Structure
Article

Functional Linkages Can Reveal Protein Complexes for Structure Determination

Sul-Min Kim,[1] Peter M. Bowers,[1,2] Debnath Pal,[1,2,4] Michael Strong,[1] Thomas C. Terwilliger,[3] Markus Kaufmann,[1] and David Eisenberg[1,2,*]
[1]Department of Chemistry and Biochemistry
[2]Howard Hughes Medical Institute, Institute for Genomics and Proteomics
University of California, Los Angeles, Los Angeles, CA 90095, USA
[3]Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[4]Present address: Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, Karnataka, India
*Correspondence: david@mbi.ucla.edu
DOI 10.1016/j.str.2007.06.021

SUMMARY

In the study of protein complexes, is there a computational method for inferring which combinations of proteins in an organism are likely to form a crystallizable complex? Here we attempt to answer this question, using the Protein Data Bank (PDB) to assess the usefulness of inferred functional protein linkages from the Prolinks database. We find that of the 242 nonredundant prokaryotic protein complexes shared between the current PDB and Prolinks, 44% (107/242) contain proteins linked at high confidence by one or more methods of computed functional linkages. Similarly, high-confidence linkages detect 47% of known *Escherichia coli* protein complexes, with 45% accuracy. Together these findings suggest that functional linkages will be useful in defining protein complexes for structural studies, including for structural genomics. We offer a database of inferred linkages corresponding to likely protein complexes for some 629,952 pairs of proteins in 154 prokaryotes and archaea.

INTRODUCTION

The premise underlying high-throughput proteomics and structural genomics is a belief that cellular systems and networks can be understood from a complete knowledge of the individual components and their interactions (Hood et al., 2004; Ideker et al., 2001). Likewise, the engineering of molecular machines with designed properties require a well-described catalog of molecular components whose functions can be combined in a rational manner. A protein's function is best understood within the context of its interactions with other proteins and ligands. Interacting proteins include signaling and network proteins which play numerous transient biochemical roles, as well as permanent protein complexes representing distinct, stable modules of function that perform defined tasks. Here, we focus on the identification of long-lived protein complexes that can be studied by X-ray crystallography.

Efforts have been undertaken to describe the complete set of protein structures and interactions in a number of organisms. Currently, 11 structural genomics consortia in the USA and numerous others abroad are working on determining the unique protein structures in 12 organisms, including 7 prokaryotes and archaea (Goulding et al., 2002; Terwilliger et al., 2003; Todd et al., 2005; Zhang and Kim, 2003). These projects have added hundreds of novel structures and protein folds to the PDB database over the past several years, accounting for over a quarter of all new protein structures determined (Terwilliger, 2004). Protein interaction data are also being generated by high-throughput techniques for many model organisms and pathogens, resulting in moderately complete genome interaction maps (Butland et al., 2005; Gavin et al., 2002; Giot et al., 2003; Ho et al., 2002; Ito et al., 2001; Li et al., 2004; Uetz et al., 2000). The expectation is that this ensemble of complexes and interactions will lead to comprehension of both normal and pathological cellular functions. Consequently, effective tools for the identification of potential protein complexes, particularly for structural genomics projects, are desirable.

Crystallization and structure determination of protein complexes offers several advantages relative to studying individual polypeptide chains (Shen et al., 2005). Complexes are potentially more stable and soluble than individual proteins expressed without their physiologic partners. Characterization of complexes may confirm existing interactions derived from high-throughput experimental techniques, and as described in this paper, validate macromolecular machines detected by functional genomics studies. Structure determination of protein complexes may help to accelerate structural genomics initiatives currently cloning, expressing, purifying and crystallizing individual proteins, and aid in our detailed understanding of their component biological function (Yakunin et al., 2004; Kim et al., 2003, 2004). Specifically, some failures in structural genomics pipelines (Figure 1) at the expression, solubility, and crystallization stages may be

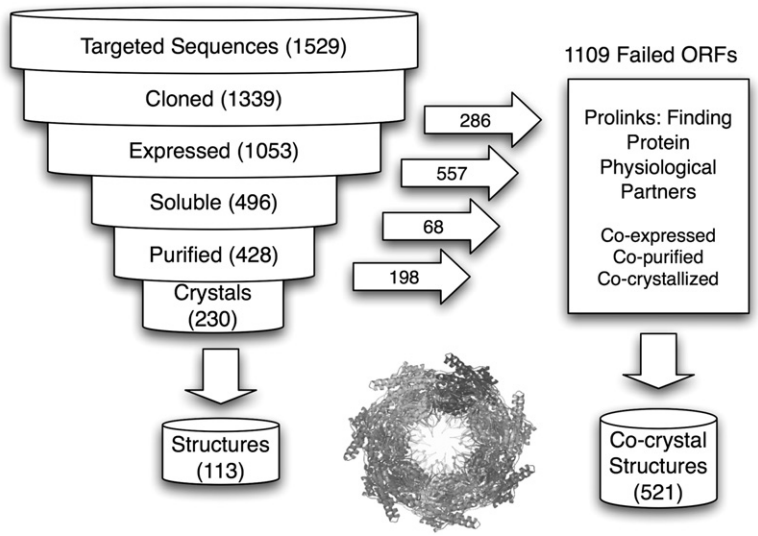

Figure 1. Prolinks: a Tool to Find Interaction Partners for Crystallization

The Prolinks database (http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav) can be used as a tool to find protein partners for inexpressible, insoluble, or uncrystallizable proteins. By forming a soluble complex between the previously recalcitrant protein and a Prolinks predicted interaction partner, a failed protein can be returned as a complex to the structural genomics pipeline in some cases. The numbers on the funnel at the left give provide the surviving proteins at each step in the *Mycobacterium tuberculosis* (Mtb) Structural Genomics Consortium pipeline. The numbers on the arrows show the losses at each step. The number 521 is the number of failed proteins that can, in principle, be rescued by partnering.

rescued by finding the appropriate protein partner for the failed protein, yielding structures that otherwise would be lost.

Functional linkages between proteins, derived from computational genomic analysis, offer one way to identify protein complexes for structural studies (Bowers et al., 2004; Dandekar et al., 1998; Enright et al., 1999; Ermolaeva et al., 2001; Lee et al., 2004; Marcotte et al., 1999a, 1999b; Pellegrini et al., 1999; Wu et al., 2003). We describe methods that can detect the subunits of stable protein complexes amenable to X-ray crystallographic structure determination, utilizing over 17 million high-confidence linkages previously assembled in the Prolinks database (Bowers et al., 2004). High-confidence Prolinks linkages identify nearly half of the prokaryotic complexes (44%) contained in the PDB database. In particular, linkages detected by the Gene Neighbor method provide excellent coverage and accuracy of 929 known *E. coli* protein complex interactions. The database of functional linkages provides a source of promising macromolecular complexes for further study by X-ray crystallography.

