Bioinformatics Resources
Accessing protein structures

Andrea Schafferhans
(Protein) 3D structure databases

- Determining structures
- PDB (Protein Data Bank)
- Visualizing structures
- Building Aquaria

Further reading:
X-ray Crystallography

- Crystallise
- Diffract
  - Measurement shows electron density
  - Not visible:
    - H
    - Flexible parts
  - Ambiguity (e.g. \( \text{NH}_2 / \text{OH} \))
- Build model

Images:
- By Thomas Splettstoesser (www.scistyle.com) - own work; images were rendered with PyMol (www.pymol.org) based on PDB id 1MBO, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1248574
- By Hydrargyrum - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17543875
- By Jeff Dahl - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=3020011
X-ray Crystallography – Values to remember

- **Resolution**: quality measure of collected data
- **R-value**: fit between measured and calculated diffraction pattern
- **R-free**: prediction power for diffractions not used in refinement

Image:
NMR (Nuclear magnetic resonance)

- Spinning nucleus absorbs radiation
- Exact frequency depends on environment → chemical shift
- J-Coupling → split
- Proteins too crowded for normal NMR!

Image:
By T.vanschaik - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=8452750
Protein NMR

- Sample preparation (<1 ml, <3 mmol/l) 
  
(\(^{15}\)N-, \(^{13}\)C-labeled samples help assign residues)
- Data collection (2D spectra)
- Resonance assignment
- Restraint generation
  - Distance (peak intensity)
  - Angle (coupling magnitude)
- Model building
  \(\rightarrow\) ensemble of solutions
- No standard quality measures

Images:
Electron Microscopy

- Sample preparation (mostly cyro)
- Measurement:
  - Electron diffraction
  - Electron tomography
- Model building:
  - Reconstructing electron density
  - Fitting atoms (e.g. from X-ray)
- No standard quality measures
Organic compounds

- 1965 CSD (Cambridge Structure Database) started by Dr Olga Kennard by collecting published crystal structure data for small molecules
- 1987 CCDC as an independent non-profit company
- 2003 COD (Crystallography Open Database) founded

Images:
- https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/
Structural biology – History

• Started around 1950 (William Laurence Bragg, Max Perutz)
• Alpha helix, beta sheet: Linus Pauling (1951)
• DNA structure: Crick, Watson with data from Wilkins and Franklin (1953)
• First protein structures: Myoglobin, Hemoglobin (1957, 1959 Kendrew, Perutz)

Images:
• Various Wikipedia contributors
• By © MRC Laboratory of Molecular Biology, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=3923453
Protein Data Bank – History

• 1968: Brookhaven RAs ter Display (BRAD)
• 1969: Edgar Meyer (Brookhaven National Laboratory): file format for atomic coordinates
• 1971: Symposium Structure and the Three-Dimensional Level
• 1971: remote access with SEARCH program (Meyer)
• 1998: transfer to RCSB (Research Collaboratory for Structural Biology)
• 2003: formation of wwPDB (PDBe, RCSB, PDBj, BMRB)
Figure 1. Timeline of Key PDB Events and Structural Biology Highlights, 1971–2011

(left) Key events in the evolution of the PDB.
(right) selected key structures in the field of structural biology (Ban et al., 2000; Carter et al., 2000; Schluenzen et al., 2000; Henderson et al., 1990; Driscoll et al., 1989; Drew et al., 1981; Wang et al., 1979; Kim et al., 1973; Robertus et al., 1974).

Structure 20, March 7, 2012 © 2012 Elsevier Ltd All rights reserved

PDB references


• http://www.rcsb.org/
## Current composition (2016/06)

<table>
<thead>
<tr>
<th>Experimental Method</th>
<th>Proteins</th>
<th>Nucleic Acids</th>
<th>Protein / Nucleic Acid complexes</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray diffraction</td>
<td>99437</td>
<td>1729</td>
<td>5041</td>
<td>4</td>
<td>106211</td>
</tr>
<tr>
<td>NMR</td>
<td>10004</td>
<td>1141</td>
<td>232</td>
<td>8</td>
<td>11385</td>
</tr>
<tr>
<td>Electron microscopy</td>
<td>745</td>
<td>30</td>
<td>266</td>
<td>0</td>
<td>1041</td>
</tr>
<tr>
<td>Hybrid</td>
<td>89</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td>Other</td>
<td>173</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>196</td>
</tr>
<tr>
<td>Total</td>
<td>110448</td>
<td>2907</td>
<td>5547</td>
<td>26</td>
<td>118928</td>
</tr>
</tbody>
</table>
## Current composition (2017/05)

<table>
<thead>
<tr>
<th>Experimental Method</th>
<th>Proteins</th>
<th>Nucleic Acids</th>
<th>Protein / Nucleic Acid complexes</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray diffraction</td>
<td>109064</td>
<td>1836</td>
<td>5549</td>
<td>4</td>
<td>116453</td>
</tr>
<tr>
<td>NMR</td>
<td>10366</td>
<td>1202</td>
<td>241</td>
<td>8</td>
<td>118117</td>
</tr>
<tr>
<td>Electron microscopy</td>
<td>1098</td>
<td>30</td>
<td>386</td>
<td>0</td>
<td>1514</td>
</tr>
<tr>
<td>Hybrid</td>
<td>100</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>106</td>
</tr>
<tr>
<td>Other</td>
<td>189</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>212</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>120817</td>
<td>3075</td>
<td>6184</td>
<td>26</td>
<td>130102</td>
</tr>
</tbody>
</table>

Unique chains: 95987
PDB statistics

• See
  http://www.rcsb.org/pdb/home/home.do#Subcategory-search_drilldown
Searching PDB online

- RSCB – example entry
- PDBe – example entry
- PDBj – example entry
File formats

- Traditional: PDB
- Latest (last) PDB file format version: Contents Guide Version 3.30 (Nov. 21, 2012) 190 pages!
- Problem: large structures
- New standard (2104): PDBx/mmCIF see mmcif.wwpdb.org
PDB file overview

• Header
  – Protein information
  – Citation
  – Details of structure resolution

• Coordinates / Connectivity
PDB file format (general)

- **allowed characters:**
  
  `abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ1234567890` `\` `-` `[` `]` `;` `,' `. `~` `@` `#` `$` `%` `^` `&` `*` `(` `)` `_` `+` `{` `}` `|` `:` `"` `<>` 

- each line is 80 characters wide including EOL

- lines self-identifying:
  - col 1-6: record name
  - col 7: blank

- Records in defined order

- Some records are mandatory
### Excerpt from PDB record list

<table>
<thead>
<tr>
<th>RECORD TYPE</th>
<th>EXISTENCE</th>
<th>CONDITIONS IF OPTIONAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEADER</td>
<td>Mandatory</td>
<td></td>
</tr>
<tr>
<td>OBSLTE</td>
<td>Optional</td>
<td>Mandatory in entries that have been replaced by a newer entry.</td>
</tr>
<tr>
<td>TITLE</td>
<td>Mandatory</td>
<td></td>
</tr>
<tr>
<td>SPLIT</td>
<td>Optional</td>
<td>Mandatory when large macromolecular complexes are split into multiple PDB entries.</td>
</tr>
<tr>
<td>CAVEAT</td>
<td>Optional</td>
<td>Mandatory when there are outstanding errors such as chirality.</td>
</tr>
<tr>
<td>COMPND</td>
<td>Mandatory</td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>Mandatory</td>
<td></td>
</tr>
<tr>
<td>KEYWDS</td>
<td>Mandatory</td>
<td></td>
</tr>
<tr>
<td>EXPDTA</td>
<td>Mandatory</td>
<td></td>
</tr>
</tbody>
</table>
# PDB section overview

