Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2008 November 3; 2008: 33–40. doi:10.1109/

BIBM. 2008. 4686206

An Integrated Database for Complex Protein Structure Modeling

Qiang Wang and **Roland L. Dunbrack Jr.**

Fox Chase Cancer Center

Qiang Wang: qiang.wang@fccc.edu; Roland L. Dunbrack: roland.dunbrack@fccc.edu

Abstract

In homology modeling of protein structures, it is typical to find templates through a sequence search against a database of proteins with known structures. In more complicated modeling cases, such as modeling a protein structure in contact with a ligand, sequence information itself may not be enough and more biological information is required for a successful modeling process. SCOP and PFAM are two databases providing protein domain information which can be utilized in complex protein structure modeling. However, due to the manually-curated nature of both databases, they fail to provide timely coverage of protein sequences existing in the Protein Data Bank (PDB). In this paper, we introduce a new relational database, IDOPS, which integrates sequence and biological information extracted from remediated PDB files and protein domain information generated with HMM profiles of PFAM families. With a carefully designed protocol, this database is updated regularly and the coverage rate of PDB entries is guaranteed to be high.

Background

In recent years, molecular modeling has become a major field in biomedical research, contributing significantly to the understanding of protein function through structure. Among different ways of protein structure modeling, homology modeling [1] (also known as comparative modeling) is a class of methods known for its simplicity and effectiveness in constructing an atomic-resolution structure for a query sequence. The successful application of homology modeling is based on the observation that protein tertiary structure is much more conserved than amino acid sequence. Proteins that diverge dramatically and share only low similarity in sequence can still share common structural properties. On the other hand, the expensive cost of experimental methods such as X-ray crystallography and protein NMR [2] also contributes to the increasing popularity of homology modeling. As of July 22, 2008, UniProtKB/Swiss-Prot contains about 400K sequence entries while the Protein Data Bank (PDB) contains only a little more than 50K experimentally determined proteins structures.

While different homology modeling methods vary in details, they usually follow a basic protocol that consists of several steps: 1) for a target sequence of unknown structure, identify a template structure with sequence related to the target and align the target sequence to the template sequence and structure; 2) for core secondary structures and all well-conserved parts of the alignment, borrow the backbone coordinates of the template according to the sequence alignment of the target and template; 3) build side chains onto the backbone model according to the target sequence; 4) for segments of the target sequence for which coordinates cannot be borrowed from the template because of insertions and deletions in the alignment (usually in loop regions of the protein) or because of missing coordinates in the template, rebuild these regions using loop modeling methods or other ab initio structure prediction methods; 5) refine the structure, modeling likely differences in the relative positions of helices, sheet strands, and other elements of structure. While every step is important and may affect the final model, the step of identifying a template structure should

be treated with more caution, since it essentially decides the direction of the modeling process.

The identification step usually involves sequence alignment: a BLAST [3] search is performed against a database of sequences with known structure. To find more remotely related sequences, a PSI-BLAST (Position-Specific Iterative Blast) search can be performed instead. After a template whose sequence similarity with the target is above a certain threshold (e.g. 25%) is identified successfully, we can continue with the other steps. However, if no such template is found, there may be more difficulty. Moreover, there may be more complicated modeling tasks where simple sequence search may fail. For example, if more biological information is given for the target sequence, it is ideal to find a template with similar biological properties. Another scenario is that instead of a single-chain protein, we may want to build the model for a protein complex, either a protein in contact with a certain ligand, or a biological unit consisting of multiple peptide chains. In order to accomplish these modeling tasks, a more comprehensive biological database becomes necessary, enabling more complex queries.