RESULTS

Can Known Protein Complexes Be Selected by Functional Linkages?

We sought to establish whether functional linkages could detect interactions between the polypeptide chains of previously determined protein complexes. The utility of functional linkages in identifying protein complexes can be described by their coverage and accuracy, where coverage describes the absolute number of "true" or correctly identified PDB protein complex interactions, and accuracy describes the number of "true" interactions detected relative to the total number of linkages predicted ("true"/ "true + false"). We define a "true" or correct prediction as a high-confidence functional linkage between polypeptide chains (of different amino acid sequence) within the same PDB complex. "False" linkages are defined as high-confidence linkages detected between sequences not yet observed to interact in the PDB. The confidence metric used to select potential PDB interactions from the Prolinks linkages was determined from a keyword analysis of proteins with existing functional annotation (Bowers et al., 2004).

Prolinks linkages are able to identify many of the protein-protein interactions contained in the PDB database (Figure 2). Of the 24,475 PDB structures represented in the PDB sequence database (PDBAA), 3924 complexes contain at least two different polypeptide chains. We focused our initial study on a subset of 782 prokaryotic complexes, as the Prolinks linkage methods rely on factors such as intergenic distance and phylogenetic distribution, and are better suited to prokaryotic organisms. Within the set of prokaryotic PDB complexes, we identified 365 structures from 74 prokaryotic source organisms with sequences shared in common with the Prolinks database. By crossreferencing sequences in both databases via BLAST sequence alignments, we find 202 prokaryotic complexes containing Prolinks linkages, providing 55% coverage (202/365) of the shared structures. As a more sensitive measure of coverage, we also examined Prolinks coverage of structures in the nonredundant PDB database. True links are found in 107 out of 242 nonredundant prokaryotic structures (44%). If *E. coli* structures, which are overrepresented in the PDB, are excluded, coverage of unique complexes increases to 53% (47/88). A majority of true PDB interactions in nonredundant structures are determined by a single method (Figure 2, lower right box): Gene Neighbor linkages. The 202 correctly predicted complexes containing 1375 prokaryotic interactions are listed in Table S1 (see the Supplemental Data available with this article online).

Each of these assessments indicates that high-confidence linkages can predict as much as ~50% of the known polypeptide chains of PDB complexes that

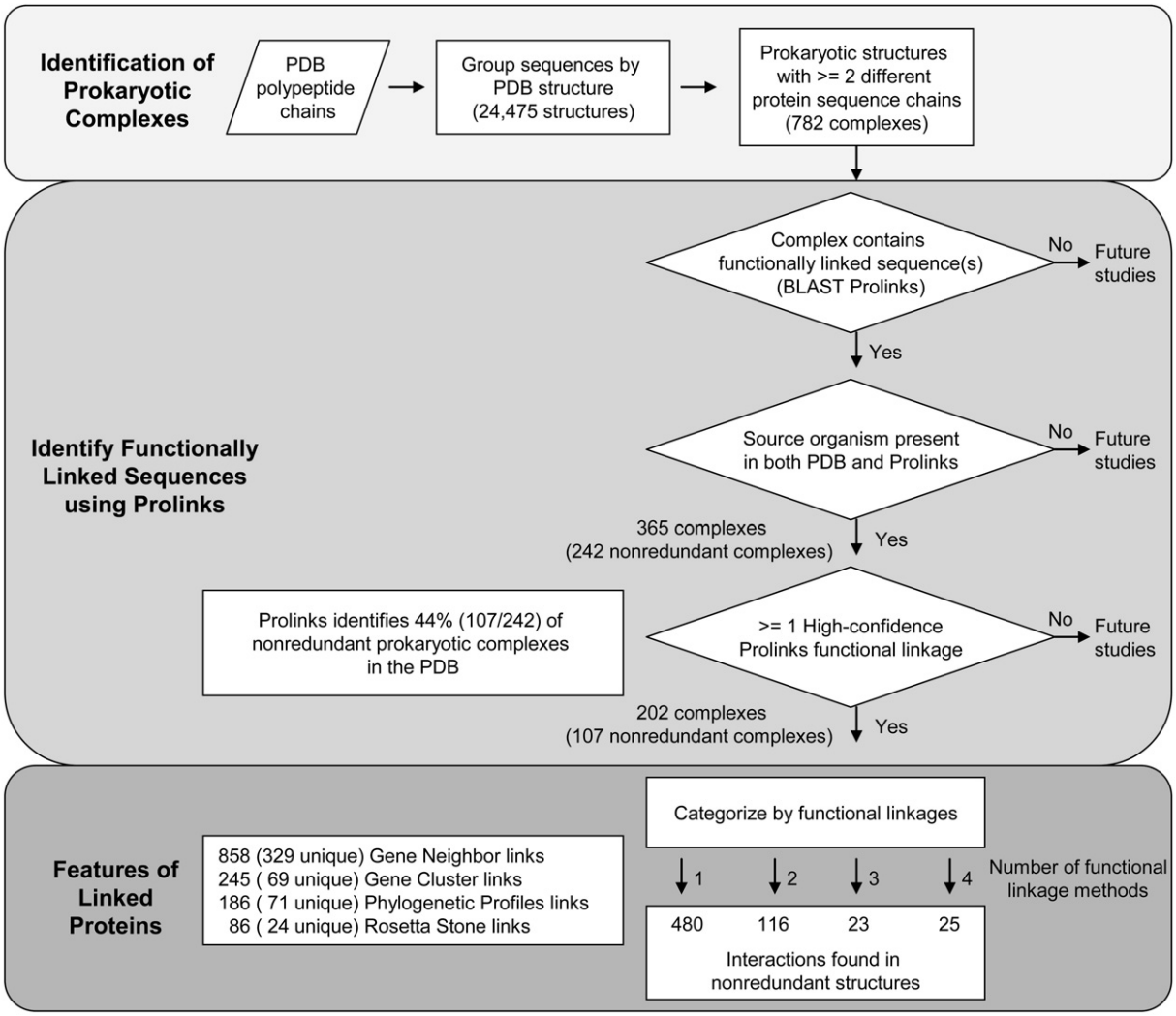

Figure 2. Flowchart to Identify Prokaryotic Prolinks Functional Linkages in PDB Structures

