Structural Alignments

28.5.2013

Katharina Hembach

Outline

1. Recap: SCOP and CATH databases
2. Structural alignment methods and scores:
    1. SSAP
    2. TopMatch
    3. CE
    4. LGA

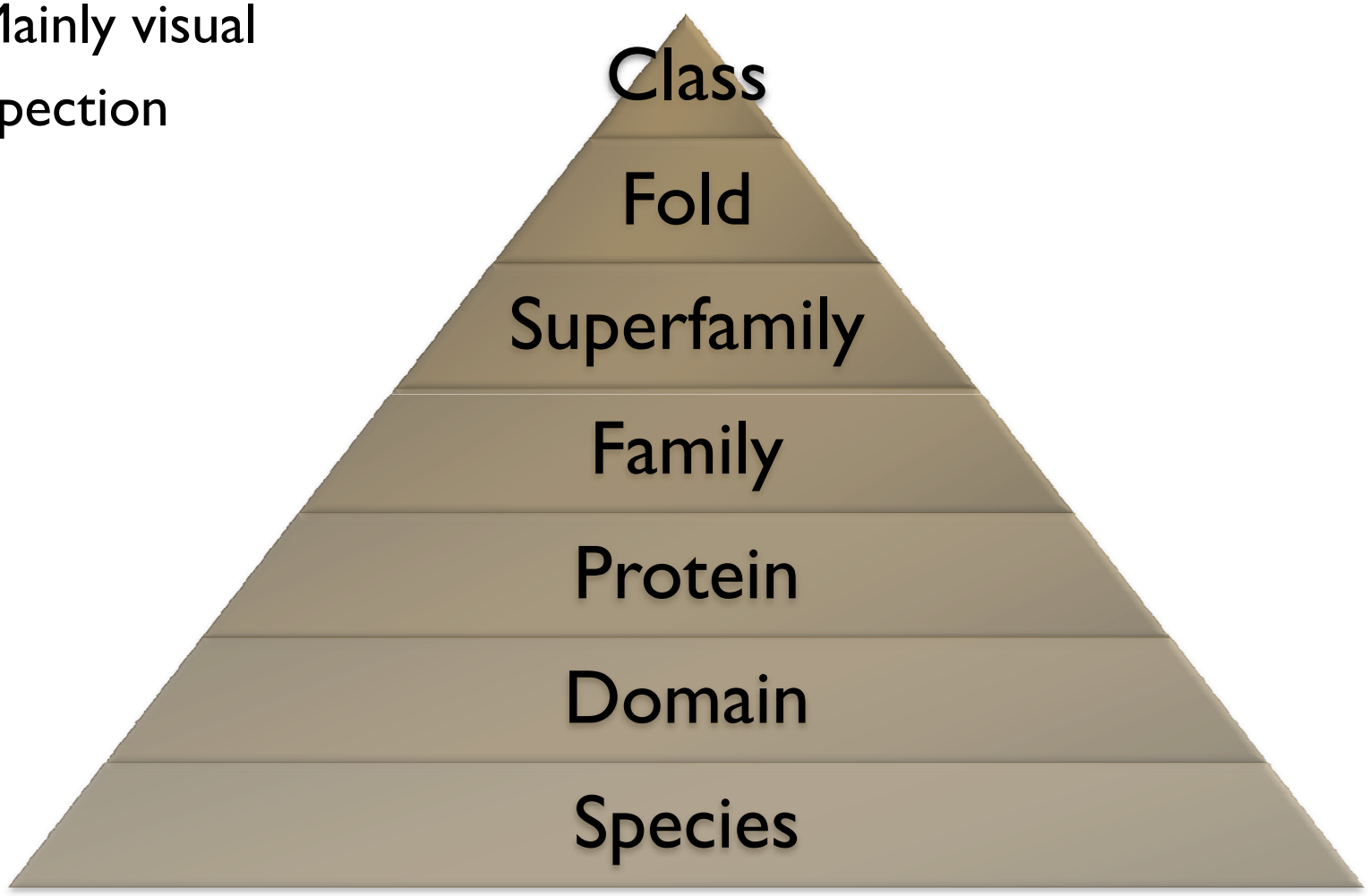
3D classification

- tells us about
  - → function
  - → evolution
- of unknown proteins

- helps to annotate proteins

SCOP
(structural classification of proteins)

