How well are protein structures annotated in secondary databases?

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · SEPTEMBER 2005

Impact Factor: 2.63 · DOI: 10.1002/prot.20520 · Source: PubMed

3 AUTHORS, INCLUDING:

Ulf Leser
Humboldt-Universität zu Berlin
212 PUBLICATIONS **2,362** CITATIONS

SHORT COMMUNICATION

How Well Are Protein Structures Annotated in Secondary Databases?

Kristian Rother,[1]* Elke Michalsky,[1] and Ulf Leser[2]

[1]*Berlin Center of Genome-Based Bioinformatics (BCB), Institute of Biochemistry at the Charité, Humboldt Universität Berlin, Berlin, Germany*

[2]*Berlin Center of Genome-Based Bioinformatics (BCB), Institute of Computer Science, Humboldt Universität Berlin, Berlin, Germany*

ABSTRACT **We investigated to what extent Protein Data Bank (PDB) entries are annotated with second-party information based on existing cross-references between PDB and 15 other databases. We report 2 interesting findings. First, there is a clear "annotation gap" for structures less than 7 years old for secondary databases that are manually curated. Second, the examined databases overlap with each other quite well, dividing the PDB into 2 well-annotated thirds and one poorly annotated third. Both observations should be taken into account in any study depending on the selection of protein structures by their annotation. Proteins 2005;60:571–576.** © 2005 Wiley-Liss, Inc.

Key words: cross-references; data quality; database annotation; databases; protein structure

BACKGROUND

The number of protein structures solved and deposited in the Protein Data Bank[1] (PDB) is increasing quickly (see http://www.rcsb.org/pdb/holdings.html). Looking at the structure alone is not sufficient for many scientific questions. Rather, structures need to be associated to second-party data, thus putting them into their specific biological context.[2] Examples of such second-party data are the protein sequence and its features, folding classification, functional annotation, taxonomic information, active sites, and the role of the protein in signaling or metabolic processes.

When viewing a structure in the PDB on the Web, some of this information is available in the form of hyperlinks. In the other direction, secondary databases link to PDB entries related to their database objects. Database curators, both of the PDB and of each second-party database, who are trying to keep these hyperlinks consistent and up to date, are struggling with the rapid growth of available data. Despite advances in automated approaches,[3] manual inspection is often necessary due to the complexity of the matter. If annotators cannot keep pace, missing and false links accumulate, thus lowering the utility of database cross-references and increasing the danger of deriving false conclusions.

For a biologist user, it is not obvious what a missing link, for instance, from a PDB structure to a SWISS-PROT sequence, really means: It could be that, although a corresponding SWISS-PROT entry exists, this link is not included in the PDB file (missing link), or it could be that no such sequence entry has been deposited yet into SWISS-PROT (missing data), or that SWISS-PROT does not store sequences of this type (not applicable, as for many immunoglobulins in SWISS-PROT). Typically, users assume "missing data," although both other cases will be shown to occur. This fact must be taken into account when conclusions about sets of structures are derived that depend on second-party annotation.

A major motivation for our study was drawn from our recent initiative to build the Columba database of protein structure annotations. Columba currently integrates PDB entries consisting of 1 or more polymer chains with annotations from 12 second-party databases.[4] Using Columba, users can quickly create subsets of the PDB for which given conditions on various sequence and structure features hold. A typical question may combine conditions on the taxonomy of a structure, its resolution, its participation in metabolic pathways, and its fold classification. However, if links between the PDB and the other databases are incomplete, such queries inherently filter away many PDB entries, not because the conditions would not fit, but because of the missing cross-references. This problem is pertinent for data integration; it affects any system depending on cross-references between objects in different databases, such as the Institute of Molecular

Grant sponsor: German Ministry of Education and Research (BMBF); Grant number: 0312705B.

