Published in final edited form as:

Curr Protoc Protein Sci. ; 73: 3.9.1–3.9.9. doi:10.1002/0471140864.ps0309s73.

Computational large-scale mapping of protein-protein interactions using structural complexes

Benjamin Shoemaker, Stefan Wuchty, and Anna R. Panchenko

National Institutes of Health, Bethesda, MD 20894, USA

Abstract

Although the identification of protein interactions by high-throughput methods progresses at a fast pace, ‘interactome’ datasets still suffer from high rates of false positives and low coverage. To map the interactome of any organism, this unit presents a computational framework to predict protein-protein or gene-gene interactions utilizing experimentally determined evidence of structural complexes, atomic details of binding interfaces and evolutionary conservation.

INTRODUCTION

Analysis of protein interactions is crucial for our understanding of the mechanisms of cell function, and recent advances in experimental methods for the identification of protein-protein interactions have provided extensive data. In some organisms, such as yeast, the map of protein interactions is close to complete with a relatively high reliability of interactions (Reguly et al, 2006). However, protein interaction data contain a lot of false positives while the coverage still remains low in many other organisms. For example, less than 10% of all human protein interactions have been experimentally determined (Stumpf et al, 2008). One way to fill this gap and provide a more reliable and comprehensive bio-molecular interaction network is to employ computational methods for protein interaction prediction and verification. Many different computational approaches to predict protein interactions exist that are based on genomic context, co-evolution, co-expression or co-occurrence patterns of potentially interacting proteins and their genes (Shoemaker & Panchenko, 2007). Another group of methods rely on similarities between proteins with unknown interactions and homologous proteins with experimentally observed interactions, methods that are called homology inference approaches (Matthews et al, 2001; Walhout et al, 2000). However, it has been shown that interaction partners may only be reliably inferred for close homologs (Dayhoff et al, 2010; Mika & Rost, 2006; Yu et al, 2004). Furthermore, annotations transferred from one homologous protein to another may lead to incorrect interolog assignment even for close homologs if they have different binding specificities. Since binding specificity is usually determined by the structural and sequence features of protein interaction interfaces, the correct detection and transfer of interaction interfaces and binding sites is essential. Current binding site prediction methods use either evolutionary conservation of binding site sequence motifs, information about structures of available complexes, or docking approaches if no such data is available (Campbell et al, 2003).

Although there are many different computational approaches to predict protein interactions, comprehensive interactome mapping for many organisms is still far from complete. The knowledge of protein structure may facilitate and improve the annotation of protein function and the characterization of protein binding partners and binding sites. Since the number of structures of protein complexes increases by a few hundred every month, low-throughput and high-resolution X-ray and NMR methods can be utilized to complement and verify interactions obtained from high-throughput screens and to infer interactions for unknown proteins. 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 (Pazos & Sternberg, 2004; Teichmann et al, 2001).

A few attempts have been made to map the interactome of yeast on a large scale using structural complexes. For example, Aloy and Russell used a homology inference approach to predict interactions in yeast and evaluated reliability of predicted interactions with knowledge based pairwise contact potentials (Aloy & Russell, 2002). Later Kim et al. addressed the same topic by mapping the yeast interactome using sequence similarity between proteins from high-throughput interactions and known protein complexes (Kim et al, 2006). Most recently a PRISM protocol was introduced for the prediction of protein-protein interactions on the proteome scale utilizing known template protein-protein interfaces (Ogmen et al, 2005; Tuncbag et al, 2011). The idea of this approach relies on the similarity of interfaces (not necessarily folds) of interacting proteins. The abovementioned structure based networks found several applications for the determination of disease related genes and annotation of missing protein functions (Huang et al, 2008; Kar et al, 2009; Kuzu et al, 2012; Singh et al, 2010; Tyagi et al, 2012a).

