

The PDBFinder II Database

Elmar Krieger, Rob W.W. Hooft, Sander B. Nabuurs and Gert Vriend

Center for Molecular and Biomolecular Informatics (CMBI)

Toernooiveld 1, NL-6525ED Nijmegen, the Netherlands

Elmar.Krieger@cmbi.kun.nl

The PDBFinder II database provides uniform access to data typically required by applications in the fields of protein structure analysis and prediction. These include the sequences of experimentally determined structures, chain breaks, assigned secondary structure (DSSP), residue variability and entropy, hot-spots for insertions and deletions (HSSP), accessibilities, crystal contacts, B-factors, quality indicators, as well as general information parsed from PDB files: experimental methods, resolution, R-factor, authors, compounds, chains, hetgroups and many more. The quality indicators consist of normality Z-scores describing bonds, angles, torsions, planarity, chirality, packing, and inside/outside distribution. Unusual backbone conformations, unsatisfied hydrogen bond donors/acceptors, and flips of Asn, Gln and His side-chains are also reported.

The database is updated weekly and available as a flat, human readable text file from [ftp.cmbi.kun.nl/pub/molbio/data/pdbfinder2](ftp://cmbi.kun.nl/pub/molbio/data/pdbfinder2) or via SRS from servers listed at www.cmbi.kun.nl/gv/pdbfinder.

A Python plugin for the molecular modeling program YASARA allows to load PDB structures and visualize the data by coloring residues accordingly. A separate Python module to directly access the PDBFinder II is also available from www.yasara.org/plugins.

INTRODUCTION

While the Protein Data Bank provides the central resource for protein structures, a lot of additional information can be found in accessory databases spread all over the world. At the CMBI, we maintain the DSSP(1) (assigned secondary structure, residue accessibilities, hydrogen bonds), HSSP(2) (alignments of the PDB sequence against

Swissprot and TrEMBL), PDBFinder(3) (all important information from PDB files, including the sequences) and PDBReport(4) databases (a detailed structural analysis with a focus on potential problems). Researchers in structural bioinformatics often need quick, automated access to all these data. So far, this either required separate downloads and parsers for each file format, or was simply impossible, e.g. the PDBReports are only available in human-, but not computer-readable form (www.cmbi.kun.nl/gv/pdbreport/). The PDBFinder II database offers quick access to all these data in a format that is easily readable for humans as well as computer programs. Applications dealing with structure prediction can quickly determine which residues have experimental coordinates, they can correct initial sequence-based alignments by considering secondary structure elements, conserved and buried residues, as well as regions with a high probability of insertions and deletions. The per-residue quality indicators allow to identify the less reliable parts of a structure and build hybrid models consisting of experimentally well determined fragments in multiple templates. Researchers developing new modeling methods can use the PDBFinder II to generate reliable test-sets, that exclude residues involved in crystal contacts or problematic structures altogether. Finally, everyone analyzing a structure can easily map the information stored in the PDBFinder II onto the protein with the help of a plugin for YASARA, that colors residues according to their properties.

RESULTS

The PDBFinder II file format

The PDBFinder II entry for crambin (PDB ID 1CRN) is shown in Figure 1. In addition to information parsed from the PDB header, which is taken from the PDBFinder database(3), the following fields have been added at the chain-level:

- **Sequence:** the sequence of all residues with experimentally determined coordinates. Contrary to the original PDBFinder, the sequence contains chain break markers ‘-’. This allows to easily align it with the complete sequence deposited in Swissprot or given in the SEQRES field of the PDB file.
- **DSSP:** the secondary structure assigned by DSSP(1).
- **Cryst-cont:** residues involved in crystal contacts are flagged with a ‘+’ sign.

In the remaining fields, numbers or Z-scores are mapped to the range [0..9]. The formulas used are described in the header of the PDBFinder II file.

- **Nalign**: the number of aligned sequences at this position in the HSSP-file.
- **Nindel**: the sum of insertions and deletions.
- **Entropy**: and **Cons-weight**: the sequence entropy and conservation weights as defined in (2).
- **Access**: residue accessibilities, '0' is completely buried and '9' is maximally exposed.
- **Quality**: overall estimator for the chain quality, obtained by averaging the 'Phi/Psi', 'Backbone' and 'Packing 1' fields described below (which are among the most reliable quality indicators). High resolution X-ray structures reach values around 0.75, while NMR structures determined from only a few experimental restraints can be found around 0.3.

