Journal of Cellular Biochemistry Supplement 37:106–109 (2001)

Computational Method to Assign Microbial Genes to Pathways

Matteo Pellegrini,* Michael Thompson, Joseph Fierro, and Peter Bowers

Protein Pathways, 1145 Gayley Ave, Suite 304, Los Angeles, California

Abstract We present techniques that mine fully sequenced microbial genomes for functional relationships between genes. We show that genes related by one of four techniques are more likely to belong to the same cellular pathways. Furthermore, we demonstrate that the pathway of an uncharacterized gene may be inferred from those of its functionally related partners. Therefore, we are now able to assign most of the genes within bacteria to cellular pathways. J. Cell. Biochem. Suppl. 37: 106–109, 2001. © 2002 Wiley-Liss, Inc.

Key words: pathways; comparative genomics; functional genomics; phylogenetic analysis

During the past few years the ability to sequence an entire microbial genome has become routine. The sequence of pathogens opens the door to the discovery of novel vaccines and therapeutics. One of the primary challenges facing the discovery of these is the identification of the functions of the genes coded by these genomes.

Typically the functions of proteins are assigned by homology. That is, if a new protein sequence is deemed similar to that of a protein whose function has already been assigned, it is likely that they share a similar function. This technique has allowed scientists to assign on average, functions to about half the proteins coded by new genomes. Remaining sequences either have no homologs in protein sequence databases, or are homologous to uncharacterized proteins. The challenge lies in assigning functions to the remainder of these proteins.

We have developed several techniques specifically designed to overcome the function annotation challenges. The techniques we outline below cluster together proteins of related function that are likely to belong to a common cellular pathway. The functional relationships between proteins may be inferred through four techniques that utilize the fully sequenced genomes of all the available organisms to date. Towards the end of the year 2001 this collection constitutes approximately 100 organisms, of which most are prokaryotic, the remainder being archaebacteria and eukaryotes.

The first method that we utilize to study protein pathways involves the reconstruction of phylogeny [Huynen and Bork, 1998; Pellegrini et al., 1999]. Since the accessible genomes allow us to catalogue all the proteins that are expressed, we can determine the pattern of the presence and absence of protein families across organisms. These patterns are termed phylogenetic profiles, and are encoded in the form of bit vectors. A schematic representation of the construction of phylogenetic profiles is shown in Figure 1.

An extension of the phylogenetic profile method considers the position of the genes on the genome [Dandekar et al., 1998; Overbeek et al., 1998, 1999]. In this analysis, for any pair of genes, we calculate the nucleotide distance between homologs of these genes across multiple organisms. For instance, in Figure 2, the A and B genes are found nearby on four genomes, while the C gene is not coded near these in three organisms. Analysis generates a pattern of distances between all pairs of genes coded by a genome. As in the case of phylogenetic profiles, we find that genes that are coded nearby on multiple genomes tend to belong to the same pathway.

An extreme case of the gene distance analysis arises when two genes that are coded as separate proteins in one organism are fused in

*Correspondence to: Matteo Pellegrini, Protein Pathways, 1145 Gayley Ave, Suite 304, Los Angeles, CA.
E-mail: matteope@proteinpathways.com

Received 8 October 2001; Accepted 8 October 2001

DOI 10.1002/jcb.10071

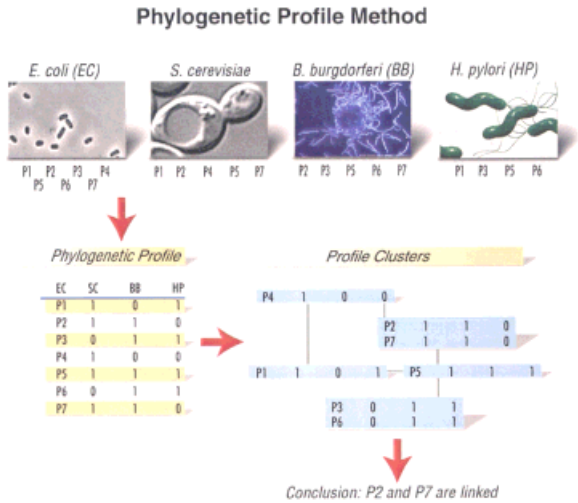

Fig. 1. Flow chart illustrating the construction of phylogenetic profiles. **a**: We begin with four fully sequenced genomes from which the protein sequences have been predicted. **b**: The first sequence, P1, within *E. coli* is compared to that of the proteins coded by the other genomes and homologs are identified. If the genome contains a homolog of P1, a 1 is placed in the corresponding phylogenetic profile position, a 0 otherwise. A similar analysis is performed on all the *E. coli* proteins. **c**: Phylogentic profiles are clustered based on similarity. **d**: Genes with similar phylogenetic profiles are likely to participate in the same pathway.

the other (see Fig. 3). The fusion protein is termed as the "Rosetta Stone" protein, because it allows us to infer that the two original proteins are functionally related [Enright et al., 1999; Marcotte et al., 1999a]. As in previous methods, pairs of proteins that are linked via a "Rosetta Stone" protein typically belong to the same cellular pathway. Often, enzymes within a metabolic pathway are fused to achieve greater catalytic activity.

