Comparative sequence analyses reveal frequent occurrence of short segments containing an abnormally high number of non-random base variations in bacterial rRNA genes

Yue Wang and Zhenshui Zhang

Author for correspondence: Yue Wang. Tel: +65 7783207. Fax: +65 7791117.
e-mail: mcbwangy@imcb.nus.edu.sg

Microbial Collection and Screening Laboratory,
Institute of Molecular and Cell Biology, 30 Medical Drive, Singapore 117609

rRNA genes are thought unlikely to be laterally transferred, because rRNA must coevolve with a large number of cellular components to form the highly sophisticated translation apparatus and perform protein synthesis. In this paper, the authors first hypothesized that lateral gene transfer (LGT) might occur to rRNA genes via replacement of gene segments encoding individual domains of rRNA: the 'simplified complexity hypothesis'. Comparative sequence analyses of the 16S and 23S rRNA genes from a large number of actinomycete species frequently identified rRNA genes containing short segments with an abnormally high number of non-random base variations. These variations were nearly always characterized by complementing covariations of several paired bases within the stem of a hairpin. The nature of these base variations is not consistent with random mutations but satisfies well the predictions of the 'simplified complexity hypothesis'. The most parsimonious explanation for this phenomenon is the lateral transfer of rRNA gene segments between different bacterial species. This mode of LGT may create mosaic rRNA genes and occur repeatedly in different regions of a gene, gradually destroying the evolutionary history recorded in the nucleotide sequence.

Keywords: complexity hypothesis, lateral gene transfer, mosaic gene, phylogeny, evolution

INTRODUCTION

The rRNA gene was once considered the ideal chronometer to record the evolutionary history of life, potentially all the way back to the last common ancestor (Woese, 1987, 1998). rRNA sequence comparisons led to the construction of a 'universal tree of life', dividing all life on Earth into three equidistant domains: *Eukarya*, *Bacteria* and *Archaea* (Woese *et al.*, 1990). It is believed that this tree provides the basis for a natural classification system of organisms. Recently, the availability of complete genome sequences of organisms representing the three domains has allowed biologists to use many other gene sets to verify the topology of the universal tree. Such analyses have frequently produced

trees with branching patterns rather different from that of the universal tree (Woese, 1998; Doolittle, 1999). Sometimes organisms representing different domains were found intermixed (Woese, 1998). Currently, the popular explanation for this confusion is that a considerable fraction of the genome of any organism is the products of lateral gene transfer (LGT) (Doolittle, 1999). Lawrence & Ochman (1997) calculated that approximately 18% of the genome of *Escherichia coli* was acquired via LGT during the 100 million years after its divergence from the *Salmonella* lineage. Nelson *et al.* (1999) also reported that 24% of the genome of the bacterium *Thermotoga maritima* was introduced from *Archaea*. These numbers do not yet include genes whose origins have been obscured in the long process of amelioration after transfer or those that are from relatively closely related organisms (Lawrence & Ochman, 1997). With such a high intensity of LGT, one has to ask whether there is any gene that may still record faithfully the entire or a significant part of the evol-

Abbreviation: LGT, lateral gene transfer.

The GenBank accession numbers for the 23S rRNA sequences determined in this study are AF192136–AF192150.

Fig. 1. For legend see facing page.

utionary history of life. Analyses of different gene sets also revealed that not all genes are equally susceptible to LGT (Rivera *et al.*, 1998; Jain *et al.*, 1999). When genes are grouped according to biological functions, those involved in transcription, translation and DNA replication are seldom identified as products of LGT (Jain *et al.*, 1999; Koonin *et al.*, 1999; Nelson *et al.*, 1999). In contrast, a considerable fraction of metabolic genes are believed to be laterally transferred. In the *Thermotoga maritima* genome, only six out of 124 translation-related genes exhibit higher similarities to *Archaea* than to *Bacteria*, while 92 out of 181 genes involved in electron transport are more closely related with *Archaea* than with *Bacteria* (Nelson *et al.*, 1999). The hypothesis interpreting the greatly unequal probability of lateral transfer of genes with different functions is that the chance of successful transfer of a gene is inversely correlated with the complexity of interactions its product makes with other cellular components (Woese, 1998; Rivera *et al.*, 1998; Jain *et al.*, 1999). Jain *et al.* (1999) called this the 'complexity hypothesis'. The rRNA genes may be among those least likely to be laterally transferred, because rRNAs must interact in a highly coordinated manner with more than 100 other gene products to form the ribosomes and carry out protein synthesis.