<table>
<thead>
<tr>
<th>SECTION</th>
<th>DESCRIPTION</th>
<th>RECORD TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Summary descriptive remarks</td>
<td>HEADER, OBSLTE, TITLE, SPLIT, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, NUMMDL, MDLTYP, AUTHOR, REVDAT, SPRSDE, JRNL</td>
</tr>
<tr>
<td>Remark</td>
<td>Various comments about entry annotations in more depth than standard records</td>
<td>REMARKs 0–999</td>
</tr>
<tr>
<td>Primary structure</td>
<td>Peptide and/or nucleotide sequence and the relationship between the PDB sequence and that found in the sequence database(s)</td>
<td>DBREF, SEQADV, SEQRES MODRES</td>
</tr>
<tr>
<td>Heterogen</td>
<td>Description of non-standard groups</td>
<td>HET, HETNAM, HETSYN, FORMUL</td>
</tr>
<tr>
<td>Secondary structure</td>
<td>Description of secondary structure</td>
<td>HELIX, SHEET</td>
</tr>
<tr>
<td>Connectivity annotation</td>
<td>Chemical connectivity</td>
<td>SSBOND, LINK, CISPEP</td>
</tr>
</tbody>
</table>
## PDB section overview (2)

<table>
<thead>
<tr>
<th>SECTION</th>
<th>DESCRIPTION</th>
<th>RECORD TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connectivity annotation</td>
<td>Chemical connectivity</td>
<td>SSBOND, LINK, CISPEP</td>
</tr>
<tr>
<td>Miscellaneous features</td>
<td>Features within the macromolecule</td>
<td>SITE</td>
</tr>
<tr>
<td>Crystallographic</td>
<td>Description of the crystallographic cell</td>
<td>CRyst1</td>
</tr>
<tr>
<td>Coordinate transformation</td>
<td>Coordinate transformation operators</td>
<td>ORIGXn, SCALEn, MTRIXn,</td>
</tr>
<tr>
<td>Coordinate</td>
<td>Atomic coordinate data</td>
<td>MODEL, ATOM, ANISOU, TER, HETATM,</td>
</tr>
<tr>
<td>Connectivity</td>
<td>Chemical connectivity</td>
<td>CONECT</td>
</tr>
<tr>
<td>Bookkeeping</td>
<td>Summary information, end-of-file marker</td>
<td>MASTER, END</td>
</tr>
</tbody>
</table>
PDB record format

- Individually defined per record type
- May have substructure (e.g. Remark)
- E.g. Resolution:

<table>
<thead>
<tr>
<th>COLUMNS</th>
<th>DATA TYPE</th>
<th>FIELD</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 6</td>
<td>Record name</td>
<td>&quot;REMARK&quot;</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>LString(1)</td>
<td>&quot;2&quot;</td>
<td></td>
</tr>
<tr>
<td>12 - 22</td>
<td>LString(11)</td>
<td>&quot;RESOLUTION.&quot;</td>
<td></td>
</tr>
<tr>
<td>24 - 30</td>
<td>Real(7.2)</td>
<td>resolution</td>
<td>Resolution.</td>
</tr>
<tr>
<td>32 - 41</td>
<td>LString(10)</td>
<td>&quot;ANGSTROMS.&quot;</td>
<td></td>
</tr>
</tbody>
</table>

Example:

```
REMARK 2
REMARK 2 RESOLUTION. 1.74 ANGSTROMS.
REMARK 2
REMARK 2 RESOLUTION. NOT APPLICABLE.
REMARK 2
REMARK 2 RESOLUTION. 7.50 ANGSTROMS.
```
Coordinates: ATOM/HETATM

- May be grouped into MODEL/ENDMDL sections (NMR)
- ATOM: amino acids and nucleotides
- HETATM: everything else
- TER: ends a chain
**ATOM Records**

Overview:
The ATOM records present the atomic coordinates for standard amino acids and nucleotides. They also present the occupancy and temperature factor for each atom. Non-polymer chemical coordinates use the HETATM record type. The element symbol is always present on each ATOM record; charge is optional.

Changes in ATOM/HETATM records result from the standardization at atom and residue nomenclature. This nomenclature is described in the Chemical Component Dictionary (ftp://ftp.wwpdb.org/pub/pdb/data/monomers).

### Record Format

<table>
<thead>
<tr>
<th>COLUMNS</th>
<th>DATA TYPE</th>
<th>FIELD</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 6</td>
<td>Record name</td>
<td>&quot;ATOM &quot;</td>
<td></td>
</tr>
<tr>
<td>7 - 11</td>
<td>Integer</td>
<td>serial</td>
<td>Atom serial number.</td>
</tr>
<tr>
<td>13 - 16</td>
<td>Atom</td>
<td>name</td>
<td>Atom name.</td>
</tr>
<tr>
<td>17</td>
<td>Character</td>
<td>altLoc</td>
<td>Alternate location indicator.</td>
</tr>
<tr>
<td>18 - 20</td>
<td>Residue name</td>
<td>resName</td>
<td>Residue name.</td>
</tr>
<tr>
<td>22</td>
<td>Character</td>
<td>chainID</td>
<td>Chain identifier.</td>
</tr>
<tr>
<td>23 - 26</td>
<td>Integer</td>
<td>resSeq</td>
<td>Residue sequence number.</td>
</tr>
<tr>
<td>27</td>
<td>AChar</td>
<td>iCode</td>
<td>Code for insertion of residues.</td>
</tr>
<tr>
<td>31 - 38</td>
<td>Real(8.3)</td>
<td>x</td>
<td>Orthogonal coordinates for X in Angstroms.</td>
</tr>
<tr>
<td>39 - 46</td>
<td>Real(8.3)</td>
<td>y</td>
<td>Orthogonal coordinates for Y in Angstroms.</td>
</tr>
<tr>
<td>47 - 54</td>
<td>Real(8.3)</td>
<td>z</td>
<td>Orthogonal coordinates for Z in Angstroms.</td>
</tr>
<tr>
<td>55 - 60</td>
<td>Real(6.2)</td>
<td>occupancy</td>
<td>Occupancy.</td>
</tr>
<tr>
<td>61 - 66</td>
<td>Real(6.2)</td>
<td>tempFactor</td>
<td>Temperature factor.</td>
</tr>
<tr>
<td>77 - 78</td>
<td>LString(2)</td>
<td>element</td>
<td>Element symbol, right-justified.</td>
</tr>
<tr>
<td>79 - 80</td>
<td>LString(2)</td>
<td>charge</td>
<td>Charge on the atom.</td>
</tr>
</tbody>
</table>

Details:

