Introductory Review

Phylogenetic profiling

Matteo Pellegrini, Todd O. Yeates and David Eisenberg

Howard Hughes Medical Institute, University of California at Los Angeles, Los Angeles, CA, USA

Sorel T. Fitz-Gibbon

IGPP Center for Astrobiology, University of California at Los Angeles, Los Angeles, CA, USA

1. Introduction

Biology has been profoundly changed by the development of techniques to sequence DNA. The advent of rapid sequencing in conjunction with the capability to assemble sequence fragments into complete genome sequences enables researchers to read and analyze entire genomes of organisms. Parallel progress has been made in algorithms to study the evolutionary history of proteins. The techniques rely on the ability to measure the similarity of protein sequences in order to determine the likelihood that different proteins are descended from a common ancestor. It is therefore possible to reconstruct families of proteins that share a common ancestor.

Combining these two capabilities, we can now not only determine which proteins are coded within an organism's genome but we can also discover the evolutionary relationships between the proteins of multiple organisms. Phylogenetic profiling is the study of which protein types are found in which organisms.

In order to perform phylogenetic profiling, one must first establish a classification of proteins into families. An example of such a classification scheme across a broad range of fully sequenced organisms is the Clusters of Orthologous Groups (Tatusov, 1997), where an attempt is made to group together proteins that perform a similar function. Next, each organism is described in terms of which protein families are coded or not coded in its genome.

As we will see in this review, this simplified representation is useful for exploring the evolutionary history of an organism as well as for studying the function of protein families and how they may be related to observable phenotypes.

2. Genome phylogeny

Species phylogenies have traditionally been constructed by measuring the evolutionary divergence in a particular family of proteins or RNAs (Fitch, 1967). The most commonly used sequence for such phylogenetic reconstructions is that of the small subunit ribosomal RNA. The advantages of using this RNA gene are that it is found in all organisms, and it has evolved relatively slowly, thus permitting the construction of phylogenies between distant organisms.

Access to the complete genomes of organisms offers a new approach to phylogenetic reconstruction. Rather than looking at the evolution of a single protein or RNA family, it is now possible to compare the gene content of two organisms. This general approach to phylogenetic reconstruction has been applied in a variety of ways (Fitz-Gibbon, 1999; Snel, 1999; Tekaia, 1999; Lin, 2000; Montague, 2000; Wolf, 2001; Bansal, 2002; Clarke, 2002; House, 2002; Li, 2002).

Several metrics have been used to measure the similarity of two organisms on the basis of their gene contents, including the percentage of genes shared by the two species. Furthermore, phylogenetic trees may be reconstructed using several techniques including distance-based phylogenies and parsimony. In general, the trees constructed using whole genome comparisons are similar to those using small subunit rRNA sequences, with occasional discrepancies of interest (Figure 1).

3. Coevolution of protein families

Before fully sequenced genomes became available, the computational study of protein function relied entirely on the detection of sequence similarity. The general notion upon which these studies are based is that proteins with detectable sequence similarity are likely to have evolved from a common ancestor and thus by definition are homologs. Furthermore, such proteins are likely to have preserved common structure and function. Therefore, similarity detection may be used to assign a putative structure and function to proteins that have a sufficient degree of sequence similarity to an experimentally characterized protein. The definition of "sufficient degree" of similarity has been at the center of much research. Depending on the methodology used to determine sequence similarity, various statistical tests have been devised to determine whether two proteins have truly evolved from a common ancestor.

Although techniques based on sequence similarity are powerful, they are unable to inform us about a possible structure or function of a protein family that does not contain experimentally characterized members. This is a significant limitation because a large fraction of all protein families currently fall within this category. Phylogenetic profiling may be used to address this problem, and give us at least partial functional information on these protein families by determining the pathway or complex to which a protein belongs.

