FACULTY OF SCIENCE
Charles University

Klara Hlouchova
Research Group

Synthetic biology | Faculty of Science, Charles University

Carolina Rocha

Klara Hlouchova Research group

Department of Cell Biology
Faculty of Science
Charles University

Introduction to web tools for protein structure classification

Basic concepts 20 min

- Protein structure and classification

- Intrinsically disordered proteins and regions

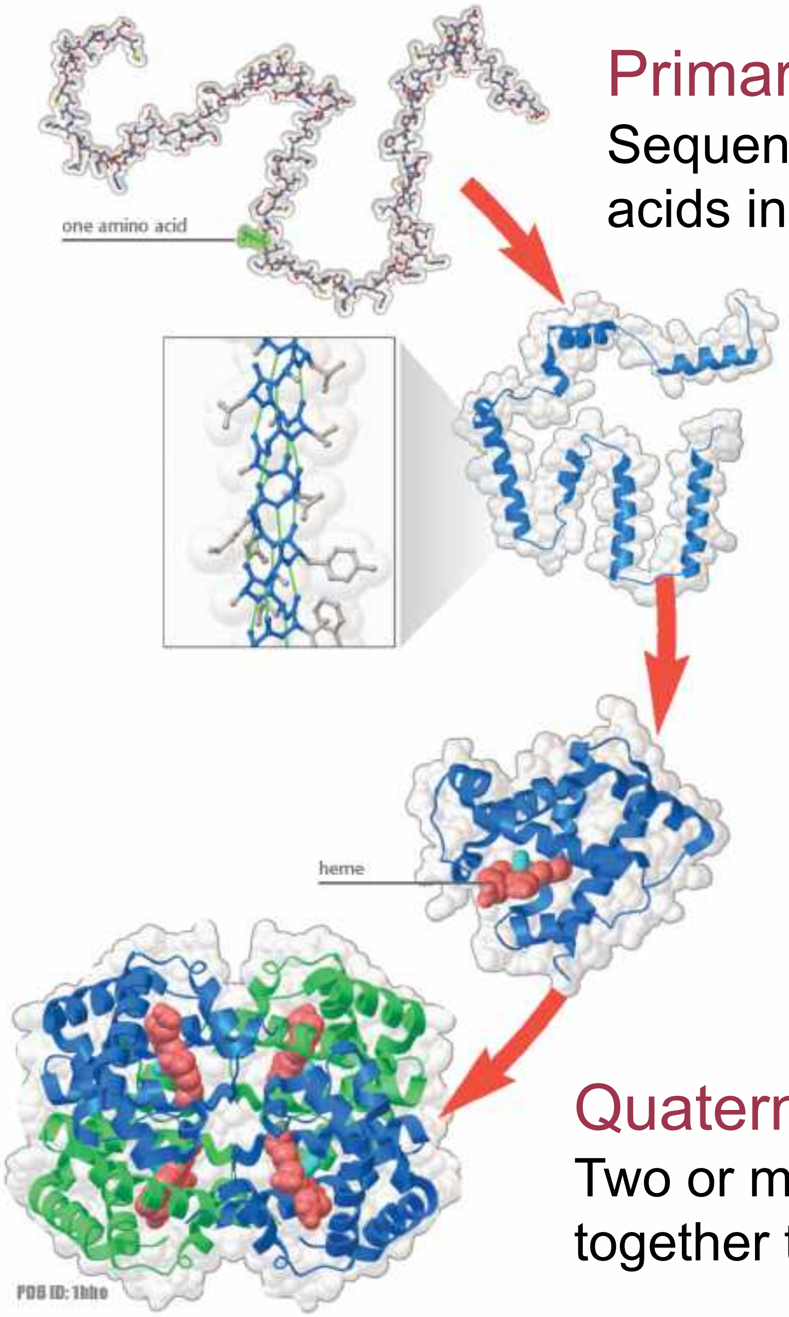
- Protein Data Bank (PDB)

Structural analysis 20 min

- Protein Data Bank (PDB) exploration

-Retrieving structural data

What is a protein?



Primary structure

Sequential arrangement of proteinogenic amino acids in a polypeptide chain.

Secondary structure

Local spatial arrangement of the polypeptide backbone.

Hydrogen bonds between amino acids form two stable structural elements:

- Alpha helices
- Beta strands

Tertiary structure

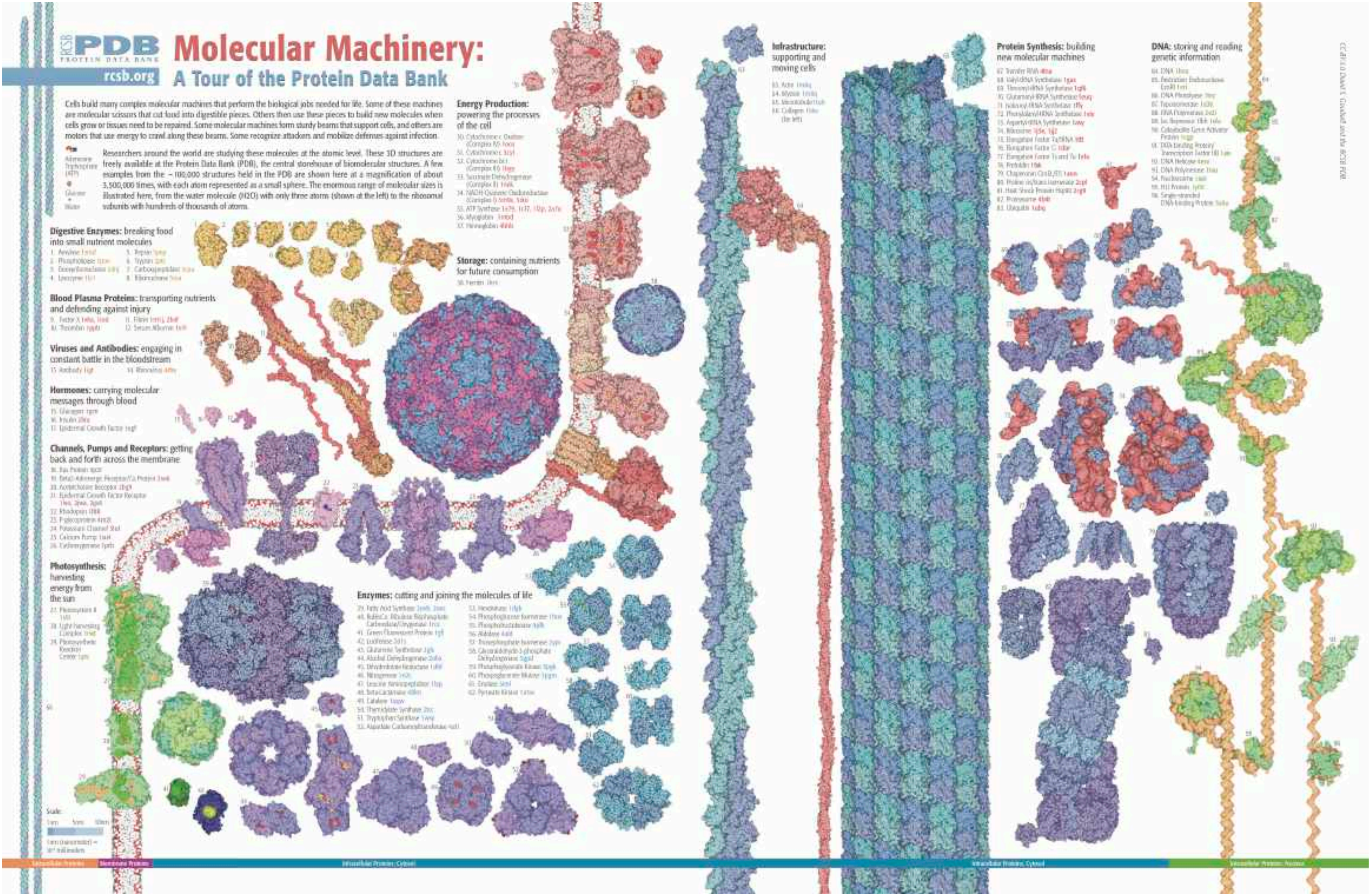
Overall spatial arrangement of atoms in a protein. Folding of a polypeptide chain.

Quaternary structure

Two or more polypeptide chains can come together to form one functional molecule.

Berman H, Westbrook J, Feng Z, Gilliland T, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. Nucleic Acids Research. 2000. 28: 235-242. doi:10.1093/nar/28.1.235.
PDB101.rcsb.org.

How does proteins look like?



