

A phylogenomic approach to microbial evolution

Thomas Sicheritz-Pontén and Siv G. E. Andersson*

Department of Molecular Evolution, Evolutionary Biology Center, Uppsala University, 752 36 Uppsala, Sweden

Received August 3, 2000; Revised and Accepted November 6, 2000

ABSTRACT

To study the origin and evolution of biochemical pathways in microorganisms, we have developed methods and software for automatic, large-scale reconstructions of phylogenetic relationships. We define the complete set of phylogenetic trees derived from the proteome of an organism as the *phylome* and introduce the term phylogenetic connection as a concept that describes the relative relationships between taxa in a tree. A query system has been incorporated into the system so as to allow searches for defined categories of trees within the *phylome*. As a complement, we have developed the *pyphy* system for visualising the results of complex queries on phylogenetic connections, genomic locations and functional assignments in a graphical format. Our phylogenomics approach, which links phylogenetic information to the flow of biochemical pathways within and among microbial species, has been used to examine more than 8000 phylogenetic trees from seven microbial genomes. The results have revealed a rich web of phylogenetic connections. However, the separation of Bacteria and Archaea into two separate domains remains robust.

INTRODUCTION

The classification of microorganisms represents a major challenge in biology (1). Molecular phylogenetics based on rRNAs and selected proteins have laid the foundation for a modern classification system, conceptually represented by the 'universal tree of life' (2). However, microbial genomes are highly dynamic in structure and horizontal gene transfer events have been suggested to occur much more frequently than was previously thought (3). The acquisition of foreign DNA combined with intra-genomic rearrangement and duplication events may provide an explanation for the remarkable ability of bacteria to constantly explore new growth habitats. However, a continuous flow of genetic material within and among bacterial species is problematical in the sense that conflicting evolutionary relationships are to be expected from phylogenetic reconstructions based on individual gene sequences (4).

Indeed, an analysis of the complete genome of the hyperthermophilic bacterium *Thermotoga maritima* has shown that

about a quarter of the genes are most similar to their homologues in Archaea (5). Similarly, it has been suggested that almost 20% of the *Escherichia coli* genes are of recent foreign origin (6,7). Thus, individual gene trees may not necessarily reflect the 'correct' species tree. To quantify the frequency at which horizontal gene transfer events occur in bacteria we need to compare phylogenetic data at the genomic level and relate results based on thousands of individual gene sequences to functional annotations and metabolic information. The term 'phylogenomics' refers to such large-scale, genomic approaches to phylogenetic analyses (8). A series of important scientific issues have to be addressed in these global analyses.

First, we need a better understanding of the distribution of horizontal gene transfer events on an evolutionary time scale. A few alternative hypotheses have been proposed to explain the complex patterns of sequence relationships observed in microbial genomes. The 'continual horizontal transfer' hypothesis suggests that gene acquisitions are ongoing processes in microorganisms (9), whereas the 'early massive transfer' hypothesis proposes that massive exchanges occurred early in prokaryotic evolution, long before the diversification of modern microbial species (10). We also have to determine whether genes are equally amenable to horizontal gene transfer or whether some genes are more suitable for tracing evolutionary relationships than others. Genes related to processes that are essential to life, such as replication, transcription and translation, have long been thought to be less likely to be horizontally transferred than genes of importance only for growth in highly specialised milieus (9,11).

The third important issue concerns the methods used for analysis and how the results of these methods are interpreted. With the rapidly accumulating number of sequences in the public databases, sequence similarity is most often defined as the closest match in database searches (best hit) using programs such as BLAST (12). This method is fast and simple and can easily be automated for the analysis of thousands of genes. The so-called 'best hits' have therefore routinely been used as a basis for gene annotations in genome sequencing projects. By analogy, examples of 'odd similarities' have been taken as indications of horizontal gene transfer events (5,13). However, indications of unexpected relationships based on sequence similarity searches may be affected by factors such as gene duplication and divergence and/or by differences in nucleotide substitution rates, which are not taken into account in these simple methods. Therefore, great caution should be exercised when trying to infer functional equivalence and evolutionary relationships solely from the results of simple database searches or from pair-wise sequence similarity measures (14).

*To whom correspondence should be addressed. Tel: +46 18 471 43 74; Fax: +46 18 471 6404; Email: siv.andersson@ebc.uu.se
Present address:

Thomas Sicheritz-Pontén, Department of Biotechnology, the Technical University of Denmark, DK-2800 Lyngby, Denmark

