Structural bioinformatics KFC/STBI

What is structural bioinformatics?

Karel Berka

Miroslav Krepl

Requirements

- Project:
  - Structure analysis, docking, comparison of proteins, prediction of properties from structure, ...
  - 1(max. 2) page-long report with
    - Hypothesis
    - Brief Methodology
    - Conclusions

  ev. ChannelsDB – doplnění 5 struktur do databáze
- Exam:
  - Project-like Questions – problem + discussion about its possible resolution from you side

Content

- Structural bioinformatics, Biomolecules, Structural hierarchy
- Structure determination (X-Ray,NMR,EM), Structure file formats
- Structural databases (PDB, CATH, SCOP, Drugbank)
- Vizualization of structure, structural alignment
- Structure prediction, CASP, AlphaFold ML revolution
- Function prediction, CASA
- Binding prediction – protein-ligand and protein-protein docking
- Challenges of structural bioinformatics - membrane proteins, nucleic acids, protein-protein interactions prediction
- Examples: SARS-CoV-2, Switchable proteins

Bioinformatics

(Molecular) **bio** – informatics: bioinformatics is conceptualising **biology in terms of molecules** (in the sense of physical chemistry) and applying "***informatics techniques***" (derived from disciplines such as applied maths, computer science and statistics) to understand and ***organise*** the ***data and information*** associated with these molecules, on a ***large scale***.

In short, bioinformatics is a management information system for molecular biology and has many ***practical applications***.

Oxford English Dictionary

Structural bioinformatics

Use of structure

- Databases, classification
  - proteins, NA, drugs
- Patterns
  - Active sites, allosteric sites, ...
- Prediction
  - structure, function, active site, channels…
- Docking
  - Fitting of small molecules into the active site -> in silico drug design
- Simulations
  - What if…

Problems of structural bioinformatics

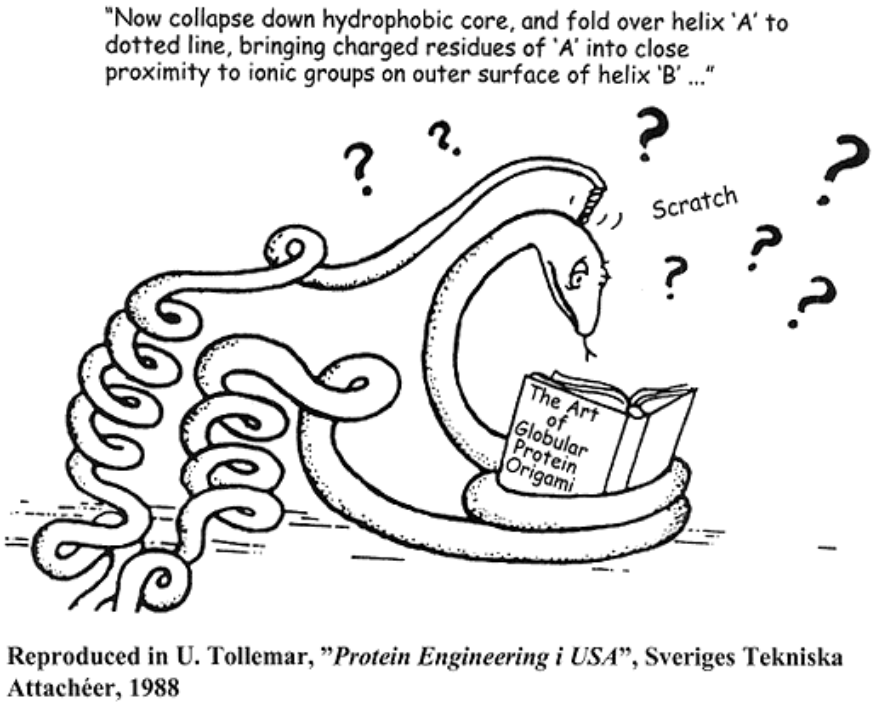
- Structural data are hard to work with:
  - Nonlinear
  - Imprecise from experiment (resolution of structure)
  - 3D representation (3D search)
  - Visualization is not trivial
  - More conserved than sequence data (genomics)
  - Structural genomics prepare structures without annotation
  - Most structures are water soluble globular proteins (most drug targets are membrane proteins)

Challenges

- Target selection
  - Large structures are resource intensive, maybe just one domain might be enough
- Structure methods
  - XRay – crystalisation is not easy
  - NMR – size problem – indistinguishable peaks
  - EM – only recently with atomistic detail
- Validation and Annotation
- Databases
- Correlation of structural data with experimental data

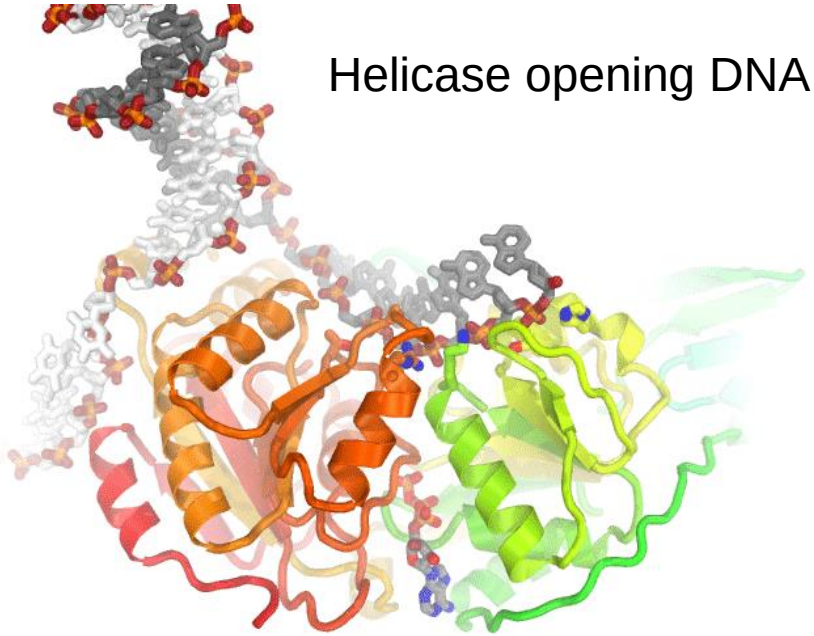
Example 1 : Prediction of protein structure

- Tertiary structure
  - Fold recognition
  - Homolog modelling
    - Structural alignment
  - ab initio modelling
  - ML methods
- Function prediction
  - active sites, channels, pores, allosteric sites, conformations…



Reproduced in U. Tollemar, "*Protein Engineering i USA*", Sveriges Tekniska Attachéer, 1988