Molecular Machinery: A Tour of the Protein Data Bank

Cells build many complex molecular machines that perform the biological jobs needed for life. Some of these machines are molecular scissors that cut food into digestible pieces. Others then use these pieces to build new molecules when cells grow or tissues need to be repaired. Some molecular machines form sturdy beams that support cells, and others are motors that use energy to crawl along these beams. Some recognize attackers and mobilize defenses against infection.

Researchers around the world are studying these molecules at the atomic level. These 3D structures are freely available at the Protein Data Bank (PDB), the central storehouse of biomolecular structures. A few examples from the ~100,000 structures held in the PDB are shown here at a magnification of about 3,500,000 times, with each atom represented as a small sphere. The enormous range of molecular sizes is illustrated here, from the water molecule (H2O) with only three atoms (shown at the left) to the ribosomal subunits with hundreds of thousands of atoms.

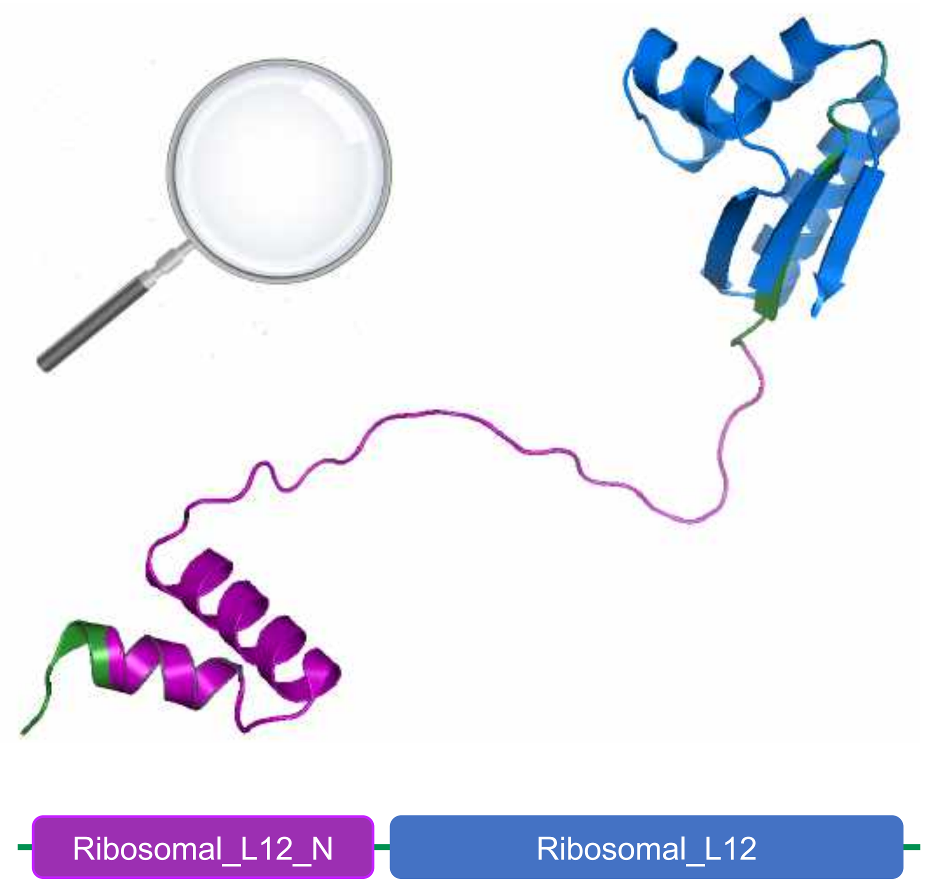
PDB
PROTEIN DATA BANK
rcsb.org
Adenosine Triphosphate (ATP)
Glucose
Water
Digestive Enzymes: breaking food into small nutrient molecules
Blood Plasma Proteins: transporting nutrients and defending against injury
Viruses and Antibodies: engaging in constant battle in the bloodstream
Hormones: carrying molecular messages through blood
Channels, Pumps and Receptors: getting back and forth across the membrane
Photosynthesis: harvesting energy from the sun
Energy Production: powering the processes of the cell
Storage: containing nutrients for future consumption
Enzymes: cutting and joining the molecules of life
Infrastructure: supporting and moving cells
Protein Synthesis: building new molecular machines
DNA: storing and reading genetic information
Scale:

Berman H, Westbrook J, Feng Z, Gilliland T, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. Nucleic Acids Research. 2000. 28: 235-242. doi:10.1093/nar/28.1.235. PDB101.rcsb.org.

Protein domains

Proteins are composed of domains, autonomous structural, functional and/or evolutionary units in a protein, therefore they can acquire a fold and function on its own.

Tompa P. Structure and Function of Intrinsically Disordered Proteins. CRC Press, Boca Raton. 2010. pp:10-11.



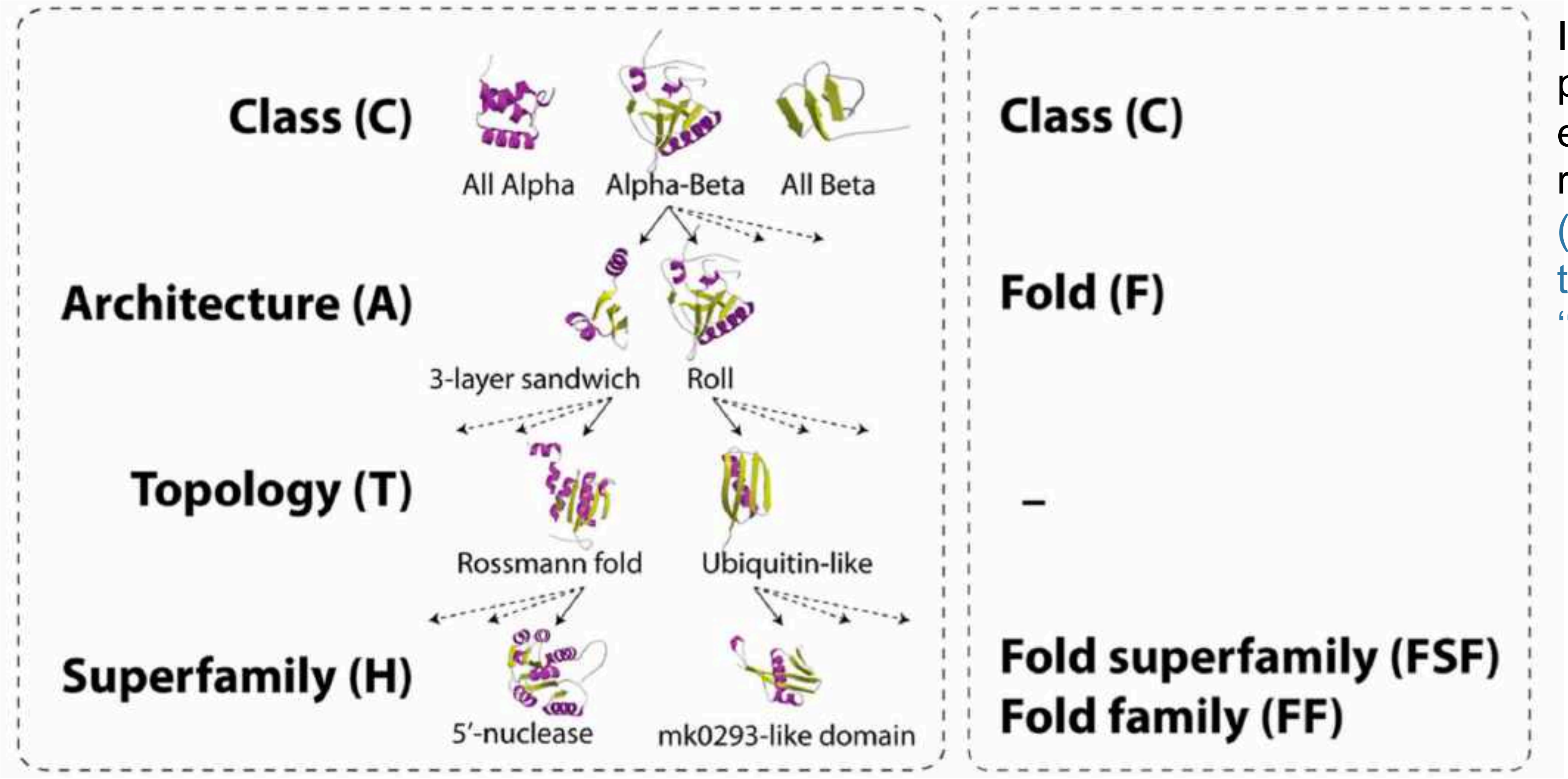
PDB ID: 1RQU. Ribosomal protein L7 from *Escherichia coli*.
Bocharov E, Sobol A, Pavlov K, Korzhnev D, Jaravine V, Gudkov A, Arseniev A. J Biol Chem. 2004. 279: 17697-17706. DOI: 10.1074/jbc.M313384200

