Research

Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns

Galina V Glazko* and Arcady R Mushegian*†

Addresses: *Stowers Institute for Medical Research, 1000 E 50th Street, Kansas City, MO 64110, USA. †Department of Microbiology, Molecular Genetics, and Immunology, University of Kansas Medical Center, Kansas City, KS 66160, USA.

Correspondence: Galina V Glazko. E-mail: gvg@stowers-institute.org

Published: 27 April 2004

Genome ***Biology*** 2004, **5:**R32

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/5/R32

Received: 11 November 2003
Revised: 19 February 2004
Accepted: 31 March 2004

Abstract

Background: Phyletic patterns denote the presence and absence of orthologous genes in completely sequenced genomes and are used to infer functional links between genes, on the assumption that genes involved in the same pathway or functional system are co-inherited by the same set of genomes. However, this basic premise has not been quantitatively tested, and the limits of applicability of the phyletic-pattern method remain unknown.

Results: We characterized a hierarchy of 3,688 phyletic patterns encompassing more than 5,000 known protein-coding genes from 66 complete microbial genomes, using different distances, clustering algorithms, and measures of cluster quality. The most sensitive set of parameters recovered 223 clusters, each consisting of genes that belong to the same metabolic pathway or functional system. Fifty-six clusters included unexpected genes with plausible functional links to the rest of the cluster. Only a small percentage of known pathways and multiprotein complexes are co-inherited as one cluster; most are split into many clusters, indicating that gene loss and displacement has occurred in the evolution of most pathways.

Conclusions: Phyletic patterns of functionally linked genes are perturbed by differential gains, losses and displacements of orthologous genes in different species, reflecting the high plasticity of microbial genomes. Groups of genes that are co-inherited can, however, be recovered by hierarchical clustering, and may represent elementary functional modules of cellular metabolism. The phyletic patterns approach alone can confidently predict the functional linkages for about 24% of the entire data set.

Background

Completely sequenced genomes and their gene repertoires are an important resource for studying biological evolution and cellular function. A crucial step in genome analysis, and the foundation of evolutionary and metabolic reconstructions, is determination of orthologous relationships between genes in different genomes [1]. In 1997, Tatusov, Koonin and Lipman combined orthologs and their lineage-specific duplicates into clusters of orthologous groups (COGs) and proposed the first practical algorithm for finding orthologs on a large scale [2]. They introduced phyletic patterns as a representation of the distribution of COGs across genomes, useful for tracking the evolutionary events such as vertical gene inheritance, gene loss and horizontal transfer.

Pellegrini and co-workers [3] emphasized the idea that phyletic patterns can also be used as a post-homology method of predicting protein function, on the premise that genes/COGs encoding functionally linked proteins are co-inherited (simultaneously present or simultaneously absent) in the same subsets of genomes. A functional link between two proteins can be understood either as physical interaction between them, or, more broadly, as their involvement in the same metabolic pathway or functional system, and phyletic patterns are coded as strings of bits, standing for presences or absences of homologs in different genomes. It has been proposed that the Hamming distance of 3 bits or less between phyletic patterns is a useful similarity threshold for detecting functionally linked genes [3]. *Ad hoc* application of the method produced several experimentally validated predictions, such as a novel type of isopentenyl pyrophosphate isomerase in archaea and some bacteria [4,5], several participants in the 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway of isoprenoid biosynthesis in bacteria and plants [6], and new components of queuosine biosynthesis pathway in Gram-positive bacteria [7].

Even with complete genome sequencing and high-throughput determination of gene function, many central metabolic pathways remain only partially characterized. The candidate genes filling the 'missing' steps are sought, and phyletic patterns may be used to identify many more such candidates. In practice, this approach is usually combined with other homology and post-homology methods, such as measurement of gene coexpression, prediction of coexpression from operon structure, and identification of multidomain fusions [8-10]. We do not know how many functional connections between genes/COGs can be inferred solely from their co-inheritance. On a more general note, co-occurrence of genes in genomes is one measure of their association in gene networks, and quantification of this association is needed for any system-wide study of gene function and evolution.

To utilize fully the information offered by phyletic patterns, and to understand their limitations, we seek a better understanding of general properties of patterns and distances between them. A possible limitation of the phyletic-pattern method is that lineage-specific gains and losses of genes, thought to be pervasive in microbial evolution [11], will corrupt the similarity, increasing distance between functionally linked genes. One example of a pathway teeming with differential gains and losses is the tricarboxylic acid (TCA) cycle, which is present in its 'full' (that is, *E. coli*-like) form in only a few species, mostly within the proteobacterial clade, but is rearranged in other microbial lineages, presumably in connection with adaptation to changes in the redox status of the environment (Figure 1 and [12]).

A special case of gene gain/loss is gene displacement, when the same function is performed by non-orthologous genes in different species [13]. For example, most enzymes from the triose part of the glycolytic pathway are present in almost every species, but one activity, phosphoglycerate mutase, can be carried out by three non-orthologous genes, and the pattern for each of these COGs is not a good match to the rest of the pathway (Figure 1). Phyletic patterns themselves, however, may be used to track displacements, by assuming that the alternative isofunctional genes display negative correlation, or 'complementarity'. A recent example of such an approach is the discovery of the novel type of thymidylate synthase, flavin-dependent ThyX, deduced by reversing presences and absences in a pattern of the conventional, folate-dependent thymidylate synthase ThyA [14]. As with positive correlations, the complementary relationship is obscured by asynchronous gains and losses and by functional redundancy, when two genes performing the same molecular function are encoded by the same genome (Figure 1).

Recent attempts at a more quantitative understanding of phyletic patterns include devising a scoring function for negative correlation, which has helped to find displacements of thiamine biosynthesis genes among the candidates shortlisted by other methods [15], and development of significance tests for similarities between two patterns [16,17]. It has also been proposed to improve the sensitivity of phyletic pattern matching by combining binary information of gene presence/absence and phylogenetic distance between orthologs [18,19].

Figure 1 *(see following page)*
Phyletic patterns are corrupted by gene gains and losses. The consensus phylogenetic tree on top is the species' tree based on genomic content [26]. Small black and white squares indicate, respectively, presences and absences of genes in each species. (**a**) TCA cycle. Blue box indicates the 'canonical' cycle, as known from saprophytic Enterobacteriaceae with large genomes. (**b**) Glycolysis. The green box indicates omnipresent COGs in the evolutionarily ancient bottom part of glycolysis, and the red box indicates three COGs coding for phosphoglycerate mutase activity. None of the patterns in the red box is close to the patterns in the green box, even though all these COGs are functionally linked. (**c**) Most genomes have just one of the two types of thymidylate synthase, but the blue boxes indicate several exceptions to this rule. **(d)** The full names of the species listed along the top of (a) and the TCA enzymes corresponding to the COGs shown in (a-c).

