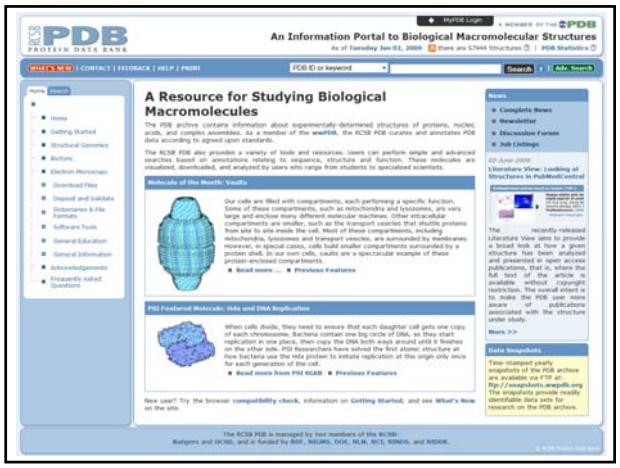
Computergestützte Strukturbiologie
(Strukturelle Bioinformatik)

The Protein Data Bank (PDB)

Sommersemester 2009

Peter Güntert



PDB Current Holdings Breakdown

		Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
Exp. Method	X-ray	46383	1147	2141	17	49688
	NMR	6864	856	146	6	7872
	Electron Microscopy	168	16	59	0	243
	Hybrid	13	1	1	1	16
	Other	108	4	4	9	125
	Total	53536	2024	2351	33	57944

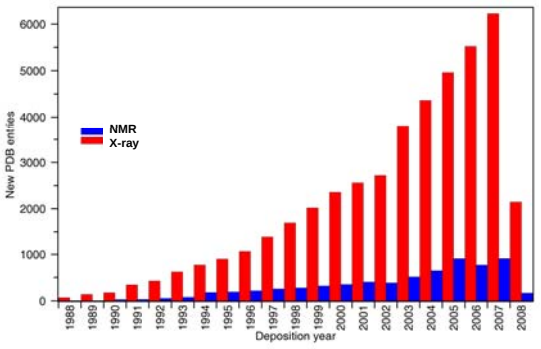
(Click on any number to retrieve the results from that category.)

Please note that theoretical models have been removed, effective July 02, 2002, as per **PDB policy**.

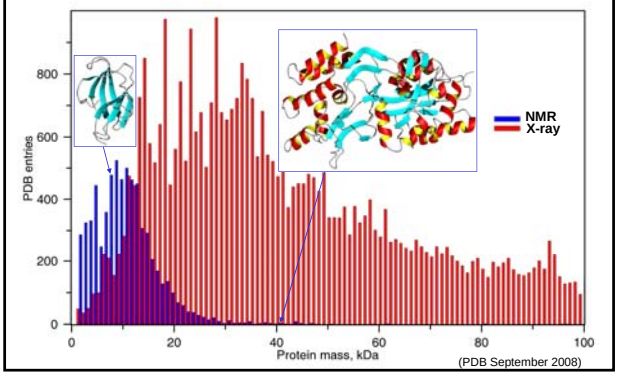
38858 structures in the PDB have a structure factor file.
5156 structures in the PDB have an NMR restraint file.

June 2, 2009

Number of released X-ray and NMR structures in the PDB (September 2008)

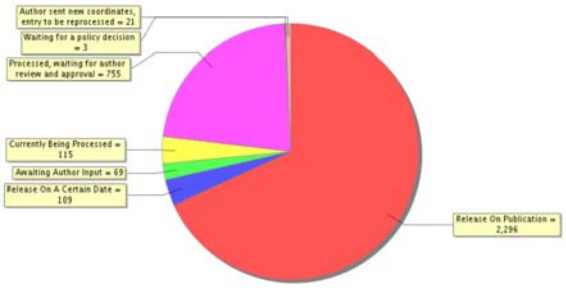


Protein structures in the PDB

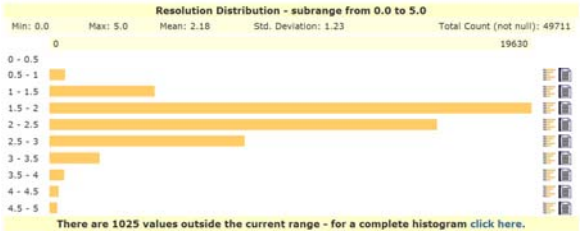


(PDB September 2008)