Hierarchical classification of protein domain structures

CATH

SCOP

ECOD



doi: 10.1093/nar/gky1097

doi.org/10.1093/nar/gkz1064

It groups domains primarily by evolutionary relationships (homology), rather than topology (or "fold").

doi.org/10.1371/journal.pcbi.1003926

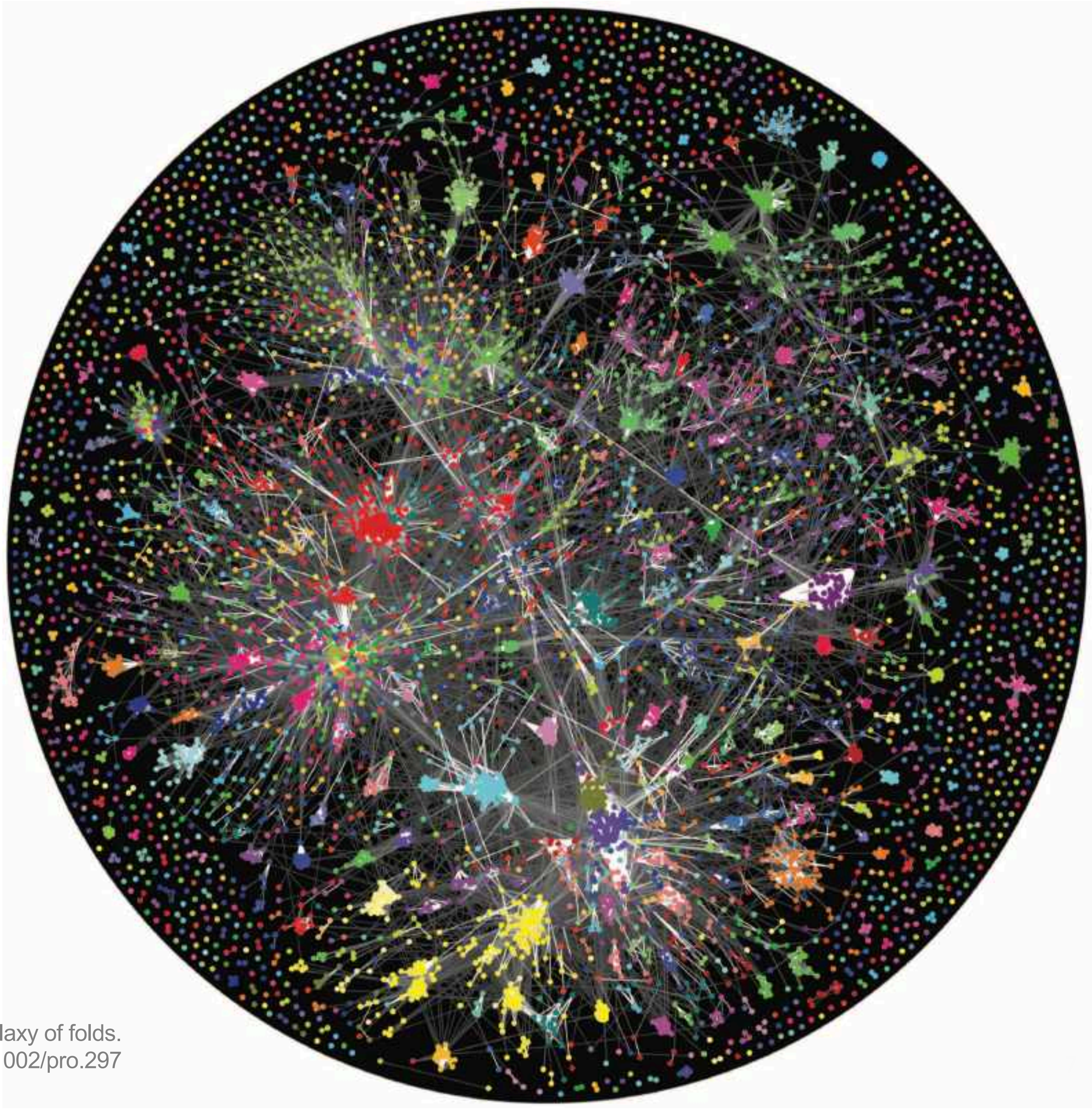
Bukhari S, Caetano-Anollés G. Origin and Evolution of Protein Fold Designs Inferred from Phylogenomic Analysis of CATH Domain Structures in Proteomes. PLoS Computational Biology. 2013. : doi:10.1371/journal.pcbi.1003009.

A Galaxy of folds

It has been described that there are in fact homologous relationships between protein superfamilies that in the past were classified as non-homologous.

…our galaxy of folds summarizes most known and many yet undescribed homologous relationships between protein superfamilies, providing new insights into the evolution of protein domains.

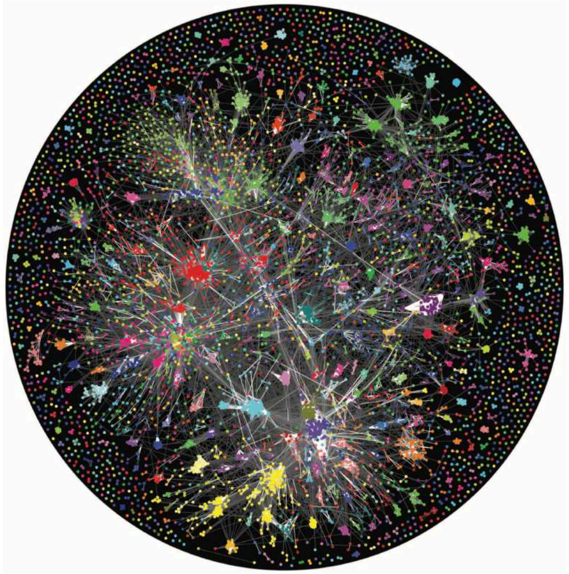
Proteins may not have had as many independent origins as hitherto assumed



Alva V, Remmert M, Biegert A, Lupas A, Söding J. A galaxy of folds.
PROTEIN SCIENCE. 2010. 19:124-130. doi:10.1002/pro.297

The Dark Proteome

Distribution of dark matter, galaxies, and hot gas in the core of the merging galaxy cluster Abell 520.
https://science.nasa.gov/astrophysics/focus-areas/what-is-dark-energy



Alva V, Remmert M, Biegert A, Lupas A, Söding J. A galaxy of folds. PROTEIN SCIENCE. 2010. 19:124-130. doi:10.1002/pro.297