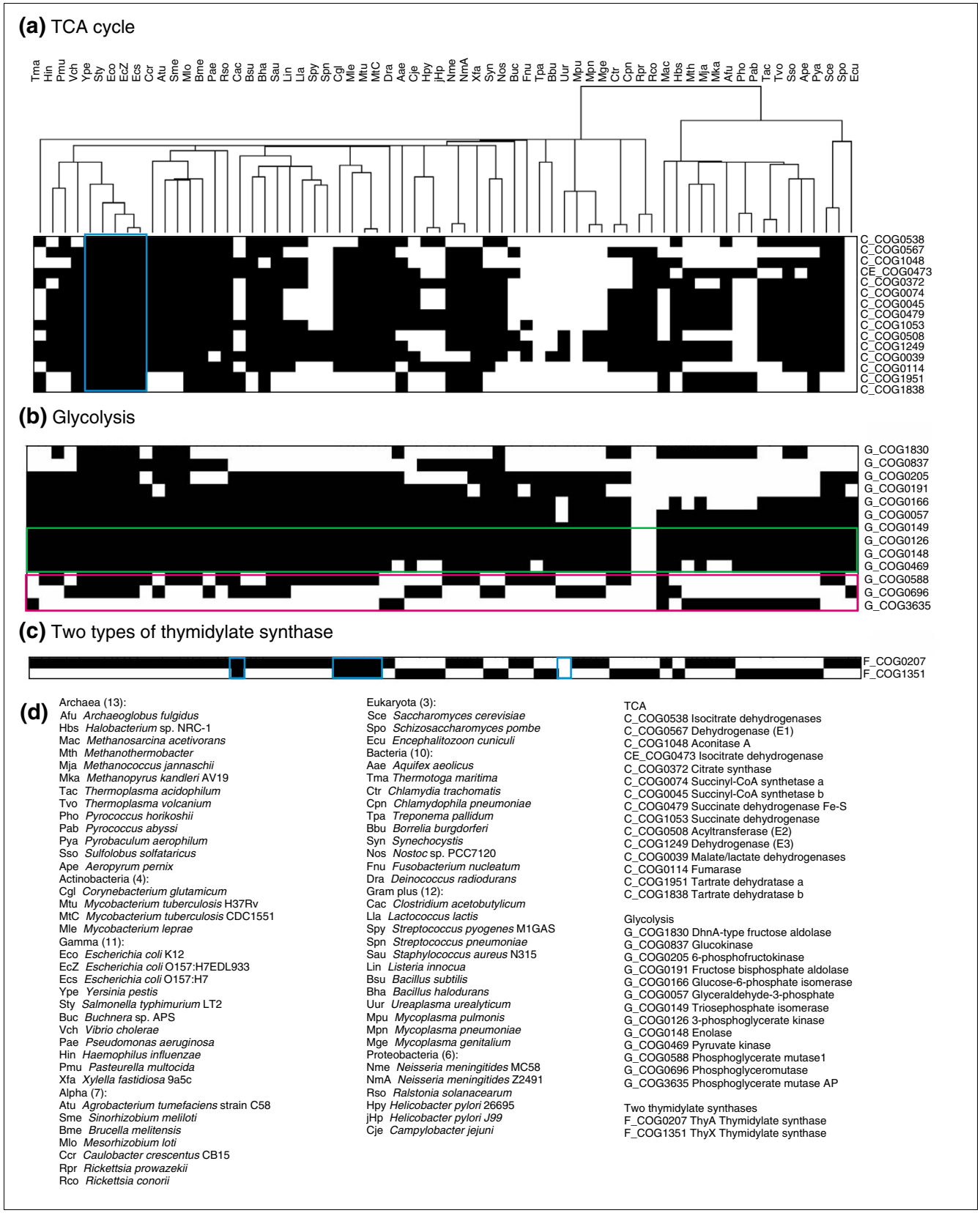

Figure 1 *(see legend on previous page)*

In this work, we characterize the relationships between functionally linked genes/COGs across multiple genomes, and ask what can be inferred, in a systematic way, about the metabolism and evolution of prokaryotes, on the basis of phyletic patterns alone. Four main components of our quantitative analysis are: distance between patterns; method for producing graphs based on the distance data; method for partitioning the graph into subsets; and estimation of error rate in predicting functional links. Generally speaking, phyletic patterns are binary vectors in species space, and distance between them can be measured in many ways. Patterns and the set of distances between each pair of them define a graph, in which one may discern subgraphs, or clusters, of similar pattern vectors. The quest for finding functionally linked genes/COGs then amounts to constructing a graph in which the number of automatically identifiable, biologically relevant clusters is maximized.

Results

Hierarchical clustering of phyletic patterns

The key question in any clustering is the choice of the appropriate combination of distance measure and clustering algorithm. We investigated the effect of various distances between patterns, of different clustering approaches, and of several methods of tree splitting on the recovery of functionally linked proteins.

Several measures of distance between phyletic patterns have been proposed [3,18,20-22]. Most of them do not address a crucial requirement, which we illustrate in the following example. Consider two pairs of proteins (x_1, y_1) and (x_2, y_2), with patterns x_1 = (1011110), y_1 = (0111110), x_2 = (1000000), y_2 = (0000001). We are interested in whether there is a functional link between x_1 and y_1, and between x_2 and y_2. Clearly, only in the case (x_1, y_1) can it be said that 'two proteins tend to be found together'. Yet, most distances, including Euclidean and other l_p-norms, Hamming distance, and J-divergence, are the same in both cases (see Materials and methods for details). The two cases are nevertheless readily distinguishable by the mutual information (MI) measure, and are placed even further apart when using complement of correlation coefficient $d_r = 1 - \mathrm{corr}(\vec{x}, \vec{y})$, or its modifications, such as squared anticorrelation, also called diametric distance [23], and absolute anticorrelation (see Materials and methods). MI- and correlation-based measures are further compared in the next section.

To derive clusters of related patterns from their pairwise distances, we have explored several unsupervised clustering techniques, of both agglomerative and divisive type. Divisive algorithms, such as K-means clustering and bisection [24], need an *a priori* fixed number of clusters, which is unknown. When we fixed this number using the average number of the UPGMA clusters, we found that divisive methods under-per-form compared to agglomerative algorithms. Therefore, agglomerative approaches were used throughout most of the study, in particular the hierarchical clustering methods UPGMA (unweighted pair-group method with arithmetic mean) and neighbor joining (NJ). The programs that we used produce a tree-like graph of phyletic patterns, exemplified in Figure 2.

The large tree-like graph produced by hierarchical clustering of phyletic patterns has to be partitioned into smaller graphs in order to find groups of functionally linked proteins. The splitting criterion can be chosen beforehand, by, say, deciding on the upper level of distance at which two phyletic patterns are still considered similar, or by controlling the size or number of clusters. Instead of making a more or less arbitrary choice of such parameters, we used the distribution of similarities between patterns to infer the threshold at which a similarity becomes significantly higher than average [25]. We settled on a threshold at which about 90% of the entire dataset was included in the partitioned clusters (see Materials and methods for details).

Species themselves can be seen as vectors in the COG space, and distances between such vectors can be used to build the species' phylogeny [26]; this aspect is not considered in the current work, except for illustrative purposes (Figure 2, tree at the top).