Amino Acid sequences in the PDB are used to identify Prolinks sequences, identifying top scoring hits (e value < $1e^{-20}$; local alignment $\geq$ 75%). PDB sequences containing significant Prolinks matches are grouped by their PDB structure and filtered for functional linkages between pairs of prokaryotic sequences found in different chains of the structure. These functional linkages are categorized by linkage method and number of links and defined as true positives for predicting multisubunit complexes.

are in direct physical contact. For example, in the *E. coli* succinate dehydrogenase complex (PDB code: 1NEK), responsible for the oxidation of succinate to fumarate in aerobic respiration, all four chains are functionally linked to one another by Prolinks linkages, as shown in Figure 3A. Of the six interactions within this complex, three are linked by a single functional linkage method, two are linked by two methods, and one interaction is linked by three methods. *Wolinella succinogenes* fumarate reductase (1QLB), shown in Figure 3B as a dimer, catalyzes the reverse reaction of succinate dehydrogenase in anaerobic environments. The interaction graph of this complex shows linkages from FrdA to FrdB and from FrdA to FrdC. Linkage between FrdA to FrdB is verified by three methods, while a single method identifies FrdA to FrdC linkage. In a similar case, the *Bordetella pertussis* toxin structure (1PRT) contains linkages predicted by the Gene Cluster method only (Figure 3C). Cα atoms from adjacent polypeptides chains in these structures are all within 8 Å of each other. In each of the examples, we find evidence that functional linkage methods not only pinpoint protein pairs found in structural complexes, but also detect proteins pairs that form physical interactions.

The Gene Neighbor Method Effectively Detects Prokaryotic PDB Complexes

We used four linkage algorithms which exhibit varied levels of coverage and accuracy in predicting protein complexes, as determined by our recovery of PDB complexes, with the Gene Neighbor method providing good

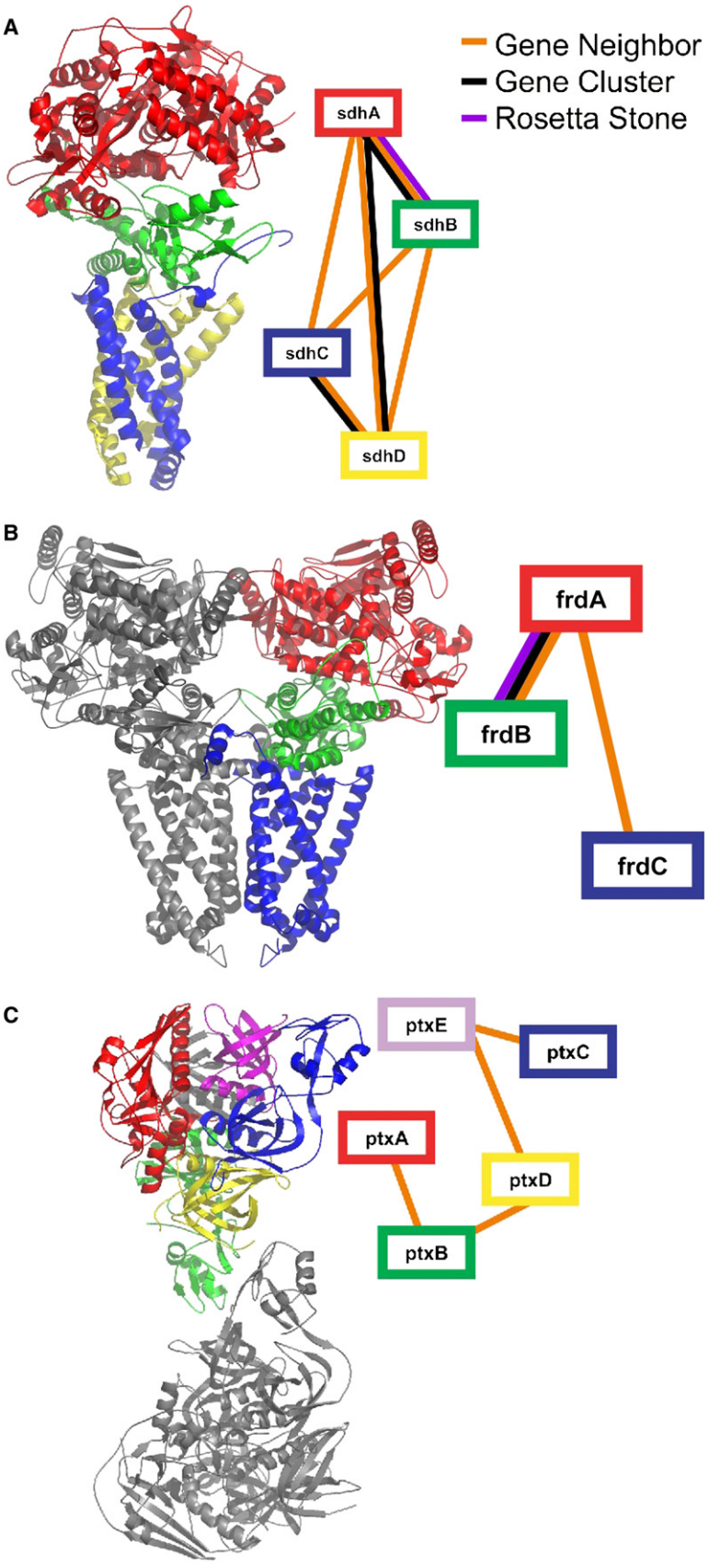

Figure 3. Functional Linkage Found within the Prokaryotic PDB Structures

High-confidence functional linkages are illustrated in the structures in which they are found and also by linkage graphs. Corresponding polypeptide chains in the 3D model and graphs are shown by the color of the chain. (A) *E. coli* succinate dehydrogenase subunit (PDB code: 1NEK). (B) *W. succinogenes* fumarate reductase dimer (1QLB). (C) *B. pertussis* toxin (1PRT).

