Structural Bioinformatics
Lecture 8

Comparison of protein 3D structures.

Alignment of protein 3D structures

HIPS

UNIVERSITÄT
DES
SAARLANDES

ZBI ZENTRUM FÜR
BIOINFORMATIK

Topics of lectures updated

1. 20.10.: Introduction to properties and structure of biological macromolecules
2. 27.10.: Experimental techniques in structural biology; Protein Data Bank
3. 3.11.: Protein structural organization, classification of proteins by structure
4. 10.11.: Prediction of structural features from sequence
5. 24~~17~~.11.: Evolution and comparison of protein sequences
6. 1.12.~~24.11.~~: Advanced sequence similarity search using hidden Markov models
7. 8~~1~~.12.: Homology-based modelling of protein 3D structure
8. 15~~8~~.12.: Comparison of protein 3D structures and of models to experimental structures.
9. 5.01.: Modelling of protein 3D structure by threading
10. 12~~5~~.01.: Fragment-based prediction of protein 3D structure
11. 19~~12~~.01.: Prediction of inter-residue contacts and implications for protein 3D structure prediction
12. 26~~19~~.01.: Introduction to molecular dynamics simulations

~~13. 26.01.: Backup~~

14. 2.02.: Q&A

~~15. 9.02.: Exam, 1st attempt~~ (2nd attempt: end of March / beginning of April)

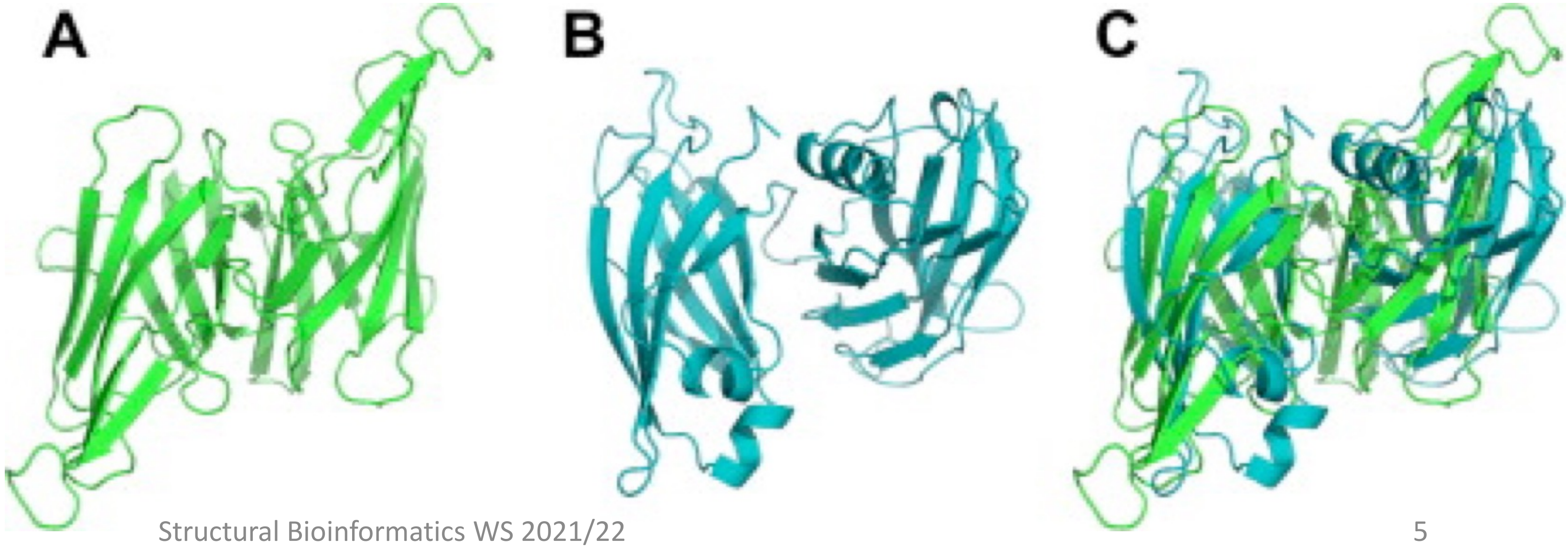
Outline

- Comparison of a structural model to an experimentally resolved 3D protein structure
- Comparison of significantly different protein 3D structure
- Alignment of protein 3D structures

Comparison of a structural model to an experimentally resolved 3D protein structure

Intuitive concept of protein structure alignment

- A **transformation** (rotation + transition)
- **Superimpose**
  - coordinates of C_α atoms of **one** proteins structure
  - coordinates of C_α atoms of **another** protein structure
- In an **optimal way**



Intuitive measure of protein structural similarity

- RMSD: **root mean square deviation** of the coordinates of Cα's of n superimposed amino acids
- If $\mathbf{v} = (v_1, \ldots v_n), v_i = (v_{ix}, v_{iy}, v_{iz})$ are the coordinates of the corresponding residues in the first structure, and $\mathbf{w} = (w_1, \ldots w_n)$ likewise from the second structure,

$$\begin{aligned}\mathrm{RMSD}(\mathbf{v}, \mathbf{w}) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|v_i - w_i\|^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}\end{aligned}$$