Table 1. Statistics based on the *phylomes* of seven microbial genomes

| Organism | Genes/genome | Trees | Genes/tree | | A (<i>Apure</i>) | E (<i>Epure</i>) | B (<i>Bpure</i>) |
|--------------------------|--------------|-------|------------|-----|--------------------|--------------------|--------------------|
| | | | Average | Max | | | |
| <i>R.prowazekii</i> | 837 | 686 | 47 | 377 | 35 (16) | 176 (121) | 534 (464) |
| <i>C.pneumoniae</i> | 1052 | 731 | 29 | 263 | 37 (18) | 146 (85) | 501 (421) |
| <i>B.burgdorferi</i> | 850 | 748 | 38 | 315 | 50 (24) | 120 (85) | 541 (492) |
| <i>Synechocystis</i> sp. | 3166 | 2011 | 39 | 542 | 212 (121) | 590 (471) | 1318 (1131) |
| <i>A.aeolicus</i> | 1522 | 1202 | 46 | 618 | 266 (173) | 129 (74) | 933 (812) |
| <i>T.maritima</i> | 1849 | 1392 | 47 | 513 | 418 (318) | 107 (59) | 989 (859) |
| <i>P.abysssi</i> | 1765 | 1436 | 35 | 501 | 1003 (893) | 74 (43) | 440 (341) |

A, all connections that include Archaea; E, all connections that include Eukarya; B, all connections that include Bacteria; *Apure*, pure connections to Archaea; *Epure*, pure connections to Eukarya; *Bpure*, pure connections to Bacteria.

To quantify and authenticate horizontal gene transfers it is necessary to establish that the transferred genes are positioned deeply within the presumed donor lineage with high support, as can only be determined by phylogenetic methods. However, phylogenetic methods are time consuming and difficult to automate and therefore have not until recently been regarded as realistic alternatives to the much simpler sequence similarity measures for large-scale genome analyses (8). Phylogenetic methods rely heavily on the accuracy of the underlying multiple alignment, which means that trivial problems such as errors in the alignments and/or bad choices of positions or species may result in a set of trees that are inconsistent with each other. This makes the entire process of choosing homologues, making and editing the alignment, reconstructing phylogenetic relationships and finally testing and evaluating the statistical support for the individual clusters into a very complex and time consuming task that has not been considered suitable for high throughput analysis of genomic sequences.

In this paper we describe the development of an automatic tool for rapid, large-scale phylogenetic reconstruction of multiple genomes with a minimum of manual intervention. To illustrate the use of our phylogenomics approach, we have compared the complete set of phylogenetic trees, the *phylomes*, for seven microbial species. The resulting tree database, which consists of more than 8000 phylogenetic trees, has been sorted according to the functional categories to which the individual gene sequences belong.

MATERIALS AND METHODS

The data set

Genome sequences as well as annotation tables were obtained from the KEGG WWW site (15). The KEGG system contains annotations of genes together with their relative positions in the genome and their presumed roles in metabolic systems. Seven complete eubacterial and archaeal genomes (Table 1) were analysed and integrated in the pyphy system.

Retrieval and identification of homologues

The complete set of genes was compared to a non-redundant protein database consisting of SWISS, TrEMBL and TrEMBL_New (16) with the help of BLAST 2.0 (12). The set of

proteins with significant sequence similarities to the submitted queries formed the basis for the selection of homologous sequences to be included in the alignment, as previously described (17). If one homologue was associated with a particular SWISS ID, all additional protein sequences in the SWISS database with the same ID were extracted to expand the initial data set. That is to say, if one BLAST hit matched EFTU_ECOLI, all entries in SWISS-PROT starting with EFTU_ were included. Entries shorter than 50% or longer than 200% of the length of the query sequence were removed from the final set of homologues.

Phylogenetic reconstructions

All potential homologues were automatically retrieved and aligned with CLUSTALW (18). Phylogenetic trees were constructed with the PAUP* program (19) using neighbour joining (20) and maximum parsimony methods with 100 bootstrap and jack-knife steps, respectively. The neighbour joining method was our method of first choice because of the rapid rate at which it operates. However, using speed optimised parameters for the jack-knifing method (A.Backlund, personal communication) we were also able to reconstruct phylogenetic trees with comparable speed by the parsimony criterion. The trees were reconstructed from 30–50 protein sequences on average (Table 1). The minimum number of homologous sequences per tree was by definition four proteins, whereas the maximum number of sequences per tree often exceeded 500 (Table 1). In total, more than 100 000 phylogenetic trees were reconstructed for test purposes.

Tracing phylogenetic connections

The BLAST score is normally used as a criterion for defining the most similar sequence (best hit) among a long list of potentially similar sequences. The procedure is simple and consists of a set of defined parameters that can be easily automated. However, phylogenetic trees provide topological information rather than sequence similarity measures. Here we have defined the proximity of the query sequence to a set of other sequences in the tree based on nearest neighbour relationships. However, identifying the nearest neighbour is not always straightforward.

The nearest neighbour is formally defined as a set of operational taxonomic units (OTUs) that are connected by the same parental node in the phylogenetic tree (20). The nearest neighbours may be two terminal nodes, i.e. a single OTU may share a node with one other OTU. The nearest neighbours may also be represented by a composite OTU, i.e. an individual OTU may share a node with a subset of other OTUs. Finally, a subset of OTUs may be the nearest neighbour to another subset of OTUs. In this case all individual sequences in one subset of OTUs are equally related to all sequences in the other subset of OTUs.