Table 1. Assessment of Prolinks Methods in Identifying True Positive Interactions in PDB Complexes

Method	No. True Positives	Coverage of Total Interactions (%)	No. Unique True Positives	Coverage of Unique Interactions (%)	No. Unique False Positives	Accuracy of Unique Interactions (%)
Gene neighbor	858	62	329	67	205	63
Gene cluster	245	18	69	14	2	97
Phylogenetic profiles	186	14	71	14	47	60
Rosetta stone	86	6	24	5	3	89
Total	1375		493			

An inferred linkage is accepted as a "true positive" if the two linked proteins are also found within a complex in the PDB. The performance of Prolinks functional links in predicting prokaryotic PDB protein interactions can be gauged by weighing two factors, coverage of unique interactions and accuracy of unique interactions. Coverage is the percentage of one method's true positives divided by total true positives. Accuracy of a method is determined by dividing a method's true positives by the sum of its true positives and false positives. The Gene Neighbor method provides the optimal balance between both characteristics in detecting prokaryotic PDB complexes.

coverage and accuracy of PDB interactions. Out of a total of 1375 high-confidence PDB linkages found between distinct polypeptides chains, there are 493 nonredundant linkages. A majority of these linkages are found in large ribosomal structures consisting of numerous polypeptide chains, such that 51% (708/1375) of total true links and 61% (300/493) of unique true links are derived from ribosomal proteins. The Gene Cluster and Gene Neighbor methods, which use gene proximity and operon structure to determine linkage between proteins, detect 69 (14%) and 329 (67%) of the 493 unique true interactions, respectively. By comparison, the Rosetta Stone method identified 24 (5%) of the unique true links and the Phylogenetic Profile method identified 71 (14%) of the unique true links (Table 1). We can also assess the accuracy with which functional linkages detect protein complex interactions in the PDB. The descending order of accuracy for each method is: Gene Cluster (97%), Rosetta Stone (89%), Gene Neighbor (63%), and Phylogenetic Profiles (60%). Analysis excluding ribosomal complexes did not reveal qualitatively different results. Our work confirms prior analyses that found that the Gene Cluster method, while limited in its coverage of known PDB interactions, was the most accurate measure in identifying multi-protein complexes (Bowers et al., 2004).

Detecting Novel Macromolecular Complexes for Structure Determination

Having demonstrated that Prolinks linkages detect physical interactions within PDB complexes, we sought to select a set of computational linkages corresponding to novel protein complexes that might facilitate future determination of macromolecular structures. We began by identifying a large training set of known macromolecular contacts in the PDB database (as described above) and experimentally validated protein complexes annotated from the scientific literature (Keseler et al., 2005). The four Prolinks computational methods—Gene Neighbor, Phylogenetic Profile, Rosetta Stone, and Gene Cluster—detect 31,716 high-confidence functional linkages for the complete set of *E. coli* proteins. The set of high-confidence Prolinks functional linkages, taken in its entirety, was able to identify 639 of the 929 (>68% coverage) annotated *E. coli* complex interactions identified from the EcoCyc and PDB databases. We can calculate the probability of this number of linkages being shared between *E. coli* Prolinks linkages and the PDB set of interactions, assuming that each interaction set was chosen randomly from among the possible ~9.3 million interaction pairs representing 4319 unique *E. coli* proteins. A simple calculation reveals that the probability of randomly observing $\geq$639 interactions in common between these two sets is $p < 1.0x10^{-200}$, showing definitively that Prolinks linkages are massively enriched for known *E. coli* protein complexes.

Subsets of Prolinks linkages were analyzed for their ability to accurately recover known intermolecular protein complex interactions (Table 2; Figure 4, upper left). Of the four computational methods, the Gene Neighbor method provided the best combination of coverage and accuracy. The top 3000 Gene Neighbor linkages, corresponding to a functional confidence cut-off of 70%, were able to detect 438 of the 929 macromolecular interactions in the training set (47% coverage) with accuracy of >14% (438/3000) (Figure 4, upper right). The top 3000 Gene Cluster linkages, by contrast, identify 209 of the 929 intermolecular, with both lower accuracy and coverage of the training set. The top 2000 Phylogenetic Profile and Rosetta Stone linkages identified 65 and 44 of the 929 interactions, respectively.

We also explored whether alternate methods for selecting and combining linkages might help improve the predictive properties of Prolinks for detecting protein complexes, and found that completely connected subgraphs (cliques) were not able to further improve accuracy for members of known macromolecular complexes (Table 2) (Carraghan and Pardalos, 1990). The set of cliques contained within the top 4000 Gene Neighbor linkages detected 286 of 929 known *E. coli* complex interactions (30% coverage), limiting the total number of predicted

Table 2. Finding the Best Method of Functional Linkage for Inferring Protein Complexes in the PDB

Prolinks Selected	Clique[a]	A	B	C	No. of Proteins
> = 1 Link, > 0.4 confidence		**31077**	*639*	290	4319
All GN links		**14024**	*590*	339	3385
Top 6000 GN link by p-value		**5480**	*525*	404	2391
(Top 3000 GN links by p-value)		**(2565)**	*(438)*	(491)	(1522)
> = 1 Link, > 0.4 confidence	X	**16803**	*366*	563	3315
Top 5000 GN links by p-value	X	**3116**	*320*	609	1480
> = 1 Link, > 0.6 confidence	X	**6675**	*304*	625	2223
Top 4000 GN links by p-value	X	**2321**	*286*	643	1257
Top 4500 links by P log odds		**4128**	*373*	556	3973
Top 3000 GN links by p-value	X	**1664**	*273*	656	1018
Top 2500 links by P log odds		**2327**	*173*	756	3482
Top 3000 GC links by p-value		**2801**	*209*	720	3929
Top 4500 links by P log odds	X	**1273**	*189*	740	1030
>1 Link, > 0.4 confidence	X	**1549**	*170*	759	916
>1 Link, > 0.4 confidence		**3547**	*287*	642	2895
Top 2000 PP links by p-value		**1940**	*65*	864	1320
Top 2000 RS links by p-value		**1969**	*44*	885	1285
High-throughput experimental		**698**	*32*	897	749
Top 2500 links by P log odds	X	**122**	*26*	903	573

Different combinations of methods and linkages were used in order to selectively recover known and novel physical interactions between components of protein complexes. The first column describes the method used to select functional linkages. Column A describes the number of functional linkages selected that did not correspond to a known intermolecular contact. Column B denotes the number of functional linkages that had corresponded to a known interaction in a macromolecular complex. Column C describes the number of interactions in the training set that were not predicted by a functional linkage. Note that the data in columns A, B, and C match the data used in Figure 4 (bold, italic, and underlined data here corresponding to blue, red, and yellow, respectively in Figure 4), corresponding to predicted and known interactions. The final column lists the number of unique proteins contained within the set of linkages. The method listed in parentheses was chosen as the optimal method to detect protein complexes.
[a] Column denotes whether clique analysis was performed on the dataset.

complex interactions to 2601 predicted pairs, or 11% accuracy. Linkages predicted by multiple methods were able to provide increased accuracy, but at the cost of coverage of the known training set of complex interactions. For instance, linkages detected by two or more high-confidence methods identified 287 interactions from known macromolecular complexes, from a total of 3834 total links, somewhat poorer than the Gene Neighbor method alone. Naive Bayesian methods for computationally combining the linkage methods also failed to improve upon the performance of the Gene Neighbor method (Table 2) (Jansen et al., 2003; Lu et al., 2005), where the first 4500 Bayesian-combined linkages (P log odds) predicted 373 known protein complex interactions from *E. coli*, compared to the first 3000 Gene Neighbor linkages alone, which identified 438 members of the training set. We conclude that Gene Neighbor functional linkages, relative to other individual or combined linkage methods, effectively detects protein complexes and that gene proximity, conserved in multiple genomes, is an key feature of long-lived macromolecular complexes.

