3D Structural Alignment of Proteins

part of “Bioinformatik von RNA- und Proteinstrukturen”

Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, SS 2011

Why Align Protein Structures?

- functional cues rather from structure than sequence
- low sequence similarity but high structural similarity
- about 1000 known protein fold classes
- most newly discovered proteins fall into these classes
- ...

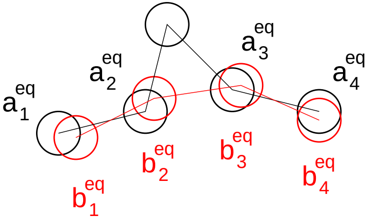
Data – the protein backbone

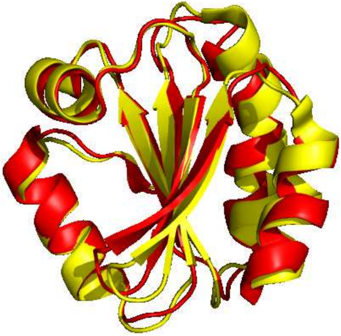
Given an amino acid sequence A of length m, let A_α be the sequence of all C_α atoms in A, $A_\alpha = \{a_1, a_2, ..., a_m\}$. Then a_i is a vector $\vec{a_i} = (x_i, y_i, z_i)$ specifying the coordinates of the C_α atom of the amino acid at position i in sequence A.
This information can be extracted from a PDB file (here: 1E7K):

- column [0]: “ATOM”
- column [2]: atom type, e.g. “CA” – C_α
- column [3]: amino acid in three letter code
- column [4]: chain
- column [5]: position of the amino acid in the sequence
- columns [6-8]: x, y, and z coordinates

Problem of Structural Alignment

Given two amino acid sequences *A* and *B* find a super-positioning of the atom coordinates such that it minimizes the distance between a_i^{eq} and b_i^{eq} for all pairs *i* of equivalent amino acids.





General Strategy

- chose the atom type appropriate to answer the question to be asked (C_α atoms for protein fold super-positioning)
- get a set of **equivalent atoms** as seed for super-positioning
- define a **distance or similarity measure** to evaluate the quality of a super-positioning
- minimize distance or maximize similarity for equivalent atoms
- infer **translation vector** and **rotation matrix** for super-positioning
- return coordinates of superposed molecules

Quality and Extent of Structural Matches

- root mean square deviation (RMSD)
  - minimize the RMSD
- topological equivalent atoms
  - two atoms are equivalent if they are closer than a distance *d*
  - distance between $C_\alpha(i)$ and $C_\alpha(i+1)$ is 0.35nm
  - difference in atom position of measurements is 0.03nm
  - maximize the number of equivalent atoms
- GDT – global distance test
- TM-score – template modeling score
- structural alignment score
  - result of dynamic programming algorithm for overall optimal structural super-positioning

Root Mean Square Deviation (RMSD)

Euclidian Distance

Given equivalent atom postions a_i^{eq} and b_i^{eq} the distance D is:

$$D_i = \sqrt{(x(a_i^{eq}) - x(b_i^{eq}))^2 + (y(a_i^{eq}) - y(b_i^{eq}))^2 + (z(a_i^{eq}) - z(b_i^{eq}))^2}$$

RMSD

The average level of deviations over all equivalent atoms is:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} D_i^2}$$

Rigid-Body Super-Positioning

Given is a set of N equivalent atoms for structures A and B and their atom coordinates.

Translation Vector

Calculate the center of mass T_A and T_B for A and B, resp.

$$T_A = \sum_{i=1}^{N} \frac{a_i^{eq}}{N} = \begin{pmatrix} (x_1 + x_2 + ... + x_N)/N \\ (y_1 + y_2 + ... + y_N)/N \\ (z_1 + z_2 + ... + z_N)/N \end{pmatrix}$$

Translate both structures so that their centers of mass are located at the origin of the coordinate system.

$$a_j^{all}(new) = a_j^{all}(old) - T_A = \begin{pmatrix} x_j(old) - x(T_A) \\ y_j(old) - y(T_A) \\ z_j(old) - z(T_A) \end{pmatrix}$$

Rigid-Body Super-Positioning

Solve the pairwise least square problem.

Rotation Matrix
To obtain the rotation matrix *R* minimize the residual δ:

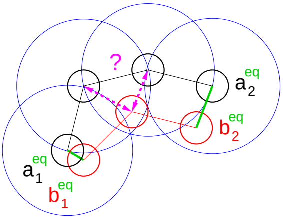
$$\delta = \sum_{i=1}^{N} (a_i^{eq} - Rb_i^{eq})^2$$

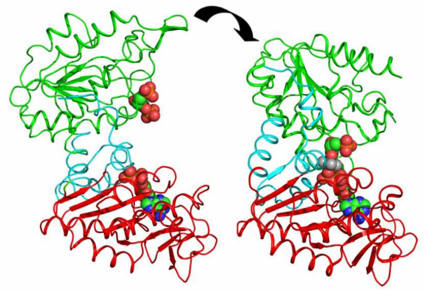
Apply the roation matrix *R* to all atoms in structure *B* to super-position *A* and *B*.

$$b_j^{all}(new) = Rb_j^{all}(old)$$

Multiple Matches

After rigid-body structural comparison, atoms of the two structures are **considered equivalent** if the Euclidian **distance is lower than a cut-off** (usually 0.3nm, in general in the range of 0.25-0.45nm.)





- multiple matches result from the process
- not useful for super-positioning of distantly related structures

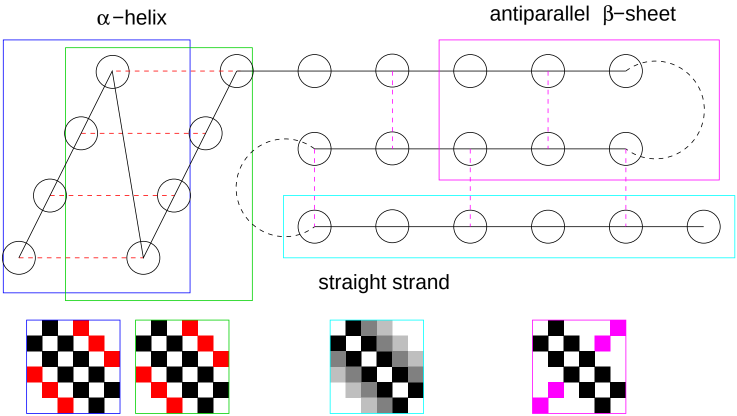
Finding Equivalent Atoms Initially

- suply a minimum of three equivalent atoms
- super-position ligand to study ligand binding
- align active site residues to study surrounding structure
- use sequence alignment if sequence and structure are similar
- use secondary structure alignment if only structure is similar

followed by rigid-body super-positioning. Than improve by

- **dynamic programming method** calculates an alignment (with gaps) minimizing paired atom distances
- **clique detection method and match list approach** compare graphs from intra-molecular C_α relationships with C_α atoms as nodes and spacial distances as vertices (e.g. FAST, DALI)

Distance Alignemnt Matrix Method (DALI)



6 × 6 distance matrices from local motifs are compared (α-helix is similar to α-helix, see Figure.)

Literature

Bioinformatics: Sequence, structure and databanks. A practical approach. Des Higgins and Willie Taylor. Oxford University Press, 2000.