

Inferred Biomolecular Interaction Server—a web server to analyze and predict protein interacting partners and binding sites

Benjamin A. Shoemaker, Dachuan Zhang, Ratna R. Thangudu, Manoj Tyagi, Jessica H. Fong, Aron Marchler-Bauer, Stephen H. Bryant, Thomas Madej* and Anna R. Panchenko*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received August 15, 2009; Revised September 16, 2009; Accepted September 21, 2009

ABSTRACT

IBIS is the NCBI Inferred Biomolecular Interaction Server. This server organizes, analyzes and predicts interaction partners and locations of binding sites in proteins. IBIS provides annotations for different types of binding partners (protein, chemical, nucleic acid and peptides), and facilitates the mapping of a comprehensive biomolecular interaction network for a given protein query. IBIS reports interactions observed in experimentally determined structural complexes of a given protein, and at the same time IBIS infers binding sites/interacting partners by inspecting protein complexes formed by homologous proteins. Similar binding sites are clustered together based on their sequence and structure conservation. To emphasize biologically relevant binding sites, several algorithms are used for verification in terms of evolutionary conservation, biological importance of binding partners, size and stability of interfaces, as well as evidence from the published literature. IBIS is updated regularly and is freely accessible via <http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.html>.

INTRODUCTION

Proteins function by interacting with other biomolecules, and a complete protein functional annotation is impossible without knowledge of the protein interactions. Mapping biomolecular interactions is invaluable in deciphering the interactome, the entire set of molecular interactions in a cell. Recent advances in the experimental and computational tools for identifying proteins and their

complexes have spawned a wealth of information that encourages such a mapping (1,2).

The most successful function prediction methods rely on evolutionary relationships between proteins and the conservation of their molecular function; they look for sequence similarities between unknown queries and functionally annotated proteins (3,4). A similar approach has been used to infer protein interaction partners from a set of homologous proteins, where an interaction between two proteins is predicted if this interaction has been observed between orthologs (interologs) in other species (5). Homology inference methods have certain limitations, though. Common descent does not necessarily imply similarity in function or interactions, and annotations transferred from one homologous protein to another may result in incorrect functional or interolog assignment at larger evolutionary distances (3,6–8). To verify and guide annotations, it is often essential to detect functionally important binding sites. Current binding site prediction methods can be subdivided into several major categories: those which use evolutionary conservation of binding site motifs, those which use information about a structure of a complex, and docking methods (9).

The knowledge of protein structure may facilitate and improve the annotation of protein function and the characterization of protein binding partners and binding sites. Structure-based methods use detailed knowledge of the protein structure to identify binding sites on the basis of the physico-chemical properties of individual residues, their electrostatic contribution, and their location in the 3D structure (10–14). A number of servers have been developed for predicting protein binding sites from structures by locating the binding pockets, by identifying sequence and structural features of homologous proteins which are important for binding, or by using threading and other approaches (14–22).

*To whom correspondence should be addressed. Tel: +1 301 435 5891; Fax: +1 301 480 4637; Email: panch@ncbi.nlm.nih.gov
Correspondence may also be addressed to Thomas Madej. Tel: +1 301 435 5998; Fax: +1 301 480 4637; Email: madej@ncbi.nlm.nih.gov

We have developed a new database and server called IBIS (Inferred Biomolecular Interaction Server), which provides tools to investigate biomolecular interactions observed in a given protein structure together with the complex set of interactions inferred from its close homologs. IBIS identifies and predicts a protein's interaction partners together with the locations of the corresponding binding sites on the protein query. It does not focus on one specific type of interacting molecule, but provides annotations of binding sites for proteins, small chemicals, nucleic acids and peptides (interactions with ions are currently under development). This may allow the mapping of a comprehensive biomolecular interaction network for a given query, depending on the data available for its protein family.

