Detecting Protein Function and Protein-Protein Interactions from Genome Sequences

Edward M. Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W. Rice, Todd O. Yeates, David Eisenberg*

A computational method is proposed for inferring protein interactions from genome sequences on the basis of the observation that some pairs of interacting proteins have homologs in another organism fused into a single protein chain. Searching sequences from many genomes revealed 6809 such putative protein-protein interactions in *Escherichia coli* and 45,502 in yeast. Many members of these pairs were confirmed as functionally related; computational filtering further enriches for interactions. Some proteins have links to several other proteins; these coupled links appear to represent functional interactions such as complexes or pathways. Experimentally confirmed interacting pairs are documented in a Database of Interacting Proteins.

The lives of biological cells are controlled by interacting proteins in metabolic and signaling pathways and in complexes such as the molecular machines that synthesize and use adenosine triphosphate (ATP), replicate and translate genes, or build up the cytoskeletal infrastructure (*1*). Our knowledge of protein-protein interactions has been accumulated from biochemical and genetic experiments, including the widely used yeast two-hybrid test (*2*). Here we ask if protein-protein interactions can be recognized from genome sequences by purely computational means.

Some interacting proteins such as the Gyr A and Gyr B subunits of *Escherichia coli* DNA gyrase are fused into a single chain in another organism, in this case the topoisomerase II of yeast (*3*). Thus, the sequence similarities of Gyr A (804 amino acid residues) and Gyr B (875 residues) to different segments of the topoisomerase II (1429 residues) might be used to predict that Gyr A and Gyr B interact in *E. coli*.

To find other such putative protein interactions in *E. coli*, we searched the 4290 protein sequences of the *E. coli* genome (*4*) for these patterns of sequence homology (*5*). We found 6809 pairs of nonhomologous sequences, both members of the pair having significant similarity (*6*) to a single protein in some other genome that we term a Rosetta Stone sequence because it deciphers the interaction between the protein pairs. The 4290 proteins could form at most $(4290)^2/2 = 9 \times 10^6$ pair interactions, but we would expect many fewer interactions in a functioning cell; roughly 2 to 10 interactions for each protein does not seem unreasonably many.

Each of these 6809 pairs is a candidate for a pair of interacting proteins in *E. coli*. Five such candidates are shown in Fig. 1. The first three pairs of *E. coli* proteins were among those easily determined from the biochemical literature in fact to interact. The final two pairs of proteins are not known to interact. They are representatives of many such pairs whose putative interactions at this time must be taken as testable hypotheses.

We devised three independent tests of interactions predicted by the method we term domain fusion analysis, each showing that a reasonable fraction may in fact interact. The first method uses the annotation of proteins given in the SWISS-PROT database (*7*). For cases where the interacting proteins have both been annotated, we compare their annotations, looking for a similar function for both members of the pair. Similar function would imply at least a functional interaction. Of the 3950 *E. coli* pairs of known function, 2682 (68%) share at least one keyword in their SWISS-PROT annotations (ignoring the keyword "hypothetical protein"), suggesting related functional roles. When pairs of annotated *E. coli* proteins are selected at random, only 15% share a keyword. In short, of the *E. coli* pairs that the domain fusion analysis turns up as candidates for protein-protein interactions, more than half have both members with a similar function; the method therefore seems to be a robust predictor of protein function. Where the function of one member of a protein pair is known, the function of the other member can be predicted. Performing a similar analysis in yeast turns up 45,502 protein pairs. Of the 9857 pairs of known function, 32% share at least one keyword in their annotations compared with 14% when proteins are selected at random.

The second test of the interactions predicted by the domain fusion analysis uses as confirmation the Database of Interacting Proteins (*8*). This database is a compilation of protein pairs that have been found to interact in some published experiment. As of December 1998, the database contained 939 entries, 724 of which have both members of the pair listed in the ProDom database. Of these 724 pairs, we found 46 or 6.4% linked by Rosetta Stone sequences. We expect this percentage to rise as more genomes are sequenced.