- ATOM records for proteins are listed from amino to carboxy terminus.
- Nucleic acid residues are listed from the 5' to 3' terminus.
- Alignment of one-letter atom name such as C starts at column 14, while two-letter atom name such as FE starts at column 13.
### Example for Coordinates

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1234567</td>
<td>890123456</td>
<td>7890123456</td>
<td>7890123456</td>
<td>7890123456</td>
<td>7890123456</td>
<td>7890123456</td>
<td>7890123456</td>
</tr>
<tr>
<td>ATOM</td>
<td>601</td>
<td>N</td>
<td>LEU A</td>
<td>75</td>
<td>-17.070</td>
<td>-16.002</td>
<td>2.409</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>602</td>
<td>CA</td>
<td>LEU A</td>
<td>75</td>
<td>-16.343</td>
<td>-16.746</td>
<td>3.444</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>603</td>
<td>C</td>
<td>LEU A</td>
<td>75</td>
<td>-16.499</td>
<td>-18.263</td>
<td>3.300</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>604</td>
<td>O</td>
<td>LEU A</td>
<td>75</td>
<td>-16.645</td>
<td>-18.789</td>
<td>2.195</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>605</td>
<td>CB</td>
<td>LEU A</td>
<td>75</td>
<td>-16.776</td>
<td>-16.283</td>
<td>4.844</td>
<td>1.00</td>
</tr>
<tr>
<td>TER</td>
<td>606</td>
<td></td>
<td>LEU A</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>1185</td>
<td>O</td>
<td>LEU B</td>
<td>75</td>
<td>26.292</td>
<td>-4.310</td>
<td>16.940</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>1186</td>
<td>CB</td>
<td>LEU B</td>
<td>75</td>
<td>23.881</td>
<td>-1.551</td>
<td>16.797</td>
<td>1.00</td>
</tr>
<tr>
<td>TER</td>
<td>1187</td>
<td></td>
<td>LEU B</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HETATM</td>
<td>1188</td>
<td>H2</td>
<td>SRT A1076</td>
<td></td>
<td>-17.263</td>
<td>11.260</td>
<td>28.634</td>
<td>1.00</td>
</tr>
<tr>
<td>HETATM</td>
<td>1189</td>
<td>HA</td>
<td>SRT A1076</td>
<td></td>
<td>-19.347</td>
<td>11.519</td>
<td>28.341</td>
<td>1.00</td>
</tr>
<tr>
<td>HETATM</td>
<td>1190</td>
<td>H3</td>
<td>SRT A1076</td>
<td></td>
<td>-17.157</td>
<td>14.303</td>
<td>28.677</td>
<td>1.00</td>
</tr>
<tr>
<td>HETATM</td>
<td>1191</td>
<td>HB</td>
<td>SRT A1076</td>
<td></td>
<td>-15.110</td>
<td>13.610</td>
<td>28.816</td>
<td>1.00</td>
</tr>
<tr>
<td>HETATM</td>
<td>1192</td>
<td>O1</td>
<td>SRT A1076</td>
<td></td>
<td>-17.028</td>
<td>11.281</td>
<td>31.131</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>295</td>
<td>HB2</td>
<td>ALA A</td>
<td>18</td>
<td>4.601</td>
<td>-9.393</td>
<td>7.275</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>296</td>
<td>HB3</td>
<td>ALA A</td>
<td>18</td>
<td>3.340</td>
<td>-9.147</td>
<td>6.043</td>
<td>1.00</td>
</tr>
<tr>
<td>TER</td>
<td>297</td>
<td></td>
<td>ALA A</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENDMDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sequences in PDB files

- **SEQRES**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>21</td>
<td>GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>21</td>
<td>TYR GLN LEU GLU ASN TYR CYS ASN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>30</td>
<td>PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>30</td>
<td>ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>30</td>
<td>THR PRO LYS ALA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>21</td>
<td>GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>21</td>
<td>TYR GLN LEU GLU ASN TYR CYS ASN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>30</td>
<td>PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>30</td>
<td>ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>30</td>
<td>THR PRO LYS ALA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Implicit from the ATOM definition
- **Careful!** : Not all residues from SEQRES have coordinates
Parsing PDB files

- Myriad of parsers exist, writing your own not too difficult
- e.g. Bioperl (Bio::Structure):

```perl
$structio = Bio::Structure::IO->new( -file => "1XYZ.pdb" );
$struc = $structio->next_structure;  # returns an Entry object
$pseq = $struc->seqres;             # returns a PrimarySeq object
@atoms = $struc->get_atoms($res);   # Atom objects, given a Residue
@xyz = $atom->xyz;                 # the 3D coordinates of the atom
```

- e.g. Biopython (Bio.PDB):

```python
parser = PDBParser()
structure = parser.get_structure ( 'PHA-L', '1FAT.pdb' )
    for model in structure:
        for chain in model:
            for residue in chain:
                for atom in residue:
                    print atom
```
Note about software architecture

• Interfaces and encapsulation help for format changes!
• e.g. biopython:
  Bio.PDB.PDBParser and Bio.PDB.MMCIFParser return the same Bio.PDB.Structure objects
PDBx/mmCIF overview

- Standard format for PDB since 2014
- Originated in the crystallographic community
- The STAR file: a new format for electronic data transfer and archiving
  

- Advantages of new format:
  - Extensible
  - Flexible with respect to order
  - Few syntax rules
  - Facilitates automatic validation (with mmCIF dictionary)
STAR (PDBx/mmCIF) syntax

- **Data name**: any text string starting with an underline (unique) (e.g. `chemical_formula`)
- **Data item**: any text string not starting with an underline, preceded by data name (e.g. ‘C23 H36 O7’)
- **Data loop**: list of data names, preceded by ‘loop_’, followed by repeated data items
  
  ```
  loop_
  _citation_author.citation_id
  _citation_author.ordinal
  _citation_author.name
  primary 1 'Fitzgerald, P.M.D.'
  primary 2 'McKeever, B.M.'
  primary 3 'Van Middlesworth, J.F.'
  ```
- **Data block**: collection of data, preceded by ‘data_***’
Concepts in mmCIF

- **Entity**: polymer / non-polymer / water
- **Chemical component**: blocks that build entities (e.g. non-standard residue)
- **Structural Component**: structural features, e.g. helix
- **Asymmetric Unit Component**: (chain), two components can refer to same entity
- **Biological Component**: sub- and super-components of the structure
Finding data in mmCIF

- **Header:**

  ```
  HEADER    PLANT SEED PROTEIN     11-OCT-91    1CBN
  ```

  becomes

  ```
  _struct.entry_id             '1CBN'
  _struct.title               'PLANT SEED PROTEIN'
  _struct_keywords.entry_id   '1CBN'
  _struct_keywords.text       'plant seed protein'
  _database_2.database_id     PDB
  _database_2.database_code   1CBN
  _database_PDB_rev.num       1
  _database_PDB_rev.date_original 1991-10-11
  ```

- **Resolution:** `_refine.ls_d_res_high`
- **Reference to sequence database:** `_struct_ref, _struct_ref_seq`
The enzymatically competent form of HIV protease is a dimer. This entity corresponds to one monomer of an active dimer. The structure of the closely related compound, isobutyryl-pepstatin (pepstatin A) in complex with rhizopus pepsin.
Sequence data in mmCIF