- Mainly visual inspection



SCOP
(structural classification of proteins)

- classes
  - alpha
  - beta
  - alpha and beta (a/b)
  - alpha plus beta (a+b)
  - multi-domain proteins
  - membrane and cell-surface proteins and peptides

  - small proteins
  - coiled coil proteins
  - low-resolution protein structures
  - peptides
  - designed proteins

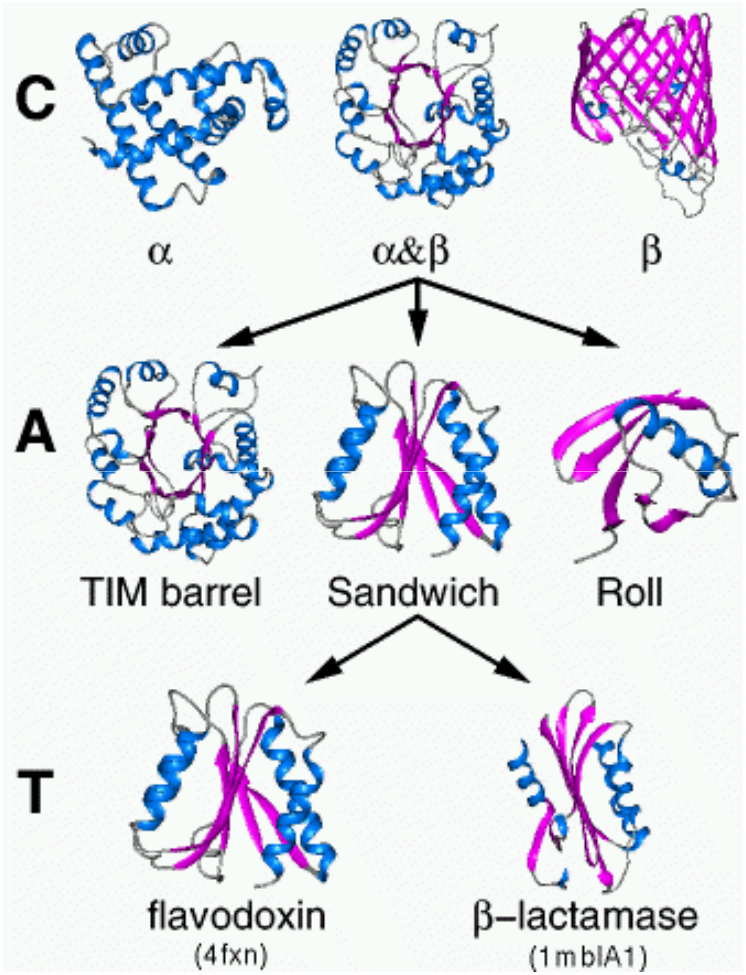
CATH

Class: mainly-alpha, mainly-beta, alpha-beta, low secondary structure

Architecture: shape of domain structure

Topology (Fold family): shape and connectivity of secondary structures

Homology: groups families that have a common ancestor



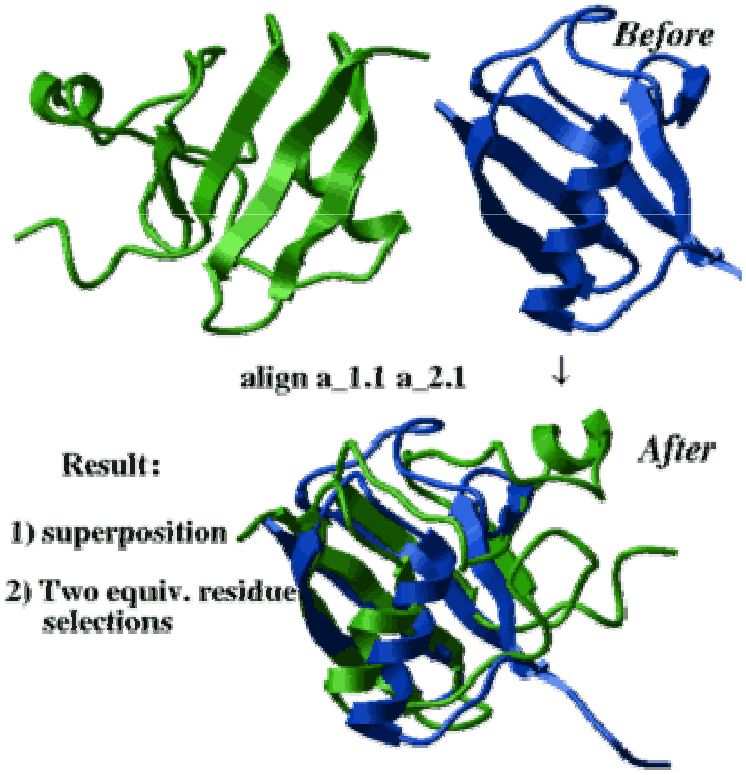
http://protein.hbu.cn/cath/cathwww.biochem.ucl.ac.uk/latest/cath_info.html