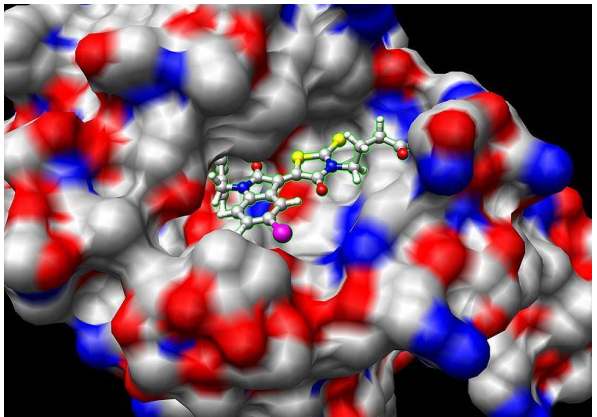
Example 2: Molecular graphics



Helicase opening DNA

- We make nice figs!
- Simulations
  - Structure => Energy
  - Time => Dynamics

- Docking – binding
  - ligands
  - Protein-protein



GOLD docking of compound to acetyltransferase

Structure Description

Coordinate systems

- XYZ (cartesian)
- Inner coordinates (bond lengths, bond angles, torsion angles)
- object representation (secondary structure)

Structure comparison:

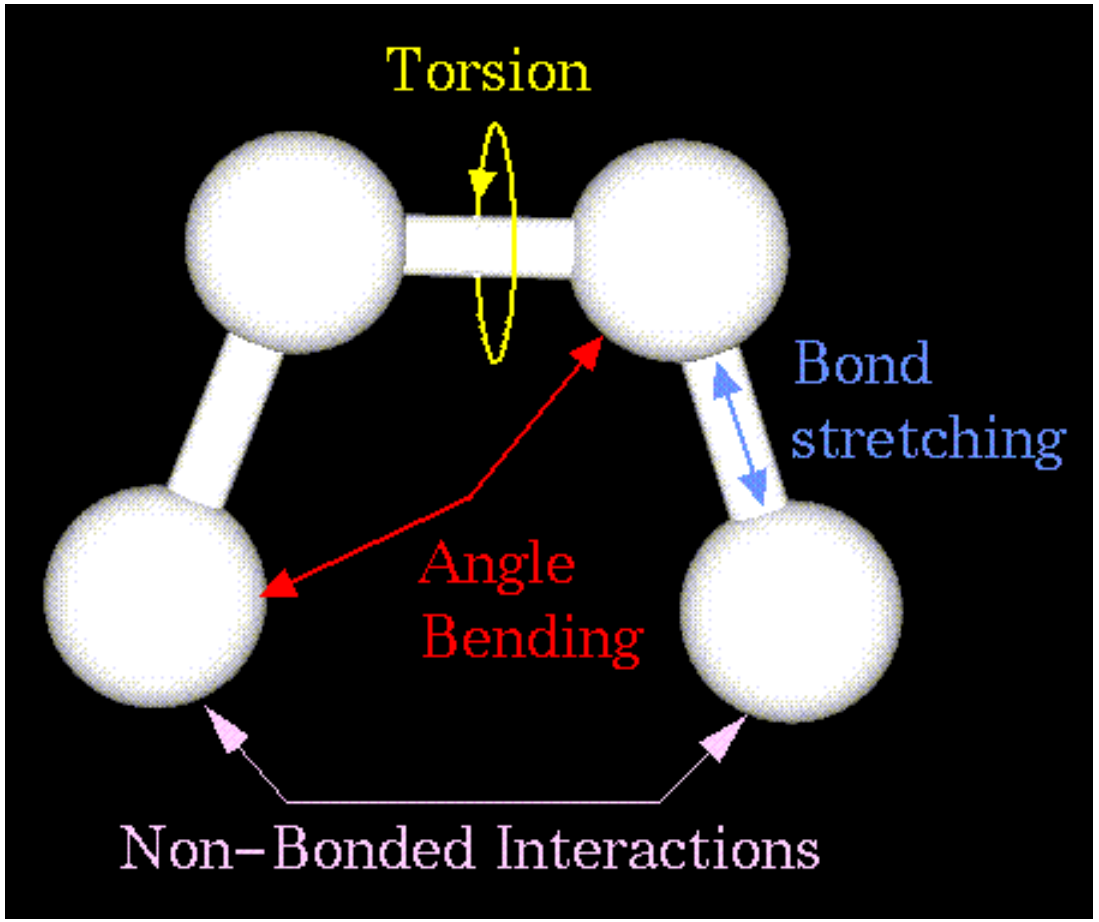
RMSD – root mean square distance

Typical geometrical operations

Bond lengths

Bond angles

Torsions
(dihedral angles)



Bond Lengths

- function of position of 2 atoms
- Bond length is almost constant
- Type of bond
  - simple C-C
  - double C=C
  - triple C≡C
- Minimal - 1.09 Å (C–H)
- Typical - 1.54 Å (C–C)
- Longer – heteroatoms (sulphur, halogens, metal ions)

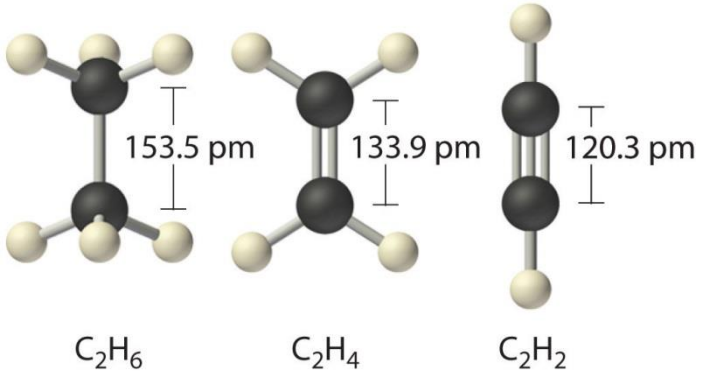


Table 9.2 Average Bond Energies (kJ/mol) and Bond Lengths (pm)

Bond	Energy	Length	Bond	Energy	Length	Bond	Energy	Length	Bond	Energy	Length
Single Bonds											
H—H	432	74	N—H	391	101	Si—H	323	148	S—H	347	134
H—F	565	92	N—N	160	146	Si—Si	226	234	S—S	266	204
H—Cl	427	127	N—P	209	177	Si—O	368	161	S—F	327	158
H—Br	363	141	N—O	201	144	Si—S	226	210	S—Cl	271	201
H—I	295	161	N—F	272	139	Si—F	565	156	S—Br	218	225
			N—Cl	200	191	Si—Cl	381	204	S—I	~170	234
C—H	413	109	N—Br	243	214	Si—Br	310	216			
C—C	347	154	N—I	159	222	Si—I	234	240	F—F	159	143
C—Si	301	186							F—Cl	193	166
C—N	305	147	O—H	467	96	P—H	320	142	F—Br	212	178
C—O	358	143	O—P	351	160	P—Si	213	227	F—I	263	187
C—P	264	187	O—O	204	148	P—P	200	221	Cl—Cl	243	199
C—S	259	181	O—S	265	151	P—F	490	156	Cl—Br	215	214
C—F	453	133	O—F	190	142	P—Cl	331	204	Cl—I	208	243
C—Cl	339	177	O—Cl	203	164	P—Br	272	222	Br—Br	193	228
C—Br	276	194	O—Br	234	172	P—I	184	246	Br—I	175	248
C—I	216	213	O—I	234	194				I—I	151	266
Multiple Bonds											
C═C	614	134	N═N	418	122	C≡C	839	121	N≡N	945	110
C═N	615	127	N═O	607	120	C≡N	891	115	N≡O	631	106
C═O	745 (799 in CO_2)	123	O_2	498	121	C≡O	1070	113			