*Correspondence to: Kristian Rother, Institute of Biochemistry at the Charité, Humboldt Universität Berlin, Monbijoustr. 2, 10117 Berlin, Germany. E-mail: kristian.rother@charite.de

Received 1 September 2004; Revised 25 January 2005; Accepted 8 February 2005

Published online 14 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20520

TABLE I. Data Sources Examined in This Study

Database	Version	No. of entries	No. of referenced PDB entries	Coverage [%]	Description
PDB	10/2004	27,489	27,489	100	Protein structures
SCOP	12/2003	20,572	20,572	74.8	Fold/family classification
CATH	01/2004	17,095	17,095	62.2	Fold/family classification
DALI	05/2003	17,451	17,451	63.5	Fold/family classification
CE	02/2003	17,478	17,478	63.6	Fold/family classification
HSSP	10/2004	25,829	25,690	93.4	Fold/family classification
HOMSTRAD	10/2004	14,940	10,593	38.5	Fold/family classification
SWISS-PROT	10/2004	162,897	20,252	73.7	Protein sequences
PDBSprotEC	10/2004	6515	20,939	76.2	Protein sequence links
UniProt	10/2004	153,713	20,255	73.7	Protein sequences
InterPro	07/2004	153,325	14,940	54.3	Protein sequences
ENZYME	06/2004	4290	12,264	44.6 (92.2)	Enzyme database
KEGG	10/2004	1988	6971	25.4 (52.4)	Enzyme/pathway database
BRENDA	10/2004	4376	11,046	40.2 (83.1)	Enzyme database
PSE.ENZYME	10/2004	1091	12,179	44.3 (91.6)	Enzyme links
GOA	10/2004	5,031,759	20,794	75.7	Functional annotation
NCBI	09/2004	230,559	20,932	76.1	Taxonomic annotation

Column 3 gives the total number of entities contained in the particular databases, that is, protein structures in the fold/family classification block, SWISS-PROT entries in the sequence block, and distinct EC numbers in the enzyme block. Column 4 contains the number of PDB entries linked by this annotation source. Column 5 shows the percentage of these entries on the whole PDB. The numbers in brackets refer to the total number of enzymes given by the PDB (13,296). The PSE.ENZYME line corresponds to all enzyme references from PDBSprotEC.

Biology (IMB) Image Library,[5] InterPro,[6] PDBSum,[7] and also general-purpose data integration systems such as the Sequence Retrieval System (SRS).[8]

In this study we investigated to what extent PDB entries are covered by second-party annotation. We examined linking of 27,489 PDB entries to 6 fold/family classification databases, 4 protein sequence resources, 4 enzyme and pathway catalogs, and 2 databases of functional and taxonomic classification.

ANALYZED DATA

We analyzed 27,489 PDB entries containing 63,611 polymer chains, as of October 2004.[1] The PDB archive forms a global stockpile of protein structures. There are no "releases," as common to many other databases, and the frequency of different types of proteins in the PDB is highly nonuniform. According to Structural Classification of Proteins (SCOP) superfamilies, there were 820 lysozymelike PDB entries, 791 immunoglobulins, 712 trypsinlike proteases, and 420 globins already composing 10% of the entire database. We found 1395 entries that did not contain a polypeptide chain, and 160 entries resolved with a nonstandard method (not X-ray and not NMR).

We parsed links from the 15 second-party databases to PDB entries (and not vice versa), as described in the following. The data sources and the number of referenced entries are listed in Table I. Our reason for disregarding links from PDB entries to second-party databases is the very nature of the PDB, which essentially is an uncurated archive of structures. Responsibility of the content of an entry is by the authors, not by the PDB. Therefore, links are often not updated after the initial submission. Note that PDB is currently changing this policy, at least in part, and there are attempts to curate PDB entries in certain aspects.[9,10]

The 6 fold/family classification databases CATH,[3] SCOP,[11] DALI,[12] HSSP,[13] CE,[14] and HOMSTRAD[15] reference PDB chains in the respective files directly. The tabular data available for CE and DALI was about 9 months older than the data available through the projects' dynamic websites. At time of writing, 47 references from SCOP primary data were out-of-date because the corresponding PDB structures had become obsolete in the PDB. Similar numbers affected the other data sources.