The third test of domain fusion predictions is by another computational method for predicting interactions (*9*), the method of phylogenetic profiles, which detects functional interactions by analyzing correlated evolution of proteins. This method was applied to the 6809 interactions predicted by the domain fusion analysis for *E. coli* proteins. Some 321 of these predictions (~5%) were suggested by the phylogenetic profile method to interact, more than eight times as many interactions in common as for randomly cho-

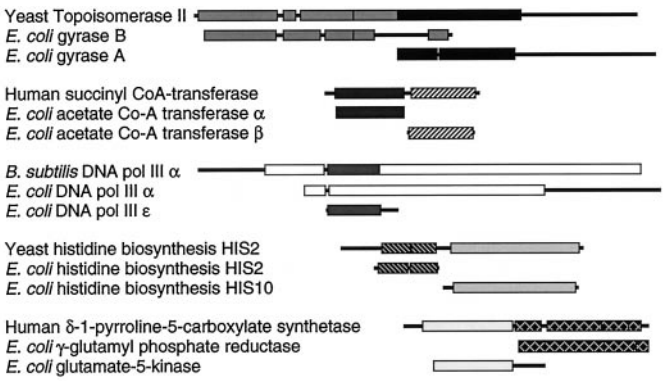

Fig. 1. Five examples of pairs of *E. coli* proteins predicted to interact by the domain fusion analysis. Each protein is shown schematically with boxes representing domains [as defined in the ProDom domain database (*17*)]. For each example, a triplet of proteins is pictured: The second and third proteins are predicted to interact because their homologs are fused in the first protein (called the Rosetta Stone protein in the text). The first three predictions are known to interact from experiments (*18*). The final two examples show pairs of proteins from the same pathway (two nonsequential enzymes from the histidine biosynthesis pathway and the first two steps of the proline biosynthesis pathway) that are not known to interact directly.

UCLA–Department of Energy Laboratory of Structural Biology and Molecular Medicine, Departments of Chemistry and Biochemistry and Biological Chemistry, Box 951570, University of California at Los Angeles, Los Angeles, CA 90095–1570, USA.

*To whom correspondence should be addressed: E-mail: david@mbi.ucla.edu

sen sets of interactions. Given that the domain fusion method and the phylogenetic profile method rest on entirely different assumptions, this level of overlap of predictions tends to support the predictive power of both methods.

The discovery of many possible pair interactions between proteins of *E. coli* encouraged us to look for coupled interactions, where A is predicted to interact with B and B with C, and so forth. That is, we asked if the domain fusion method can turn up complexes of proteins or protein pathways. As Fig. 2 shows, suggestive information on both pathways and complexes did emerge from linked pairs of *E. coli* proteins. The pathways for shikimate biosynthesis and purine biosynthesis are shown in Fig. 2 (pathways A and C, respectively). The enzymes in these pathways for which links were found to other members of the same pathway are shown in bold type. The precise links suggested by Rosetta Stone sequences are shown in Fig. 2 (B and D). Some of these discovered links are between sequential enzymes in the pathway, and others are between more distant members, perhaps suggesting a multienzyme complex. An alternative explanation of the same findings is that enzymes in the pathway are expressed in a fused form in some organisms as an aid in regulation of expression; in this case, linked members of a pair would not necessarily bind to each other (see below).

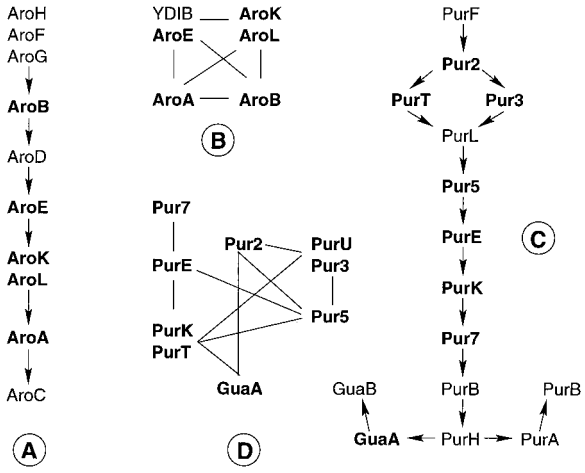

Fig. 2. Reconstruction of two metabolic pathways in *E. coli*, with only interactions predicted by the domain fusion method. Pathways A and C are the known pathways for biosynthesis of shikimate and purine, respectively; they are ordered by the traditional method of successive action of the enzymes on the known metabolites. Pathways B and D are constructed from the proteins in pathways A and C with connections predicted by the domain fusion method. In both cases, more than half of the proteins in the biochemical pathway are predicted by the domain fusion method to interact with other proteins of the pathway. It is possible that these groupings represent multiprotein complexes. Enzymes stacked together (for example, AroK and AroL) are homologs.

