Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics

Your unique id g404104

Phylogenetic Profiling

Matteo Pellegrini, Sorel T Fitz-Gibbon, Todd Yeates, David Eisenberg

Howard hughes Medical Institute
Institute for Genomics and Proteomics
Box 951570
UCLA
Los Angeles, CA 90095-1570
Tel: (310) 825-3754
Fax: (310) 206-3914
david@mbi.ucla.edu

Keywords

Phylogentic Profiling, Phylogentic profiles, phylogeny, functional genomics, phenotyipic profiles, comparative genomics

Abstract

Phylogenetic profiling involves the study of the occurrence of gene families across fully sequenced genomes. For any gene, it is possible to create a phylogenetic profile which encodes the presence or absence of homologs of the gene across organisms. The comparison of phylogenetic profiles may be used to measure the evolutionary distance between organisms leading to the reconstruction of phylogenetic trees. The pair-wise comparison of phylogenetic profiles allows us to identify genes that co-occur across organism and are likely to participate within the same pathway or protein complex. Finally, by comparing phylogenetic profiles to phenotypic profiles that encode the presence or absence of phenotypes across organisms, it is possible to infer that a gene is partially responsible for establishing a phenotype. These techniques are currently used to survey about one hundred fully sequenced genomes that are currently determined. However, they promise to become more informative as the number of fully sequenced genomes increases by orders of magnitude.

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 genomes sequences enables researches 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 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 in its genome.

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

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 based on 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).

Co-evolution 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 of 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 co-absence of pairs of protein families across genomes (see figure 2). The underlying assumption of this method is that pairs of non-homologous proteins that are present together in genomes, or absent together, are likely to have co-evolved. 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 co-evolved 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.

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 1999), 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 one and its absence with a zero.

There is no need to restrict phylogenetic profiles to contain only entries of ones and zeroes. 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).

Estimating the Probability of Co-evolution

Once the phylogenetic profiles have been computed, one needs to determine the likelihood that two proteins have co-evolved based on 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 a 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 co-evolve by using the hypergeometric distribution. If we assume that the two proteins **A** and **B** do not co-evolve, we can compute the probability of observing a specific overlap between their two profiles by chance by using the hypergeometric distribution:

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

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 co-evolve, $1 - P(k > k')$ is then the probability that they do co-evolve.

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

$$MI(A,B) = H(A) + H(B) - H(A,B)$$

where

$$H(A) = -\sum p(a) \ln p(a)$$

and

$H(A,B) = -\sum p(a,b) \ln p(a,b)$. 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 non-binary phylogenetic profiles, whereas the hypergeometric function cannot.

Recovery of Pathways and Complexes

Protein pairs that co-evolve 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). In figure 3 we show that proteins that are likely to have co-evolved (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).

Combining all pairs of co-evolving proteins with significant P values we can generate a vast network. This is because if protein A is found to co-evolve 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 co-evolving. This is not surprising given the intimately coupled function of flagella and chemotaxis within the cell.

Phenotype Profiling

We have discussed the use of phylogenetic profiling to study the evolution of genomes and to study the co-evolution 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 2004).

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 cataloguing 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 thermo-stability (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.

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 co-evolution 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.

Acknowledgements

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

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.

Bansal AK, Meyer TE. Evolutionary analysis by whole-genome comparisons. J. Bacteriol. 2002 Apr;184(8), 2260–72.

Bowers PM, Pellegrini P, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. PROLINKS: A Database of Protein Functional Linkages Derived from Co-evolution, Genome Biology, in press.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 2004 Jan 1;32(1):D262-6.

Clarke GD, Beiko RG, Ragan MA, Charlebois RL. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J Bacteriol. 2002 Apr; 184(8): 2072-80.

Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, Tiedje JM. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res 2003 Jan 1;31(1):442-3

Date SV, Marcotte EM. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol. 2003 Sep;21(9):1055-62. Epub 2003 Aug 17.

Fitch WM, Margoliash E. Construction of phylogenetic trees. Science. 1967 Jan 20;155(760):279-84.

Fitz-Gibbon ST, House CH. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. 1999 Nov 1;27(21):4218-22.

Gaasterland T, Ragan MA. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microb Comp Genomics. 1998;3(4):199-217.