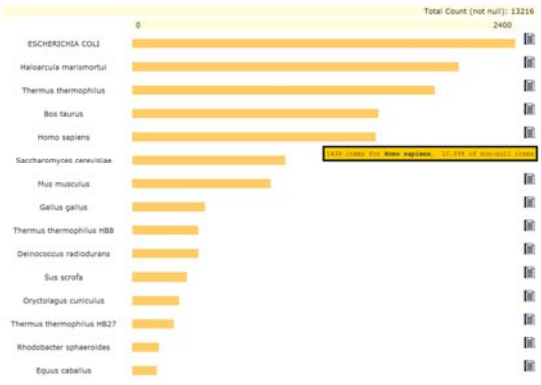
PDB: Status of unreleased entries



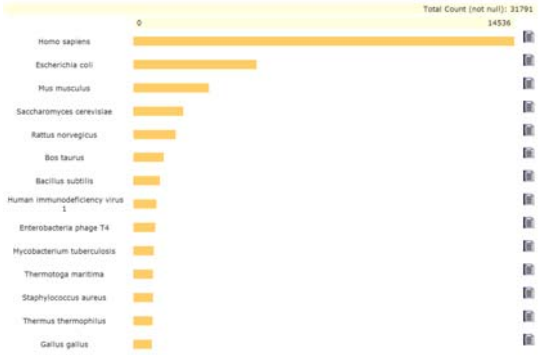
PDB: Resolution distribution



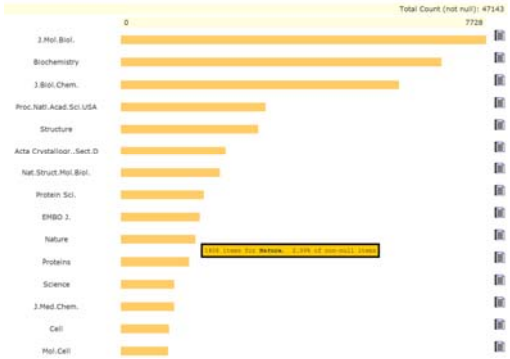
Distribution by natural source organism



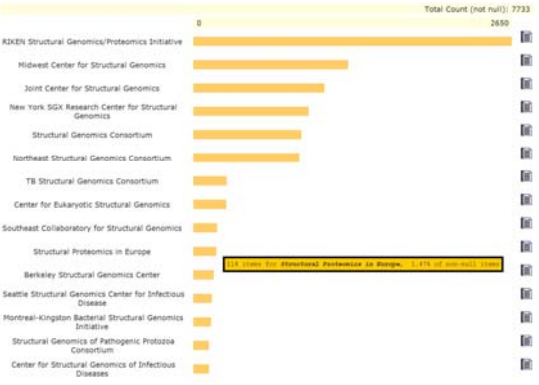
Distribution by gene source organism



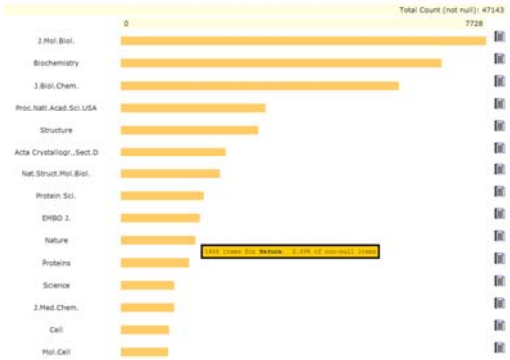
Distribution by primary cited journal



Distribution by structural genomics center

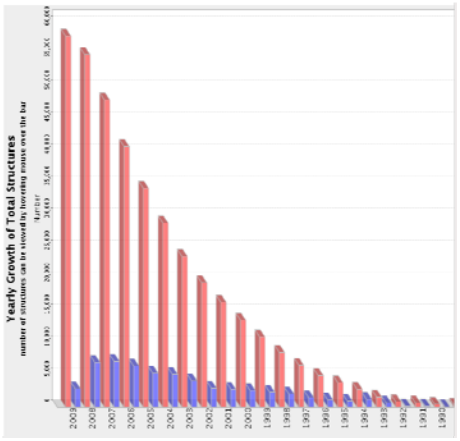


Distribution by primary cited journal

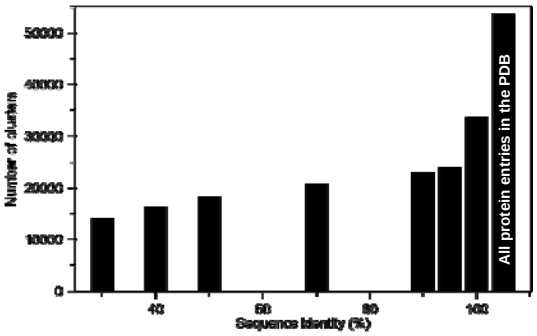


PDB

total structures

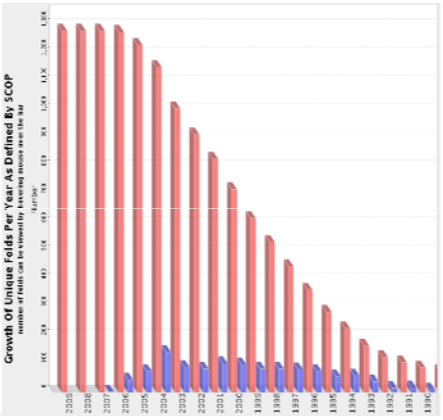


PDB redundancy



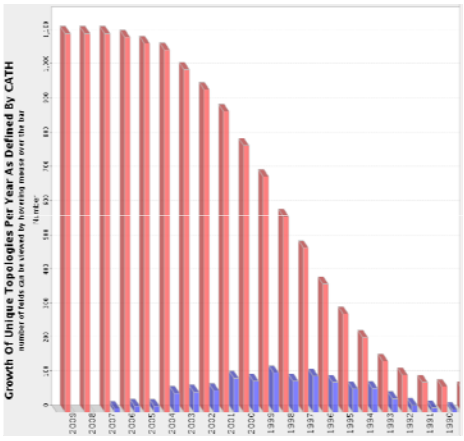