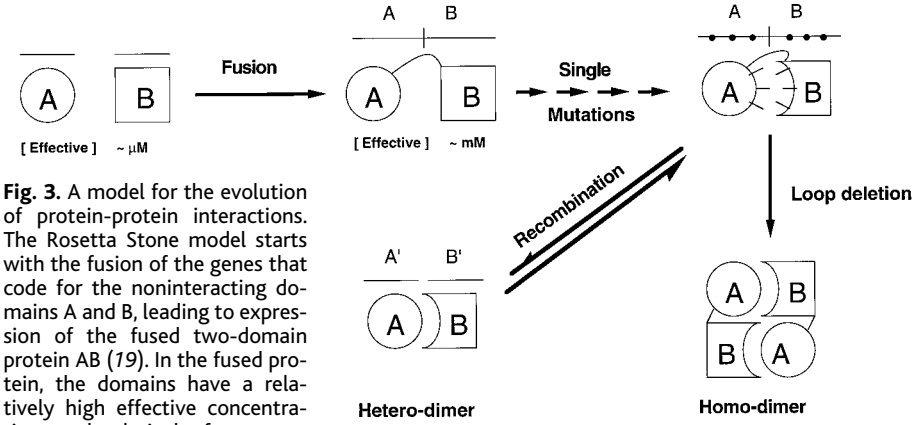

Fig. 3. A model for the evolution of protein-protein interactions. The Rosetta Stone model starts with the fusion of the genes that code for the noninteracting domains A and B, leading to expression of the fused two-domain protein AB (*19*). In the fused protein, the domains have a relatively high effective concentration, and relatively few mutations create a primitive binding site between the domains that is optimized by successive mutations. In the second line, the interacting domains are separated by recombination with another gene to create an interacting pair of proteins A and B. An interacting pair of proteins A and B can be created by fission of a protein, so that the preliminary fusion step is not essential to the Rosetta Stone hypothesis. The lower righthand step shows another possible mutation, a loop deletion that leads to a domain-swapped homodimer. This evolutionary path to homooligomers has been discussed earlier (*20*) and is the analog for homooligomers of the evolutionary path suggested here for heterooligomers.

To evaluate the reliability of domain fusion predictions of protein interactions, it is helpful to consider why the method should work in the first place. This emerges from considerations of protein affinity. It follows from the laws of thermodynamics that the fusion of protein domains A and B into a single protein chain can profoundly enhance the affinity of A for B. The reason for this is that fusion greatly reduces the entropy of dissociation of A with B, thereby reducing the association free energy of A to B (*10*). This reduction in entropy is often expressed as an increase in the effective concentration of A with respect to B. The concentrations of proteins in *E. coli* cells tend to be on the order of micromolar (*11*), whereas the effective concentrations of fused proteins can be ~millimolar or even greater (*12*). Put another way, the standard free energy of dissociation for protein subunits from a complex is typically 8 to 20 kcal/mol at 27°C (corresponding to dissociation constants of 10^{-6} to 10^{-14} M) (*13*) and can be reduced by ~10 kcal/mol when the subunits are fused into a single protein chain. Because affinity between proteins A and B is greatly enhanced when A is fused to B, some interacting pairs of proteins may have evolved from primordial proteins that included the interacting domains A and B on the same polypeptide, as shown in Fig. 3. We term this pathway the Rosetta Stone hypothesis for evolution of protein interactions. Also in support of the Rosetta Stone pathway is the observation that protein-protein interfaces have strong similarity to interdomain interfaces within single protein molecules (*14*).

It is important to realize that the domain fusion analysis makes two distinct predictions. First, it predicts protein pairs that have related biological functions—that is, proteins that participate in a common structural complex, metabolic pathway, or biological process. Prediction of function is robust: For *E. coli*, general functional similarity was observed in over half the testable predictions. Second, the method predicts potential protein-protein interactions. For this more specific prediction, the considerations of protein affinity and evolution aid understanding in which cases the domain fusion method will miss pairs of interacting proteins (false negatives) and in which cases it will turn up false candidates for interacting pairs (false positives). One reason for missing interactions is that many protein-protein interactions may have evolved through other mechanisms, such as gradual accumulation of mutations to evolve a binding site. In these cases, there never was a fusion of the interacting proteins, and so no Rosetta Stone sequence can be found. Second, even in other cases where the interacting partners were once fused, the fused protein may have disappeared during the course of evolution, and so there is no Rosetta Stone relic remaining to decipher binding partnerships. As more genomes are sequenced, however, there is a higher chance of finding Rosetta Stone sequences.