Links pointing from sequence-related databases to the PDB were retrieved from 4 sources: The official SWISS-PROT release[16] (without TrEMBL), UniProt,[17] InterPro,[6] and the PDBSprotEC[18] link database. All but the last one suffer from the drawback that only entire PDB entries, but not chains, are being referenced. Tracking these cross-references down to individual protein chains is not trivial, since given protein sequences may differ between SWISS-PROT and PDB entries. We performed alignments of 6648 SWISS-PROT sequences with all chains of the referenced PDB entries using CLUSTAL W[19] with a sequence identity threshold of 0.95. This way, 20,252 links from SWISS-PROT to PDB entries could be verified. In 89 cases, the corresponding structures had become obsolete. For the 9 remaining references, large loops that had not been crystallized and a few obvious errors spoiled the alignment. Note

that an all-against-all sequence alignment would not be reasonable, because not all sequences with a high similarity share the same annotation (e.g., come from different organisms, tissues, etc.). While InterPro and UniProt provide additional information, the PDBSprotEC database contains carefully curated links to the SWISS-PROT database. They outnumber the links given in SWISS-PROT itself, showing that for a number of SWISS-PROT entries, "missing link" rather than "missing data" applies.

Of the 4 enzyme databases, ENZYME,[20] KEGG,[21] BRENDA,[22] and PDBSprotEC,[18] only PDBSprotEC links to individual PDB chains. From BRENDA, links to PDB entries were used. For ENZYME and KEGG, references were created by matching the 4-number enzyme classification [Enzyme Commission (EC numbers)] with EC numbers given in the PDB header. However, as several enzymes are known to have more than 1 EC number, it is not certain that every protein structure can be associated to all biologically relevant reactions by only matching EC numbers. Of the 1161 E.C. numbers used in PDB entries, 22 were not contained in the current ENZYME release. We found that these included 7 obsolete EC numbers and 15 typographical errors. For analysis, the cross-references for all data sources were compared to the full set of PDB entries and compared to each other.

RESULTS

Time Dependency

We observe that coverage of many data sources is time-dependent (see Fig. 1). The number of available secondary-database annotation clearly depends on the deposition date of a structure, and there is a time-dependent annotation gap for many of the second-party databases.

As can be seen from Figure 1(a), the coverage of all fold/family databases but HSSP starts declining from 1995 through 2000. HSSP—which is calculated automatically for each new PDB entry—remains constant throughout the years. SCOP stays at a high coverage for a long time, followed by CE, DALI, and CATH. The HOMSTRAD project has lost pace with the recent growth in data. Even though there were manual updates in the beginning of 2003, HOMSTRAD covers less than 40% of the PDB entries.

For sequence databases, the situation is similar [see Fig. 1(b)]. Structures deposited after the mid-1990s are affected with respect to SWISS-PROT and UniProt. For InterPro and PDBSprotEC, coverage drops only after 2001, where the drop is less severe for PDBSprotEC. It has to be pointed out that the curves for both pairs SWISS-PROT/UniProt and PDBSprotEC/InterPro are somewhat similar, suggesting that these databases share their data at least partially.

The curves for the ENZYME and KEGG databases [Figure 1(c)] basically show that the amount of enzymes in the PDB has remained constant for about 10 years. In the 1980s, enzyme classification was enforced less strictly by the PDB maintainers, and the amount of links was smaller then. BRENDA fails to annotate some enzyme structures published since 1998; PDBSprotEC has missed only a few since 2001. Comparing the latter 2 with the ENZYME links, which were created using the EC numbers from the