House CH, Fitz-Gibbon ST. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J Mol Evol. 2002 Apr;54(4):539-47.

Huynen MA, Bork P. Measuring genome evolution. Proc Natl Acad Sci U S A. 1998 May 26;95(11):5849-56.

Jim K, Parmar K, Singh M, Tavazoie S. A cross-genomic approach for systematic mapping of phenotypic traits to genes. Genome Res. 2004 Jan;14(1):109-15.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004 Jan 1;32(1):D277-80.

Levesque M, Shasha D, Kim W, Surette MG, Benfey PN. Trait-to-gene: a computational method for predicting the function of uncharacterized genes. Curr Biol. 2003 Jan 21;13(2):129-33.

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

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

Montague MG, Hutchison CA. Gene content phylogeny of herpesviruses.
Proc Natl Acad Sci U S A. 2000 May 9; 97(10): 5334-9.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999 Apr 13;96(8):4285-8.

Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content.
Nat Genet. 1999 Jan;21(1):108-10.

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 Dec 15;31(24):7099-109.

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families.
Science. 1997 Oct 24;278(5338):631-7.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003 Sep 11;4(1):41.

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

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 U S A. 2003 Dec 23;100(26):15428-33. Epub 2003 Dec 12.

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

Wu J, Kasif S, DeLisi C. Identification of functional links between genes using phylogenetic profiles. Bioinformatics. 2003 Aug 12;19(12):1524-30.

Figure 1

Phylogenetic trees of prokaryotes, based on gene content (upper tree, House 2002), and small subunit ribosomal RNA sequence (lower tree), constructed using online 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 bold taxa.

Figure 2

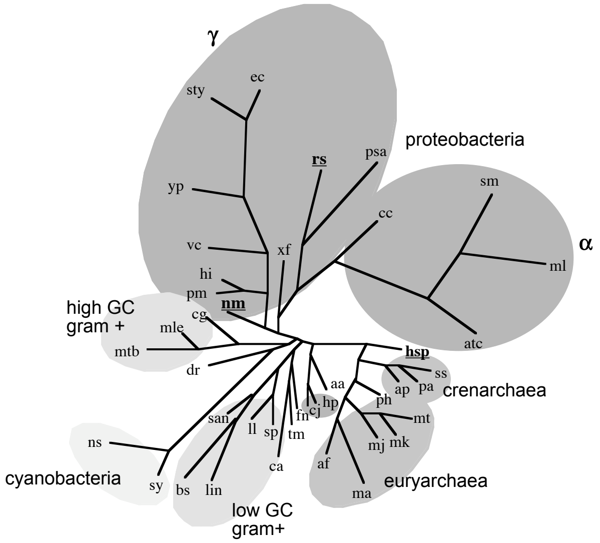
Clustered phylogenetic profiles of human HMBS (hydroxymethylbilane synthase), ALDH3 (aldehyde dehydrogenase) and FTHFD (formyltetrahydrofolate dehydrogenase) genes. The profiles are computed over 83 organisms shown on the top. Red indicates that a homolog of the human gene was found in the corresponding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