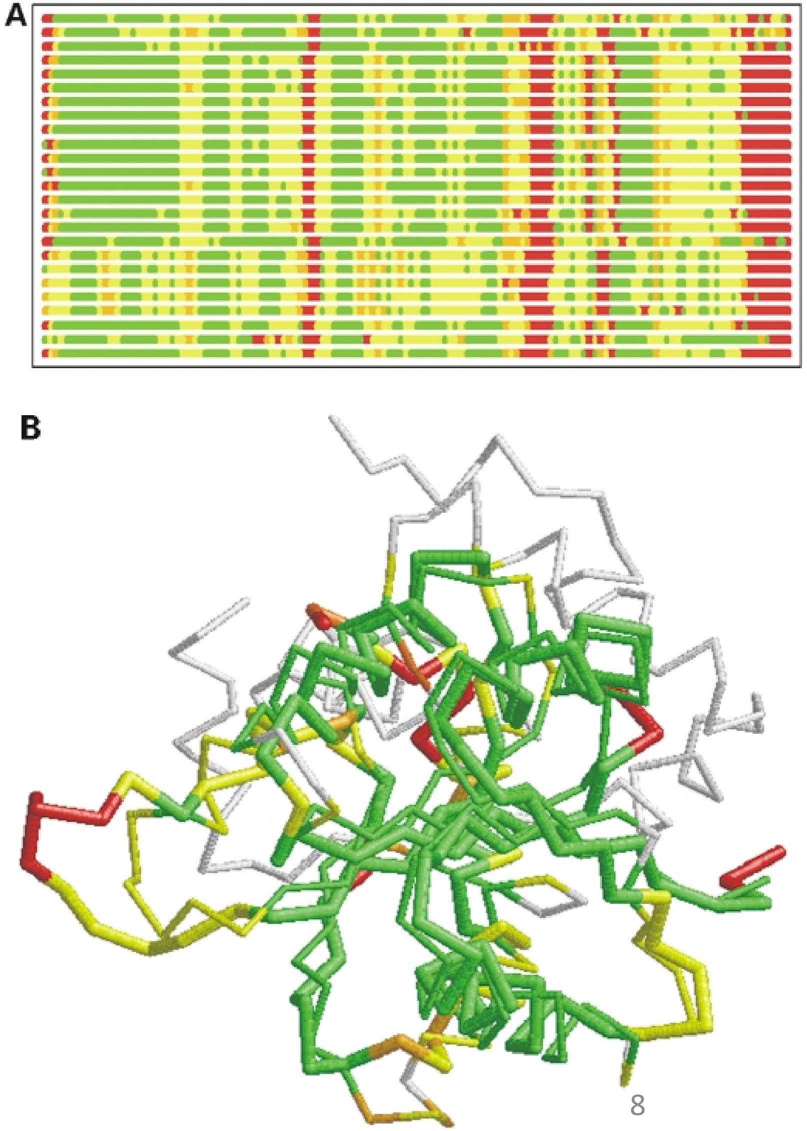
Some drawbacks of RMSD

- All atoms equally weighted => sensitive to local structure deviations
- Does not take into account the length of the alignment (the shorter the alignment, the better is RMSD)

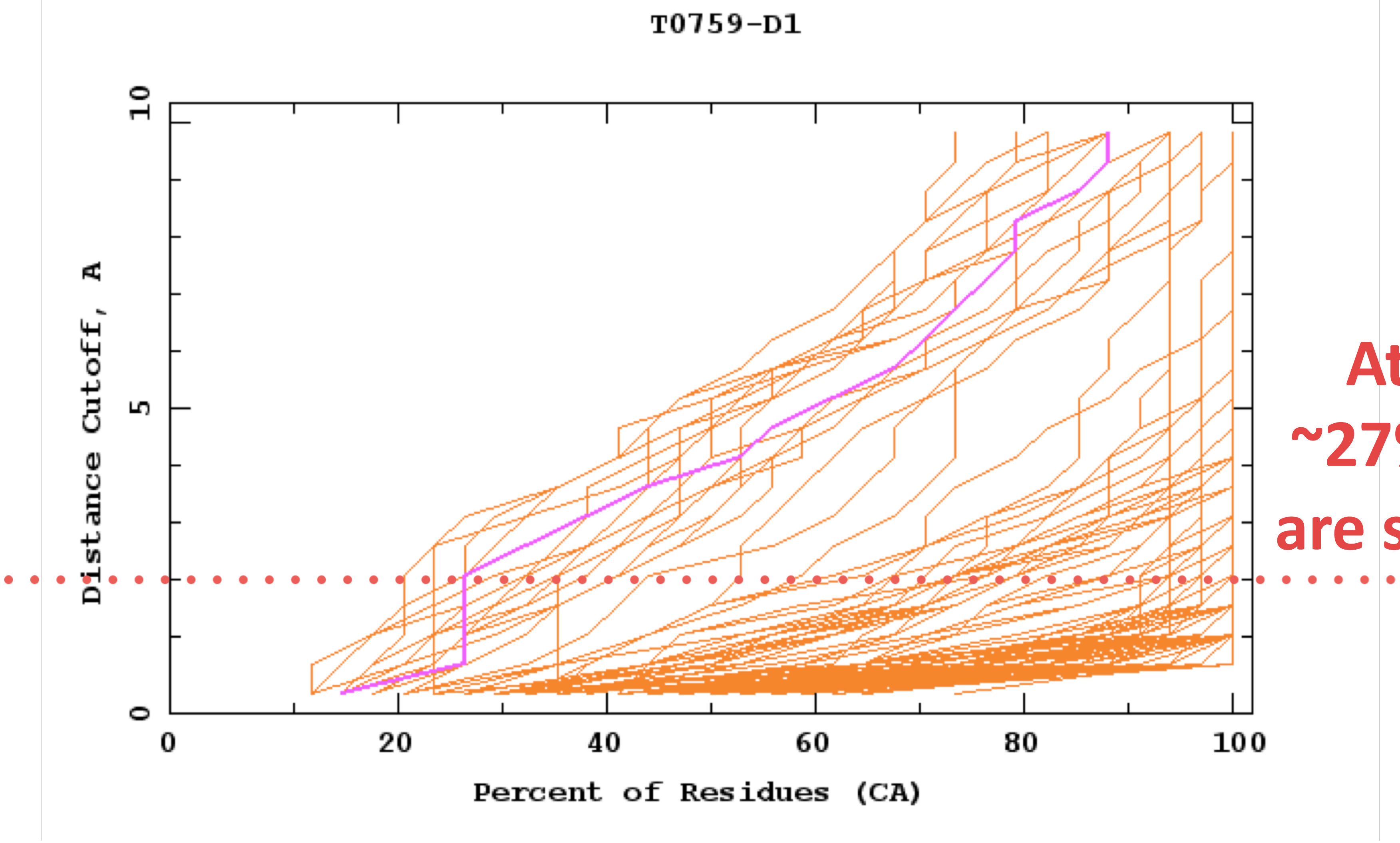
- => other measures needed that combine RMSD and alignment length

GDT_TS score (*Kryshtafovych et al., 2007, Proteins*)

- Used as the major model quality measure in CASP
- **Global Distance Test, Total Score**
    - Average proportion of Cα that have a distance <1, 2, 4, and 8 Å after optimal superimposition (trivial for models)
    - Largest set of superimposible residues is identified for every cutoff