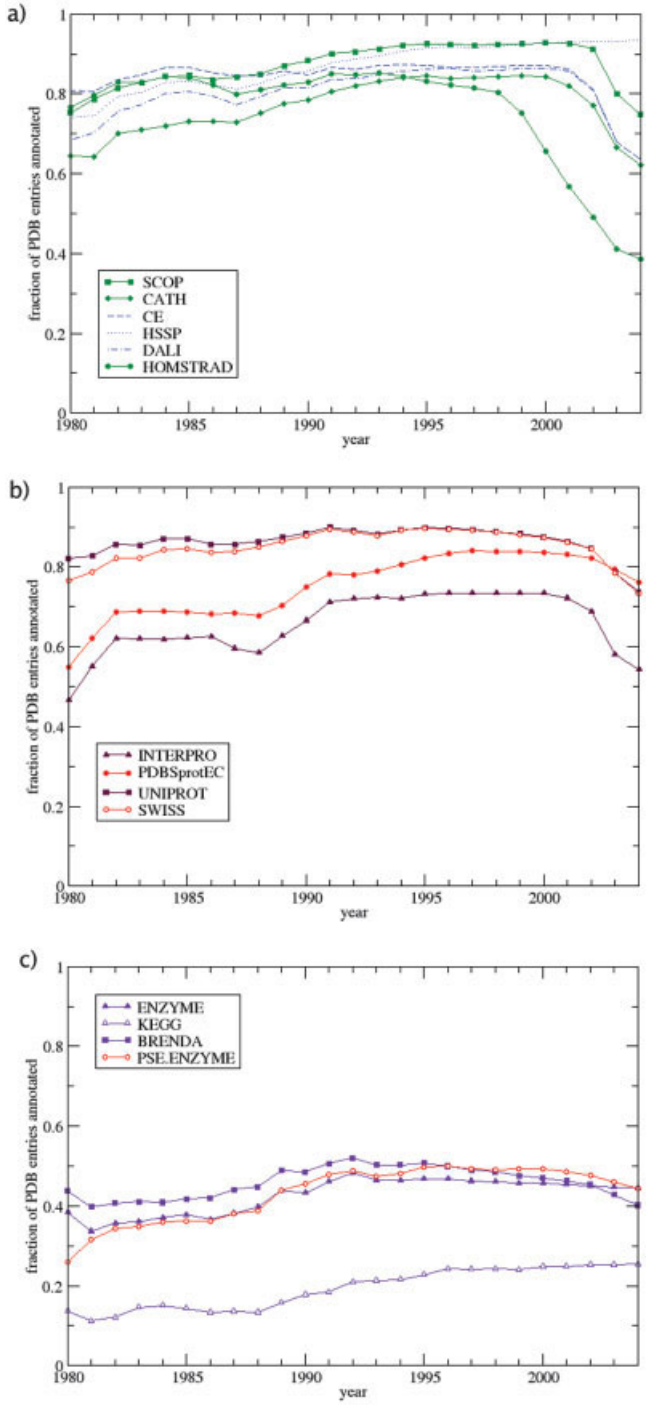

Fig. 1. Fraction of the PDB entries published up to a certain year for which links to a specific secondary-party database exists. The coverage over time is shown for (**a**) fold/family classification databases, (**b**) sequence databases, and (**c**) enzyme databases. During the 1970s, the PDB was very small and a large variations were observed. After 1997, many data sources for which the creation of cross-references requires manual interaction were not able to keep pace with the rapid growth of the PDB. The PSE.ENZYME curve corresponds to all enzyme references from PDBSprotEC.

PDB, it becomes clear that the slight decrease of the BRENDA and PDBSprotEC curves is really due to a lack of database links, and not because the relative amount of enzymes in the PDB has dropped.

Comparison of Databases

There is considerable variation between second-party databases in the fraction of the PDB they annotate. As mentioned above, one cannot expect most databases to reach 100% coverage. For instance, there will never be links from enzyme to nonenzyme structures in the PDB; CATH does not consider disordered structures or peptides, while SCOP does; and SWISS-PROT excludes immunoglobulins. The maximal coverage of data sources thus varies greatly depending on the nature of a database and the policy of its maintainers. However, computing a maximal coverage for each secondary database is very difficult. In the following, we shall try to contrast the absolute coverage with the informal thoughts on the expected maximal coverage wherever reasonable.