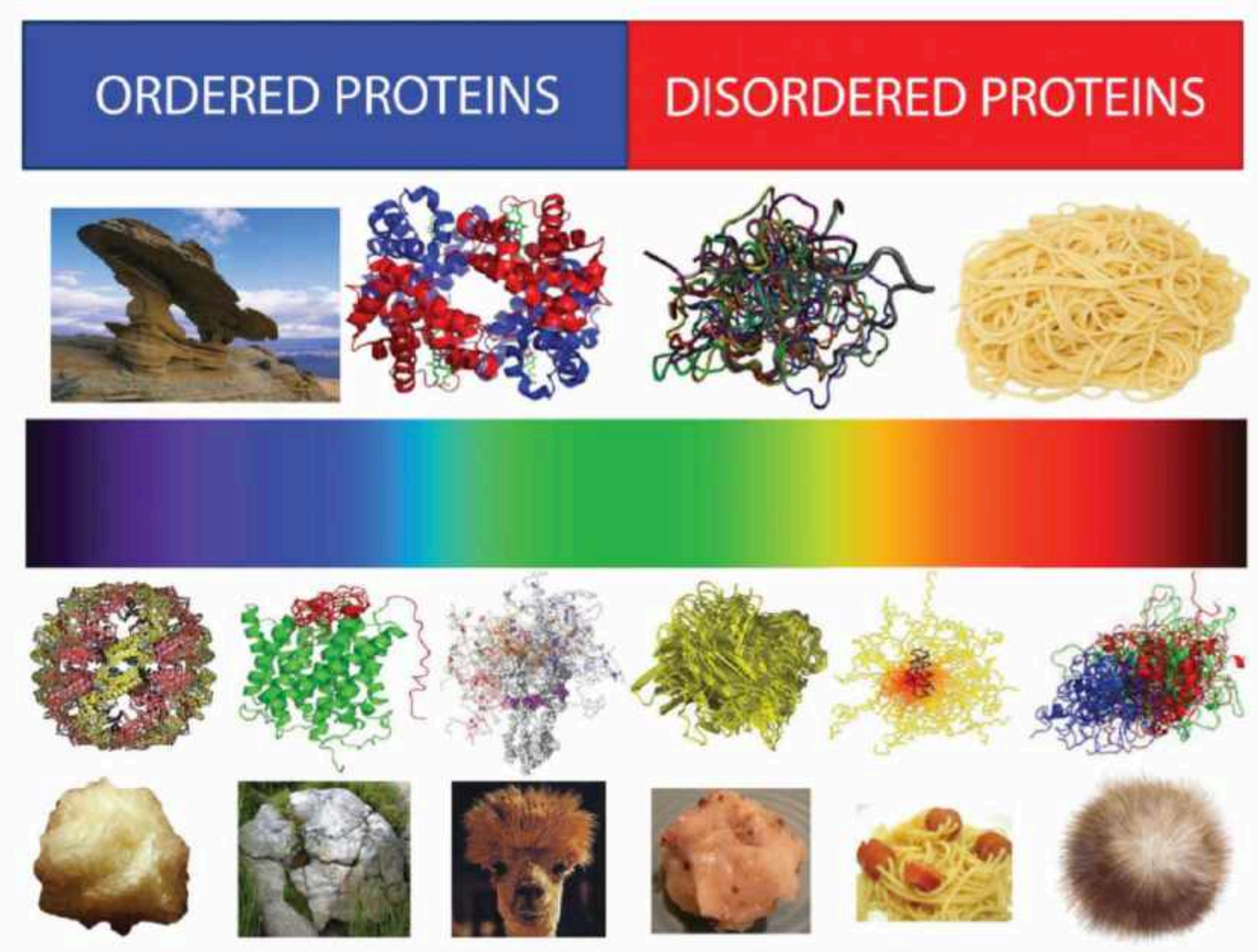
There are some proteins that do not adopt a dominant well-folded structure, and therefore have remained "unseen" by traditional structural biology methods. Those macromolecules conform the Dark Proteome of the protein universe on Earth.

The Dark proteome is conformed by intrinsically disordered proteins and others in which the entire sequence lacked similarity to any known structure.

Asmit Bhowmick, David H. Brookes, Shane R. Yost, H. Jane Dyson, Julie D. Forman-Kay, Daniel Gunter, Martin Head-Gordon, Gregory L. Hura, Vijay S. Pande, David E. Wemmer, Peter E. Wright, and Teresa Head-Gordon. Finding Our Way in the Dark Proteome. Journal of the American Chemical Society. 2016. 138 (31),9730-9742. doi:10.1021/jacs.6b06543

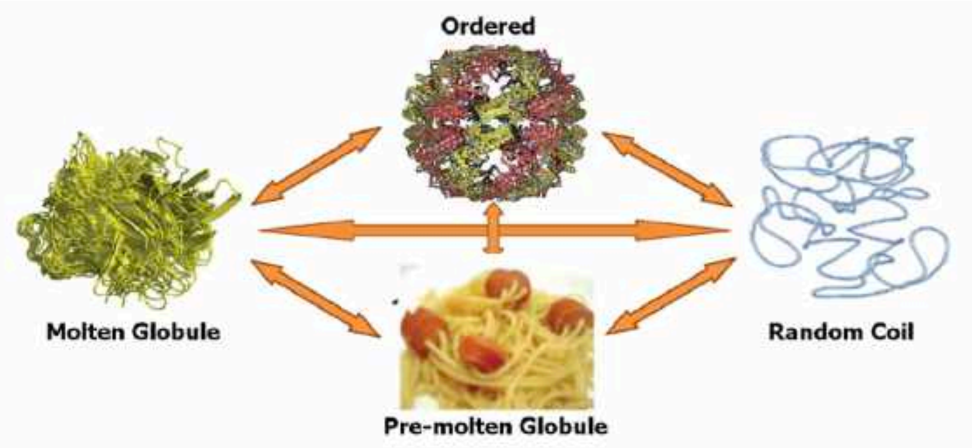
Perdigão N, Heinrich J, Stolte C, Sabir K, Buckley M, Tabor B, Signal B, Gloss B, Hammang B, Rost B, Schafferhans A, O'Donoghuec S. Unexpected features of the dark proteome. PNAS. 2015. 112:15898–15903. doi: 10.1073/pnas.1508380112

What are Intrinsically disordered proteins?



Uversky V. A decade and a half of protein intrinsicdisorder: Biology still waits for physics. PROTEIN SCIENCE .2013. 22:693-724. doi:10.1002/pro.2261

Intrinsically disordered proteins (IDPs) or regions (IDRs) do not have a unique 3-D structure in their functional states. This flexible proteins exist as a heterogeneous, highly dynamic set of conformers.

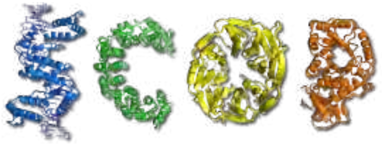


van der Lee et al. Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. 2014. 114: 6589−6631. doi: 10.1021/cr400525m.
Uversky V. Introduction to Intrinsically Disordered Proteins (IDPs). Chem. Rev. 2014. 114:6557−6560. doi.org/10.1021/cr500288y.

Intrinsically disordered proteins databases

PDB ID: 2JU4. *Bos Taurus*. NMR structure of the gamma subunit of cGMP phosphodiesterase,. Song J, Guo L, Muradov H, Artemyev N, Ruoho A Markley J. PNAS. 2018. 105:1505-1510. doi:10.1073/pnas.0709558105

Hierarchical classification of disordered proteins domains*



doi:10.1093/nar/gkz1064

Protein disorder databases

DisProt

doi:10.1093/nar/gkz975

Experimental characterization and the **functionalities** of IDRs and IDPs.

IDEAL

doi: 10.1093/nar/gkt1010

Experimentally verified IDPs. Regions that undergo coupled folding and binding upon interaction with other proteins.

pE-DB

doi:10.1093/nar/gkt960

Deposition of **structural ensembles**.

MobiDB

a database of protein disorder and mobility annotations

doi:10.1093/nar/gkx1071

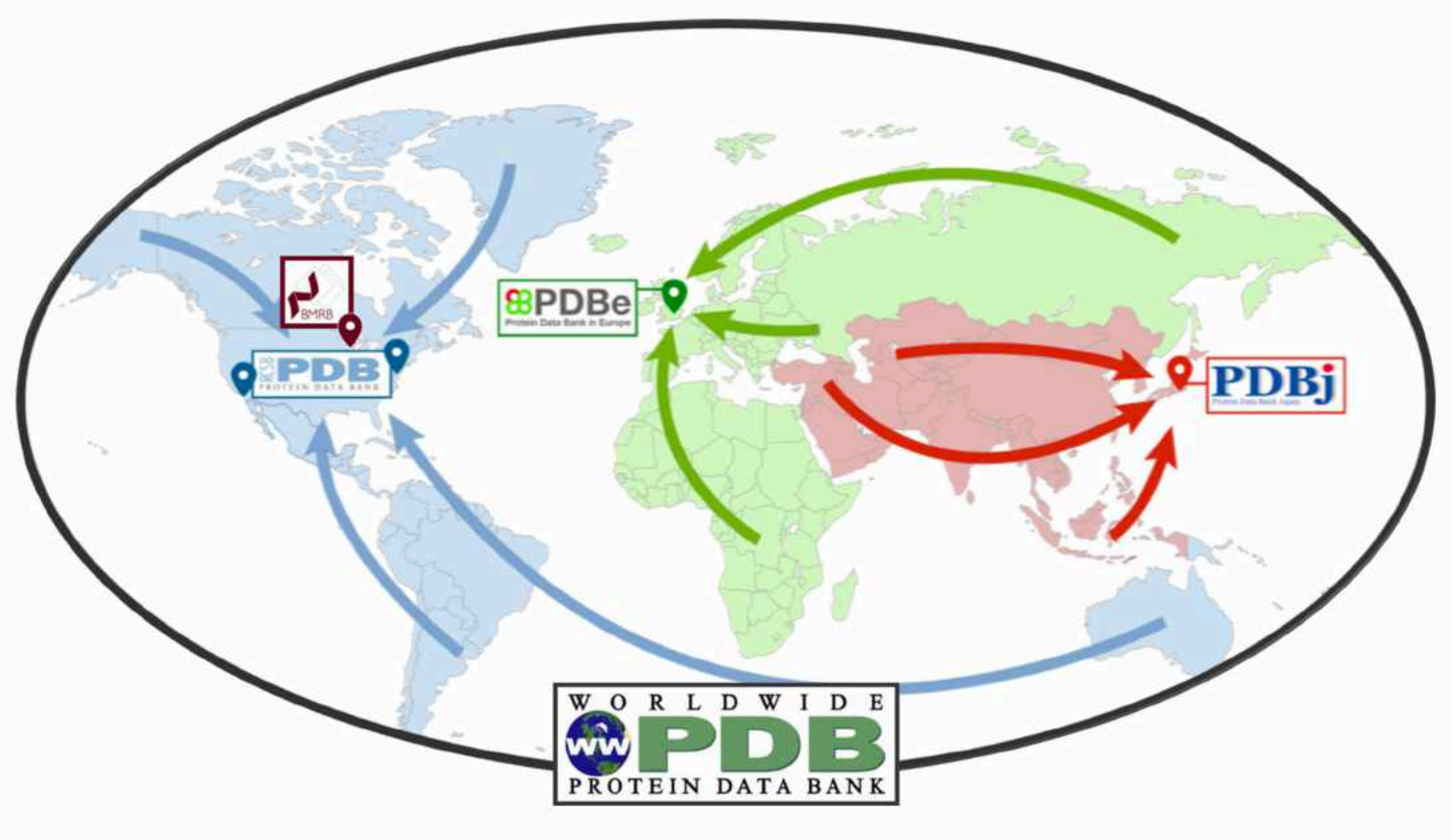
Experimental characterization of IDRs and it also stores **disorder prediction** data from three methods.

