Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 4285–4288, April 1999
Biochemistry

Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles

(genomic/bioinformatics/metabolic pathways/structural complexes)

MATTEO PELLEGRINI*, EDWARD M. MARCOTTE*, MICHAEL J. THOMPSON, DAVID EISENBERG, AND TODD O. YEATES†

Molecular Biology Institute and Departments of Energy Laboratory of Structural Biology and Molecular Medicine, and Chemistry and Biochemistry, University of California, Box 951570, Los Angeles, CA 90095-1570

Contributed by David S. Eisenberg, January 20, 1999

ABSTRACT **Determining protein functions from genomic sequences is a central goal of bioinformatics. We present a method based on the assumption that proteins that function together in a pathway or structural complex are likely to evolve in a correlated fashion. During evolution, all such functionally linked proteins tend to be either preserved or eliminated in a new species. We describe this property of correlated evolution by characterizing each protein by its phylogenetic profile, a string that encodes the presence or absence of a protein in every known genome. We show that proteins having matching or similar profiles strongly tend to be functionally linked. This method of phylogenetic profiling allows us to predict the function of uncharacterized proteins.**

The fully sequenced genomes of numerous organisms offer large amounts of information about cellular biology (see the genomes listed at the web site of The Institute for Genome Research: www.tigr.org). It is a central challenge of bioinformatics to use this information in discovering the function of proteins. Functional assignments of genes come primarily from biochemical experimentation, which can be extended by matching recently sequenced proteins to those that have already been characterized (1). For the exceptionally well studied genome of *Escherichia coli* (2), these and related techniques (3, 4) have lead to tentative functional assignments of slightly more than half of its proteins (5). The problem of assigning functions to the remaining proteins is addressed here.

Our computational method detects proteins that participate in a common structural complex or metabolic pathway. Proteins within these groups are defined as *functionally linked*. The underlying hypothesis is that functionally linked proteins evolve in a correlated fashion, and, therefore, they have homologs in the same subset of organisms. For instance, we expect to find flagellar proteins in bacteria that possess flagella but not in other organisms. In short, we show that if two proteins have homologs in the same subset of fully sequenced organisms, they are likely to be functionally linked. We exploit this property systematically to map links between all the proteins coded by a genome. In general, pairs of functionally linked proteins have no amino acid sequence similarity with each other and, therefore, cannot be linked by conventional sequence-alignment techniques.

METHODS

To represent the subset of organisms that contain a homolog, we constructed a phylogenetic profile for each protein. This profile is a string with n entries, each one bit, where n corresponds to the number of genomes (16 in the present article). We indicate the presence of a homolog to a given protein in the nth genome with an entry of unity at the nth position. If no homolog is found, the entry is zero. Proteins are clustered according to the similarity of their phylogenetic profiles. Similar profiles show a correlated pattern of inheritance and, by implication, functional linkage. The method predicts that the functions of uncharacterized proteins are likely to be similar to characterized proteins within a cluster (Fig. 1).

We computed phylogenetic profiles for the 4,290 proteins encoded by the genome of *E. coli* by aligning (6) each protein sequence (P_i) with the proteins from 16 other fully sequenced genomes (listed at the web site of The Institute for Genome Research: www.tigr.org). Proteins coded by the nth genome are defined as including a homolog of P_i if they align to P_i with a score that is deemed statistically significant.‡

RESULTS AND DISCUSSION

To test whether proteins with similar phylogenetic profiles are functionally linked, we examined the phylogenetic profiles for two proteins that are known to participate in structural complexes, the ribosome protein RL7 and the flagellar structural protein FlgL, as well as a protein known to participate in a metabolic pathway, the histidine biosynthetic protein HIS5. We first identified all other *E. coli* ORFs with phylogenetic profiles identical to those of these three proteins and then those ORFs with profiles that differ by one bit. The results are shown in Fig. 2.

Homologs of ribosome protein RL7 are found in 10 of 11 eubacterial genomes and in yeast but not in archaeal genomes. We find that more than half of the *E. coli* proteins with the RL7 phylogenetic profile or profiles that differ from it by one bit have functions associated with the ribosome (Fig. 2*A*). Because none of these proteins have significant amino acid sequence similarity to RL7, the functional relationships to the ribosome—had they not been determined already—could not be inferred by sequence comparisons. This finding supports the idea that proteins with similar profiles are likely to belong to a common group of functionally linked proteins. Several other proteins with these profiles have no assigned function and are listed accordingly as hypothetical. The testable prediction of

Abbreviation: EcoCyc, *Encyclopedia of Escherichia coli Genes and Metabolism*.

