The phylogenetic placement of the non-phototrophic, Gram-positive thermophile '*Thermobaculum terrenum*' and branching orders within the phylum '*Chloroflexi*' inferred from gene order comparisons

Takashi Kunisawa

Correspondence
Takashi Kunisawa
kunisawa@rs.noda.tus.ac.jp

Department of Applied Biological Sciences, Science University of Tokyo, 2641 Yamasaki, Noda 278-8510, Japan

The phylogenetic position of an anaerobic, non-spore-forming thermophile '*Thermobaculum terrenum*' was investigated on the basis of gene order data from completely sequenced bacterial genomes. Gene order data can be an excellent source of phylogenetic information. Shared gene arrangements are unlikely to have arisen by chance convergence. They are likely to reflect common ancestry. '*Thermobaculum terrenum*' was found to share three gene arrangements that are present uniquely in genomes of members of the phylum '*Chloroflexi*', indicating convincingly that '*Thermobaculum terrenum*' is a member of this phylum. Branching orders within the phylum '*Chloroflexi*' were inferred by identifying monophyletic groups of species, which were circumscribed by characteristic gene arrangements. The branching orders thus inferred were in good agreement with previously reported phylogenies based on single 16S rRNA gene sequences and on multiple protein sequences. The gene order comparisons revealed a close phylogenetic affinity of '*Thermobaculum terrenum*' to *Sphaerobacter thermophilus* and *Thermomicrobium roseum*.

INTRODUCTION

'*Thermobaculum terrenum*' was originally isolated from an extreme thermal soil in Yellowstone National Park, USA (Botero *et al.*, 2004). The initial phylogenetic analysis of 16S rRNA gene sequences placed this anaerobic, non-spore-forming hyperthermophilic bacterium as being most closely related to the phyla '*Chloroflexi*' and '*Thermomicrobia*' (Botero *et al.*, 2004). The phylum '*Thermomicrobia*' was later emended to be a class within the phylum '*Chloroflexi*' on the basis of more sequence data from the 16S rRNA genes (Hugenholtz & Stackebrandt, 2004). This phylogenetic replacement has been supported by a recent phylogenomic analysis of *Thermomicrobium roseum* based on a concatenated alignment of 31 housekeeping gene products (Wu *et al.*, 2009). The genome of '*Thermobaculum terrenum*' ATCC BAA-798 was recently completely sequenced (GenBank accession no. NC_013525, S. Lucas and others); the genomic sequence data were produced by the US Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov/>). This bacterium is annotated as an unclassified bacterium in the NCBI genome database. Thus, it is unclear whether '*Thermobaculum terrenum*' is a member of the

phylum '*Chloroflexi*', as in the case of *Thermomicrobium roseum*, or whether this species constitutes its own phylum.

Phylogenetic relationships have commonly been inferred based on comparisons of sequence data from a single protein (or gene). Recently, it has widely been accepted that genome trees reconstructed based on a concatenated alignment of multiple protein sequences are less susceptible to stochastic errors than those based on a single protein or gene and provide more reliable phylogenetic information (e.g. Delsuc *et al.*, 2005). Identification of insertion/deletion sequences characteristically conserved among a group of species (signature sequences) provides another line of phylogenetic information that is obtained independently of the genome tree method (Gupta, 1998, 2010). Signature proteins that are present in a specific group of bacteria also provide phylogenetic information (see references cited in Gupta, 2010). The other approach is to compare the order of genes among different genomes (Sankoff *et al.*, 1992). If genome rearrangements occur at random positions on a bacterial genome harbouring more than 1000 genes, then the probability of finding gene B next to a given gene A is approximately 1/1000. Thus, the shared gene adjacency reflects common ancestry rather than chance convergence. Although gene order is generally not a conserved property, restricted portions of bacterial

Abbreviation: COG, clusters of orthologous groups.

genomes show gene arrangements that have been conserved between distantly related phyla. In previous papers, I have described four sets of gene arrangements that are found uniquely in a group of the major bacterial phyla, 'Firmicutes', 'Actinobacteria', 'Thermotogae', 'Deinococcus-Thermus', 'Fusobacteria' and 'Chloroflexi' (Kunisawa, 2006, 2010). In contrast, other phyla, such as the 'Proteobacteria', 'Chlorobi', 'Verrucomicrobia' and 'Acidobacteria', show alternative gene arrangements. From these comparisons, the phylum 'Chloroflexi', for instance, is inferred to be grouped within the former group of phyla. It is to be noted that, unlike the genome tree and indel methods, gene order comparison does not rely on multiple sequence alignments and therefore can provide an independent measure of the reliability of evolutionary trees obtained using aligned sequences. The accuracy of multiple sequence alignment programs has been the object of numerous computational studies (Blackshields *et al.*, 2006; Edgar & Batzoglou, 2006; Notredame, 2007).