A practical problem is that the nearest neighbours may be sequences from the same species (e.g. one and the same gene may be represented by several slightly different database entries) or from closely related species (e.g. the same gene may have been sequenced in several different strains). In both of these two cases we need to expand the definition of the nearest neighbour to exclude OTUs that are too closely related to each other. This can be easily done by using different levels of the taxonomic description attached to the entries, e.g. we can exchange the species name *T.maritima* for the genus name *Thermotoga* if the taxonomic descriptor contains 'Bacteria; Thermotogales; Thermotoga'. Similarly, we can exchange the species name *Rickettsia prowazekii* for the genus name *Rickettsia* if the taxonomic descriptor contains 'Bacteria; Proteobacteria; alpha subdivision; Rickettsiales; Rickettsiaceae; Rickettsiae; Rickettsia' or for 'Proteobacteria; alpha subdivision' if the resolution into different kingdoms is more important than the resolution into different bacterial groups.

In this paper we introduce the term 'phylogenetic connection', which we define as nearest neighbours that are taxonomically distinct at any defined level in this hierarchical structure. The identification of a phylogenetic connection is performed in a three-step process. This includes (i) selection of OTUs to be tested for a phylogenetic connection, (ii) identification of the level in the taxonomic hierarchy to which these OTUs belong and (iii) compilation of a list of species or genera that will be ignored in the analysis. The last step allows the search to proceed to a higher hierarchical structure if a phylogenetic connection is denied because it includes too closely related variants, strains or species. A phylogenetic connection is established either by identification of a set of neighbours connected by the query node or by the nearest identifiable node which connects a set of taxonomically distinct nearest neighbours.

The phylogenetic connection provides indications about the selected level in the taxonomic hierarchy to which the query belongs. Any individual gene sequence may have phylogenetic connections to species belonging to several different hierarchical structures. Using simple set theory we can generate lists of genes that are represented by pure hierarchical connections in the *phylome*. For the purpose of this paper we have defined the hierarchical structure at the level of the domain (2). Thus, genes have been classified according to three hierarchical structures, A, B and E, that represent Archaea, Bacteria and Eukarya, respectively. Here a list of archaeal genes (*Apure*) represents genes with phylogenetic connections such that the nearest neighbours consist solely of archaeal genes. However, it should be emphasised that the hierarchical structure in the *phylome* can be defined at any taxonomic level. For example, the bacterial phylogenetic connection (*Bpure*) can be further

subdivided into phylogenetic connections consisting of cyanobacteria, proteobacteria, spirochetes, etc.

Availability of databases and software

For automatic processing of these data sets we have constructed the *pyphy* system, which is a set of modules in the programming language Python (www.python.org), also used for graphical comparison and visualisation of the *phylome* maps. The databases and the software are available at <http://www.cbs.dtu.dk/thomas/pyphy>.

RESULTS

Comparative *phylome* maps

The *phylome* map represents a one-dimensional schematic network of phylogenetic trees and BLAST hits arranged according to the genomic locations of the gene sequences used as queries for the analysis. To present a quick overview of the results, phylogenetic trees and BLAST hits have been colour coded according to their phylogenetic connections. The colour coding system is currently based on the most inclusive taxonomic category, i.e. the domain (2). The circular organisation of the colour coded trees around the genome makes it possible for the user to visually examine the extent to which trees with similar phylogenetic connections are clustered in the genome. Here we have chosen to illustrate the *phylome* map concept with the help of the *R.prowazekii* genome sequence (Fig. 1).

The *R.prowazekii* genome is expected to contain many genes with phylogenetic relationships to mitochondrial proteins. Indeed, a significant fraction of the *R.prowazekii* genes have eukaryotic connections and these are rather evenly distributed around the genome (Fig. 1). By selecting Eukarya, a graphic representation of all genes associated with phylogenetic connections to the Archaea will be highlighted in the *phylome* map of *R.prowazekii*. By selecting a gene tree of interest in the *phylome* map, the user can then display and inspect the topology of the phylogenetic tree from which the phylogenetic connection was inferred. The taxa in the tree are labelled according to their IDs in the non-redundant database (see Materials and Methods). To facilitate a more detailed inspection of the tree topology, the standardised taxon labels can be automatically exchanged for taxonomic names, gene names and/or domain connections. The input data files used for reconstruction of the phylogenetic tree (i.e. the BLAST results, the EMBL files, the multiple alignments and the tree files) can also be inspected via an additional mouse pop-up menu.