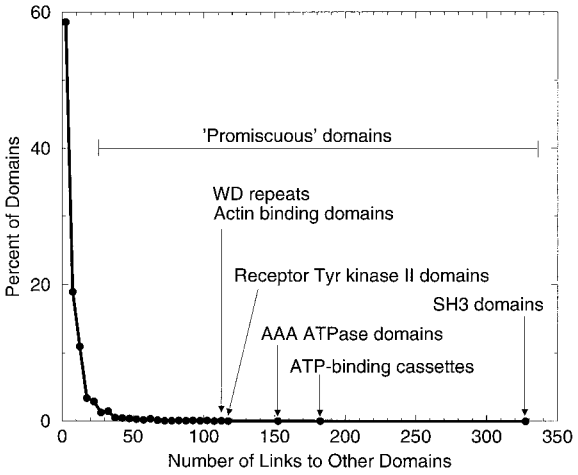

Fig. 4. The detection of "promiscuous" domains for filtering of false interactions by the domain fusion method. For each protein domain (as defined in the ProDom database), we calculated the number of Rosetta Stone links that could be found to other domains. Plotting this distribution shows that for most domains (~95%), only a few Rosetta Stone links are found. For the remaining ~5% of domains, many links are found. These "promiscuous" domains are domains such as the SH3 domains and ATP-binding cassettes that are found in many otherwise unrelated proteins.

False predictions of physical interactions may be made by the domain fusion analysis in cases where domains are fused but not interacting. This may be so when proteins have been fused to regulate coexpression or protein signaling. For these cases, the "interaction" of the proteins can be a functional interaction rather than a physical interaction. Other false predictions can arise because the domain fusion analysis cannot distinguish between homologs that bind and those that do not. As an example, consider the signaling domains SH2 and SH3. The kinase domain and the SH2 and SH3 domains of the src homology kinase interact with one another in the src molecule (*15*), but homologs of these domains are found in many other proteins, and it is certainly untrue that all SH2 domains interact with all SH3 domains. A similar problem crops up with epidermal growth factor and immunoglobulin domains. The false positive rate in *E. coli* due to the inability to distinguish homologs is about 82% (*16*). That is, although the domain fusion analysis gives a robust prediction of protein function of the form "A is functionally linked to B," only a subset of these putative interactions represent physical interactions between proteins.

To quantify and reduce errors in predicting protein-protein interactions, we calculated the occurrence of "promiscuous" domains such as SH3 that are present in many otherwise different proteins. These domains can be identified and removed during domain fusion analysis. In the ProDom database of domains, we counted the number of other domains that each domain could be linked to using the domain fusion method. As shown in Fig. 4, about 95% of the domains are linked to only a few other domains. For the 7842 domains in the ProDom domain database for which we can find Rosetta Stone links, only about 5% are "promiscuous," making more than 25 links to other domains. That is, by filtering of only 5% of all domains from our domain fusion predictions, we can remove the majority of falsely predicted interactions. When this type of filtering is applied to the 3531 Rosetta Stone links of *E. coli* found with the ProDom analysis, the number is reduced to 749. Although dropping the number of predictions, this filtration step increases the likelihood that predicted links represent true physical interactions by 47% over the unfiltered predictions. Also, after filtering out promiscuous domains, the average false positive rate in *E. coli* due to the inability to distinguish homologs drops to 65%. The practical result of domain fusion analysis is that many protein interactions can be predicted from genome sequences, permitting experimentalists to focus on promising interactions.

In summary, genomic information opens new paths to biochemical discoveries. The finding in a genome of many pairs of protein sequences A′ and B′ that are both homologs to a single sequence A^B in another genome suggests the possibility that A′ and B′ are binding partners and provides robust functional information about A′ and B′. Systematic searches of this sort may lead to identifications of new pathways and protein complexes in organisms.