- Reference to database (_struct_ref, _struct_ref_seq)
- Polymer definition in _entity_poly_seq
- Sequence in _entity_poly.pdbx_seq_one_letter_code_can
  - _entity_poly_seq.entity_id
  - _entity_poly_seq.num
  - _entity_poly.entity_id
  - _entity_poly.type
  - _entity_poly.nstd_linkage
  - _entity_poly.nstd_monomer
  - _entity_poly.pdbx_seq_one_letter_code
  - _entity_poly.pdbx_seq_one_letter_code_can
  - _entity_poly.pdbx_strand_id
  - 'polypeptide(L)' no no TFGSGEADCGLRPLFEKKSLEDKTERELLESYIDGR
  - TFGSGEADCGLRPLFEKKSLEDKTERELLESYIDGR L
Parsing mmCIF files

- Python modules:
  - PDBx
  - PDBeCIF
  - mmLIB
  - bioPython
- Java: bioJava
- C++: cifparse-obj
Molecular Graphics Programs

• See links @PDB
• Famous tools:
  – PyMol
  – Jmol
  – Chimera
  – VMD
  – Yasara
Views of a PDB structure

- **Representation**
  - Wireframe
  - Spacefill
  - Backbone
- **Colouring**
- **Surface**

Images:
Aquaria: simplifying discovery and insight from protein structures

Seán I. O’Donoghue\textsuperscript{1,2,3}, Kenneth S. Sabir\textsuperscript{2,3}, Maria Kalemanov\textsuperscript{4}, Christian Stolte\textsuperscript{1}, Neil Saunders\textsuperscript{1}, Bill Wilson\textsuperscript{1}, Benjamin Wellmann\textsuperscript{4}, Vivian Ho\textsuperscript{1,6}, Manfred Roos\textsuperscript{4}, Nelson Perdigão\textsuperscript{5}, Fabian A. Buske\textsuperscript{2,6}, Julian Heinrich\textsuperscript{1}, Burkhard Rost\textsuperscript{4} & Andrea Schafferhans\textsuperscript{4}

\textsuperscript{1}. CSIRO, 2. Garvan, 3. University of Sydney, 4. Technische Universität München, 5. Universidade de Lisboa, 6. UNSW

Overview

• The idea behind Aquaria
• What does it look like
• History of Aquaria
• Scientific background
• Technical background
MOLECULAR MACHINERY: A Tour of the Protein Data Bank

110,688
(70,953)

06/2015
An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity.

DOI: 10.2210/pdb3q01/pdb

Primary Citation

An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity.

Petty, T.J., Emamzadah, S., Costantino, L., Petkova, I., Stavridi, E.S., Savva, J.G., Vauthey, E., Halazonetis, T.D.


PubMed: 21522129
PubMedCentral: PMC3117648
DOI: 10.1038/emboj.2011.127
Search Related Articles in PubMed

PubMed Abstract:

The p53 tumour suppressor gene, the most frequently mutated gene in human cancer, encodes a transcription factor that contains sequence-specific DNA binding and homo-tetramerization domains. Interestingly, the affinities of p53 for specific and non-specific DNA sites differ by only one... [Read More & Search PubMed Abstracts]

Molecular Description

Classification: Antitumor Protein
Structure Weight: 53020.76

Molecule: Cellular tumor antigen p53
Polymer: 1
Length: 233
Growth of protein sequence data

# entries in TREMBL (total sequence number)

- 5E6
- 4E6
- 3E6

# entries in Swissprot (curated sequences)

- 5E5
- 1E5

7E4 known structures
Please specify a Human protein.

Or, specify a new default organism.
An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity.

Petty et al., EMBO J (2011)

Abstract: The p53 tumour suppressor gene, the most frequently mutated gene in human cancer, encodes a transcription factor that contains sequence-specific DNA binding and homo-tetramerization domains. Interestingly, the affinities of p53 for specific and non-specific DNA sites differ by only one order of magnitude, making it hard to understand how this protein recognizes its specific DNA targets in vivo. We describe here the structure of a p53 polypeptide containing both the DNA binding and oligomerization domains in complex with DNA. The structure reveals that sequence-specific DNA binding...
Aquaria: P53, 93% Sequence Identity to PDB 3q01, chain A

Abstract: The p53 tumour suppressor gene, the most frequently mutated gene in human cancer, encodes a transcription factor that contains sequence-specific DNA binding and homo-oligomerization domains. Interestingly, the affinities of p53 for specific and non-specific DNA sites differ by only one order of magnitude, making it hard to understand how this protein recognizes its specific DNA targets in vivo. We describe here the structure of a p53 polyproteome containing both the DNA binding and oligomerization domains.

3kz8-A: A polyprotein containing both the DNA binding and oligomerization domains.

3kz8, chain A: Cellular tumor antigen P53 (Tumor suppressor p53, phosphoprotein P53, antigen nyco-1:3) Diversity in DNA recognition by P53 revealed by crystal structures with high sequence pairs (P53-DNA complex 3) (X-ray diffraction)
Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs.


Abstract: p53 binds as a tetramer to DNA targets consisting of two decameric half-sites separated by a variable spacer. Here we present high-resolution crystal structures of complexes between p53 core-domain tetramers and DNA targets consisting of contiguous half-sites. In contrast to previously reported p53-DNA complexes that show standard Watson-Crick base pairs, the newly reported structures show noncanonical Hoogsteen base-pairing geometry at the central A-T doublet of each half-site. Structural and computational analyses show that the Hoogsteen geometry distinctly modulates the DNA...
### Mouse Controls

**Selection**

<table>
<thead>
<tr>
<th>Action</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Click on:</strong></td>
<td></td>
</tr>
<tr>
<td>Atom</td>
<td>Selects atom</td>
</tr>
<tr>
<td>Residue</td>
<td>Selects residue</td>
</tr>
<tr>
<td>Annotation lane (H)</td>
<td>Activate annotation coloring &amp; linking</td>
</tr>
<tr>
<td>Feature (F)</td>
<td>Selects feature</td>
</tr>
<tr>
<td>Element in sequence overview (s)</td>
<td>Selects secondary structure element</td>
</tr>
<tr>
<td>Background</td>
<td>Clears selection</td>
</tr>
<tr>
<td><strong>Ctrl + click on:</strong></td>
<td></td>
</tr>
<tr>
<td>Selected object</td>
<td>Extends selection to next level in hierarchy</td>
</tr>
<tr>
<td>Unselected object</td>
<td>Adds object to selection</td>
</tr>
<tr>
<td><strong>Ctrl + Shift + click on:</strong></td>
<td></td>
</tr>
<tr>
<td>Selected object</td>
<td>Removes object from selection</td>
</tr>
<tr>
<td>Deselected object</td>
<td>Removes next level of hierarchy from selection</td>
</tr>
<tr>
<td><strong>Shift + click</strong></td>
<td>Selects a range of residues</td>
</tr>
<tr>
<td><strong>Alt + click</strong></td>
<td>Selects whole chain</td>
</tr>
</tbody>
</table>