Phylogenetic connections and other gene tags

Each gene has a characteristic set of attributes that we refer to as tags. These provide descriptions of the gene at both the biochemical and phylogenetic levels. The standard tags in the *phylome* maps include: (i) the annotation of the gene as provided by the EMBL/GenBank file; (ii) the functional category to which the gene belongs; (iii) the metabolic pathway in which the gene product participates; (iv) the phylogenetic connection assigned to the gene. The user can pose queries and interact with the complex information contained in these tags in a variety of different ways. For example, genes with a given set of features can be identified by selecting tags of interest and choosing a suitable operator such as Union, Intersection,

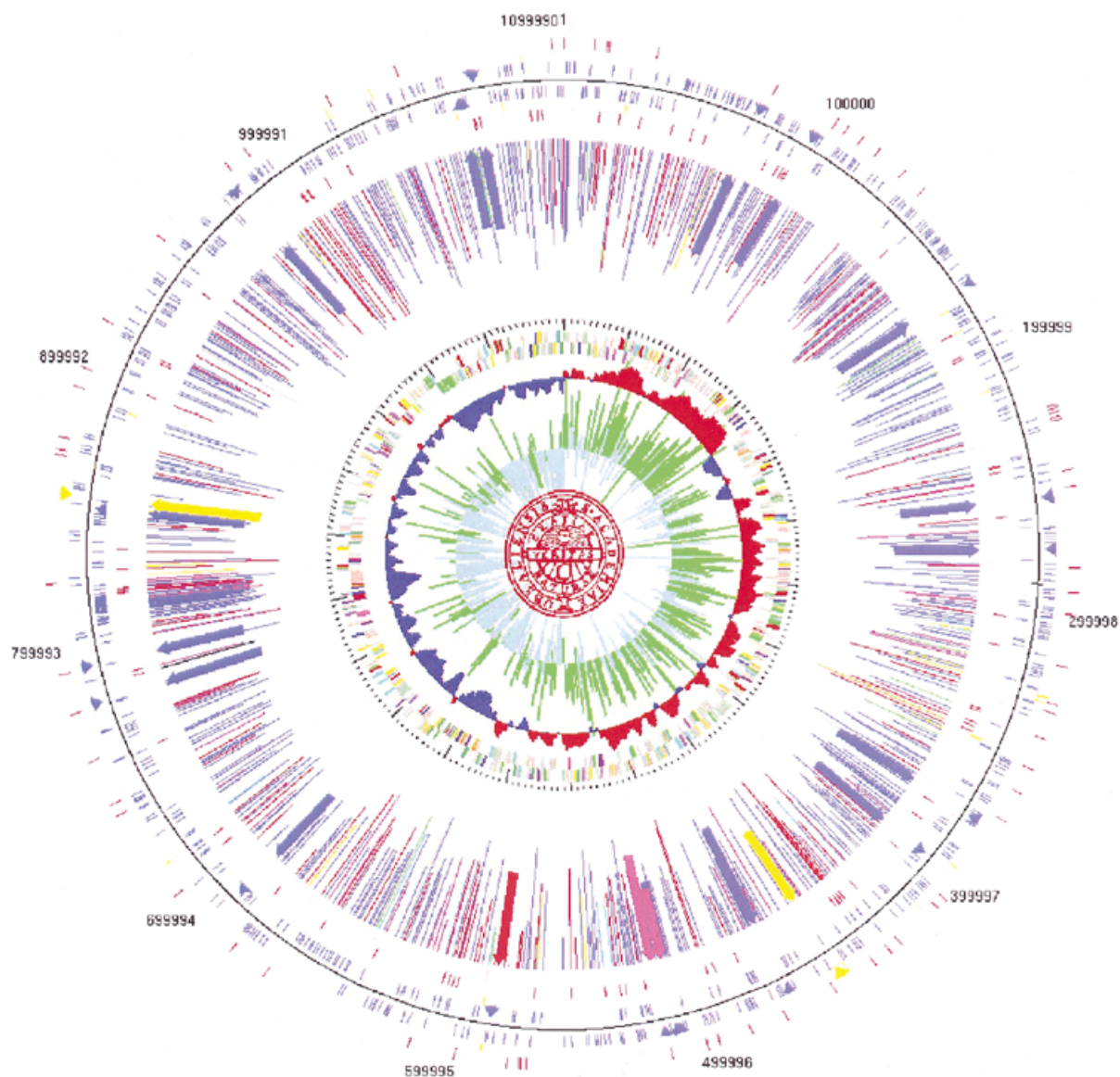


Figure 1. The *phylome* map of *R.prowazekii* (25). The outer circles represent the best hits obtained in searches against public databases using the program BLAST (12). The inner circles represent the phylogenetic connections inferred from phylogenetic reconstructions using the neighbour joining algorithm (20) and the program PAUP* (19). The lengths of the lines in the inner circle are proportional to the number of taxa in the phylogenetic trees. The key to the colour coding system is shown in Figure 2. The thick arrows in the *R.prowazekii* *phylome* map show the location of the aminoacyl-tRNA synthetases.

Difference and Exclusive Disjunction to describe their internal relationships.