In this work, the phylogenetic placement of '*Thermobaculum terrenum*' is investigated on the basis of gene order comparison of completely sequenced bacterial genomes. According to the NCBI Genome Database, at the time of writing, 11 genomic sequence data of six genera, '*Dehalococcoides*', '*Roseiflexus*', '*Chloroflexus*', '*Herpetosiphon*', '*Sphaerobacter*', '*Thermomicrobium*', are available within the phylum 'Chloroflexi'. If one identifies gene arrangements that are shared characteristically by 'Chloroflexi' and '*Thermobaculum terrenum*' genomes but are not found elsewhere, then '*Thermobaculum terrenum*' is likely to be a member of the phylum 'Chloroflexi'. I describe three such gene arrangements uniquely found in genomes of members of the phylum 'Chloroflexi', which serve to define and circumscribe that phylum. I also attempt to reconstruct branching orders within the phylum 'Chloroflexi' by identifying a monophyletic group of species from gene transfer-suggesting arrangements (Kunisawa, 2001, 2003). The branching orders thus inferred from gene order comparisons are consistent with the recently reported genome tree based on a concatenated alignment of sequences from species of the phylum 'Chloroflexi' (Wu *et al.*, 2009), although '*Thermobaculum terrenum*' is not included in their analysis. The gene order comparisons indicate that '*Thermobaculum terrenum*' is most closely related to members of the genera *Sphaerobacter* and *Thermomicrobium*.

METHODS

Gene order data. Complete genomic sequence data of bacteria were downloaded from the ftp server at NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). Gene order data for 590 bacterial genomes were prepared as described in a previous paper (Kunisawa, 2010), in which a substantial portion of the genomes compared in this study was listed. The downloaded genomic sequence data of the phylum 'Chloroflexi' were as follows: '*Dehalococcoides ethenogenes*' 195 (GenBank accession no. CP00027, Seshadri *et al.*, 2005), '*Dehalococcoides*' sp. CBDB1 (NC_007356, Kube *et al.*, 2005) and

BAV1 (NC_009455, A. Copeland and others), '*Roseiflexus castenholzii*' DSM 13941^T (CP000804, A. Copeland and others) and '*Roseiflexus*' sp. RS-1 (NC_009523, A. Copeland and others), '*Chloroflexus aurantiacus*' J-10-fl (NC_010175, A. Copeland and others), '*Chloroflexus aggregans*' DSM 9485^T (NC_011831, S. Lucas and others), '*Chloroflexus*' sp. Y-400-fl (NC_012032, S. Lucas and others), '*Herpetosiphon aurantiacus*' ATCC 23779^T (NC_009972, A. Copeland and others), '*Sphaerobacter thermophilus*' DSM 20745^T (NC_013523, S. Lucas and others) and '*Thermomicrobium roseum*' DSM 5159^T (NC_011959, Wu *et al.*, 2009). Most of these sequence data were produced by the US Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov/>).

Using the Clusters of Orthologous Groups of Proteins (COG) Database (Tatusov *et al.*, 2001) in which homologous proteins are uniquely labelled by a four digit number, the gene order of a genome was first represented as an order of COG numbers. Then, genes encoding tRNA or rRNA were inserted into the COG order according to their base coordinates. Gene orders thus prepared are ready for comparison between different genomes.

Computer methods. A computer program was designed in order to search for specific gene arrangements, *a-x-b* on some genomes, say genomes 1 and 2, but *c-x-d* on other genomes 3 and 4. Here gene *x* is located between genes *a* and *b*, or between genes *c* and *d*. If such arrangements are found, then it is likely that gene *x* left the neighbourhood of genes *a* and *b* and arrived at another genomic position located adjacent to gene *c* and *d* in a common ancestor of genomes 3 and 4, or vice versa. Similar reasoning can be applied to the gene arrangements, *x-y* in some genomes but *a-x* and *b-y* in others, indicating a split (or formation) of the adjacency of genes *x* and *y*. If genome 1 can be regarded as an outgroup for the rest of the genomes, then it is likely that genome 2 diverged first and then a transfer of gene *x* has occurred in a common ancestor of genomes 3 and 4. Therefore, if such gene transfer-suggesting arrangements are identified, it can be inferred that genome 2 diverged prior to the split of genomes 3 and 4 (Kunisawa, 2001, 2003). Repeated computer searches for such gene arrangements led to the reconstruction of the branching orders of the genomes under comparison. The computer search was restricted to gene *x* that was present in a single copy on the '*Thermobaculum terrenum*' genome. In many cases the search conditions were relaxed to *a-x-b* in some genomes, but *c-x* (or *x-d*) in others. Furthermore, alternative gene arrangements, *a-x* or *c-x*, are discussed in this work.