Currently there are two major database containing protein structure and related biological information: the Protein Data Bank (PDB) [4] and the Macromolecular Structure Database (MSD). The PDB is an archival database for 3-D structural data of proteins and nucleic acids. Original found in 1971, the PDB was transferred in 1998 to the members of the Research Collaboratory for Structural Bioinformatics (RCSB) and now it is managed through the Worldwide Protein Data Bank (wwPDB) [5], an international research collaboration. Besides atomic level 3D coordinate data, the PDB archive also provide other biological details, such as polymer sequence, ligand chemistry and author's annotations. Deposited structures in the PDB are kept in separate files, available in several different formats. These files can be downloaded either through the ftp server or the website of wwPDB members. The RCSB website also provides an interface for advanced searches (<http://www.rcsb.org/pdb/search/advSearch.do>). As of August 12, 2008, 52,402 structures have been deposited into the PDB. The other database, MSD, maintained by the European Bioinformatics Institute (EBI), is the European project for the collection, management and distribution of macromolecular structure data. The EBI shares the same structural data as the RCSB PDB but offers a different interface to the data. A very useful feature of MSD is that it provides various ways of searching protein structural data. On its website (<http://www.ebi.ac.uk/msd/Services.html>), users can easily find interesting structures they have in mind based on different search criteria.

At the beginning of a homology modeling process (the identification step), any homology modeling program needs to search a database which contains all the sequences with known structure. This database can be generated through extracting sequence information from the PDB. One widely used sequence database, PDBAA [6], is provided by the Dunbrack Lab at Fox Chase Cancer Center. Although in most simple modeling tasks, a template can be identified merely based on the sequence similarity, things can become complicated when more limitations have been added to the potential templates. When given a target sequence with certain biological properties, e.g. binding an adenosine nucleotide molecule, it would be ideal to find a template that has both high sequence similarity and the required biological properties. Both the PDB and the MSD provide web services for user-customized database search. However, to our knowledge, neither of these two protein structure archives provides a comprehensive relational database, or no such a relational database is publicly available. In an interactive homology modeling environment, such as our MolIDE program[7,8], the integration of a relational database containing both sequence and biological information of all known protein structures is very helpful.

There are also situations where a single template is not enough and we have to make use of multiple templates [9,10]. One example is that there are hits in the database with high sequence similarity with the target sequence but lacking certain biological properties, such as particular ligands, while other structures with the required biological properties do not show obvious sequence similarity. Another example is that several protein structures may have different desired biological properties but none of them has all properties. In either case, the modeling program has to find a way to combine the structure information from different structure templates. When multiple templates are to be used, it is essential to make sure that they come from the same protein family and share common functions.

Since a single protein chain may contain one or more S domains, the basic unit of protein structure and function, the knowledge of domain assignment of each protein peptide would bring insights in finding appropriate templates. The existing databases providing protein domain definitions fall in two major categories: structure-based and sequence-based, with SCOP [11] and PFAM [12] as representatives respectively. Manually curated, the SCOP database provides a detailed and comprehensive description of the structural and evolutionary relationships of proteins. Based on the known structure information, proteins are classified with a hierarchy of a five levels: species, protein, family, superfamily and fold. PFAM, on the other hand, is a database built through multiple sequence alignments. As SCOP, PFAM is also manually curated and each domain family is defined with a hidden Markov model (HMM) profile that is generated through aligning a group of carefully selected seed sequences. A recent study [13] shows that although the SCOP and the PFAM databases define domain families based on different sources of information, there is a general agreement between them. During the past decades, both of these two databases have been utilized by many applications, but manual creation has prevented them from being up-to-date as many more entries are being deposited into the PDB. The current version of SCOP (release 1.73) was released in November 2007 and covers only 34,494 PDB entries, while the current PFAM (version 22.0) was released in July 2007 and covers only 30,082 PDB entries. Since there are more than 50,000 protein structures in the PDB, the coverage rates of PDB entries of SCOP and PFAM are only about 70% and 60%, respectively.

Method

In order to provide a reliable, comprehensive, and most importantly, up-to-date information for protein homology modeling, we develop an integrated relational database, Integrated Database of Protein Structure (IDOPS), which integrates the sequence, biological and evolutionary information from various resources. Specifically, current IDOPS extracts the sequence and biological information for all proteins with known structures from the PDB archive, and generates protein domain assignments for all PDB entries with the help of the protein family profiles tables have a relationship with `pdb_entry` either directly or indirectly through the field `pdb_id`, which is the four-letter id for PDB entries.