Prolinks Linkages Detect Known and Putative *E. coli* Transport Complexes

To illustrate the ability of Prolinks to predict protein complex interactions, we highlight a portion of the entire interaction network containing members of the ABC transport proteins. The ABC superfamily of proteins is typically composed of four components. A periplasmic binding protein binds a heterodimeric channel. Transport of the solute is facilitated by a fourth component ATPase protein. The maltose transport system in *E. coli*, highlighted in Figure 4 (lower left), is composed of these components, including periplasmic maltose-binding protein MalE, the transmembrane channel made up of proteins MalF and MalG, and two copies of the ATPase subunit MalK (Davidson and Nikaido, 1990). Members of this complex form a clique, with Gene Neighbor interactions detected between each of the subcomponents of the complex.

YhbN and YhbG are proteins believed to compose a portion of an ABC transport complex which transports sugars across the membrane (connected by a blue edge, Figure 4, lower left), and offer an example of selecting

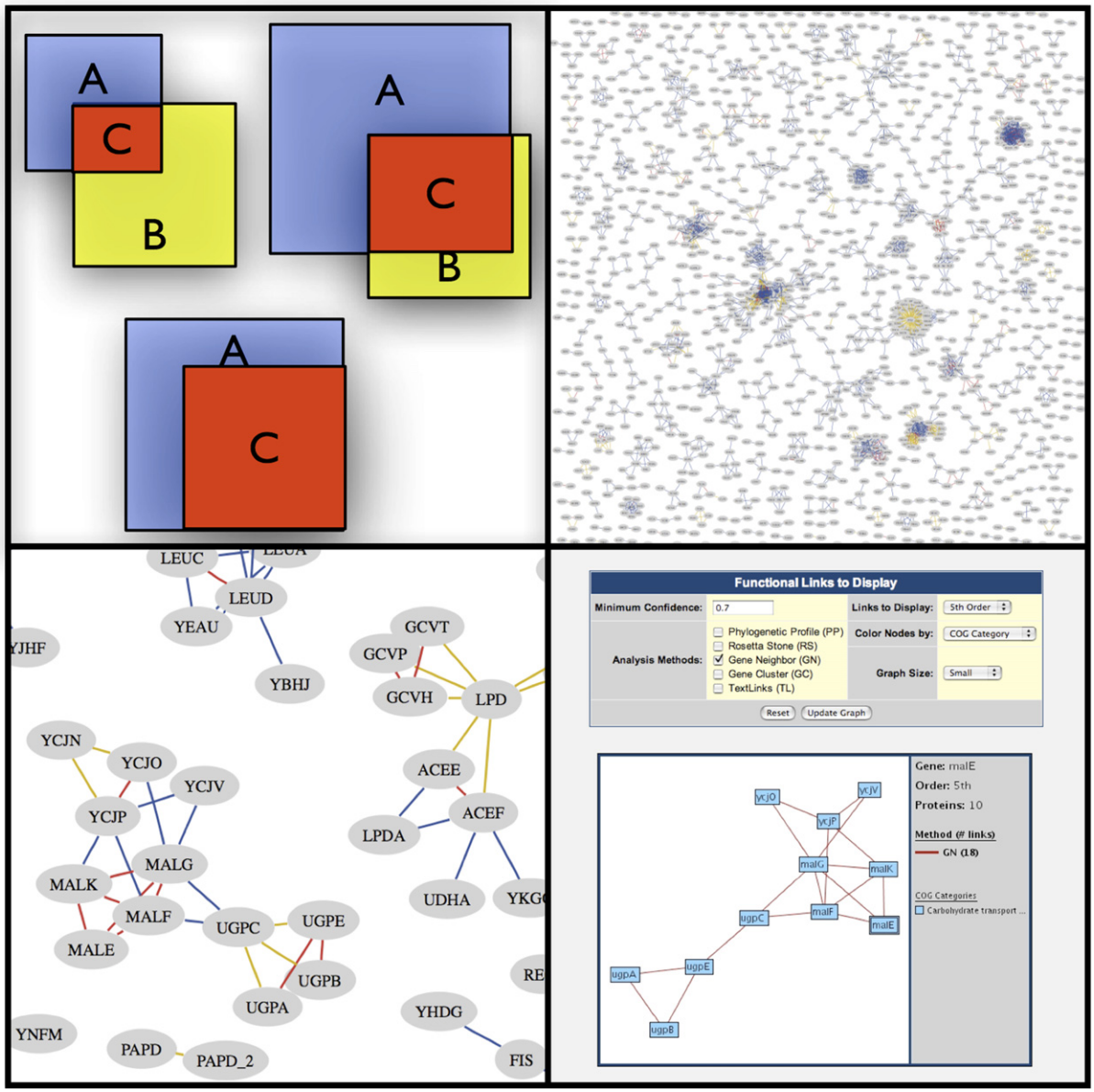

Figure 4. Potential Protein Complexes in *E. coli* Predicted by the Gene Neighbor Method