RESULTS

Phylogenetic placement of the genus '*Thermobaculum*'

The computer search identified three gene arrangements that were present uniquely in species of the phylum 'Chloroflexi' and '*Thermobaculum terrenum*'. The first example was found around the *nusB* gene (COG0781) encoding a transcription termination protein. In Fig. 1 the gene arrangements of the species of the phylum 'Chloroflexi' are compared with those of other species from the phyla 'Thermotogae', 'Dictyoglom?', 'Deinococcus-Thermus', 'Fusobacteria', 'Actinobacteria' and 'Firmicutes'. Genes are represented by their four digit COG number and common gene names are shown at the bottom of the gene alignment. Homologous genes are vertically aligned and contiguous genes are connected by dashes, while split segments are interrupted by forward slashes. For the

Thermotogae (11/11)									
Thermotoga maritima	0006/ 0231	-----	n-----1302--0781	-----	0190--1570--1722--1154/		n	0331/ 1028/	0236
Kosmotoga olearia	0006--0231	-----	n-----1302--0781	-----	0190-----n-----1154/		/	0332--0331	0236
Dictyoglomi (2/2)									
Dictyoglomus thermophilum	0006--0231--0511	-----	0439--1302--0781	-----	n-----n-----1154/		0333--0332--0331--1028		0236
Dictyoglomus turgidum	0006--0231--0511	-----	0439--1302--0781	-----	n-----n-----1154/		0333--0332--0331--1028		0236
Deinococcus-Thermus (8/8)									
Deinococcus geothermalis	0006/ 0231--0511	-----	0439/ 1302--0781	-----	0190/ 1570/ n / 1154/		/	0332--0331--1028	0236
Thermus thermophilus HB27	0006/ 0231--0511	-----	0439--1302--0781	-----	0190/ n / n / 1154/		0333--0332--0331--1028		0236
Fusobacteria (4/4)									
Fusobacterium nucleatum	0006/ 0231/ 0511	/ n	1302--0781	/	0190/ 1570/ 1722/ 1154/		/	0332--0331	0236
Leptotrichia buccalis	0006/ 0231/ 0511	/	0439--1302--0781	/	0190/ 1570/ n / 1154/		0333--0332--0331/ 1028/		0236
Actinobacteria (75/78)									
Mycobacterium tuberculosis	0006--0231	-----	-----0781	-----	0190/ 1570--1722/ 1154/		/	0332/ 0331	0236
Atopobium parvulum	0006--0231--n	-----	n-----1302--0781	-----	0190/ 1570--1722/ 1154/		0333--n-----n		0236
Firmicutes (111/113)									
Thermoanaerobacter sp. X514	0006--0231/ n	n	1302--0781	-----	0190--1570--1722--1154/		0333--0332--0331--1028		0236
Staphylococcus aureus	0006--0231--0511	-----	0439--1302--0781	-----	0190/ 1570--1722/ n /		/	0332/ 0331--1028	0236
	pepQ efp accB		accC yqhY nusB		folD xseA xseB dxs		rpmF fabH fabD fabG		acpP
Chloroflexi									
Dehalococcoides ethenogenes	0006--0231/ n				n n /		0333--n-----0331--1028	n-----n-----0781	0236
Dehalococcoides sp. CBDB1	0006--0231/ n				n n /		0333--n-----0331--1028	n-----n-----0781	0236
Dehalococcoides sp. BAV1	0006--0231/ n				n n /		0333--n-----0331--1028	n-----n-----0781	0236
Herpetosiphon aurantiacus	0006--0231--0511	-----	-----	-----	1570/ 1722/		0333--0332--0331--1028	0439--1302--0781	0236
Roseiflexus sp. RS-1	0006/ 0231/ 0511	-----	-----	-----	1570--1722/		0333--0332/ 0331/ 1028	n-----n-----0781	0236
Roseiflexus castenholzii	0006/ 0231/ 0511	-----	-----	-----	1570--1722/		0333--0332/ 0331/ 1028	n-----n-----0781	0236
Chloroflexus aurantiacus	0006/ 0231/ 0511	-----	-----	-----	1570--1722/		0333--0332/ 0331/ 1028	0439--n-----0781	0236
Chloroflexus aggregans	0006/ 0231/ 0511	-----	-----	-----	1570--1722/		0333--0332/ 0331/ 1028	0439--n-----0781	0236
Chloroflexus sp. Y-400-fl	0006/ 0231/ 0511	-----	-----	-----	1570--1722/		0333--0332/ 0331/ 1028	0439--n-----0781	0236
Sphaerobacter thermophilus	0006--0231--0511	-----	-----	-----	1570--1722/		0333--0332--0331--1028	0439--n-----0781	0236
Thermomicrobium roseum	0006--0231--0511	-----	-----	-----	1570--1722/		0333--0332--0331--1028	0439--n-----0781	0236
Thermobaculum terrenum	0006--0231--0511	-----	-----	-----	1570/ n /		0333--0332--0331--1028	0439--n-----0781	0236
	pepQ efp accB				xseA xseB		rpmF fabH fabD fabG accC yqhY nusB		acpP

