BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

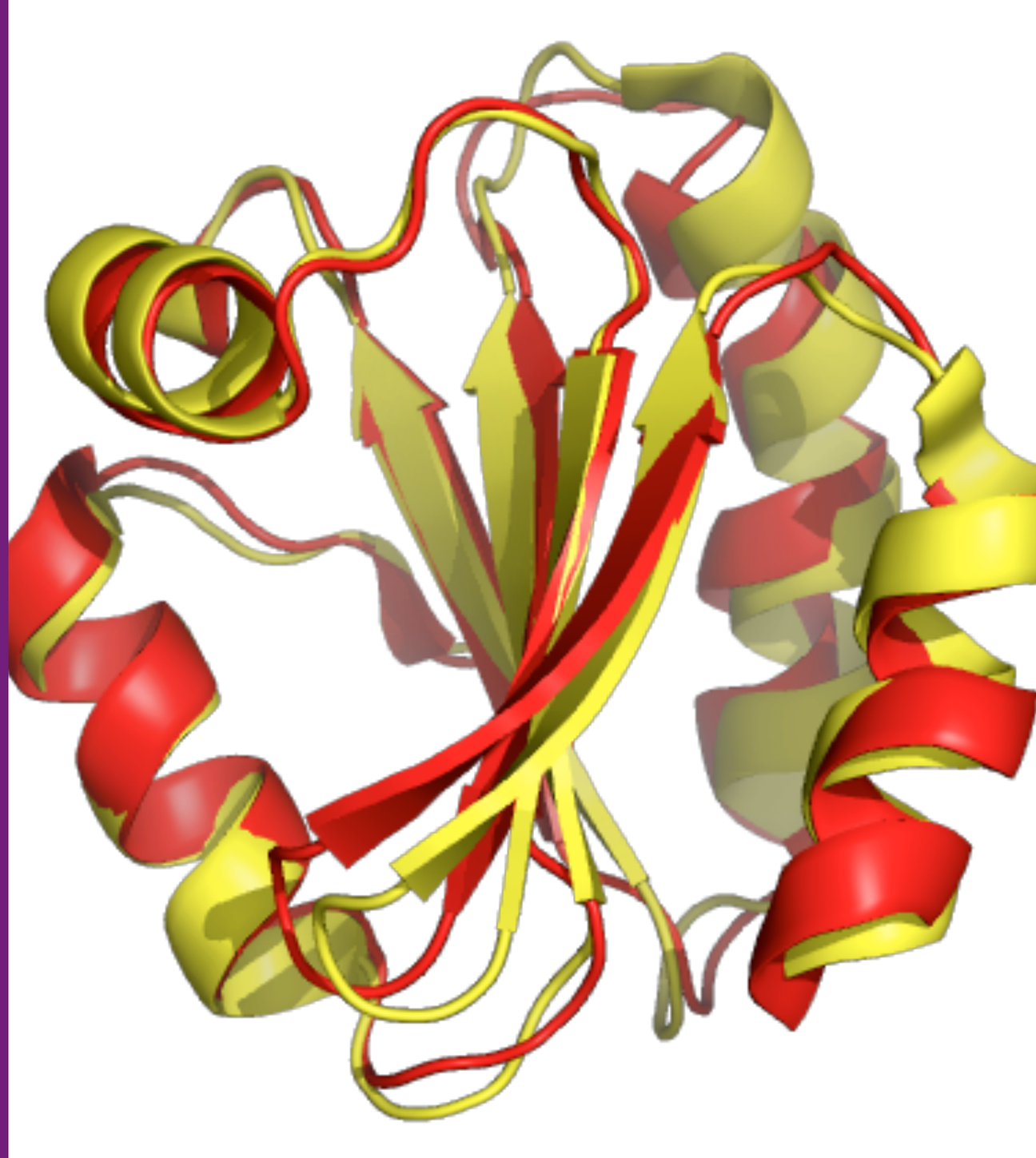
MPCS56420
SESSION 6

THE UNIVERSITY OF CHICAGO

PROTEIN STRUCTURE ALIGNMENT

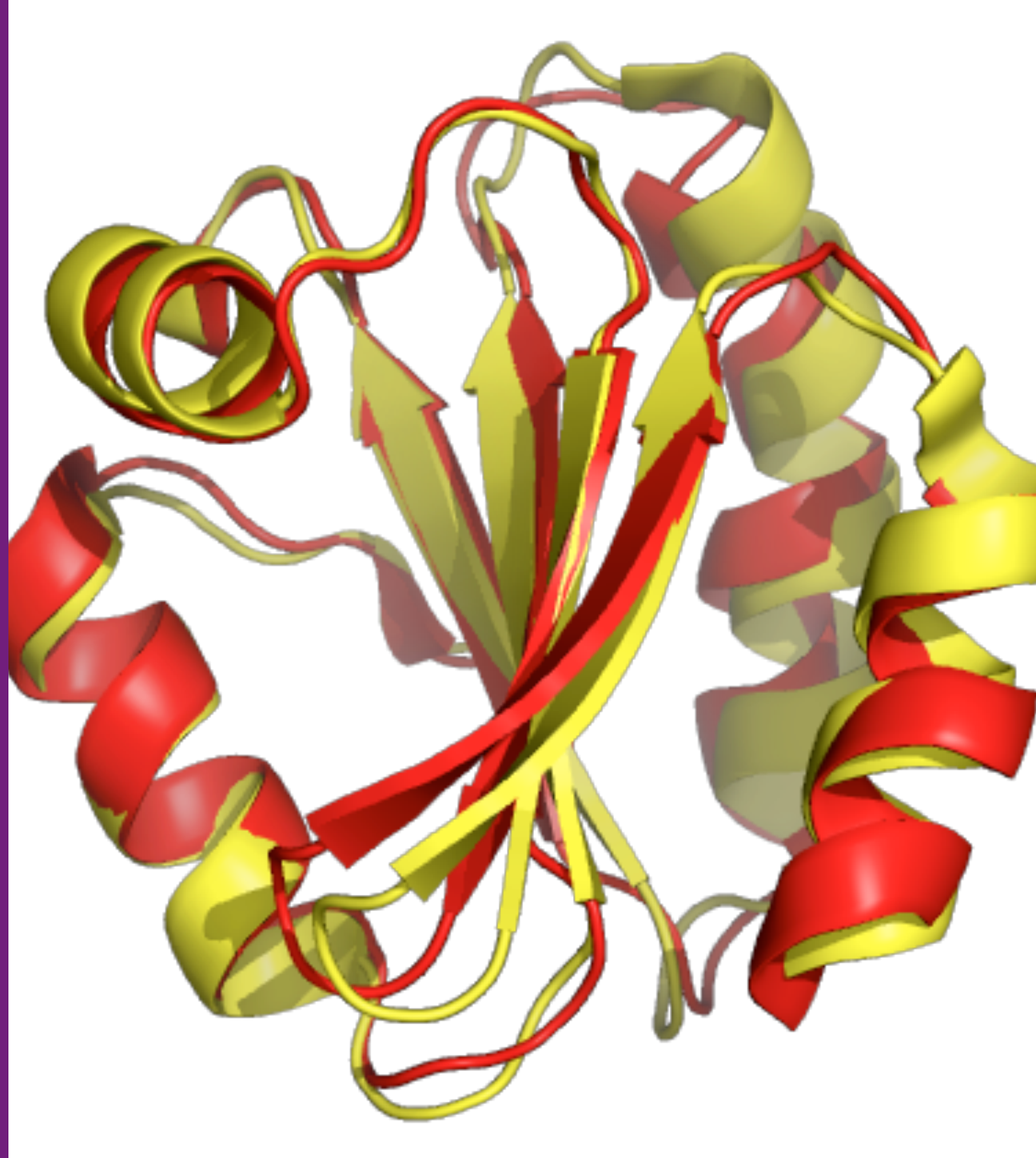
PROTEIN STRUCTURE ALIGNMENT

- Structural alignment
  - Attempts to establish homology based on their 3D structure
  - Requires no a priori knowledge of equivalent positions
  - Will detect similarity in absence of sequence similarity
    - Shared motifs (e.g helix-turn-helix for DNA binding)



