Inferring Protein-protein Interactions based on Conservation of Interfaces in Homologs

Manoj Tyagi, Ratna R. Thangudu, Benjamin A. Shoemaker, Stephen H. Bryant, Thomas Madej, Anna R. Panchenko

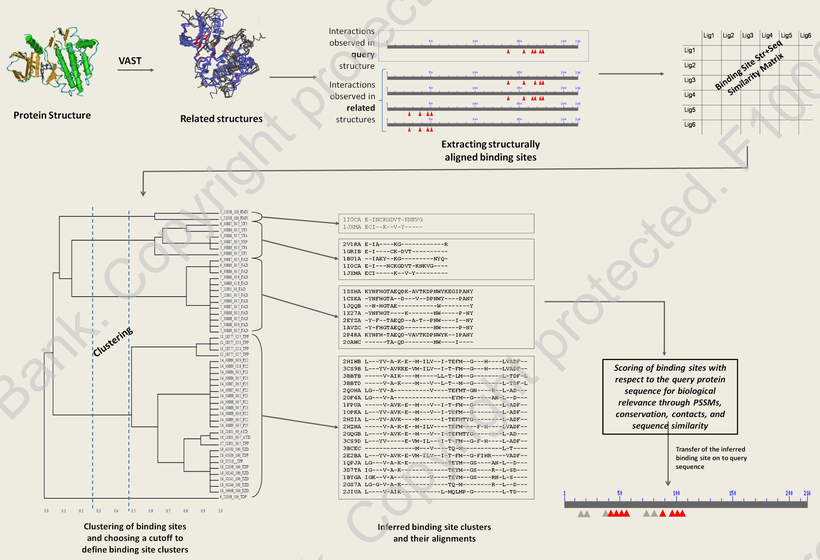
National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA

Abstract

Elucidating protein-protein interactions is an important step towards the understanding of protein functions and the implications for biological processes. Using our newly developed method/server, Inferred Biomolecular Interaction Server (IBIS), we predict and annotate protein-protein interaction partners and binding sites by inference from homologs. This allows us to annotate binding sites on protein structures and/or protein sequences without known structure. To ensure biological relevance of binding sites, our method clusters similar binding sites found in homologous proteins based on their sequence and structure conservation. Binding sites coming from stable assemblies of structural homologs and evolutionarily conserved among non-redundant sets of homologs are given higher priority. 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 protein and to better gauge their biological relevance. The method also facilitates a succinct and informative representation of observed and inferred binding sites from homologs, thereby providing the means to analyze conservation and diversity of binding modes.

The IBIS annotation were validated by using a set of manually curated protein-protein binding site annotations. We show that for nearly 95% of queries, the annotated sites appear among the three most highly ranked binding sites. Furthermore, using a set of known crystal packing interfaces and known biological protein-protein interactions we show that our method can distinguish very well between the two sets and achieves specificity and sensitivity up to 89% and 88% respectively. With the availability of more structural data our approach will evolve as a discovery tool for new protein-protein interactions.

Overview of IBIS



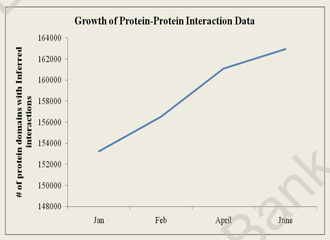
IBIS has been designed to provide annotations for different types of interaction partners (**protein**, **chemical**, **nucleic acid**, **peptide**, and **ions**) of a protein and thus enables us to map comprehensive biomolecular interaction network for a given protein query. IBIS reports interactions observed in experimentally determined structural complexes of a given protein, and at the same time it infers binding sites/interacting partners by inspecting protein complexes formed by homologous proteins. The table shown below gives the current count of observed and inferred interactions for each interaction type and figure shows growth of PPIs over a period of time. IBIS is freely accessible at

http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi

Type of interaction	No of protein domains/chains with Observed interactions	No of protein domains/chains with Inferred interactions	Average no of inferred binding site clusters
P-DNA	3435	11025	2
P-RNA	8253	16148	2
P-P	131545	162965	9
P-Chemical	63908	109936	8
P-Peptide	5011	27237	3
P-Ion	35913	92245	8

1st June 2010

An **observed interaction** is one that actually occurs in the experimental data. An **inferred interaction** is one that is inferred through a homologous structure neighbor. The inferred counts in the table include observed interactions.



Methods

Defining Observed Protein-Protein Interactions:

To record biologically meaningful protein-protein interaction we chose to use the domain as the unit of interaction instead of the full length protein. We annotated domains on each protein chain using the Conserved Domain Search service (CD-Search). If a complete protein chain has multiple domains, domain-domain interaction annotations are provided separately for each domain identified on this query.