The last analysis we perform on bacterial genomes is the reconstruction of operons. Genes within the same operon tend to have very short intergenic spacing, while genes at the boundary

Fig. 2. The second method for deducing functional relationships between genes relies on the observation that certain gene pairs are coded nearby on multiple genomes. In the figure we see that genes A and B are coded nearby on all four genomes, while the C gene is not coded near A and B in three out of the four genomes. We therefore conclude that the A and B genes are functionally related to each other but not to the C gene.

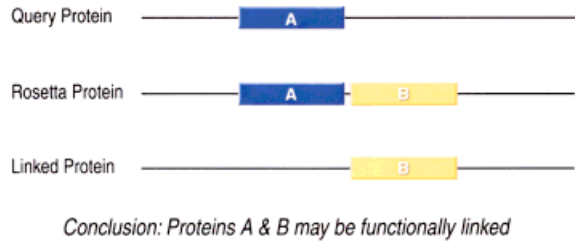

Fig. 3. The Rosetta Stone method searches for gene fusion events. In the figure we see that the A and B proteins are expressed as separate proteins in one organism. However, in a second organism a sequence exists that represents the fusion of the two proteins. The fusion protein is termed the Rosetta Stone protein since it allows us to deduce that the A and B proteins are typically functionally related.

of two operons tend to be further spaced. Thus one can use the intergenic distance to estimate whether two consecutive, co-directional genes are in the same operon or not. Once again, the fact that two or more genes are part of the same operon is an indication that they are likely to be part of the same pathway.

The observation that these four techniques cluster proteins together that are in the same pathway is in some sense not surprising. It seems reasonable to assume that bacterial cells have evolved to efficiently express genes that have related functions by placing them in a clustered fashion on the genome, and that these clusters are conserved to some degree across species.

Since each of these methods is statistical in nature, we can compute a probability associated with each pair-wise relationship. For instance, we can compute the probability of observing a certain similarity between two phylogenetic profiles, by measuring how often this degree of similarity would be found if we randomly assign gene families to organisms. Furthermore, since the *E. coli* genome has been extensively annotated, we can directly measure how often genes related by one of these four methods with a certain probability are found in the same pathway, as seen in Figure 4.

The ability to find pathway associations between all the proteins in a genome allows us to construct genome-wide pathway maps. One way to visualize such maps is by linkage anazlysis. We consider pairs of proteins that have a greater than 50% chance of belonging to the same pathway by the above criteria to be linked. By combining all the links between the proteins in a genome we are able to construct a protein

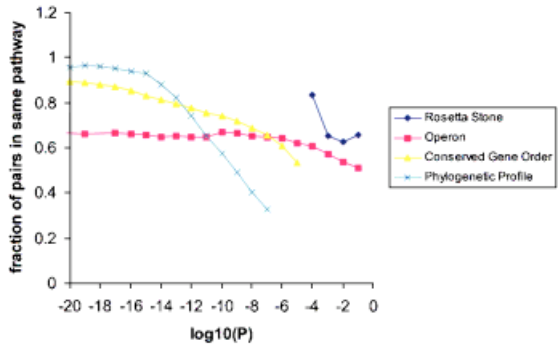

Fig. 4. We have described four methods to associate pairs of proteins that are functionally related: phylogentic profiles, conserved gene order, Rosetta Stone, and operon analysis. We first associate a probability with the pair-wise relationship inferred from each method. This probability represents the likelihood of finding the relationship between the two genes when the genomes are randomized. In the figure we relate the log base 10 of the probability to the likelihood that the two proteins are in the same pathway, as determined by the COG annotation [Tatusov et al., 1997].

function network (see in Fig. 5 the network for Mycoplasma Genitalium). This network is a first attempt to understand a cell on a genome-wide scale.