**Rotation and Translation**

<table>
<thead>
<tr>
<th>Action</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drag</td>
<td>Rotates structure around X &amp; Y axes</td>
</tr>
<tr>
<td>Ctrl + drag (left &amp; right)</td>
<td>Rotates structure around Z axis</td>
</tr>
<tr>
<td>Right button drag</td>
<td>Translates structure along X &amp; Y axes</td>
</tr>
</tbody>
</table>

**Zooming**

<table>
<thead>
<tr>
<th>Action</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shift + drag (up &amp; down)</strong></td>
<td>Zooms out &amp; in</td>
</tr>
<tr>
<td>Double click</td>
<td>Auto zooms to object</td>
</tr>
<tr>
<td>Shift + double click</td>
<td>Auto zooms to residue range or ligand</td>
</tr>
<tr>
<td><strong>Alt + double click</strong></td>
<td>Auto zooms to whole chain</td>
</tr>
</tbody>
</table>

---

### Keyboard Controls

<table>
<thead>
<tr>
<th>Action</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-Z or *</td>
<td>Finds text patterns in sequences</td>
</tr>
<tr>
<td>Esc</td>
<td>Clears selection</td>
</tr>
<tr>
<td>Enter</td>
<td>Auto zooms to selection; if no selection is made, zooms to whole structure</td>
</tr>
<tr>
<td>Shift + Enter</td>
<td>Auto zooms to whole structure</td>
</tr>
<tr>
<td>Right arrow</td>
<td>Selects next residue in sequence</td>
</tr>
<tr>
<td>Left arrow</td>
<td>Selects previous residue in sequence</td>
</tr>
<tr>
<td>Shift + right arrow</td>
<td>Adds next residue to selection</td>
</tr>
<tr>
<td>Shift + left arrow</td>
<td>Adds previous residue to selection</td>
</tr>
<tr>
<td>Up arrow</td>
<td>Selects next highest level in hierarchy</td>
</tr>
<tr>
<td>Tab</td>
<td>Selects next object, depending on selected object</td>
</tr>
<tr>
<td>Shift + Tab</td>
<td>Selects previous object</td>
</tr>
<tr>
<td>2</td>
<td>Rotates about X axis (anticlockwise)</td>
</tr>
<tr>
<td>8</td>
<td>Rotates about X axis (clockwise)</td>
</tr>
<tr>
<td>4</td>
<td>Rotates about Y axis (clockwise)</td>
</tr>
<tr>
<td>6</td>
<td>Rotates about Y axis (anticlockwise)</td>
</tr>
<tr>
<td>Ctrl + 2, 4, 6 or B</td>
<td>Rotates in 45° steps (axes as above)</td>
</tr>
<tr>
<td>Alt + 2, 4, 6 or B</td>
<td>Translates (axes as above)</td>
</tr>
<tr>
<td>+</td>
<td>Zooms in stepwise</td>
</tr>
<tr>
<td>-</td>
<td>Zooms out stepwise</td>
</tr>
<tr>
<td>Ctrl + C</td>
<td>Copies selection to paste buffer</td>
</tr>
<tr>
<td>Ctrl + V</td>
<td>Pastes a text string from another program into the search buffer</td>
</tr>
<tr>
<td>Ctrl + L</td>
<td>Selects all ligands</td>
</tr>
<tr>
<td>Ctrl + R</td>
<td>Resets view to initial state</td>
</tr>
<tr>
<td>Ctrl + S</td>
<td>Saves view state</td>
</tr>
<tr>
<td>Ctrl + Shift + L</td>
<td>Labels selection (toggle)</td>
</tr>
<tr>
<td>Ctrl + Shift + W</td>
<td>Toggles wireframe representation</td>
</tr>
<tr>
<td>Alt + Ctrl + PrtSc</td>
<td>Prints screen image to paste buffer</td>
</tr>
</tbody>
</table>

---

Aquaria 3D Viewer is based on SHS 3D Viewer, originally developed at LION bioscience AG.
### Applet Menu

**FILE**

<table>
<thead>
<tr>
<th>Menu option</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Save</td>
<td>Saves view state (automatically used when structure is next loaded)</td>
</tr>
<tr>
<td>Reload</td>
<td>Reloads saved view state</td>
</tr>
<tr>
<td>Revert</td>
<td>Reverts the entire view (rotation, representation, coloring etc.) to the initial state.</td>
</tr>
</tbody>
</table>

**Export Image...** Save the current view as a PNG image. You can specify image dimensions and background color (white/grey/black).

**Print...** Opens the current view in a separate window for printing. You can specify image dimensions, background color, and toggle color legend and selection overlays.

**EDIT**

<table>
<thead>
<tr>
<th>Menu option</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy</td>
<td>Copies residues in ClustalW format, for pasting into text processor (e.g. Notepad)</td>
</tr>
<tr>
<td>Paste</td>
<td>Pastes text string from another program; finds occurrences of string in all sequences</td>
</tr>
<tr>
<td>Select All</td>
<td>Selects all objects</td>
</tr>
<tr>
<td>Select Ligands</td>
<td>Selects all ligands</td>
</tr>
<tr>
<td>Select Proximity</td>
<td>Selects everything within 4Å of current selection</td>
</tr>
<tr>
<td>Select Water</td>
<td>Selects all water molecules in the structure</td>
</tr>
</tbody>
</table>

**COLOR**

<table>
<thead>
<tr>
<th>Menu option</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence Similarity</td>
<td>Colors selection by sequence similarity (default)</td>
</tr>
<tr>
<td>Secondary Structure</td>
<td>Colors selection by secondary structure</td>
</tr>
<tr>
<td>Molecule</td>
<td>Colors selection by molecule type</td>
</tr>
<tr>
<td>Chains</td>
<td>Colors selection by chain</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>Colors selection by hydrophobicity index</td>
</tr>
<tr>
<td>Polarity</td>
<td>Colors selection by charge type</td>
</tr>
<tr>
<td>Temperature</td>
<td>Colors selection by B-factor (high B-factor is correlated with high temperature)</td>
</tr>
<tr>
<td>Element</td>
<td>Colors selected atoms by chemical element type (CPK standard).</td>
</tr>
</tbody>
</table>

**Choose Color...** Opens a color picker to apply chosen color to the selection

**REPRESENTATION**

<table>
<thead>
<tr>
<th>Menu option</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visible</td>
<td>Displays / hides selection</td>
</tr>
<tr>
<td>Transparent</td>
<td>Makes selection slightly transparent</td>
</tr>
<tr>
<td>Ribbon</td>
<td>Cartoon representation of the backbone</td>
</tr>
<tr>
<td>C-alpha</td>
<td>Straight lines connecting all Co atoms</td>
</tr>
<tr>
<td>Wireframe</td>
<td>Atoms connected by straight lines</td>
</tr>
<tr>
<td>Spacefill</td>
<td>Atoms as spheres with van der Waals radii</td>
</tr>
<tr>
<td>Ball &amp; Stick</td>
<td>Atoms as spheres, bonds as sticks</td>
</tr>
<tr>
<td>Surface</td>
<td>Calculates the molecular surface of the selection</td>
</tr>
<tr>
<td>Remove All Surfaces</td>
<td>Removes any calculated surfaces</td>
</tr>
</tbody>
</table>