The quality of clustering solutions

We studied the quality of eight clustering solutions produced by combining two clustering algorithms - UPGMA and NJ - and four variants of correlation-based distance (all graphs are available from the authors on request). We were interested in two criteria of quality: sensitivity and percentage of lost data. Sensitivity of a solution is defined as the percentage of genes/COGs that belong to the same pathway or functional system and are assigned to the same cluster, counted for each pathway and averaged. The percentage of lost data is the fraction of COGs that belong to the set of known pathways, but were not included into our clusters at a given similarity threshold. In these tests, we used 52 pathways and functional systems, containing 716 COGs altogether, from the COG database [27]. The sensitivity and the percentage of lost data both depended on the clustering algorithm and distance measure. Use of diametric distance resulted in the major improvement in sensitivity and the lowest percentage of the lost data (Figure 3a).

Functional inference on the basis of phyletic patterns has been benchmarked by von Mering *et al.* [28]. They used smaller set of species and distances based on mutual information, and evaluated the performance of the method by comparing linked pairs of genes in *E. coli* and their co-occurrence in the KEGG metabolic maps (see Figure 2 in [28] for details). We compared their and our approach in the context of the current dataset, by clustering our phyletic patterns using their MI-derived distance, and interrogating KEGG maps

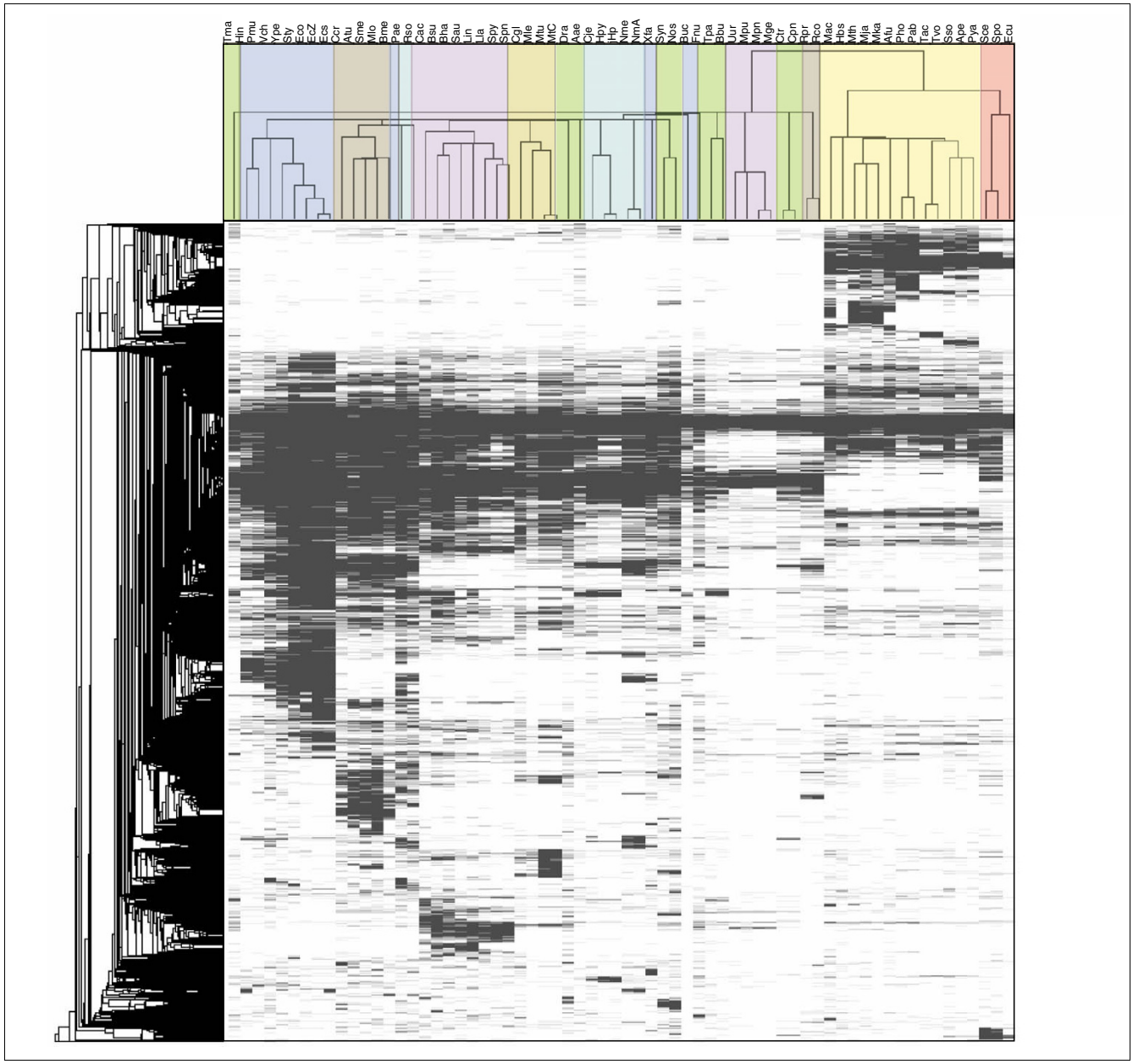

Figure 2
Groups of phyletic patterns and COGs revealed by hierarchical clustering of patterns in species space. The presentation is similar to Figure 1, but the black and white squares are vertically compressed in order to show all 4,589 COGs in one figure. The full tree of COGs is shown at the left; at 170 COGs per 1 mm height, it is not particularly suitable for visual consumption, but some closely linked clusters (short branches) can be discerned.

with *E. coli* proteins. The performances of correlation and MI-based distances were quite similar in this test (Figure 3b). Apparently, however, correlation distance is more accurate in assigning to genes to metabolic pathways as defined in the COG database than to larger KEGG charts (Figure 3a and G.V.G. and A.R.M., unpublished data).

Analysis of the clustering quality within the individual pathways and functional systems indicate that they tend to fall into three broad categories. For some pathways, such as heme biosynthesis or TCA cycle, the specificity of clustering was similar and low, regardless of the methods. Other pathways and systems were confidently clustered regardless of the protocol. These include, for example, the MEP pathway of terpenoid biosynthesis, lipid A biosynthesis, and the NADH-ubiquinone oxidoreductase complex. The third, and largest, category included pathways for which recovery in a cluster was dependent on the clustering method. Perhaps predictably, the percentage of correctly extracted genes in a pathway correlates significantly ($p < 0.05$) with its average