Extracting data from remediated PDB files

Except for table `pdb_pfam`, data in all tables is extracted from the remediated PDB files downloaded from the PDB. As the result of a remediation project, the PDB archive now has remediated data files in three different formats: mmCIF, PDBML-XML and PDB File format. The remediated data files were officially released on August 1st, 2007 and there were a lot improvements in terms of the data quality, such as more detailed chemical description of non-polymer and monomer chemical components, standardized atom nomenclature, updated sequence database references, etc. To extract protein structure information, we decide to use the PDBML-XML files which can be downloaded through the ftp services of the wwPDB. While using the same logical data organization as the traditional PDB file format, PDBML-XML format has a few advantages: structured data format,

complementary information, standardized data fields, and straightforward mappings between XML data categories and IDOPS tables.

To fulfill the task of data extraction, we developed a program named PdbmlParser. This program was written in C++ and two libraries were utilized. The Expat XML parser library [14] was used to parse the PDBML-XML data files, and the other library, mysql++ [15], provided the functions of database operations for MySQL. To further improve the XML-parsing efficiency, for each PDB entry, instead of the single complete XML file, we used two separate files, one for coordinate information (under the XML-extatom directory on the PDB ftp website) and one for biological information (under the XML-noatom directory on the PDB ftp website). With PdbmlParser, it takes only a few hours to finish reading data for all entries in the PDB archive. Due to the fact that not all data categories are available in all XML files, some PDB entries may be missing in some tables in the IDOPS database.

Generating domain assignments

The latest version of PFAM (Version 22.0) was released in July, 2007 and totally 9,318 domain families are defined. In a relational database published at the same time, PFAM provided pre-calculated domain assignments from PFAM. We also introduce an update protocol so that IDOPS can be updated regularly, as more protein structures are being deposited into the PDB. IDOPS is implemented with MySQL and currently consists of 23 tables, as shown in figure 1. Table `pdb_entry` contains the basic description about PDB entries and all the other for PDB entries. However, while the number of PDB entries keeps increasing and now is over 52,000 (as of August 12, 2008), only a little more than 30,000 PDB entries have their domains defined in the PFAM database. The coverage rate is only about 60 percent.

While PFAM fails to update at the same pace as the PDB, it provides the HMM profiles for all the 9,318 domain families. This fact makes it possible to assign domain definitions for any protein sequences. In the software package HMMER [16] developed by Washington University in St. Louis, there is a program named `hmmpfam` which has the ability to search a single sequence against an HMM database.

By running `hmmpfam` for each polypeptide sequence in the PDB against the PFAM HMM database, we obtain the potential domain definitions for all potential sequences. However, before these domains can be used with confidence, two issues need to be addressed: overlapped domain assignments and quality of domain definitions. As an example, Figure 2(a) demonstrates the potential domain definitions for chain E of PDB entry 2OXX. In order to remove the overlap and domain assignments with low quality, we used a straightforward greedy algorithm which gives preference to domain definitions with better (smaller) E-values and allows a user defined threshold for tolerance of overlap (e.g. 10%). The better defined domain definitions for chain E of 2OXX is shown in figure 2(b).

Although every domain assignment found with the `hmmpfam` program is accompanied with an E-value, there is no accurate conclusion on a threshold which can be set as the quality criterion. By default, `hmmpfam` uses 10 as the threshold. While this default value guarantees finding most appropriate domain assignments, it also introduces many false positive, as the ones we have seen in the example shown in figure 2(a).

To evaluate the effects of the threshold value, we test whether or not two proteins structures assigned with the same PFAM domain(s) actually come from the same family. This is done through structural alignment with the FATCAT [17] program. When this test was performed, we held a mild assumption that all domain assignments with an E-value less than 0.001 are reliable. With this assumption, we were able to find at least one protein structure

for 4,256 PFAM families. Then within each of these 4,256 families, we compared all structures with the one with the lowest E-value. If two proteins have similar structure, which is judged based on the p-value returned by FATCAT, they are considered coming from the same family and the domain assignment is reliable. Or else, the domain assignment is considered a false positive. In our experiments, a threshold of 0.05 was used for p-value in FATCAT. Figure 3 shows the precision of domain assignment and coverage of PDB entries when different E-value thresholds are used. It turned out that domains with an E-value lower than 0.5 can be generally considered correct, with a precision over 80 percent, while the coverage of PDB entries stays high at 96.2%.