*M.P. and E.M.M. contributed equally to this work.

†To whom reprint requests should be addressed. e-mail: yeates@mbi.ucla.edu.

‡The statistical significance of an alignment score is described by the probability (P) of obtaining a higher score when the sequences are shuffled. To compute a P value threshold, we first consider the total number of sequence comparisons that we are performing. If there are n proteins in *E. coli* and m in all other genomes, this number is $n \times m$. If we were to compare this number of random sequences, we would expect one pair to yield a P value of $1/(n \times m)$ by chance. We, therefore, set this P value as our threshold.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

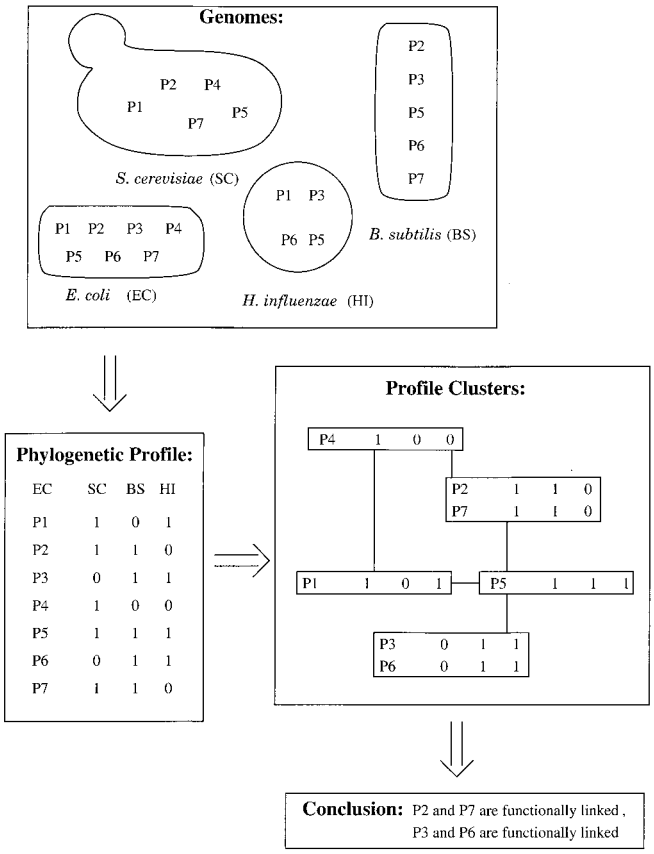

FIG. 1. Our method of analyzing protein phylogenetic profiles is illustrated schematically for the hypothetical case of four fully sequenced genomes (from *E. coli*, *Saccharomyces cerevisiae*, *Haemophilus influenzae*, and *Bacillus subtilis*) in which we focus on seven proteins (P1–P7). For each *E. coli* protein, we construct a profile, indicating which genomes code for homologs of the protein. We next cluster the profiles to determine which proteins share the same profiles. Proteins with identical (or similar) profiles are boxed to indicate that they are likely to be functionally linked. Boxes connected by lines have phylogenetic profiles that differ by one bit and are termed *neighbors*.

the clustering of phylogenetic profiles is that these as yet uncharacterized proteins have functions associated with the ribosome.

The comparisons of the phylogenetic profiles of flagellar proteins (Fig. 2*B*) further support the idea that proteins with similar profiles are likely to be functionally linked; 10 flagellar proteins share a common profile. Their homologs are found in a subset of five bacterial genomes: those of *Aquifex aeolicus, Borrelia burgdorferi, B. subtilis, Helicobacter pylori,* and *Mycobacterium tuberculosis*. Other proteins that appear in neighboring clusters (groups of proteins that share a common profile) include various flagellar proteins and cell-wall maintenance proteins. Flagellar and cell-wall maintenance proteins may be biochemically linked, because flagella are inserted through the cell wall. For example, the lytic murein transglycosylase MltD has a phylogenetic profile that differs by only one bit from that of the flagellar structural protein FlgL. This transglycosylase cuts the cell wall for unknown reasons. Therefore, another prediction is that this enzyme may participate in flagellar assembly.