PDB

unique folds as defined by SCOP



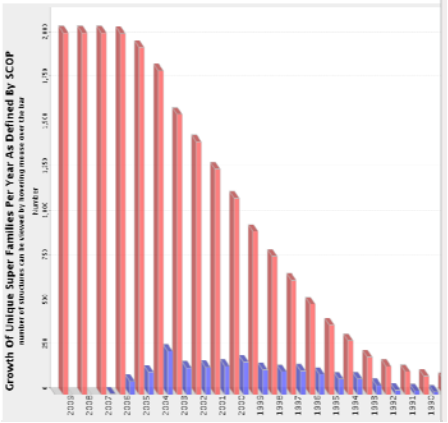
PDB

unique topologies as defined by CATH



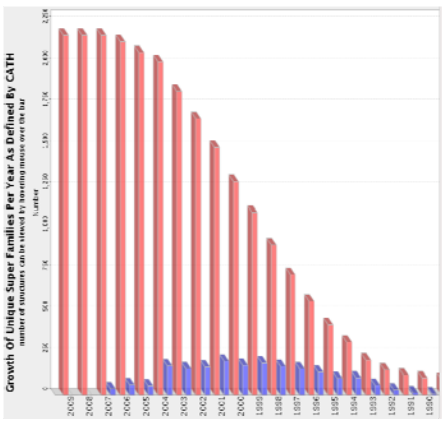
PDB

unique superfamilies as defined by SCOP

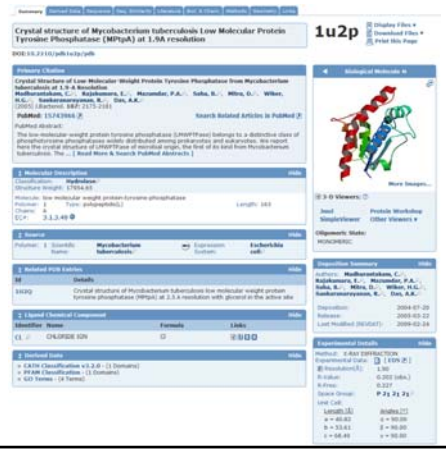


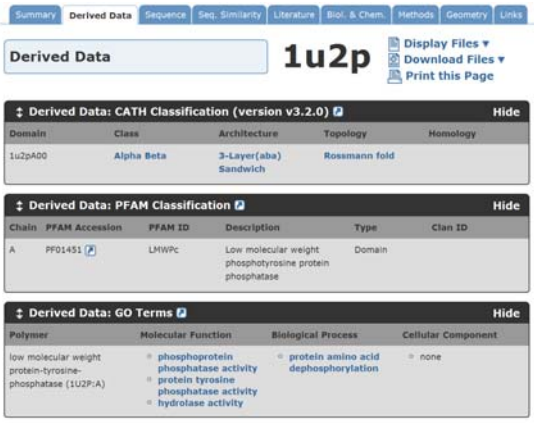
PDB

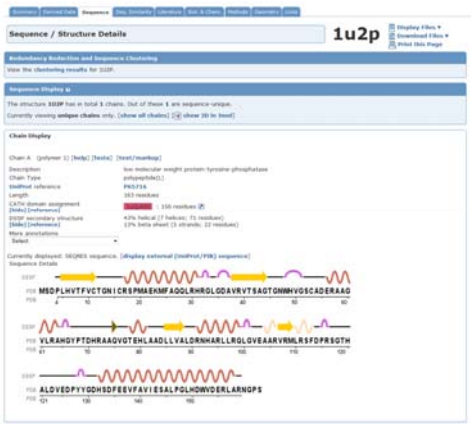
unique superfamilies as defined by CATH



PDB entry 1U2P







PDB entry 1U2P: Header, source

```
HEADER    HYDROLASE                               20-JUL-04   1U2P
TITLE     CRYSTAL STRUCTURE OF MYCOBACTERIUM TUBERCULOSIS LOW
TITLE    2 MOLECULAR PROTEIN TYROSINE PHOSPHATASE (MPTPA) AT 1.9A
TITLE    3 RESOLUTION
COMPND    MOL_ID: 1;
COMPND    2 MOLECULE: LOW MOLECULAR WEIGHT PROTEIN-TYROSINE-
COMPND    3 PHOSPHATASE;
COMPND    4 CHAIN: A;
COMPND    5 SYNONYM: PTPASE;
COMPND    6 EC: 3.1.3.48;
COMPND    7 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE    2 ORGANISM_SCIENTIFIC: MYCOBACTERIUM TUBERCULOSIS;
SOURCE    3 ORGANISM_TAXID: 1773;
SOURCE    4 GENE: MPTPA;
SOURCE    5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE    6 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE    7 EXPRESSION_SYSTEM_STRAIN: SG13009;
SOURCE    8 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE    9 EXPRESSION_SYSTEM_PLASMID: PQE30
KEYWDS    HYDROLASE, TYROSINE PHOSPHATASE, MYCOBACTERIUM
```