Calculation of atom distance

In Cartesian coordinates:
For two points with coordinates (x_1,y_1,z_1) and (x_2,y_2,z_2)

$$d_{2-1} = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]}$$

Some distances within protein backbone are **constant** even if not in direct bond:

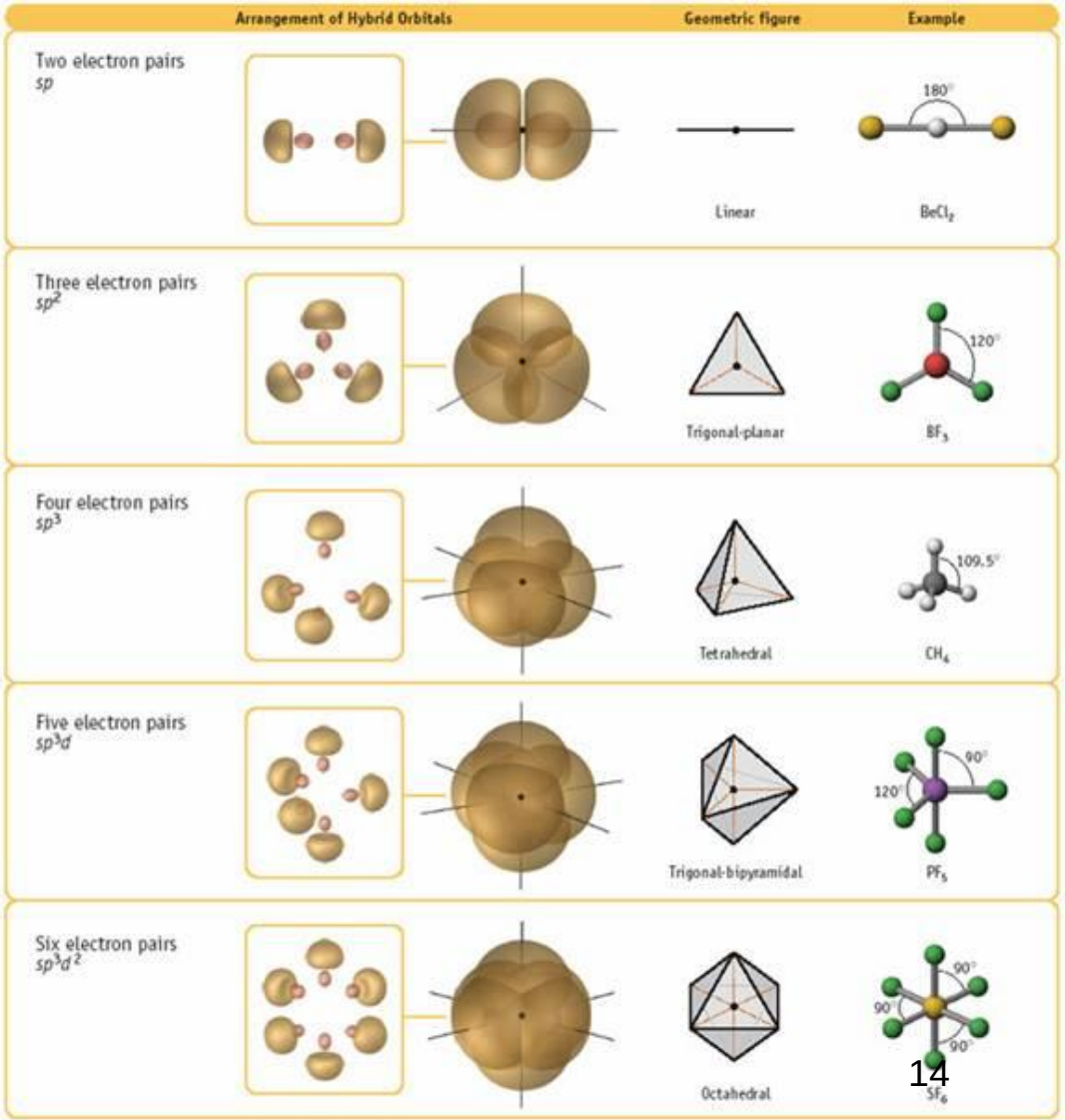
Cα – Cα distance between consecutive amino acids is 3.8 Å

Bond Angles

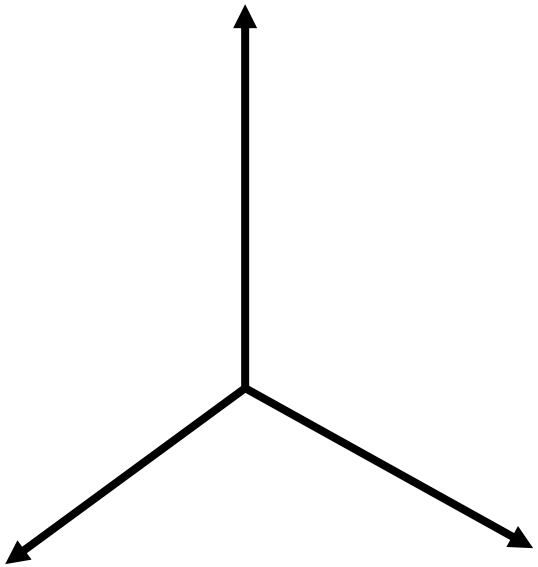
- function of position of 3 atoms
  - Almost constant for given combination of type of atoms

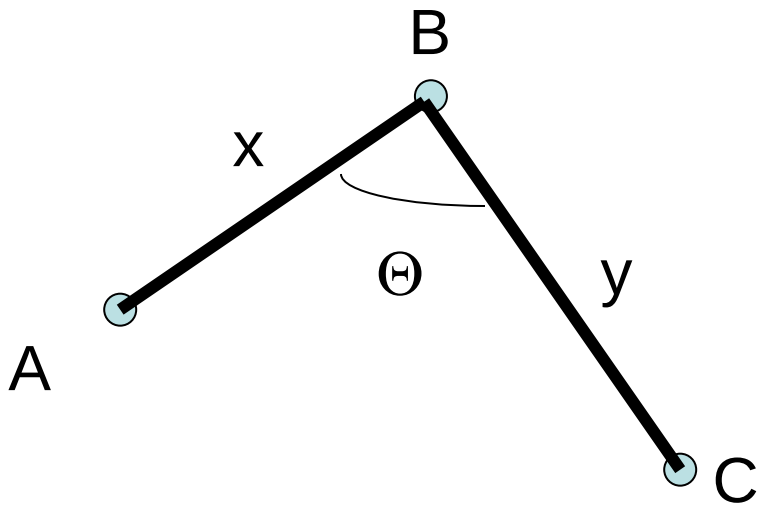
- Depend on atom type and number of electrons in bonding