This figure outlines a solution for identifying contacts between proteins in macromolecular complexes by using functional linkages, graphical representations of the resulting complex predictions. As shown in the Venn diagram in the upper left, genomic inference methods ("A" in upper-left panel) detect functional relationships between proteins, some of which correspond to known contacts between proteins in macromolecular complexes ("C" in upper-left panel). Some of the known complexes, however, may not be predicted by inference methods ("B" in upper-left panel). It is trivial to increase coverage of known protein complex contacts by selecting additional functional linkages (middle diagram), but the fraction of linkages corresponding to complex contacts (accuracy) decreases. We sought an optimal selection of functional linkages (bottom, upper left panel) that would maximize coverage of known protein complexes, while simultaneously improving the reliability or accuracy of new macromolecular contact predictions. The upper right panel displays the set of Gene Neighbor linkages corresponding to predicted protein complex interactions in *E. coli*. In the top graph, links were determined by the set of Gene Neighbor linkages (number of linkages = 3000; functional confidence > 0.7). Links (edges) shown in blue correspond to functional linkages not observed among the set of known EcoCyc or PDB protein complexes. Linkages shown in yellow correspond to the set of known *E. coli* protein complexes not predicted by our selected Gene Neighbor functional linkages. Edges shown in red are both predicted by Gene Neighbor functional linkages and are observed within our training set of experimentally determined protein complexes. The lower right panel shows a local region of the graph, which contains known *E. coli* complexes detected by the Gene Neighbor method, including the MalEFGK maltose ABC transport complex. The lower right panel demonstrates the use of the Prolinks database to select likely protein complexes. Specifying a confidence metric of 0.7, Gene Neighbor linkages only, starting protein MalE from *E. coli*, and graph degree setting of 5, we obtain an equivalent network from Prolinks (http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav), colored by COG functional categories. Three subnetworks are related, each containing a specific ABC transport complex family. Each family sublocalizes within the network, providing an intuitive summary of the *E. coli* multiprotein complexes and their relationships.

complex partners. The members of the complex that compose the heterodimeric transport channel remain undetermined, but our high-confidence linkages suggest that YrbK, a hypothetical protein of unknown function, may fulfill this functional role. The lower right panel of Figure 4 shows an equivalent view of the ABC transport complexes as viewed by the Prolinks database (http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav). Starting the search with the *E. coli* protein MalE, on the main page, Gene Neighbor linkages can be selected with a functional confidence value of 0.70 and graphed, resulting in the network shown Figure 4, lower right. This protein network highlights an approach for the experimentalist selecting complex partners for further crystallographic characterization.

DISCUSSION

The Gene Neighbor Method Identifies Protein Complexes

The Gene Neighbor method for detecting functionally related proteins, which utilizes conserved operon structure in multiple genomes, was found to identify known protein complexes in *E. coli*, as judged by recovery of protein interactions from the PDB and EcoCyc databases. The first 3000 Gene Neighbor linkages were able to identify almost half of all the physical interactions, 438 of 929 (47%), in the training set. A pessimistic assessment would conclude that all remaining 2562 (= 3000 − 438) Gene Neighbor linkages are likely to be false predictions of physical interactions. We note that many of the remaining predicted interacting pairs, however, involve proteins for which neither partner has been characterized in the PDB or Ecocyc databases. If we include in our analysis only those Gene Neighbor linkages for which each protein has been structurally characterized, we again find 438 corrected predicted interactions, and only 532 linkages involving PDB/Ecocyc proteins not observed to interact, yielding an accuracy of 45% (438/532). We anticipate that a similar fraction of the remaining 2030 linkages (438 + 532 + 2030 = 3000) correspond to *E. coli* protein complexes that have yet to be characterized, making these linkages a valuable asset when selecting crystallization targets.

Prolinks functional linkages may identify protein complex interactions not observed by high-throughput experiments. We compared our method of predicting physical interactions against high-throughput experimentally determined complexes as a gauge of success and potential utility to crystallographic structure determination. We found that a proteomic data set of experimentally determined physical interactions in *E. coli* (Butland et al., 2005), consisting of 730 tandem affinity purified (TAP) multisubunit complexes, confirmed by bidirectional baits, identified only 32 of the 929 known *E. coli* protein complexes documented in literature (Table 2). Likewise, only 42 of the 730 HTP *E. coli* interactions were confirmed by the first 3000 Gene Neighbor linkages. It is difficult to determine whether these findings result from a bias in the *E. coli* proteins selected for the study, or from the false-positive rate of the determined interactions within the dataset, or from the nature of physical interactions being identified by the technique.

Prolinks linkages may be able to detect additional physical interactions within PDB complexes beyond the structures discussed here. We detected, by BLAST sequence alignment, numerous functional linkages in almost half of expected nonredundant PDB complexes (107 out of 242) that are common to sequences in both datasets. We used a selective BLAST e value cut-off, determined by our goal to create high-confidence cross-references between Prolinks sequences derived from sequenced genomes, and PDB sequences which are optimized for structural studies via mutations and excisions of problematic domains. Because BLAST e values are a function of alignment length, alignments between shorter or distantly homologous Prolinks and PDB sequences may have been overlooked in our studies.

To What Extent Are Functional Linkages Useful in Structural Biology and Genomics?

From our data, we can offer a crude estimate of how many inferred complexes a practicing structural biologist will need to express to have a 95% probability of crystallizing a complex. For binary complexes, we can estimate the number of constructs that must be prepared to ensure success in identifying physiological partners as $(1-0.45)^n \sim (1-S)$, where n is the number of constructs tested, S is the fractional likelihood of one or more successful partner predictions, and 0.45 is the estimated true positive rate for predicting a complex interaction, as discussed in the results section. Based on these assumptions, and the further assumption of independence of functional linkages, the exploration of 5 pair-wise constructs yields a ~95% likelihood of correctly predicting one or more complexes. Likewise, identifying potential multisubunit complexes for characterization must be guided by both Gene Neighbor linkages and the operon structure of the experimental organism in question. A recent application of this method to proteins, resistant to individual expression and crystallization, has shown that use of Gene Neighbor predictions can lead to successful structure determination of protein complexes (Strong et al., 2006).

Rescued Structures

It is increasingly found that cytosolic proteins are stalled in the structural genomics pipeline at the steps of soluble expression and crystallization. Because cytosolic proteins in the cell exist in complex mixtures of macromolecules and metabolites, we assume the cases of insoluble and unfolded and partially folded proteins encountered in vitro are missing their natural partners. Therefore we have focused our study on the problem of finding the natural protein partners for proteins that fail in vitro to be soluble and crystallizable. Application of the methods proposed here will permit a test of our hypothesis that supplying missing protein partners may diminish pipeline attrition in structural genomics. How many proteins might be effectively rescued by such an approach? In the current *M. tuberculosis*

structural genomics consortium (http://www.doe-mbi.ucla.edu/TB/), there are 1529 defined crystallization targets (Figure 1). Due to attrition in expression, solubility, purification and crystallization phases of structure determination, only 230 of the 1339 cloned proteins have been crystallized. This means that over 1109, or two-thirds of the starting *M. tuberculosis* protein targets have failed to crystallize in the absence of their physiologic partner, and perhaps for other reasons. Based on the earlier determined confidence estimates for Gene Neighbor prediction of complex partners, our methods could, in practice, yield many hundreds of rescued structures ($1109 \times 0.45 \approx 499$), representing over a third of the *Mtb* pipeline ($499/1529 \approx 33\%$).