Structural alignment

- Superposition of two structures

1. find corresponding positions
2. compute best superposition

- Calculate score



http://www.molsoft.com/man/align3d.png

RMSD (root mean square deviation)

- squared distance between corresponding positions (typically Cα) of two superimposed proteins A and B

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}$$

- where $$d_i^2 = (a_i - b_i)^2$$

- d_i is the distance between two corresponding points a_i and b_i

Structural alignment methods

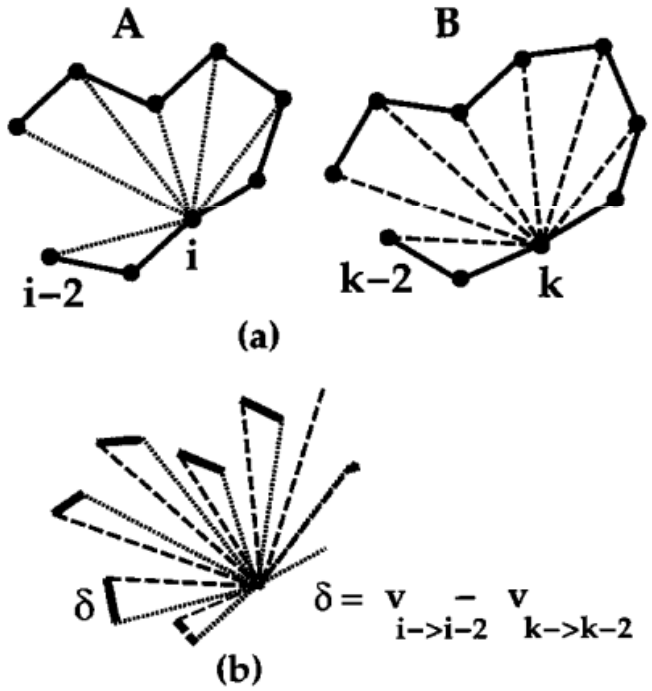
- For example:

- SSAP
- TopMatch
- CE
- LGA

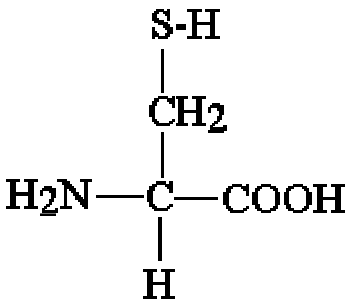
SSAP
(sequential structural alignment programm)

1. compute residue view of each residue:

   set of distance vectors from Cβ to Cβ of all other residues



WR Taylor & CA Orengo (1996) Meth. in Enzym. 266:617-635



cysteine

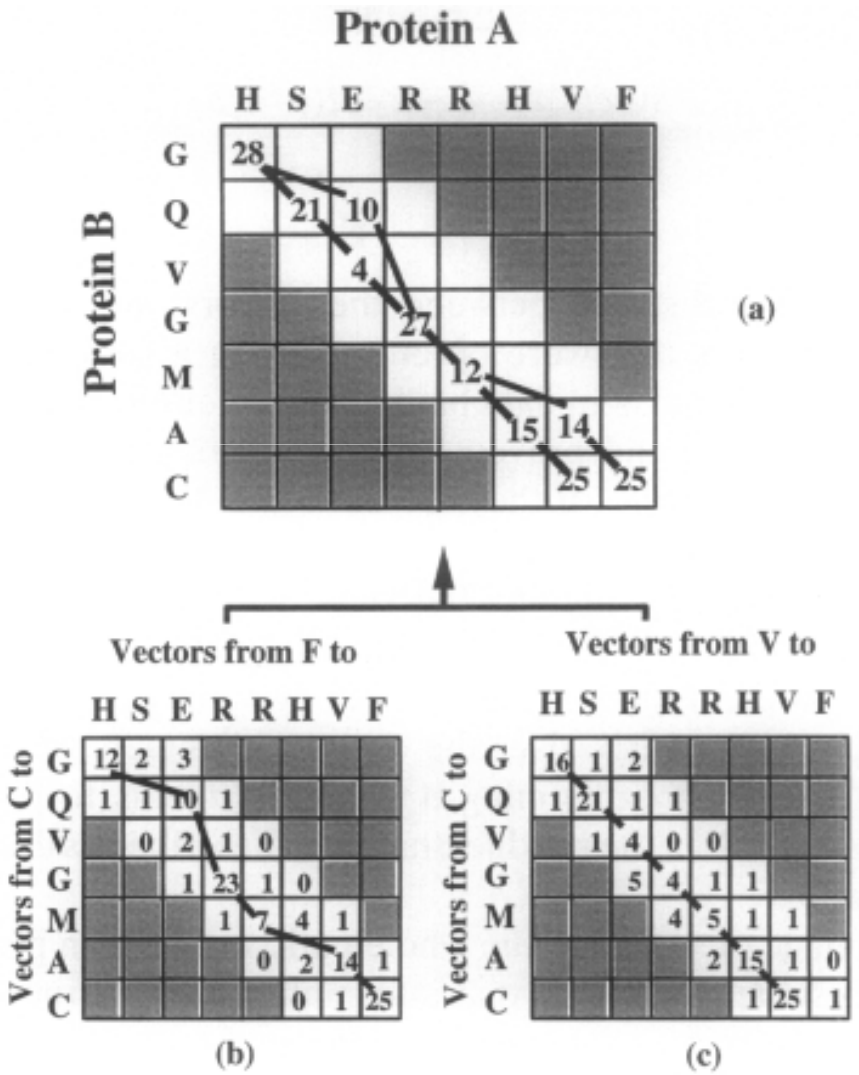
http://groups.molbiosci.northwestern.edu/holmgren/Glossary/Definitions/Def-C/Cysteine.html

very similar structures
→ compare by subtracting equivalent vectors

SSAP

2. double dynamic programming

- Align all residue views to find corresponding residues.
- Align the residues and find optimal path trough summary matrix.



WR Taylor & CA Orengo (1996) Meth. in Enzym. 266:617-635

TopMatch

- query structure Q is aligned to target structure T

→ ranked list of possible alignments

→ which are combined to composite alignment

- structures represented by Cα atoms
- Multiple chains joined to single chain

TopMatch score

- L = length of A, not counting gaps
- Root mean square error : $E_r = \sqrt{\frac{1}{L}\sum r_i^2}$
- where $r_i^2 = (x_i - y_i)^2$
- Similarity: $S = \sum_i^L e^{-r_i^2/\sigma^2}$ with scaling factor σ
- Similarity per residue: $s = \frac{S}{L} = \frac{1}{L}\sum_i^L e^{-r_i^2/\sigma^2}$
- Distance error S_r: $S_r = \sigma\sqrt{-\ln s}$

Example

- Composite alignment between

HFE protein (1A6Z,A) and

MHC class I molecule (1BII,A)

- Query: blue (orange),
- Target: green (red)

- L = 266
- S = 226
- S_r = 2.81
- E_r = 2.98

CE (combinatorial extension)

1. Compute all possible AFPs (aligned fragment pairs): confer structural similarity
   - Based on local geometry
   - Fixed size m (e.g. 8)