Fig. 2 *A* and *B* includes proteins in structural complexes, whereas Fig. 2*C* shows proteins involved in amino acid metabolism. We find that more than half of the proteins with phylogenetic profiles similar (within one bit) to that of the histidine synthesis protein His5 are involved in amino acid metabolism. With the 16 currently available fully sequenced genomes, however, phylogenetic profiles are not able to separate the metabolic pathways of specific amino acids. Instead, because of the limitations of currently available data, a histidine biosynthesis protein seems to have the same profile as a tryptophan, arginine, and cysteine synthesis protein. It is probable that, as more genomes are fully sequenced and the number of entries in phylogenetic profiles is increased, similar but distinct amino acid metabolic pathways will cluster separately in phylogenetic-profile spaces.

Table 1. Phylogenetic profiles link protein with similar keywords

Keyword	No. proteins	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin and Molybdenum, and molybdoterin	12	6	1
Hypothetical†	1,084	108,226	8,440

Proteins grouped on the basis of similar keywords in SwissProt have more similar phylogenetic profiles than random proteins. Column 2 gives the number of nonhomologous proteins in the keyword group. Column 3 gives the number of protein pairs in the keyword group with profiles that differ by less than 3 bits. These pairs are called neighbors. Column 4 lists the number of neighbors found on average for a random group of proteins of the same size as the keyword group.

*Only membrane proteins without uniformly zero phylogenetic profiles were included.

†Unlike the other rows of the table, the hypothetical proteins do contain homologous pairs.

The examples included in Fig. 2 show that proteins with phylogenetic profiles similar to a query protein are likely to be functionally linked with it. We next show the converse: that groups of proteins known to be functionally linked often have similar phylogenetic profiles. As shown in Table 1, we chose groups of *E. coli* proteins that share a common keyword in their SwissProt (7) annotation, reflecting well known families of functionally linked proteins. Because homologous proteins coded by the same genome necessarily have similar profiles, they were eliminated from the groups. For each group, we computed the number of protein pairs that are neighbors;

Table 2. Phylogenetic profiles link proteins in EcoCyc classes

EcoCyc class	No. proteins	No. neighbors in EcoCyc class	No. neighbors in random group
Carbon compounds	88	798	60
Anaerobic respiration	66	275	30
Aerobic respiration	28	39	6
Electron transport	26	91	5
Purine biosynthesis	21	11	3
Salvage nucleosides	15	10	1
Fermentation	19	17	3
Tricarboxylic acid cycle	16	6	1
Glycolysis	14	5	1
Peptidoglycan biosynthesis	12	10	1

Proteins grouped according to metabolic function on the basis of EcoCyc classes have more similar phylogenetic profiles than random proteins. Column 2 gives the number of proteins in the EcoCyc class. Column 3 gives the number of protein pairs in the EcoCyc class with profiles that differ by less than 3 bits. These pairs are called neighbors. Column 4 lists the number of neighbors found on average for a random group of proteins of the same size as the keyword group.