Unlike the application of phylogenetic profiling to genome phylogeny where we were interested in measuring the similarity of organisms based on their profile of gene families, here we wish to measure the similarity between the profiles of the families themselves. To accomplish this, we measure the co-occurrence or coabsence of pairs of protein families across genomes (see Figure 2). The underlying

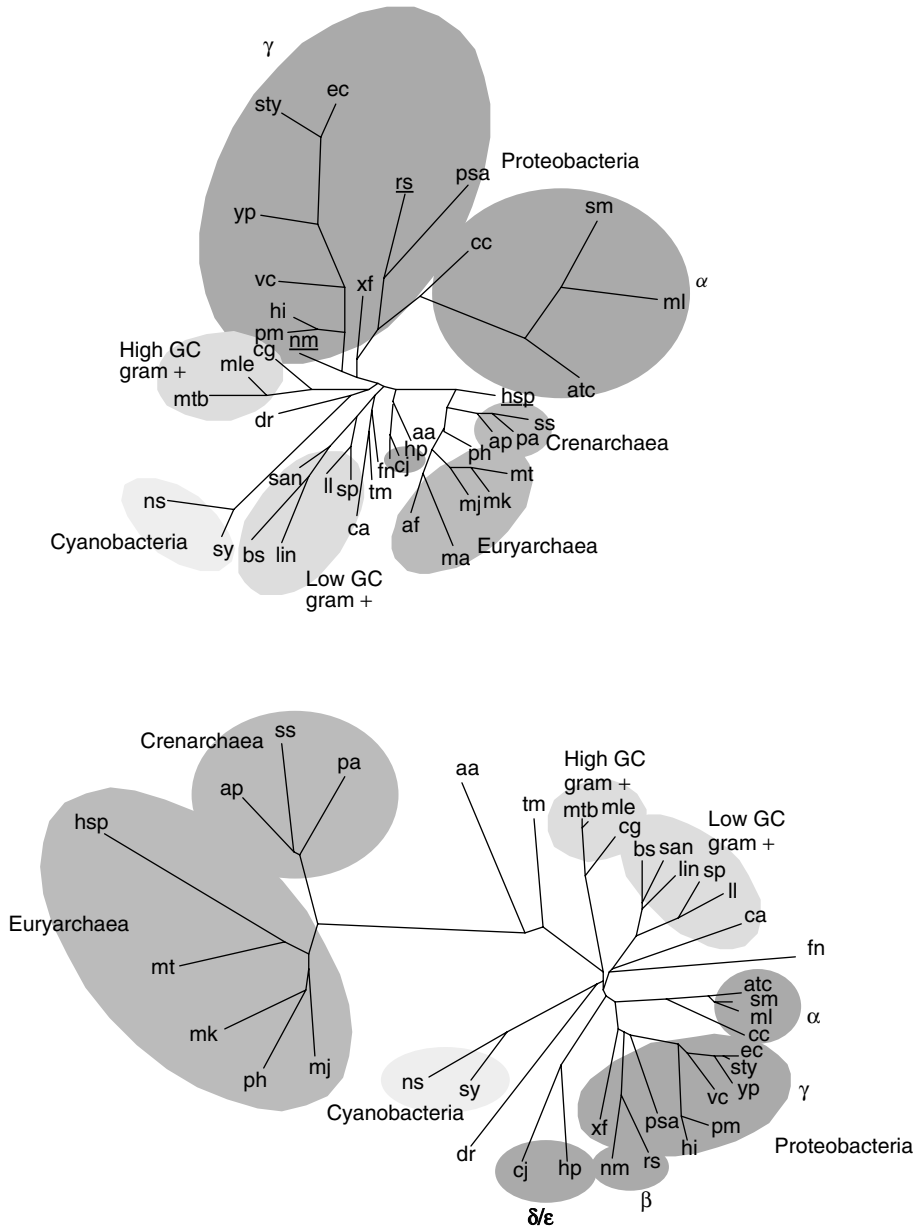

Figure 1 Phylogenetic trees of prokaryotes, based on gene content (upper tree; House, 2002), and small subunit ribosomal RNA sequence (lower tree), constructed using on-line analysis tools at the Ribosomal Database Project (**http://rdp.cme.msu.edu**) (Cole, 2003). A few notable discrepancies are shown in the gene content tree as underlined taxa