For the following quality-related fields, the mapping is chosen such that '9' corresponds to 'perfect' and '0' to 'requires attention'. Exceptions are the fields 'Torsions', 'Phi/Psi', 'Chi-1/2', 'Packing 1' and 'Packing 2', where '5' indicates about the average of high resolution X-ray structures, '9' corresponds to 'suspiciously good' and '0' to 'treat with caution', according to the WHAT_CHECK output(4).

- **Present**: 9 minus the number of missing atoms per residue.
- **B-Factors**: average crystallographic B-factor per residue.
- **Bonds** and **Angles**: absolute Z-score of the largest bond or angle deviation per residue according to the Engh&Huber parameters(5).
- **Torsions**: Average Z-score of the torsion angles per residue. This one and the following Z-scores including 'in/out' are calculated from the distributions found in the internal WHAT IF database of high resolution X-ray structures(6).
- **Phi/Psi**: Ramachandran Z-score per residue. No value can be determined for the N- and C-terminal residues, which is indicated by a question mark '?'.
and
- **Planarity**: Z-score of the side-chain planarity. Residues without a planar side-chain always score '9'.
- **Chirality**: Average absolute Z-score of all 'improper dihedrals' per residue, defined by one central and three bound heavy atoms, excluding planar groups. Glycine always scores '9'.

- **Backbone:** Number of similar backbone conformations found in the database, determined by superimposing stretches of five residues. No score can be obtained for the N- and C-terminal two residues. If less than 10 hits are found, there are not sufficient data to perform the following two checks for peptide plane flips and rotamers (indicated by question marks).
- **Peptide-PI:** RMS distance of the backbone oxygen from the oxygen in similar backbone conformations found in the database. Low scores indicate that the peptide-plane may have been flipped.
- **Rotamer:** Probability that the side-chain rotamer (chi-1 only) is correct. Glycine, alanine and proline always score '9'.
- **Chi-1/Chi-2:** Z-score for the side-chain chi-1/chi-2 combination.
- **Packing 1:** Three-dimensional packing quality Z-score(7).
- **Packing 2:** Second packing quality Z-score.
- **In/Out:** Absolute Z-score for the residue accessibility.
- **Bumps:** Sum of bumps (i.e. difference between VdW and actual distance) per residue.
- **H-Bonds:** 9 minus the number of unsatisfied hydrogen bonds. Additional penalties: 1 is subtracted for buried backbone nitrogens, 4 for unsatisfied side-chain hydrogen bonds.
- **Flips:** Asparagine, glutamine and histidine side-chains that need to be flipped in the optimum hydrogen bonding network(8) are scored with '0', all other residues with '9'.

Unless indicated otherwise, numbers at the right border (separated with a pipe symbol '|') are the average over the chain, multiplied with 0.9. This average is calculated before the individual residue values are saturated to fit into the interval [0..9] and can therefore lie outside the corresponding interval [0..1].

Database interfaces

A Python plugin for the molecular modeling program YASARA (www.yasara.org/plugins) allows to load PDB structures, automatically retrieve the

corresponding PDBFinder II entry via HTTP and map the information onto the structure. An example is shown in figure 1, where trypsin has been colored according to the HSSP conservation weights. A separate Python module to access the PDBFinder II directly is available from www.yasara.org/biotools. Both are licensed under the GNU GPL.