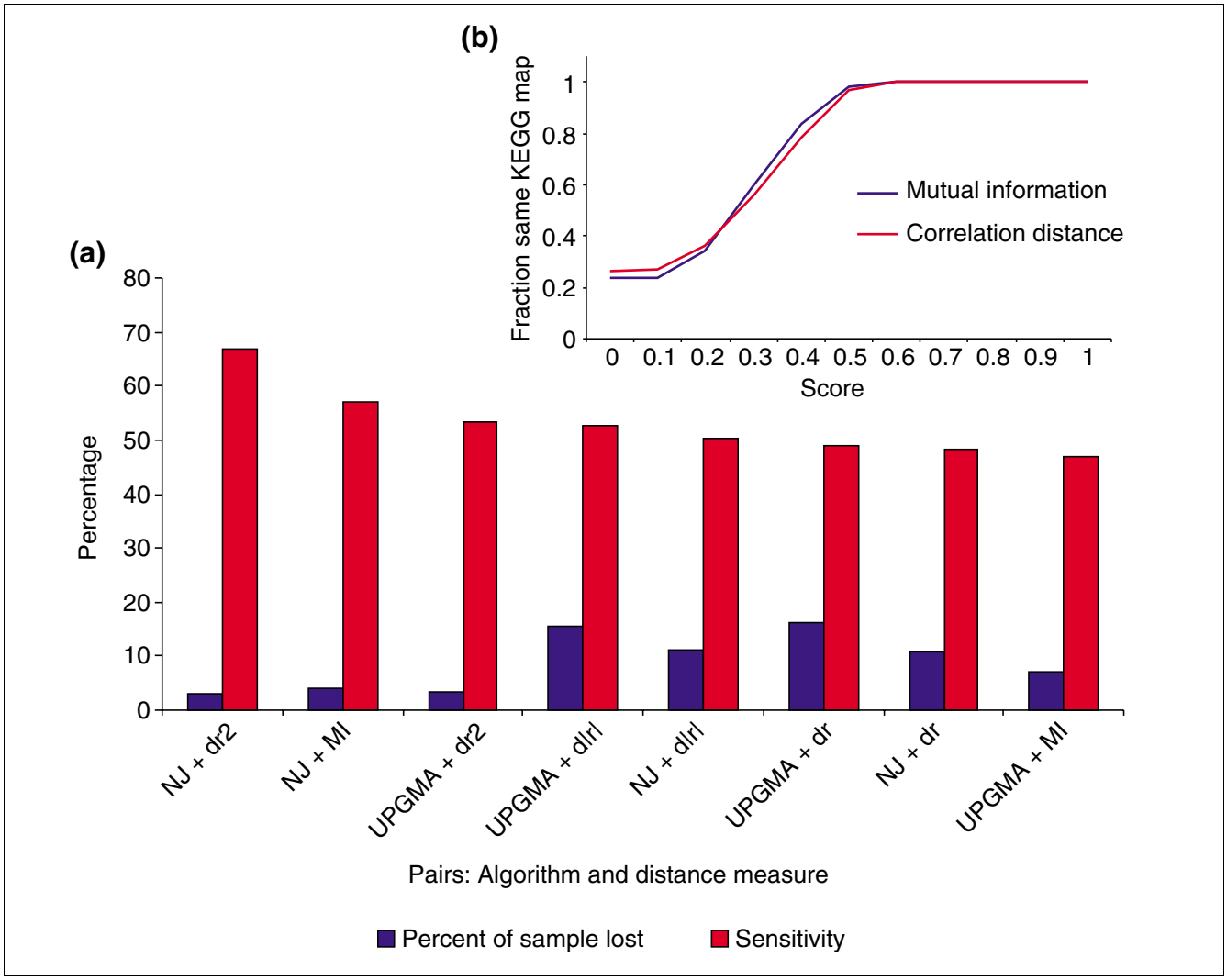

Figure 3
Comparison of distance measures and clustering algorithms. (**a**) Diametric distance combined with NJ clustering results in the highest sensitivity and the smallest percentage of lost data. (**b**) The effect of selected distance measures between phyletic patterns on the recovery of functionally linked pairs of genes. The criteria of functional linkages on the basis of the KEGG maps, as well as the values for mutual information are as in [28].

information content; that is, with the conservation of phyletic patterns among the members of a pathway (Additional data file 1). It seems likely that, given the genes already assigned to a partially characterized pathway or function, one might be able to estimate the probability that phyletic patterns will be helpful in finding the functionally linked genes.

Partitioning of the best clustering solution

The NJ algorithm in combination with diametric distance between phyletic patterns produced the clustering solution in which the known pathways were optimally recovered. To detect more relationships between phyletic patterns, and novel functional links between genes, we analyzed that clustering solution manually, by studying all clusters of similar phyletic patterns within the graph.

One obvious result of our analysis is the existence of several large clusters of co-inherited COGs, seen as prominent rectangles of black and white (Figure 2). Inspection of the corresponding phyletic patterns indicates mostly phylogenetic, rather than functional, relationships, namely the presence of these COGs in all species, or only in bacteria, or only in archaea/eukarya. The former type of pattern reflects the minimal gene set compatible with modern-type cell [29]. The latter two patterns apparently indicate extreme divergence of some pathways between bacteria and archaea/eukarya, and the independent origin of other pathways in these domains of life [30].

Each of these three clusters contains COGs from more than one functional system. The minimal gene set (about 70 COGs)

is dominated by proteins involved in translation and transcription, and also includes components of other systems, such as protein maturation and nucleotide salvage [31]. Archaea and eukarya share, to the exclusion of bacteria, many ribosomal proteins, basic machinery for DNA replication and transcription, some factors of RNA transcription, translation and decay, and a few metabolic enzymes (about 55 COGs [30,31]). Forty-seven COGs found in all bacteria but not in archaea have roles in replication, transcription, translation and protein secretion. Thus, if an uncharacterized protein has a phyletic pattern similar to any of these three patterns, this would suggest a shortened list of functional possibilities, but would not be sufficient to pinpoint the pathway.

We removed these large clusters and focused on identifying every small cluster that consisted of proteins with experimentally established functional connections. We called these functionally linked clusters of genes 'PP-clusters', because genes in these clusters share similar phyletic patterns. There were 223 PP-clusters, ranging in size from two to 23 COGs, with diametric distance from zero to 0.4, and including 890 COGs (24% of the entire dataset) altogether (see the list of PP-clusters in Additional data file 2).

To estimate the probability of obtaining these functional connections by chance, all COGs were randomly assigned to clusters, so that the average size was the same as the average PP-cluster size (327 random clusters, 14 COGs per cluster on average). The ratios of experimentally established functional connections observed within PP-clusters and at random were computed for 100 independent replicates of random clusters. The probability of getting, in random trials, as many or more functional connections as found in PP-clusters was estimated to be less than 3%. Thus, the functional linkage of COGs in PP-clusters was highly significant.

We were next interested in how many of these tightly linked PP-clusters could be derived automatically, without manual inspection. We computed the range of the average within PP-cluster branch lengths, which, in the case of diametric distance, were found to vary from 0 to 0.4, and derived clusters in one step, by cutting the graph in Figure 2 at several fixed lengths within this range. Cutting at two different branch lengths produced the same number (89) of automatically derived PP-clusters, but the number of COGs included in these clusters was different (Table 1). The number of false positives, estimated as the percent of automatically derived PP-clusters that were not presented in manually derived PP-clusters, was less than 20% in each case.

The largest distance between two COGs in one PP-cluster (0.36) was observed for two subunits of NADH:ubiquinone oxidoreductase - COG1143 and COG1894 - linked into PP-cluster both manually and automatically. Among the 66 genomes, 25 contain both these COGs, and 15 genomes either one or the other, giving a Hamming distance of 15. Although this is an extreme case, many COGs in other PP-clusters were separated by Hamming distances as high as 8 to 10. Thus, hierarchical clustering with diametric distance can detect functional links in the zone where more simple measures were not particularly helpful.