GDT_TS plot



At 2 Å cutoff, ~27% of residues are superimposed

Comparison of significantly different protein structure

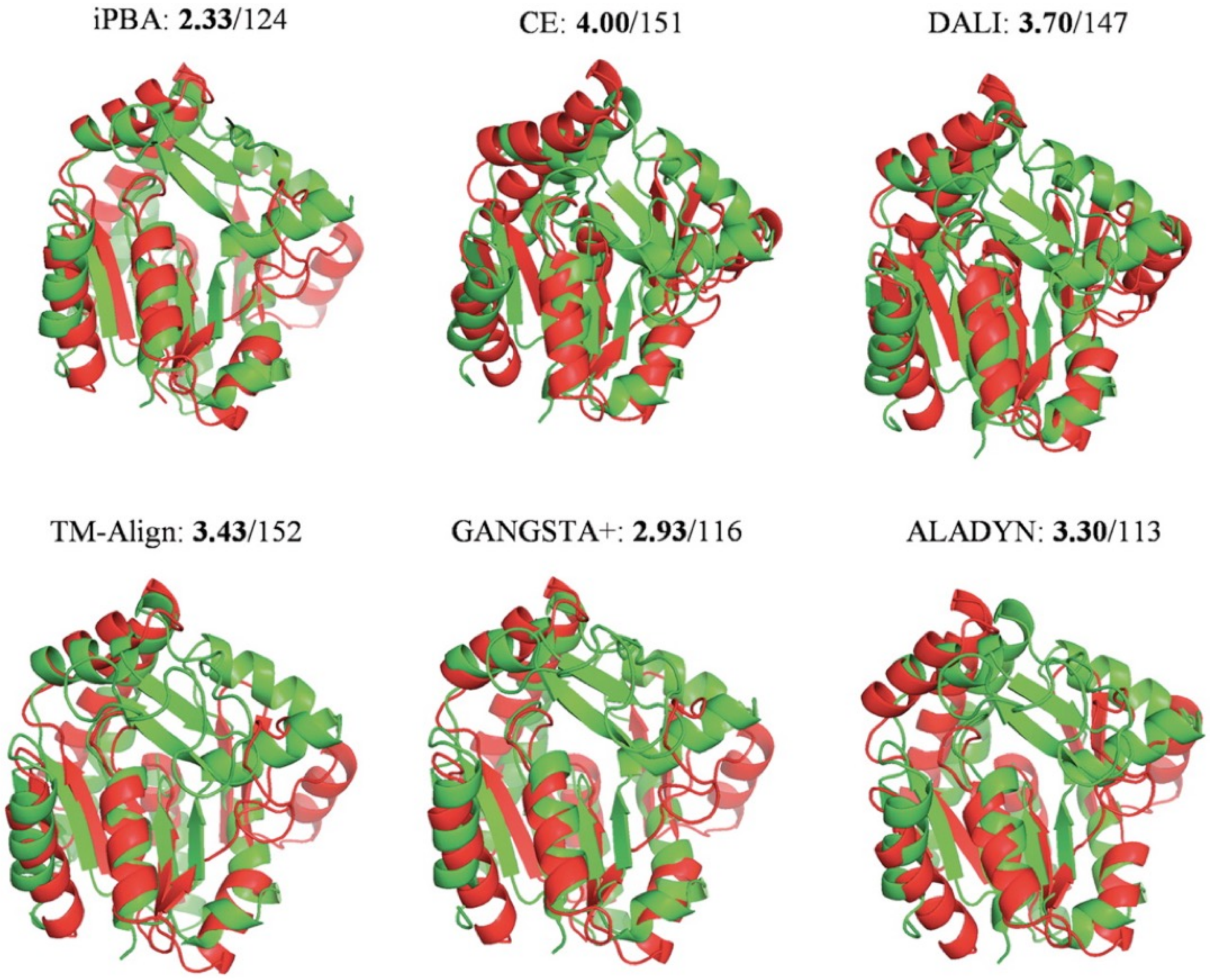
Intuitive measure of protein structural similarity

- RMSD: **root mean square deviation** of the coordinates of Cα’s of n superimposed amino acids
- If $\mathbf{v} = (v_1, \dots v_n), v_i = (v_{ix}, v_{iy}, v_{iz})$ are the coordinates of the corresponding residues in the first structure, and $\mathbf{w} = (w_1, \dots w_n)$ likewise from the second structure,

$$\begin{aligned} \mathrm{RMSD}(\mathbf{v}, \mathbf{w}) &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|v_i - w_i\|^2} \\ &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} ((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)} \end{aligned}$$

Some properties of RMSD

- The shorter are the superimposed segments, the better is RMSD
- A tradeoff between RMSD and the alignment length (length of the superimposed segments)



iPBA: **2.33**/124

CE: **4.00**/151

DALI: **3.70**/147

TM-Align: **3.43**/152

GANGSTA+: **2.93**/116

ALADYN: **3.30**/113

Why “significantly different” is important?

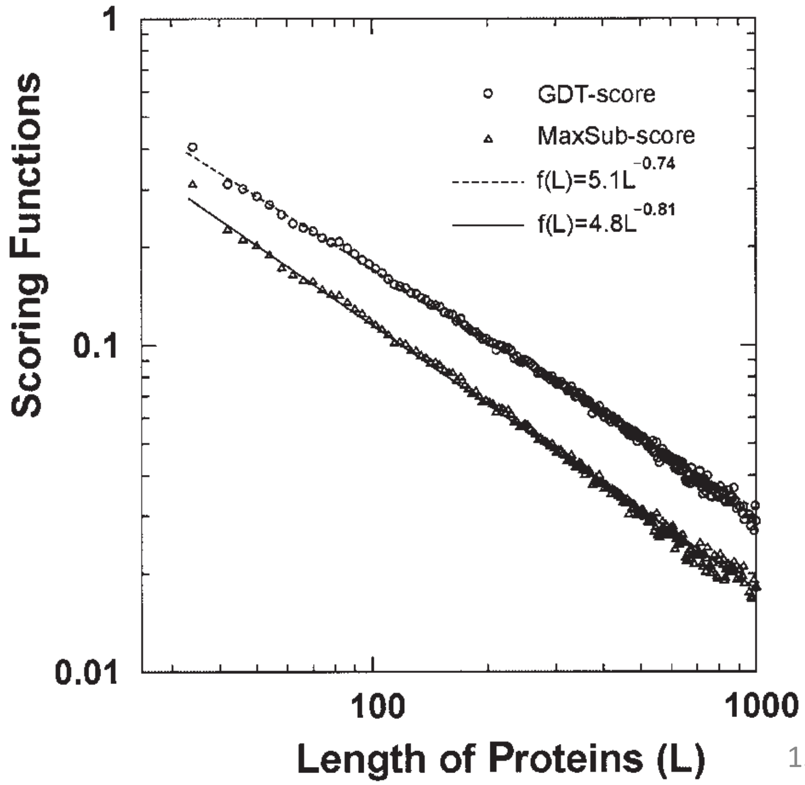
- Comparing 3D structures of very similar proteins is somewhat trivial
  - Align sequences
  - Based on alignment, create a superimposition that minimizes cumulative distance between Cα’s
  - Report that minimum
- …E.g. for comparing a 3D model to an experimentally resolved 3D structure, the alignment is straightforward
  - **Yet, this is not the whole story, cf. CASP (lecture 10)**

Structure comparison vs. alignment

- Like for sequences, it‘s **not the same thing!**
- **Comparison**: two protein are similar in 3D
- **Alignment**: mapping between amino acid residues
- Alignment $\Rightarrow$ comparison, comparison $\nRightarrow$ alignment
- However, (all) comparison tools work by constructing alignment first