To focus on biologically relevant binding sites, IBIS clusters similar binding sites found in homologous proteins based on the sites' conservation of sequence and structure. Binding sites which appear evolutionarily conserved among non-redundant sets of homologous proteins are given higher priority in the displays. Additionally, binding site clusters are validated by comparing them with binding site annotations from a manually curated subset of the Conserved Domain Database (CDD) (23), if available. In the case of protein-protein binding sites, IBIS also compares its findings to binding interfaces confirmed by the PISA algorithm (24), which estimates the stability of protein-protein interfaces observed in crystal structures. After binding sites are clustered, position specific score matrices (PSSMs) are constructed from the corresponding binding site alignments. Together with other measures, the PSSMs are subsequently used to rank binding sites to assess how well they match the query, and to gauge the biological relevance of binding sites with respect to the query.

MATERIAL AND METHODS

The current release of the Molecular Modeling Database (MMDB) (25), an automatically parsed and validated derivative of the Protein Data Bank (PDB) (26) hosted by the National Center for Biotechnology Information (NCBI) is used in this study. MMDB addresses several issues in interpreting PDB's 3D structure data and provides standardized structural information. For example, MMDB attempts to fix atom name ambiguity, establishes chemical graphs that contain explicit bonding information, extracts biopolymer sequences and small molecules that get deposited into corresponding databases, and cross-references its entries to GenBank, PubMed, the NCBI taxonomy database and PubChem.

Defining a unit of interaction

Protein-protein interactions are identified and analyzed on the level of domains. The Conserved Domain Search service (CD-Search) provides domain annotation for query sequences and pre-computed annotation for the majority of all entries in NCBI's Entrez Protein database (27). If a complete protein chain is used as a

query, protein-protein interaction annotations are provided separately for each domain identified on this query. Interactions between protein domains may occur on the same protein chain, not involving any other molecule. For other types of binding partners (chemicals, nucleic acids and peptides), interactions are defined for a complete protein chain regardless of its domain annotations and always involve another molecule. One effort that is made to reduce nonbiological contacts regards the case of a chemical that interacts with multiple chains. If contacts to that chemical are dominated by one of the chains (>75%), then its interactions with the other proteins are not considered; otherwise each protein interaction with the chemical will be listed separately.

Defining interactions and binding site residues

An interaction and binding site is defined if a protein has at least five residues in contact with another protein, chemical, DNA, RNA or peptide. Contacts are defined if any of the heavy-atom inter-atomic distances is shorter than 4Å. The binding site is defined as a group of residues which make a contact with a given type of interaction partner. For protein-DNA interactions each DNA strand is considered separately. In the case of protein-chemical interactions, chemical ligands are all validated and standardized (if possible) by the PubChem databases and have explicit links to PubChem (28) which may provide extensive information on their known biological activities. There are two types of interactions and binding sites recorded in IBIS: '*observed*' from experimental structures and '*inferred*' from homologs.

Clustering and Inferring binding sites

A flowchart that summarizes the inference of binding partners and sites on a query is presented in Figure 1. First we collect homologs with known structures and higher than 30% identity to the query. To ensure good quality alignments, the VAST structure-structure comparison algorithm is used (29). If a query does not have a structure, the BLAST heuristic is applied to find the most closely related structure (30). The closest homolog (with an *E*-value <0.01) is picked with a conservative threshold for alignment extent, requiring 80% or more of the query sequence to be aligned.

A binding site cluster represents a collection of structures which are related to the query, and where all members of the cluster contain similar overlapping binding sites when mapped onto the query. Similarity between binding sites is measured in terms of sequence similarity, and those positions which overlap structurally are assigned an additional weight. Binding sites are clustered by a hierarchical complete linkage clustering procedure. To decide on the cutoff for clustering, we use a recently described energy function which maximizes the mean similarity of members within a cluster and minimizes the complexity of the description provided by cluster membership (number of bits required to describe the data) (31). Clusters which contain an actual interaction observed in the query structure are marked by the letter