Table 1

Manually and automatically derived PP-clusters*

Procedure of PP-cluster definition	Number of PP-clusters	Total number of COGs in all clusters	COGs shared with manually derived PP-clusters	Average number of COGs in a cluster	Number of clusters absent in manually derived PP-clusters	Number of pure RS† clusters	FPs‡
Manual annotation	223	890	N/A	4.1	N/A	-	-
Automated tree cutting at average branch length 0.2	89	1,774	315	19.9	38	20	0.19
Automated tree cutting at average branch length 0.3	89	3,960	395	44.5	26	12	0.16

*PP-clusters, clusters of COGs functionally linked on the basis of similar phyletic patterns. †RS clusters: clusters containing only COGs annotated as 'poorly characterized' in COGs database, where R stands for 'general function prediction only' and S stands for 'function unknown'. ‡The number of false positives (FPs) is the proportion of clusters that were not presented in manually derived PP-clusters.

Case-by-case analysis of phyletic pattern hierarchy: known and new functional connections

The PP-clusters are dominated by groups of COGs from the same metabolic pathway or functional system. In 56 cases, however, a PP-cluster contained component(s) without an established functional connection to the rest of the cluster. In 17 cases, such COGs were the ingroups within the PP-cluster; that is, the distance between a COG and the rest of the PP-cluster was smaller than between some of the functionally linked PP-cluster members. In 23 cases, the connection between the 'unexpected' COG and the rest of the PP-cluster could be tentatively proposed. Examples of such novel functional connections follow (see Additional data file 2 for complete listing of COGs and additional predictions).

PP-cluster new005

PP-cluster new005 (genes found in archaea, eukarya and gammaproteobacteria) is a multienzyme system probably involved in RNA maturation. It contains RNA 3'-terminal phosphate cyclase (COG0430), pseudouridylate synthase distantly related to TruB (COG0585), and a multifunctional protein (COG1444) that is found in the rRNA processosome [32] and contains an uncharacterized enzymatic domain with a Rossmann-like fold, a Walker-type ATPase domain, a GNAT-type acetyltransferase and a putative nucleic acid-binding domain [33].

PP-cluster new023

PP-cluster new023 links cell-shape determination genes *mreA* (COG1077), *mreB* (COG1792), *ccmA* (COG1664) and COGs involved in flagellum biosynthesis and chemotaxis (in diverse bacteria, including spirochetes, proteobacteria and cyanobacteria).

PP-cluster new015

PP-cluster new015 suggests novel activities involved in the MEP pathway (most bacteria, except Gram-positives) and links it to the biosynthesis of cell-wall components (COG0860, COG0791).

PP-cluster new012

PP-cluster new012 from proteobacteria links a component of the N-end rule protein degradation pathway - Leu/Phe-tRNA-protein transferase (COG2360) - to the putative executive components of the pathway, two metalloproteases (COG2377 and COG0339).

PP-cluster new001 and PP-cluster new006

Two specialized systems consist of divalent cation transporters and enzymes predicted to require these cations for activity. PP-cluster new001 (diverse bacteria, archaea and some fungi) contains a zinc transporter (COG0053) and membrane zinc-dependent hydrolase (COG2220). PP-cluster new006 (many bacteria and some archaea) contains thymidine phosphorylase (COG0213) and two proteins transporting cobalt or similar divalent cation (COG0619 and COG 1122); an unidentified cation has been detected in thymidine phosphorylase crystals and is thought to be involved in enzyme function [34].

Other PP-clusters

There were 16 putative PP-clusters composed only of COGs that had at best only a very generic functional prediction ('putative hydrolase') or none at all. These clusters may represent pathways and systems that we still have to discover.

Finally, a distinct type of PP-clusters is recovered by two of the distances we used in this study, $d_{|r|}$ and d_{r2}. Both of these distances approach zero not only when two patterns are similar, but also when they are close to complementarity. This can indicate mutual exclusion between two COGs, as often observed with non-orthologous gene displacements [14]. We found 12 PP-clusters that included COGs with complementary patterns (Additional data file 2). Some of them represent the well-known pairs of mutually displacing COGs, for example, the two types of thymidylate synthases (COG0207 and COG1351), two ribose 5-phosphate isomerases (COG0120 and COG0698), or two classes of lysyl-tRNA synthetases (COG1190 and COG 1384). Other PP-clusters of that type seem to predict previously unknown gene displacements, such as the HD-superfamily phosphohydrolase (COG1078), probably substituting for some function of the Holiday junction resolvase RuvC (COG0817). Thus, the diametric distance is not only the most sensitive distance measure for PP-cluster definition, but it also has an advantage of finding some gene displacements. We expect to detect many more as our methodology of pattern comparison improves.

Hierarchical clustering decomposes pathways and systems into blocks of genes with tight co-inheritance

One result of this work is that, whatever we tried, most of the PP-clusters recovered only fragments of the known pathways and functional complexes. This fragmentation affects all classes of processes - biosynthesis and degradation of all classes of molecules, signal transduction, cell division, and so on - and was especially evident in the case of long biosynthetic pathways. Indeed, of the 52 pathways represented among the PP-clusters, only MEP pathway, lipid A biosynthesis and the aerobic branch of cobalamine biosynthesis were completely covered by one specific PP-cluster each (Additional data file 2), whereas most of the other pathways were distributed among two, three or four PP-clusters, and some of their components may not be included in any PP-cluster at all.

One reason for this fragmentation may be a rigid hierarchical clustering procedure, which forces each COG into a cluster once and for all. For example, the path of riboflavin biosynthesis was split between PP-cluster 211 and PP-cluster 220, and the latter cluster also included the components of two pathways for biosynthesis of several different amino acids; there are no obvious links between biosynthesis of all those compounds (unless one resorts to the general arguments of

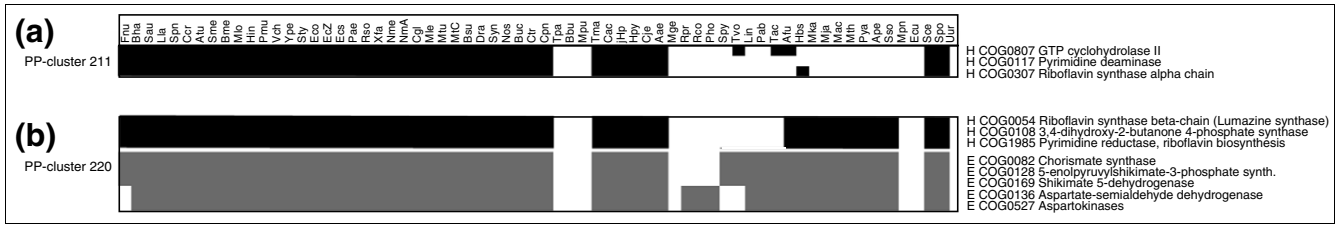

Figure 4
Fragmentation of riboflavin biosynthesis. (**a**) PP-cluster 211 contains the volatile part of riboflavin biosynthesis that is mostly missing in archaea (COGs 0307, 0117, 0807). (**b**) PP-cluster 220 contains the evolutionary most conservative part of the pathway (COGs 1985, 0108, 0054). Gray shading indicates enzymes in PP-cluster 220, unrelated to riboflavin biosynthesis.

carbon-pool availability). Because a COG cannot be included in more than one PP-cluster, there is also a possibility of COG misplacement, which may happen more readily in the case of the larger COGs that include paralogs with different functions. This phenomenon deserves further investigation.