MaxSub score (*Siew et al., 2000, Bioinformatics*)

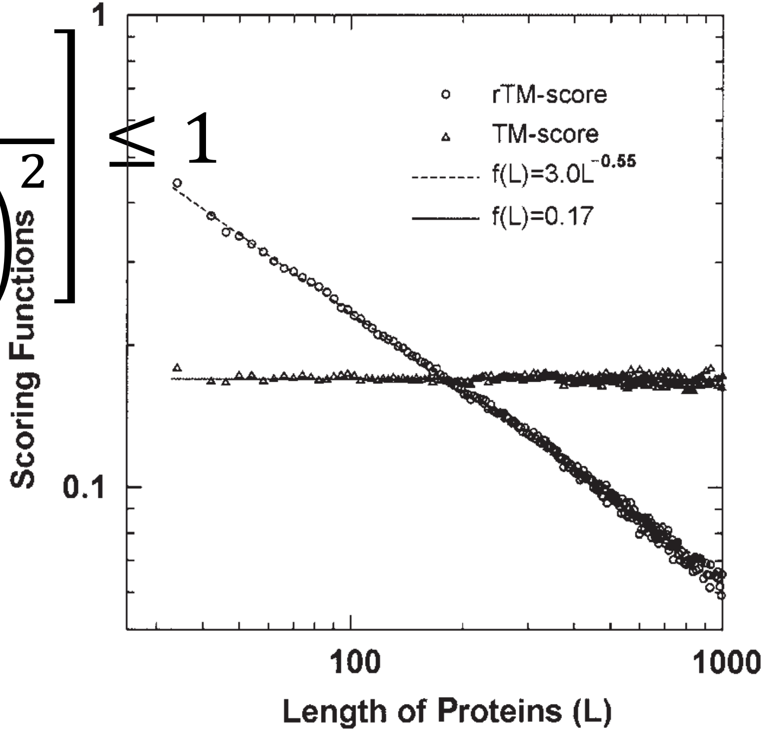
- Size of the largest substructure that can be superimposed with an RMSD under a threshold (3.5 Å by default) divided by protein length
- Non-continuous segments
- Similar in spirit to GDT_TS
- **Depends on the protein length** (GDT_TS as well)



TM-score (Zhang and Skolnik, 2004, *Proteins*)

- Eliminate *ad hoc* cutoffs => Sum over all aligned residue pairs:

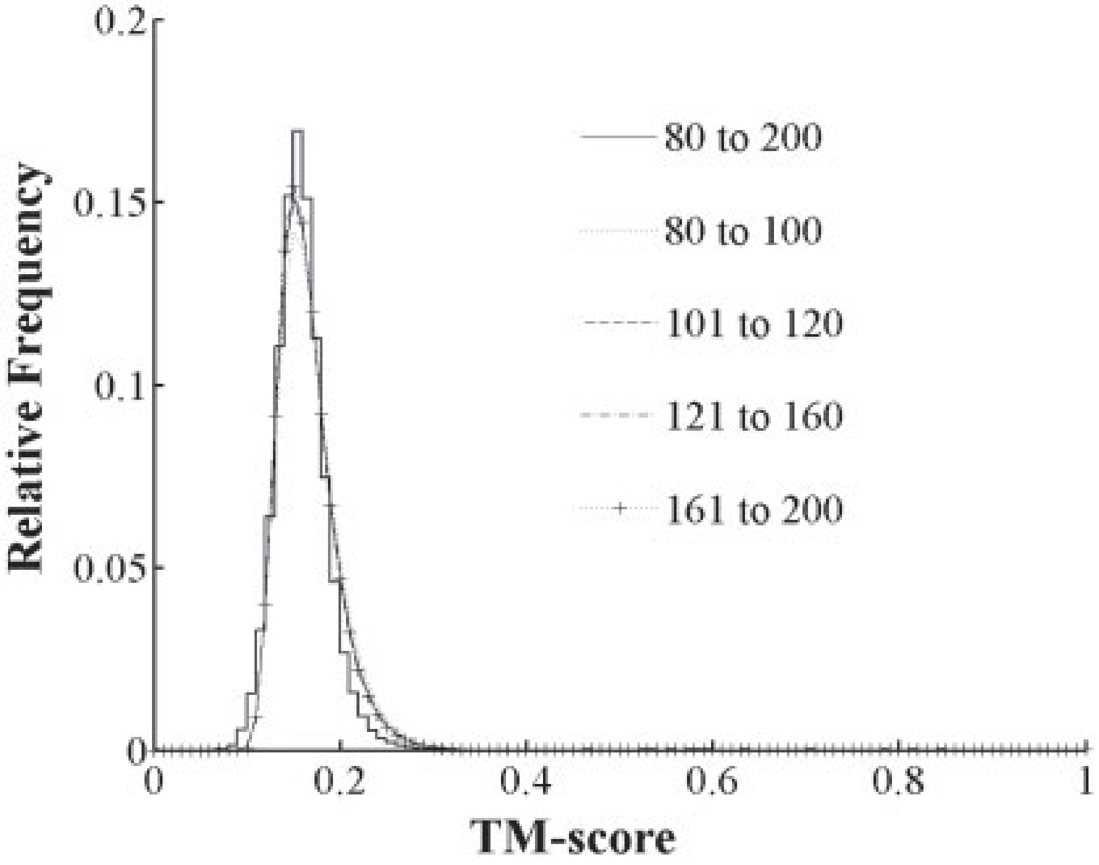
$$0 \leq \mathrm{TM{-}score} = \max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \leq 1$$



  - L_N: length of the sequence in structure,
    L_T: number aligned residues,
    d_i: distance between the *i*-th aligned pair,
    d_0 normalization factor
- Designed for model-template comparison => **non-symmetric**
- Get rid of length dependence =>Flexible $d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8$

TM-scores agree well with evolutionary relatedness (Xu and Zhang, 2010*, Bioinformatics*)

- TM-score is distributed following the extreme value distribution on non-homologous superimposable protein structure (Topology level in CATH, Fold in SCOP)
- Can be converted to P-values: **significance** of a given observed TM-score