Fig. 1. Gene arrangements suggesting a transfer of the *nusB* (COG0781) gene. Genes are represented by the four-digit COG number and common gene names are shown at the bottom. Homologous genes are vertically aligned. Contiguous genes are connected by '-', while split segments are interrupted by '/'. For the maximum alignment, open reading frames are represented simply as '.', and 'n' denotes cases where the corresponding gene is not identified. Gene orders are shown for all the members of the phylum 'Chloroflexi', whereas conserved orders are illustrated for only two members belonging to other phyla. The transferred genes are boxed. Note that 'Thermobaculum terrenum' shows a gene arrangement similar to those present in the species of the phylum 'Chloroflexi'.

maximum alignment, open reading frames are represented simply as dots, and 'n' is denoted when the corresponding gene is not identified. For example, the *pepQ* (COG0006) gene is located immediately upstream of the *efp* gene (COG0231) in *Kosmotoga olearia*, while in another member of the phylum 'Thermotogae', *Thermotoga maritima*, the *pepQ* gene is not adjacent to the *efp* gene and is located elsewhere on its genome. The *xseB* gene (COG1722) is not identified in the genome of *K. olearia*, and 'n' is indicated at the corresponding position in the gene alignment. As can be seen in Fig. 1, in those phyla shown, i.e. 'Thermotogae', 'Dictyoglomi', 'Deinococcus-Thermus', 'Fusobacteria', 'Actinobacteria' and 'Firmicutes' (superphylum 1), a conserved gene cluster (*accC*)-*yqhY*-*nusB* is found between the *efp* (COG0231) and *folD* (COG0190) genes. In order to show the degree of gene order conservation, the number of species showing the indicated or similar order and the number of genomes compared are indicated as the numerator and denominator, respectively, in the parentheses following the phylum name. In contrast to the arrangements depicted in Fig. 1, the other group of phyla, i.e. 'Aquificae', 'Nitrospirae', 'Acidobacteria', 'Verrucomicrobia'

and 'Proteobacteria' (superphylum 2), the *nusB* gene tends to be located immediately downstream of the *ribH* gene (COG0054), as described in previous papers (Kunisawa, 2006, 2010). Some of the species of the phylum 'Chloroflexi' show the adjacency *accC*-*nusB* as seen in superphylum 1. Based on this gene adjacency, the phylum 'Chloroflexi' has been inferred to be a member of superphylum 1 (Kunisawa, 2006, 2010). However, the gene location of *nusB* is different between the phylum 'Chloroflexi' and other members of superphylum 1. In species of the phylum 'Chloroflexi', the entire cluster *accC*-*yqhY*-*nusB* or its fragments, *accC*-*nusB* or *nusB* alone, are present between the *fabG* (COG1028) and *acpP* (COG0236) genes, while in other superphylum 1 species, the *fabG* and *acpP* genes are adjacent to each other. Since the *nusB* gene is present in a single copy in almost all the bacterial genomes, the most straightforward explanation of this observation may be that the *accC*-*yqhY*-*nusB* cluster was translocated in evolution from the region adjacent to the *efp* and *folD* genes to a different region neighbouring the *fabG* and *acpP* genes within a common ancestral genome of the species of the phylum 'Chloroflexi'. In 'Thermobaculum terrenum', the *accC*-*nusB* arrangement is present between the

fabG and *acpP* genes just as in the species of the phylum 'Chloroflexi', indicating convincingly that '*Thermobaculum terrenum*' is a member of the phylum 'Chloroflexi'.