HSSP is getting nearest to full PDB coverage (93.5%), followed by SCOP (73.0%). CATH, DALI, and CE are about equal around 63%. These 5 databases cover mostly the same structures, and 55.3% of the PDB is covered by all of them.

Of all PDB entries, 73.7% are linked by the SWISS-PROT sequence database, which is almost the same number of links as UniProt. However, by using PDBSprotEC (76.2%), this amount still can be increased. InterPro contains much fewer references (54.3%), which could be a natural effect given that InterPro only links to sequences containing certain domains. Two-thirds of the structures (65.7%) were annotated by both PDBSprotEC and either SCOP or CATH, suggesting that these data sources overlap to a very high degree.

We also analyzed the coverage of 2 databases annotating SWISS-PROT entries, which in turn are linked to the PDB via PDBSprotEC: functional annotation from the Gene Ontology [Gene Ontology Annotation[23] (GOA)] and taxonomic classification from the National Center for Biotechnology Information (NCBI).[24] Both of them annotate (indirectly) almost as many structures as their corresponding sequence database entries, showing the importance of having high-quality SWISS-PROT references.

Almost half of the PDB structures (13,296) are designated as enzymes. Of these, 92.2% have an EC number indicating a specific enzymatic function described in the ENZYME database. BRENDA annotates 83.1%, and PDBSprotEC annotates 91.6% of the enzymes in the PDB, while only 52.4% of these enzymes could be linked to the detailed pathway descriptions from KEGG. The enzymes missed by ENZYME, BRENDA, and PDBSprotEC have probably been judged as enzymes by the first lines of the PDB header, not by EC numbers. These 3 sources annotate 70.3% of all enzymes in common. For the ENZYME and KEGG references, the EC numbers from the PDB were used. This is an exception to our general rule of using links from secondary databases into the PDB. The EC numbers are no references pointing in the other way, either. However, the additional data allow a better estimate of how well the other enzyme databases perform.

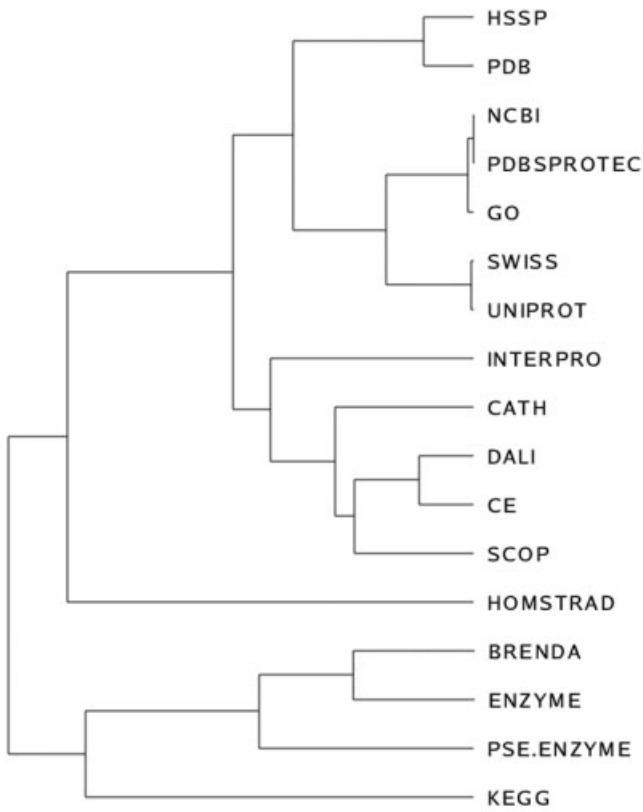

Fig. 2. Degree of overlap between the sets of cross-referenced PDB entries for the 15 secondary-party databases we studied. The data sources are arranged in a tree computed using the UPGMA method. Distances were calculated as the number of entries in the intersection divided by the union of 2 data sources. The best overlapping data sources were assigned a common node first. The branch lengths correspond to the calculated dissimilarity of 2 nodes. The PSE.ENZYME leaf corresponds to all enzyme references from PDBSprotEC.