PDB entry 1U2P: Authors

```
EXPDTA    X-RAY DIFFRACTION
AUTHOR    C.MADHURANTAKAM,E.RAJAKUMARA,P.A.MAZUMDAR,B.SAHA,D.MITRA,
AUTHOR   2 H.G.WIKER,R.SANKARANARAYANAN,A.K.DAS
REVDAT   2   24-FEB-09 1U2P    1       VERSN
REVDAT   1   22-MAR-05 1U2P    0
JRNL        AUTH   C.MADHURANTAKAM,E.RAJAKUMARA,P.A.MAZUMDAR,B.SAHA,
JRNL        AUTH 2 D.MITRA,H.G.WIKER,R.SANKARANARAYANAN,A.K.DAS
JRNL        TITL   CRYSTAL STRUCTURE OF LOW-MOLECULAR-WEIGHT PROTEIN
JRNL        TITL 2 TYROSINE PHOSPHATASE FROM MYCOBACTERIUM
JRNL        TITL 3 TUBERCULOSIS AT 1.9-A RESOLUTION
JRNL        REF    J.BACTERIOL.                  V. 187  2175 2005
JRNL        REFN                   ISSN 0021-9193
JRNL        PMID   15743966
JRNL        DOI    10.1128/JB.187.6.2175-2181.2005
REMARK   1
REMARK   2
REMARK   2 RESOLUTION.    1.90 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : CNS 1.1
REMARK   3   AUTHORS     : BRUNGER,ADAMS,CLORE,DELANO,GROS,GROSSE-
REMARK   3               : KUNSTLEVE,JIANG,KUSZEWSKI,NILGES, PANNU,
REMARK   3               : READ,RICE,SIMONSON,WARREN
```

PDB entry 1U2P: Refinement

```
REMARK   2 RESOLUTION.    1.90 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : CNS 1.1
REMARK   3   AUTHORS     : BRUNGER,ADAMS,CLORE,DELANO,GROS,GROSSE-
REMARK   3               : KUNSTLEVE,JIANG,KUSZEWSKI,NILGES, PANNU,
REMARK   3               : READ,RICE,SIMONSON,WARREN
REMARK   3
REMARK   3  REFINEMENT TARGET : ENGH & HUBER
REMARK   3
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 1.90
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 24.96
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   DATA CUTOFF HIGH         (ABS(F)) : 1161871.740
REMARK   3   DATA CUTOFF LOW          (ABS(F)) : 0.0000
REMARK   3   COMPLETENESS (WORKING+TEST)   (%) : 99.6
REMARK   3   NUMBER OF REFLECTIONS             : 12309
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD           : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION   : RANDOM
REMARK   3   R VALUE            (WORKING SET) : 0.202
REMARK   3   FREE R VALUE                      : 0.227
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : 5.000
REMARK   3   FREE R VALUE TEST SET COUNT       : 616
REMARK   3   ESTIMATED ERROR OF FREE R VALUE   : 0.009
```

PDB entry 1U2P: Missing residues

```
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
REMARK 465     MET A     1
REMARK 465     SER A     2
REMARK 465     ASP A     3
REMARK 465     ASN A   160
REMARK 465     GLY A   161
REMARK 465     PRO A   162
REMARK 465     SER A   163
```