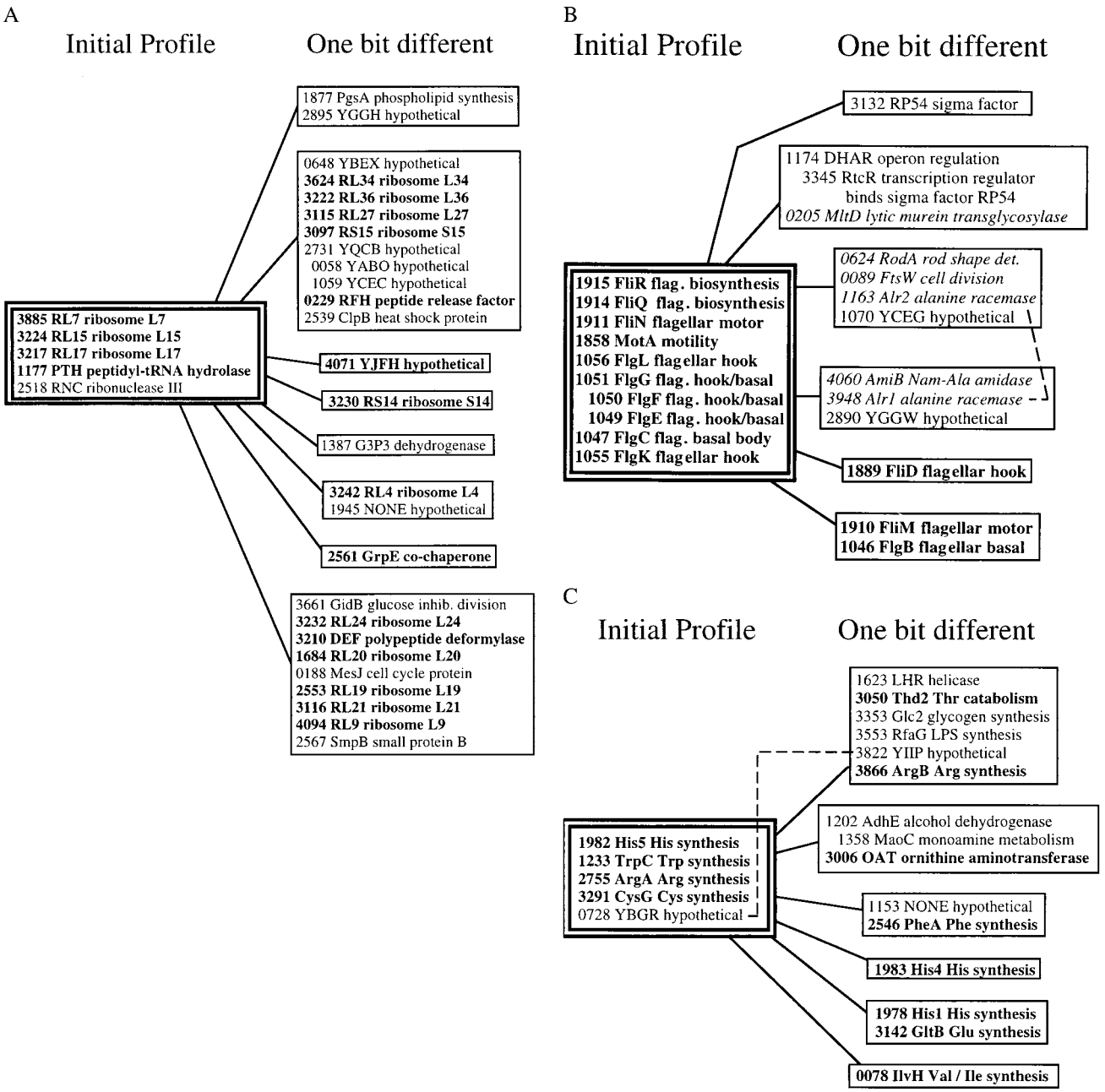

FIG. 2. Proteins with phylogenetic profiles in the neighborhood of ribosomal protein RL7 (*A*), flagellar structural protein FlgL (*B*), and histidine biosynthetic protein His5 (*C*). In each case, we first found all proteins with profiles identical to our query proteins; the proteins we found are shown in the double boxes. We then found all the proteins with profiles that differed from our query proteins by one bit; these are shown in the single boxes. Proteins in bold participate in the same complex or pathway as the query protein, and proteins in italics participate in a different but related complex or pathway. Proteins with identical profiles are shown within the same box. Single lines between boxes represent a one-bit difference between the two profiles. All neighboring proteins whose profiles differ by one bit from the query protein are shown. Homologous proteins are connected by a dashed line or are indented. Each protein is labeled by a four-digit *E. coli* gene number, a SwissProt gene name, and a brief description. Note that proteins within a box or in boxes connected by a line have similar functions. Hypothetical proteins (i.e., those of unknown function) are prime candidates for functional and structural studies. Proteins in the double boxes in *A*, *B*, and *C* have 11, 6, and 10 ones, respectively, in their phylogenetic profiles, of a possible 16 for the 17 genomes presently sequenced.

neighbors were defined as proteins whose profiles differ by less than 3 bits. For a group of n proteins there are at most $[n(n-1)]/2$ possible neighbors.