Although the complexity hypothesis is largely correct, it overlooks one important aspect of LGT: the gene is not the smallest unit of transfer. Sawyer (1989) deduced that the average units of LGT are DNA patches of 100–200 bp, suggesting that transfer of parts of genes is common in nature. When a gene is broken down to small pieces, the complexity will be greatly simplified. For example, a gene fragment encoding a domain, such as a hairpin of an rRNA molecule, may replace a corresponding region of the host gene via a single recombinational event. A new hairpin in an rRNA molecule may not be detrimental to the organism as long as the local secondary structure remains identical or similar (Gutell *et al.*, 1994; Van de Peer *et al.*, 1996; Asai *et al.*, 1999). This process may occur repeatedly, involving different parts of a gene, leading to gradual changes of the nucleotide sequence. This 'simplified complexity hypothesis' predicts a mosaic nature for rRNA genes (as well as for other genes). Unlike the replacement of an entire gene, the step-by-step segmental replacement gradually obscures the evolutionary history documented in the nucleotide sequence. Although there have been previous reports of mosaic genes arising from local gene exchange between related bacterial species (Smith *et al.*, 1991; Groisman *et al.*, 1992), it is not clear whether and how frequently rRNA genes are affected by the lateral transfer of gene segments.

In this study, we investigated the distribution of nucleotide variations in the alignments of the 16S rDNA sequences of actinomycete species belonging to several well-defined genera. We frequently found the occurrence of two or three types of drastically different sequences in short and rather conserved regions of rRNA genes among species of the same genus. The nucleotide substitutions involve three to five base pairs of compensating covariations and are found within the stems of individual hairpins in the rRNA molecule. These observations strongly support our 'simplified complexity hypothesis' to explain how LGT may affect the highly conserved genes. We also demonstrate how this segmental gene transfer may lead to confusion in rDNA sequence-based phylogenetic analysis.

METHODS

Organisms and culture conditions. The actinomycete strains used in this study were purchased from IFO (Institute for Fermentation, Osaka, Japan) and JCM (Japan Collection of Microorganisms, Wako, Japan). The cells were grown according to the suppliers' instructions.

Preparation of genomic DNAs. The genomic DNA of the actinomycetes was prepared as previously described (Wang *et al.*, 1996).

PCR amplification, cloning and sequencing. The 5' one-third of the 23S rRNA genes was amplified by PCR using a pair of primers, one targeting a conserved region at the end of the 16S rRNA gene and the other a conserved block within the 23S rRNA gene. The PCR product also includes the spacer region. The sequences of the two oligonucleotides are as follows: 5'-GGTTGGATCCACCTCCTT-3', corresponding to nt 1525–1542 of *E. coli* 16S rRNA (Brosius *et al.*, 1978); and 5'-ACCAGTGAGCTATTAGCG-3' (nt 1090–1107). After the cloning of the PCR products, the M13 forward and reverse universal primers were used for sequencing the ends of each clone. The internal regions were sequenced in both orientations by using the following two sets of oligonucleotide primers targeting two conserved sequences within 23S rDNA. The first set of primers, targeting nt 45–60 of *E. coli* 23S rRNA gene, are 23S-40f (5'-CCGATGAAGGACGTGGGA-3') and 23S-40r (5'-TCCCACGTCCTTCATCGG-3'); and the second set of primers, targeting nt 456–472, are 23S-460f (5'-CCTTTCCTCACGGTACT-3') and 23S-460r (5'-AGTACCGTGAGGGAAAGG-3').