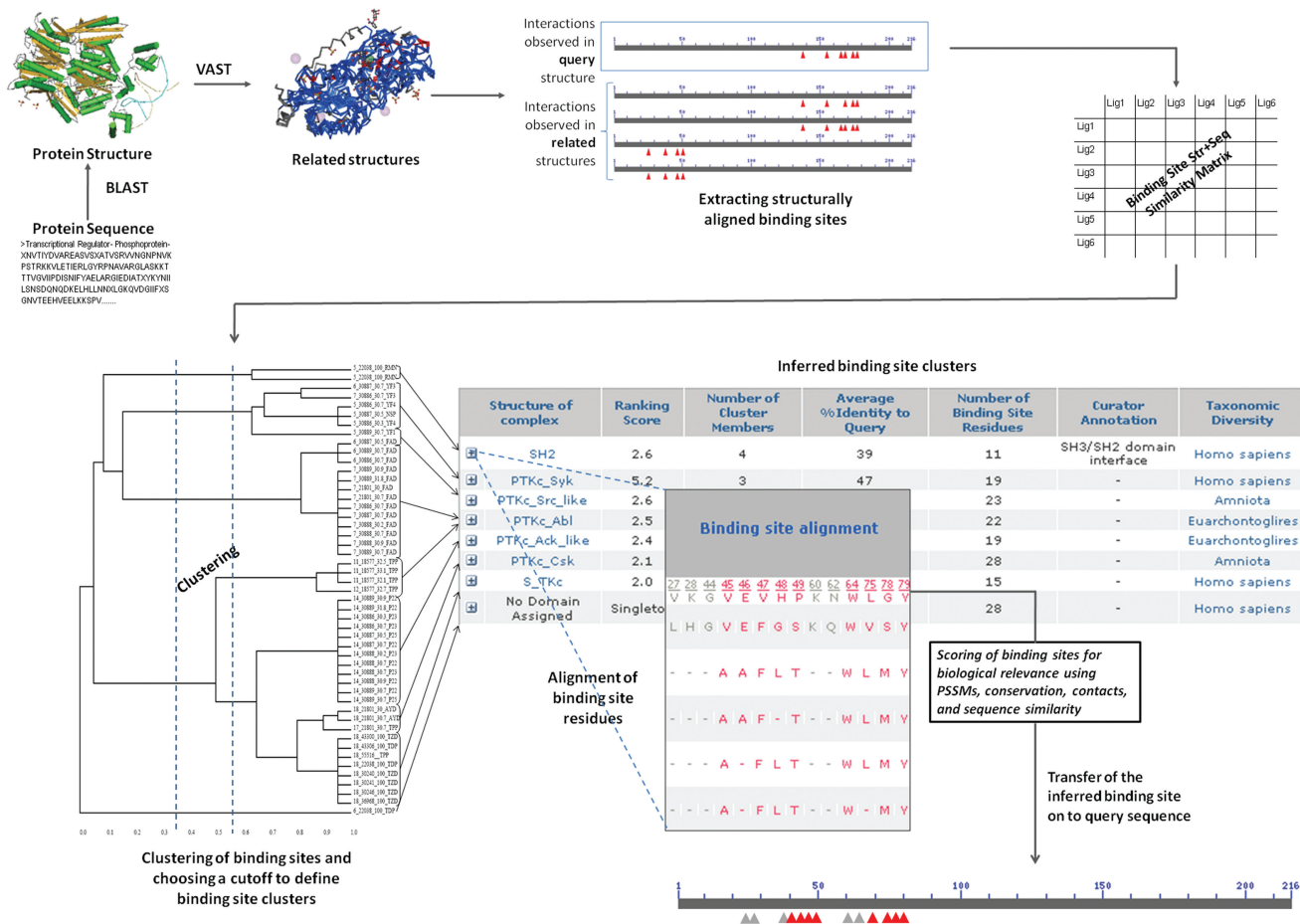


Figure 1. Overview of the binding site annotation procedure in IBIS.

‘O’. By expanding the cluster one can see additional information about its members.

All binding site clusters are ranked in terms of their predicted biological relevance and similarity to the query. The components of the ranking score are the sequence-PSSM score; the average sequence identity between the query and cluster members calculated over the whole structure–structure alignment; the number of interfacial contacts and the average sequence conservation of binding site alignment columns. All components of the ranking score are then normalized and all clusters are ranked with respect to the Z-scores.