References and Notes

1. B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland, New York, ed. 3, 1994); H. Lodish *et al.*, *Molecular Cell Biology* (Scientific American Books, New York, ed. 3, 1995).
2. S. Fields and O. K. Song, *Nature* **340**, 243 (1989).
3. J. M. Berger, S. J. Gamblin, S. C. Harrison, J. C. Wang, *ibid.* **379**, 225 (1996).
4. F. R. Blattner *et al.*, *Science* **277**, 1453 (1997).
5. The triplets of proteins are found with the aid of protein domain databases such as the ProDom or Pfam databases (*17*). Here, a list of all ProDom domains in every one of the 64,568 SWISS-PROT proteins was prepared, as well as a list of all proteins that contain each of the 53,597 ProDom domains. Then every protein in ProDom was considered for its ability to be a linking (or Rosetta Stone) member in a triplet. All pairs of domains that are both members of a given protein P were defined as being linked by protein P, if we could find at least one protein with only one of the two domains. By this method, we found 14,899 links between the 7843 ProDom domains. Then in a single genome (such as *E. coli*), we found all nonhomologous pairs of proteins containing linked domains. These pairs are linked by the Rosetta Stone proteins. For *E. coli*, this method finds 3531 protein pairs. An alternate method for discovering protein triplets uses amino acid sequence alignment techniques to find two proteins that align to a Rosetta Stone protein such that the alignments do not overlap on the Rosetta Stone protein. For *E. coli*, this method finds 4487 protein pairs, 1209 of which were also found by the ProDom search method (even though different sequence databases were searched for each method). All predictions are available on the World Wide Web at www.doe-mbi.ucla.edu.
6. Two amino acid sequences are said to be similar when the sequences align with a statistically significant alignment score. The significance is described by the probability of obtaining a higher alignment score when comparing shuffled sequences, with the acceptable probability threshold set by considering the total number of sequence comparisons performed. That is, if n proteins in *E. coli* are compared with m proteins in other genomes, $n \times m$ total comparisons are performed. We set a probability of $1/(n \times m)$ as the threshold as this is the lowest value that could be obtained by comparing $n \times m$ random sequences. For the ProDom-based identification of homologs, definitions of sequence similarity are as in the ProDom database.
7. The SWISS-PROT database is available at www.expasy.ch/sprot/.
8. The Database of Interacting Proteins is available on the Web at http://www.doe-mbi.ucla.edu.
9. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4285 (1999).
10. H. P. Erickson, *J. Mol. Biol.* **206**, 465 (1989); A. D. Nagi and L. Regan, *Folding Design* **2**, 67 (1997).
11. S. Pederson, P. S. Bloch, S. Reen, F. C. Neidhardt, *Cell* **14**, 179 (1978).
12. C. R. Robinson and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 5929 (1998).
13. N. Horton and M. Lewis, *Protein Sci.* **1**, 169 (1992); J. Janin, *Biochimie* **77**, 497 (1995).
14. C. J. Tsai and R. Nussinov, *J. Mol. Biol.* **260**, 604 (1996).
15. W. Xu, S. C. Harrison, M. J. Eck, *Nature* **385**, 595 (1997); F. Sicheri, I. Moarefi, J. Kuriyan, *ibid.*, p. 602.
16. The error in predicting protein-protein interactions due to the inability to distinguish homologs was estimated as 1–T, where T is the mean percentage of potential true positives calculated for all domain pairs in *E. coli*. For each domain pair linked by a Rosetta Stone protein, there are n proteins with the first domain but not the second and m proteins with the second domain but not the first. The percentage of true positives T is therefore estimated as the smaller of n or m divided by n times m.
17. F. Corpet, J. Gouzy, D. Kahn, *Nucleic Acids Res.* **26**, 323 (1998); A. Bateman *et al.*, *ibid.* **27**, 260 (1999).
18. A. Sugino, N. P. Higgins, N. R. Cozzarelli, *ibid.* **8**, 3865 (1980); W. K. Yeh and L. N. Ornston, *J. Biol. Chem.* **256**, 1565 (1981); C. S. McHenry and W. Crow, *ibid.* **254**, 1748 (1979).
19. See Table II of J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981). Note also that eukaryotic genes, in contrast to prokaryotic genes, often code for multidomain proteins [W. J. Netzer and F. U. Hartl, *Nature* **388**, 343 (1997)].
20. M. J. Bennett, S. Choe, D. Eisenberg, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 3127 (1994).

21. Supported by the following grants: Department of Energy (DOE) DE-FC03-87ER-60615, NIH PO1 GM 31299, and NSF MCB 94 20769. E. M. was supported by a DOE Hollaender fellowship. We thank M. K. Baron for her work with the Database of Interacting Proteins.

28 December 1998; accepted 23 June 1999