Genome Analysis

Eubacterial phylogeny based on translational apparatus proteins

Céline Brochier, Eric Bapteste, David Moreira and Hervé Philippe

Lateral gene transfers are frequent among prokaryotes, although their detection remains difficult. If all genes are equally affected, this questions the very existence of an organismal phylogeny. The complexity hypothesis postulates the existence of a core of genes (those involved in numerous interactions) that are unaffected by transfers. To test the hypothesis, we studied all the proteins involved in translation from 45 eubacterial taxa, and developed a new phylogenetic method to detect transfers. Few of the genes studied show evidence for transfer. The phylogeny based on the genes devoid of transfer is very consistent with the ribosomal RNA tree, suggesting that an eubacterial phylogeny does exist.

The completion of many genome sequence projects has revealed the fundamental importance of lateral gene transfers (LGTs) in prokaryotic evolution [1–3]. Even the very existence of an organismal phylogeny has been questioned. However, phylogenetic trees based on gene content [4] are remarkably similar to the phylogeny based on ribosomal RNA (rRNA), which is currently the most widely accepted [5]. This indicates that LGT, although frequent, has not completely erased the phylogenetic signal.

It has been proposed that genes encoding proteins involved in multiple interactions are less likely to be transferred (the complexity hypothesis) and therefore represent a conserved core of genes that defines the prokaryotic phylogeny [6]. Here, we have studied all proteins classified as involved in translation and inferred a bacterial phylogeny of 45 species. Among these proteins, we selected 57 proteins that were present in at least 42 species and that have no (or very few) duplicated copies. We concatenated the sequences of the 57 genes into a large fusion (~ 9000 amino acid positions). The phylogeny based on this fusion is very similar to that inferred from rRNA and gene content. Detailed analysis revealed that 13 out of the 57 gene phylogenies were INCONGRUENT (see Glossary) with the phylogeny based on the fusion of the 57 genes, either due to methodological tree-reconstruction problems or to a few recent LGTs. A true organismal phylogeny for Bacteria seems to exist, which could be fully resolved by the analysis of a core group of very rarely transferred genes.

Phylogenetic analysis of a large protein fusion

For our analysis, we retrieved from the public databanks and from ongoing

Glossary

Congruence and incongruence: Congruence is the agreement between phylogenies obtained using different datasets or different reconstruction methods. Trees are perfectly congruent if they display the same topology; that is, they reflect the same evolutionary history. By contrast, incongruent trees show conflicting robust nodes, which could be due to different evolutionary histories (e.g. lateral gene transfers) or tree reconstruction problems.

Γ law: Traditional models of sequence evolution assume that all positions in the sequences are equally likely to undergo a substitution, which reduces the complexity of these models. However, in reality, positions in sequences are more or less 'free' to vary; that is, they have different probabilities of undergoing substitutions. This limits the biological realism of traditional models and their efficiency for phylogenetic reconstruction. The variation of substitution rates is commonly approximated using a gamma distribution, also known as a Γ law, which has a shape parameter α that specifies the range of rate variation [a]. Small α values result in an L-shaped distribution with extreme variation of rates (most sites are invariable, but a few have very high substitution rates). As α gets larger, the range of variation diminishes, until α approaches infinity and all sites have the same substitution rate.

HKY model: The Hasegawa, Kishino and Yano [b] model of sequence evolution is a merger of the Felsenstein [c] and the Kimura two-parameter models [d], which allows transitions and transversions to occur at different rates and base frequencies to vary during the course of evolution, respectively.

Jack-knife analysis: A statistical method to evaluate the robustness of an inference. It is based on the construction of random sub-samples of the original alignment by taking a fraction of the positions without replacement (in contrast to the bootstrap method, which allows replacement). Usually, trees are reconstructed with the random sub-samples and the robustness of each node is estimated as the number of its occurrences among these trees [e].

Log–Det method: A method to evaluate evolutionary distances that are consistent for sequences with different nucleotide or amino acid composition [f]. This approach is required because other methods tend to group sequences on the basis of their composition, irrespective of their evolutionary history.

Kishino–Hasegawa test: A test used for the estimation of incompatibility between alternative tree topologies with the same taxonomic sampling but obtained using different datasets [g]. Two tree topologies are significantly different if the differences of their likelihood values (expressed as the $\Delta \ln L$, where L is the likelihood) is larger than 1.96 standard error in the estimation of likelihood. For a recent criticism of this test see Ref. [h].