Update protocol

In order to keep up with the pace of the PDB archive, an update protocol was designed to update IDOPS regularly and automatically. This protocol is implemented with a series of Shell and Perl scripts and its major steps are shown in figure 4. At the beginning, a local version of PDB archive is updated. This is done through anonymous rsync access to the remediated site of wwPDB. Then the list of PDB entries is updated and a list of new PDB entries is generated. For each new PDB entry, its sequence and biological information is extracted from the XML files and domain assignments are generated. All the information is finally inserted into the database. Besides the insertion of new entries, the protocol also checks the lists of modified and obsolete entries and makes corresponding changes. Currently, IDOPS database is updated weekly, a few hours after the update of the PDB archive.

Results

Statistics of IDOPS

As of August 12, 2008, there are 52,402 structures in the PDB archive. Table 1 shows the numbers of entries and total records stored in all tables of IDOPS database. It also shows that there are missing entries in several tables because of the absence of corresponding information in the PDBML files. Even though only 49,621 entries have their domains defined in table `pdb_pfam`, considering the fact that only 50,473 PDB entries have one or more polypeptide chains in their structures, the coverage rate is 98.3%.

Some issues with PDB XML files

With IDOPS, it is possible for us to find the answers to various interesting questions quickly. As an example, here we discuss several issues with the quality of PDB XML files identified through a few queries that we ran against IDOPS. These issues were first identified by one of the authors three years ago with complicated Perl scripts. The new query results reconfirmed them.

- There are quite a lot intra-model (within the same asymmetric unit) connection records which do not involve water molecules. The total 992,759 records come from 28,889 pdb entries.

```
mysql> select count(*) from pdb_link
where ptrn1_label_comp_id!='HOH' and
ptrn2_label_comp_id!='HOH' and
ptrn1_symmetry=ptrn2_symmetry;
+-----+
|count(*)|
+-----+
```

```
|992759|
+-----+
mysql> select count(distinct pdb_id) from
pdb_link where ptrn1_label_comp_id!='HOH'
and ptrn2_label_comp_id!='HOH' and
ptrn1_symmetry=ptrn2_symmetry;
+-----+
|count(distinct pdb_id)|
+-----+
|28889|
+-----+
```

- Only 11 of the 2887 polymer sugars in PDB are labeled “polysaccharide” in entity_polyCategory.

```
mysql> select count(*) from pdb_compnd
where description like '%sugar%' and type
='polymer';
+-----+
|count(*)|
+-----+
|2887|
+-----+
mysql> select count(*) from pdb_seq where
type like '%polysaccharide%';
+-----+
|count(*)|
+-----+
|11|
+-----+
```

- Of the 1,079 entries with NAG-NAG covalent bonds 1,039 have sugars labeled “polymer” and 636 have sugars labeled “non-polymer”. Only 5 sugar entities with ‘NAG’ are given their polymer sequence.

```
mysql> select count(distinct pdb_id) from
pdb_link where ptrn1_label_comp_id =
'NAG' and ptrn2_label_comp_id='nag';
+-----+
|count(distinct pdb_id)|
+-----+
|1079|
+-----+
mysql> select count(distinct pdb_id) from
pdb_compnd where description like
'%sugar%' and type='polymer' and pdb_id
in (select distinct pdb_id from pdb_link
where ptrn1_label_comp_id = 'NAG' and
ptrn2_label_comp_id='nag');
+-----+
|count(distinct pdb_id)|
```



```

+-----+
|1039|
+-----+
mysql> select count(distinct pdb_id) from
pdb_compnd where description like
'%sugar%' and type='non-polymer' and
pdb_id in
(select distinct pdb_id from pdb_link
where ptrn1_label_comp_id = 'NAG' and
ptrn2_label_comp_id='nag');
+-----+
|count(distinct pdb_id)|
+-----+
|636|
+-----+
mysql> select count(*) from pdb_seq where
seq_code like '%(NAG)%';
+-----+
|count(*)|
+-----+
|5|
+-----+