2. Heuristics used to define a set of optimal paths joining AFPs with gaps as needed

3. Optimization of the path with the best RMSD using dynamic programming

→ Optimal alignment

example

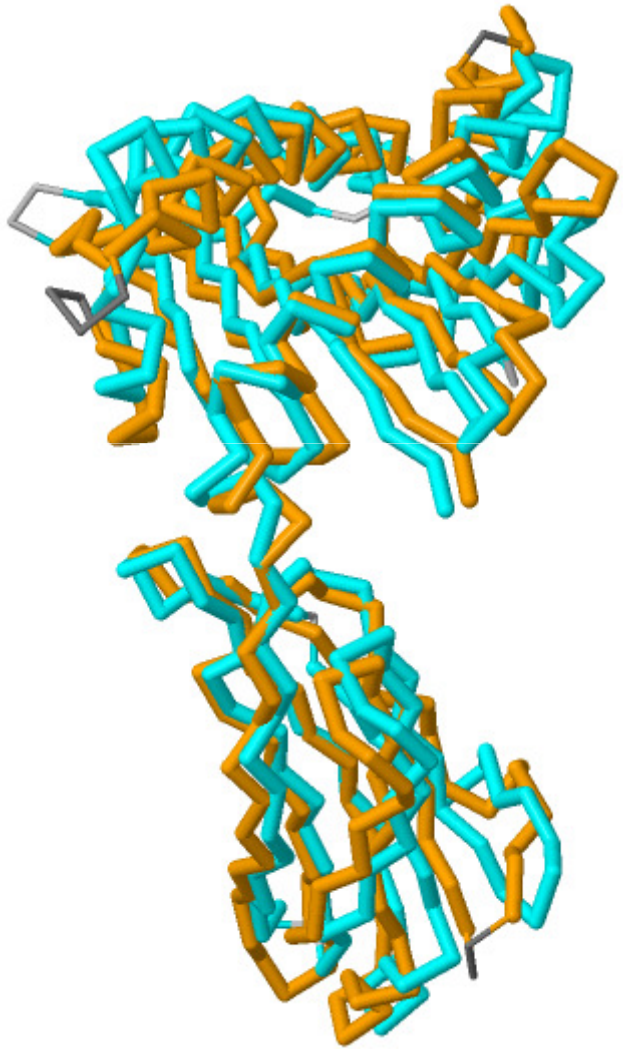
- alignment between

HFE protein (1A6Z,A) and

MHC class I molecule (1BII,A)

- HFE in orange
- MHC class I molecule in cyan

- RMSD = 2.63



LGA (local global alignment)

- Takes both local and global structure superpositions into account

- 2 methods:
  1. LCS (longest continuous segments)
  2. GDT (global distance test)

- Used to detect regions of local and global structural similarity

LCS

- Localize and superimpose the longest continuous segments of residues under a RMSD cutoff
- cutoff = 1, 2, 5 Å

- LCS_vi = % of continuous residues that can fit under RMSD cutoff vi

→ identifies local regions of similarity

GDT

- finds largest set of corresponding residues deviating no more than distance cutoff
- cutoff = 0.5, 1.0, 1.5, .., 10 Å

- GDT_vi = % of residues (largest set) that can be superimposed under the distance cutoff of vi

→ sequence continuity not maintained

→ global level of similarity

- Scoring function LGA_S combines all LCS_vi and GDT_vi

LGA – example result

- alignment between HFE protein (1A6Z,A) and

MHC class I molecule (1BII,A):

#CA	N1	N2	DIST	N	RMSD	Seq_Id	LGA_S	LGA_Q
SUMMARY(LGA)	272	274	5.0	253	2.37	36.36	63.924	10.228

- N = number of superimposed residues under distance cutoff 5 Å

Thany you.

Any questions?

References

1. http://scop.mrc-lmb.cam.ac.uk/scop/
2. A Murzin (1995), SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures , JMB 247:536-540
3. http://protein.hbu.cn/cath/cathwww.biochem.ucl.ac.uk/latest/cath_info.html
4. WR Taylor & CA Orengo (1989), Protein structure alignment, JMB 208:1-22
5. WR Taylor & CA Orengo (1996), SSAP: Sequential Structure Alignment Program for Protein Structure Comparison, Methods in Enzymology 266:617-635
6. MJ Sippl & M Wiederstein (2012), Detection of spatial correlations in protein structures and molecular complexes, Structure 20:718-728
7. MJ Sippl & M Wiederstein (2008), A note on difficult structure alignment problems, Bioinformatics 24(3):426–427

8. IN Shindyalov & P Bourne (1998), Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, Prot Engineering 1:739-47
9. http://source.rcsb.org/jfatcatserver/help.jsp
10. A Zemla (2003), LGA: a method for finding 3D similarities in protein structures, Nucleic Acids Research 31:3370-3374
11. http://proteinmodel.org/AS2TS/LGA/lga_format.html