Sequence alignment and phylogenetic analysis. Multiple sequence alignment of the rRNA gene sequences was carried out by the CLUSTAL method of the DNASTAR program (DNASTAR, Inc.) and verified according to the consensus secondary structure model (Gutell *et al.*, 1994). Phylogenetic trees were constructed by using both a distance method (Saitou & Nei, 1987) and a parsimony method (Swofford & Begle, 1993). The two methods generated very similar trees and only the neighbour-joining trees are shown in this paper. The confidence level of phylogenetic tree topology was

Fig. 1. Localized sequence variations in rRNA genes of closely related species. (a) Nucleotide sequence alignment of rRNA gene segments from actinomycete species of different genera. The varied bases are boxed. Some sequences from species of distant genera are included in each alignment to show potential sources for LGT. (b) Positions and secondary structure of the gene segments shown in (a). The consensus secondary structure was drawn according to previous publications (Gutell *et al.*, 1994; Van de Peer *et al.*, 1996) and the sequence of *E. coli* 16S rRNA (Brosius *et al.*, 1978).

evaluated by the bootstrap method (Felsenstein, 1985). The softwares for the tree construction and bootstrap analysis are contained in the CLUSTAL V (Higgins *et al.*, 1992) and the PAUP (Swofford & Begle, 1993) phylogenetic analysis software packages.

RESULTS

Occurrence of non-random base variations within the hairpin regions of rRNA

The simplified complexity hypothesis predicts the existence in some rRNA genes of short regions containing abnormally high numbers of non-random base variations in comparison with the rRNA genes of closely related organisms. The location of such regions should correspond to individual hairpins in the rRNA molecule. It also predicts that sequences identical or almost identical to some of the variants can be found in species belonging to a distant taxon, the potential source of the introgressed DNA fragment. We started our search for such regions in 16S rDNAs from species belonging to the actinomycete genera *Actinomadura*, *Nonomuraea*, *Micromonospora*, *Streptosporangium*, *Dactylosporangium* and *Gordona*, because there is a sufficiently large

number of species in each of these genera and the complete 16S rDNA sequences of most of the valid species have been determined (Chun & Goodfellow, 1995; Koch *et al.*, 1996; Wang *et al.*, 1996; Ward-Rainey *et al.*, 1996; Rheims *et al.*, 1998; Zhang *et al.*, 1998). A 16S rDNA sequence alignment was generated for each genus for visual inspection. We focused our attention on regions that are normally quite conserved within a genus. In nearly all the sequence alignments short regions (20–70 bases) were found where the sequences could be divided into two or three distinct types differing at 6 to 10 positions (Fig. 1a). The pattern of base variations in these regions distinguished them from the random mutations occurring in the known highly variable regions. First, the base variations occurred in a short gene segment outside the known highly variable regions. Second, nearly all of the base variations are compensating covariations of paired bases corresponding to the stem of a hairpin in the 16S rRNA molecule (Fig. 1b). Third, besides the two or three distinct types of sequences, there are no other types of base changes in these regions that may represent the transitional states between the distinct types. For example, *Actinomadura* species have two types of sequences in an approximately

Fig. 2. Neighbour-joining trees of actinomycete species belonging to the family Streptosporangiaceae. The numbers at nodes are bootstrap values based on 1000 resamplings. The bars at the top represent one inferred substitution per 100 nt. Both trees were rooted using *Thermobispora bispora* sequence as outgroup. The abbreviations for the genus names are: *S.*, *Streptosporangium*; *N.*, *Nonomuraea*; *M.*, *Microbispora*; *Mt.*, *Microtetraspora*. The 16S rRNA sequences were determined previously (Wang *et al.*, 1996; Zhang *et al.*, 1998). The GenBank accession numbers for the 23S rRNA sequences determined in this study are AF192136–AF192150.

Fig. 3. The 16S rRNA genes of some *Nonomuraea* species contain segments matching genes from species belonging to other genera of the same family. The sequences representing all or a majority of the species of a genus are shown at the top. The sequences of some *Nonomuraea* species are given at the bottom. The varied base pairs are boxed. The lines with arrows at the ends connect identical sequence motifs.

30 bp segment corresponding to nt 1301–1335 of *E. coli* rRNA (Fig. 1a). The two sequences differ at three pairs of positions, but no other sequence exists within the genus that has other base variations, indicating the insusceptibility of this region to random base changes. To determine whether sequences identical or highly similar to the different sequence variants are present in species of other taxa, we searched the database using the BLAST program and readily found such sequences (examples are shown below). In summary, the discovery of and all the properties displayed by the localized sequence variations in rRNA genes fulfil the predictions of the ‘simplified complexity hypothesis’.