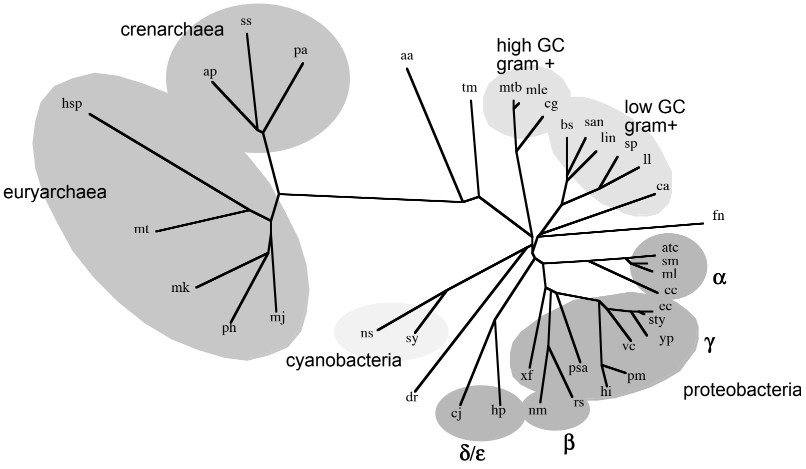
Figure 3

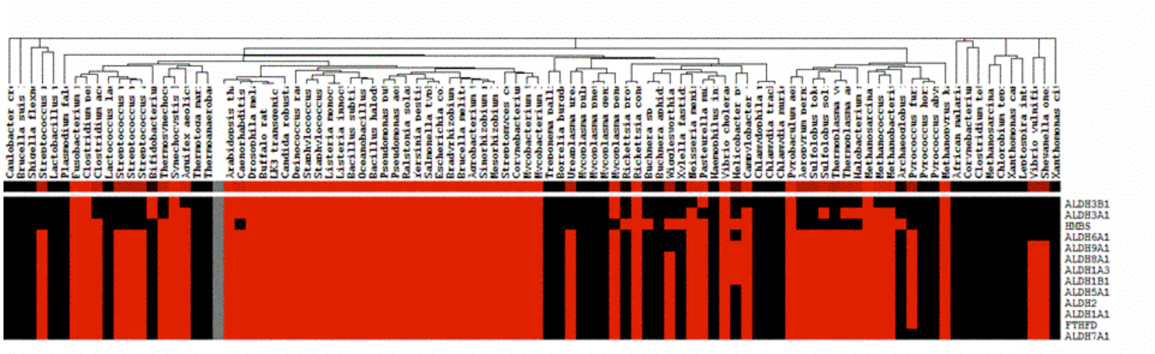
The probability that two genes have co-evolved as a function of their likelihood to belong to the same pathway. The probability is computed using the hypergeometric function (see text). The pathways are obtained from the COG databases (Tatusov 2003). Pairs of genes with significant P values (on left)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