PDB entry 1U2P: Ramachandran plot outliers

```
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: TORSION ANGLES
REMARK 500
REMARK 500 TORSION ANGLES OUTSIDE THE EXPECTED RAMACHANDRAN REGIONS:
REMARK 500 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN IDENTIFIER;
REMARK 500 SSEQ=SEQUENCE NUMBER; I=INSERTION CODE).
REMARK 500
REMARK 500 STANDARD TABLE:
REMARK 500 FORMAT:(10X,I3,1X,A3,1X,A1,I4,A1,4X,F7.2,3X,F7.2)
REMARK 500
REMARK 500 EXPECTED VALUES: GJ KLEYWEGT AND TA JONES (1996). PHI/PSI-
REMARK 500 CHOLOGY: RAMACHANDRAN REVISITED. STRUCTURE 4, 1395 - 1400
REMARK 500
REMARK 500  M RES CSSEQI        PSI       PHI
REMARK 500    CYS A  16      -83.04   -122.74
```

PDB entry 1U2P: Sequence

```
DBREF  1U2P A    1   163  UNP    P65716   PTPA_MYCTU       1    163
SEQRES   1 A  163  MET SER ASP PRO LEU HIS VAL THR PHE VAL CYS THR GLY
SEQRES   2 A  163  ASN ILE CYS ARG SER PRO MET ALA GLU LYS MET PHE ALA
SEQRES   3 A  163  GLN GLN LEU ARG HIS ARG GLY LEU GLY ASP ALA VAL ARG
SEQRES   4 A  163  VAL THR SER ALA GLY THR GLY ASN TRP HIS VAL GLY SER
SEQRES   5 A  163  CYS ALA ASP GLU ARG ALA ALA GLY VAL LEU ARG ALA HIS
SEQRES   6 A  163  GLY TYR PRO THR ASP HIS ARG ALA ALA GLN VAL GLY THR
SEQRES   7 A  163  GLU HIS LEU ALA ALA ASP LEU LEU VAL ALA LEU ASP ARG
SEQRES   8 A  163  ASN HIS ALA ARG LEU LEU ARG GLN LEU GLY VAL GLU ALA
SEQRES   9 A  163  ALA ARG VAL ARG MET LEU ARG SER PHE ASP PRO ARG SER
SEQRES  10 A  163  GLY THR HIS ALA LEU ASP VAL GLU ASP PRO TYR TYR GLY
SEQRES  11 A  163  ASP HIS SER ASP PHE GLU GLU VAL PHE ALA VAL ILE GLU
SEQRES  12 A  163  SER ALA LEU PRO GLY LEU HIS ASP TRP VAL ASP GLU ARG
SEQRES  13 A  163  LEU ALA ARG ASN GLY PRO SER
HET     CL  A 164       1
HETNAM      CL CHLORIDE ION
FORMUL   2   CL    CL 1-
FORMUL   3  HOH   *152(H2 O)
```

PDB entry 1U2P: Secondary structure

```
HELIX    1   1 CYS A   16  ARG A   32  1                           17
HELIX    2   2 ASP A   55  HIS A   65  1                           11
HELIX    3   3 GLY A   77  ALA A   82  1                            6
HELIX    4   4 ASP A   90  LEU A  100  1                           11
HELIX    5   5 GLU A  103  ALA A  105  5                            3
HELIX    6   6 ARG A  111  ASP A  114  5                            4
HELIX    7   7 ASP A  131  ARG A  159  1                           29
SHEET    1   A 4 VAL A  38  GLY A  44  0
SHEET    2   A 4 LEU A   5  CYS A  11  1  N  LEU A   5   O  ARG A  39
SHEET    3   A 4 LEU A  85  ALA A  88  1  O  VAL A  87   N  THR A   8
SHEET    4   A 4 VAL A 107  MET A 109  1  O  ARG A 108   N  LEU A  86
SITE     1 AC1  4 THR A  12  GLY A  13  ARG A  17  HOH A 171
```

PDB entry 1U2P: Crystal data

```
CRYST1   40.816   53.610   68.486  90.00  90.00  90.00 P 21 21 21    4
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.024500  0.000000  0.000000        0.00000
SCALE2      0.000000  0.018653  0.000000        0.00000
SCALE3      0.000000  0.000000  0.014602        0.00000
```