PROTEIN STRUCTURE ALIGNMENT

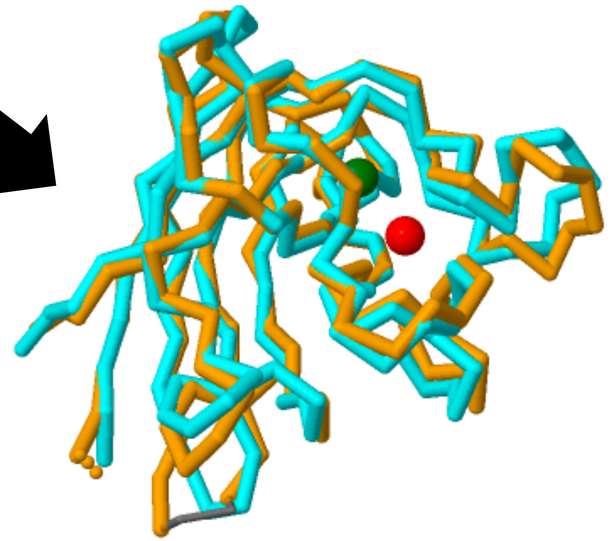
- Structural superposition
  - Uses knowledge of at least some equivalent residues to guide a rigid body superposition
  - Requires a pre-calculated alignment as input to determine which of the residues in the sequence to use
    - Sequence alignments
    - Conserved residues



PROTEIN STRUCTURE ALIGNMENT

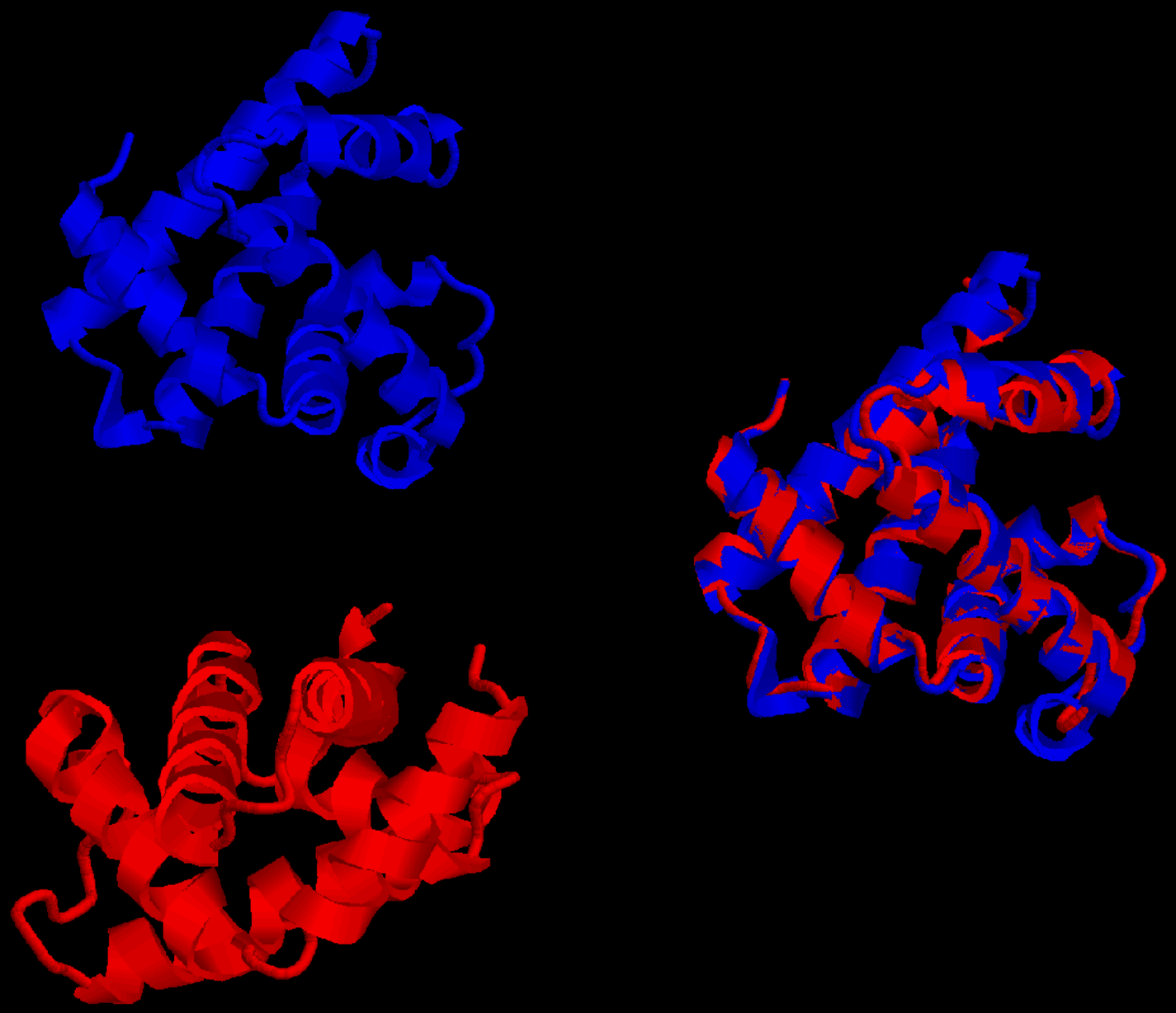
- Straightforward superposition based on sequence alignment
  - Not always useful, can dominate (and obscure remote) relationships
- Due to computational complexity, most structural alignments are pairwise, but multiple alignment methods do exist
  - Global and local approaches

```
Align 1SPD.B.pdb Length1: 153 with PDP:3F7LAa.pdb Length2: 151
Z-score 6.58
Equ: 151
RMSD: 1.25
Score: 296.75
Align-len: 153
Gaps: 2 (1.31%)

Identity: 61.59%
Similarity: 74.83%

1:B       10:B      20:B      30:B      40:B      50:B      60:B      70:B
|         |    .    |    .    |    .    |    .    |    .    |    .    |
ATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHVHEFGDNTAGCTSAGPHFNPLSRK
|..|||||||.||.|.|...:..  ..|.|.|.|.|||.|.||||||||||||.||||||.||||..:.
AIHAVCVLKGDSPVTGTIHLKEEG--DMVTVTGEITGLTPGKHGFHVHEFGDNTNGCTSAGGHFNPHGKE
|         |    .    |      .    |    .    |    .    |    .    |    .
1:A       10:A      20:A        30:A      40:A      50:A      60:A


          80:B      90:B      100:B     110:B     120:B     130:B     140:B
    .     |    .    |    .    |    .    |    .    |    .    |    .    |
HGGPKDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHCIIGRTLVVHEKADDLGKGGNEESTKTGNA
||.|.||.||.||||||.|..||.|.:...|..:.|.|...:||||||||...||||:||.|:|..||||
HGAPEDENRHAGDLGNVVAGEDGKAVINMKDKLVKLTGPDSVIGRTLVVHVDEDDLGRGGHEQSKITGNA
 |    .    |    .    |    .    |    .    |    .    |    .    |    .
  70:A      80:A      90:A      100:A     110:A     120:A     130:A


          150:B
    .     |
GSRLACGVIGIAQ
|.|||||||||..
GGRLACGVIGITK
 |    .    |
```