Fig. 2 shows two other gene arrangements shared uniquely by the species of the phylum 'Chloroflexi' and '*Thermobaculum terrenum*'. In this identification, computer search conditions on gene arrangements were relaxed to those of *a-x-b* in some genomes and *c-x* (or *x-d*) in other genomes. While the *rpmE* (COG0254) gene encoding ribosomal protein L31 is located between the *rho* (COG1158) and *hemK* (COG2890) genes in most of the members of the phyla 'Firmicutes' and 'Actinobacteria' as shown in Fig. 2a, the ribosomal protein gene is found immediately downstream of a gene cluster *rplU* (COG0261)–*rpmA* (COG0211) in the species of the phylum 'Chloroflexi' (Fig. 2a), suggesting a transfer of the *rpmE* gene in their common ancestor. Since the *rplU*–*rpmA*–*rpmE* gene arrangement is present in the genome of '*Thermobaculum terrenum*', but is not found in other species, '*Thermobaculum terrenum*' is most likely to be a member of the phylum 'Chloroflexi'. Another example of such a gene transfer is inferred from the gene arrangements shown in Fig. 2b; while the *mazG* (COG1694) gene is present between the *mfd* (COG1197) and *himA* (COG0776) genes in some species from the phyla 'Actinobacteria' and 'Firmicutes', its orthologue is found downstream of the *lepA* (COG0481) gene or *rpsT* (COG0268) gene in the phylum 'Chloroflexi', accompanying the *surA* (COG0760) gene in some cases. Here, the *rpsT* gene is encoded on the complementary strand and is denoted with an asterisk. Thus, the *mazG* gene, probably together with the *surA* gene, is likely to have been

transferred in a common ancestor of the species of the phylum 'Chloroflexi' with a reservation about *H. aurantiacus*, in which the synteny is disrupted. As can be seen in Fig. 2b, the gene arrangement around the *mazG* gene in '*Thermobaculum terrenum*' is identical or similar to those found in other members of the phylum 'Chloroflexi'. This observation provides empirical evidence that '*Thermobaculum terrenum*' should be included in the phylum 'Chloroflexi'.

Branching orders within the phylum 'Chloroflexi'

It is attempted here to infer branching orders among the species of the phylum 'Chloroflexi' on the basis of gene order comparisons.

(i) '*Dehalococcoides*' as a deepest branching genus.

The genus '*Dehalococcoides*' appears to be a most deeply branching genus among the completely sequenced species of the phylum 'Chloroflexi' on the basis of the following gene arrangements that are uniquely found only in a specific group of the 'Chloroflexi' species. As shown in Fig. 3a, the *tatC* gene (COG0805) coding for a Sec-independent protein secretion pathway component is located immediate downstream of the *nusB* (COG0781)–*acpP* (COG0236) cluster in the species of the phylum 'Chloroflexi' other than those of the genus '*Dehalococcoides*'. In contrast, *tatC* orthologues are not located adjacent to the *acpP* gene, but are located downstream of the *tatA* (COG1826) gene in the genus '*Dehalococcoides*'. The adjacency of *tatA*–*tatC* can be recognized in various bacteria from the phyla 'Firmicutes', 'Actinobacteria' or '*Deinococcus*–*Thermus*'. Therefore, the

Fig. 2. Gene arrangements present uniquely in the species of the phylum 'Chloroflexi'. Suggested transfers of gene(s) in a common ancestor of the phylum 'Chloroflexi'; (a) *rpmE* (COG0254), (b) *surA* (COG0760)–*mazG* (COG1694). The *rpmE* gene (COG0254) resides on the complementary strand and is denoted with an asterisk.

Fig. 3. Gene arrangements shared among the species of the genus '*Dehalococcoides*' and bacterial species belonging to other phyla. (a) Transfer of the *tatC* (COG0805) gene. (b) Transfer of the *recA* (COG0468)-*oraA* (COG2137) gene arrangement.

tatA-tatC arrangement is likely to be ancestral, and the alternative *acpP-tatC* arrangement is thought to have been derived through evolution. The latter arrangement is thus inferred to have been created in the lineage leading to *H. aurantiacus*, *Roseiflexus* species, *Chloroflexus* species, *S. thermophilus*, *Thermomicrobium roseum* and '*Thermobaculum terrenum*', after the divergence of the genus '*Dehalococcoides*'. Another conserved adjacency between the two-gene clusters *recA* (COG0468)-*oraA* (COG2137) and *yddR* (COG2220)-*pyrG* (COG0504) is found in the members of the genera *Roseiflexus* and *Chloroflexus*, *S. thermophilus*, *Thermomicrobium roseum* and '*Thermobaculum terrenum*' (Fig. 3b). In contrast, the species of the genus '*Dehalococcoides*' show that the former gene cluster is present immediately upstream of the *ymdA* gene (COG1418) as in other bacteria from the phyla '*Thermotogae*', '*Actinobacteria*' and '*Firmicutes*', suggesting a transfer of *recA-oraA* in a common ancestor of the species of the phylum '*Chloroflexi*' other than species of the genus '*Dehalococcoides*'. Since *H. aurantiacus* forms a monophyletic clade with the members of the genera *Roseiflexus* and *Chloroflexus*, as will be described below, *recA* (COG0468)-*oraA* (COG2137), *yddR* and *pyrG* are likely to have been dispersed on the genome of *H. aurantiacus* after the adjacency of *recA-oraA* to *yddR* (COG2220)-*pyrG* (COG0504) was once created. Thus, the gene moves of *tatC* and *recA-oraA* consistently suggest an early divergence of the species of the genus '*Dehalococcoides*' within the phylum '*Chloroflexi*'.