There were 11,054 structures (40.2%) annotated by all of the following databases: SCOP, CATH, HSSP, PDBSprotEC, UniProt, InterPro, GOA, NCBI taxonomy, and, for enzymes, ENZYME and BRENDA. On the contrary, 2196 structures are annotated by only HSSP or ENZYME/KEGG, most of them being recent entries. We found 1462 structures having no secondary annotation at all, including 1357 nucleic acid structures, 41 small molecules, 52 proteins partially with unknown function, and 12 low-resolution structures.

To display the interdependencies between the data sources, we computed the overlap of any 2 sets of linked PDB entries as the number of entries in the intersection of both sets divided by the number of entries in their union. Based on these overlaps, we computed a tree using the Unweighted Pair Group Method With Arithmetic Mean (UPGMA).[25] The tree visualizes the degree of overlap in the content of the different databases (see Fig. 2). It becomes clear that the fold/family classifications and the sequence databases group together well. Only those data sources having comparatively high (HSSP) or low coverage (HOMSTRAD and INTERPRO) locate in different regions.

The enzyme-related data sources are located on a different branch, since they have much fewer links and do not overlap with the other sets that much. The 2 additional sources, GOA and NCBI, are almost identical to the PDBSprotEC links that were used to link them.

Finally, we analyzed whether representative sets of structures offered on the Web show a better coverage by secondary annotation than the average PDB entry. We compared the representative structures offered by the PDB[26] and from the PISCES server[27] to randomly created subsets of the PDB having the same size. Surprisingly, we find that random structures are annotated by an average of 9.5 data sources of the 15 we considered, whereas the representative structures are annotated by only 9.0 (PDB) and 9.3 (PISCES) data sources. This may be due to both methods preferring more recent structures in the representative sets, which tend to be less well annotated, as shown above.

CONCLUSIONS

Protein structures deposited in the PDB before 1997 are generally well annotated. The fraction of entries covered is constantly around 90% for SCOP and HSSP, 80% for several other fold/family classifications, around 80% for sources containing SWISS-PROT references, and 40% for enzyme databases. The score for SCOP is higher than the score of CATH, since SCOP contains classifications for peptides and nonstandard proteins not considered by CATH. These figures are probably the highest possible given that PDB also contains structures that are not addressed by these databases, such as nonproteins and synthetic peptides. DALI and CE are similar to SCOP.

Roughly two-thirds of the PDB entries are linked to both a sequence database and 1 or more fold/family classifications. This coincides with earlier results on the coverage of fold classification databases,[28] suggesting that the availability of annotations in this respect has not changed much in the last few years. References to ENZYME or KEGG are more frequent after 1990, probably because the PDB started to enforce submission of EC numbers from that time on. After 2000, references to the folding classifications except HSSP became less abundant (see Fig, 1). The same applies for sequence-related databases and BRENDA from 1997 on. Note that this affects a time range in which about 75% of the PDB entries fall. Classification by HSSP, ENZYME, and KEGG is available even for very recent structures, since it is calculated automatically. For 1462 structures, no annotation at all exists yet, but they are mostly nucleic acids.

We draw the following conclusions. First, we suggest that the availability of cross-references should be considered in the creation of representative sets, as a structure is more useful, the more secondary annotation is available.

Second, researchers should be aware that choosing a set of structures based on properties that are only available in second-party annotations may introduce a strong bias into the selection. This bias suppresses peptides, nonproteins, and unfolded and exotic structures, which probably is an effect that most researchers will welcome rather than be concerned about, but unpleasantly also suppresses recent structures, where "recent" stretches over a period of approximately 4 years. This is a particularly annoying effect, since the quality of revealed protein structures is generally improving over the years owing to improvements in the underlying technology.

Third, the structures referenced by different sources are often the same. For structures chosen based on sequence features, very likely also fold classification annotation is available. However, combining sequence features and pathway data immediately erases many structures, because the latter information is not available.