Comparison of the rDNA sequences of the members of different genera of the family Streptosporangiaceae

The family Streptosporangiaceae encompasses four major genera: *Streptosporangium*, *Microtetraspora*, *Microbispora* and *Nonomuraea*. Extensive 16S rDNA sequence-based phylogenetic analyses have been carried out previously, showing the phylogenetic coherence of each genus of the family except for two *Strepto-*

sporangium species, *S. corrugatum* and *S. claviforme* (Wang *et al.*, 1996; Ward-Rainey *et al.*, 1996; Zhang *et al.*, 1998). The result of a partial 23S rDNA (~1200 bases, nt 1–1108 of *E. coli* numbering) sequence analysis conducted in this study largely confirms the phylogenetic integrity of each genus, but demonstrates significant discrepancy at the subgenus level. For example, Fig. 2 shows that the members of *Nonomuraea* form a very stable clade in both 16S and 23S rRNA trees. In the 16S rDNA tree, *N. pusilla* and *N. ferruginea* form one subclade with the deepest branch; and *N. roseola*, *N. africana* and *N. recticatena* form another subclade separated from the third containing the rest of the *Nonomuraea* species. However, drastically different species relationships are reconstructed in the 23S rRNA tree. To seek possible explanations for the observed discrepancy, we visually scanned the sequence alignments and found that the *Nonomuraea* species can be grouped by the sharing of covaried nucleotide substitutions involving 10 base pairs located in two stems within a short region (nt 587–753) in the 16S rRNA sequences. These base variations occur outside of the highly variable regions and their numbers are comparable to or even higher than those observed in

Fig. 4. Exclusion of the short region of sequence variations in the construction of phylogenetic tree affects the cluster pattern of the *Nonomuraea* species. The 16S rDNA tree shown in Fig. 2 was reconstructed. In tree (a), nt 668–672 and 734–738 (see Fig. 3) were excluded from all the sequences; in tree (b), nt 612–614 and 624–628 were excluded from all the sequences. The trees were constructed by using the neighbour-joining method.

these regions. The occurrence of a large number of non-random base variations in a rather stable region of the gene is most likely the consequence of lateral transfer of short segments of rRNA genes. Supporting this hypothesis, we found identical sequences matching each variation in species of other genera in the same family. These results are summarized in Fig. 3. Such a gene segment replacement is not expected to have much effect on the local and the overall secondary or tertiary structure of the rRNA and the function of ribosomes. According to the nucleotide variability map of 16S rRNA (Van de Peer *et al.*, 1996), this region is highly conserved in secondary structure but variable in nucleotide sequences characterized largely by co-variation of paired bases in the helices, indicating the importance of the secondary structure of this region for ribosomal function.

To see whether the exclusion of the short stretches of sequences in the construction of phylogenetic trees affects the cluster patterns of the *Nonomuraea* species, two 16S rDNA trees were reconstructed using all the sequences represented in the 16S rDNA tree shown Fig. 2. In the first tree (Fig. 4a), the 10 nucleotides nt 668–672 and 734–738 were excluded from all the sequences. The overall topology of this new tree is very similar to that of the tree shown in Fig. 2, except that the species cluster pattern within the *Nonomuraea* clade is significantly different. Although *N. roseola*, *N. recticatena* and *N. africana* still aggregate tightly, they are now placed among other *Nonomuraea* species. In the second tree (Fig. 4b), where nt 612–616 and 624–628 were excluded, the *N. pusilla* and *N. ferruginea* clade and the *N. roseola*, *N. recticatena* and *N. africana* clade switched positions in comparison with the tree shown in Fig. 2. The use of the maximum-parsimony method on the same sets of sequences also showed a similar change of

cluster patterns of *Nonomuraea* species between different trees (trees not shown). Despite the fact that some species clusters are not supported by sufficiently high bootstrap values, this result nevertheless shows that the occurrence of such short stretches of varied nucleotide sequences will have a significant effect on the topology of the phylogenetic trees.

We and others have previously reported the presence of distinct types of 16S rRNA genes in an organism. Thus it is necessary to ask whether the distinct types of sequence variants found in the 16S rRNA genes of different *Nonomuraea* species are present in one organism. We used oligonucleotides specific for the different sequence variants in PCR amplifications from the genomic DNA of all the *Nonomuraea* species. In all the PCR reactions only one type of sequence variant was detected in each organism (Fig. 5). In addition, the approximately 1000 bp 16S rDNA fragments PCR-amplified by the pair of universal primers (PU5' and PU3'; Fig. 5) were cloned and five independent clones from each species picked for sequence analysis. Again, all the sequences from each species are identical except those from *N. pusilla*. Three sequences of *N. pusilla* were found differing between each pair at one or two positions (nt 1248 A or G, nt 1267 T or C, and nt 1296 G or C; *E. coli* numbering), which are outside of the regions of concern (alignment not shown).