```

A complex modeling example

In this subsection, we use a real homology modeling example to demonstrate the usefulness of IDOPS in the modeling process of a complex structure. The target sequence in this example is the C-terminal domain of human cystathionine β -synthase (CBS). CBS catalyzes the covalent linkage of serine and the methionine metabolite homocysteine to produce cystathionine. Cystathionine is subsequently converted into cysteine. High levels of homocysteine are strongly linked to heart disease [18], and patients with homocystinuria have been found to have mutations in the gene that codes for CBS [19].

The enzyme domain of human CBS is preceded by a 75 amino acid proline-rich region that binds heme, and is followed by a 155 amino acid region C-terminal domain. This domain consists of two copies of a ~60 amino acid motif which itself is known as a CBS domain. This tandem pair of CBS domains is now often referred to as a Bateman domain. In many proteins, the function of the Bateman domain appears to be to bind adenosine nucleotides, AMP and ATP, thereby modulating the activity of the associated protein. Human CBS is activated by the binding of S-adenosyl methionine (SAM). Elimination of the Bateman domain region produces an enzyme that is constitutively active and unresponsive to SAM [20].

While there are a few structures in the PDB whose sequences are remote homologs of the Human CBS Bateman domain (CBS-BD), none of them succeeds to provide every detail that is necessary for a reliable model. However, with multiple templates, we are able to include borrowing of loop conformations from different templates and binding of SAM to CBS using information from several structures that contain AMP, ADP, ATP or SAM. Before we can proceed with the modeling, we need to find all the known structures of CBS domain which are in contact with an adenosine nucleotide molecule. With the help of IDOPS, we can easily find a list of potential templates with an SQL query:

```
mysql> select distinct pdb_id from pdb_pfam
where model like '%CBS' and e_value<0.5 and
pdb_id in
(select distinct pdb_id from pdb_ligand
where comp_id='ATP' or comp_id='ADP' or
comp_id='AMP' or comp_id='SAM');
```

In current IDOPS, this query returns a list of 16 PDB entries: 2J9L, 2JA3, 2OOX, 2OOY, 2QR1, 2QRC, 2QRD, 2RIF, 2UV4, 2UV6, 2UV7, 2V8Q, 2V92, 2V9J, 2YZQ, 3DDJ. This list has been verified by a manual examination. Without the help of IDOPS, we would get a much larger list (as shown in the query below) of CBS domains through a regular sequence search against current databases, and it would take a even longer time to manually check if any of the hits are in contact with the above mentioned ligands.

```
mysql> select count(distinct pdb_id) from
pdb_pfam where model like '%CBS' and
e_value<0.5;
+-----+
|count(distinct pdb_id)|
+-----+
|58|
+-----+
```

Conclusion

In this paper, we introduce a relational database, IDOPS, which integrates sequence, biological, and domain information for all known protein structures. The sequence and biological information is extracted from the XML files downloaded from the PDB archive. With the help of HMMER tools and HMM profiles provided by PFAM, we are able to generate a classification of domains in the PDB. With the help a well designed protocol, implemented with Shell and Perl scripts, IDOPS is always kept up-to-date and it guarantees a high coverage rate of PDB entries.

IDOPS can be utilized in many applications of structural bioinformatics. Specifically, the easy access to comprehensive and up-to-date structural information will be beneficial in complex protein structural modeling in a future implementation of our modeling program MolIDE. IDOPS is still under development and will be formally published with the newer version of MolIDE. Interested readers can contact the authors to obtain an _ version of the database.