An interaction site is defined if a domain has at least 5 residues in contact with another domain. We define a residue to be in contact if there is at least one (heavy) atom of the residue within 4.0Å of some atom from the other domain. All of the residues making a contact constitutes a domain-domain binding site.

All the protein complexes present in the NCBI Molecular Modeling Database (MMDB) are scanned and pairs of interacting domains in a complex are recorded as observed interactions.

Collecting Homologs with Observed Interactions:

We collect structurally similar protein domains (structural neighbors) based on the VAST algorithm and having at least 30% sequence identity to the query. Then we retrieve observed domain-domain interactions for all structure neighbors (including the query domain). Since the alignments may contain gaps, we retain only those instances where at least 75% of the binding site residues (based on 4.0Å contact radii) occur within the structure alignment footprint of the query and neighbor.

Measuring Binding Site Similarity:

To capture the similarity of the binding sites, the similarity measure includes both structural equivalence and sequence similarity terms. The similarity score between two binding sites is defined as:

$$S = \sum S_{ij}$$

$$S_{ij} = H(a_i, a_j)\Delta_{ij} + \boldsymbol{\theta}\Delta_{ij} + w(1 - \Delta_{ij})$$

.

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

$$CS = \frac{S_{(A,B)}}{Max(S_{(A,A)}, S_{(B,B)})}$$

where H is BLOSUM62 score for aligned amino acids in positions i and j; Δ_{ij} is equal to 1 if i and j are aligned or 0 otherwise. θ is an additional weight of "+1" for each structurally equivalent position. w is a gap penalty of "-4". This raw score is converted to a bit score with the statistical parameters λ and K previously defined in the BLOSUM scoring system

The similarity score is then converted into a conservation score CS by dividing by the maximum of the bit scores when the binding sites are scored against themselves.

Clustering of Binding Sites:

Based on the calculated similarity matrix, the binding sites of the homologs are clustered using a complete-linkage clustering algorithm. A distance cutoff value to define the clusters is chosen using a free energy function defined previously. This function F is formulated to maximize the mean similarity of members within a cluster and minimize the complexity of the description provided by cluster membership (Slonim, Atwal et al. 2005).

$$F = \frac{1}{N}\{\sum_{C}\frac{1}{|C|}\sum_{i,j\in C} S(i,j) + T\sum_{C}|C|\log|C| - TN\log N\}$$

Here T is the temperature factor, $S(i,j)$ is the similarity score between binding site i and binding site j in each cluster, C represents a cluster, $|C|$ is the number of binding sites in the cluster C, and N is the total number of binding sites clustered. The temperature T is a parameter (constant) that is chosen so as to correctly balance the energy-like and entropy-like terms in the function.

Biological Relevance: Ranking binding sites

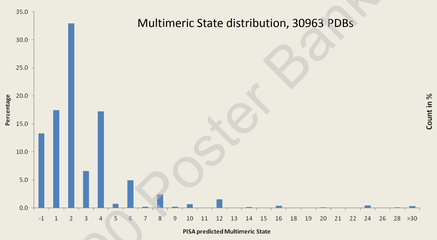
Binding site clusters are ranked in terms of predicted biological relevance and similarity to the query. The ranking score is a weighted sum of four Z-scores: (1) the PSSM score for the query against the PSSM for the cluster; (2) a conservation score, which is a measure of residue conservation for the members of the cluster; (3) a contact count score, which gives a higher score to larger binding sites; and (4) a sequence identity score, which measures the overall sequence similarity between the cluster members and the query sequence/structure.

In addition preference is given to binding sites:

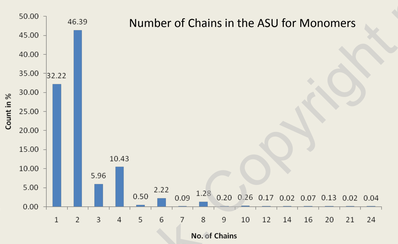
- occuring in two or more non-redundant homologs.
- validated by the PISA algorithm (Krissinel & Henrick,*2007*).
- overlapping with the CDD-defined curated binding site.

Crystallization effect on Interaction Data

A large number of protein structures contain crystal packing interactions due to the definition of the asymmetric unit of the structure. In IBIS we have used a tool called Protein Interfaces, Surfaces and Assemblies (PISA) to eliminate fallacious interactions introduced by crystal packing. PISA performs interface analysis and checks the thermodynamic stability of various assemblies.

