* 2001), *Rickettsia prowazekii* (Andersson *et al.* 1998), *Borrelia burgdorferi* (Fraser *et al.* 1997), *Campylobacter jejuni* (Parkhill *et al.* 2000a), *Helicobacter pylori* 26695 (Tomb *et al.* 1997), *H. pylori* J99 (Alm *et al.* 1999), *E. coli* (Blattner *et al.* 1997), *Salmonella typhi* (Parkhill *et al.* 2001a), *Vibrio cholerae* (Heidelberg *et al.* 2000), *Yersinia pestis* (Parkhill *et al.* 2001b), *Neisseria meningitidis* (Parkhill *et al.* 2000b), *Haemophilus influenzae* (Fleischmann *et al.* 1995) and *Treponema pallidum* (Fraser *et al.* 1998). The following genomic data of plasmids were used: *S. aureus* pN315B (Kuroda *et al.* 2001) and pTY4 (Yamaguchi *et al.* 2001), *B. subtilis* pTA1040 (Meijer *et al.* 1995) and p1414 (Thorsted *et al.* 1999), *C. perfringens* pCP13 (Shimizu *et al.* 2002) and pIP404 (Garnier & Cole 1988), *Bo. burgdorferi* cp32-1 and cp32-7 (Casjens *et al.* 2000), *E. coli* pB171 (Tobe *et al.* 1999) and pO157 (Makino *et al.* 1998), *Y. pestis* pCD1 and pMT1 (Hu *et al.* 1998), and *Mycob. avium* pVT2 (Kirby *et al.* 2002), pCLP (Picardeau & Vincent 1998).

3. RESULTS AND DISCUSSION

The patterns of amino acid compositions based on the complete genomes of various eubacteria are 'star shaped', as shown in figure 1.

To characterize the pattern of amino acid compositions and to classify eubacteria, we focused on particular amino acids. Concentrations of Arg, Ala and Lys changed markedly between two groups; 'S-type' represented by *S. aureus* and 'E-type' represented by *E. coli*, as shown in figure 2. Concentrations of Arg and Ala were 4.19% or less and 7.88% or less, respectively, in 'S-type', whereas they were 4.53% or more and 8.31% or more, respectively, in 'E-type' (figure 2). Concentrations of Lys were 7.10% or more in 'S-type', and 6.39% or less in the 'E-type'. The two groups are separated from each other by these values.

Further, the ratios of the next two amino acid compositions in the radar chart, [Thr]/[Arg], [Leu]/[Ile] and [Lys]/[Phe], seem to be useful factors to characterize the

Figure 1. Radar charts of amino acid compositions determined from the complete genomes of various eubacteria. The values represent the percentage of total amino acids. However, tryptophan, which was decomposed in the amino acid analysis and whose concentrations were less than 1% of the total amino acids, was neglected as described previously (Sorimachi 1999). Blue represents Gram-positive bacteria, red represents Gram-negative bacteria, while pink represents mycoplasmas, which lack a cell wall.

calculated from their complete genome (figures 1 and 2). The pattern of amino acid composition of *Mycoplasma* obviously resembled that of *S. aureus*, compared with that of *E. coli*. In Gram-positive bacteria, the patterns of the amino acid compositions of *S. aureus*, *Str. pyogenes*, *C. perfringens*, *B. subtilis* and *L. monocytogenes*, resembled each other, but those of *Mycob. tuberculosis* and *Mycob. leprae*, which resembled an 'E-type', differed. However, in Gram-negative bacteria, the patterns of the amino acid compositions of *E. coli*, *Sa. typhi*, *V. cholerae*, *Y. pestis*, *N. meningitidis*, *H. influenzae* and *T. pallidum* resembled each other (figure 1), but those of *Ca. jejuni*, *R. prowazekii* and *Bo. burgdorferi* differed. *Treponema pallidum* and *Bo. burgdorferi*, which belong to the same Spirochaetaceae, have each been classified into different types. The pattern of the amino acid composition of mycoplasmas obviously resembled that of *S. aureus*, compared with that of *E. coli*. These results indicate that eubacteria are classifiable into the following two groups: *S. aureus* 'S-type'; and *E. coli* 'E-type', based on their genomic structures, and that the Mycoplasmataceae are 'S-type'.