- Interval from 90 to 180



Calculation of bonding angle





$X.Y = |X|.|Y|.\cos(\Theta)$

$\Theta = \arccos(X.Y/|X|.|Y|)$

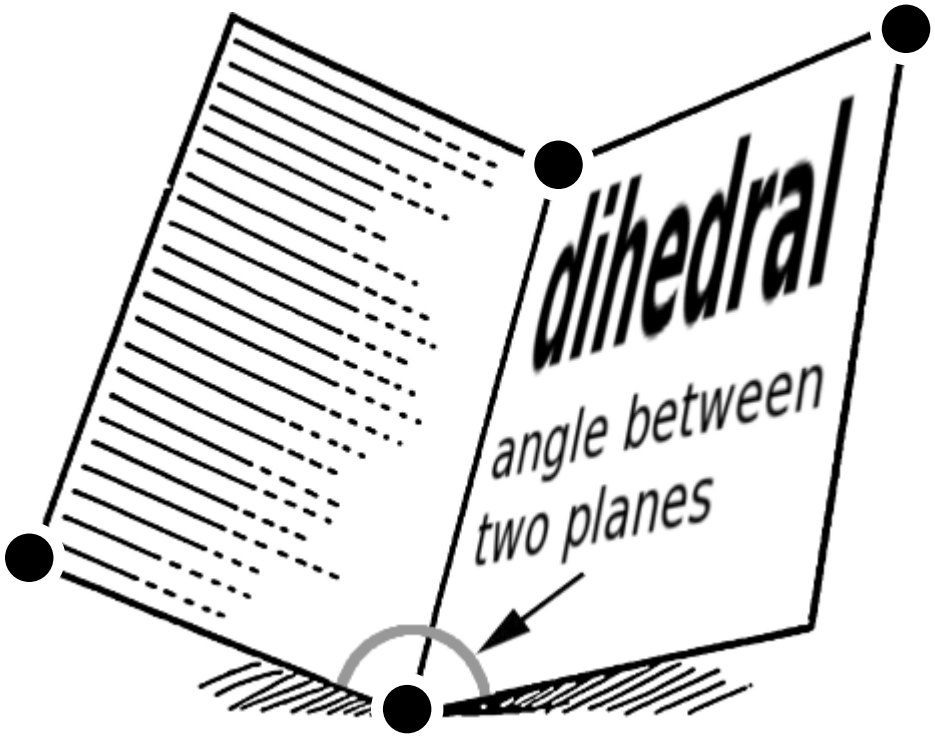
Arccosin of angle between two vectors BA and BC

Dihedral Angle

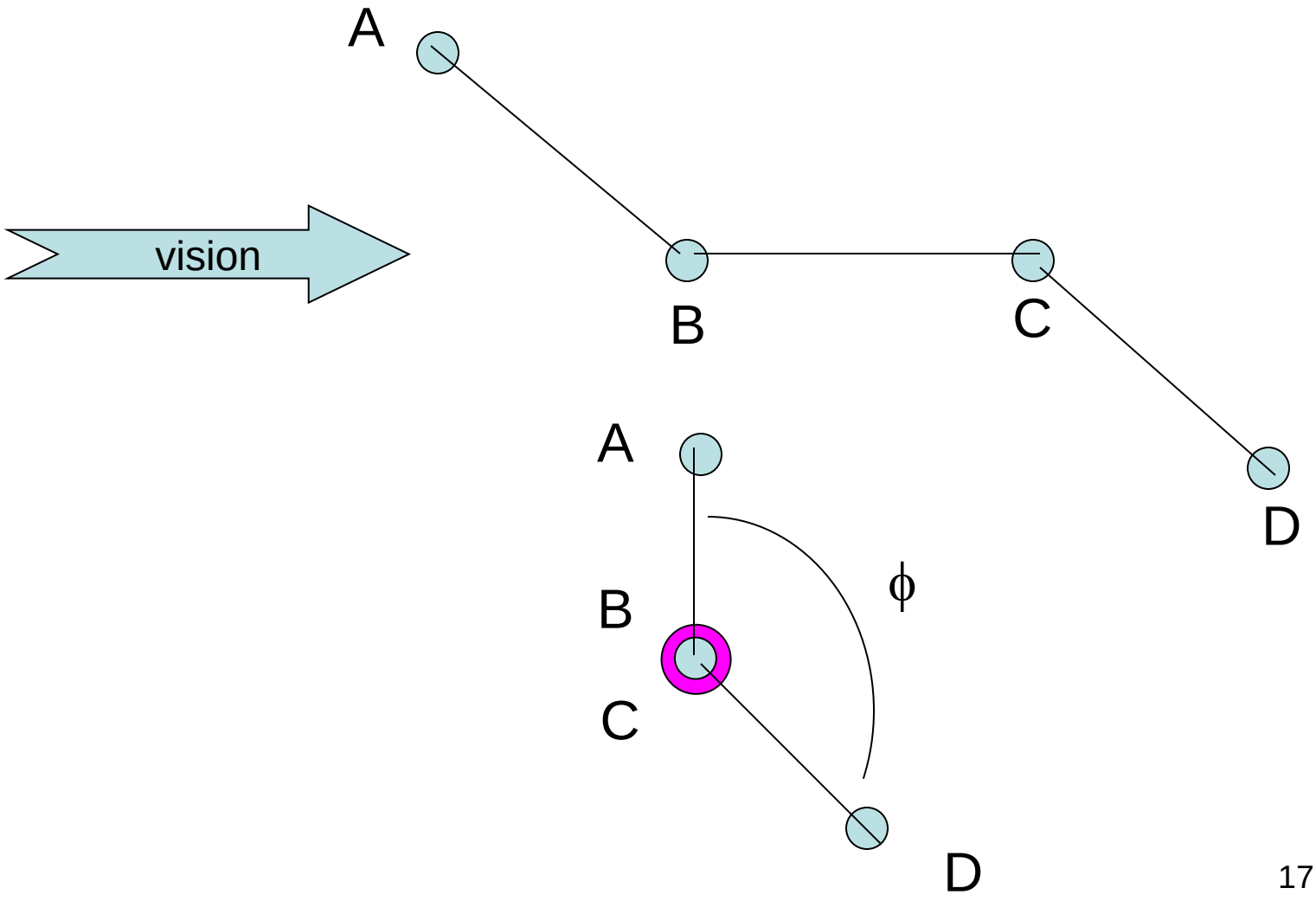
• function of position of 4 atoms

- Quite variable (0 to 360°)