Figure 2 Clustered phylogenetic profiles of human HMBS (hydroxymethylbilane synthase), ALDH3 (aldehyde dehydrogenase), and FTHFD (formyltetrahydrofolate dehydrogenase) genes. The profiles are computed over 83 organisms shown on the top. Red indicates that a homolog of the human gene was found in the corresponding organism and black that it was not. The profiles have been clustered using hierarchical clustering (Eisen, 1998)

assumption of this method is that pairs of nonhomologous proteins that are present together in genomes, or absent together, are likely to have coevolved; that is, the organism is under evolutionary pressure to encode both or neither of the proteins within its genome and encoding just one of the proteins lowers its fitness.

It has been observed that coevolved protein families are likely to be members of the same pathway or complex (Huynen, 1998; Pellegrini, 1999). This is not surprising since it is more efficient for an organism to retain all or none of the subunits of a complex, or members of a pathway, since preserving only a fraction of these would not retain the function of the complex or pathway yet would entail their wasteful synthesis. Phylogenetic profiling has therefore emerged as a powerful method to group proteins together into cellular complexes and pathways.

Notice that protein families clustered on the basis of their phylogenetic profiles need not possess any sequence similarity. Therefore, phylogenetic profiling is able to determine functions for proteins families with no experimentally characterized members, thus going beyond the capabilities of conventional sequence similarity–based techniques.

4. Computing phylogenetic profiles

To compute phylogenetic profiles for each protein coded within a genome, one can use several approaches. One of these is to first define orthologous proteins across genomes. Orthologs are proteins that have descended from a common ancestor by way of speciation. Although the actual calculation of orthologs is not trivial, an estimate of groups of orthologous proteins has been compiled in the Clusters of Orthologous Groups (COG) database (Tatusov, 1997). Armed with these clusters, a profile may be trivially calculated by enumerating the organisms that are represented in each COG.

Another approach to establishing a phylogenetic profile is to identify homologs of a protein using a sequence alignment technique. Along these lines, a popular method is to define a homolog of a query protein to be present in a secondary genome if the alignment, using BLAST (Altschul, 1997), of the query protein

with any of the proteins encoded by the secondary genome generates a significant alignment. The result of this calculation across N genomes yields an N-dimensional phylogenetic profile of ones and zeroes for the query protein. At each position in the phylogenetic profile, the presence of a homolog in the corresponding genome is indicated with a 1 and its absence with a 0.

There is no need to restrict phylogenetic profiles to contain only entries of 1's and 0's. Various methods have been used in which the entries of the phylogenetic profile measure the similarity of two proteins. As an example, one method uses the inverse of the log of the E value from a BLAST search as the similarity metric (Date, 2003).

5. Estimating the probability of coevolution

Once the phylogenetic profiles have been computed, one needs to determine the likelihood that two proteins have coevolved on the basis of the similarity of their profiles. A variety of techniques have been reported to compute these probabilities. Here, we briefly review a few of them.

The first approach is the computation of the similarity between two phylogenetic profiles using the Hamming distance (Pellegrini, 1999). The Hamming distance is the number of bits that differ between the two profiles. Although this is a simple measure to compute, it is limited by not providing a probability estimate of observing this distance.

It is possible to obtain such an estimate of the probability that two proteins coevolve by using the hypergeometric distribution. If we assume that the two proteins A and B do not coevolve, we can compute the probability of observing a specific overlap between their two profiles by chance by using the hypergeometric distribution:

$$P(k'|n, m, N) = \frac{\binom{n}{k'}\binom{N-n}{m-k'}}{\binom{N}{m}} \tag{1}$$

where N represents the total number of genomes analyzed, n the number of homologs for protein A, m the number of homologs for protein B, and k' the number of genomes that contain homologs of both A and B (Wu, 2003). Because P represents the probability that the proteins do not coevolve, $1 - P(k > k')$ is then the probability that they do coevolve.