The lack of annotation, especially for recent structures, is certainly worrisome. Since we believe that the data production pace increases faster than the budgets of databases relying on manual curation, it is likely that the gap will become larger rather than smaller, although we have no data to prove this claim. This gap can only be closed by increasing the amount of resources invested into manual inspection and annotation of protein structures, unless methods for automated structure annotation are improving significantly.

ACKNOWLEDGMENTS

We thank Cornelius Frömmel and Robert Preissner for helpful advice during the project.

WEB LINKS

BRENDA enzyme information system: http://www.brenda.uni-koeln.de

CATH protein structure classification: http://www.biochem.ucl.ac.uk/bsm/cath/

Combinatorial Extension structure alignment (CE): http://cl.sdsc.edu/ce.html

DALI domain directory: http://www.ebi.ac.uk/dali

Enzyme nomenclature database (ENZYME): http://www.expasy.org/enzyme

Gene Ontology Annotation (GOA): http://www.ebi.ac.uk/goa/

Homologous Structure Alignment Database (HOMSTRAD): http://www-cryst.bioc.cam.ac.uk/~homstrad

Homology-Derived Secondary Structure of Proteins (HSSP): http://swift.cmbi.kun.nl/swift/hssp

InterPro database: http://www.ebi.ac.uk/interpro

Kyoto Encyclopedia of Genes and Genomes (KEGG): http://www.genome.ad.jp/kegg

NCBI Taxonomy: http://www.ncbi.nlm.nih.gov/Taxonomy

PDBSprotEC database: http://www.bioinf.org.uk/pdbsprotec

Protein Data Bank (PDB): http://www.pdb.org

Structural Classification of Proteins (SCOP): http://scop.berkeley.edu

SWISS-PROT protein knowledge base: http://www.expasy.org/sprot

Universal Protein Resource (UniProt): http://www.ebi.ac.uk/uniprot

REFERENCES

1. 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.
2. Carugo O, Pongor S. The evolution of structural databases. Trends Biotechnol 2002;20:498–501.
3. Orengo CA, Pearl FM, Thornton JM. The CATH domain structure database. Methods Biochem Anal 2003;44:249–271.
4. Rother K, Mueller H, Trissl S, Koch I, Steinke T, Preissner R, Froemmel C, Leser U. COLUMBA: multidimensional data integration of protein annotations: DILS conference on databases in life sciences. LNBI 2004;2994:156–171.
5. Reichert J, Suhnel J. The IMB Jena Image Library of Biological Macromolecules: 2002 update. Nucleic Acids Res 2002;30:253–254.
6. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJA, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 2003;31:315–318.
7. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res 2001;29:221–222.
8. Zdobnov EM, Lopez R, Apweiler R, Etzold T. The EBI SRS server—new features. Bioinformatics 2002;18:1149–1150.
9. Bhat TN, Bourne P, Feng Z, Gilliland G, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H, Westbrook J, Berman HM. The PDB data uniformity project. Nucleic Acids Res 2001;29:214–218.
10. Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W. E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. Nucleic Acids Res 2003;31:458–462.
11. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32(Database issue):D226–D229.
12. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res 2001;29:55–57.
13. Dodge C, Schneider R, Sander C. The HSSP database of protein structure–sequence alignments and family profiles. Nucleic Acids Res 1998;26:313–315.
14. Shindyalov IN, Bourne PE. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Res 2001;29:228–229.
15. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
16. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370.
17. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2004;32:115–119.
18. Martin AC. PDBSprotEC: a web-accessible database linking PDB chains to EC numbers via SwissProt. Bioinformatics 2004;20:986–988.
19. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
20. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–305.
21. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32(Database issue):D277–D280.
22. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 2004;32(Database issue):D431–D433.
23. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in UniProt with Gene Ontology. Nucleic Acids Res 2004;32(Database issue):D262–D266.
24. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2000;28:10–14.
25. Sokal RR, Rohlf FJ. Biometry. San Francisco: W. H. Freeman; 1981. 859 p.
26. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17:282–283.
27. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
28. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Struct Fold Des 1999;7:1099–1112.