Principal component analysis (PCA): This involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Principal components are obtained by projecting the multivariate data vectors on the space spanned by the eigen vectors.

References

a Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–370

b Hasegawa, M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174

c Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376

d Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. U. S. A.* 78, 454–458

e Lapointe, F.J. *et al.* (1994) Jackknifing of weighted trees: validation of phylogenies reconstructed from distance matrices. *Mol. Phylogenet. Evol.* 3, 256–267

f Lockhart, P. *et al.* (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11, 605–612

g Kishino, H. and Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* 29, 170–179

h Goldman, N. *et al.* (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49, 652–670

http://tig.trends.com 0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(01)02522-7

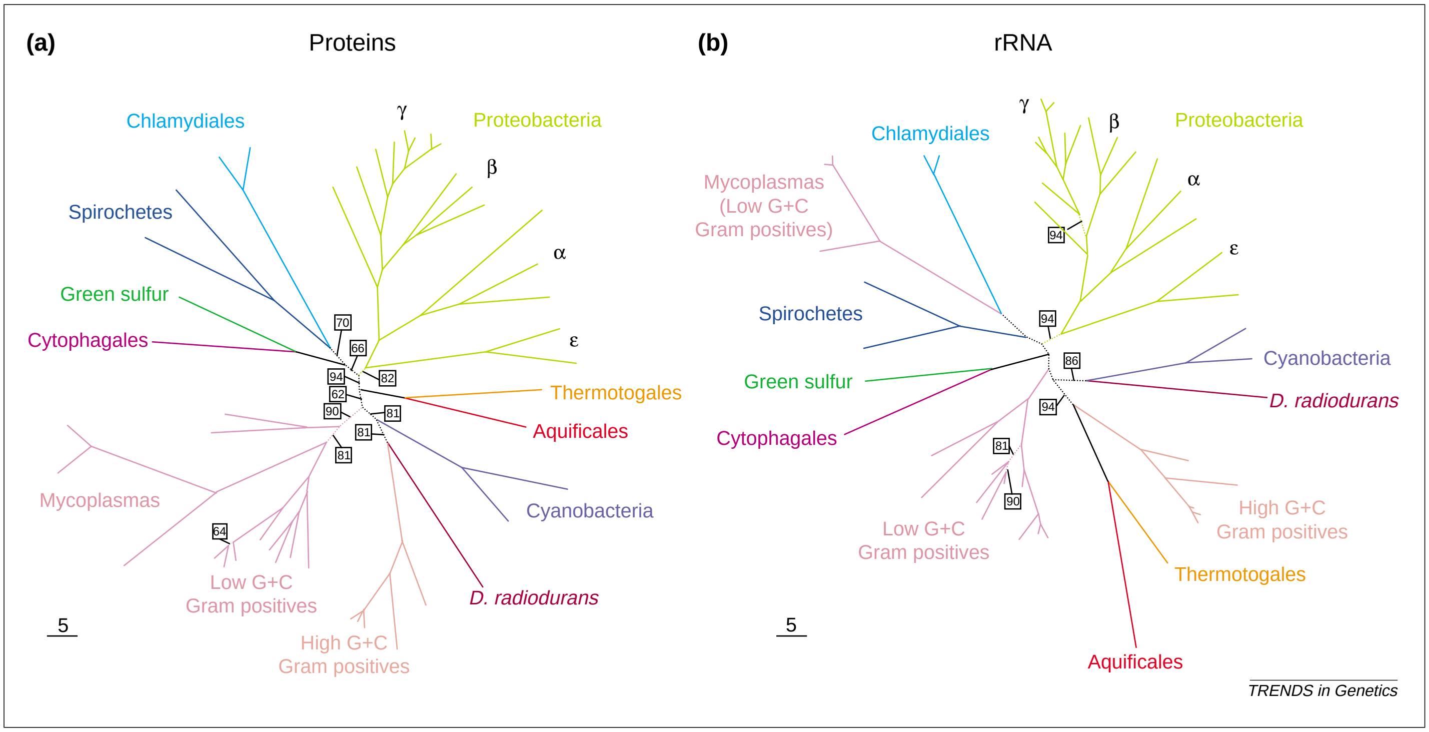

Fig. 1. Unrooted maximum-likelihood (ML) phylogenetic trees constructed for the protein fusion P1 (a) and the rRNA fusion R (b). The currently accepted taxonomic groups are indicated with different colours. Numbers at nodes are bootstrap proportions. Note that the use of almost twice as many positions for fusion P1 (8857) than for fusion R (3704) does not significantly increase the resolution of intergroup relationships, which might be reflecting a strong level of mutational saturation in sequences and/or a rapid diversification (radiation) at the origin of these bacterial phyla. The scale bar corresponds to five changes per 100 positions for a unit branch length. The trees with all the species names are available upon request. The fusions and alignments are available at http://sorex.snv.jussieu.fr/translation/translation.html.

genome projects sequences homologous to all *Escherichia coli* proteins classified as involved in translation in the Cluster of Orthologous Genes (COG) database [7], as well as the 16S and 23S rRNAs. We aligned 76 proteins from 45 bacterial species, having eliminated any proteins that are present only in a restricted sample of phyla (see http://sorex.snv.jussieu.fr/translation/translation.html). In addition, as a sample of transferred genes, we used the tRNA synthetases (tRS), most of which are known to have undergone numerous LGTs (perhaps related to antibiotic resistance [8,9]). The 76 genes were analysed individually, and 19 of them were excluded from further analyses because they were: (1) difficult to align reliably, (2) present in less than 42 of the 45 species, or (3) have more than one copy for certain phyla (indicating possible