D^2P^2

doi:10.1093/nar/gks1226

Stores **disorder predictions** made by nine different predictors.

Protein Data Bank (PDB)



Nucleic Acids Research, 2019, Vol. 47, Database issue

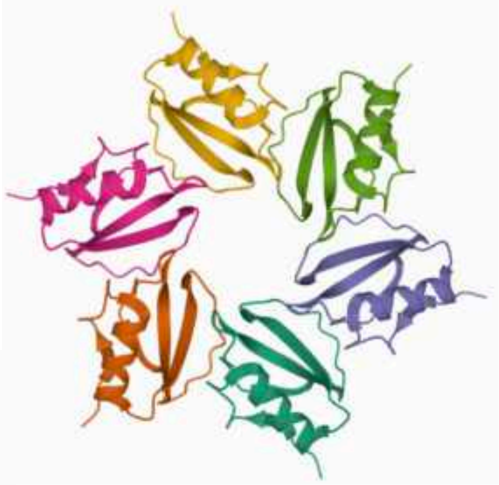
Protein Data Bank: the unique repository of structural data

The PDB was established in 1971 at Brookhaven National Laboratory (USA) under the leadership of Walter Hamilton.

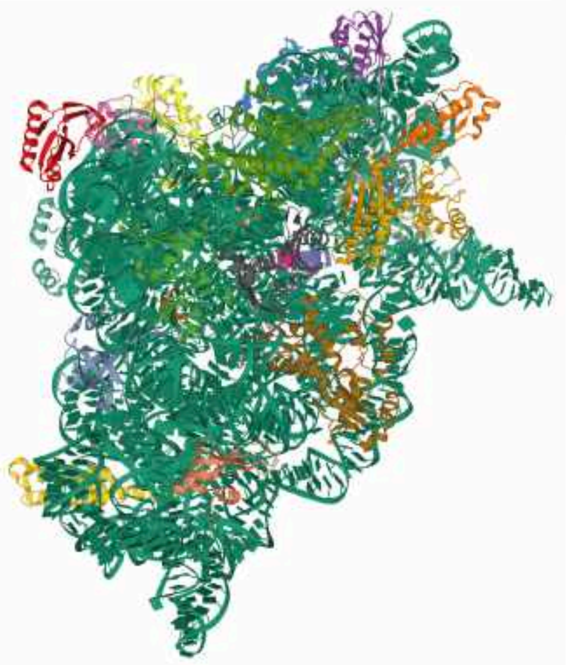
The PDB stores solved biological macromolecules; each had its own information recorded in coordinate files that list the atoms in each structure and their 3D location in space.

WORLDWIDE PDB PROTEIN DATA BANK

Nucleic acids (NA)



Proteins



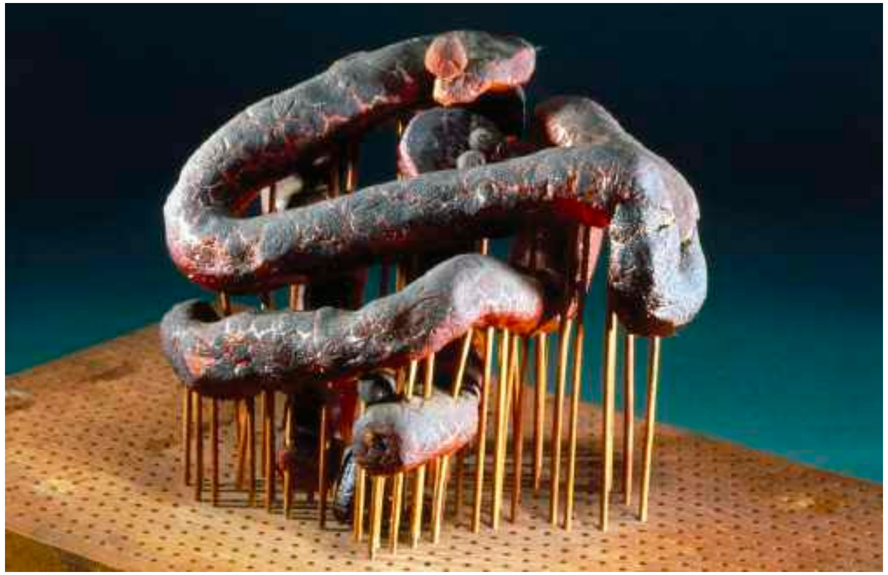
Complexes of Protein and NA

Berman H, Westbrook J, Feng Z, Gilliland T, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. Nucleic Acids Research. 2000. 28: 235-242. doi:10.1093/nar/28.1.235

How did it all started?

In 1958, Sir John Kendrew and his coworkers solved the first atomic structure of a protein, the myoglobin, revealing how it stores oxygen in muscle cells. The structure was a huge brass model.

The resolution for data collection was set to 6 Å. In further experiments the resolution was increased to 2 Å, which helped in establishing the secondary structure, with α-helices seen for the first time.

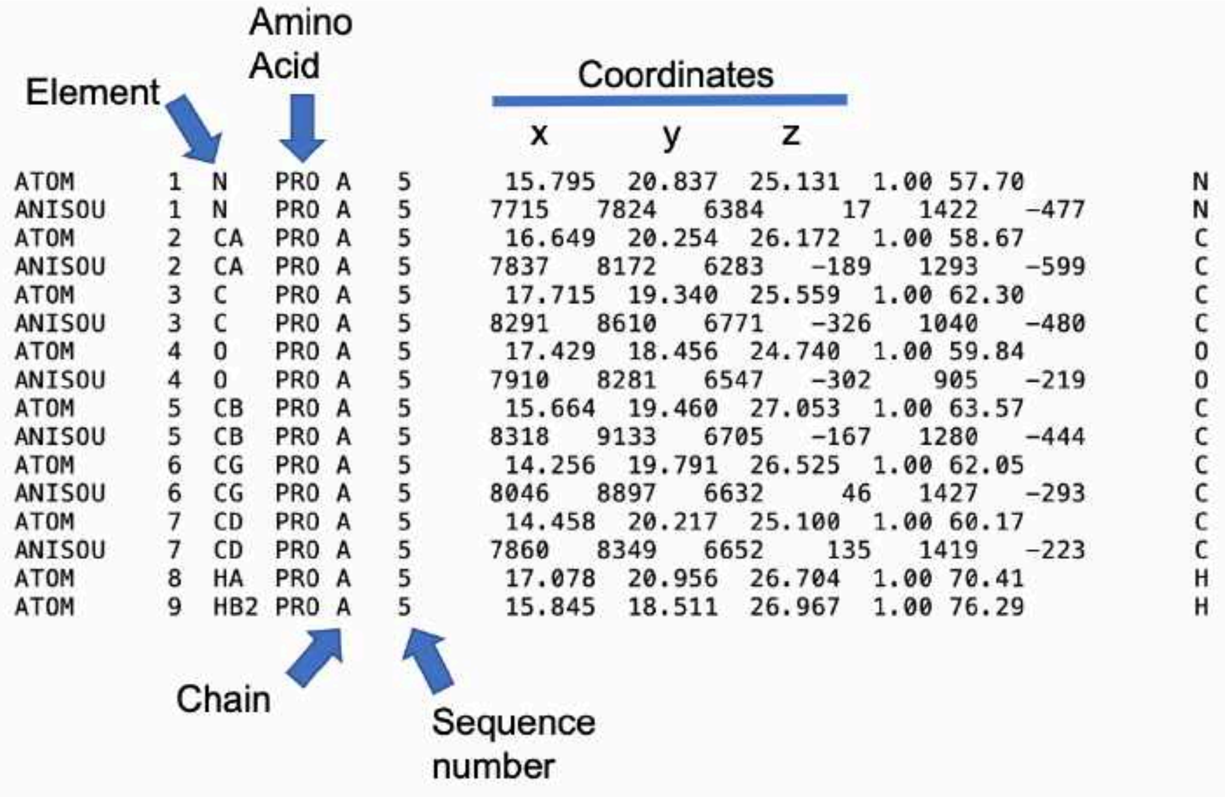


Myoglobin, 6 Å

It all started with myoglobin. 2019. https://www.ebi.ac.uk/pdbe/about/news/it-all-started-myoglobin