- its change change conformations



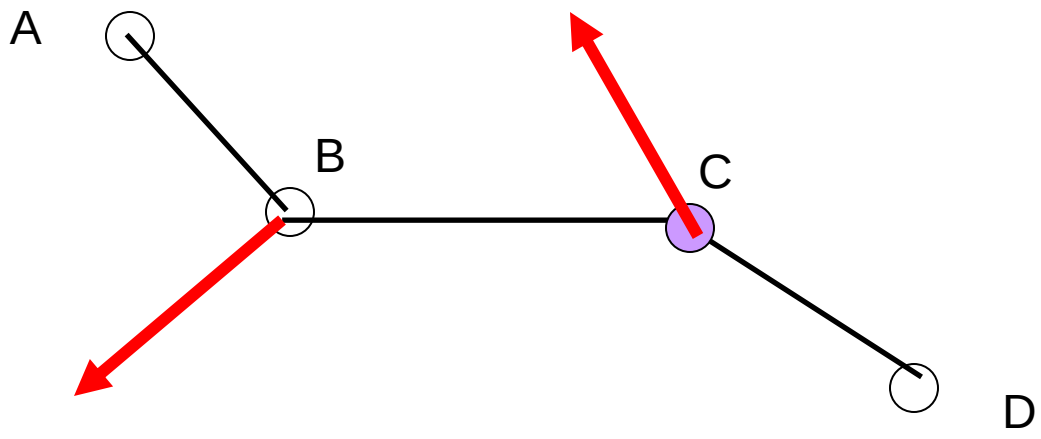
Dihedral Angle



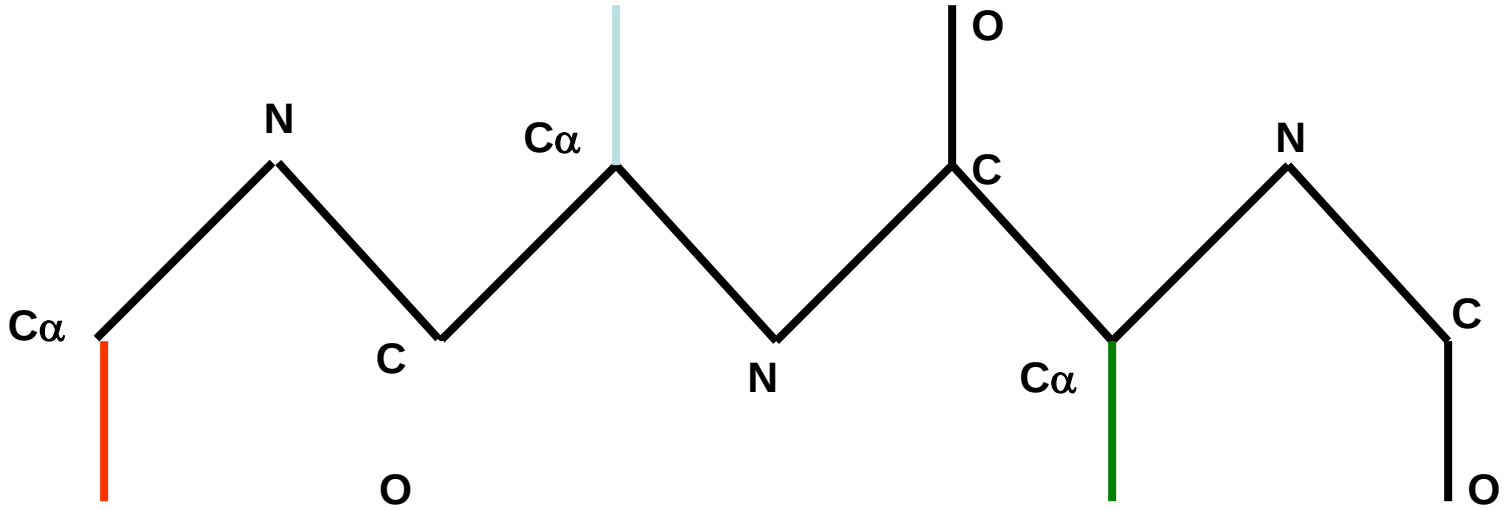
Calculation of dihedral angle

Dihedral angle = Angle between vectors orthogonal to planes defined by vectors:

1) Plane 1 - Vectors BA and CB
2) Plane 2 - Vectors CB and DC



Important dihedral angles in proteins



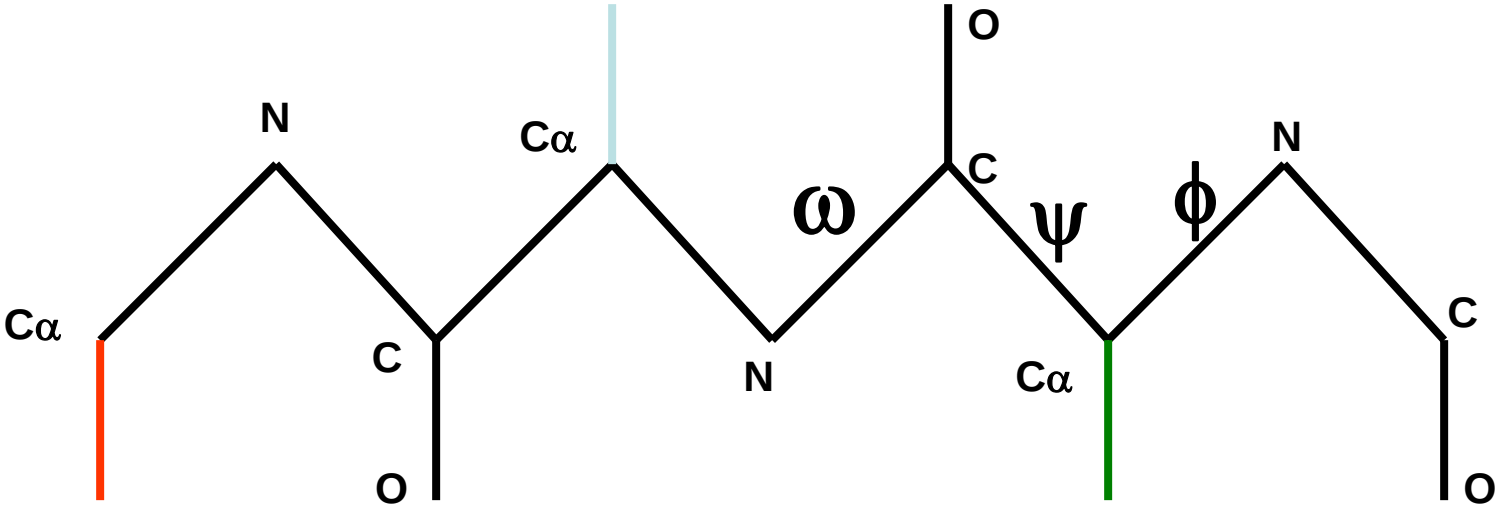
omega ω Cα - Cα

psi ψ N - N

phi ϕ C - C

Important dihedral angles in proteins

- Omega ω is constant = 180 (C-N do not rotate)
- Phi Φ ,Psi Ψ intervals (Cα-N, C-Cα can rotate) restricted to certain areas due to following amino acids



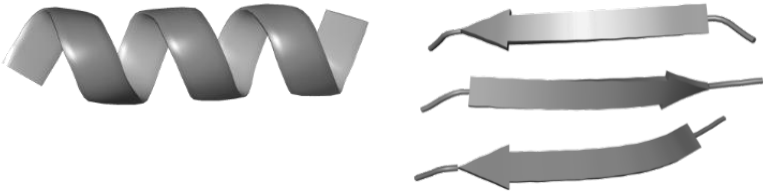
Ramachandran plot

- Typical values of dihedral angles define individual secondary structure elements:

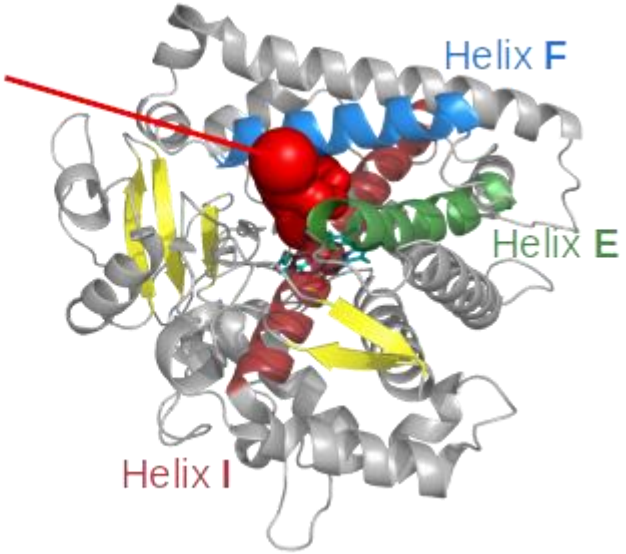
  - α-helix phi = - 57, psi = - 47
  - 3-10 helix phi = - 49, psi = - 26
  - Parallel β-sheet phi = - 119, psi = 113
  - Antiparallel β-sheet phi = - 139, psi = 135

Secondary structure

Helices and β-strands = Secondary Structure Elements (SSEs)



- Quite conserved arrangement within a protein family
- Can serve as landmarks, which
  - Help us orient in the structure
  - Help us locate the key regions (active sites, channels…)



Other Coordinate Systems

Cartesian coordinates are orthogonal (x,y,z)

-> used most often

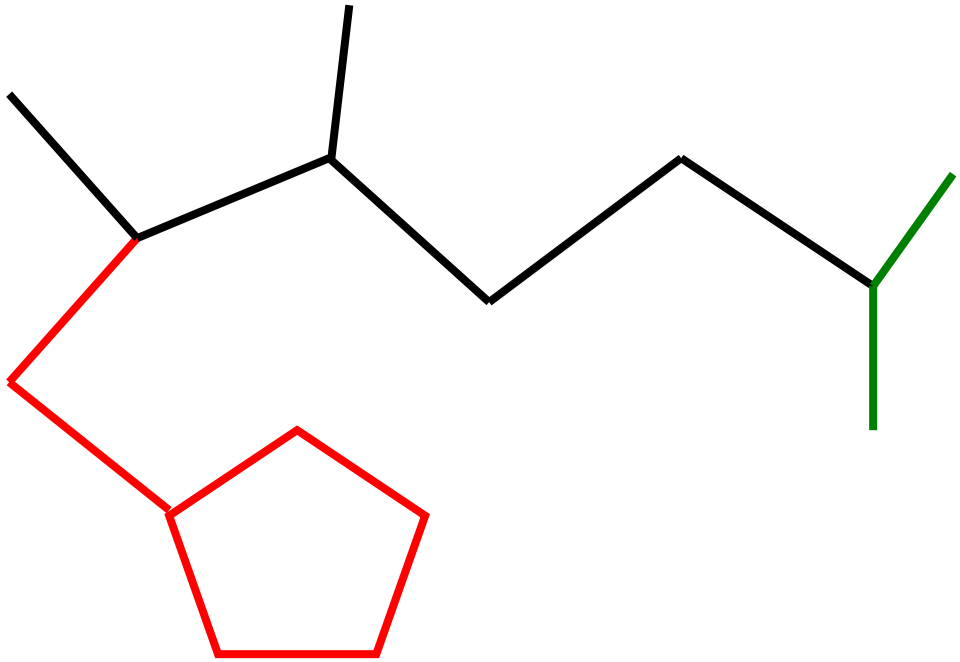
If bond lengths and bond angles are constant -> reduction of coordinates -> only dihedral angles =>

Inner coordinates

If some part of structure can be defined by “rigid” structural element -> solid objects =>

Object-based coordinates

Advantages of Inner Coordinates



3 peptide units = 12 atoms = 36 coordinates OR 6 dihedral angles
3 sidechains = 12 atoms = 36 souřadnic OR 5 dihedral angles

72 cartesians versus 11 inners

Disadvantages of Inner Coordinates

Some calculations are more difficult

Atom-atom distance

Closest atoms toward a point in space

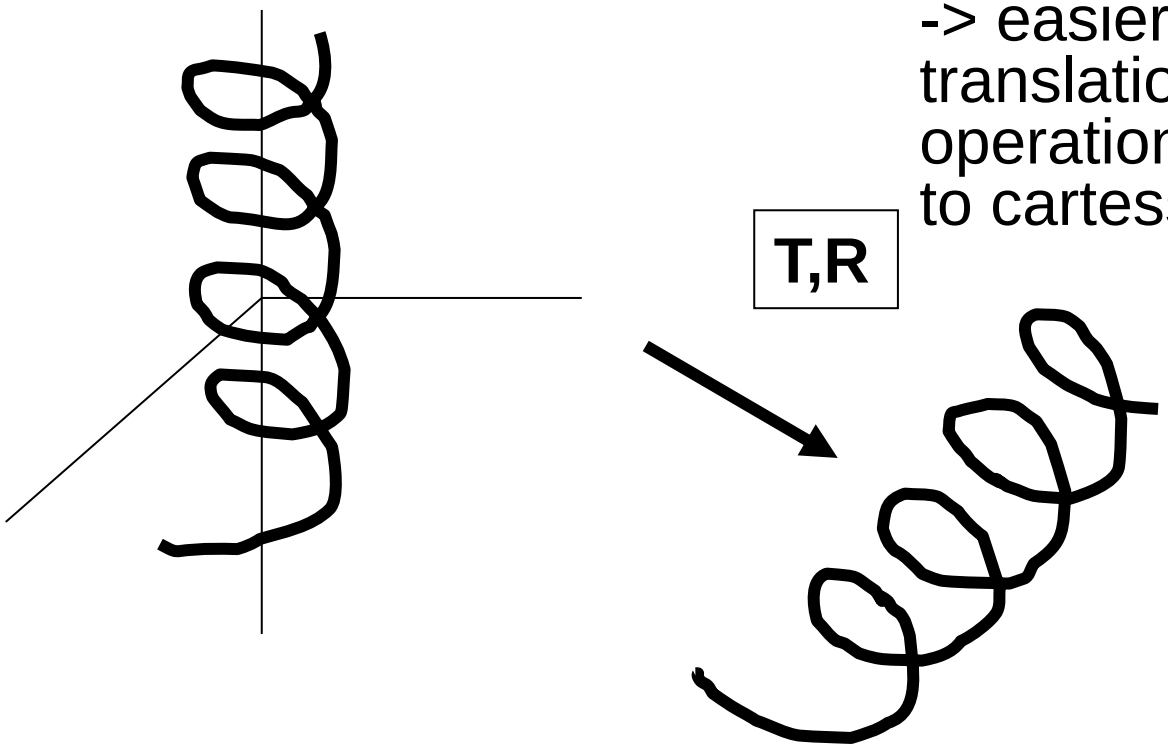
Hard comparison of independent objects (two molecules)