ancient duplications and losses, and/or LGTs). The remaining 57 genes, after elimination of ambiguously aligned regions (alignments available on our website), were concatenated for the 45 bacterial species into a large fusion of 8857 amino acids (fusion P1). Most of the best-known bacterial phyla were represented, of which we had a broad taxonomic sampling for Proteobacteria and Gram-positive bacteria. We do not use Archaea as an outgroup, because this would seriously reduce the number of aligned positions and the long archaeal branch could generate a tree reconstruction artefact [e.g. 10,11].

All distance (NJ), maximum parsimony (MP) and maximum-likelihood (ML) phylogenetic methods retrieved very similar trees from this fusion. The monophyly of most accepted phyla (e.g. Proteobacteria, low-GC-content Gram-positives, high-GC-content Gram-positives, Spirochaeta and Cyanobacteria) was retrieved with high statistical support (Fig. 1a). The topology was almost CONGRUENT with that obtained for the same 45 species using a fusion of 16S and 23S rRNAs as phylogenetic marker (fusion R, 3704 nucleotide positions; Fig. 1b). Relationships within phyla were almost identical for both phylogenies (see website) and the most important difference was the position of mycoplasmas, which emerged far from the low-GC Gram-positive clade in the rRNA tree, whereas in the protein-fusion tree they emerged within low-GC Gram-positives with a high bootstrap support (bootstrap proportion [BP] of 90%). This confirmed previous reports suggesting that protein sequences reflect more accurately the phylogenetic position of mycoplasmas [12]. The resolution of the interphyla relationships was poor (BP < 90%), as frequently observed [4,5], despite the use of ~9000 amino acid positions (Fig. 1a).

One exception, both for the protein and rRNA fusions, was the high statistical support (BP of 100%) obtained for two interesting clades. The first grouped the hyperthermophiles *Aquifex aeolicus* and *Thermotoga maritima*, and the second, the green sulfur bacterium *Chlorobium tepidum* and the cytophagal *Porphyromonas gingivalis*. However, the proximity of *Aquifex* and *Thermotoga* was probably due to a bias in amino acid composition because of their thermophilic lifestyle [13]. The sisterhood of green sulfur bacteria and cytophagales, which has been observed previously and led to the proposal for a new bacterial kingdom [9], was more reliable because amino acid composition seemed unbiased. The grouping of high-GC Gram-positives with *Deinococcus radiodurans* (BP of 81%) is more surprising (Fig. 1a) because neither clear biological features nor genomic ones support this grouping [14]. However, their genomes are all GC-rich, which biases the