PDB entry 1U2P: Coordinates

```
ATOM      1  N   PRO A   4       6.719 -12.134  26.603  1.00 18.91     N
ATOM      2  CA  PRO A   4       6.735 -10.746  27.122  1.00 18.45     C
ATOM      3  C   PRO A   4       6.209  -9.735  26.108  1.00 16.72     C
ATOM      4  O   PRO A   4       6.701  -9.658  24.983  1.00 16.64     O
ATOM      5  CB  PRO A   4       8.174 -10.427  27.495  1.00 20.82     C
ATOM      6  CG  PRO A   4       8.942 -11.387  26.584  1.00 20.17     C
ATOM      7  CD  PRO A   4       8.093 -12.664  26.557  1.00 22.00     C
ATOM      8  N   LEU A   5       5.207  -8.963  26.521  1.00 16.15     N
ATOM      9  CA  LEU A   5       4.605  -7.937  25.674  1.00 14.51     C
ATOM     10  C   LEU A   5       5.700  -6.960  25.244  1.00 14.38     C
ATOM     11  O   LEU A   5       6.564  -6.600  26.042  1.00 15.34     O
ATOM     12  CB  LEU A   5       3.513  -7.204  26.458  1.00 13.81     C
ATOM     13  CG  LEU A   5       2.639  -6.180  25.737  1.00 14.69     C
ATOM     14  CD1 LEU A   5       1.815  -6.864  24.656  1.00 15.29     C
ATOM     15  CD2 LEU A   5       1.725  -5.506  26.754  1.00 15.24     C
```

- Keyword
- Atom index
- Atom name
- Residue name
- Chain identifier
- Residue number
- *x* coordinate
- *y* coordinate
- *z* coordinate
- Occupancy
- *B*-factor
- Element

Secondary structure assignment: DSSP algorithm

Kabsch, W., Sander, C. *Biopolymers* 22, 2577–2637 (1983)

The definitions of H-bonded features form a hierarchy:

1. H-bonds are defined.
2. Based on them, turns and bridges.
3. Based on them, α-helices and β-ladders, including common imperfections such as helical kinks and β-bulges.

Each structural feature is defined independently of the others and structural overlaps are resolved by defining a secondary structure summary that assigns a single state to each residue.

DSSP-Algorithm: H-bonds

Hydrogen bonds in proteins have little wave-function overlap and are well described by an electrostatic model. We calculate the electrostatic interaction energy between two H-bonding groups by placing partial charges on the C,O *($+q_1$, $-q_1$) and N,H ($-q_2$,$+q_2$) atoms, i.e.,*

$$E = q_1 q_2 (1/r(\mathrm{ON}) + 1/r(\mathrm{CH}) - 1/r(\mathrm{OH}) - 1/r(\mathrm{CN}))*f$$

with $q_1 = 0.42e$ *and* $q_2 = 0.20e$, e being the unit electron charge and r(AB) the interatomic distance from A to B. In chemical units, r is in Å, the dimensional factor f = 332, and E is in kcal/mol. A good H bond has about -3 kcal/mol binding energy. We choose a generous cutoff to allow for bifurcated H bonds and errors in coordinates and assign an H bond between C=O of residue i and N-H of residue j if E is less than the cutoff, i.e.,

"Hbond(ij)=: [E < -0.5kcal/mole]."

Elementary H-Bond Pattern: *n*-Turn

The basic turn pattern is a single H-bond of type *(i, i + n). We* assign an *n*-turn at residue *i* if there is an H bond from CO(*i*) to NH(*i + n),* i.e.,

"*n*-turn(*i*)=: Hbond(*i*,*i + n*), *n* = 3,4,5."

When the pattern is found, the ends of the H bond are indicated by using ">" at *i* and "<" at *i + n*; the residues bracketed by the H-bond are noted "3," "4," or "5" unless they are also the end points of other H-bonds. Coincidence of ">" and "<" at one residue is indicated by "X."