Conclusions

We investigated the presence of functionally linked proteins in PDB structures. In predicting proteins that interact to form complexes, the Gene Neighbor interactions were the most powerful at detecting interactions within multiprotein structures, while the Gene Cluster and Rosetta Stone algorithms were the most accurate. The Gene Neighbor method provides the optimal balance of coverage and selectivity in our benchmark of solved structures. The conclusion is the same in the case of computationally inferred potential complexes, where Gene Neighbor linkages performed best in an analysis of known *E. coli* complexes. This investigation suggests that functional linkages may outperform recent high-throughput datasets in detecting *E. coli* complexes. The present study of 242 complexes in the PDB suggests that systematic experimental study of complexes inferred from functional linkages could cut the attrition of the pipeline of structural genomics projects.

EXPERIMENTAL PROCEDURES

Identifying Prolinks Interactions in PDB Structures

To identify the structures of complexes in the PDB detected by the Prolinks database of functional linkages, we aligned the August 2004 version of the PDBAA FASTA database (http://ftp.ncbi.nih.gov/blast/db/FASTA/), which contains 17,844 nonredundant PDB chains, against sequences found in Prolinks (595,675 sequences) using BLAST (Altschul et al., 1997). To ensure an accurate cross reference between PDB and Prolinks sequences, an e value upper-bound threshold of $1e^{-20}$ and a local alignment identity greater than or equal to 75% was used. The PDB sequences were grouped by the PDB structure they are found in. Every prokaryotic PDB structure having two or more Prolinks BLAST matches to separate chains was saved for further analysis. Hybrid structures made of a combination of prokaryotic and eukaryotic/viral/synthetic proteins are not included.

A true positive in this study is taken to be a high-confidence (≥ 0.4 confidence) functional linkage reported in the Prolinks database between a pair of different amino acid sequences found within the same PDB structure. If a PDB structure had at least two significant Prolinks BLAST alignments from the same source organism with a high-confidence link, then this pair was considered a true positive functional linkage identified by Prolinks. Sequences were categorized by source organism and number of linkages found between the pairs of sequences.

Detecting Novel Protein Complexes

Four computational inference methods, available from the Prolinks database, were used to detect physical interactions within multisubunit protein complexes. These include the Rosetta Stone, Phylogenetic Profile, Gene Cluster, and Gene Neighbor methods, each of which are described in detail by Bowers et al. (2004). Computationally derived linkages for 168 fully sequenced organisms are available at http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav.

The Phylogenetic Profile method uses the co-occurrence or absence of pairs of nonhomologous genes across genomes to infer relatedness. The underlying model for this method is that pairs of proteins that are often present or absent together within genomes are likely to have coevolved, and may therefore be functionally related. Sequenced genomes allow us to catalog the proteins encoded in each organism, allowing us to determine the pattern of presence and absence of a protein by searching for its orthologs across organisms. The result of such a homology search is an N-dimensional vector of ones and zeroes for the query protein that we call a phylogenetic profile. If we assume that the two proteins, A and B, have evolved independently, we can compute the probability of observing a specific profile overlap by chance by using the hypergeometric distribution. We compute this probability for all pairs of proteins within a genome.

The Gene Cluster method utilizes the fact that proteins with closely related functions are often encoded within a genome in close physical proximity. Operons contain two or more genes, the transcription of which is controlled by a single promoter. Various methods have been developed to identify operon structure within microbial genomes, relying on intergene distance as a predictor (Bowers et al., 2004; Moreno-Hagelsieb and Collado-Vides, 2002; Overbeek et al., 1999a, 1999b). We have found that gene start positions can be modeled by a Poisson distribution, with each nucleotide position having the same probability of being a start site. The probability that a gene starts at any position is given by $P(\text{start}) = me^{-m}$, where m is the total number of genes divided by the number of intergenic nucleotides within the genome. It follows that we can estimate the probability that two sequentially encoded genes are separated by a distance of less than N nucleotides as:

$$P(separation < N) = \int_0^x me^{-mN} = 1 - e^{-mx}$$

The conservation of operon structure across many genomes provides additional evidence that two genes encoded in close physical proximity are functionally coupled and perhaps components of a protein complex. We have developed a novel algorithm (Dandekar et al., 1998), the Gene Neighbor method, that generates a P value for the likelihood that two proteins are coded within a conserved operon. The method first computes the probability that two genes are separated by fewer than d genes:

$$P(\leq d) = \frac{2d}{N-1}$$

where N is the total number of genes in the genome. If the two genes have homologs in other organisms we compute the product of the above probability across these organisms:

$$X = \prod_{i=1}^{m} P_i(\leq d_i) = \prod_{i=1}^{m} \frac{2d_i}{N_i - 1}$$

where m is the number of organisms that contain homologs of the two genes of interest. It can be shown that the probability that two genes are components of a conserved operon is given by:

$$P_m(\leq X) = 1 - P_m(>X) \approx X \sum_{k=0}^{m-1} \frac{(-\ln X)^k}{k!}$$

Occasionally, two proteins expressed separately in one organism can be found as a single chain in the same or a secondary genome. Analysis of gene fusion/division events to infer functional relatedness, commonly known as the Rosetta Stone method, has been described in

detail elsewhere (Dandekar et al., 1998; Ermolaeva et al., 2001; Lee et al., 2004). Proteins comprising consecutive metabolic steps or components of molecular complexes are often expressed as a single polypeptide chains to maximize kinetic or expression efficiency.

The four inference methods were applied to the *E. coli* K-12 genome, yielding 31,716 high-confidence linkages (for 4310 unique protein entities), as determined by keyword recovery on known protein functional categories (Bowers et al., 2004; Marcotte et al., 1999b).

EcoCyc Pairs

A total of 89 protein complexes from the EcoCyc database were used to validate our Prolinks selection criteria, consisting of 713 distinct intracomplex interactions. Self-interactions between homodimers, homotrimers, and the like were excluded from further analysis. All other protein chains contained within the same macromolecular complex, independent of direct physical contact, were considered to form macromolecular contacts for the purpose of our analysis. PDB contacts and Ecocyc interactions were combined to form a training set of 929 protein interactions.

Clique Analysis

Selected sets of Prolinks linkages were analyzed to identify completely connected subgraphs contained within the entire graph of interactions (Carraghan and Pardalos, 1990). A version of the clique analysis algorithm was modified to accept a binary and symmetric matrix of Prolinks linkages augmented with self-interactions (corresponding to the diagonal of the matrix).