A similar approach attempts to compute the likelihood of coevolution using the mutual information between two phylogenetic profiles (Date, 2003; Wu, 2003):

$$MI(A, B) = H(A) + H(B) - H(A, B) \tag{2}$$

where

$$H(A) = -\sum p(a) \ln p(a) \tag{3}$$

and

$$H(A, B) = -\sum p(a, b) \ln p(a, b) \tag{4}$$

Here, the sums are over the possible states that the profiles can assume. If two profiles are identical, their mutual information is zero. Dissimilar profiles have positive mutual information scores. One advantage of the mutual information approach is that it can be applied to nonbinary phylogenetic profiles, whereas the hypergeometric function cannot.

6. Recovery of pathways and complexes

Protein pairs that coevolve are likely under some evolutionary pressure because their functions are coupled: preserving one without the other disables their combined function. This scenario may occur if the proteins are subunits of cellular complexes or components of pathways.

It is possible to test this hypothesis starting from pathway annotation. Several databases have been developed that through extensive manual curation have categorized proteins into pathways (Tatusov, 2003; Kanehisa, 2004; Camon, 2004).

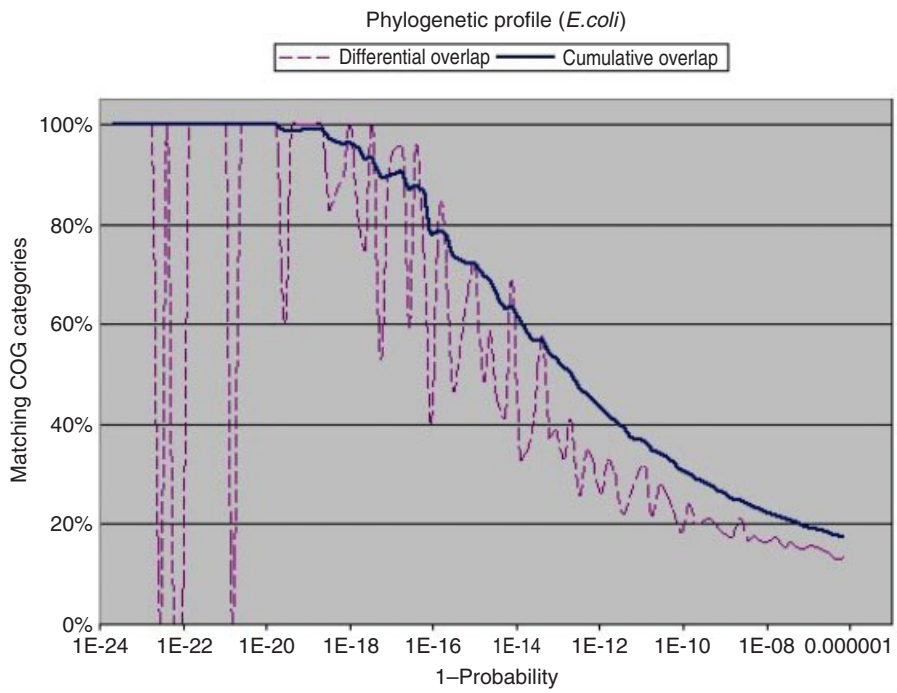

Figure 3 The probability that two genes have coevolved as a function of their likelihood to belong to the same pathway. The probability is computed using the hypergeometric function (see text). The pathways are obtained from the COG databases (Tatusov, 2003). Pairs of genes with significant *P*-values (on left) are nearly always found to belong to the same pathway