At least in some cases, however, the fragmentation of pathways into PP-clusters seems also to reflect different functional roles and evolutionary fates of COGs within the same pathway. Indeed, further inspection of the split between the components of riboflavin pathway (Figure 4) indicates that PP-cluster 211 contains the components of the pathway that are missing from most archaea (archaeal protein with riboflavin synthase activity [35] belongs to COG1731, which appears to be a distant paralog of COG0054; A.R.M., unpublished data). COGs 1985, 0108 and 0054 (PP-cluster 220) define the evolutionarily most conserved core of the pathway, whereas the entrance into it (COGs 0807 and 0117), as well as the last step, enabled by COGs 0307 or 1731, but also known to occur spontaneously [36], are more variable and prone to gene displacements.

In another example, the bacterial type IV secretion apparatus came out as four PP-clusters, one of which (PP-cluster 067) consisted of genes *virB8*, *virB9*, *virB10* and *virB4* (the names are from the operon involved in transfer of plasmid DNA in *Agrobacterium*). Recent studies indicate that the *virB7-virB8-virB9-virB10* subset of the VirB operon is indeed a module sufficient for DNA uptake by the recipient, but some of the VirB1-VirB4 components are additionally required for maximum recipient activity [37].

These two examples may represent two facets of pathway decomposition into PP-clusters. In the case of the type IV secretion apparatus, at least some of the components of the system appear to represent a functionally and perhaps structurally discrete subsystem, which may be inherited semi-autonomously and retain its own phyletic pattern. In the case of riboflavin biosynthesis, evolutionary variation at the first step of the pathway remains unexplained, while a non-orthologous gene displacement appears to have perturbed the phyletic pattern of the last step.

Discussion

Here we have examined the quantitative aspects of deducing functional links between proteins on the basis of their simultaneous presences and absences in completely sequenced genomes. Whereas the post-homology methods, including definition of operons, multidomain proteins and phyletic patterns, work quite well when combined with each other [8-10,28,30], very little is known about the efficiency and limitations of each method. It has been noted that a high 'co-occurrence score' (essentially, the distance between phyletic patterns based on the complement of mutual information) is less indicative of a functional link than chromosomal proximity of genes or translational fusion of domains [28]. We were interested in whether the comparison of phyletic patterns can be improved, in order to detect functional links and to separate them from the phylogenetic signal [2].

One notable result of our investigation is that the use of correlation-based measures, and, in particular, of the diametric distance between patterns, substantially improves recovery of functional links between genes. This choice of distance measure appears to distinguish well between true co-inheritance versus pairs of rare genes whose patterns are dominated by zeros. Moreover, diametric distance groups not only patterns that are close to identity, but also those that are close to complementarity, thus helping to detect gene displacements. We focused on the algorithms producing the hierarchical trees, that is, directed acyclic graphs. Other, non-hierarchical types of graph have also been used to represent the relationships between proteins; for example, pairwise linkage graphs with scale-free properties have been used to describe the network of protein-protein interactions [38] and the space of protein structures [39]. Some of these approaches may complement our pattern-clustering procedure, and different types of graphs may discover different subsets of functional links.

Another result of our study is the evidence that the co-inheritance of functionally linked genes is constantly perturbed by differential gains, losses and displacements of orthologous genes. This volatility of phyletic patterns reflects the high plasticity and rapid evolution of gene content in microbial

genomes [40] and calls for improving the techniques for phyletic pattern comparison. When this manuscript was under revision, Snel and Huynen [41] reported a similar set of observations of perturbation of gene co-inheritance in microbial evolution. It did not escape our attention that the two-dimensional image of clustered patterns is similar to the now-familiar presentation of whole-genome gene-expression arrays, and that our PP-cluster discovery process is akin to inferring functional links from co-activation and co-inhibition of gene activity. The analysis of gene expression makes extensive use of hierarchical clustering of gene-expression patterns, and many techniques involved will be the same as in the case of phyletic patterns [42]. We note, however, that there is currently no clear quantitative model of the process that produces gene-expression values. In contrast, in our case, phyletic patterns and distances between them can be understood, in quantitative detail, in terms of gene gains and losses in the course of genome evolution [6].

Materials and methods

The data

Gene presences and absences are summarized in the COG database [43]. There were 4,873 COGs from 66 complete genomes of unicellular organisms in the COG database, as of 21 September, 2003 [44]. After exclusion of 284 fungus-specific COGs, we have 3,372 patterns containing one COG and 316 patterns containing two or more COGs, 4,589 COGs in total. Each *i*th COG (i = 1,..., 4,589) is a vector, where the *j*th coordinate (j = 1,..., 66) is set at 1 if it is represented in the *j*th genome, and at 0 if it is not. This vector is equivalent to what has been called 'phylogenetic pattern' in [2] and 'phylogenetic profile' in [3]. We feel 'phyletic' is preferential to 'phylogenetic', because a pattern explicitly tells us what is going on in each phylum, whereas phylogeny of a set of species is not necessarily recoverable from a pattern or even from a set of patterns.

Some COGs contain a mix of orthologs and lineage-specific gene duplications [2]. In some cases, functions of genes within such enlarged COG diverge substantially, which may produce artifacts in the process of functional inference [10]. In our final set of PP-clusters (see Results and Discussion sections for details) there were only 26 (3%) of these 'multifunctional' COGs. An average COG in PP-clusters contained 1.2 genes per species, and 85% of all COGs in the database had less than two genes per species (counting in the denominator only species that had genes in this COG - that is, the 'ones' in the phyletic pattern). The impact of large, functionally heterogeneous COGs on our analysis thus appears to be slight.

The choice of distance measure

The successful discovery of a relationship between phyletic patterns depends on the way the distance and similarity between two pattern vectors $\vec{x},\vec{y} \in \{0,1\}^n$ are measured. We considered a variety of distance measures. These include: l_p norm (that is,

$$d_p = \left[\sum_i |x_i - y_i|^p \right]^{1/p},$$

where p = 1: Manhattan; p = 2: Euclidean; $p = \infty$: Chebyshev distance); Hamming distance, that is, the number of mismatched vector coordinates between two patterns, $d_H = \#(x_i \neq y_i)$; the complement d_{MS} = (1 - J) of Jaccard's similarity index J, which is the cardinality of vectors' intersection divided by the cardinality of their union, $J = \#(x_i \cap y_i)/\#(x_i \cup y_i)$ (this is also known as the Marczewski-Steinhaus distance); the complement of the correlation coefficient, $d_r = 1 - \mathrm{corr}(\vec{x},\vec{y})$, where

$$cor(\vec{x},\vec{y}) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}$$

is the Pearson correlation coefficient; squared anticorrelation, or diametric distance [23], $d_{r2} = 1 - \mathrm{corr}^2(\vec{x},\vec{y})$; absolute anticorrelation distance, $d_{|r|} = 1 - |\mathrm{corr}(\vec{x},\vec{y})|$; mutual information

$$M(i,j) = \sum_{i,j} P_{ij} \log P_{ij} / P_i P_j \text{ [9,45]},$$

where P_i, P_j, P_{ij} are the frequencies of occurrences for, respectively, genes i,j and gene pairs (i,j) in two genomes; Kullback-Leibler (KL) distance and J-divergence. KL is the relative entropy of two probability mass functions $p(x)$ and $q(x)$ over the random variable X

$$d_{KL}(p||q) = \sum p(x) \log_2 \frac{p(x)}{q(x)} \text{ [46]}.$$

The average of two KL distances between two distributions (J-divergence) is symmetric and therefore more applicable for clustering [47].