(ii) The HRC clade. The present computer search suggested a monophyletic grouping of *H. aurantiacus*, members of the genera *Roseiflexus* and *Chloroflexus* (HRC clade), exclusive of the other species of the phylum '*Chloroflexi*'. As shown in Fig. 4, the *purH* gene (COG0138) encoding phosphoribosyl aminoimidazole carboxy formyl formyltransferase and the *hisI* gene (COG0139) encoding phosphoribosyl-AMP cyclohydrolase are located adjacent to each other in this HRC clade. In contrast, the genes *purH* and *hisI* are present separately at the 3' region of the *purM* (COG0150) or *purN* (COG0299) gene and of the *hisA* (COG0106) or *hisF* (COG0107) gene, respectively, in the species of the genus '*Dehalococcoides*', *S. thermophilus*, *Thermomicrobium roseum* and '*Thermobaculum terrenum*' as well as in some species from other phyla, the '*Firmicutes*', '*Actinobacteria*', '*Fusobacteria*' and '*Thermotogae*'. The *purH-hisI* arrangement is thus found uniquely in the HRC clade, which indicates that this gene adjacency has been created in a common ancestor of the HRC clade.

The gene arrangement *purL* (COG0047)-*purQ* (COG0046) or its reverse arrangement is found in the species of the genus '*Dehalococcoides*', *H. aurantiacus*, *S. thermophilus*, *Thermomicrobium roseum* and '*Thermobaculum terrenum*' as well as in other bacteria from the phyla '*Firmicutes*', '*Actinobacteria*' and so on, while this adjacency is disrupted in the members of the genera *Roseiflexus* and *Chloroflexus*, as seen in Fig. 5a. Furthermore, the indicated arrangements in the genera *Roseiflexus* and *Chloroflexus* are not found

Fig. 4. Gene adjacency of *purH* (COG0138)-*hisI* (COG0139) in the HRC clade.

elsewhere. Therefore, this disruption is likely to have occurred in a common ancestor of members of the genera *Roseiflexus* and *Chloroflexus* exclusive of the other species of the phylum 'Chloroflexi'. Similarly, the *gmk* (COG0194)-*yqeY* (COG1610) arrangement is thought to have been created in a common ancestor of the species of the genera *Roseiflexus* and *Chloroflexus*, since this gene adjacency is not found in other species (see Fig. 5b).

(iii) The STT clade. The present comparisons suggest another monophyletic grouping of *S. thermophilus*, *Thermomicrobium roseum* and '*Thermobaculum terrenum*' (STT clade). As shown in Fig. 6a, in the members of this group a gene cluster *clpA* (COG0542)-*sms* (COG1066) is located immediately downstream of the *rpsA* gene (COG0539) and another cluster *yacL* (COG1855)-*ispD* (COG1211) is present at the 3' region of the COG0613 gene, namely, these two clusters are located separately on their genome. The adjacencies of *clpA* to *rpsA* and of *yacL* to COG0613 are both uniquely found in this group. In contrast, these two clusters are located adjacent to each other in the species of the genus '*Dehalococcoides*' and HRC clade within the phylum 'Chloroflexi', as well as in other phyla, such as the '*Firmicutes*' and '*Dictyoglomi*', suggesting a split into the two clusters in the lineage leading to the STT clade exclusive of other members of the phylum 'Chloroflexi'.

Similarly, a unique gene adjacency *ruvC* (COG0817)-*yqx*C (COG1189) is likely to have been created in a common ancestor of the STT clade, as can be inferred from Fig. 6b. The gene arrangement *yeeI* (COG0217)-*ruvC* (COG0817) is found not only in the phyla '*Actinobacteria*' and '*Dictyoglomi*'

but also in members of the genus '*Dehalococcoides*', members of the genera *Roseiflexus* and *Chloroflexus*, and '*Thermobaculum terrenum*', while in *S. thermophilus* and *Thermomicrobium roseum* the *yeeI* (COG0217) gene is located separately from the *yqx*C (COG1189) gene and is present at the 3' region of the *srnB* gene (COG0513). From these observations, it is possible to infer that the gene adjacency between *yeeI*-*ruvC* and *yqx*C has been created in a common ancestor of the STT members and then the *yeeI* gene has been transferred to downstream of *srnB* in a common ancestor of *S. thermophilus* and *Thermomicrobium roseum*. Besides *S. thermophilus* and *Thermomicrobium roseum*, the *srnB*-*yeeI* arrangement can also be seen in *Mycoplasma penetrans* as a very rare exceptional case; 15 other completely sequenced mycoplasma genomes do not show this arrangement. This move of the *yeeI* gene suggests that '*Thermobaculum terrenum*' is most closely related to *S. thermophilus* and *Thermomicrobium roseum*.