The amino acid composition encoded by the plasmid was also determined from its complete genome, as shown in electronic Appendix A (figure 3) (available on The Royal Society's Publications Web site). The pattern of amino acid compositions based on plasmids resembled that based on the complete genome of the parent bacterium because the number of amino acid residues encoded by many plasmids was more than 3000. Even though the numbers of amino acid residues encoded by the *B. subtilis* plasmids p1414 (1900) and pTA1040 (1438), *C. perfringens* plasmid pIP404 (2613) and *Mycob. tuberculosis* plasmid pVT2 (2769) were smaller than 3000, the amino acid compositions based on these plasmids resembled those based on their parent genomes. However, a degree of similarity of the amino acid composition between the plasmid and the parent genome reduces naturally with the reduction of plasmid size. Evidently, the rule that the amino acid composition based on a gene assembly encoding more than 3000 amino acid residues represents that based on the complete genome (Sorimachi & Okayasu 2003) is applicable not only to the complete parent genome but also to the plasmid. These results suggest that a plasmid might be derived from its parent chromosome, or that it might be transferred from the bacteria belonging to the same genus. If the number of amino acid residues encoded by a plasmid increases, the amino acid compositions of the complete parent genome and the plasmid will coincide with each other. In fact, amino acid compositions based on each chromosome coincide with each other and with that based on the complete genome of *Sac. cerevisiae* (Sorimachi & Okayasu 2003).

'star-shape' pattern. However, their values are not good for separating the 'S-types' and 'E-types' (data not shown).

Eubacteria can be classified into two groups, 'S-type' and 'E-type', based on the amino acid compositions

Figure 2. Concentrations of Arg, Ala and Lys in various eubacteria. The grey columns represent mycoplasmas.

The present study shows that eubacteria are classified into two groups, 'S-type' and 'E-type', based on their complete genomes. This approach is applicable to the classification of eubacteria, if their genomes have been analysed.

- Alm, R. A. (and 22 others) 1999 Genomic sequence comparison of two unrelated isolates of human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180.
- Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M., Naslund, A. K., Eriksson, A. S., Winkler, H. H. & Kurland, C. G. 1998 The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140.
- Blattner, F. R. (and 16 others) 1997 The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.
- Brooks, G. F., Butel, J. S. & Morse, S. A. 2001 *Jawetz, Melnick & Adelberg's medical microbiology*, 22nd edn. New York: McGraw-Hill.
- Casjens, S. (and 14 others) 2000 A bacterial genome in flux: the twelve linear and circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol. Microbiol.* **35**, 490–516.
- Chambaud, I. (and 11 others) 2001 The complete genome sequence of the murin respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acid Res.* **29**, 2145–2153.
- Cole, S. T. (and 42 others) 2001 Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011.
- Ferretti, J. J. (and 22 others) 2001 Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl Acad. Sci. USA* **98**, 4658–4663.
- Fleischmann, R. D. (and 39 others) 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
- Fleischmann, R. D. (and 25 others) 2002 Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184**, 5479–5490.
- Fraser, C. M. (and 24 others) 1995 The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.
- Fraser, C. M. (and 37 others) 1997 Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586.
- Fraser, C. M. (and 32 others) 1998 Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388.
- Garnier, T. & Cole, S. T. 1988 Complete nucleotide sequence and genetic organization of the bacteriocinogenic plasmid, pIP404, from *Clostridium perfringens*. *Plasmid* **19**, 134–150.
- Glaser, P. (and 54 others) 2001 Comparative genomics of *Listeria* species. *Science* **294**, 849–852.