Combining Prolinks Linkages

E. coli linkages detected by the four computational methods described above were combined into a single metric for each protein linkage pair (Jansen et al., 2003). The four prediction sets were assumed to be independent, such that a naive Bayesian approach could be used to combine probabilities to arrive at a single probability that two proteins coevolve:

$$O_{post} = \left(\prod_{i=1}^{4} \frac{P(f_i|pos)}{P(f_i|neg)} \right) \frac{P(pos)}{P(neg)}$$

each linkage method for each linkage pair is described by an odds ratio, which describes the ratio of the likelihood of correctly detecting a physical interaction relative to that of incorrectly detecting a physical interaction. Positive pairs are proteins with common complex annotation, and negative pairs are proteins with different complex annotation, as defined by the EcoCyc database. The product of the odds ratios corresponding to the Prolinks predictions for each protein pair is used as the ranking metric for identifying likely physical interactions.

Functional Benchmarking

We assessed keyword category recovery for the four individual methods as a measure of our confidence in the inferred linkages. The confidence measure describes the likelihood that the pair of linked proteins is acting within the same COG pathway (Tatusov et al., 1997), reflecting the number of COG annotated pairs that lie within the same pathway, relative to the total number of annotated pairs. *E. coli* protein pairs used in this paper had a COG pathway confidence recovery of > 0.4, corresponding to a 40% cumulative accuracy of recovering matching pathway annotation.

We calculated the likelihood that two interaction sets would produce k or more pairs in common, under the assumption that interaction sets m and n were chosen randomly from a global interaction set containing a total of N pairs. In this instance, $N = P \times (P - 1)/2$ where P is equal to the total number of distinct *E. coli* proteins contained in both interaction sets. Given these assumptions, we can compute the probability of observing a specific overlap between two interaction sets by chance by using the hypergeometric distribution:

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

Supplemental Data

Supplemental Data include one table that is available online at http://www.structure.org/cgi/content/full/15/9/1079/DC1/.

ACKNOWLEDGMENTS

We thank the Department of Energy–OBER, Howard Hughes Medical Institute, and National Institutes of Health for Support.

Received: November 18, 2005
Revised: May 25, 2007
Accepted: June 1, 2007
Published: September 11, 2007

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., and Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. *5*, R35.

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., et al. (2005). Interaction network containing conserved and essential protein complexes in *Escherichia coli*. Nature *433*, 531–537.

Carraghan, R., and Pardalos, P.M. (1990). An exact algorithm for the maximum clique problem. Oper. Res. Lett. *9*, 375–382.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. *23*, 324–328.

Davidson, A.L., and Nikaido, H. (1990). Overproduction, solubilization, and reconstitution of the maltose transport system from *Escherichia coli*. J. Biol. Chem. *265*, 4254–4260.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86–90.

Ermolaeva, M.D., White, O., and Salzberg, S.L. (2001). Prediction of operons in microbial genomes. Nucleic Acids Res. *29*, 1216–1221.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of *Drosophila melanogaster*. Science *302*, 1727–1736.

Goulding, C.W., Apostol, M., Anderson, D.H., Gill, H.S., Smith, C.V., Kuo, M.R., Yang, J.K., Waldo, G.S., Suh, S.W., Chauhan, R., et al. (2002). The TB structural genomics consortium: providing a structural foundation for drug discovery. Curr. Drug Targets Infect. Disord. *2*, 121–141.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature *415*, 180–183.

Hood, L., Heath, J.R., Phelps, M.E., and Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. Science *306*, 640–643.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. Annu. Rev. Genomics Hum. Genet. *2*, 343–372.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA *98*, 4569–4574.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449–453.

Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., and Karp, P.D. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res. *33*, D334–D337.

Kim, S.-H., Shin, D.-H., Choi, I.-G., Schulze-Gahmen, U., Chen, S., and Kim, R. (2003). Structure-based functional inference in structural genomics. J. Struct. Funct. Genomics *4*, 129–135.

Kim, Y., Dementieva, I., Zhou, M., Wu, R., Lezondra, L., Quartey, P., Joachimiak, G., Korolev, O., Li, H., and Joachimiak, A. (2004). Automation of protein purification for structural genomics. J. Struct. Funct. Genomics *5*, 111–118.

Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. Science *306*, 1555–1558.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. (2004). A map of the interactome network of the metazoan *C. elegans*. Science *303*, 540–543.

Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. Genome Res. *15*, 945–953.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751–753.

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. Nature *402*, 83–86.

Moreno-Hagelsieb, G., and Collado-Vides, J. (2002). A powerful nonhomology method for the prediction of operons in prokaryotes. Bioinformatics *18* (*Suppl 1*), S329–S336.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999a). Use of contiguity on the chromosome to predict functional coupling. In Silico Biol. *1*, 93–108.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999b). The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA *96*, 2896–2901.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA *96*, 4285–4288.

Shen, W., Yun, S., Tam, B., Dalal, K., and Pio, F.F. (2005). Target selection of soluble protein complexes for structural proteomics studies. Proteome Sci. *3*, 3.

Strong, M., Sawaya, M.R., Wang, S., Phylip, M., Cascio, D., and Eisenberg, D. (2006). Towards the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. USA *103*, 8060–8065.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631–637.

Terwilliger, T.C., Park, M.S., Waldo, G.S., Berendzen, J., Hung, L.W., Kim, C.Y., Smith, C.V., Sacchettini, J.C., Bellinzoni, M., Bossi, R., et al. (2003). The TB structural genomics consortium: a resource for *Mycobacterium tuberculosis* biology. Tuberculosis (Edinb.) *83*, 223–249.

Terwilliger, T.C. (2004). Structures and technology for biologists. Nat. Struct. Mol. Biol. *11*, 296–297.

Todd, A.E., Marsden, R.L., Thornton, J.M., and Orengo, C.A. (2005). Progress of structural genomics initiatives: an analysis of solved target structures. J. Mol. Biol. *348*, 1235–1260.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature *403*, 623–627.

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

Yakunin, A.F., Yee, A.A., Savchenko, A., Edwards, A.M., and Arrowsmith, C.H. (2004). Structural proteomics: a tool for genome annotation. Curr. Opin. Chem. Biol. *8*, 42–48.

Zhang, C., and Kim, S.-H. (2003). Overview of structural genomics: from structure to function. Curr. Opin. Chem. Biol. *7*, 28–32.