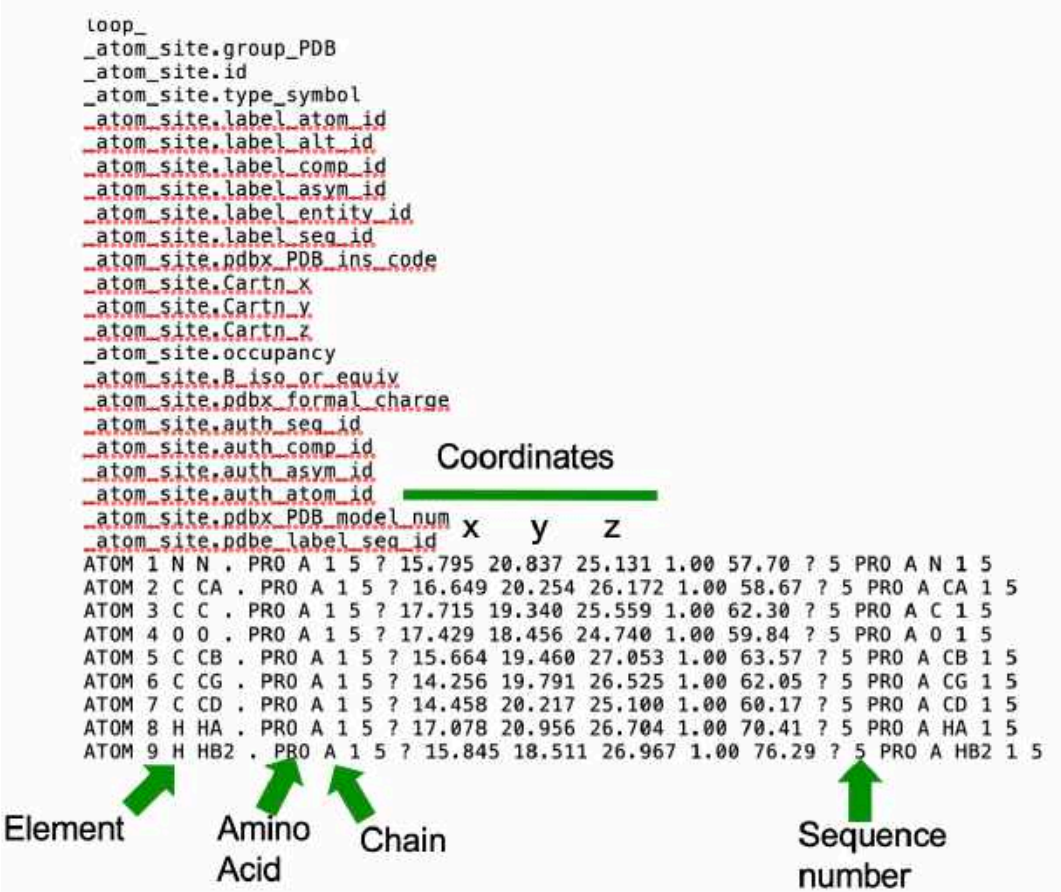
Introduction to coordinate file formats: Atomic-level data

PDB format



```
ATOM      1 N   PRO A   5     15.795  20.837  25.131  1.00 57.70           N
ANISOU    1 N   PRO A   5    7715    7824    6384      17    1422    -477      N
ATOM      2 CA  PRO A   5     16.649  20.254  26.172  1.00 58.67           C
ANISOU    2 CA  PRO A   5    7837    8172    6283    -189    1293    -599      C
ATOM      3 C   PRO A   5     17.715  19.340  25.559  1.00 62.30           C
ANISOU    3 C   PRO A   5    8291    8610    6771    -326    1040    -480      C
ATOM      4 O   PRO A   5     17.429  18.456  24.740  1.00 59.84           O
ANISOU    4 O   PRO A   5    7910    8281    6547    -302     905    -219      O
ATOM      5 CB  PRO A   5     15.664  19.460  27.053  1.00 63.57           C
ANISOU    5 CB  PRO A   5    8318    9133    6705    -167    1280    -444      C
ATOM      6 CG  PRO A   5     14.256  19.791  26.525  1.00 62.05           C
ANISOU    6 CG  PRO A   5    8046    8897    6632      46    1427    -293      C
ATOM      7 CD  PRO A   5     14.458  20.217  25.100  1.00 60.17           C
ANISOU    7 CD  PRO A   5    7860    8349    6652     135    1419    -223      C
ATOM      8 HA  PRO A   5     17.078  20.956  26.704  1.00 70.41           H
ATOM      9 HB2 PRO A   5     15.845  18.511  26.967  1.00 76.29           H
```

mmCIF format



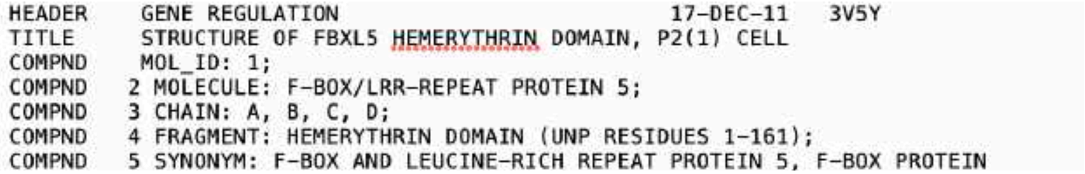
```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
_atom_site.pdbe_label_seq_id
ATOM 1 N N . PRO A 1 5 ? 15.795 20.837 25.131 1.00 57.70 ? 5 PRO A N 1 5
ATOM 2 C CA . PRO A 1 5 ? 16.649 20.254 26.172 1.00 58.67 ? 5 PRO A CA 1 5
ATOM 3 C C . PRO A 1 5 ? 17.715 19.340 25.559 1.00 62.30 ? 5 PRO A C 1 5
ATOM 4 O O . PRO A 1 5 ? 17.429 18.456 24.740 1.00 59.84 ? 5 PRO A O 1 5
ATOM 5 C CB . PRO A 1 5 ? 15.664 19.460 27.053 1.00 63.57 ? 5 PRO A CB 1 5
ATOM 6 C CG . PRO A 1 5 ? 14.256 19.791 26.525 1.00 62.05 ? 5 PRO A CG 1 5
ATOM 7 C CD . PRO A 1 5 ? 14.458 20.217 25.100 1.00 60.17 ? 5 PRO A CD 1 5
ATOM 8 H HA . PRO A 1 5 ? 17.078 20.956 26.704 1.00 70.41 ? 5 PRO A HA 1 5
ATOM 9 H HB2 . PRO A 1 5 ? 15.845 18.511 26.967 1.00 76.29 ? 5 PRO A HB2 1 5
```

Useful information: As the PDBx/mmCIF format continues to evolve, PDB format files will become outdated.