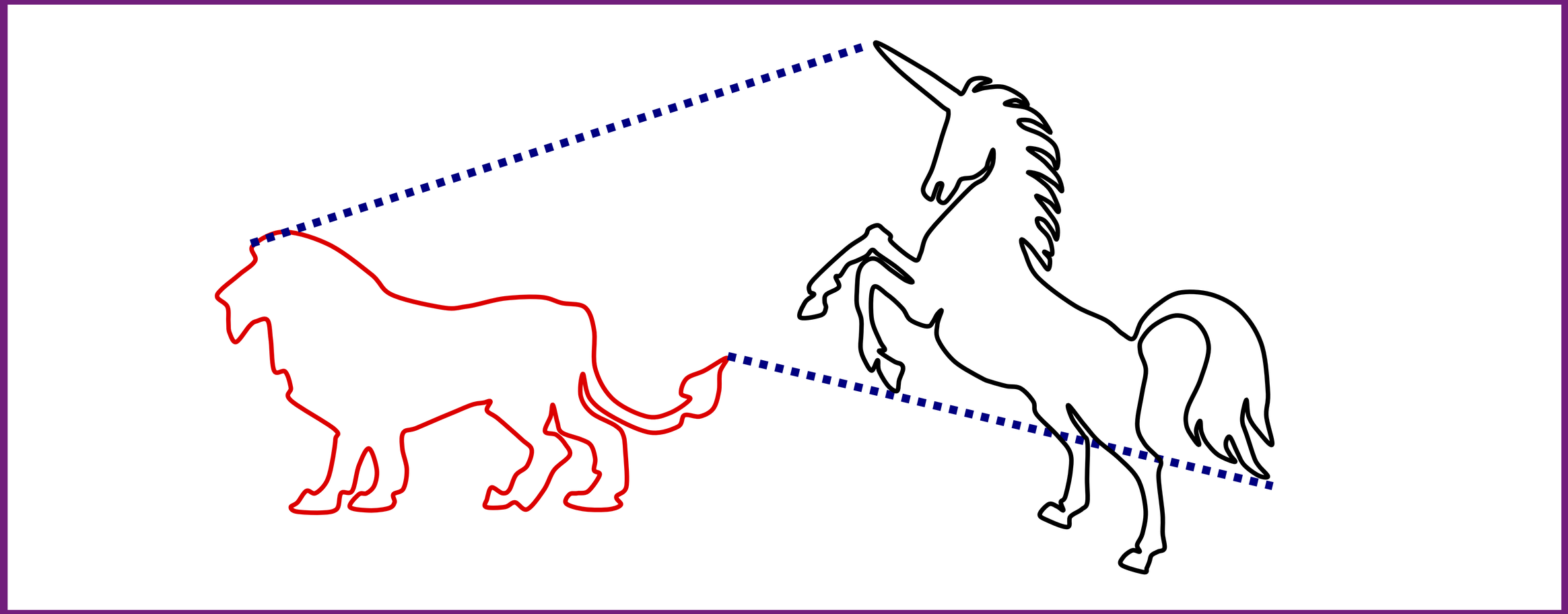
PROTEIN STRUCTURE ALIGNMENT

- Human Hemoglobin alpha-chain
  - pdb:1jebA
- Human Myoglobin
  - pdb:2mm1



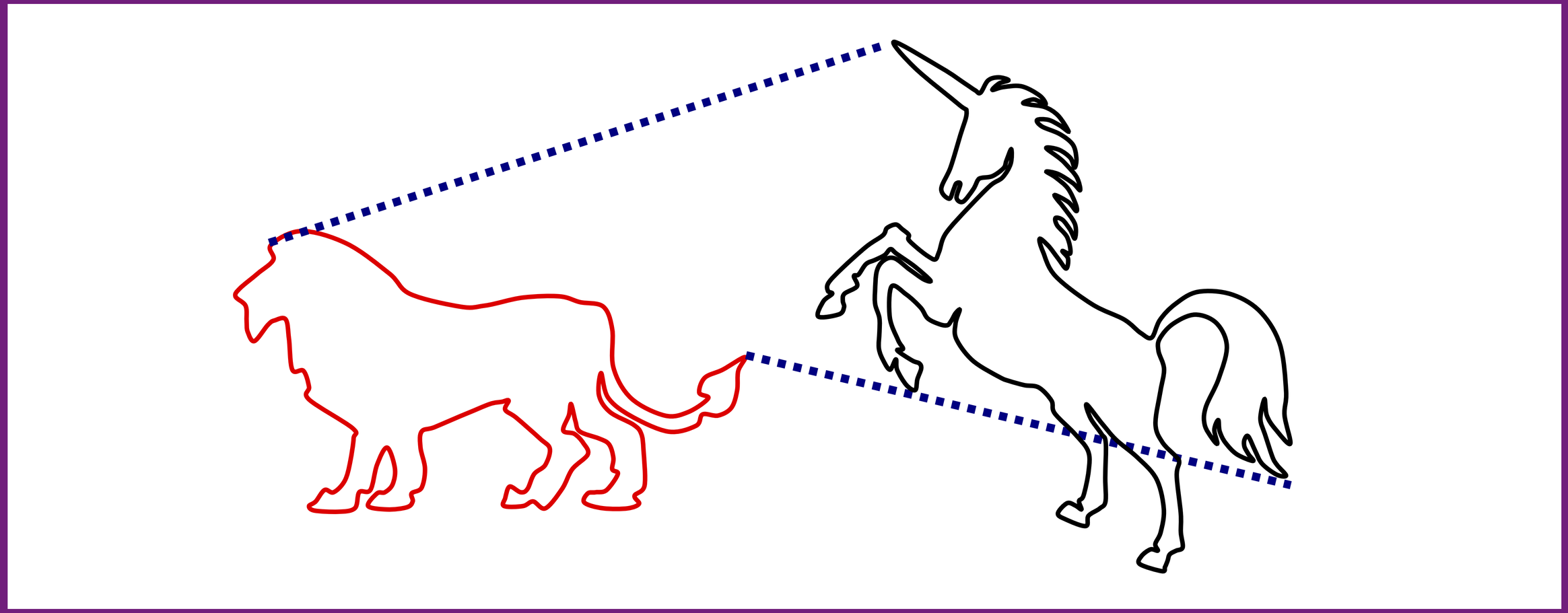