This unit describes how we implement the homology inference approach to annotate the biological interfaces of protein-protein interactions of any given organism. We illustrate the method and framework by mapping the human interactome. In Figure 1, we present a schematic overview of the rationale behind our procedure, and the details are listed in the Commentary section. The strength of our approach is that it ensures close evolutionary relationships between structural complexes and target proteins, and verifies the interactions and binding interfaces by several means. First, it examines the evolutionary conservation between homologous complexes under the assumption that it is more likely to be biologically relevant and not only lineage specific if the binding site is conserved among non-redundant homologs. Second, our approach uses algorithms to infer correct biological units and, finally, applies a rigorous scheme to rank binding sites with respect to their relevance to the target protein. Our method is called IBIS (Inferred Biomolecular Interactions Server) (Shoemaker et al, 2010; Shoemaker et al, 2012) and can be accessed at <http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi>. The IBIS user interface is designed to allow quick access to detailed information about binding sites while providing a comprehensible overview of the complex interaction data at the same time. The Basic Protocol that follows describes the download and processing procedures necessary to map IBIS interaction data inferred from structures onto a complete set of protein sequences from an organism of choice.

Basic Protocol

COMPUTATIONAL PREDICTION OF PROTEIN INTERACTIONS ON THE ORGANISM SCALE USING IBIS

A flowchart of the step-by-step protocol is shown in Fig. 2, and the databases and packages used for this protocol are listed in Table 1. The steps of the protocol are the following.

1. Download a complete set of FASTA protein sequences for a given organism from a protein sequence database. The NCBI RefSeq database (Pruitt et al, 2012) can be used as it provides a non-redundant, comprehensive and curated set of sequences. To obtain FASTA file of RefSeq proteins, go to the NCBI Protein database and search it with the string 'srcdb_refseq[prop] AND "Homo sapiens"[ORGN]' replacing "Homo sapiens" with the desired organism. On the obtained Results page towards the upper-right-side, choose "Send to > File" pull-down menu with Format set to "FASTA" to create a FASTA-formatted file for download.

2. Optional step. Obtain the corresponding NCBI Gene record for each protein sequence; for simplicity, select only the longest protein product from each gene. While gene records are not necessary for mapping protein interaction data onto a proteome, they facilitate gene analysis and are useful to group a gene's protein products and other related RefSeq records. This can be achieved using the NCBI E-utilities retrieval interface which uses URL queries to make Entrez search requests. One convenient wrapper for this interface can be found within the BioPerl package (Stajich, 2007). It has at least two relevant modules, `Bio::DB::EUtilities` and `Bio::SeqIO`, which can be used in a Perl script to help download Gene information of interest from the FASTA input and to help select the longest protein product from each RefSeq gene.
3. Create a local, organism-specific BLAST database from this set of RefSeq proteins for your organism by downloading the *blast+* package (Camacho et al, 2009) and using the program *makeblastdb*.
4. Download the latest interaction data 'ibisdown.tar.gz' from the IBIS FTP site (<ftp://ftp.ncbi.nih.gov/pub/mmdb/ibis/>). A detailed explanation of column definitions can be found on the FTP site in the README.txt file along with a sample file for browsing. The second column gives a code for the type of interaction; only "PPI" (protein-protein interaction) need to be extracted for this protocol. Each line of each unzipped file with the flag "o" (stands for "observed") in column 14 corresponds to the actual experimentally observed interaction, and only these rows should be extracted. PDB codes of interacting proteins can be found in column 13 (four-letter structural complex identifier plus one-letter chain identifiers are separated by an underscore, for example, "1VI6D_1VI6A" corresponds to an interaction between proteins/chains D and A of protein complex 1VI6). For this protocol, however, the other columns also provide potentially useful annotations and criteria by which some interactions might be avoided. If both interacting chains in column 13 are the same, the interaction is intra-chain and may in general be skipped. In addition, selection criteria can include: evolutionary conservation of interaction found in column 8 ("0" for conserved and "1" for non-conserved interaction), biounit validation by PISA algorithm (see next section) found in column 9 ("1" for validated and "0" for non-validated). Column 11 gives the name of corresponding protein/domain family if binding site residues overlap more than 50% with the Conserved Domain Database (CDD) (Marchler-Bauer et al, 2011) annotated sites. The enforcement of each additional criterion would result in a smaller subset of interactions depending on the type or level of evidence required.
5. Using selected lines/rows of the ftp file, compile the pairs of identifiers from column 13 into a list of single protein identifiers (ex: "1VI6D, 1VI6A") with each identifier in an upper-case letter on a separate line with the duplicates removed. Next, parse the binding site locations on both proteins/chains in a complex; namely, look for the corresponding identifier (ex: "1VI6D") in the first column of the ftp file and extract the third column of binding site locations separated by space (these are locations where the residue numbering starts with the first residue of PDB (Berman et al, 2000) sequence). Do the same procedure for "1VI6A" protein in a complex.
6. Using this compiled list of IBIS proteins as queries, search for matches against the local, organism-specific BLAST database with the *blastp* standalone program (suggested E-value threshold = 0.01).
7. Filter BLAST search results to keep only the most significant match for each query which passes the requirements. Namely, ensure that all or almost all binding sites