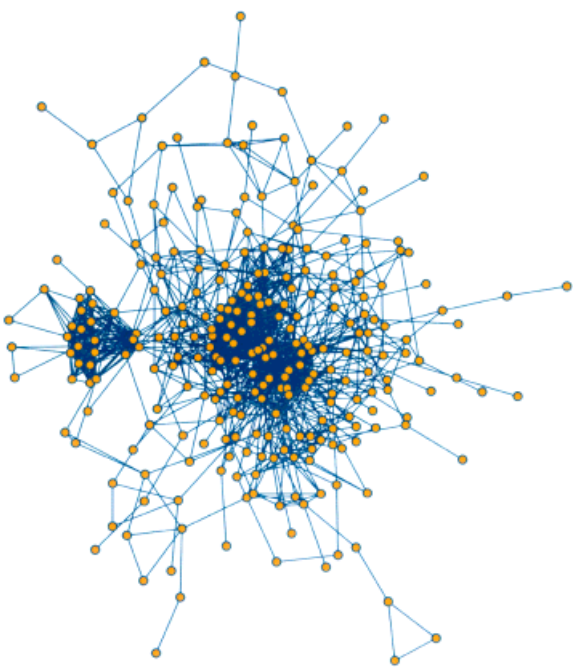

Fig. 5. In Figure 4 we are able to assign a confidence measure to the likelihood that a pair of proteins is acting within the same cellular pathway. If the confidence exceeds 0.5 we consider the two proteins linked. We are able to generate such links for all the related proteins with a genome. In the figure we represent the linkage map for the proteins contained within Mycoplasma Genitalium, the organism with the smallest of all the available genomes.

One possible use of these networks is to assign uncharacterized genes to pathways [Marcotte et al., 1999b; Huynen et al., 2000]. In the case of *E.coli*, for example, about half the genes have been placed within a pathway according to the NCBI annotation scheme. This leaves about 2000 genes that cannot be placed within a pathway using any of the standard homology based bioinformatics techniques (for some of these, however, the biochemical function may be inferred from homology to characterized proteins). We can use our methods to assign genes to the pathway which is most represented among the proteins it is linked to. This concept has often been called "guilt by association": if we know a gene is linked to histidine biosynthesis proteins, it is likely to be part of this pathway.

As before, in *E. coli* we can measure our ability to recover the pathway of previously annotated genes, and use this knowledge to estimate the accuracy with which we assign uncharacterized ORFs to pathways. As shown in Figure 6, we find that for most ORFs our assignment accuracy is about 70%, while for a few hundred it is much higher. Therefore, we can use this technique to significantly enhance the pathway annotation of a complete genome.

The ability to assign uncharacterized proteins to cellular pathways should allow scientists to mine genomes for potential vaccines and

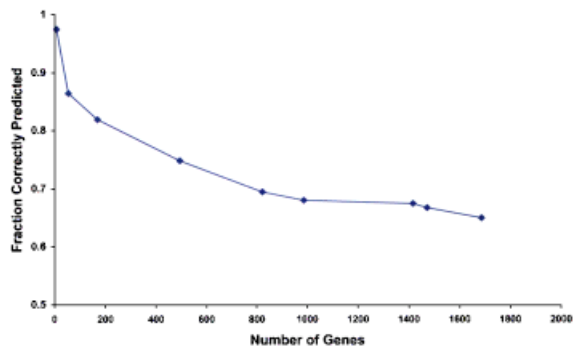

Fig. 6. We attempt to assign a COG (1) category to genes that are un-annotated within the COG scheme. To do this we consider all the links, by our four methods, to the un-annotated proteins. The protein is assigned the most highly represented COG category of the linked proteins. For proteins whose COG category has been assigned we may then check to see the accuracy of our assignments using this scheme. In the figure we plot various points corresponding to different thresholds for determining a link. On the right hand side all links above 0.1 confidence are retained, while on the left hand side only links above 0.9 confidence are used. As we vary the confidence threshold from 0.1 to 0.9 we are able to assign fewer genes (X-axis) to their COG categories, but with higher accuracy (Y-axis).

targets. Inevitably, discovery of new vaccines and targets in microbial genomes will require that computational techniques be coupled with high-throughput experimental techniques. It is now possible to systematically knock-out all genes within a genomes to discover which are essential for the organism's survival and which are not. The combination of experimental information with pathway annotation promises to greatly enhance the ability of pharmaceutical companies to mine genomes for anti-microbial targets. This in turn should facilitate the discovery of new anti-microbial compounds that may be used to combat these pathogens.

REFERENCES

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biuochem Sci 23:324–328.

Enright AJ, Ilioupolos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90.

Huynen MA, Bork P. 1998. Measuring genome evolution. Proc Natl Acad Sci USA 95:5849–5856.

Huynen M, Snel B, Lathe W, Bork P. 2000. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. Genome Research 10: 1204–1210.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999a. Detecting protein function and protein–protein interactions from genome sequences. Science 285(5428):751–753.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999b. A combined algorithm for genome-wide prediction of protein function. Nature 402:83–86.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1998. Use of contiguity on the chromosome to predict functional coupling. In Silico Biology 1:9.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Scie USA 96:2896–2901.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci USA 96(8):4285–4288.

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