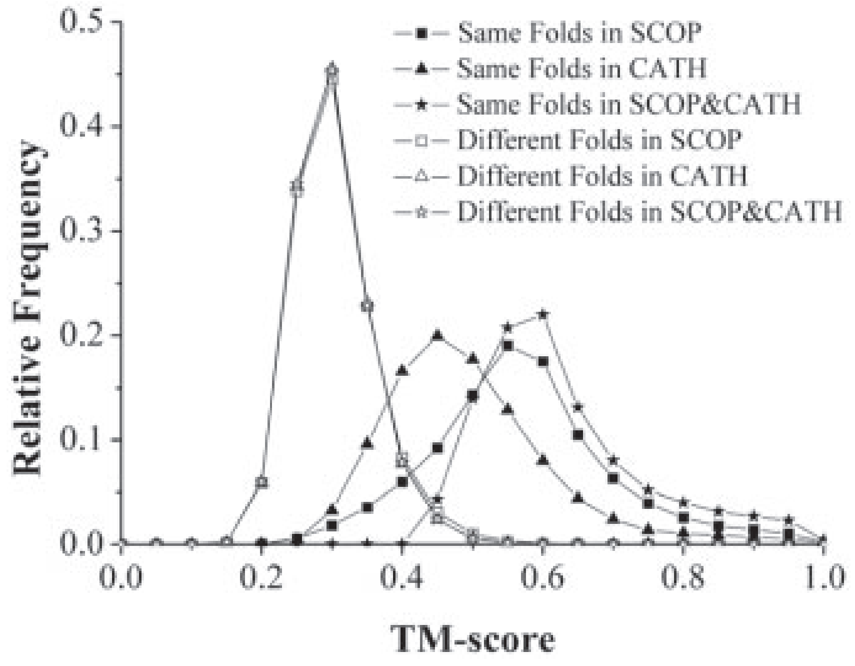
TM-scores agree well with evolutionary relatedness

- It is possible to define probability of two protein structures with a given TM-score to **belong to the same fold ($P(F|\mathrm{TM})$) and different folds ($P(\overline{F}|\mathrm{TM})$)**

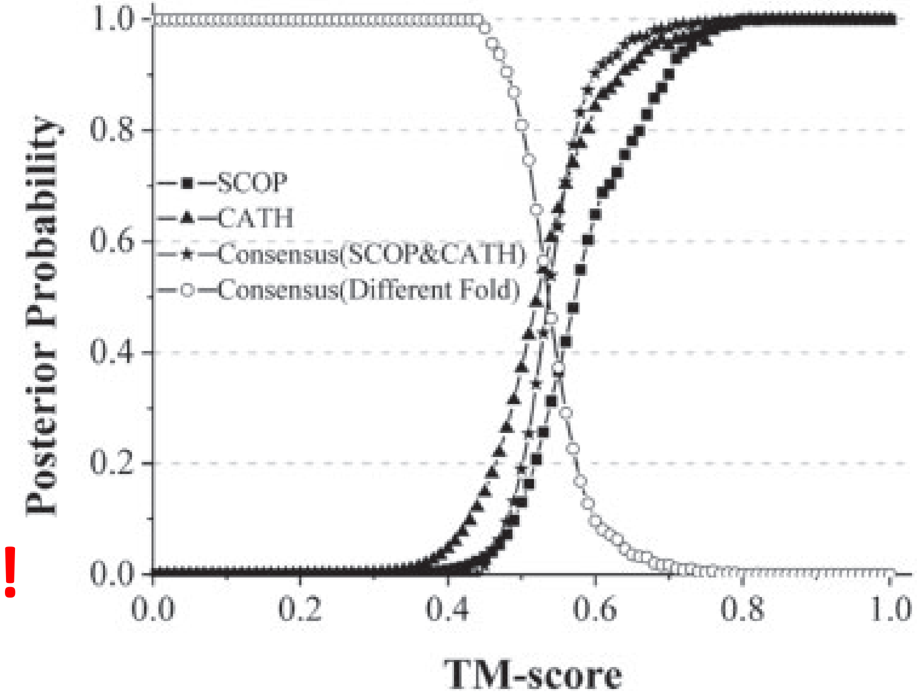
$$\begin{cases} P(F|\mathrm{TM}) = \dfrac{P(\mathrm{TM}|F)P(F)}{P(\mathrm{TM}|F)P(F)+P(\mathrm{TM}|\overline{F})P(\overline{F})} \\ P(\overline{F}|\mathrm{TM}) = \dfrac{P(\mathrm{TM}|\overline{F})P(\overline{F})}{P(\mathrm{TM}|F)P(F)+P(\mathrm{TM}|\overline{F})P(\overline{F})} \end{cases}$$

$$\begin{cases} P(\mathrm{TM}|F) = \dfrac{N(\mathrm{TM})}{\sum N(\mathrm{TM})} & \text{\# pairs with a certain TM-score within the same fold} \\ P(\mathrm{TM}|\overline{F}) = \dfrac{\overline{N}(\mathrm{TM})}{\sum \overline{N}(\mathrm{TM})} & \text{\# pairs with a certain TM-score in different folds} \end{cases}$$

$$\begin{cases} P(F) = \dfrac{N(F)}{N(F)+N(\overline{F})} \\ P(\overline{F}) = 1 - P(F) \end{cases}$$