By selecting the *R.prowazekii* *phylome* map and the tag 'aminoacyl-tRNA synthetase', 21 annotated aminoacyl-tRNA synthetases are highlighted in the *phylome* map (Fig. 1). By combining the tag 'aminoacyl-tRNA synthetase' with each of the three possible phylogenetic connections (A, B and E), 16 of the 21 enzymes were found to have pure bacterial connections (*Bpure*). One enzyme had a pure eukaryotic connection (*Epure*), two had mixed bacterial and eukaryotic connections (*BEmix*) and two had pure archaeal connections (*Apure*). The latter category consists of the valyl and lysyl aminoacyl-tRNA synthetases and represents putative horizontal transfers to or from the Archaea. Indeed, a more rigorous phylogenetic analysis

provides strong evidence to suggest that the *R.prowazekii* valyl and lysyl aminoacyl-tRNA synthetases are closely related to those of the Archaea. Thus, by combining tags with intelligent match criteria, the pyphy system can function as a tool for filtering out interesting phylogenetic information from the *phylome* maps.

Schematic representation of the *phylome* maps for seven microorganisms

To obtain an overview of the distribution of phylogenetic connections among the three domains we have analysed 11 041 genes from seven microbial genomes with the help of the pyphy system (Table 1). In total, 8206 phylogenetic trees were automatically constructed based on all alignments with a

significant number of homologues. We observed that in 7021 (85.5%) of these phylogenetic trees the query OTUs had pure phylogenetic connections to Bacteria, Archaea or Eukarya. In these trees the interpretation of the domain to which the query sequence belongs is straightforward. The remaining phylogenetic trees (14.5%) were more complex. Here each individual query clustered with species from more than one domain and, accordingly, the cluster included evolutionarily distinct groups.

It can be assumed that the predominant phylogenetic connection for each species reflects the evolution of the species, whereas atypical connections may represent putative horizontal gene transfer events. Indeed, a predominant fraction of the phylogenetic trees reconstructed based on the six genomes previously classified as Bacteria linked the query gene to other bacterial genes (72.1%). As many as 62.7% on average had pure bacterial connections (Table 1). The *phylomes* derived from the early diverging species *Aquifex aeolicus* and *T.maritima* contained bacterial connections to the same extent as the *phylomes* from other Bacteria, with 64.6% of the query sequences having pure bacterial connections. This supports the placement of these two species within the Bacteria. In contrast, the highest fraction of phylogenetic connections in trees reconstructed from the genome of *Pyrococcus abyssi* were to other archaeal species (69.8%), with 62.2% of the genes having pure archaeal connections. We conclude that the deep divergence of Bacteria and Archaea as inferred from rRNA sequence data is supported by a majority of the reconstructed phylogenetic trees in our analysis.

However, a closer inspection of the *phylomes* revealed a striking variation with respect to the identities of the atypical phylogenetic connections. For example, *Rickettsia*, *Chlamydia* and *Borrelia* were found to have much higher frequencies of genes with pure eukaryotic connections (14–20%) than of genes with pure archaeal connections (3–4%). These trends are reversed in *A.aeolicus* and *T.maritima*, which have much higher fractions of genes with pure archaeal connections (16–26%) than of genes with pure eukaryotic connections (5–7%). In contrast, the archaeon *P.abysssi* has a higher fraction of genes with atypical connections to bacteria (27%) than to eukaryotes (3.4%). It is noteworthy that *R.prowazekii* and *Synechocystis* contain the highest fraction of atypical phylogenetic connections to eukaryotic genomes, 20 and 27%, respectively. This is most likely a remnant of the transfer of bacterial genes to eukaryotic genomes during the origin and evolution of mitochondria and chloroplasts, respectively (21).

Distribution of phylogenetic connections in functional categories

The complexity hypothesis suggests that operational proteins with few interactions to other proteins are more likely to be horizontally transferred than informational proteins that are often members of large, complex protein systems (9). To obtain an overview of the distribution of phylogenetic connections, we sorted the phylogenetic trees into a few different functional categories. The connectivity plots for six *phylomes* are schematically shown in Figure 2. Here, each *phylome* is characterised by a colour coded system that reflects all of the theoretically possible phylogenetic connections (A, B, E, AB, AE, BE and ABE). By visually scanning the connectivity plots it can be seen that the

atypical phylogenetic connections are far from being equally distributed among the functional categories (Fig. 2).

A more detailed inspection of the connectivity plots shows that the operational genes display a much broader spectrum of phylogenetic affinities than the informational genes. This effect is particularly strong in *T.maritima*, which has a high fraction of operational genes with atypical connections to archaeal species. *Thermotoga maritima* does not have any known pathway that is completely of non-bacterial origin and only a limited number of atypical phylogenetic connections in the informational category. For example, 97 of the 98 genes coding for translational proteins have pure bacterial connections. Similarly, most genes involved in cellular processes in *T.maritima* are of bacterial origin, in contrast to the many genes involved in energy metabolism and transport functions with phylogenetic connections to the archaea.