on IBIS queries are aligned to the organism-specific sequences. Moreover, sequence identity of the alignment should be high enough; we suggest using more than 50% identity. Next disregard alignments/matches that do not pass these thresholds. This produces a map between sequences from a given organism and interacting IBIS proteins, which allows the transfer of interactions/edges from IBIS interacting pairs to the proteins of interest. At the same time, it produces the annotations of binding sites on the organism-specific sequences.

8. Transfer the identifier from the best-matching organism-specific sequences onto the IBIS interacting pairs and remove those interactions in which one or both sequences do not find sufficient similarity to a sequence in the local BLAST database. Optionally record from the IBIS interaction data which interactions are conserved, “biounit” validated, or use other criteria outlined above. This optional data can be used to subset results for varying levels of confidence.
9. Optional step. If gene interaction network is desired, map protein to gene identifiers to create a gene-gene interaction map as described in Step 2.

COMMENTARY

Background Information

Defining interactions and binding site residues—An interaction and binding site is defined if a protein has at least 5 residues in contact with another protein, and two residues are said to be in contact if any of heavy-atom inter-atomic distances is shorter than 4 Å. The binding site is defined as a group of residues which make a contact with a given type of interaction partner. There are two types of interactions and binding sites recorded in IBIS: “*observed*” from experimental structures and “*inferred*” from homologs.

Inferring interactions from homologs—To ensure their biological relevance binding sites are clustered and their sequence and structural similarity is assessed. An overview of the process is shown in Figure 1. Here we present the important details concerning the main steps in the processing.

1. Collecting homologous complexes: A flowchart that summarizes the inference of binding partners and sites on a query is presented in Figure 1. First, IBIS collects homologs with known structures and sequence identity above a certain threshold (25% is used on the webserver). To ensure good quality alignments the VAST structure-structure comparison algorithm is used (Gibrat et al, 1996). To infer interactions based on homology, proteins which are structurally similar to a given query protein are selected (they are referred to as “homologous structure neighbors”). No sequence redundancy filter is applied to remove structures since we showed that removing redundant structures from the consideration decreases the predictive power of the method (Thangudu et al, 2010). If an interaction is inferred from a homooligomeric complex (complex of identical chains), it is called a self-interaction.

2. Measuring similarity between binding sites of different homologs: To cluster the binding sites of the homologous structure neighbors, we construct their structure-structure alignment and define a similarity measure. To capture the similarity of the binding sites, the similarity measure includes both the structural equivalency and sequence similarity. The similarity score between two positions i and j of two binding sites is defined as:

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

where H is the element of the BLOSUM62 matrix corresponding to the aligned amino acids in positions i and j ; Δ_{ij} is equal to 1 if two positions are aligned, and 0 otherwise. θ is an additional weight of “+1” added to each structurally equivalent position. w is a gap penalty of “-4”, to mimic the most unfavorable substitution score from BLOSUM62 matrix, which showed the best performance in our preliminary studies. The overall similarity score between two binding sites is calculated by summing up S_{ij} over all positions in the gapped alignment. To facilitate comparison of scores from different alignments, the raw score is converted to a bit score. The bit-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.

$$CS = \frac{S_{(A,B)}}{\text{Max}(S_{(A,A)}, S_{(B,B)})} \quad (2)$$

3. Clustering and ranking of binding sites: A binding site cluster represents a collection of homologous proteins which contain similar overlapping binding sites. Based on the calculated conservation score CS , the binding sites of the homologs are clustered using a complete-linkage clustering algorithm, which considers the distance between two clusters to be equal to the maximum distance between their members. A distance cutoff value to define the clusters is chosen using a free energy function defined previously (Slonim et al, 2005). This function F is defined to maximize the mean similarity of members within a cluster and minimize the complexity of the description provided by cluster membership (Slonim et al, 2005).

$$F = \frac{1}{N} \left\{ \sum_C \frac{1}{|C|} \sum_{i,j \in C} S(i,j) + T \sum_C |C| \log |C| - TN \log N \right\} \quad (3)$$

where 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 to correctly balance the energy-like and entropy-like terms in the function (Slonim et al, 2005).