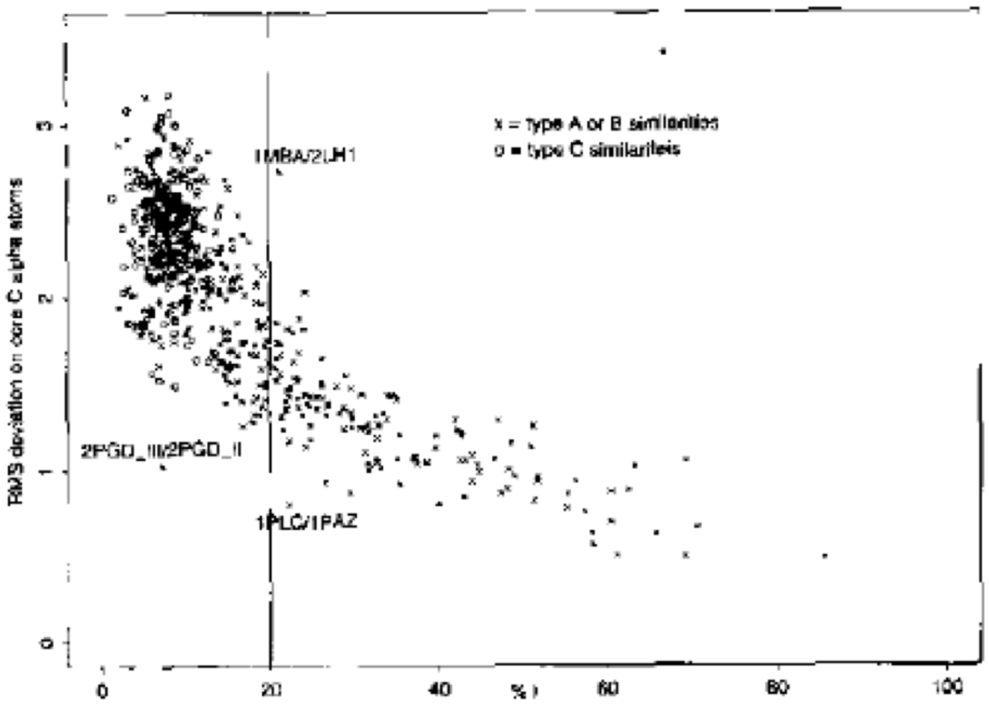


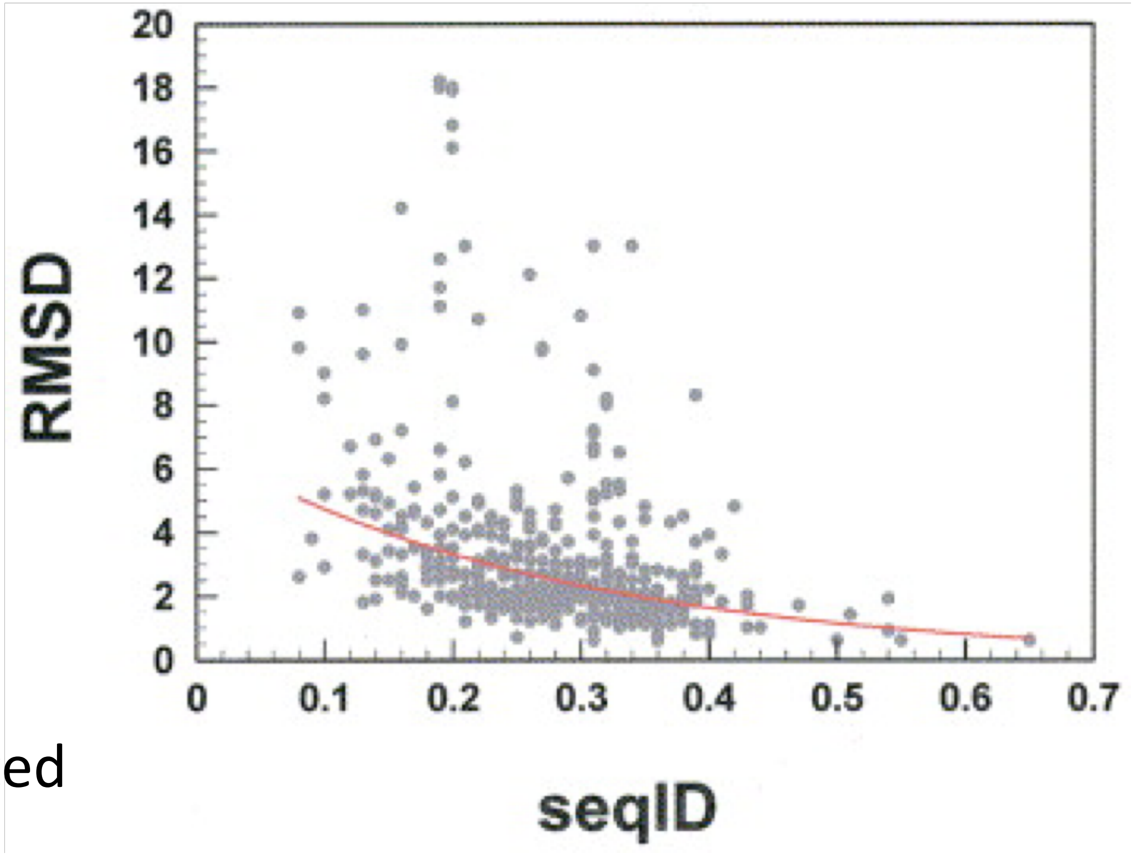
TM-score = 0.5 is a great cutoff!



Alignment of protein 3D structures

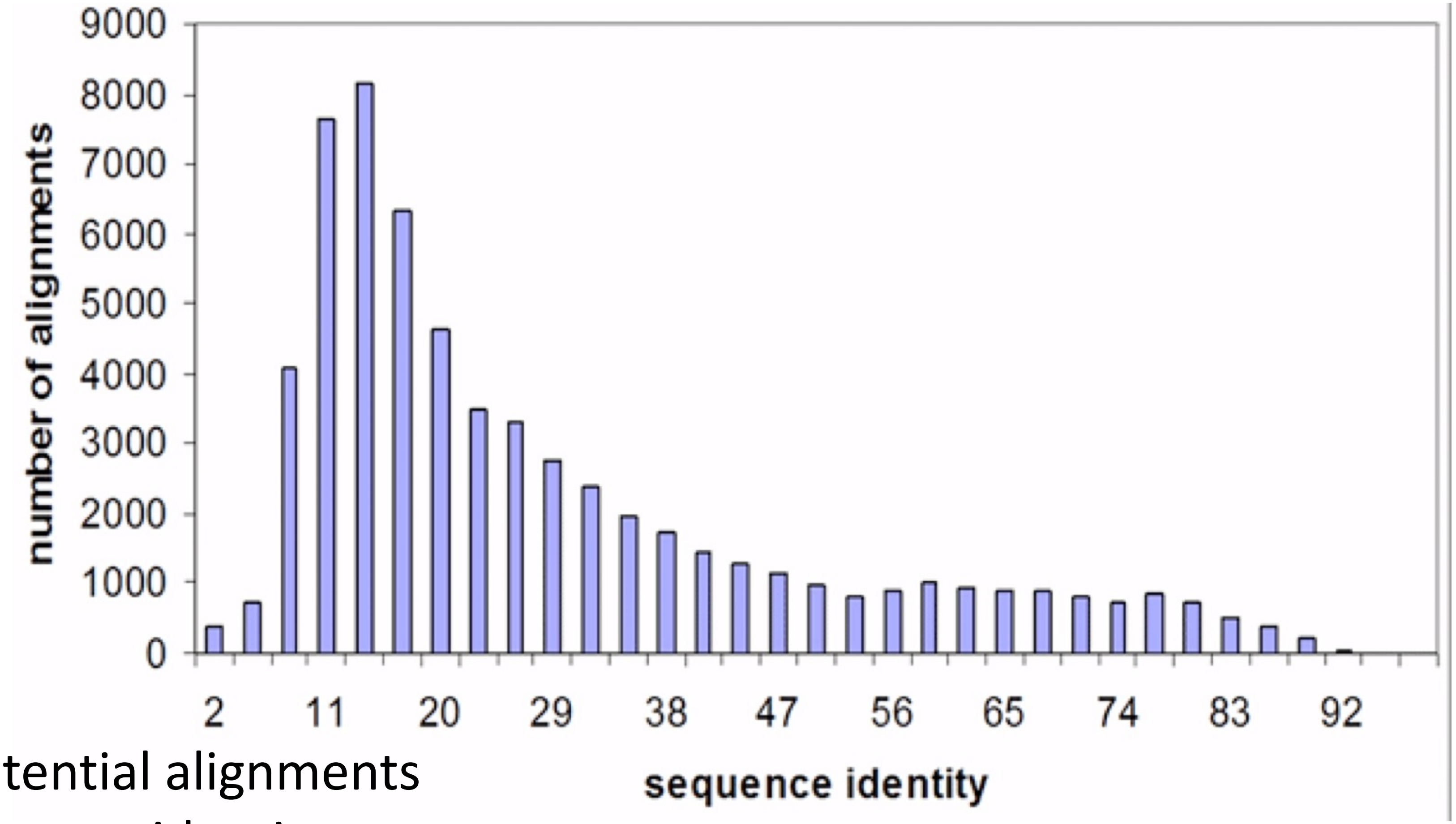
Why bother?