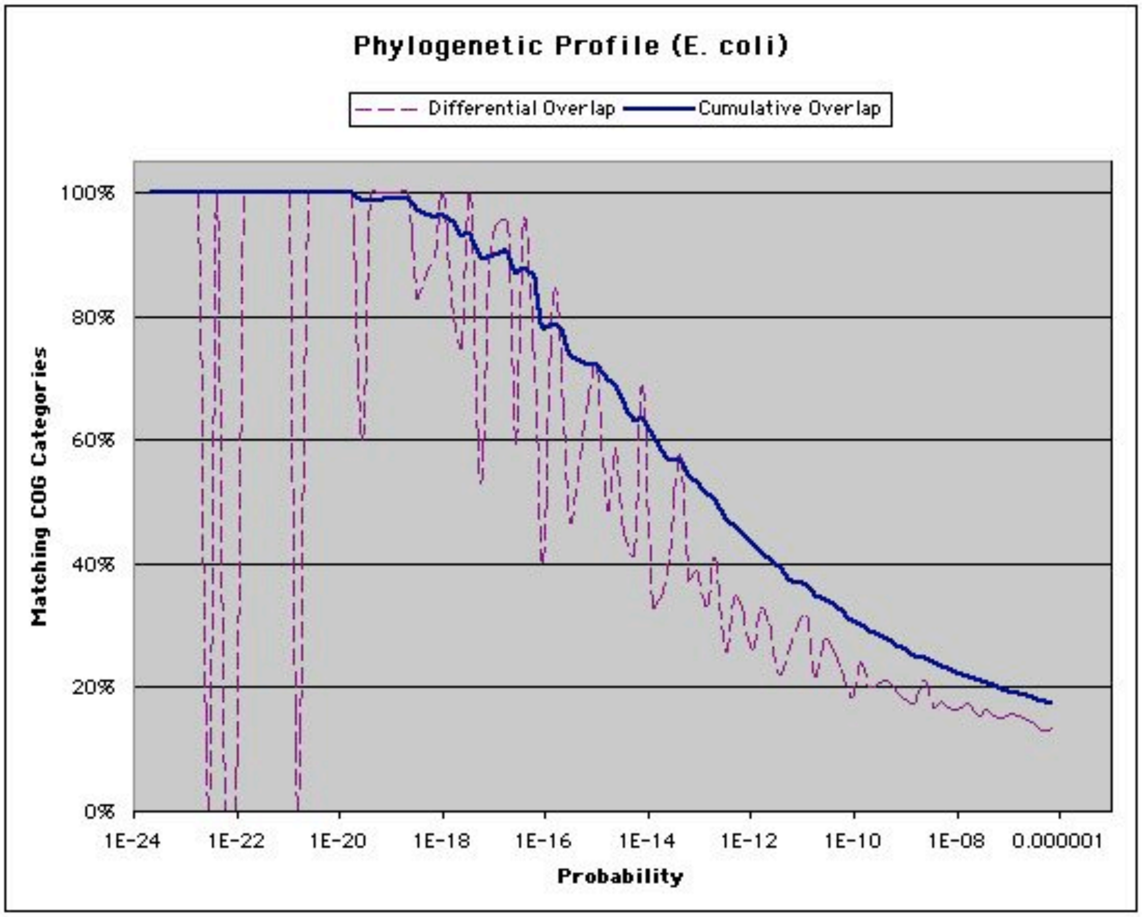
Figure 4

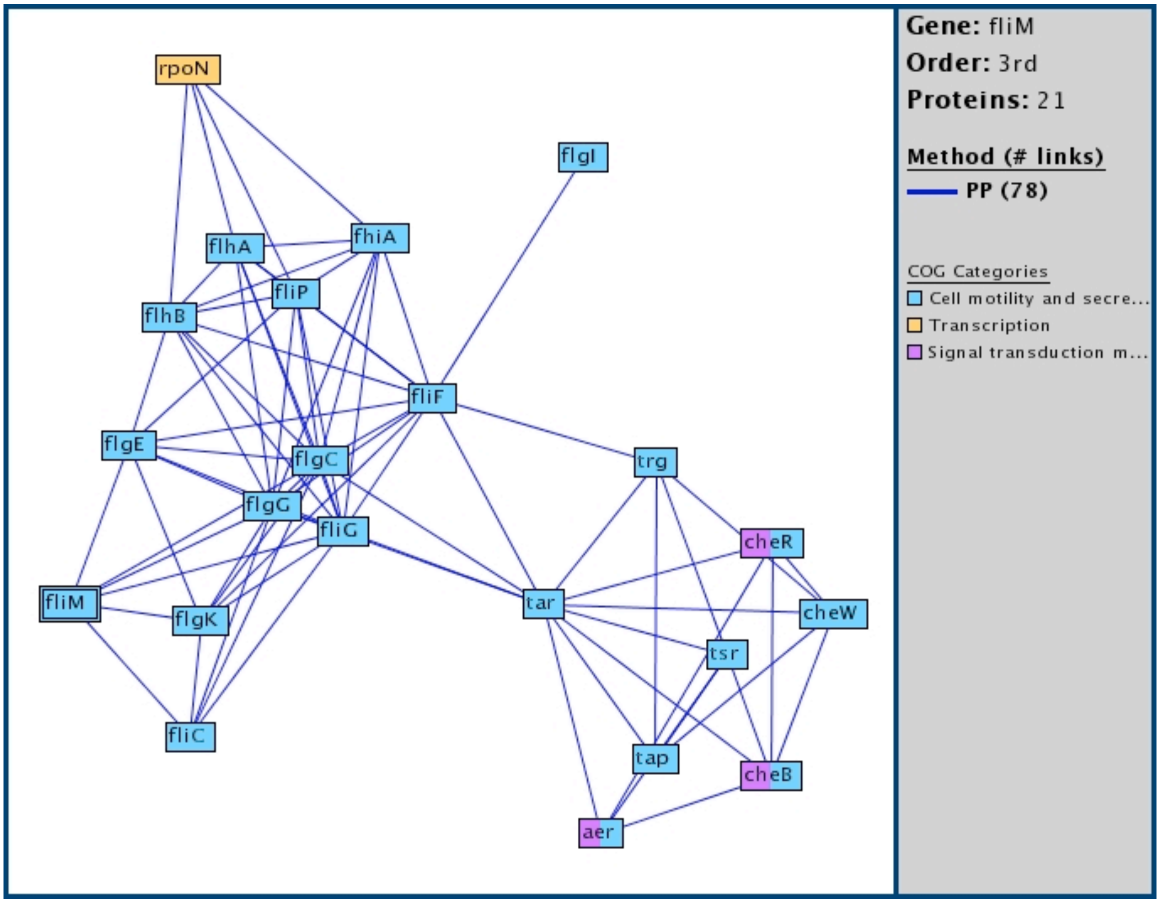
Clusters of *E coli* proteins that are predicted to co-evolve by the phylogenetic profile analysis, and which form a large network. The network shows a cluster of proteins (flg and flh genes) that are components of the bacterial flagella. A second cluster includes components of the chemotaxis pathway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