Besides genes coding for proteins, a tRNA-specifying gene has been suggested as a candidate gene for transfer. A threonine-tRNA gene with the anticodon sequence GGT (Tggt) is present between a tyrosine-tRNA gene with the anticodon sequence GTA (Ygta) and an elongation factor encoding gene *tufB* (COG0050) in many bacteria, such as members of the phyla '*Firmicutes*', '*Actinobacteria*' and '*Dictyoglomi*' as shown in Fig. 7a. Although similar gene arrangements are found in most of the species of the phylum 'Chloroflexi', *S. thermophilus* and *Thermomicrobium roseum* possess their own unique arrangement; Tggt is not found between Ygta and the *tufB* gene but is located upstream of another threonine-tRNA gene with the anticodon sequence CGT (Tcgt). These observations

Fig. 5. Gene transfers in a common ancestor of members of the genera *Roseiflexus* and *Chloroflexus* (a) Disruption of adjacency *purL* (COG0047)-*purQ* (COG0046). (b) Transfer of *yqeY* (COG1610)-*yqfF* (COG1480).

suggest a transfer of *Tggt* in a common ancestor of *S. thermophilus* and *Thermomicrobium roseum*. While the *dnaX* and *yaaK* genes are located adjacent to each other in various bacteria including members of the genus '*Dehalococcoides*', the HCR clade, and '*Thermobaculum terrenum*', the *yaaK*-COG2141 and *yabC*-*dnaX*-*gatC* arrangements are both only found in *S. thermophilus* and *Thermomicrobium roseum* (Fig. 7b). This observation suggests that a disruption of the gene adjacency *dnaX* (COG2812)-*yaaK* (COG0718) must have occurred in a

common ancestor of *S. thermophilus* and *Thermomicrobium roseum* exclusive of the other species of the phylum '*Chloroflexi*'.

DISCUSSION

In this work the phylogenetic placement of '*Thermobaculum terrenum*' was examined on the basis of gene order comparisons of completely sequenced bacterial genomes. Compared with the more common approach based on sequence alignment, an advantage underlying the present approach is in that gene transfers that occurred in a common ancestor of a group of bacteria can be identified and gene orders present uniquely in that group can be specified. If such gene-transfers are found, then they define and circumscribe that group. This is not always possible by drawing phylogenetic trees based on sequence alignment. Along this line, three such transfers were identified that are thought to have occurred in a common ancestor of species of the phylum '*Chloroflexi*' (see Figs 1 and 2). It is to be noted here that the present method is valid when a foreign gene is introduced into the genome of a common ancestor of the species of the phylum '*Chloroflexi*' by means of lateral gene transfer. Gene arrangements created by means of lateral transfer can also serve as genomic markers if they are present uniquely in a particular group of species.

A summary of the species branching pattern within the phylum '*Chloroflexi*' obtained in this study based on the gene order comparisons is depicted in Fig. 8. The gene transfer events inferred are mapped onto the branching orders. The gene arrangements shown in Figs 1–7 are all consistently understood on the basis of the occurrence orders of the gene transfers and species branching. It is interesting to compare the branching orders deduced here with those obtained by the use of sequence alignments. Although bootstrap statistical supports were marginal, phylogenetic trees based on 16S rRNA sequences show that the species of the genus '*Dehalococcoides*' (designated subclass 2) branched off first and then a split between the HRC clade (subclass 3) and the *S. thermophilus*/*Thermomicrobium roseum* group (subclass 3) occurred (Hugenholtz *et al.*, 1998; Hugenholtz & Stackebrandt, 2004), which is in agreement with the branching orders illustrated in Fig. 8. Although sequences of '*Thermobaculum terrenum*' were not included in the sequence analysis, recently reported phylogenomic trees based on a concatenated alignment (Wu & Eisen, 2008; Wu *et al.*, 2009), show branching orders essentially identical to those inferred here. Thus, the present results provide another line of support for the branching orders that are weakly supported by the concatenation of sequence alignments.