Elementary H-Bond Pattern: *Bridge*

Two nonoverlapping stretches of 3 residues each, *i* - 1, *i*, *i* + 1 and j - 1, *j*, *j* + 1, form either a parallel or antiparallel bridge, depending on which of two basic patterns is matched. We assign a bridge between residues *i* and *j* if there are two H-bonds characteristic of β-structure; in particular,

Parallel Bridge(i,j) =: [Hbond(*i* - 1,*j*) and Hbond(*j*,*i* + 1)] or
[Hbond(*j* - 1,*i*) and Hbond(*i*,*j* + 1)]
Antiparallel Bridge(i,j) =: [Hbond(*i*,*j*) and Hbond(*j*,*i*)] or
[Hbond(*i* - 1,*j* + 1) and Hbond(*j* - 1,i + 1)]

Parallel bridges are marked at *i* and *j* by lower-case letters, antiparallel ones by upper-case letters.

DSSP Helix

Cooperative H-Bond Pattern: Helices

A minimal helix is defined by two consecutive *n*-turns. For example, a 4-helix, of minimal length 4 from residues *i* to *i* + 3, requires 4-turns at residues *i* − 1 and *i*,

4-helix(i,i + 3)=: [4-turn(i − 1) and 4-turn(i)]

i.e., an H bond (*i* − 1,*i* +3) and an H bond (*i*,*i* + 4). Note that nothing is required about the H-bond state of residues *i* + 1 and *i* + 2. Similarly, two consecutive turns are required and a 3-helix of minimal length 3 from residue *i* to *i* + 2 and a 5-helix of minimal length 5 from residue *i* to *i* + 5:

3-helix(i,i + 2)=: [3-turn(i − 1) and 3-turn(i)]
5-helix(i,i + 5)=: [5-turn(i − 1) and 5-turn(i)]

Longer helices are defined as overlaps of minimal helices. Conventionally, these structures are called α-helix, 3_{10}-helix, and π-helix. In Table AIII, a 3-helix can be recognized by the pattern ⟩⟩3⟨⟨, a 4-helix by ⟩⟩44⟨⟨, and a 5-helix by ⟩⟩555⟨⟨. In the line SUMMARY, the residues bracketed by H bonds are labeled G, H, I, e.g.,

5-TURN			⟩⟩555⟨⟨
4-TURN		⟩⟩44⟨⟨	
3-TURN	⟩⟩3⟨⟨		
SUMMARY	GGG	HHHH	IIIII

DSSP Sheet

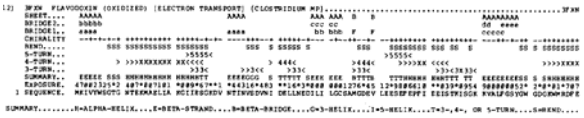
Cooperative H-Bond Patterns: β-Ladders and β-Sheets

We coin the term "ladder" and define

ladder=: set of one or more consecutive bridges of identical type
sheet=: set of one or more ladders connected by shared residues

Ladders are given letter names, where a,b,c, . . . is for parallel, A,B,C . . . for antiparallel arrangement. Along the sequence, the first ladder is named "a" or "A," the second "b" or "B," etc. Sheets are also given letter names A,B,C . . . When the alphabet is exhausted, names restart at "a" or "A." In Table AIII, each residue is labeled in line SHEET by the sheet name and in lines BRIDGE by the names of the ladders in which it participates (at most two, one on each side). In line SUMMARY, residues in single bridges (ladders of length 1) are marked "B," all other ladder residues "E" (extended). Thus, continuous stretches of "E" are β-strands. The β-sheet notation is illustrated in Fig. 4.

DSSP result for Flavodoxin



Literatur

- Anna Tramontano: *Protein Structure Prediction, Wiley-VCH*, 2006.
- Tamar Schlick: *Molecular Modeling and Simulation*, Springer, 2006.