γ
ec
sty
proteobacteria
rs
psa
sm
yp
cc
α
vc
xf
ml
hi
pm
high GC
gram +
nm
cg
mle
atc
mtb
hsp
dr
ss
aa
crenarchaea
ap
pa
hp
ph
san
cj
ll
fn
mt
sp
tm
ns
mj
mk
ca
af
cyanobacteria
euryarchaea
sy
bs
lin
ma
low GC
gram+


crenarchaea
ss
pa
ap
hsp
euryarchaea
mt
mk
ph
mj
aa
high GC
gram +
tm
mtb
mle
cg
low GC
gram+
bs
san
lin
sp
ll
ca
fn
atc
sm
ml
cc
α
ec
sty
yp
γ
vc
pm
hi
proteobacteria
ns
sy
cyanobacteria
xf
psa
rs
nm
β
dr
cj
hp
δ/ε


Caulobacter cr
Brucella suis
Shigella flexn
Streptococcus
Lactobacillus
Plasmodium fal
Fusobacterium
Clostridium pe
Clostridium ac
Lactococcus la
Streptococcus
Streptococcus
Streptococcus
Bifidobacteriu
Thermosynechoc
Synechocystis
Aquifex aeolic
Thermotoga mar
Thermoanaeroba
Arabidopsis th
Caenorhabditis
Drosophila mel
Buffalo rat
LK3 transgenic
Candida robust
Deinococcus ra
Staphylococcus
Staphylococcus
Listeria monoc
Listeria innoc
Bacillus subti
Oceanobacillus
Bacillus halod
Pseudomonas pu
Pseudomonas ae
Ralstonia sola
Yersinia pesti
Salmonella typ
Escherichia co
Bradyrhizobium
Brucella melit
Agrobacterium
Sinorhizobium
Mesorhizobium
Streptomyces c
Corynebacteriu
Mycobacterium
Mycobacterium
Treponema pall
Borrelia burgd
Ureaplasma ure
Mycoplasma pul
Mycoplasma pne
Mycoplasma gen
Mycoplasma pen
Rickettsia pro
Rickettsia con
Buchnera sp
Buchnera aphid
Wigglesworthia
Xylella fastid
Neisseria meni
Pasteurella mu
Haemophilus in
Vibrio cholera
Helicobacter p
Campylobacter
Chlamydophila
Chlamydia trac
Chlamydia muri
Pyrobaculum ae
Aeropyrum pern
Sulfolobus tok
Sulfolobus sol
Thermoplasma v
Thermoplasma a
Halobacterium
Methanosarcina
Methanococcus
Methanobacteri
Archaeoglobus
Pyrococcus fur
Pyrococcus hor
Pyrococcus aby
Methanopyrus k
African malari
Corynebacteriu
Clostridium te
Methanosarcina
Chlorobium tep
Xanthomonas ca
Leptospira int
Vibrio vulnifi
Shewanella one
Xanthomonas ci
ALDH3B1
ALDH3A1
HMBS
ALDH6A1
ALDH9A1
ALDH8A1
ALDH1A3
ALDH1B1
ALDH5A1
ALDH2
ALDH1A1
FTHFD
ALDH7A1


Phylogenetic Profile (E. coli)
Differential Overlap
Cumulative Overlap
100%
80%
60%
40%
20%
0%
Matching COG Categories
1E-24
1E-22
1E-20
1E-18
1E-16
1E-14
1E-12
1E-10
1E-08
0.000001
Probability



Gene: fliM

Order: 3rd

Proteins: 21

Method (# links)

PP (78)

COG Categories

- 🟦 Cell motility and secre...
- 🟧 Transcription
- 🟪 Signal transduction m...