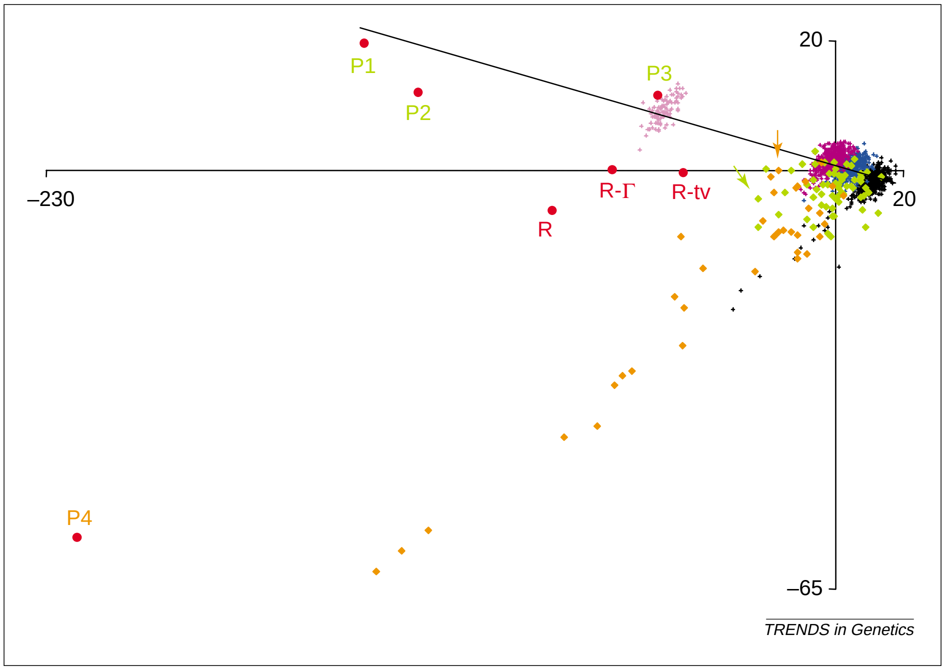

Fig. 2. Principal component analysis (PCA) of the likelihood values estimated for 375 tree topologies for different protein and rRNA datasets. The first (*x*) axis (83% of the variance) mainly corresponds to sequence size, and the second (*y*) axis (7% of the variance) to congruence among datasets. Orange, genes that were suspected to have undergone lateral gene transfer (LGT) and/or duplications; green, the 57 genes for which LGT was not suspected *a priori*; red, to protein (P1, P2, P3, and P4) and rRNA (R, R-Γ, and R-tv) fusions; crosses, random sub-samples of the P1 fusion of 50 (black), 150 (blue), 250 (purple), and 2000 (mauve) sites, respectively. The regression line for these random sub-sample points is shown. The distance to this regression line is indicative of congruence with fusion P1. A gene that was congruent despite *a priori* suspected LGT, Ala–tRS, is marked by an orange arrow, and a gene for which LGT was detected *a posteriori* because of its incongruence, EF-G, is marked by a green arrow.

amino acid composition [13]. As in the case of rRNA (Fig. 1b) and of certain proteins [15], *Deinococcus* also clustered with cyanobacteria in our gene fusion when we used the log-det method, which is less sensitive to amino acid composition bias [16] (see website).

A new tool to estimate phylogenetic congruence

We did not use the standard Kishino–Hasegawa test [17] to analyse the congruence between the individual 57 proteins, the fusion P1 and the rRNA fusion because of its recently described intrinsic flaws [18]. Instead, we have developed a new method that allows the simultaneous analysis of the congruence between many markers within a reasonable computing time (see website). Each dataset is described by its likelihood for 375 topologies as a sample of all the possible topologies, and the likelihood values for all datasets are then analysed by a principal component analysis (PCA). For our datasets, the first two axes of the PCA explained almost all the variance (83% and 7%, respectively). Figure 2, which displays these first two axes of our PCA, is therefore sufficient to describe our results adequately.

We studied the impact of stochastic effects through a jack-knife analysis. We generated 1600 random sub-samples of our large P1 alignment (8857 positions) with various sizes (50 positions, black; 150 positions, blue; 250 positions, purple; and 2000 positions, lilac, respectively, on Fig. 2). The random sub-sample dots clearly appear within four groups mainly distributed along the first (*x*) axis, which correspond to the four different sequence sizes used (the smallest on the right and the largest on the left). The regression line is computed for all the sub-samples and, as expected, the point corresponding to the fusion P1 (upper left in Fig. 2) is located near this line. This regression line is close to axis 1, which can be interpreted as mainly corresponding to sequence size, whereas axis 2 (*y*) mainly corresponds to incongruence between markers. The points corresponding to the sub-samples show a dispersion perpendicular to the regression line (i.e. level of incongruence) that is slightly inferior to that observed for real genes (except for a few 50-site sub-samples). This is not surprising, because real genes can be subjected to specific biases (e.g. fast evolutionary rates or biased amino acid compositions for some species) that do not affect to the random sub-samples, and thus they can be more incongruent owing to tree reconstruction artefacts.