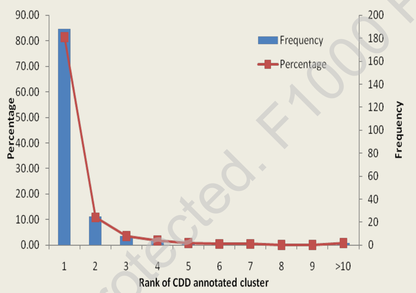


Distribution of multimeric states of structures with observed interactions in IBIS. State -1 represent structures that could not be processed by PISA or where no stable assembly was predicted.



Number of chains in the asymmetric unit for structures predicted as monomeric by PISA. Structures having more than one chain represent potential cases with crystal packing interfaces. PISA validation allows us to flag such interactions.

IBIS Performance



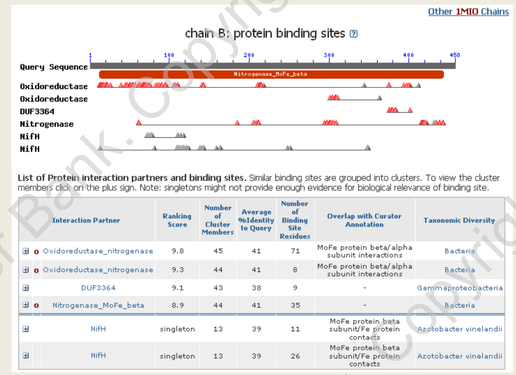
We took a non redundant set of 278 protein chains representing 418 distinct manually curated PPI features from CDD release 2.16v. Using these as queries to IBIS we were able to recover annotations for 83% (231) of the queries. The figure on the left shows the ranking function performs very well, mostly ranking CDD annotated sites within the top three positions. A binding site cluster corresponds to a CDD annotated site if there are at least 50% of the CDD-annotated residues included in the binding site.

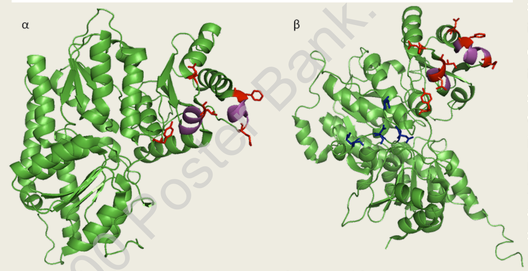
Data Set	Predicted Interaction from multiple homologs	Predicted Interactions from single or more homologs
Crystal Packing Interfaces (76)	8 (89% specificity)	25
Known Protein Interfaces (74)	54	65 (88% sensitivity)

Interactions only validated by PISA in combination with scoring function are considered.

The table on the left shows the ability of IBIS to successfully distinguish the biologically relevant interfaces from the crystal packing interfaces. On two independent datasets taken from the literature, we show that specificity ranges from 89% to 67% where majority of false positive interactions are coming from singletons. IBIS sensitivity varies from 89% to 74% where most of the known interfaces are present in multiple structures.

Fe-protein Binding Sites





Inferred Fe-protein binding sites of the α and β subunits

Nitrogenase complex is a hetero-oligomer, composed of two components, molybdenum-iron (MoFe) protein and iron (Fe) protein. MoFe-protein from *C. pasteurianum* (Cp1) has been crystallized as an α2β2 tetramer without the Fe-protein dimer (PDB 1MIO) and PDB has no complete nitrogenase complex from *C. pasteurianum.* IBIS predicts putative binding sites regions between Fe-protein (NifH domain) and MoFe-protein (β subunit) in Cp1, inferred from nitrogenase complexes of other bacteria systems. Two binding sites between the α and β subunits are also captured. Helical regions assumed to be critical for interaction are shown in magenta. Residues shown with side chains (in red color) are inferred by IBIS and match with residues predicted by Kim at al 1993. Interacting residues in blue are part of an inserted 50-residue sequence, unique to the Cp1 protein and are not predicted by IBIS.

References:
Inferred Biomolecular Interaction Server – a web server to analyze and predict protein interacting partners and binding sites, B.A. Shoemaker, D. Zhang, R.R. Thangudu, M. Tyagi, J.H. Fong, A. Marchler-Bauer, S.H. Bryant, T. Madej, A.R. Panchenko. *Nucleic Acids Res* 2010 Jan; 38 (Database issue): D518-24
Knowledge-based annotation of small molecule binding sites in proteins, R.R. Thangudu, M. Tyagi, B.A. Shoemaker, S.H. Bryant, A.R. Panchenko, T. Madej. *BMC Bioinformatics* 2010, 11:365.

Contact: tyagim@mail.nih.gov