**VIEW**

<table>
<thead>
<tr>
<th>Menu option</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Proteins &amp; DNA/RNA shown as ribbons, ligands as spacefill, disulfide bonds as ball &amp; stick. Uses Sequence Similarity coloring scheme.</td>
</tr>
<tr>
<td>Ligand Binding Sites</td>
<td>Highlights ligand &amp; binding site interactions. Polypeptide chains shown as C-alpha trace. All atoms within 4Å of ligand shown as wireframe.</td>
</tr>
<tr>
<td>Spin</td>
<td>Toggles spinning around Y axis on &amp; off (drag to change spin axis)</td>
</tr>
<tr>
<td>Zoom to Selection</td>
<td>Auto zooms to current selection</td>
</tr>
<tr>
<td>Zoom Out</td>
<td>Auto zooms to whole structure</td>
</tr>
<tr>
<td>Show PDB Sequence</td>
<td>Displays sequence track for current structure</td>
</tr>
<tr>
<td>Show Water</td>
<td>Displays all water molecules in the structure</td>
</tr>
<tr>
<td>Mouse Controls</td>
<td>Toggles mouse controls overlay on &amp; off</td>
</tr>
<tr>
<td>Color Legend</td>
<td>Toggles color legend on &amp; off</td>
</tr>
</tbody>
</table>

**TOOLS**

<table>
<thead>
<tr>
<th>Menu option</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>&gt; Create New... bases new annotation on current selection</td>
</tr>
<tr>
<td>Label</td>
<td>&gt; Molecules Labels each molecule</td>
</tr>
<tr>
<td></td>
<td>&gt; Chains Labels each chain</td>
</tr>
<tr>
<td></td>
<td>&gt; Secondary Structure Labels each secondary structure element in selection</td>
</tr>
<tr>
<td>Calculate</td>
<td>&gt; Distance Calculates the closest interatomic distance between two selected regions</td>
</tr>
<tr>
<td></td>
<td>&gt; Atomic Contacts Calculates van der Waals overlaps between selection &amp; rest of structure</td>
</tr>
<tr>
<td></td>
<td>&gt; Remove All Distances Removes any displayed distances</td>
</tr>
<tr>
<td></td>
<td>&gt; Hydrogen Bonds Calculates hydrogen bonds for current selection</td>
</tr>
<tr>
<td>Advanced Settings</td>
<td>&gt; Move as C-alpha Trace Enables fast rotation and translation</td>
</tr>
<tr>
<td></td>
<td>&gt; High Quality Display smooth curves with specular highlights and shading</td>
</tr>
<tr>
<td></td>
<td>&gt; About Display build information &amp; authors</td>
</tr>
</tbody>
</table>
Time travel
SRS 3D Team in 2003
SRS

SRS = Sequence Retrieval System

- Frontend for biological (flat file) databases
- Creates links and complex queries
- See e.g. http://www.dkfz.de/srs/
Linking sequences and structures

HSSP: lists homologous sequences for all structures; based on Maxhom

MaxHom-Filter[2]:

sequence features

- SWALL
- SwissProt Features
- InterPro
- SNPs
- exon boundary
- bioSCOUT features
- custom features

sequences

- PDB
- SWALL
- sequences

structures

sequence-to-structure alignments

- PSSH
- HSSP-align
- HSSP-chain
SRS 3D Architecture

Perl scripts
- updatePSSH
- updateHSSP2
- makePSSH
- makeHSSP2
- MaxHom2

Databases
- Uniprot
- PDB
- PSSH
- HSSP-align
- HSSP-chain

SRS scripts
- 5 parsers
- 5 loaders
- 10 views

3D viewer

5 parsers
5 loaders
10 views
Quick Search: human nuclear receptor

Swall

Swall RXRB Human P28702 Retinoic acid receptor RXR-beta

1ho2-A retinoic acid receptor

2.1 angstrom resolution crystal structure of the heterodimer of the human ralpalpha and
ppgammagamma liganded binding domains respectively bound with 9-cis retinoic acid and
p263570 and co-activator peptides.
SRS 3D

Integrating 3D structures, sequences, and features.

blk_human
Search I'm Feeling Lucky

About BioWisdom • About EMBL • About SRS 3D • Contact Us

http://srs3d.org
Entry view - PSSH:P51451

<table>
<thead>
<tr>
<th>PSSH:P51451</th>
<th>All 3D Structures with Significant Matches to Sequence P51451 (505 Residues)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description of P51451</strong></td>
<td>RecName: Full=Tyrosine-protein kinase BLK; EC=2.7.10.2; AltName: Full=B lymphocyte kinase; AltName: Full=p55-BLK;</td>
</tr>
<tr>
<td><strong>Keywords for P51451</strong></td>
<td>ATP-binding</td>
</tr>
<tr>
<td><strong>Organism</strong></td>
<td>Homo sapiens (Human)</td>
</tr>
<tr>
<td><strong>Number of Matches</strong></td>
<td>587 (Only top 400 displayed in image map)</td>
</tr>
</tbody>
</table>

**Sequence features:**

**Matching structures:**

---

http://srs3d.org
PSSH2: The database behind Aquaria
<4%
HSSP – PSSH – PSSH2

• Since 1991 (Dodge, Schneider, Sander)
  HSSP = Homology-derived Structures of Proteins

• 2003
  PSSH = Protein Sequence-to-Structure Homologies

• 2010
  Too many sequences (>500,000 Swissprot) and structures (>60,000 PDB chains)
  → faster method needed
  → PSSH2
## CASP9(2010)