We also concatenated the 44 proteins for which LGTs were not detected *a posteriori* (fusion P2, 6436 positions) and the 13 proteins for which LGTs were detected *a posteriori* (fusion P3, 2422 positions). Finally, the 39 proteins suspected *a priori* to have undergone LGTs (tRS and proteins with multiple copies for some phyla) were concatenated in the P4 fusion (11108 positions).

Possible causes of incongruence

Most of the 57 translation proteins for which extensive paralogy was not detected *a priori* (green diamonds, Fig. 2) form a compact group of points, suggesting that they are congruent (namely, they share the same evolutionary history even if the trees inferred from individual markers could be different from the one in Fig. 1a because of stochastic effects). By contrast, the 39 proteins for which LGT or paralogy were suspected *a priori* (orange diamonds, Fig. 2) appear very dispersed, which confirms the lack of congruence between them and the 57 others. Nevertheless, the two groups of points are not separated entirely. Points can overlap because:

(1) stochastic errors due to the small size of some genes (see above);
(2) lack of LGTs for some of the 39 proteins for which LGTs was *a priori* suspected;
(3) LGTs for some of the 57 other proteins;
(4) tree reconstruction artefacts.

The Ala–tRS (orange arrow, Fig. 2) and Cys–tRS are two good examples for the second case (see website). Ala–tRS seems devoid of LGT, at least within the current sample. Cys–tRS only shows LGTs from Bacteria to Archaea [9], which are not considered here, and a probable ancient duplication in high GC Gram-positives that does not bias phylogenetic inference. More interestingly, our method allowed the detection of LGTs. A striking example concerns the elongation factor G (EF-G, green arrow, Fig. 2). The analysis of an exhaustive EF-G dataset (103 sequences) reveals a very probable case of LGT between β- and γ-Proteobacteria (see website).

In addition to LGT, tree reconstruction artefacts can constitute a major problem, as is well illustrated by the comparison between the protein and rRNA fusions. When a simple model of sequence evolution is used (HKY MODEL, with equal substitution rate for all the positions [19]), the rRNA point (R in Fig. 2) appears incongruent with proteins (i.e. far from the regression line). However, when among-site rate variation is taken into account through a Γ LAW, rRNA (point R-Γ) becomes more congruent. A comparable effect is found when only transversions are taken into account (point R-tv). Transversions are less affected by compositional GC-bias [20], whereas a Γ law describes the evolution of sequences more accurately [21]. Therefore, the use of any of these approaches improves the inference of the rRNA phylogeny, and it is especially significant that both increase congruence with protein phylogeny.

Incongruence and lateral gene transfer

It has been suggested that combining datasets with different evolutionary histories could increase their explanatory power, and thus improve the phylogenetic inference (i.e. the signal but not the noise will increase) [22]. However, the concatenation of the 39 translational genes that seem *a priori* to have had different histories because of LGT events and duplications (point P4, Fig. 2) is highly incongruent with the rest of genes. This case is nevertheless extreme because of numerous LGTs, often between very distant taxa (e.g. the Val–tRS of *Rickettsia* is of archaeal origin). We have focused on the 13 genes without *a priori* LGT that were the more distant from the regression line (Fig. 2); that is, less congruent with the protein fusion P1 phylogeny. Surprisingly, a detailed analysis of their phylogeny strongly suggests LGTs in only five cases (EF-G, FMT, PTH, KsgA and Rpl7). For the remaining eight genes (IF2, Rpl5, Rpl9, Rpl15, Rpl18, Rpl24, Rpl29 and Rps5), the incongruence seems to be due to their short sequence size, gene specific bias or LGTs too ancient to be easily identified. Interestingly, the fusion of these 13 genes (2422 positions; point P3, Fig. 2) is congruent with the complete fusion, because P3 is located within the group of the 2000-position random sub-samples. Therefore, it is not reasonable to concatenate genes with different evolutionary histories unless the discrepancies are small [23] (compare P3 and P4, Fig. 2). The fusion based on the 44 congruent genes (6436 positions; point P2, Fig. 2) is logically close to the complete fusion (P1) because of a major overlap of data. Interestingly, this fusion P2 is closer to the rRNA points than P1 (i.e. it is more congruent with rRNA), indicating that the rRNA operon and these 44 proteins have undergone a similar phylogenetic history.