Protein structure conservation is correlated with sequence conservation

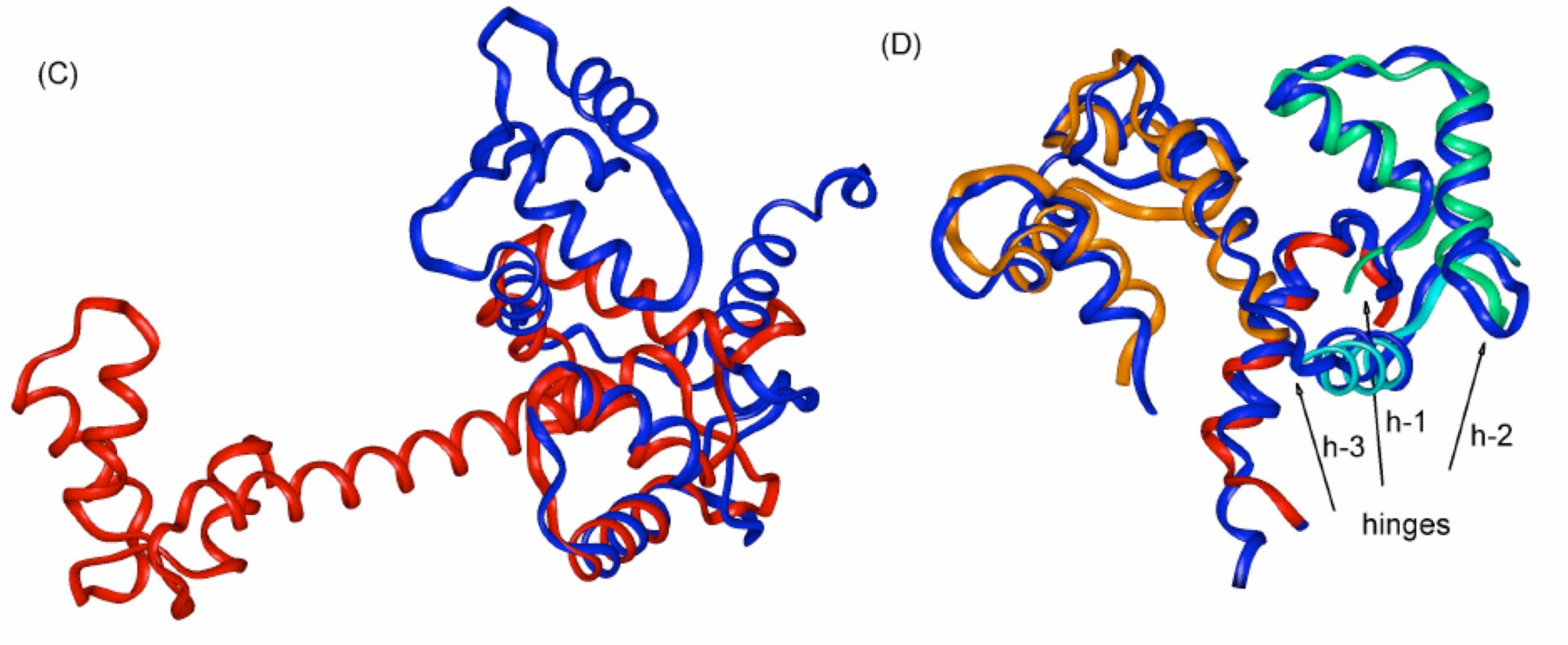
Why bother?



A lot of potential alignments
at low sequence identity

Rigid vs. flexible structural alignment

- Rigid body aligners: perform a rigid body transformation that optimizes RMSD/coverage tradeoff
- Flexible aligners: can introduce hinges/twists/structure breaks



Structural alignment tools

- DALI (http://ekhidna2.biocenter.helsinki.fi/dali/)
- TMalign (https://zhanggroup.org/TM-align/)
- SSAP (http://v3-4.cathdb.info/cgi-bin/SsapServer.pl)
- VAST (http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml)
- FATCAT (http://fatcat.burnham.org/): seems to be the only one capable of introducing hinges
- MAMMOTH (http://ub.cbm.uam.es/software/online/mamothmult.php): can do multiple alignments
- SALIGN (http://modbase.compbio.ucsf.edu/salign-cgi/index.cgi): based on sequence alignment; also can do multiple alignments
- More in Proteopedia: https://proteopedia.org/wiki/index.php/Structure_superposition_tools

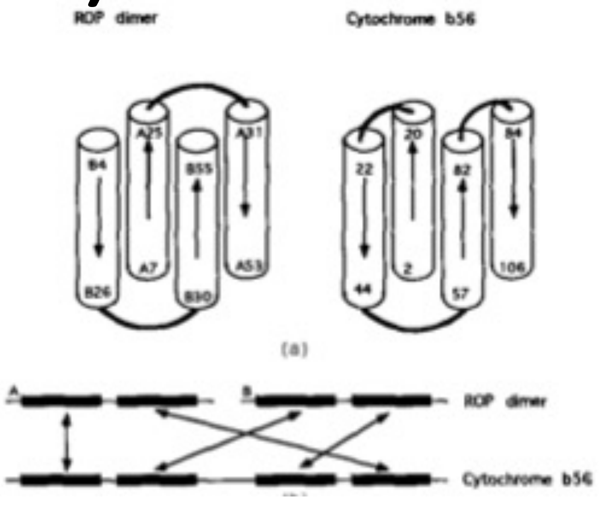
DALI (Holm and Sander, *J Mol Biol* 1993)