**HHblits**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HHpredA</td>
</tr>
<tr>
<td>2</td>
<td>HHpredB</td>
</tr>
<tr>
<td>3</td>
<td>Zhang-Server</td>
</tr>
<tr>
<td>4</td>
<td>HHpredC</td>
</tr>
<tr>
<td>5</td>
<td>QUARK</td>
</tr>
<tr>
<td>6</td>
<td>RaptorX-MSA</td>
</tr>
<tr>
<td>7</td>
<td>RaptorX</td>
</tr>
<tr>
<td>8</td>
<td>RaptorX-Boost</td>
</tr>
<tr>
<td>9</td>
<td>Seok-server</td>
</tr>
<tr>
<td>10</td>
<td>MULTICOM-CLUSTER</td>
</tr>
<tr>
<td>11</td>
<td>gws</td>
</tr>
<tr>
<td>12</td>
<td>Phyre2</td>
</tr>
<tr>
<td>13</td>
<td>BAKER-ROSETTASERVER</td>
</tr>
<tr>
<td>14</td>
<td>MULTICOM-NOVEL</td>
</tr>
<tr>
<td>15</td>
<td>MULTICOM-REFINE</td>
</tr>
<tr>
<td>16</td>
<td>chunk-TASSER</td>
</tr>
<tr>
<td>17</td>
<td>ZHOU-SPARKS-X</td>
</tr>
<tr>
<td>18</td>
<td>MULTICOM-CONSTRUCT</td>
</tr>
<tr>
<td>19</td>
<td>FAMSD</td>
</tr>
<tr>
<td>20</td>
<td>MUFOLD-Server</td>
</tr>
<tr>
<td>21</td>
<td>PconsD</td>
</tr>
<tr>
<td>22</td>
<td>PconsM</td>
</tr>
<tr>
<td>23</td>
<td>pro-sp3-TASSER</td>
</tr>
<tr>
<td>24</td>
<td>Pcomb</td>
</tr>
<tr>
<td>25</td>
<td>Bilab-ENABLE</td>
</tr>
<tr>
<td>26</td>
<td>IntFOLD-TS</td>
</tr>
<tr>
<td>27</td>
<td>Jiang_THREADER</td>
</tr>
<tr>
<td>28</td>
<td>Jiang_Assembly</td>
</tr>
<tr>
<td>29</td>
<td>FALCON-SWIFT</td>
</tr>
<tr>
<td>30</td>
<td>CLEF-Server</td>
</tr>
<tr>
<td>31</td>
<td>Pcons</td>
</tr>
<tr>
<td>32</td>
<td>ProQ2</td>
</tr>
<tr>
<td>33</td>
<td>SAM-T08-server</td>
</tr>
<tr>
<td>34</td>
<td>prdos2</td>
</tr>
<tr>
<td>35</td>
<td>circle</td>
</tr>
<tr>
<td>36</td>
<td>chuo-fams</td>
</tr>
<tr>
<td>37</td>
<td>ProfileCRF</td>
</tr>
<tr>
<td>38</td>
<td>GSmetaserver</td>
</tr>
<tr>
<td>39</td>
<td>3D-JIGSAW_V4-0</td>
</tr>
<tr>
<td>40</td>
<td>FFAS03n</td>
</tr>
<tr>
<td>41</td>
<td>Atome2_CBS</td>
</tr>
<tr>
<td>42</td>
<td>BioSerf</td>
</tr>
<tr>
<td>43</td>
<td>MUSTER</td>
</tr>
<tr>
<td>44</td>
<td>MidwayFoldingServer</td>
</tr>
<tr>
<td>45</td>
<td>FFAS03ss</td>
</tr>
<tr>
<td>46</td>
<td>3D-JIGSAW_V4-5</td>
</tr>
<tr>
<td>47</td>
<td>PconsR</td>
</tr>
<tr>
<td>48</td>
<td>FFAS03a</td>
</tr>
<tr>
<td>49</td>
<td>Distill</td>
</tr>
<tr>
<td>50</td>
<td>FFAS03</td>
</tr>
<tr>
<td>51</td>
<td>PRECORS</td>
</tr>
<tr>
<td>52</td>
<td>Wolfson-serv</td>
</tr>
<tr>
<td>53</td>
<td>MUFOLD-MD</td>
</tr>
<tr>
<td>54</td>
<td>LOOPP_Austin</td>
</tr>
<tr>
<td>55</td>
<td>MUSICS_server</td>
</tr>
<tr>
<td>56</td>
<td>panther</td>
</tr>
<tr>
<td>57</td>
<td>PROTAGORAS</td>
</tr>
<tr>
<td>58</td>
<td>ProQ</td>
</tr>
<tr>
<td>59</td>
<td>SAM-T02-server</td>
</tr>
<tr>
<td>60</td>
<td>LMUserver</td>
</tr>
<tr>
<td>61</td>
<td>YASARA</td>
</tr>
<tr>
<td>62</td>
<td>SAM-T06-server</td>
</tr>
<tr>
<td>63</td>
<td>FUGUE_KM</td>
</tr>
<tr>
<td>64</td>
<td>Ma-OPUS-server</td>
</tr>
<tr>
<td>65</td>
<td>m4t_2009</td>
</tr>
<tr>
<td>66</td>
<td>Pushchino</td>
</tr>
<tr>
<td>67</td>
<td>MUSICS-2S</td>
</tr>
<tr>
<td>68</td>
<td>rehtnap</td>
</tr>
<tr>
<td>69</td>
<td>RBO-PROTEUS</td>
</tr>
<tr>
<td>70</td>
<td>RaptorX-FM</td>
</tr>
<tr>
<td>71</td>
<td>Yang_kdd</td>
</tr>
<tr>
<td>72</td>
<td>STAT-PROTEUS</td>
</tr>
<tr>
<td>73</td>
<td>LenServer</td>
</tr>
<tr>
<td>74</td>
<td>BHAGEERATH</td>
</tr>
<tr>
<td>75</td>
<td>ConStruct</td>
</tr>
<tr>
<td>76</td>
<td>PLATO</td>
</tr>
<tr>
<td>77</td>
<td>schenk-torda-server</td>
</tr>
<tr>
<td>78</td>
<td>Fortmann_server</td>
</tr>
<tr>
<td>79</td>
<td>PHAISTOSserver</td>
</tr>
</tbody>
</table>


CASPI10 (2012)

370× slower
HHblits

1. Zhang-Server
2. QUARK
3. BAKER-ROSETTASERVER
4. RaptorX-ZY
5. RaptorX
6. TASSER-VMT
7. HPredA
8. HPredAQ
9. PMS
10. HPred-thread
11. PconsM
12. chunk-TASSER
13. Pcons-net
14. MULTICOM-REFINE
15. MUFOLD-Server
16. MULTICOM-NOVEL
17. Mufold-MD
18. MULTICOM-CLUSTER
19. Phyre2_A
20. MULTICOM-CONSTRUCT
21. Seek-server
22. ZHOU-SPARKS-X
23. PconsD
24. FALCON-TOPO
25. FALCON-TOPO-X
26. SAM-T08-server
27. hGen3D
28. Distill
29. IntFOLD2
30. samcha-server
31. NewSerf
32. FFAS03c
33. FFAS03mt
34. sibio
35. chuo-fams-server
36. 3D-JIGSAW_V5-0
37. MUFold_CRF
38. Bilab-ENABLE
39. Atome2_CBS
40. MATRIX
41. Distill_roll
42. chuo-repack-server
43. IntFOLD
44. STRINGS
45. FRESS_server
46. FFAS03hj
47. FFAS03
48. Aoba-server
49. YASARA
50. Jiang_Server
51. UGACSBL
52. GSmetaserver
53. Jiang_Threader
54. Jiang_Fold
55. sysimm
56. BhageerathH
57. PROTAGORAS
58. SAM-T08-server
59. panther
60. RBO-MBS
61. RBO-i-MBS
62. HOMER
63. RBO-i-MBS-BB
64. RBO-MBS-BB
65. RaptorX-Roll
66. Lenserver
67. Bhageerath_abinitio
68. confuzzGS
69. confuzz3d

Background on HHblits

- *query and database* sequences represented by profile hidden Markov models (HMMs)
- Profile HMM: for each sequence position probabilities to observe each of 20 amino acids in related proteins
- Similar to HHsearch, but with prefilter for speed

Remmert, M., Biegert, A., Hauser, A. & Söding, J.
HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.
HHblits profile database

- Uniprot20 (4.8 mio clusters): sequences clustered to 20% sequence identity provided there is
  - (almost) full-length alignability
  - ≥ 80% coverage
- Each cluster → MSA → profile HMM → discretized profile (219 states) for pre-filtering
PSSH2 validation

TP/FP based on TopMatch/COPS L30: ≥30% structural overlap
PSSH2 workflow

Workflow for generating PSSH2. We used HHblits to find similarities between HMMs from PDB and HMMs from SwissProt, resulting in a total of 46 million sequence alignments. Of these, 9.3 million are high confidence; 1.2 million are medium confidence; and 9.3 million are low confidence. During the medium confidence step, the alignments were clustered based on pairwise sequence identity between clusters. For each unique PDB sequence (PDB_full), 9.3 million high or medium confidence alignments were considered. Within the medium confidence zone (see below), alignments of high or medium confidence were considered to indicate alignments of files with 3D structure similar to that observed in the PDB entry used.