Conclusions

The reconstruction of a robust bacterial phylogeny could be possible only if a core of conserved markers not affected by LGT exists. We have shown that the rRNAs and the 44 translational proteins constitute an important part of this core. Nevertheless, we have found that an important subset of the translational apparatus proteins (21 out of 76, plus most tRS) does not belong to this core, some of them undergoing recurrent LGT (e.g. Rps14 [24]). Other candidate members of the conserved core should obviously be searched among the universally distributed proteins, such as DNA gyrase. The probability of the successful transfer of such a gene is very low because its orthologue is already present in the receiving genome, because the new gene has several disadvantages (i.e. poorer gene regulation, inferior translation accuracy because of non-optimal codon usage, and worse interaction with the host proteins). To replace the host orthologue, the selective advantage of the new gene (e.g. resistance to antibiotics) has to be large enough to circumvent all these disadvantages [25].

The analysis of all the genes of the conserved core should fully resolve the bacterial phylogeny, but this will require tree reconstruction methods that are able to handle major biases such as the amino acid composition [16] or the covarion structure [26,27]. Our phylogeny (Fig. 1a) should therefore be considered as a reasonable approximation of the organismal phylogeny because it is probable that specific evolutionary properties of ribosome are not correctly handled by current reconstruction methods. At any rate, the fact that three different approaches in analysing the bacterial phylogeny (rRNA phylogeny, gene content and our multiprotein phylogeny) converge towards a similar result strongly suggests that a true bacterial evolutionary history can be reconstructed. This further implies that bacterial organismal identity exists despite frequent LGTs. This conclusion has to be taken into account when developing a new definition of the much debated species concept for prokaryotes, such as the 'species genome' concept [28].

Acknowledgements

We thank Miklós Müller and Philippe Lopez for critical reading of the manuscript. We acknowledge the provision of sequences before publication by Chicago University, Genome Therapeutics, The Institute of Genome Research, Joint Genome Initiative, Oklahoma University, Sanger Centre, Washington University and Wisconsin University, are acknowledged.

References

1 Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124–2129

2 Woese, C. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6854–6859

3 Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304

4 Snel, B. *et al.* (1999) Genome phylogeny based on gene content. *Nat. Genet.* 21, 108–110

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

6 Jain, R. *et al.* (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 96, 3801–3806

7 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28

8 Wolf, Y.I. *et al.* (1999) Evolution of aminoacyl-tRNA synthetases-analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* 9, 689–710

9 Woese, C.R. *et al.* (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* 64, 202–236

10 Brown, J.R. *et al.* (2001) Universal trees based on large combined protein sequence datasets. *Nat. Genet.* 28, 281–285

11 Rujan, T. and Martin, W. (2001) How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends Genet.* 17, 113–120

12 Kamla, V. *et al.* (1996) Phylogeny based on elongation factor Tu reflects the phenotypic features of mycoplasmas better than that based on 16S rRNA. *Gene* 171, 83–87

13 Kreil, D.P. and Ouzounis, C.A. (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.* 29, 1608–1615

14 Makarova, K.S. *et al.* (2001) Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol. Mol. Biol. Rev.* 65, 44–79

15 Gupta, R.S. and Johari, V. (1998) Signature sequences in diverse proteins provide evidence of a close evolutionary relationship between the deinococcus-thermus group and cyanobacteria. *J. Mol. Evol.* 46, 716–720

16 Lockhart, P. *et al.* (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11, 605–612
17 Kishino, H. and Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* 29, 170–179
18 Goldman, N. *et al.* (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49, 652–670
19 Hasegawa, M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174
20 Woese, C. *et al.* (1991) Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. *Syst. Appl. Microbiol.* 14, 364–371
21 Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–370
22 Eernisse, D.J. and Kluge, A.G. (1993) Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Mol. Biol. Evol.* 10, 1170–1195
23 Bull, J.J. *et al.* (1993) Partitioning and combining characters in phylogenetic analysis. *Syst. Biol.* 42, 384–397
24 Brochier, C. *et al.* (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet.* 16, 529–533
25 Boucher, Y. *et al.* (2001) Microbial genome: dealing with diversity. *Curr. Opin. Microbiol.* 4, 285–289
26 Galtier, N. (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.* 18, 866–873
27 Philippe, H. and Germot, A. (2000) Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.* 17, 830–834
28 Lan, R. and Reeves, P.R. (2000) Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol.* 8, 396–401