Evaluating biological relevance of binding sites

To emphasize biologically relevant binding sites we validate sites according to a few criteria. First, we assess the evolutionary conservation of binding site clusters. Those sites which reoccur in diverse enough protein complexes are ranked higher, an idea which was previously implemented in the Conserved Binding Modes (CBM) database (32). Clusters that have only one non-redundant member (after members with >90% identity are purged) are considered ‘singletons’ and are displayed at the bottom of the interaction summary table with a low rank. Another way to evaluate binding

sites is to compare them with manually curated site annotations from the Conserved Domain Database (CDD), which have been extracted from the published literature or derived from manual interpretation of individual three-dimensional structures (23). Binding site clusters which overlap by >50% with a CDD annotation are ranked first. For protein–chemical interactions, we exclude by default chemicals such as buffers, salts, detergents, solvents and ions that are typically added for the purpose of crystallization and/or purification. Most often, these are not relevant with respect to the protein’s biological function. Finally, we employ the PISA algorithm (24) to validate protein–protein interaction interfaces and eliminate those interfaces which appear to be the result of crystal packing.

RESULTS AND DISCUSSION

Summary statistics of the IBIS database

Currently, a total of 40 716 proteins (151 887 protein chains/domains) are represented in IBIS with at least one type of interaction observed in their structural complexes. As can be seen from Figure 2, protein–protein and protein–chemical interactions are the most

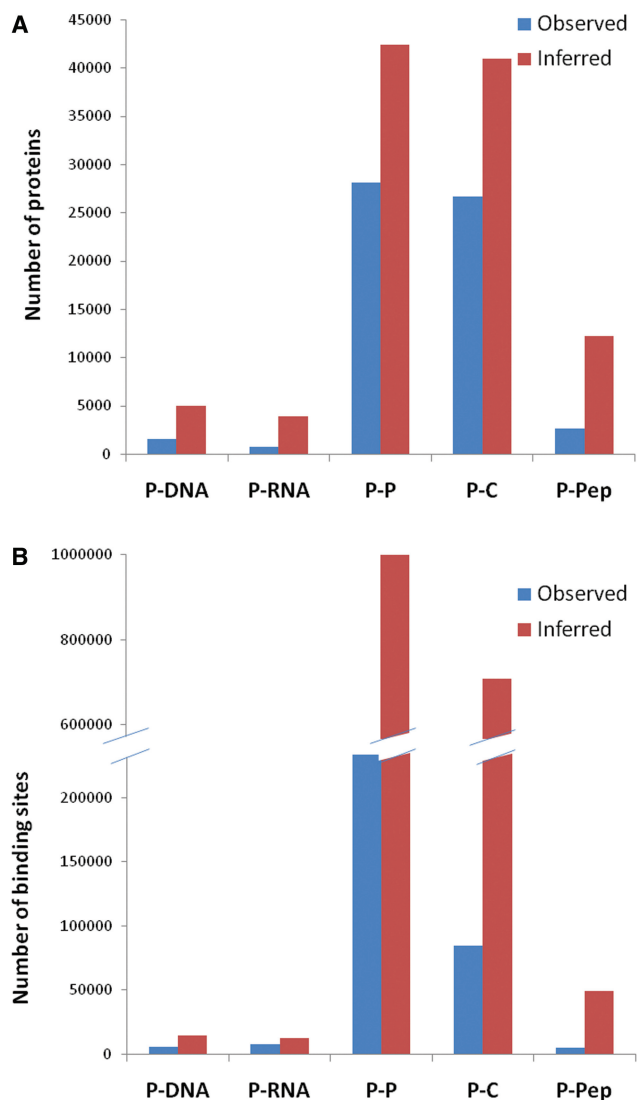


Figure 2. (A) Histogram depicting the number of proteins in PDB with observed/inferred binding sites. (B) Histogram showing the number of binding sites inferred by IBIS as compared to those observed in protein structure complexes.

frequent types of interactions observed in protein structures. Protein–protein interactions are the most prevalent interactions as reflected by the number of domains involved in interactions and the number of binding sites. The number of inferred interactions is always higher than the number of observed interactions, especially for protein–peptide and protein–nucleic acid interactions, where the number of inferred interactions exceeds the number of observed ones (in terms of the number of protein chains) almost 5-fold. This ratio is even higher for binding site clusters (Figure 2B). Altogether, IBIS provides information on binding partners and binding site locations with averages of 3.4 protein–chemical binding site clusters per chain, and eight protein–protein binding site clusters per domain. The scale of such annotations is approaching the scale of whole interactomes.