Berman H, Westbrook J, Feng Z, Gilliland T, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. Nucleic Acids Research. 2000. 28: 235-242. doi:10.1093/nar/28.1.235

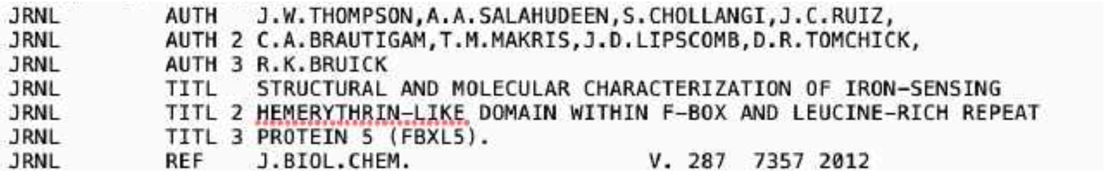
PDB format

Description of the molecule



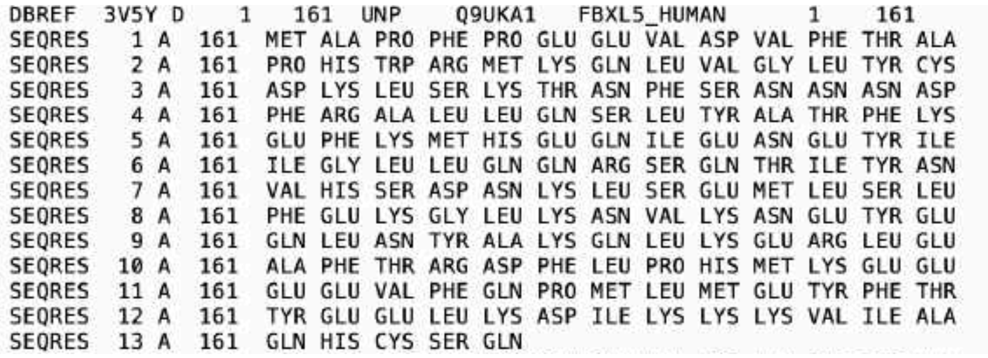
```
HEADER    GENE REGULATION                         17-DEC-11   3V5Y
TITLE     STRUCTURE OF FBXL5 HEMERYTHRIN DOMAIN, P2(1) CELL
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: F-BOX/LRR-REPEAT PROTEIN 5;
COMPND   3 CHAIN: A, B, C, D;
COMPND   4 FRAGMENT: HEMERYTHRIN DOMAIN (UNP RESIDUES 1-161);
COMPND   5 SYNONYM: F-BOX AND LEUCINE-RICH REPEAT PROTEIN 5, F-BOX PROTEIN
```

Authors information

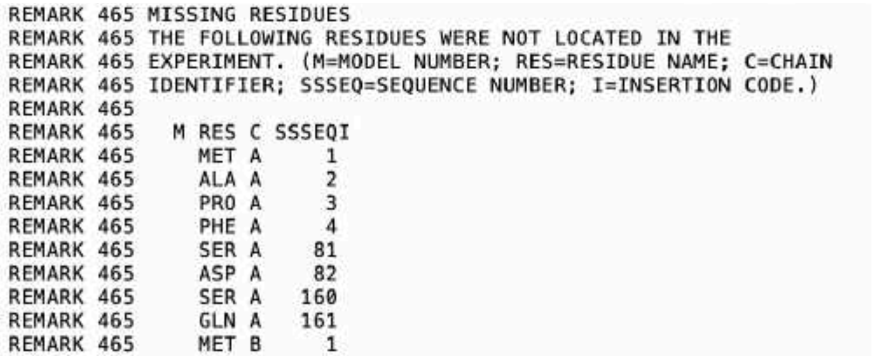


```
JRNL        AUTH   J.W.THOMPSON,A.A.SALAHUDEEN,S.CHOLLANGI,J.C.RUIZ,
JRNL        AUTH 2 C.A.BRAUTIGAM,T.M.MAKRIS,J.D.LIPSCOMB,D.R.TOMCHICK,
JRNL        AUTH 3 R.K.BRUICK
JRNL        TITL   STRUCTURAL AND MOLECULAR CHARACTERIZATION OF IRON-SENSING
JRNL        TITL 2 HEMERYTHRIN-LIKE DOMAIN WITHIN F-BOX AND LEUCINE-RICH REPEAT
JRNL        TITL 3 PROTEIN 5 (FBXL5).
JRNL        REF    J.BIOL.CHEM.                  V. 287  7357 2012
```

Sequence information (SEQRES). Sometimes some coordinate ATOM records are absent from SEQRES, those are missing residues. Ther are recorded in the REMARK 465 section.

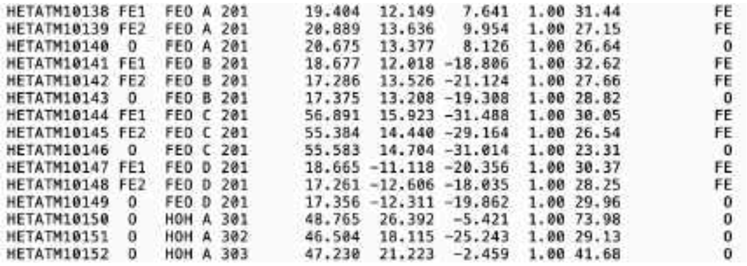


```
DBREF  3V5Y D    1   161  UNP    Q9UKA1   FBXL5_HUMAN      1    161
SEQRES   1 A  161  MET ALA PRO PHE PRO GLU GLU VAL ASP VAL PHE THR ALA
SEQRES   2 A  161  PRO HIS TRP ARG MET LYS GLN LEU VAL GLY LEU TYR CYS
SEQRES   3 A  161  ASP LYS LEU SER LYS THR ASN PHE SER ASN ASN ASN ASP
SEQRES   4 A  161  PHE ARG ALA LEU LEU GLN SER LEU TYR ALA THR PHE LYS
SEQRES   5 A  161  GLU PHE LYS MET HIS GLU GLN ILE GLU ASN GLU TYR ILE
SEQRES   6 A  161  ILE GLY LEU LEU GLN GLN ARG SER GLN THR ILE TYR ASN
SEQRES   7 A  161  VAL HIS SER ASP ASN LYS LEU SER GLU MET LEU SER LEU
SEQRES   8 A  161  PHE GLU LYS GLY LEU LYS ASN VAL LYS ASN GLU TYR GLU
SEQRES   9 A  161  GLN LEU ASN TYR ALA LYS GLN LEU LYS GLU ARG LEU GLU
SEQRES  10 A  161  ALA PHE THR ARG ASP PHE LEU PRO HIS MET LYS GLU GLU
SEQRES  11 A  161  GLU GLU VAL PHE GLN PRO MET LEU MET GLU TYR PHE THR
SEQRES  12 A  161  TYR GLU GLU LEU LYS ASP ILE LYS LYS LYS VAL ILE ALA
SEQRES  13 A  161  GLN HIS CYS SER GLN
```



```
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     ALA A     2
REMARK 465     PRO A     3
REMARK 465     PHE A     4
REMARK 465     SER A    81
REMARK 465     ASP A    82
REMARK 465     SER A   160
REMARK 465     GLN A   161
REMARK 465     MET B     1
```

HETATM record is used to identify atoms in small molecules



```
HETATM10138 FE1  FEO A 201      19.404  12.149   7.641  1.00 31.44          FE
HETATM10139 FE2  FEO A 201      20.889  13.636   9.954  1.00 27.15          FE
HETATM10140  O   FEO A 201      20.675  13.377   8.126  1.00 26.64           O
HETATM10141 FE1  FEO B 201      18.677  12.018 -18.806  1.00 32.62          FE
HETATM10142 FE2  FEO B 201      17.286  13.526 -21.124  1.00 27.66          FE
HETATM10143  O   FEO B 201      17.375  13.208 -19.308  1.00 28.82           O
HETATM10144 FE1  FEO C 201      56.891  15.923 -31.488  1.00 30.05          FE
HETATM10145 FE2  FEO C 201      55.384  14.440 -29.164  1.00 26.54          FE
HETATM10146  O   FEO C 201      55.583  14.704 -31.014  1.00 23.31           O
HETATM10147 FE1  FEO D 201      18.665 -11.118 -20.356  1.00 30.37          FE
HETATM10148 FE2  FEO D 201      17.261 -12.606 -18.035  1.00 28.25          FE
HETATM10149  O   FEO D 201      17.356 -12.311 -19.862  1.00 29.96           O
HETATM10150  O   HOH A 301      48.765  26.392  -5.421  1.00 73.98           O
HETATM10151  O   HOH A 302      46.504  18.115 -25.243  1.00 29.13           O
HETATM10152  O   HOH A 303      47.230  21.223  -2.459  1.00 41.68           O
```

PDB ID: 3v5y. Structure of FBXL5 hemerythrin domain. Thompson, J.W., Salahudeen, A.A., Chollangi, S., Ruiz, J.C., Brautigam, C.A., Makris, T.M., Lipscomb, J.D., Tomchick, D.R., Bruick, R.K.
J Biol Chem. 2012. 287: 7357-7365. doi: DOI: 10.1074/jbc.M111.308684

How can we measure the detail of the structural data?