The similarity of the phylogenetic profiles of the proteins that share a common keyword was evaluated by a statistical test; we compared the number of neighbors found in our keyword groups to the average number of neighbors found in a group of the same size but with randomly selected *E. coli* proteins. We found that, on average, the random sets contain very few neighbors compared with the keyword groups, even though the keyword groups contain only a fraction of all possible neighbor pairs. Thus, proteins that are functionally linked are far more likely to be neighbors in profile space than randomly selected proteins. However, we find only a fraction of all possible neighbors within a group. Therefore, not all functionally linked proteins have similar profiles; they may fall into multiple clusters in profile space. It is interesting to note that hypothetical proteins are also more likely to be neighbors than random proteins, suggesting that many hypothetical proteins are part of uncharacterized pathways or complexes.

A second indication that functionally linked proteins are likely to have similar phylogenetic profiles comes from the

analysis of classes of proteins obtained from the EcoCyc library (*Encyclopedia of Escherichia coli Genes and Metabolism*, ref. 8). We selected several classes that contain more than 10 members and that represent well known biochemical pathways. The results of our analysis are listed in Table 2. The conclusions that we draw from this analysis are similar to those found with the keyword groups: members of the group are far more likely to have neighboring profiles than members of a randomly selected control group.

Finally, we attempted to determine the ability of our method to predict the function of uncharacterized proteins. We equate the function of a protein with that of its neighbors in phylogenetic-profile space. This equation is accomplished by means of the keyword annotations found in the SwissProt database. To test the efficacy of this method, we compared the keywords of each characterized protein to those of the neighbors in phylogenetic-profile space. All of the neighbors, in this case, were other proteins with identical profiles. We found that on average 18% of the neighbor keywords overlapped the known keywords of the query protein. By comparison, random proteins had only a 4% overlap with the same set of neighbors. We make the rough estimate that, for more than half of *E. coli* proteins, we can assign the general function correctly by examining the functions of their phylogenetic-profile neighbors. This estimate should also hold true for the ability of phylogenetic profiles to assign functions to uncharacterized proteins.

CONCLUSION

The phylogenetic profile of a protein describes the presence or absence of homologs in organisms. Proteins that make up multimeric structural complexes are likely to have similar profiles. Also, proteins that are known to participate in a given biochemical pathway are likely to be neighbors in the space of phylogenetic profiles. These findings indicate that comparing profiles is a useful tool for identifying the complex or pathway in which a protein participates. Finally, we were able to make functional assignments of uncharacterized proteins by examining the function of proteins with identical phylogenetic profiles.

As the number of fully sequenced genomes increases, scientists will be able to construct longer and potentially more informative protein phylogenetic profiles. There are at least 100 genome projects underway that are due to be completed within the next few months. These data will enable the construction of profiles 100 bits rather than 16 bits in length. Because the number of profile patterns grows exponentially with the number of fully sequenced genomes, the results of 100-bit comparisons should be considerably more informative than those with 16 bits. Furthermore, because the newly sequenced genomes will include several eukaryotic organisms, protein phylogenetic profiles also should become a useful tool for studying structural complexes and metabolic pathways in these higher organisms.

Note Added in Proof. A data structure similar to the one described in *Methods* has been proposed independently by Regan and Gaasterland (9) for describing the distribution of proteins in genomes.

This work was supported by a postdoctoral fellowship from the Sloan Foundation and the Department of Energy (to M.P.), by a Hollaender postdoctoral fellowship from the Department of Energy and the Oak Ridge Institute for Science and Education (to E.M.), and by grants from the Department of Energy and the National Institutes of Health.

1. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998) *J. Mol. Biol.* **283**, 707–725.
2. Blattner, F. R., Plunckett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al*. (1997) *Science* **265**, 1453–1474.
3. Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996) *Curr. Biol.* **6**, 279–291.
4. Andrade, M. A. & Sander, C. (1997) *Curr. Opin. Biotechnol.* **8**, 675–683.
5. Riley, M. (1998) *Nucleic Acids Res.* **26**, 54.
6. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acid Res.* **25**, 3389–3402.
7. Bairoch, A. & Apweiler, R. (1998) *Nucleic Acids Res.* **26,** 38–42.
8. Karp, P., Riley, M., Paley, S. & Pellegrini-Toole, A. (1998) *Nucleic Acids Res.* **26**, 50–53.
9. Gaasterland, T. & Regan, M. A. (1998) *Microb. Comp. Genomics* **3,** 177–192.