- Glass, J. I., Lefkowitz, E. J., Glass, J. S., Heiner, C. R., Chen, E. Y. & Cassell, G. H. 2000 The complete sequence of the mucosal phageogen *Ureaplasma urealyticum*. *Nature* **407**, 757–762.
- Heidelberg, J. F. (and 25 others) 2000 DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477–483.
- Himmelreich, R., Hilbert, H., Plagens, H., Pirk, E., Li, B. C. & Herrmann, R. 1996 Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acid Res.* **24**, 4420–4449.
- Hu, P., Elliott, J., McCready, P., Skowronski, E., Garnes, J., Kobayashi, A., Carrano, A. V., Brubaker, R. & Garcia, E. 1998 Structural organization of virulence-associated plasmids of *Yersinia pestis*. *J. Bacteriol.* **180**, 5192–5202.
- Kirby, C., Waring, A., Griffin, T. J., Falkinham, J. O., Grindley, N. D. & Derbyshire, K. M. 2002 Cryptic plasmids of *Mycobacterium avium*: Tn552 to the rescue. *Mol. Microbiol.* **43**, 173–186.
- Krieg, N. R. 2001 Prokaryotic domains. In *Bergey's manual of systematic bacteriology*, 2nd edn, vol. 1 (ed. G. M. Garrity, D. R. Boone & R. W. Castenholz), pp. 21–25. Heidelberg: Springer.
- Kunst, F. (and 150 others) 1997 The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.
- Kuroda, M. (and 36 others) 2001 Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet* **357**, 1225–1240.
- Makino, K. (and 16 others) 1998 Complete nucleotide sequences of 93 kb and 3.3 kb plasmids of an enterohemorrhagic *Escherichia coli* O157:H7 derived from Sakai outbreak. *DNA Res.* **5**, 1–9.
- Meijer, W. J., Venema, G. & Bron, S. 1995 Characterization of single strand origins of cryptic rolling-circle plasmids from *Bacillus subtilis*. *Nucleic Acid Res.* **23**, 612–619.
- Okayasu, T., Ebara, Y. & Sorimachi, K. 2004 Similarity of amino acid composition based on gene assembly and different gene-size distributions among the 11 chromosomes in *Encephalitozoon cuniculi*. *Dokkyo J. Med. Sci.* (In the press.)
- Parkhill, J. (and 20 others) 2000a The genomic sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668.
- Parkhill, J. (and 27 others) 2000b Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–506.
- Parkhill, J. (and 40 others) 2001a Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848–852.
- Parkhill, J. (and 34 others) 2001b Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523–527.
- Picardeau, M. & Vincent, V. 1998 Mycobacterial linear plasmids have an invertebrate-like structure related to other linear replicons in actinomycetes. *Microbiology* **144**, 1981–1988.

- Shimizu, T., Ohtani, K., Hirakawa, H., Ohshima, K., Yamashita, A., Shiba, T., Ogasawara, N., Hattori, M., Kuhara, S. & Hayashi, H. 2002 Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. *Proc. Natl Acad. Sci. USA* **99**, 996–1001.
- Sorimachi, K. 1999 Evolutionary changes reflected by the cellular amino acid composition. *Amino Acids* **17**, 207–226.
- Sorimachi, K. & Okayasu, T. 2003 Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. *Mycoscience* **44**, 415–417.
- Sorimachi, K., Okayasu, T., Akimoto, K. & Niwa, A. 2000 Conservation of the basic pattern of cellular amino acid composition during biological evolution in plants. *Amino Acids* **18**, 193–196.
- Sorimachi, K., Itoho, T., Kawarabayashi, Y., Okayasu, T., Akimoto, K. & Niwa, A. 2001 Conservation of the basic pattern of cellular amino acid composition during biological evolution and the putative amino acid composition of primitive life forms. *Amino Acids* **21**, 393–399.
- Thorsted, P. B., Thomas, C. M., Poluektova, E. U. & Prozorov, A. A. 1999 Complete sequence of *Bacillus subtilis* plasmid p1414 and comparison with seven other plasmid types found in Russian soil isolates of *Bacillus subtilis*. *Plasmid* **41**, 274–281.
- Tobe, T., Hayashi, T., Han, C. G., Schoolnik, G. K., Ohtsubo, E. & Sasakawa, C. 1999 Complete DNA sequence and structural analysis of the enteropathogenic *Escherichia coli* adherence factor plasmid. *Infect. Immun.* **67**, 5455–5462.
- Tomb, J.-F. (and 41 others) 1997 The complete sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547.
- Yamaguchi, T., Hayashi, T., Takami, H., Ohnishi, M., Murata, T., Nakayama, K., Asakawa, K., Ohara, M., Komatsuzawa, H. & Sugai, M. 2001 Complete nucleotide sequence of a *Staphylococcus aureus* exfoliative toxin B plasmid and identification of a novel ADP-ribosyltransferase, EDIN-C. *Infect. Immun.* **69**, 7760–7771.

Visit www.journals.royalsoc.ac.uk and navigate to this article through *Biology Letters* to see the accompanying electronic appendix.