Synechocystis sp. and *R.prowazekii* are also exceptional in this respect because of their high frequencies of informational genes with atypical phylogenetic connections to the eukaryotes. This signal is not surprising, however, considering that mitochondria and chloroplasts are thought to have originated from the groups of bacteria to which *Rickettsia* and cyanobacteria belong, respectively (22). Thus, some of the atypical phylogenetic connections that we observe may trace their origin far back in evolutionary time.

Horizontal gene transfer versus intra-genomic expansion events

Genes with atypical phylogenetic connections may be the result of a large number of independent horizontal gene transfers. Alternatively, a small subset of horizontally transferred genes may later have expanded into larger gene families within their new host genomes, or vice versa. To study this in more detail we examined genes coding for oligopeptide ABC transport proteins that are members of large paralogous gene families. These transporters are composed of five subunits encoded by five genes that are clustered in operon-like structures.

The gene order structures of two such clusters in *P.abysssi* and four in *T.maritima* are schematically shown in Figure 3A. Phylogenetic reconstructions based on each of the five individual genes within these clusters yield identical tree topologies (Fig. 3B), suggesting that the unit of duplication and/or horizontal gene transfer has been the cluster rather than the individual gene sequences. A phylogenetic analysis based on a concatenation of the five proteins provides evidence for a complex evolutionary scenario that has involved both gene duplications and horizontal gene transfer events (Fig. 3C). An exhaustive analysis of the oligopeptide ABC transporters is beyond the scope of this paper, but they provide an example of a gene family that has resulted from several intra-genomic duplication events. The result of such expansions will be that multiple genes will show atypical phylogenetic connections, not all of which may have been horizontally transferred. Accordingly, the fractions of atypical phylogenetic connections presented in this paper (Table 1) should only be taken as upper limits for the frequencies of horizontal gene transfers.

DISCUSSION

We have here presented an automatic tool for reconstructing and visualising phylogenetic relationships based on the entire



Figure 2. Schematic representation of the functional distribution of phylogenetic connections derived from six microbial *phylomes*. In, informational genes; Op, operational genes. The following gene category abbreviations are used: A, amino acid synthesis; B, biosynthesis of cofactors; C, cell envelope proteins; E, energy metabolism; I, intermediary metabolism; L, fatty acid and phospholipid biosynthesis; N, nucleotide biosynthesis; O, other; P, cell processes; R, replication; S, transcription; T, translation; X, transport; Y, tRNA synthesis; Z, regulation.

genomes of unicellular organisms. On average 75% of the proteome had enough homologues in the public databases to support multiple alignments and phylogenetic reconstructions. For an average sized prokaryotic genome with ~1500 genes, this will result in more than 1000 phylogenetic trees. By

utilising programmed methods for parsing all trees and summarising their evolutionary information, this otherwise extremely time consuming task can be converted into a realistic method for high throughput analyses of microbial genomes. In this study we have referred to the complete set of available