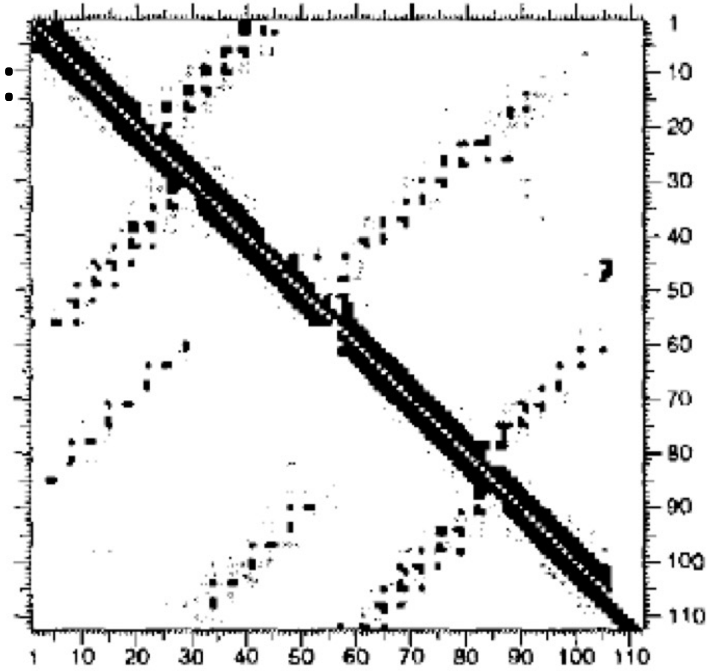
- http://ekhidna2.biocenter.helsinki.fi/dali/: structure search and pairwise comparison
- Based on pairwise distances

1. For two structures, find a set of equivalent residue pairs
   - Exhaustive with certain filters
2. Maximize a similarity measure for structures A and B:

$$S = \sum_{i=(i_A,i_B)=1}^{L} \sum_{j=(j_A,j_B)=1}^{L} \phi(i,j)\,,$$

where $\phi(i,j)$ is a distance-based similarity measure (Cα distance are taken into account)





DALI (Holm and Sander, *J Mol Biol* 1993)

Similarity between structures A and B: $S = \sum_{i=(i_A,i_B)=1}^{L} \sum_{j=(j_A,j_B)=1}^{L} \phi(i,j)$

Residue-pair score: $\boldsymbol{\phi(i,j)} = \begin{cases} \left(\boldsymbol{\theta} - \frac{\left|\boldsymbol{d_{ij}^A - d_{ij}^B}\right|}{\boldsymbol{d_{ij}^*}}\right) \boldsymbol{w\left(d_{ij}^*\right)}, \textbf{if } \boldsymbol{i \neq j}, \\ \boldsymbol{\theta}, \textbf{if } \boldsymbol{i = j} \end{cases}$

d_{ij}^A, d_{ij}^B: distances between pair i, j in A, B; d_{ij}^*: average of d_{ij}^A and d_{ij}^B

$w(r) = \exp(-r^2/\alpha^2)$: envelope function, $\alpha = 20$ Å: size of a domain

$\theta = 0.2$: zero-level similarity threshold

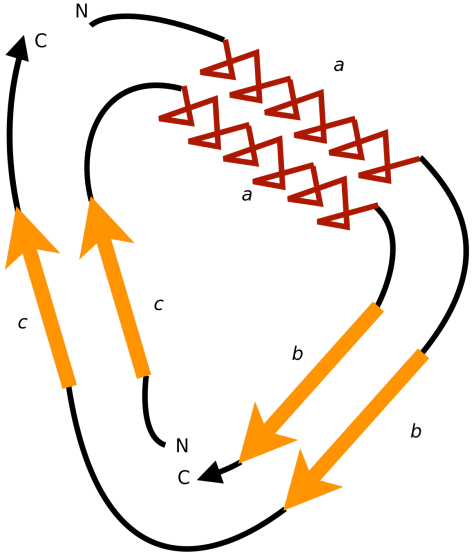
TMalign (Zhang and Skolnick, *Nucleic Acids Res* 2005)

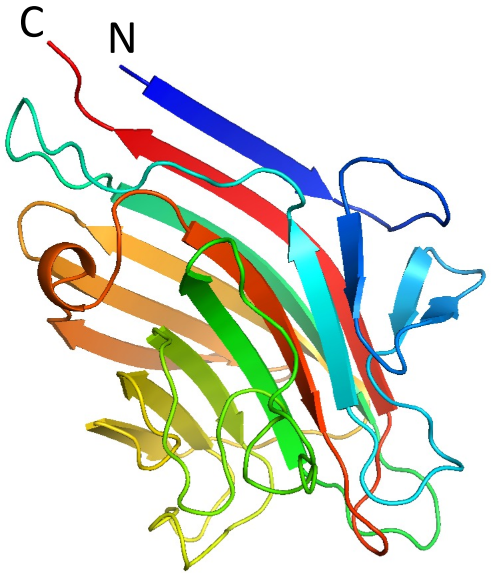
$$0 \leq \text{TM}-\text{score} = \max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \leq 1$$

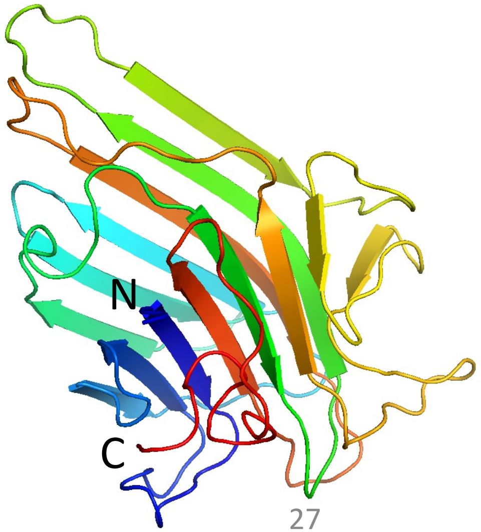
- Only distances between Cα atoms are considered (as in DALI)
- 3 rounds of dynamic programming (DP):
  1. Alignment of secondary structure elements (only exact matches)
  2. Gapless matching (*threading*) of the smalles structures against the larger structure while optimizing the TM-score
     - Only Cα coordinates are considered => **sequence-independent**
  3. Same as (2) with allowed gaps and a mixture score (Cα coordinates + secondary structure)
- Heuristic structure refinement (in theory, NP-hard): $S(i,j) = \frac{1}{1+\left(d_{ij}/d_0\right)^2}$: a new similarity matrix for DP

Circular permutations in protein structures

- **Circular permutation (CP)**: changed order of amino acids / secondary structure elements in the sequence
- DALI **can** account for CPs by design
- TMalign should not account for CPs, but a heuristic(?) was added in 2019, so it **can**
- Example: 2pel and 3cna
(two plant lectins, sugar-binding proteins)







Summary and possible exam questions

- Different measure of similarity of protein 3D structures:
  - RMSD
  - GDT_TS
  - TM-score
- Idea behind DALI protein structure alignment
- Idea behind TMalign
- Are these methods sequence-dependent?
- What about circular permutations in the aligned proteins?