All binding site clusters are then ranked in terms of their predicted biological relevance and similarity to the query. To account for evolutionary closeness of a given binding site cluster to the query we use the sequence-PSSM score and the average sequence identity between the query and all cluster members calculated over the full-chain alignment (listed in column 7 of the IBIS ftp files). To rank the larger interfaces more highly we also calculate the average number of interfacial contacts the binding site makes in the complex of the corresponding homolog.

Finally, we parse biological unit information from the PDB files (Madej et al, 2012) and employ PISA algorithm (Krissinel & Henrick, 2007). PISA is an automated method for detecting macromolecular assemblies based on the analysis of interfaces and stability of assemblies reported in crystal structures. Specifically, we use this algorithm to validate protein-protein interaction interfaces and eliminate those which appear to be the result of crystal packing (column 9).

Critical Parameters

There are certain critical points which need to be remembered before embarking on the interactome mapping and interpreting the results. First, unzipping the IBIS ftp file will produce more than 70 thousand files in different directories and require more than 1GB of hard disk space. Second, thresholds used to filter out the alignment at step 7 may greatly affect the network. Namely, using very conservative thresholds (80% sequence identity, 100% binding site overlap, only conserved binding site clusters) may produce reliable interactions but a limited number of them so the coverage of the organism interactome might be an issue. Indeed, upon interpreting the results one should keep in mind that there exists a trade-off between specificity and sensitivity (Tyagi et al, 2012b).

Finally, in a couple of steps, basic text-parsing and script-writing skills, such as for the Perl or Python languages, are necessary. Specifically in Step 2 to optionally map gene to protein sequences and to remove multiple isoforms require scripting which can be assisted with the BioPerl modules. The extraction of interacting sequences from the IBIS data and the subsequent mapping and analysis of BLAST search results with the annotated interactions also require script-writing skills.

Time considerations

Computationally, the most time-consuming step of this protocol is the sequence similarity comparison of the organism-specific sequences to the IBIS interacting sequences using the BLAST algorithm. Depending on the size of the BLAST database and number of query sequences, this search may take several hours or days to complete, but it should be reasonable to run on a desktop PC.

Acknowledgments

This work was supported by the Intramural Research Program of the National Library of Medicine at the U.S.

References

- Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*. 2002; 99:5896–5901. [PubMed: 11972061]
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000; 28:235–242. [PubMed: 10592235]
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009; 10:421. [PubMed: 20003500]
- Campbell SJ, Gold ND, Jackson RM, Westhead DR. Ligand binding: functional site location, similarity and docking. *Curr Opin Struct Biol*. 2003; 13:389–395. [PubMed: 12831892]
- Dayhoff JE, Shoemaker BA, Bryant SH, Panchenko AR. Evolution of protein binding modes in homooligomers. *J Mol Biol*. 2010; 395:860–870. [PubMed: 19879880]
- Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol*. 1996; 6:377–385. [PubMed: 8804824]
- Huang YJ, Hang D, Lu LJ, Tong L, Gerstein MB, Montelione GT. Targeting the human cancer pathway protein interaction network by structural genomics. *Mol Cell Proteomics*. 2008; 7:2048–2060. [PubMed: 18487680]
- Kar G, Gursoy A, Keskin O. Human cancer protein-protein interaction network: a structural perspective. *PLoS Comput Biol*. 2009; 5:e1000601. [PubMed: 20011507]
- Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*. 2006; 314:1938–1941. [PubMed: 17185604]
- Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol*. 2007; 372:774–797. [PubMed: 17681537]