Finally, we generated PSSH2. We created a database of HMMs for every unique protein sequence profile. To create these, we ran HHblits HH-suite to create a database of HMMs for every unique sequence. In the first step, 657,138 sequence alignments were found using HHblits against PDB. In the second step, 195,501 sequence alignments were found using HHblits against UniProt20, a database of 9.3 million protein sequence profiles. We then used the same process to create a database of HMMs for every unique sequence in each cluster, resulting in 4.8 million sequence HMMs. We defined this set of HMMs as the PSSH2 workflow. The latest version of the PSSH2 workflow contains 57,727 HMMs.

The recall, precision, and false positive rate (FPR) were then calculated by considering all matches with less than or equal to a given expected value (E-value) from HHblits. We defined recall as the True Positive Rate (Recall) = \( \frac{TP}{TP + FN} \), precision as Precision = \( \frac{TP}{TP + FP} \), where TP is the number of true positive alignments, FP is the number of false positive alignments, and FN is the number of false negative alignments. Weibull distribution: the COPS L30 dataset was used to calculate the recall and precision by considering all matches with an expected value for each sequence (value for each sequence was used). For both these steps, we used UniProt20, a database of 9.3 million protein sequence profiles. We then used the same process to create a database of HMMs for every unique sequence in PDB (derived from the PDB entries). Of these, 9.3 million are high confidence; 1.2 million are medium confidence; and 9.3 million are low confidence. During the medium confidence step, the alignments were clustered based on pairwise sequence identity between clusters.
<4%
88%

126 million sequence-to-structure alignments
160 per protein
Reinventing the technical set-up
The new user interface integrates the functions of searching for structures (A) and viewing clusters of matching structures (B) with the 3D viewer (C), which provides an interactive context linking structure, sequence, and features. In the structures, transparency is used to indicate a lack of agreement between the query sequence and the matched structure.
Redundancy reducing PDB

SEQRES - Coordinate matching
eexample from thrombin structure 1aix, chain H, with insertion codes:

1–22 : 16–37
23 : 37A
24–46 : 38–60
47 : 60A
48 : 60B
49 : 60C
50 : 60D
51 : 60E
52 : 60F
53 : 60G
54 : 60H
55 : 60I
56–72 : 61–77
[...]

SEQRES:ATOM
## PSSH2 table

<table>
<thead>
<tr>
<th>Protein_sequence_hash</th>
<th>Sequence md5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB_chain_hash</td>
<td>SEQRES md5</td>
</tr>
<tr>
<td>Repeat_domains</td>
<td>Same pair match counter</td>
</tr>
<tr>
<td>E_value</td>
<td>Hhblits E-value</td>
</tr>
<tr>
<td>Identity_Score</td>
<td>% sequence identity</td>
</tr>
<tr>
<td>Alignment</td>
<td>Concise alignment</td>
</tr>
</tbody>
</table>

### Alignment format:
sequenceRange:structureRange

**THRBS_HUMAN** vs **thrombin**

- 364–622:1–259

**THRBS_HUMAN** vs. **plasminogen**

- 106–163:356–413
- 165–246:414–495
- 247–295:497–545
- 334–344:546–556
- 359–384:557–582
Cool features
Aquaria: GLUT-4, 68% Sequence Identity to PDB 4ppy, chain A

**3D STRUCTURE** GLUT-4 sequence aligned onto SLC2A1 structure from PDB 4ppy-A (68% sequence identity)

**FUNCTION:** Insulin-regulated facilitative glucose transporter.

**SUBUNIT:** Interacts with NDUA9 (by similarity). Binds to DAXX. Interacts via... [+]

**SUBCELLULAR LOCATION:** Cell membrane; Multi-pass membrane protein (by similarity)... [+]

**TISSUE SPECIFICITY:** Skeletal and cardiac muscles; brown and white fat.

**PTM:** SUMOylated.

**DISEASE:** Diabetes mellitus, non-insulin-dependent (NIDDM)

**关于 GLUT-4**

**COLORING**
- Secondary Structure
- Identical
- Conserved
- Not conserved
- Insertion

**ABOUT PDB 4ppy**

Mapping the conformational space accessible to catechol-O-methyltransferase.


Abstract: The glucose transporter GLUT1 catalyzes facilitative diffusion of glucose into erythrocytes and is responsible for glucose supply to the brain and other organs. Dysfunctional mutations may lead to GLUT1 deficiency syndrome, whereas overexpression of GLUT1 is a prognostic indicator for cancer. Despite decades of investigation, the structure of GLUT1 remains unknown. Here we report the crystal structure of human GLUT1 at 3.2 Å resolution. The full-length protein, which has a canonical major facilitator superfamily fold, is captured in an inward-open conformation. This structure allows... [+]

Determined by: X-ray diffraction at 3.17 Å resolution

Chain A: SLC2A1 (Solute carrier family 2, facilitated glucose transporter member 1)
Mapping features

Mapping the conformational space accessible to catechol-O-methyltransferase.


Abstract: The glucose transporter GLUT1 catalyses facilitative diffusion of glucose into erythrocytes and is responsible for glucose supply to the brain and other organs. Dysfunctional mutations may lead to GLUT1 deficiency syndrome, whereas over-expression of GLUT1 is a prognostic indicator for cancer. Despite decades of investigation, the structure of GLUT1 remains unknown. Here we report the crystal structure of human GLUT1 at 3.2 Å resolution. The full-length protein, which has a canonical major facilitator superfamily fold, is captured in an inward-open conformation. This structure allows... (+) Determined by X-ray diffraction at 3.17 Å resolution

Chain: SLC2A1 (Solute carrier family 2, facilitated glucose transporter member 1)

Organism: Homo sapiens
User defined feature sets

SNAP2 – Predicted Effect of Point Mutations

Extra annotation in Aquaria

http://aquaria.ws/P14672/4pyp/A?features=... seeFeatureAPI
The Aquaria Team

Seán I. O’Donoghue¹,²,³, Kenneth S. Sabir²,³, Maria Kalemanov⁴, Christian Stolte¹, Neil Saunders¹, Bill Wilson¹, Benjamin Wellmann⁴, Vivian Ho¹,⁶, Manfred Roos⁴, Nelson Perdigão⁵, Fabian A. Buske²,⁶, Julian Heinrich¹, Burkhard Rost⁴, Andrea Schafferhans⁴


http://aquaria.ws