What causes the contradictory phylogenetic positions of *S. corrugatum* and *S. claviforme* in the 16S and 23S rRNA gene trees?

An obvious discrepancy between the 16S and 23S rRNA trees shown in Fig. 2 is the positions of two *Streptosporangium* species, *S. corrugatum* and *S. claviforme*. Independent of the outgroups used and whether the

Fig. 5. PCR reactions using specific oligonucleotides detect the presence of a single type of the sequence variant in each *Nonomuraea* species. The organisms examined are indicated at the top. The oligonucleotide primers used were: (a) two universal primers, PU5' (5'-GGGCGTAAAGAGCTCGTAGG-3', nt 567–585, *E. coli* numbering) and PU3' (5'-TACGGCTA-CCTTGTTACGACTT-3', nt 1513–1492); (b) PNI5' (5'-TAGGT/CA-GGGGCAAGT-3', nt 658–672) and PU3', specific for type I variant (see Fig. 1b); (c) PNII5' (5'-ACCGGTGGCGAAGGCGTTCT-3', nt 718–738) and PU3', specific for type II (Fig. 1b); (d) PNIII5' (5'-TGCCGTGAAAGCT/CCGCA-3', nt 600–616) and PU3', specific for all *Nonomuraea* species except *N. pusilla*; and (e) PNIV5' (5'-AAAGCTTAGGGCTTAACCTA-3', nt 608–628) and PU3' specific for *N. pusilla* (see Fig. 4).

distance (neighbour-joining) or the parsimony (PAUP) method was employed for the tree construction, the two species have been consistently shown to aggregate with the *Microtetraspora*/*Microbispora* clade in the 16S rRNA tree and with *Streptosporangium* clade in the 23S rRNA tree. The phylogenetic associations of the two species shown in both trees are supported by high bootstrap values (877 and 878). Since multiple entries of nearly identical sequences from the same organisms are available in databases, the possibility of artefact sequences is excluded. Then, why do 16S and 23S rRNA genes, which are thought to be under the same level of functional constraints, give drastically contradictory organismal phylogenies? An inspection of the sequence alignments revealed that the 16S rDNA sequences of *S. corrugatum* and *S. claviforme* can be roughly divided

into two halves, with the 5' half carrying nearly all the signatures characterizing *Microbispora*/*Microtetraspora* and the 3' half containing nucleotides specific to most *Streptosporangium* species (Fig. 6a). When two new trees were constructed by using the two halves of the 16S rRNA sequences separately, the two species were placed next to the *Microbispora*/*Microtetraspora* clade in the 5' tree, in contrast to a placement next to the *Streptosporangium* clade in the 3' tree (Fig. 6b). The position of the two *Streptosporangium* species in the 5' tree is supported by a high bootstrap value (990). The bootstrap value for the two species to cluster with *Streptosporangium* species in the 3' tree is lower (682), probably due to the presence of far fewer genus-specific nucleotides in this half of the sequence. Nevertheless, the relationship shown in the 3' tree is consistent with the result of the 23S rDNA analysis and also with the unifying morphological feature, namely the growth of sporangia by all *Streptosporangium* species. No species in other genera in the family *Streptosporangiaceae* grows sporangia. Why do the two halves of the 16S rRNA gene record conflicting evolutionary histories? The most plausible answer to this question is that the 16S rRNA genes of *S. corrugatum* and *S. claviforme* are mosaics of segments originating from different organisms. This could have happened to the two *Streptosporangium* species simply by replacing a region of the endogenous gene with a corresponding gene segment from a *Microbispora* species via a double crossover or gene conversion (Hillis *et al.*, 1990; Syvanen, 1994). In fact, nearly all the signature nucleotides specific to *Microbispora* in the 16S rRNA genes of *S. corrugatum* and *S. claviforme* are localized in the region between nt 600 and 750 encoding two connected hairpins in the rRNA (see Figs 1b and 3 for location).