In the challenge example that we discuss in Results, that of two gene pairs (x_1, y_1) and (x_2, y_2), with patterns x_1 = (1011110), y_1 = (0111110), x_2 = (1000000), y_2 = (0000001), all l_p-norm distances are the same for both pairs: for example, Euclidean, $d_2(x_1, y_1) = d_2(x_2, y_2) = \sqrt{2}$; or Hamming, $d_H(x_1, y_1) = d_H(x_2, y_2) = 2$. J-divergence is zero in both cases. The MI measure distinguishes between the two cases: $M(x_1, y_1)$ = 0.019 and $M(x_2, y_2)$ = 0.010. The difference, however, is more pronounced in the case of correlation distance d_r (0.3 and -0.16, respectively). The d_{r2} and $d_{|r|}$ distances also readily

distinguish between these cases, as well as Jaccard's similarity index ($J(x_1, y_1) = 0.5, J(x_2, y_2) = 0$). Note that, while all distances equal zero for two identical phyletic patterns, only the squared correlation and the absolute anticorrelation distances also equal zero for two complementary patterns. This is a useful property when one wants to look for gene displacements (see Results and Discussion).

Clustering and preliminary partitioning

Algorithms of supervised, parametric or partitional clustering are of limited use for our purpose, because of the lack, respectively, of a well-defined training set, a statistical model of pattern distribution, and the knowledge of underlying cluster number. We studied several algorithms of divisive clustering included in the CLUTO package [24], as well as two standard agglomerative algorithms for hierarchical clustering, familiar from the phylogenetic studies - average linkage (UPGMA) and neighbor joining (NJ) from the PAUP* 4.0b8 package [48]. Agglomerative clustering was the most sensitive and specific, as described in detail in Results. Because the divisive clustering algorithms need an *a priori* fixed number of clusters, we estimated such numbers on the basis on the average number of UPGMA clusters (from 67 to 157, depending on the parameters). The quality of clustering solution, however, was lower for K-means and other divisive algorithms (for example, repeated bisections) than in the case of agglomerative algorithms. Results of all clustering experiments are shown in Tables 1 and 2 in Additional data file 3.

To partition the space of clustered patterns into groups of functionally linked proteins, we used the cutoffs derived from comparing similarity between random patterns, as well as between the functionally linked ones. In each phyletic pattern, all ones and zeros were randomly shuffled to destroy the existing correlations. The figures in Additional data file 4 show distributions of 10,527,166 correlation coefficients among phyletic patterns and shuffled phyletic patterns from the COGs database. Among shuffled patterns, 99% of correlation coefficients were below 0.3, corresponding to the distance $d_r = 0.7$. Therefore, if we choose this distance value as a threshold, the probability that two uncorrelated patterns have a correlation coefficient more than 0.3 is less than 1%. At this threshold, however, only several huge clusters can be found. In another test, we inferred the similarity threshold from the distribution of correlation coefficients among original non-shuffled patterns. The distribution of all pairwise correlation coefficients among original patterns does not differ significantly from the normal distribution (Figure A in Additional date file 4, $\chi^2 = 1.34$) and 99% of correlation coefficients are below 0.8, corresponding to the distance $d_r = 0.2$ and $\bar{b} = 0.1$ (average branch length in cluster). At this threshold, only about 30% of the entire dataset was included in the clusters (see Table 1 in Additional data file 3).

The quality of a clustering solution

At the first step, in order to estimate the quality of each clustering solution, we introduced three empirical indices: the 'group homogeneity' (GrH), 'functional homogeneity' (FunH), 'uncertainty' (Unc), and percentage of data lost. The first two indices indicate the percentage of COGs from the same group/functional category in the cluster (we used definitions of groups and functional categories from the COGs database [27]). The Unc is computed as the percentage of poorly characterized COGs in the cluster. Statistical properties of the cluster were evaluated using three other indices, namely 'consistency' (Cons), 'average distance between cluster members' (AveD) and 'in-cluster variance' (Var) (see Additional data file 1 for computational details). For best functional parsing of the metabolic map, GrH_{Max}, $FunH_{Max}$ and $Cons_{Max}$ as well as Unc_{Min}, $AveD_{Min}$ and Var_{Min} should be found. In practice, these measures are highly correlated, for example, lower $AveD_{Min}$ is, the higher $FunH_{Max}$ is (Table 1 in Additional data file 1). Moreover, most of these indices were almost the same in all clustering solutions. The only exception was the percentage of data lost, which showed about 10% difference between solutions (Figure 3a).

The other measure of quality of a clustering solution is its sensitivity, which is the proportion of COGs from the same pathway or functional category, included in the same cluster. This measure was strongly dependent on the distance and clustering algorithm (Table 2 in Additional data file 3). Diametric distance d_{r2} tends to simultaneously minimize data loss and recovers the largest number of statistically significant clusters (Figure 3a and see also Table 1 in Additional data file 3), most likely because the square of correlation decreases its value, thus increasing the allowed distance between patterns.

The information content of a pathway $I_p = H_r - H_p$, where H_p is the sum over uncertainties of every position in patterns in a pathway:

$$H_p = -\sum_j \sum_i f_i^j \ln f_i^j$$

($j = 1,..,66$, $i = 0$ or 1). The frequency f_i^j stands for patterns 'support' for jth species, $f_0^j + f_1^j = 1$ [49]. H_r is computed similarly, but for reshuffled patterns.

Additional data files

The following additional files are included with the online version of this paper. Additional data file 1 is a figure showing correlations between the percentage of correctly predicted pathway and its information content (Additional data file 1). Additional data file 2 is a list of PP-clusters describing (1) functional predictions and gene displacements and (2) functionally linked clusters of genes, PP-clusters (Additional data file 2). Additional data file 3 contains tables describing the

results of clustering experiments: Table 1 shows the values of classification quality indices for UPGMA/NJ algorithms with different distance measures and Table 2 the performance of UPGMA/NJ algorithms with different distance measures (Additional data file 3). Additional data file 4 is a figure showing the distributions of correlation coefficients between phyletic patterns. The distributions of 10,527,166 correlation coefficients and modified correlation coefficients between original (red bars) and shuffled (blue bars) phyletic patterns from COGs database are shown (Additional data file 4).