Nonlinear relationships between coordinates => problem for optimizations and simulations

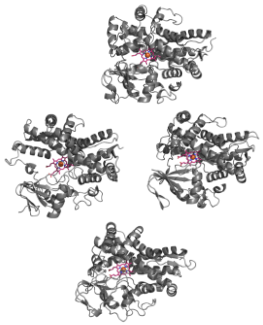
Object-based coordinates

Use of larger objects – secondary structure, subset of atoms…

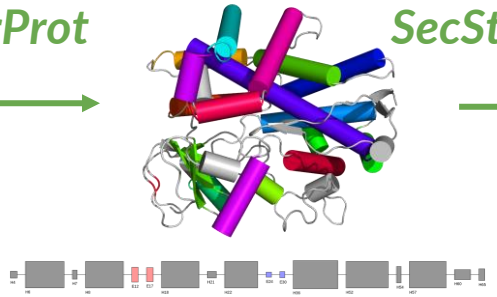
Example -> Helix can be represented as a vector with just 6 coordinates -> easier operations such as translation and rotation (final operation is later transferred to cartessian coordinates)



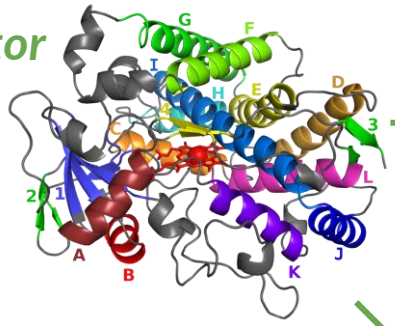
Bigger image



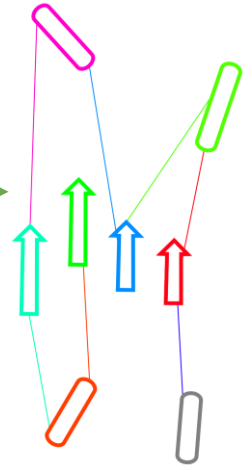
OverProt



SecStrAnnotator

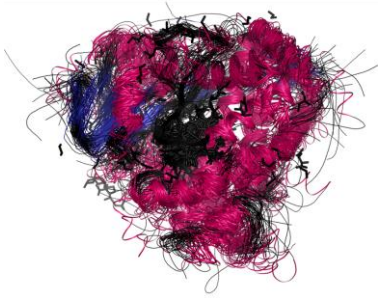


2DProt



OverProt web server

CATH/PDBe



Precomputed for CYP family and available through SecStrAPI plugin in PyMOL

Annotation of key regions

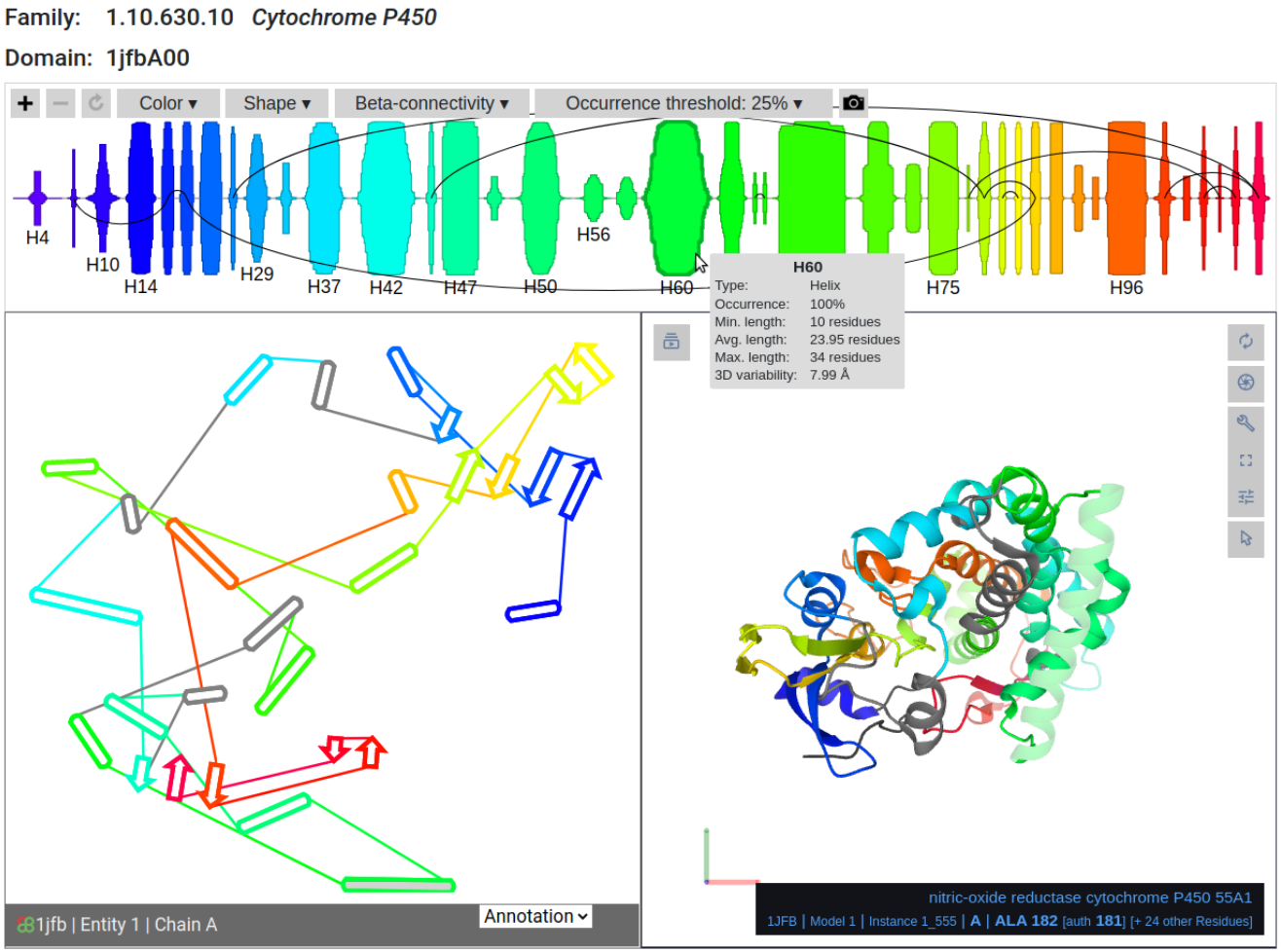
Midlik A, Hutařová Vařeková I, Hutař J, Chareshneu A, Berka K, Svobodová R: **OverProt**: secondary structure consensus for protein families, *Bioinformatics*, 38(14), July 2022, 3648–3650

Midlik A, Navrátilová V, Moturu TR, Koča J, Svobodová R, Berka K: Uncovering of cytochrome P450 anatomy by **SecStrAnnotator**. *Sci Rep* 11, 2021, 12345

Hutařová Vařeková I, Hutař J, Midlik A, Horský V, Hladká E, Svobodová R, Berka K, **2DProts**: database of family-wide protein secondary structure diagrams, *Bioinformatics*, 37(23), 2021, 4599–4601,