Céline Brochier
Eric Bapteste
David Moreira
Hervé Philippe*
Phylogénie, Bioinformatique et Génome, UMR 7622 CNRS, Université Pierre et Marie Curie, 9, quai St Bernard, 75005 Paris, France.
*e-mail: herve.philippe@snv.jussieu.fr

Eukaryotic genes in *Mycobacterium tuberculosis* could have a role in pathogenesis and immunomodulation

Junaid Gamieldien, Andrey Ptitsyn and Winston Hide

Acquisition of new genetic material through horizontal gene transfer has been an important feature in the evolution of many pathogenic bacteria. Here, we report the presence of 19 genes of eukaryotic origin in the genome of *Mycobacterium tuberculosis*, some of which are unique to the *M. tuberculosis* complex. These genes, having been retained in the genome through selective advantage, most probably have key functions in the organism and in mammalian tuberculosis. We explore the role these genes might have in manipulation of the host immune system by altering the balance of steroid hormones.

Mycobacterium tuberculosis is a Gram-positive bacterium that causes approximately three million deaths annually and infects an estimated third of the world's population, making it the most successful human pathogen. The primary site of infection is the lung, where it is initially ingested by pulmonary macrophages in the lower lobes before multiplying intracellularly.

Many bacterial pathogens have evolved the capacity to produce virulence factors that are directly involved in infection and disease. Changes in the genetic repertoire, occurring through gene acquisition and deletion, are the major events underlying the emergence and evolution of bacterial pathogens [1]. Although horizontal transfer of virulence determinants between bacteria is the most common mechanism for acquisition of new genetic material [2–5], the genomes of two obligate intracellular pathogens, *Chlamydia trachomatis* and *Rickettsia prowazekii*, harbour a number of eukaryote-like virulence genes, some of which are shared between the two organisms [6]. *M. tuberculosis* has recently been reported to have the highest number of eukaryotic–prokaryotic interkingdom gene fusions of all the sequenced bacterial genomes [7]. The individual 'fused' genes, however, do not offer immediate clues with respect to the organism's virulence mechanisms.

To determine whether the acquisition of eukaryotic genes by horizontal gene transfer is an important feature in the evolutionary history of the pathogenic mycobacteria, we developed a system of stepwise elimination that identifies eukaryotic-like genes in a bacterial genome. The system, similar to that employed by Wolf *et al.* [6], compared BLASTP [8] E-VALUES (see Glossary) for each predicted *M. tuberculosis* protein against bacterial and eukaryotic subsets of GenBank as a preliminary screen for horizontal transfer candidates. Proteins that scored higher against eukaryotes than bacteria, with at least ten orders of magnitude difference in E-values, were selected as possible candidates for horizontal gene transfer. All mycobacterial protein sequences were removed from the bacterial subset to allow us to identify more ancient gene acquisitions, as well as those unique to *M. tuberculosis*.

In the next step, candidate proteins were compared against a complete non-redundant protein database using the NCBI PSI-BLAST search engine (http://www.ncbi.nlm.nih.gov/blast [8]).

Glossary

Bootstrapping: Estimates confidence levels of inferred relationships by resampling the original data matrix through random replacement of characters and reconstruction of the phylogeny, for a predetermined number of iterations, and determining the frequency that any internal node is recovered.
E-value: The expected number of matches that would occur by chance, with an equal or better score, in the given database search.
Maximum-likelihood method: Infers phylogenetic relationships using a pre-specified model of sequence evolution.
Neighbour-joining method: Infers a branching tree diagram from a sequence similarity distance matrix by successively clustering pairs of taxa together.
Non-congruent phylogeny: Displays a tree topology that is different from the expected or known phylogeny.
Protein-parsimony method: Parsimony methods evaluate all possible trees and select the most parsimonious tree as the one with the minimum number of evolutionary changes.

0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(01)02529-X