References

1. Fitch WM: **Homology: a personal view on some of the problems.** *Trends Genet* 2000, **16:**227-231.
2. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637.
3. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.
4. Smit A, Mushegian A: **Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway.** *Genome Res* 2000, **10:**1468-1484.
5. Kaneda K, Kuzuyama T, Takagi M, Hayakawa Y, Seto H: **An unusual isopentenyl diphosphate isomerase found in the mevalonate pathway gene cluster from *Streptomyces* sp. strain CL190.** *Proc Natl Acad Sci USA* 2001, **98:**932-937.
6. Rohdich F, Kis K, Bacher A, Eisenreich W: **The non-mevalonate pathway of isoprenoids: genes, enzymes and intermediates.** *Curr Opin Chem Biol* 2001, **5:**535-540.
7. Reader JS, Metzgar D, Schimmel P, De Crecy-Lagard V: **Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine.** *J Biol Chem* 2004, **279:**6280-6285.
8. Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D: **Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach.** *Genome Biol* 2003, **4:**R59.
9. Date SV, Marcotte EM: **Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages.** *Nat Biotechnol* 2003, **21:**1055-1062.
10. von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P: **Genome evolution reveals biochemical networks and functional modules.** *Proc Natl Acad Sci USA* 2003, **100:**15428-15433.
11. Mirkin BG, Fenner TI, Galperin MY, Koonin EV: **Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes.** *BMC Evol Biol* 2003, **3:**2.
12. Huynen MA, Dandekar T, Bork P: **Variation and evolution of the citric-acid cycle: a genomic perspective.** *Trends Microbiol* 1999, **7:**281-291.
13. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12:**334-336.
14. Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U: **An alternative flavin-dependent mechanism for thymidylate synthesis.** *Science* 2002, **297:**105-107.
15. Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B, Bork P: **Systematic discovery of analogous enzymes in thiamin biosynthesis.** *Nat Biotechnol* 2003, **21:**790-795.
16. Zheng Y, Roberts RJ, Kasif S: **Genomic functional annotation using co-evolution profiles of gene clusters.** *Genome Biol* 2002, **3:**research0060.1-0060.9.
17. Wu J, Kasif S, DeLisi C: **Identification of functional links between genes using phylogenetic profiles.** *Bioinformatics* 2003, **19:**1524-1530.
18. Liberles DA, Thoren A, von Heijne G, Elofsson A: **The use of phylogenetic profiles for gene predictions.** *Curr Genomics* 2002, **3:**131-138.
19. Vert JP: **A tree kernel to analyse phylogenetic profiles.** *Bioinformatics* 2002, **18 Suppl 1:**S276-S284.
20. Marcotte EM, Xenarios I, van Der Bliek AM, Eisenberg D: **Localizing proteins in the cell from their phylogenetic profiles.** *Proc Natl Acad Sci USA* 2000, **97:**12115-12120.
21. Enault F, Suhre K, Abergel C, Poirot O, Claverie JM: **Annotation of bacterial genomes using improved phylogenomic profiles.** *Bioinformatics* 2003, **19 Suppl 1:**I105-I107.
22. Strong M, Graeber TG, Beeby M, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps.** *Nucleic Acids Res* 2003, **31:**7099-7109.
23. Dhillon IS, Marcotte EM, Roshan U: **Diametrical clustering for identifying anti-correlated gene clusters.** *Bioinformatics* 2003, **19:**1612-1619.
24. Zhao Y, Karypis G: **Evaluation of hierarchical clustering algorithmsfor document datasets.** [http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/vhcluster2.pdf].
25. Herrero J, Valencia A, Dopazo J: **A hierarchical unsupervised growing neural network for clustering gene expression patterns.** *Bioinformatics* 2001, **17:**126-136.
26. Wolf YI, Rogozin IB, Grishin NV, Koonin EV: **Genome trees and the tree of life.** *Trends Genet* 2002, **18:**472-479.
27. **COGs database: pathways and functional systems** [http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?sys=all]
28. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins.** *Nucleic Acids Res* 2003, **31:**258-261.
29. Mushegian A: **The minimal genome concept.** *Curr Opin Genet Dev* 1999, **9:**709-714.
30. Koonin EV, Galperin MY: *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics* Norwell, MA: Kluwer Academic Publishers; 2003.
31. Makarova KS, Koonin EV: **Comparative genomics of archaea: how much have we learned in six years, and what's next?** *Genome Biol* 2003, **4:**115.
32. Fatica A, Tollervey D: **Making ribosomes.** *Curr Opin Cell Biol* 2002, **14:**313-318.
33. Mushegian AR: **Evolution and function of processosome, the complex that assembles ribosomes in eukaryotes: clues from comparative sequence analysis.** *Prog Nucl Acids Mol Biol* 2004 in press.
34. Pugmire MJ, Ealick SE: **Structural analyses reveal two distinct families of nucleoside phosphorylases.** *Biochem J* 2002, **361:**1-25.
35. Eberhardt S, Korn S, Lottspeich F, Bacher A: **Biosynthesis of riboflavin: an unusual riboflavin synthase of *Methanobacterium thermoautotrophicum.*** *J Bacteriol* 1997, **179:**2938-2943.
36. Bacher A, Eberhardt S, Fischer M, Kis K, Richter G: **Biosynthesis of vitamin b2 (riboflavin).** *Annu Rev Nutr* 2000, **20:**153-167.
37. Liu Z, Binns AN: **Functional subsets of the virB type IV transport complex proteins involved in the capacity of *Agrobacterium tumefaciens* to serve as a recipient in virB-mediated conjugal transfer of plasmid RSF1010.** *J Bacteriol* 2003, **185:**3259-3269.
38. Rzhetsky A, Gomez SM: **Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome.** *Bioinformatics* 2001, **17:**988-996.
39. Dokholyan NV, Shakhnovich B, Shakhnovich EI: **Expanding protein universe and its origin from the biological Big Bang.** *Proc Natl Acad Sci USA* 2002, **99:**14132-14136.
40. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content.** *Genome Res* 2002, **12:**17-25.
41. Snel B, Huynen MA: **Quantifying modularity in the evolution of biomolecular systems.** *Genome Res* 2004, **14:**391-397.
42. Datta S, Datta S: **Comparisons and validation of statistical clustering techniques for microarray gene expression data.** *Bioinformatics* 2003, **19:**459-466.
43. **Clusters of Orthologous Groups (COGs)** [http://www.ncbi.nlm.nih.gov/COG/new]
44. Tatusov RL, Fedorova ND, Jackson JJ, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al.*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
45. Huynen M, Snel B, Lathe W 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10:**1204-1210.
46. Cover TM, Thomas JA: *Elements of Informational Theory* New York: Wiley; 1991.

47. Johnson DH, Sinanovic S: **Symmetrizing the Kullback-Leibler Distance.** [http://cmc.rice.edu/docs/docs/Joh2001Mar1Symmetrizi.pdf].
48. Swofford DL: *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4* Sunderland, MA: Sinauer Associates; 2000.
49. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: **Information content of binding sites on nucleotide sequences.** *J Mol Biol* 1986, **188:**415-431.