OverProt Server – Interactive view

- 1D of the family linked to 2D and 3D of a domain



Midlik A, Hutařová Vařeková I, Hutař J, Chareshneu A, Berka K, Svobodová R: **OverProt**: secondary structure consensus for protein families, *Bioinformatics*, 38(14), July 2022, 3648–3650

Structure Comparison

For comparison of two structures A and B we need:

1. **Which atom from A corresponds to which atom from B**
   => alignment
2. **Atom localization**
   => PDB files
3. **Comparison criteria**
   RMSD, energy

RMSD = Root Mean Square Deviation

- Atoms from A and B are taken as equivalent
- Superposition and calculation of differences in distance

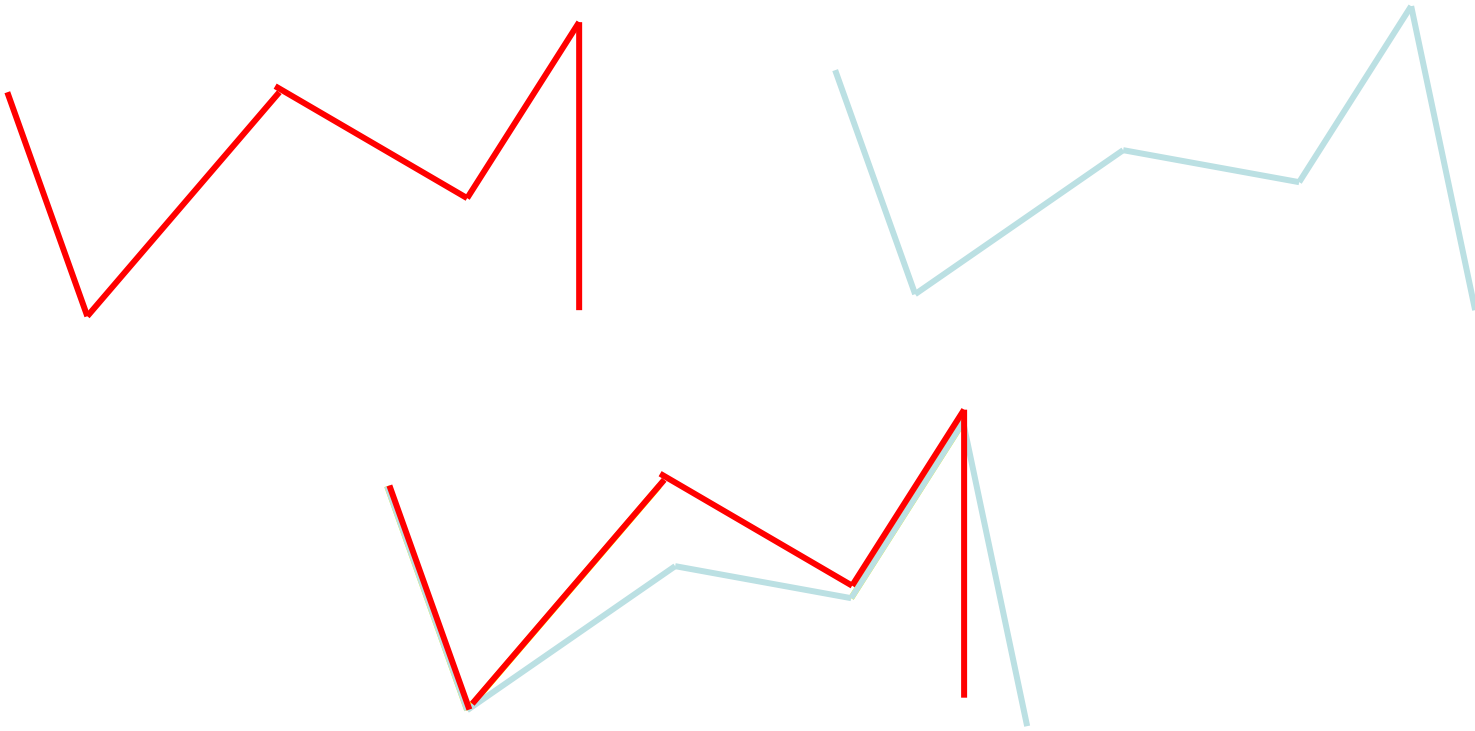
$$\mathbf{RMSD} = \sqrt{\frac{\Sigma\, \mathbf{d}^2_i}{\mathbf{N}}}$$

- If are structures identical -> RMSD = 0
- With more differences between structures -> RMSD increses

N – number of atoms
d_i – distance of two atoms with index *i* from A and B

Structure Comparison

To find minimal RMSD



Calculation of RMSD

- translate and rotate one structure with respect to the other to minimize the RMSD

- Centroid-based solutions

  (Huang,Blostein,Margerum)
- Quaternion-based solutions
  - (rotation-translation) that minimizes the RMSD between two sets of vectors

  (Faugeras a Hebert, Petitjean)
- Matrix Singularity-based methods

  (Arun, Huang, Blostein)

Arun algorithm

- Matrices of pi’ = R.pi + T + Ni
  - pi – 3x1 column matrix of positions
  - R – rotation matrix
  - T – translation vector 3x1 column matrix
  - N – noise vector

- 1) Translation over **centroids**
- 2) Singular value decomposition of matrix to obtain **rotation**

- Arun algorithm is optimal, universal and not iterative

Kabsch algorithm

- 1) Translation over **centroids**
- 2) computation of a **covariance matrix**,
- 3) the computation of the **optimal rotation matrix**.

- Kabsch algorithm is widely used as *fit* function in PyMol, or within VMD
- Algorithm do not recognise similar pairs of residues – these have to be defined iteratively (typically Cα)

Kabsch, W (1976): A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* ***A32*** *(5): 922.* With a correction in Kabsch, W (1978). "A discussion of the solution for the best rotation to relate two sets of vectors". *Acta Cryst.* ***A34*** *(5): 827–828.*

Advantages and Disadvantages of RMSD

Good behavior, identical structures RMSD = 0

Simple calculation in Cartesian coordinates

Natural units (Ångstroms)

Experience (similar structures have RMSD ~ (1 – 3 Å)

Weight of all atoms is the same

however hydrogens have much smaller effect in practice
–> RMSD only for backbone or Cα

Prone to extremities

RMSD of larger protein is larger even if the structure is almost identical

RMSD of 3 Å for 100 residue protein is really bad,
for 1000 residue protein it is sensible.

Other measures

- **global distance test (GDT)**
  - largest set of amino acid residues' Cα atoms in the model structure falling within a defined distance cutoff of their position in the experimental structure.

  ⇒Used in structure prediction assessment (CASP)
- **template modeling score (TM-score)**
  - difference between two structures by a score between (0,1]

$$\text{TM-score} = \max\left[\frac{1}{L_{\text{target}}}\sum_{i}^{L_{\text{aligned}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}\right]$$

  - TM-score = 1 - perfect match between two structures
  - TM-score > 0.5 assume roughly the same fold
  - TM-score < 0.20 - randomly chosen unrelated proteins

  ⇒Used in structure prediction assessment (CASP)