Description of the IBIS interface

IBIS may be queried by supplying either a protein NCBI GenBank identifier or PDB code (the one letter PDB chain identifier is optional). For a given query, it is possible to see different types of interactions, protein–protein, protein–chemical, protein–DNA, protein–RNA and protein–peptide, by navigating through different tabs at the top of the page (the display of protein–ion interactions is currently under development). Figure 3 illustrates an IBIS Interaction Summary page. Observed and inferred binding site clusters are sorted by the ranking score. Each row in the table corresponds to a binding site cluster and can be expanded to show the cluster members.

The main features of binding sites and interaction partners in the Interaction Summary table are as follows:

‘Interaction partner’—name of the interaction partner which interacts with either the actual query (‘observed’ interactions) or homologs of the query from within a given binding site cluster (‘inferred’ interactions). For protein–protein interactions, the CDD domain name of the binding partner is listed. For protein–chemical interactions, the column reports the name of the chemical bound to a representative member of the cluster. For protein–nucleic acid and protein–peptide interactions, the column reports the sequence of the first 20 biopolymer residues from the interaction partner of a representative cluster member.

‘Ranking score’—the score which ranks the binding site clusters in terms of their biological relevance and similarity to the query. The ranking score is not defined for the ‘singleton’ clusters.

‘Number of cluster members’—the number of cluster members. Upon cluster expansion only non-redundant cluster members are displayed (at <90% identity level). A complete list of members can also be viewed by clicking the ‘See all members’ link.

‘Average percent identity to query’—the average sequence identity between the query and the cluster members calculated over all of their structural alignments with the query.

‘Number of binding site residues’—the union of binding sites mapped from all members of the cluster to the query.

‘Number of chemicals’ (for protein–chemical interactions)—the number of unique, standardized chemicals present in a given binding site cluster.

‘Curator annotation’—binding site annotation from the CDD which overlaps by >50% with the sites annotated by IBIS. Binding site clusters with matching CDD annotation are top-ranked irrespective of their ranking score.

‘Taxonomic diversity’—the last common ancestor of the proteins from a given cluster, listed with a link to NCBI’s Taxonomy Browser, so that one can explore all taxonomic groups represented by the cluster.

The actual binding site residue alignment can be seen upon expanding the clusters, including the PDB codes corresponding to all complex structures summarized by the clusters. It is also possible to view the inferred binding sites projected onto the actual query structure using the Cn3D visualization software (<http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>). For the case of


```

>|pdb|1U59|A|S| Chain A, Crystal Structure Of The Zap-70 Kinase Domain In Complex
With Staurosporine
Length=287

Score = 356 bits (913), Expect = 6e-99, Method: Compositional matrix adjust.
Identities = 169/277 (61%), Positives = 215/277 (77%), Gaps = 2/277 (0%)

Query 9 KEVYLDKRLKLLLEDEKLSGNGFTYKGGYQMKKVKTVAKIKLKNKNEANDPALKDELLAE 68
K++L R L + D ELG GNGFVY+G Y+M+K VAKK+LK + + A +E++ E
Sbjct 2 KKLFLKRDNLLLIADIELGCGNGFQVROGVYRMRKQIDVAKIKLKNKNEANDPALKDELLAE 60

Query 69 ANVMOQLDNPYIVRMIGICEAESWMLVMEAEIIGPLMKYLOQRH-VDKNIIELVHQS 127
A +M QLDNPIYVR+IG+C+AE+ MLVMEAEIIGPL+K+L R + N+ EL+HQVS
Sbjct 61 AQIMHQLDNPYIVRLIGVCAEALMLVMEAEIIGPLHKFLVQKREEIPVSNVAELLHQVS 120

Query 128 MGHKYLEESNFVHRDLAARNWLLVTHYAKISDFGLSKALRADENNYKAQTHCKWPKWY 187
MGHKYLEE N FVHRDLAARNWLLV +HYAKISDFGLSKAL AD++YY A++ GKWP+KWY
Sbjct 121 MGHKYLEEKNFVHRDLAARNWLLVNRHYAKISDFGLSKALGADDSYTTARSAGKWPWKWY 180

Query 188 APECINYYKFSKSDVWSFGVLMWEAFSGQKPYRGMHKGSEVTAHLEKGERMHCPCAGCPR 247
APECIN+ KFSS+SDVWS+GV MWEA SYGQKPY+ HEG EV A +E+G+RM CP CP
Sbjct 181 APECINFRKFSRSRSDVWSYGVTHWEALSYGQKPYKMKKGPVMAFIEQGRMECPPECPP 240

Query 248 EMYDLMNLCWTDYVENRPGFAAEVRLRLRMYDDVNE 284
E+Y LM+ CW Y E+RP F VE R+R YY + ++
Sbjct 241 ELYALMSDCWIKWEDRPDLTVEQRMRCYYSLASK 277