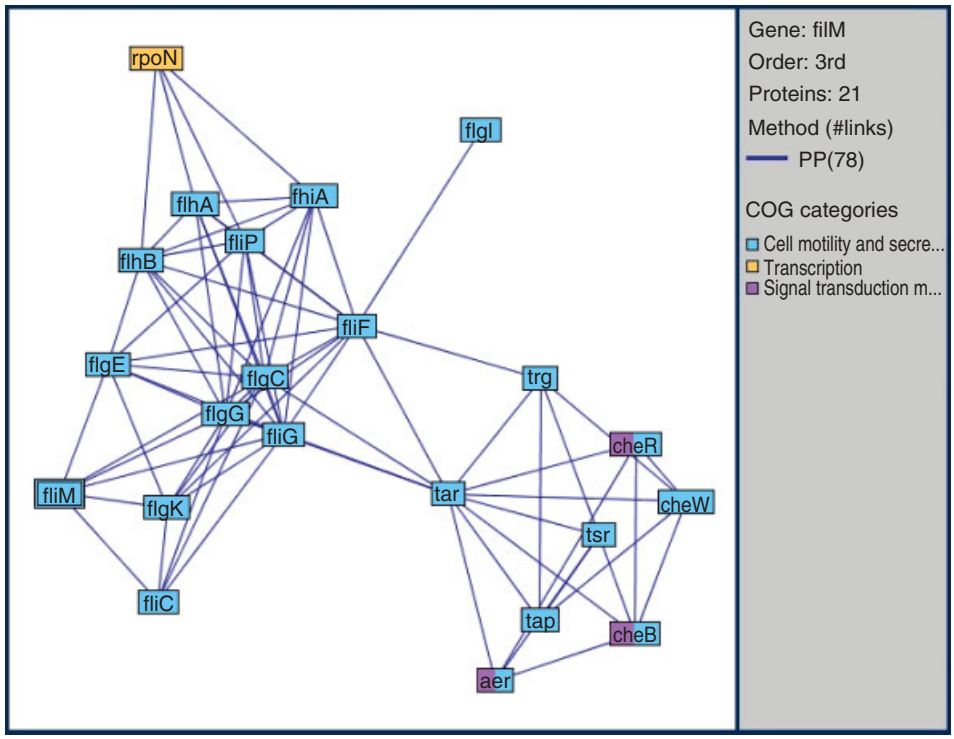

Figure 4 Clusters of *Escherichia coli* proteins that are predicted to coevolve by the phylogenetic profile analysis and that form a large network. The network shows a cluster of proteins (flg and flh genes) that are components of the bacterial flagella. A second cluster includes components of the chemotaxis pathway (che genes). These two clusters are linked to each other, indicating that flagellar and chemotaxis clusters have coevolved in bacteria

In Figure 3, we show that proteins that are likely to have coevolved (have significant *P*-values) are likely to belong to the same pathway (using the COG pathway definitions, Tatusov, 2003). In fact, we find that protein pairs with significant *P*-values nearly always belong to the same pathway. A similar curve could also be constructed using protein complexes instead of pathways, yielding similar results (Bowers, 2004).

By combining all pairs of coevolving proteins with significant *P*-values, we can generate a vast network. This is because if protein A is found to coevolve with B, and is thus said to be functionally linked to B, B may then be linked to C, C to D, and so forth. By examining clustered groups of proteins within this network, one can identify the protein components of pathways and complexes (Strong, 2003; Von Mering, 2003). An example of such a network is shown in Figure 4. Here, we see that many of the components of the flagella form a cluster, as do the components of the chemotaxis pathway. Furthermore, the network also illuminates the fact that these two clusters are coevolving. This is not surprising given the intimately coupled function of flagella and chemotaxis within the cell.

7. Phenotype profiling

We have discussed the use of phylogenetic profiling to study the evolution of genomes and to study the coevolution of encoded proteins, yielding functional clusters and networks of clusters. A third application we review is the linking of genes to phenotypes (Jim, 2004; Levesque, *et al*., 2003).

Each of the fully sequenced organisms that is used to construct phylogenetic profiles of a gene has specific phenotypes. A phenotype is any observable characteristic of the organism. Examples of phenotypes include flagella, pili, and thermosensitivity. It is possible to construct a phenotypic profile by cataloging the presence or absence of the phenotype across genomes, just as we have done for the presence or absence of genes.