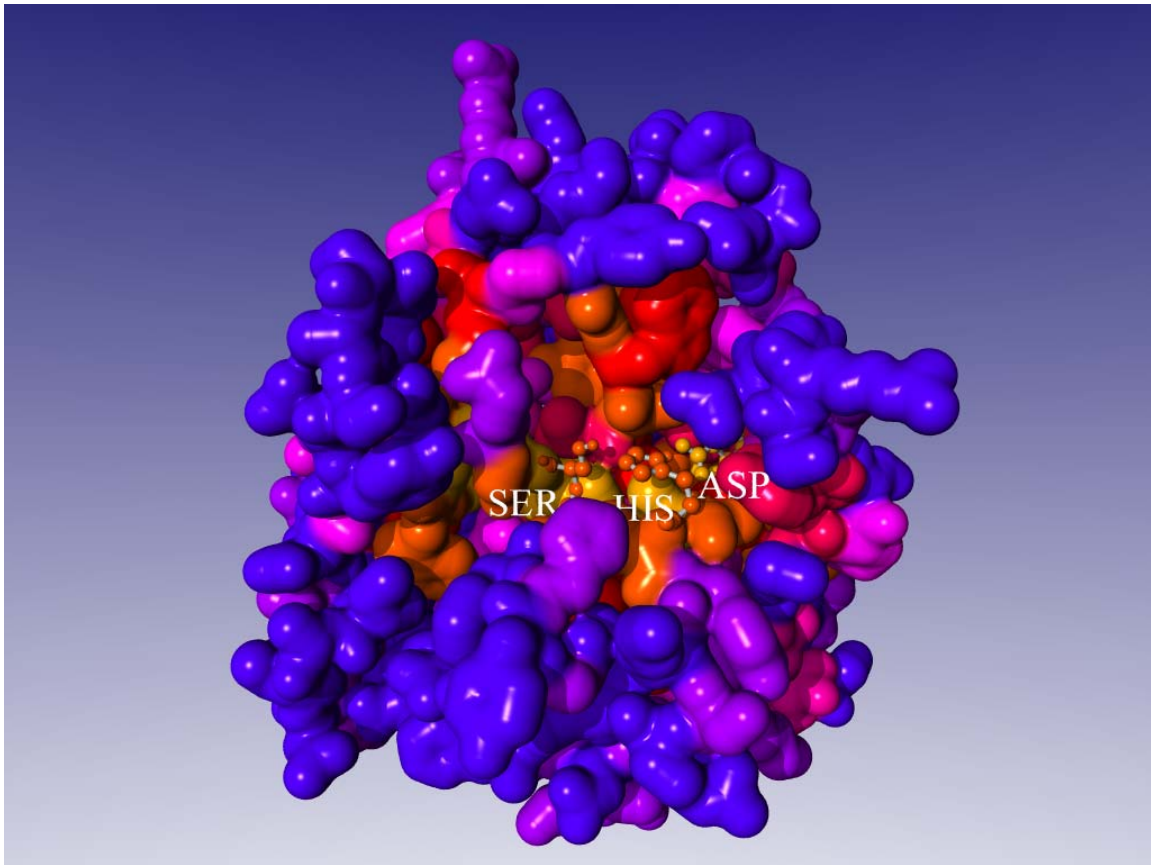


Figure 2: Crystal structure of trypsin from Atlantic salmon(9), colored by HSSP conservation weights. Blue corresponds to ‘not conserved’ and yellow to ‘completely conserved’. The Ser-His-Asp catalytic triad is indicated, the image has been created with YASARA and PovRay.

Reference List

1. Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577-2637.
2. Dodge, C., Schneider, R. and Sander, C. (1998) *Nucleic Acids Res.*, **26**, 313-315.
3. Hooft, R.W.W., Sander, C., Scharf, M. and Vriend, G. (1996)
Comput.Appl.Biosci., **12**, 525-529.
4. Hooft, R.W.W., Vriend, G., Sander, C. and Abola, E.E. (1996) *Nature*, **381**, 272-272.
5. Engh, R.A. and Huber, R. (1991) *Acta Cryst.A*, **47**, 392-400.
6. Hooft, R.W.W., Sander, C. and Vriend, G. (1996) *J.Appl.Cryst.*, **29**, 714-716.
7. Vriend, G. and Sander, C. (1993) *J.Appl.Cryst.*, **26**, 47-60.
8. Hooft, R.W.W., Sander, C. and Vriend, G. (1996) *Proteins*, **26**, 363-376.
9. Schroder, H.K., Willassen, N.P. and Smalas, A.O. (1998) *Acta Crystallogr.D Biol.Crystallogr.*, **54**, 780-798.