DISCUSSION

The concept of using gene or protein sequences to reconstruct the evolutionary history of life is facing great challenges from the discovery of extensive LGT among organisms. It is now indisputable that all organisms are evolutionary chimeras whose genomes contain both vertically inherited and laterally transferred genes, accounting for the conflicting phylogenies derived from analyses of different gene sets. Many biologists believe that it is still possible to trace, albeit roughly, the evolutionary history of life through comparative analysis of gene sequences. The argument, represented by the complexity hypothesis proposed by Jain *et al.* (1999), is that genes are not equally susceptible to LGT; genes whose products must interact with multiple components in complex cellular systems are highly unlikely to be laterally transferred. The rRNA gene is perhaps the best example of this category of genes. Thus, a large part, if not the entire, evolutionary history of life may still be inferred by comparing these genes (Woese, 1998).

We initiated this study by hypothesizing that LGT might affect complex genes, such as rRNA genes, through

Fig. 6. A mosaic structure of the 16S rRNA genes of *S. corrugatum* and *S. claviforme*. (a) Alignment of regions containing genus-specific nucleotide signatures. Sequences representing *Microbispora* species (M), *S. corrugatum* and *S. claviforme* (Sc) and other *Streptosporangium* species (S) are aligned. Signatures shared between M and Sc, or between Sc and S, are boxed. (b) Neighbour-joining trees constructed using the two halves of 16S rRNA sequences. Only *Streptosporangium* and *Microbispora* species are included, to show the contradictory positions of *S. corrugatum* and *S. claviforme* in the two trees.

gradual substitution of gene segments encoding individual domains. Though the overall structure and its interaction with other parts of the translation machinery are complex, an rRNA molecule consists of many hairpin structures, each of which may interact with only one or a few other components of the translation apparatus. In addition, it is well known that the secondary but not the primary structure of many hairpins in rRNA is essential for function (Gutell *et al.*, 1994; Van de Peer *et al.*, 1996). These two properties of rRNA molecules underscore the high possibility of the exchange of individual hairpins without damaging effect on the proper function of a ribosome. This view is supported by reports that an organism may have distinct types of rRNA genes with high levels of base variations throughout the entire gene (Gunderson *et al.*, 1987; Mylvaganam & Dennis, 1992; Carranza *et al.*, 1996; Wang *et al.*, 1997; Yap *et al.*, 1999). Niebel *et al.* (1987) demonstrated that the rRNA cistrons from *Proteus vulgaris* were expressed and the products correctly processed and assembled into ribosomes when transformed into *E. coli*. Recently, Asai *et al.* (1999)

reported a complete exchange of rRNA genes between different bacterial species and replacement of a 23S rRNA gene segment of *E. coli* by the corresponding region of yeast. If a transgressed rRNA gene or gene segment brings to the recipient a beneficial property, such as antibiotic resistance (Green *et al.*, 1997; Mankin, 1997), it will be kept and may eventually replace the endogenous gene(s) via gene conversion (Hillis *et al.*, 1990). High levels of localized intergenic base variations corresponding to the stem of a hairpin have also been reported (Ueda *et al.*, 1999; Yap & Wang, 1999). Ueda *et al.* (1999) found two types of base variations in the highly variable α region of the 16S rRNA gene and suggested that both random mutation and LGT contribute to rRNA gene heterogeneity of a bacterium.

In this study, we found high levels of base variations localized to the stem of hairpins of rRNA molecules that seldom experience random mutations between species of the same genus or closely related genera. We propose that the LGT of gene segments is the most plausible explanation of this phenomenon. We observed this

phenomenon in the 16S rRNA genes of almost all genera we examined, indicating that such events probably happen with considerable frequency in nature. Although it is difficult, if not impossible, to determine the donor of a transferred DNA fragment, identical or nearly identical sequences are always found in species often belonging to distantly related taxa. Introduction of a piece of foreign DNA will certainly corrupt the evolutionary history recorded in the affected gene, though the extent of damage will depend on the size and the number of changed bases of the region involved. In most cases we have investigated, the presence of sequence variations in one hairpin does not affect the stable aggregation in one clade of all the species of a genus, reflecting the overall evolutionary stability of 16S rRNA genes. At subgenus level, the pattern of species clustering may be affected by the presence of these localized sequence variations, as shown in the analysis of the 16S rDNA of *Nonomuraea* species. However, it is not clear whether the sharing of a short DNA sequence is a result of common ancestry, or whether the sequences were introduced in independent events. The cluster pattern of *Nonomuraea* species shown in the 16S rDNA trees, constructed by including or excluding the short regions of sequence variations, was not reproducible in the 23S rRNA tree. In the study of *Streptosporangium* species, we speculate that the lateral transfer of a larger DNA fragment encoding two adjacent hairpins was the likely cause of the confusion at genus level of the positions of *S. corrugatum* and *S. claviforme* in the 16S rRNA tree. Regardless of whether the 16S or the 23S tree is correct, the dramatically different positions of the two species in the two trees are unlikely to be the result of random mutational events.