By identifying the genes whose phylogenetic profiles are correlated with the phenotypic profiles, it is possible to associate a gene with the phenotype. For instance, about half of the fully sequenced organisms contain flagella. The genes whose phylogenetic profiles are correlated with a flagella profile are nearly all known components of the bacterial flagella (Levesque, 2003; Jim, 2004). The same approach may also be used to identify the components of pili, and the proteins that endow organisms with thermostability (Jim, 2004). In general, if a reliable phenotypic profile can be constructed for a trait that is found in a significant fraction of the sequenced genomes, this technique can identify the proteins that are most likely responsible for the trait.

8. Conclusions

The availability of fully sequenced genomes has enabled us to perform phylogenetic profiling by identifying the distribution of protein families across organisms. As we have discussed in this review, phylogenetic profiling may be used to study the evolution of genomes, the coevolution of proteins or the association between proteins and phenotypes.

Today, we have access to about 100 fully sequenced genomes. However, it is reasonable to assume that within the next decade this number will grow by orders of magnitude. As the data become available, phylogenetic profiling will become far more powerful than it is today. As a result, phylogenetic profiling will undoubtedly continue to expand our understanding of genome evolution and protein function.

Acknowledgments

We thank DOE, NIH, and Howard Hughes Medical Institute for support.

Further reading

Gaasterland T and Ragan MA (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microbial & Comparative Genomics*, **3**(4), 199–217.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Bansal AK and Meyer TE (2002) Evolutionary analysis by whole-genome comparisons. *Journal of Bacteriology*, **184**(8), 2260–2272.

Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO and Eisenberg D (2004) ProLinks: a database of protein functional linkages derived from coevolution. *Genome Biology*, **R35**, Epub 2004 Apr. 16.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R and Apweiler R (2004) The gene ontology annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Research*, **32**(1), D262–D266.

Clarke GD, Beiko RG, Ragan MA and Charlebois RL (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *Journal of Bacteriology*, **184**(8), 2072–2080.

Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, *et al*. (2003) The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Research*, **31**(1), 442–443.

Date SV and Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology*, **21**(9), 1055–1062, Epub 2003 August 17.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster Analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Fitch WM and Margoliash E (1967) Construction of phylogenetic trees. *Science*, **155**(760), 279–284.

Fitz-Gibbon ST and House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, **27**(21), 4218–4222.

House CH and Fitz-Gibbon ST (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *Journal of Molecular Evolution*, **54**(4), 539–547.

Huynen MA and Bork P (1998) Measuring genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(11), 5849–5856.

Jim K, Parmar K, Singh M and Tavazoie S (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Research*, **14**(1), 109–115.

Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**(1), D277–D280.

Levesque M, Shasha D, Kim W, Surette MG and Benfey PN (2003) Trait-to-gene: a computational method for predicting the function of uncharacterized genes. *Current Biology*, **13**(2), 129–133.

Li W, Fang W, Ling L, Wang J, Xuan Z and Chen R (2002) Phylogeny based on whole genome as inferred from complete information set analysis. *Journal of Biology Physics*, **28**, 439–447.

Lin J and Gerstein M (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research*, **10**, 808–818.

Montague MG and Hutchison CA (2000) Gene content phylogeny of herpesviruses. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(10), 5334–5339.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(8), 4285–4288.

Snel B, Bork P and Huynen MA (1999) Genome phylogeny based on gene content. *Nature Genetics*, **21**(1), 108–110.

Strong M, Graeber TG, Beeby M, Pellegrini M, Thompson MJ, Yeates TO and Eisenberg D (2003) Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchical clustering of genome wide functional linkage maps. *Nucleic Acids Research*, **31**(24), 7099–7109.

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science*, **278**(5338), 631–637.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**(1), 41.

Tekaia F, Lazcano A and Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Research*, **9**, 550–557.

Von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA and Bork P (2003) Genome evolution reveals biochemical networks and functional modules. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(26), 15428–15433, Epub 2003 December 12.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL and Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology*, **1**, 8.

Wu J, Kasif S and DeLisi C (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, **19**(12), 1524–1530.