```

Conserved binding residues in the inferred ATP binding site of 1U59	Query (1XBB) residues overlapping with conserved inferred ATP-binding site residues of 1U59	Native Imatinib contacts to 1XBB Chain A in the ATP binding site
L18	L	L5
		S27
V26	V	V33
A41	A	A48
K43	K	
V73	V	
M88 E89 M90 A91	M E M A	M96 E97 M98 A99
G94	G	E100
		G102
		P103
L142	L	L149
S152 D153	S D	

Figure 4. Mapping of the 1U59 inferred ATP binding site onto the sequence of Syk tyrosine kinase (1XBB chain A) and its agreement with the observed binding site in Syk + complex with imatinib. MMDB residue numbering is used which starts from the beginning of the corresponding GenBank protein sequence.

significant comparative analysis. Here IBIS groups and identifies an ATP-binding site, and provides a list of various chemicals, among them many kinase inhibitors, which might potentially bind to and inhibit the query protein. All binding sites observed in the actual structure complex with the anticancer drug imatinib (1XBB) are correctly annotated by IBIS (see table in Figure 4). Interestingly, imatinib binds not only to the ATP-binding site but also to a regulatory myristoylation site on the C-terminus (from the binding site cluster #8) that can be annotated on the query sequence.

In addition to chemical binding sites, it is also possible to predict protein interaction partners for the Syk protein. For example, binding site cluster #1 under protein-protein interactions points to a potential SH2 domain binding site which is further validated by CDD curator annotation, although no structural complexes have been solved between Syk and SH2.

DISCUSSION

In this paper, we presented a comprehensive, web-accessible database, which organizes, analyzes and predicts different types of interaction partners and binding sites in proteins. For proteins with or without known binding partners, IBIS provides a succinct and informative representation of observed binding sites and binding sites inferred from homologs with known 3D structure. It provides analysis of how well a binding site is conserved across members of a homologous protein family. Several structures of the same protein or close homologs with different binding partners may be available

in the Protein Data Bank, or the same protein may have been crystallized under different physiological conditions. In such cases, the IBIS database facilitates a detailed classification and analysis of binding sites. IBIS also attempts to validate binding sites by assessing their biological relevance and ranks them accordingly. It can be used to annotate oligomeric states by inferring relevant homo-oligomer interfaces and should prove useful in studying the evolution of protein interactions.

IBIS is updated regularly (currently on a biweekly schedule) to account for the growth of the GenBank, PDB/MMDB, VAST and CDD databases. Recently, it was estimated that almost half of all sequences in the GenBank database have at least one structure homolog with an extensive alignment and at least 30% identical residues (34). As the on-going structural genomics initiative continues to close the sequence-structure gap, IBIS serves as a powerful knowledge-based annotation system for proteins of unknown structure.

ACKNOWLEDGEMENTS

The authors would like to thank Yanli Wang and Lewis Geer for useful discussions and Eugene Krissinel for help with the PISA software.