References

1. Bourne, PE.; Weissig, H. Structural Bioinformatics. John Wiley & Sons Canada, Ltd; 2003.
2. Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? *Proteins: Structure, Function, and Bioinformatics* 2005;60:139–147.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215(3):403–410. [PubMed: 2231712]
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research* 2000;28:235–242. [PubMed: 10592235]

5. Berman, H.; Henrick, K.; Nakamura, H.; Markley, JL. Nucleic Acids Research. Vol. 35. 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data; p. D301-D303.
6. Wang, G.; Dunbrack, RL, Jr. PDBAA: Sequence Files Representing the Whole PDB. http://dunbrack.fccc.edu/Guoli/pisces_download.php
7. Canutescu AA, Dunbrack RL Jr. MolIDE: a homology modeling framework you can click with. *Bioinformatics* 2005;21:2914–2916. [PubMed: 15845657]
8. Wang Q, Canutescu A, Dunbrack RL. SCWRL and MolIDE: Computer programs for sidechain conformation prediction and homology modeling. *Nature Protocols* 2008;3 (11):1–16.
9. Larsson P, Wallner B, Lindahl E, Elofsson A. Using multiple templates to improve quality of homology models in automated homology modeling. *Protein Science* 2008;17:990–1002. [PubMed: 18441233]
10. Dalton JAR, Jackson RM. An evaluation of automated homology modelling methods at low target-template sequence similarity. *Bioinformatics* 2007;23 (15):1901–1908. [PubMed: 17510171]
11. Andreeval A, Howorth D, Chandonia J, Brenner S, Hubbard TJP, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research* 2007;1–7.
12. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz H, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A. The Pfam protein families database. *Nucleic Acids Research* 2008;36(Database issue):D281–D288. [PubMed: 18039703]
13. Zhang Y, Chandonia J, Ding C, Holbrook SR. Comparative mapping of sequence-based and structure-based protein domains. *BMC Bioinformatics* 2005;6:77. [PubMed: 15790427]
14. The Expat XML Parser. <http://expat.sourceforge.net/>
15. MySQL++: a C++ wrapper for MySQL's C API. <http://www.tangentsoft.net/mysql++/>
16. HHMI Janelia Farm Research Campus. HMMER: biosequence analysis using profile hidden Markov models. <http://hmmer.janelia.org/>
17. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 2003;19(Suppl 2):ii246–ii255. [PubMed: 14534198]
18. Refsum H, Ueland PM, Nygard O, Vollset SE. Homocysteine and cardiovascular disease. *Annu Rev Med* 1998;49:31–62. [PubMed: 9509248]
19. Kraus JP, Janosik M, Kozich V, Mandell R, Shih V, Sperandio MP, Sebastio G, de Franchis R, Andria G, Kluijtmans LA, Blom H, Boers GH, Gordon RB, Kamoun P, Tsai MY, Kruger WD, Koch HG, Ohura T, Gaustadnes M. Cystathionine beta-synthase mutations in homocystinuria. *Hum Mutat* 1999;13:362–375. [PubMed: 10338090]
20. Shan X, Kruger WD. Correction of disease-causing CBS mutations in yeast. *Nat Genet* 1998;19:91–93. [PubMed: 9590298]

Figure 1.
The design of IDOPS database (generated with the Toad Data Modeler).

Figure 2. Domain assignments for a chain E of PDB entry 200X. (a) overlapping assignments. (b) non-overlapping assignments.

Figure 3.
Evaluation of different E-values in domain assignments.

Figure 4.
Flowchart of the update protocol for IDOPS.

Table 1

Numbers of PDB entries and total records in IDOPS tables.

Table	Number of PDB entries	Number of total records
pdb_entry	52402	52402
pdb_author	52402	265111
pdb_asu	52402	429869
pdb_cell	48594	48594
pdb_bu	38713	42825
pdb_bu_details	52402	367505
pdb_citation	52361	85449
pdb_compnd	52402	191336
pdb_component	52402	1114036
pdb_seq	52380	73515
pdb_seqadv	27311	53356
pdb_dbref	49416	65729
pdb_dbref_item	49414	114807
pdb_het	36569	246954
pdb_link	29242	1045036
pdb_sheet	44539	220448
pdb_sheet_order	44406	755500
pdb_sheet_range	44539	980070
pdb_source	42353	47488
pdb_struct_conf	48661	1086450
Pdb_ligand	45141	114850
pdb_coordinate	52402	237668418
pdb_pfam	49621	291075