Conclusion

Complementing the complexity hypothesis, the simplified complexity hypothesis suggests that due to the overall functional constraints, the lateral transfer of parts of genes is likely to be the main mode of LGT occurring to the genes encoding components of complex systems. This mode of LGT results in gradual corruption of the evolutionary history written in the nucleotide sequences. The rate of sequence corruption of a gene should be a function of the complexity of interactions with other cellular components by the gene product. Even highly conserved genes like rRNA genes may only roughly record part of the evolutionary history of life, though how far back is not known. There is probably an increasingly high restriction for successful transfer of gene segments with widening evolutionary distance, which may explain why the three-domain separation of life largely holds when proteins of the information-processing machinery are used to derive phylogeny (Woese, 1998; Jain *et al.*, 1999). The results of this study caution that in bacterial taxonomy the effect of LGT on the result of sequence-based phylogenetic analysis should be considered. Analysis of more than one gene set together with the examination of major phenotypic characteristics should be practised in future. Given that LGT is well established, it is now important to identify

the principles that govern gene distribution patterns across prokaryotic genomes to determine the values of genes of different functional groups in bacterial taxonomy.

ACKNOWLEDGEMENTS

This work was supported by the Institute of Molecular and Cell Biology. We thank Dr W. H. Yap for critically reading the manuscript and S. M. Ali for help with PCR amplification and cloning of rRNA genes.

REFERENCES

- Asai, T., Zaporozets, D., Squires, C. & Squires, C. L. (1999). An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *Proc Natl Acad Sci USA* **96**, 1971–1976.
- Brosius, J., Palmer, J. J., Kennedy, J. P. & Noller, H. F. (1978). Complete nucleotide sequence of a 16S ribosomal gene from *Escherichia coli*. *Proc Natl Acad Sci USA* **75**, 4801–4805.
- Carranza, S., Giribet, G., Riberat, C., Baguna, J. & Riutort, M. (1996). Evidence that two types of 18S rDNA coexist in the genome of *Dugesia (Schmidtea) mediterranea* (Platyhelminthes, Turbellaria, Tricladida). *Mol Biol Evol* **13**, 824–832.
- Chun, J. & Goodfellow, M. (1995). A phylogenetic analysis of the genus *Nocardia* with 16S rRNA gene sequences. *Int J Syst Bacteriol* **45**, 240–245.
- Doolittle, W. F. (1999). Phylogenetic classification and universal tree. *Science* **284**, 2124–2128.
- Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **46**, 159–173.
- Green, R., Samaha, R. R. & Noller, H. F. (1997). Mutations at nucleotides G2251 and U2585 of 23S rRNA perturb the peptidyl transferase center of the ribosome. *J Mol Biol* **266**, 40–50.
- Groisman, E. A., Saier, M. H., Jr & Ochman, H. (1992). Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the *Salmonella* genome. *EMBO J* **11**, 1309–1316.
- Gunderson, J. H., Sogin, M. L., Wollet, G., Hollingdale, M., de la Cruz, V. F., Waters, A. P. & McCutchan, T. F. (1987). Structurally distinct, stage-specific ribosomes occur in *Plasmodium*. *Science* **238**, 933–937.
- Gutell, R. G., Larsen, N. & Woese, C. R. (1994). Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol Rev* **58**, 10–26.
- Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Comput Appl Biosci* **8**, 189–191.
- Hillis, D. M., Moritz, C., Porter, C. A. & Baker, R. J. (1990). Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science* **251**, 308–310.
- Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**, 3801–3806.
- Koch, C., Kroppenstedt, R. M., Rainey, F. A. & Stackebrandt, E. (1996). 16S ribosomal DNA analysis of the genera *Micromonospora*, *Actinoplanes*, *Catellatospora*, *Catenuloplanes*, *Couchioplanes*, *Dactylosporangium*, and *Pilimelia* and emendation of the family *Micromonosporaceae*. *Int J Syst Bacteriol* **46**, 765–768.

- Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1999). Comparison of archaeal and bacterial genomes: computer analysis of protein sequences suggests a chimeric origin for the archaea. *Mol Microbiol* **25**, 619–637.
- Lawrence, J. G. & Ochman, H. (1997). Amelioration of bacterial genome: rates of change and exchange. *J Mol Evol* **44**, 383–397.
- Mankin, A. S. (1997). Pactamycin resistance mutations in functional sites of 16S rRNA. *J Mol Biol* **274**, 8–15.
- Mylvaganam, S. & Dennis, P. P. (1992). Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaeobacterium *Halobacterium marismortui*. *Genetics* **130**, 399–410.
- Nelson, K. E., Clayton, R. A., Gill, S. R. & 22 other authors (1999). Evidence for lateral gene transfer between *Archaea* and *Bacteria* from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329.
- Niebel, H., Dorsch, M. & Stackebrandt, E. (1987). Cloning and expression in *Escherichia coli* of *Proteus vulgaris* genes for 16S ribosomal RNA. *J Gen Microbiol* **133**, 2401–2409.
- Rheims, H., Schumann, P., Rohde, M. & Stackebrandt, E. (1998). *Verrucosipora gifhornensis* gen. nov., sp. nov., a new member of the actinobacterial family *Micromonosporaceae*. *Int J Syst Bacteriol* **48**, 1119–1127.
- Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* **95**, 6239–6244.
- Saitou, N. & Nei, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.
- Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol Biol Evol* **6**, 526–538.
- Smith, J. M., Dowson, C. G. & Spratt, B. G. (1991). Localized sex in bacteria. *Nature* **349**, 29–31.
- Swofford, D. L. & Begle, D. P. (1993). *Phylogenetic Analysis Using Parsimony (Version 3.1), User's Manual*. Champaign, IL: Smithsonian Institution Laboratory of Molecular Systematics.
- Syvanen, M. (1994). Horizontal gene transfer: evidence and possible consequences. *Annu Rev Genet* **28**, 237–261.
- Ueda, K., Seki, T., Kudo, T., Yoshida, T. & Kataoka, M. (1999). Two distinct mechanisms cause heterogeneity of 16S rRNA. *J Bacteriol* **181**, 78–82.
- Van de Peer, Y., Chapelle, S. & Wachter, R. D. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res* **24**, 3381–3391.
- Wang, Y., Zhang, Z. S. & Ruan, J. S. (1996). A proposal to transfer *Microbispora bispora* (Lechevalier 1965) to a new genus, *Thermobispora* gen. nov., as *Thermobispora bispora* comb. nov. *Int J Syst Bacteriol* **46**, 933–938.
- Wang, Y., Zhang, Z. S. & Ramanan, N. (1997). The actinomycete *Thermobispora bispora* contains two distinct types of transcriptionally active 16S rRNA genes. *J Bacteriol* **179**, 3270–3276.
- Ward-Rainey, N., Rainey, F. A. & Stackebrandt, E. (1996). The phylogenetic structure of the genus *Streptosporangium*. *Syst Appl Microbiol* **19**, 50–55.
- Woese, C. R. (1987). Bacterial evolution. *Microbial Rev* **51**, 221–271.
- Woese, C. R. (1998). The universal ancestor. *Proc Natl Acad Sci USA* **95**, 6854–6859.
- Woese, C. R., Kandler, O. & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains *Archaea*, *Bacteria* and *Eucarya*. *Proc Natl Acad Sci USA* **87**, 4576–4579.
- Yap, W. H. & Wang, Y. (1999). Molecular cloning and comparative sequence analyses of rRNA operons in *Streptomyces nodosus* ATCC 14899. *Gene* **232**, 77–85.
- Yap, W. H., Zhang, Z. S. & Wang, Y. (1999). Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol* **181**, 5201–5209.
- Zhang, Z. S., Wang, Y. & Ruan, J. S. (1998). Reclassification of *Thermomonospora* and *Microtetraspora*. *Int J Syst Bacteriol* **48**, 411–422.

Received 8 February 2000; revised 22 July 2000; accepted 7 August 2000.