FUNDING

National Institutes of Health/DHHS (Intramural Research program of the National Library of Medicine). Funding for open access charge: National Institutes of

Health/DHHS (Intramural Research program of the National Library of Medicine).

Conflict of interest statement. None declared.

REFERENCES

- 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.
- 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.
- Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences—where are the bottlenecks? *Nat. Genet.*, **18**, 313–318.
- Rentzsch,R. and Orengo,C.A. (2009) Protein function prediction—the power of multiplicity. *Trends Biotechnol.*, **27**, 210–219.
- Matthews,L.R., Vaglio,P., Reboul,J., Ge,H., Davis,B.P., Garrels,J., Vincent,S. and Vidal,M. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or ‘interologs’. *Genome Res.*, **11**, 2120–2126.
- Gerlt,J.A. and Babbitt,P.C. (2000) Can sequence determine function? *Genome Biol.*, **1**, REVIEWS0005.
- Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.D., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.
- Hegyi,H. and Gerstein,M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.*, **288**, 147–164.
- Campbell,S.J., Gold,N.D., Jackson,R.M. and Westhead,D.R. (2003) Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.*, **13**, 389–395.
- Jones,S. and Thornton,J.M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–232.
- Teichmann,S.A., Murzin,A.G. and Chothia,C. (2001) Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.*, **11**, 354–363.
- Landgraf,R., Xenarios,I. and Eisenberg,D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.
- Pazos,F. and Sternberg,M.J. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA*, **101**, 14754–14759.
- Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA*, **105**, 129–134.
- Hernandez,M., Ghersi,D. and Sanchez,R. (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.*, **37**, W413–W416.
- Huang,B. and Schroeder,M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol.*, **6**, 19.
- Laurie,A.T. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.
- Qin,S. and Zhou,H.X. (2007) meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, **23**, 3386–3387.
- Talavera,D., Laskowski,R.A. and Thornton,J.M. (2009) WSSas: a web service for the annotation of functional residues through structural homologues. *Bioinformatics*, **25**, 1192–1194.
- Snyder,K.A., Feldman,H.J., Dumontier,M., Salama,J.J. and Hogue,C.W. (2006) Domain-based small molecule binding site annotation. *BMC Bioinformatics*, **7**, 152.
- Chen,Y.C., Lo,Y.S., Hsu,W.C. and Yang,J.M. (2007) 3D-partner: a web server to infer interacting partners and binding models. *Nucleic Acids Res.*, **35**, W561–567.
- Stein,A., Panjkovich,A. and Aloy,P. (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.*, **37**, D300–D304.
- Marchler-Bauer,A., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M. *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, **37**, D205–210.
- Krisinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.
- Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) MMDB: Entrez’s 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.
- Sussman,J.L., Lin,D., Jiang,J., Manning,N.O., Prilusky,J., Ritter,O. and Abola,E.E. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 1078–1084.
- Marchler-Bauer,A. and Bryant,S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.*, **32**, W327–W331.
- Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
- Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
- Wang,Y., Bryant,S., Tatusov,R. and Tatusova,T. (2000) Links from genome proteins to known 3-D structures. *Genome Res.*, **10**, 1643–1647.
- Slonim,N., Atwal,G.S., Tkacik,G. and Bialek,W. (2005) Information-based clustering. *Proc. Natl Acad. Sci. USA*, **102**, 18297–18302.
- Shoemaker,B.A., Panchenko,A.R. and Bryant,S.H. (2006) Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein Sci.*, **15**, 352–361.
- Atwell,S., Adams,J.M., Badger,J., Buchanan,M.D., Feil,I.K., Froning,K.J., Gao,X., Hendle,J., Keegan,K., Leon,B.C. *et al.* (2004) A novel mode of Gleevec binding is revealed by the structure of spleen tyrosine kinase. *J. Biol. Chem.*, **279**, 55827–55832.
- Wang,Y., Address,K.J., Chen,J., Geer,L.Y., He,J., He,S., Lu,S., Madej,T., Marchler-Bauer,A., Thiessen,P.A. *et al.* (2007) MMDB: annotating protein sequences with Entrez’s 3D-structure database. *Nucleic Acids Res.*, **35**, D298–D300.