- Kuzu G, Keskin O, Gursoy A, Nussinov R. Constructing structural networks of signaling pathways on the proteome scale. *Curr Opin Struct Biol.* 2012
- Madej T, Address KJ, Fong JH, Geer LY, Geer RC, Lanczycki CJ, Liu C, Lu S, Marchler-Bauer A, Panchenko AR, Chen J, Thiessen PA, Wang Y, Zhang D, Bryant SH. MMDb: 3D structures and macromolecular interactions. *Nucleic Acids Res.* 2012; 40:D461–464. [PubMed: 22135289]
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 2011; 39:D225–229. [PubMed: 21109532]
- Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or “interologs”. *Genome Res.* 2001; 11:2120–2126. [PubMed: 11731503]
- Mika S, Rost B. Protein-protein interactions more conserved within species than across species. *PLoS Comput Biol.* 2006; 2:e79. [PubMed: 16854211]
- Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. PRISM: protein interactions by structural matching. *Nucleic Acids Res.* 2005; 33:W331–336. [PubMed: 15991339]
- Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A.* 2004; 101:14754–14759. [PubMed: 15456910]
- Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 2012; 40:D130–135. [PubMed: 22121212]
- Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskaya OG, Ideker T, Dolinski K, Batada NN, Tyers M. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J Biol.* 2006; 5:11. [PubMed: 16762047]
- Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol.* 2007; 3:e43. [PubMed: 17465672]
- Shoemaker BA, Zhang D, Thangudu RR, Tyagi M, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR. Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites. *Nucleic Acids Res.* 2010; 38:D518–524. [PubMed: 19843613]
- Shoemaker BA, Zhang D, Tyagi M, Thangudu RR, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res.* 2012; 40:D834–840. [PubMed: 22102591]
- Singh R, Park D, Xu J, Hosur R, Berger B. Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic Acids Res.* 2010; 38:W508–515. [PubMed: 20513650]
- Slonim N, Atwal GS, Tkacik G, Bialek W. Information-based clustering. *Proc Natl Acad Sci U S A.* 2005; 102:18297–18302. [PubMed: 16352721]
- Stajich JE. An Introduction to BioPerl. *Methods Mol Biol.* 2007; 406:535–548. [PubMed: 18287711]
- Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. *Proc Natl Acad Sci U S A.* 2008; 105:6959–6964. [PubMed: 18474861]
- Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol.* 2001; 11:354–363. [PubMed: 11406387]
- Thangudu RR, Tyagi M, Shoemaker BA, Bryant SH, Panchenko AR, Madej T. Knowledge-based annotation of small molecule binding sites in proteins. *BMC Bioinformatics.* 2010; 11:365. [PubMed: 20594344]
- Tuncbag N, Gursoy A, Nussinov R, Keskin O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc.* 2011; 6:1341–1354. [PubMed: 21886100]

- Tyagi M, Hashimoto K, Shoemaker BA, Wuchty S, Panchenko AR. Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep.* 2012a; 13:266–271. [PubMed: 22261719]
- Tyagi M, Thangudu RR, Zhang D, Bryant SH, Madej T, Panchenko AR. Homology inference of protein-protein interactions via conserved binding sites. *PloS One.* 2012b
- Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science.* 2000; 287:116–122. [PubMed: 10615043]
- Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* 2004; 14:1107–1118. [PubMed: 15173116]

Figure 1.

The biological rationale allowing the prediction of potential protein-protein interactions from structural protein complexes with known binding sites. In the first step, we collect protein complexes with known structures. In a second step, we use a query protein sequence and subsequently superimpose or align the query sequence on the homologous protein complexes. Assuming that an alignment covers a binding site of the known underlying protein structure, we infer potential protein interactions partners from a homolog to a query protein.

Figure 2.

Flowchart of the computational inference of protein interactions. In particular, we indicate all steps that are described in the basic protocol in detail.

Table 1

Software and data resources used in the protocol

Name	Location	Notes
NCBI Protein (includes RefSeq)	http://www.ncbi.nlm.nih.gov/protein/	Protein sequences with automated and curated annotations
NCBI Gene (includes RefSeq)	http://www.ncbi.nlm.nih.gov/gene/	Gene records including RefSeq, pathways, variations, phenotypes
IBIS Web Interface	http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi	Server for inferring biomolecular interactions from structure
IBIS FTP	ftp://ftp.ncbi.nlm.nih.gov/pub/mmdb/ibis/	Regularly updated full IBIS data dump
NCBI E-utilities	http://www.ncbi.nlm.nih.gov/books/NBK25501/	Programmatic interface to NCBI Entrez search with fixed URL syntax
BioPerl	http://www.bioperl.org/	Open source Perl tools for bioinformatics
blast+ package (blastp, makeblastdb)	http://www.ncbi.nlm.nih.gov/books/NBK52640/	Standalone programs for running BLAST locally
Conserved Domains (CDD, CD-Search)	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	Annotation & classification services from protein domain families
Protein DataBank (PDB)	http://www.rcsb.org/pdb/home/home.do	Databank of biological macromolecular structures
Molecular Modeling Database	http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml	Database of biological structures, mirrors PDB