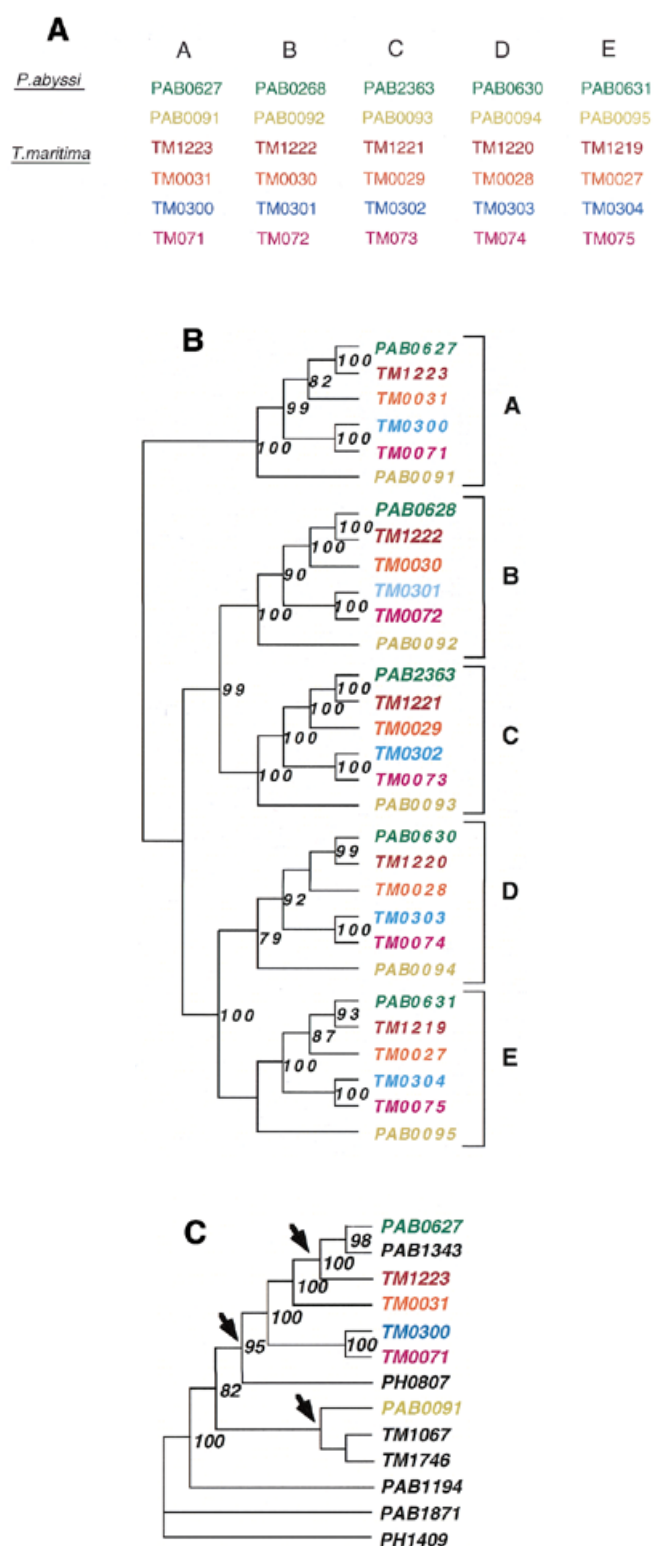


Figure 3. Phylogenetic analyses of oligopeptide ABC transporters in *T.maritima* (TM), *Pabyssi* (PAB) and *Pyrococcus horikoshii* (PH). Schematic picture showing the co-location (A) and phylogenetic relationships based on the individual (B) and combined (C) protein sequences of the oligopeptide ABC transporters. Neighbour joining (NJ) and maximum parsimony (MP) methods gave similar topologies. Values at nodes indicate the percentage of 1000 neighbour joining bootstraps. Values <70% are not shown. Arrows indicate sites of putative horizontal gene transfer events.

phylogenetic trees for an organism as the *phylome* and their graphical representation on the genome as the *phylome* map.

There are two important applications of the *phylome* maps. One concerns the practical use of this information for gene annotations in genome sequencing projects (8). Currently, a majority of the annotated gene functions have been inferred solely from sequence similarities to genes in the public databases. However, paralogous genes need not necessarily have the same function and cannot be distinguished by simple pair-wise sequence similarity measures. This is an exponentially growing problem since an ever increasing number of gene functions are being inferred from pair-wise sequence similarities to genes whose functions have been inferred from pair-wise sequence similarities, etc. Provided that automatic methods for phylogenetic reconstructions become available, the evolutionary information contained in phylogenetic trees can be of great help, particularly for functional predictions of large gene families (8). In the future it would be highly desirable if experimentally associated gene functions could be tagged to individual branches in the phylogenetic trees and used as reference points for standardised methods of gene annotation.

A second important application of the *phylome* maps concerns the gathering of phylogenetic information. By systematically parsing phylogenetic trees and sorting phylogenetic connections according to defined criteria, a simple database describing the *phylome* of a genome can be generated. This allows the user to pose a series of questions to the *phylome* database by combining different sets of evolutionary and biochemical queries. It may also help resolve questions about whether there is a 'universal tree of life' (2) or whether horizontal gene transfers have occurred to such an extent that no universal classification is possible (4).

The prevalent view of microbial phylogeny is based on molecular sequence data, particularly rRNA sequences (2). However, many phylogenetic trees based on other gene sequences have yielded conflicting results and it has been argued that the history of life may more appropriately be described as a cyclic network (4). It is well established that horizontal gene transfer of antibiotic resistance and virulence traits occurs at regular intervals in nature (3). Taken to the extreme, it has been suggested that we might have to abandon phylogenetic classifications because of the extent to which horizontal gene transfer events have taken place (4). Alternatively, we might consider viewing the domains as taxonomic descriptors that are based on shared genes, rather than on shared ancestries (4).

However, our results suggest that there is a backbone of genes with an evolutionary history that is compatible with the main features of the 'universal tree of life' (2). This is also consistent with results from much simpler analyses based on BLAST searches (13) and gene contents (23,24). However, unlike these methods, our results provide indications of shared ancestry rather than merely supplying lists of taxonomic descriptors. Although our analysis does not examine whether there is a natural scheme for all levels in the hierarchy, it suggests that a majority of genes have phylogenetic signals compatible with the domain structure of the 16S rRNA tree (2). This distinction was also observed in a previous analysis based on 203 operational and 109 informational genes, 65 and 92% of which generated trees that supported separation of the bacterial and the archaeal domains (9). It remains to be seen whether this

backbone of conserved genes is the same for all microbial genomes.