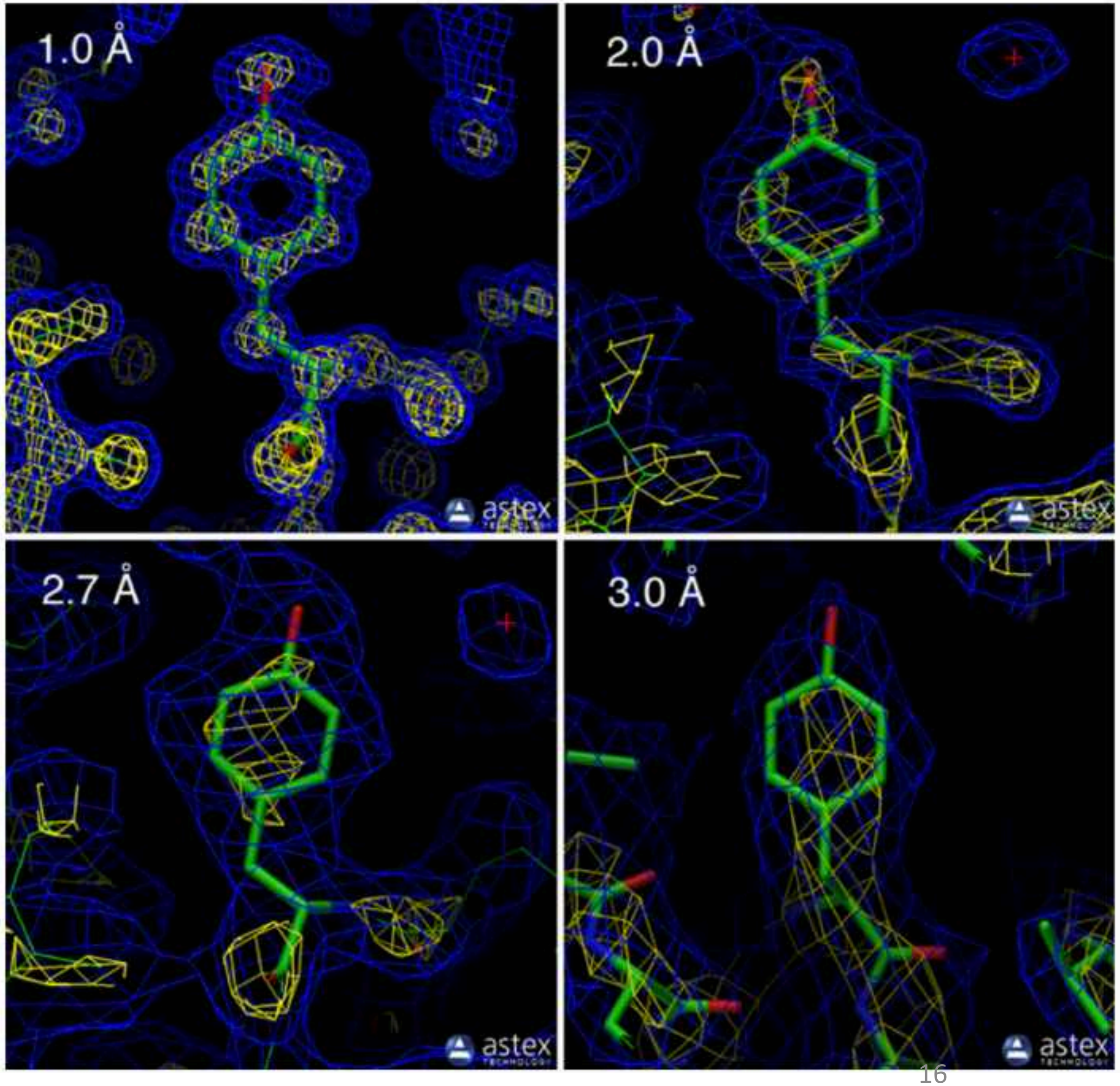
Resolution is a measure of the detail of the data.

High-resolution structures, with resolution values of 1 Å or so, are highly ordered and it is easy to see every atom in the electron density map.

- High = 1.0 - 1.8 A
- Medium = 1.8 - 3.0 A
- Low = > 3.0 A

Note: Not all parts of the structure are at the same resolution.

Berman H, Westbrook J, Feng Z, Gilliland T, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. Nucleic Acids Research. 2000. 28: 235-242. doi:10.1093/nar/28.1.235



Validation reports to measure structure quality

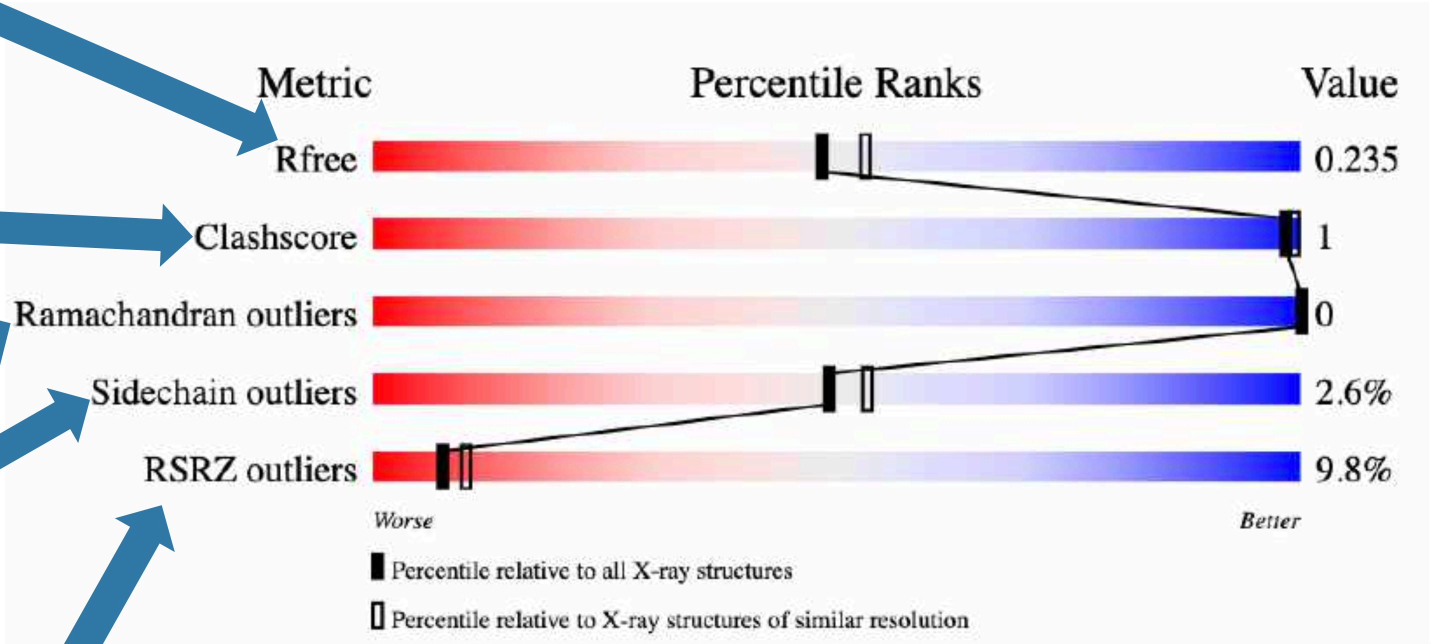
How well a simulated diffraction pattern matches the experimental one?

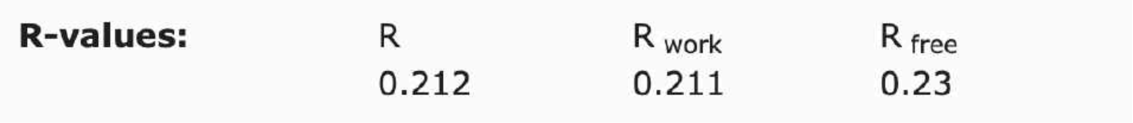
Number of atoms unusually too-close to each other

Unusual bond angles of the polymer residues

Percentage of residues with an unusual backbone conformation

Fraction of residues that do not fit the electron density





R-values:	R	R_{work}	R_{free}
	0.212	0.211	0.23

R-value is the measure of the quality of the atomic model obtained from the crystallographic data. When solving the structure of a protein, the researcher first builds an atomic model and then calculates a simulated diffraction pattern based on that model. The R-value measures how well the simulated diffraction pattern matches the experimentally-observed diffraction pattern.

Berman H, Westbrook J, Feng Z, Gilliland T, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. Nucleic Acids Research. 2000. 28: 235-242. doi:10.1093/nar/28.1.235