The present analysis of gene orders revealed that '*Thermobaculum terrenum*' is most closely related to *S. thermophilus* and *Thermomicrobium roseum*. This phylogenetic affinity of '*Thermobaculum terrenum*' is also supported by a BLAST best-hit analysis. Since genes involved

Fig. 8. A summary diagram showing the relationships between gene transfers and species divergence. Note that a total of 12 gene transfers indicated are consistent with gene arrangements shown in Figs 1–7.

with those from the members of the phylum ‘*Chloroflexi*’, and 12 sequences from members of the phylum ‘*Firmicutes*’ gained the highest scores. Among the 35 ‘*Chloroflexi*’ proteins, *S. thermophilus* showed the largest number of the best hits; 13 best-hit sequences were from *S. thermophilus* followed by seven sequences from *Thermomicrobium roseum*. Thus, the BLAST analysis also suggests that ‘*Thermobaculum terrenum*’ is most closely related to *S. thermophilus* or *Thermomicrobium roseum*.

Taxonomic classification has traditionally been carried out on the basis of morphological features. In this work, instead of morphological features, gene orders were used. Since genomic sequence data of bacteria are accumulating rapidly, the method employed here is expected to be applicable to various phylogenetic problems.

ACKNOWLEDGEMENTS

In this work many unpublished genomic sequence data were used. The author would like to thank those who produced the genomic sequence data.

REFERENCES

- Blackshields, G., Wallace, I. M., Larkin, M. & Higgins, D. G. (2006). Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biol* **6**, 321–339.
- Botero, L. M., Brown, K. B., Brumefield, S., Burr, M., Castenholz, R. W., Young, M. & McDermott, T. R. (2004). *Thermobaculum terrenum* gen. nov., sp. nov.: a non-phototrophic gram-positive thermophile representing an environmental clone group related to the ‘*Chloroflexi*’ (green non-sulfur bacteria) and *Thermomicrobia*. *Arch Microbiol* **181**, 269–277.
- Delsuc, F., Brinkmann, H. & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**, 361–375.
- Edgar, R. C. & Batzoglou, S. (2006). Multiple sequence alignment. *Curr Opin Struct Biol* **16**, 368–373.
- Gupta, R. S. (1998). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* **62**, 1435–1491.
- Gupta, R. S. (2010). Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. *Photosynth Res* **104**, 357–372.
- Hugenholtz, P. & Stackebrandt, E. (2004). Reclassification of *Sphaerobacter thermophilus* from the subclass *Sphaerobacteridae* in the phylum *Actinobacteria* to the class *Thermomicrobia* (emended description) in the phylum ‘*Chloroflexi*’ (emended description). *Int J Syst Evol Microbiol* **54**, 2049–2051.
- Hugenholtz, P., Goebel, B. M. & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**, 4765–4774.
- Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* **96**, 3801–3806.
- Kube, M., Beck, A., Zinder, S. H., Kuhl, H., Reinhardt, R. & Adrian, L. (2005). Genome sequence of the chlorinated compound-respiring bacterium *Dehalococcoides* species strain CBDB1. *Nat Biotechnol* **23**, 1269–1273.
- Kunisawa, T. (2001). Gene arrangements and phylogeny in the class *Proteobacteria*. *J Theor Biol* **213**, 9–19.
- Kunisawa, T. (2003). Gene arrangements and branching orders of gram-positive bacteria. *J Theor Biol* **222**, 495–503.
- Kunisawa, T. (2006). Dichotomy of major bacterial phyla inferred from gene arrangement comparisons. *J Theor Biol* **239**, 367–375.
- Kunisawa, T. (2010). Evaluation of the phylogenetic position of the sulfate-reducing bacterium *Thermodesulfovibrio yellowstonii* (phylum *Nitrospirae*) by means of gene order data from completely sequenced genomes. *Int J Syst Evol Microbiol* **60**, 1090–1102.
- Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLOS Comput Biol* **3**, e123.
- Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F. & Cedergren, R. (1992). Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc Natl Acad Sci U S A* **89**, 6575–6579.
- Seshadri, R., Adrian, L., Fouts, D. E., Eisen, J. A., Phillippy, A. M., Methe, B. A., Ward, N. L., Nelson, W. C., Deboy, R. T. & other authors

(2005). Genome sequence of the PCE-dechlorinating bacterium *Dehalococcoides ethenogenes*. *Science* **307**, 105–108.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. & Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22–28.

Wu, M. & Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**, R151.

Wu, D., Raymond, J., Wu, M., Chatterji, S., Ren, Q., Graham, J. E., Bryant, D. A., Robb, F., Colman, A. & other authors (2009). Complete genome sequence of the aerobic CO-oxidizing thermophile *Thermomicrobium roseum*. *PLoS ONE* **4**, e4207.