Although our analysis has been designed to meet demands on automation as well as sophistication, it has certain limitations. First, automatic handling of the data means that the results can only be taken as indications of evolutionary relationships. The more than 8000 alignments and phylogenetic trees reconstructed in this study should ideally be manually inspected and edited, which is beyond the scope of this study. It should also be emphasised that our current view of phylogenetic connections is inferred from sequences currently available in the public databases. It cannot be excluded that organisms with a closer relationship to the query sequences will be described in the future. Another problem concerns the difficulty of distinguishing paralogous genes that were retained differently in different lineages from horizontal gene transfer events across lineages. Thus, phylogenetic connections to atypical lineages may have evolutionary explanations other than horizontal gene transfer.

For example, consider the bacterium *T.maritima*, which has a remarkably high fraction of archaeal connections. We can think of at least three possible evolutionary scenarios to explain this observation. First, that all genes with atypical phylogenetic connections were independently transferred from the Archaea into *T.maritima*, or vice versa. Second, that a small number of genes were transferred from the Archaea into *T.maritima*, or vice versa, and these were later expanded into large gene families within the recipient genome. Third, that the genes originated in the last universal common ancestor but were differentially lost or evolved so fast in some lineages that they are no longer recognisable as homologues.

To a first approximation we can assume that the atypical phylogenetic connections identified in this study represent a mixture of these three scenarios. To distinguish between a transfer from *T.maritima* to the Archaea or vice versa, the trees need to have a broad representation of homologues within both of the two domains, which is most often not the case. If a small number of archaeal genes have expanded into large gene families within the *T.maritima* genome we would expect to find phylogenetic trees which contain clusters of *T.maritima* genes, all of which have a deeper phylogenetic connection to the Archaea rather than to the bacteria. Indeed, at least some operons with archaeal connections seem to have duplicated in the *T.maritima* genome subsequent to their integration, as illustrated by the cluster of genes coding for the oligopeptide ABC transporters.

Despite these obstacles, we believe that the overall statistics presented in this manuscript are accurate enough to provide a glimpse of 'the big picture'. To further explore explanations for atypical phylogenetic connections we need to examine the *phylome* maps in much greater detail. Insights into the complex evolution of prokaryotic genomes can only be gained by overlaying the *phylome* maps with metabolic pathways, functional categories and other annotations. This will allow us to test hypotheses about microbial evolution and identify, if possible, a core set of genes that represent the majority classification system.

ACKNOWLEDGEMENTS

This study was supported by the Fourth Framework programme of the European Commission (BIO4-CT95-0130).

We thank Desmond Higgins, Mikael Tholleson, Anders Backlund and Charles Kurland for stimulating discussions and David Ardell for helpful comments on the manuscript. No living trees were harmed during this analysis.

REFERENCES

1. Woese, C.R. (1994) There must be a prokaryote somewhere: microbiology's search for itself. *Microbiol. Rev.*, **58**, 1–9.
2. Woese, C.R., Kandler, O. and 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.
3. Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovations. *Nature*, **405**, 299–304.
4. Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2128.
5. Nelson, K.E., Clayton, R.A., Gill, S.R., Gwin, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A. *et al.* (1999) Evidence for lateral gene transfer between Archaea and bacteria from the genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.
6. Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, **44**, 383–397.
7. Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA*, **95**, 9413–9417.
8. Eisen, J. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.
9. Jain, R., Rivera, M.C. and Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA*, **96**, 3801–3806.
10. Woese, C.R. (1998) The universal ancestor. *Proc. Natl Acad. Sci. USA*, **95**, 6854–6859.
11. Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.*, **51**, 221–271.
12. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R. and Koonin, E. (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.*, **14**, 442–444.
14. Logsdon, J.M. and Faguy, D.M. (1999) *Thermotoga* heats up lateral gene transfer. *Curr. Biol.*, **9**, R747–R751.
15. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.
16. Bairoch, A. and Apweiler, R. (1999) The SWISS-PROT protein sequence databank and its supplement. *Nucleic Acids Res.*, **27**, 49–54.
17. Duret, L., Perriere, G. and Gouy, M. (1998) *Molecular Biology Databases*. Kluwer Academic Press, The Netherlands.
18. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
19. Swofford, D.L. (1998) *PAUP: Phylogenetic Analysis Using Parsimony (and Other Methods)*, Version 4. Sinauer Associates, Sunderland, MA.
20. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 672–674.
21. Karlberg, O., Canbäck, B., Kurland, C.G. and Andersson, S.G.E. (2000) The dual origin of the yeast mitochondrial proteome. *Yeast Comp. Funct. Genomics*, **17**, 170–187.
22. Gray, M.W., Burger, G. and Lang, B.F. (1999) Mitochondrial evolution. *Science*, **283**, 1476–1481.
23. Fitz-Gibbon, S.T. and House, C.H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.*, **27**, 4218–4222.
24. Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content. *Nature Genet.*, **21**, 108–110.
25. Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C.M., Podowski, R.M., Näsälund, A.K., Eriksson, A.-S., Winkler, H.H. and Kurland, C.G. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 133–140.