Alignment Transformations

TRANSFORMATIONS



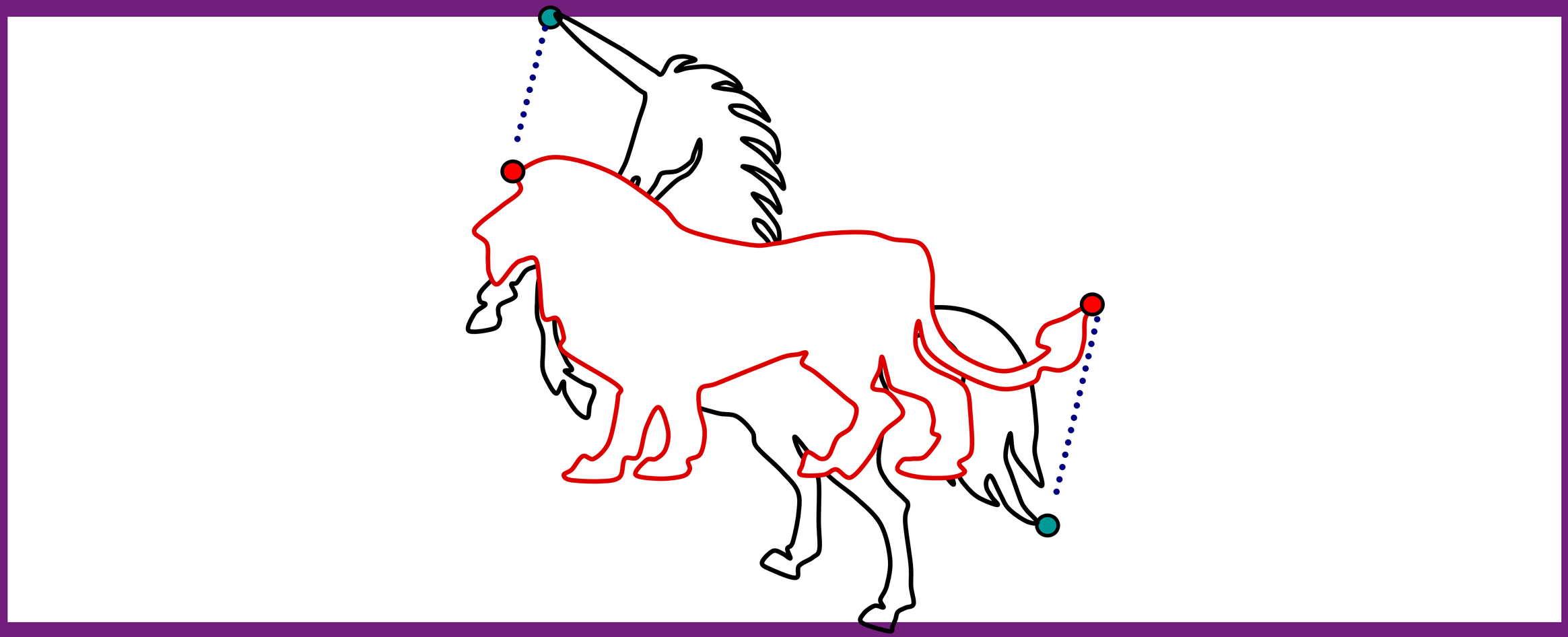
- What is the best transformation that superimposes the unicorn on the lion?

TRANSFORMATIONS



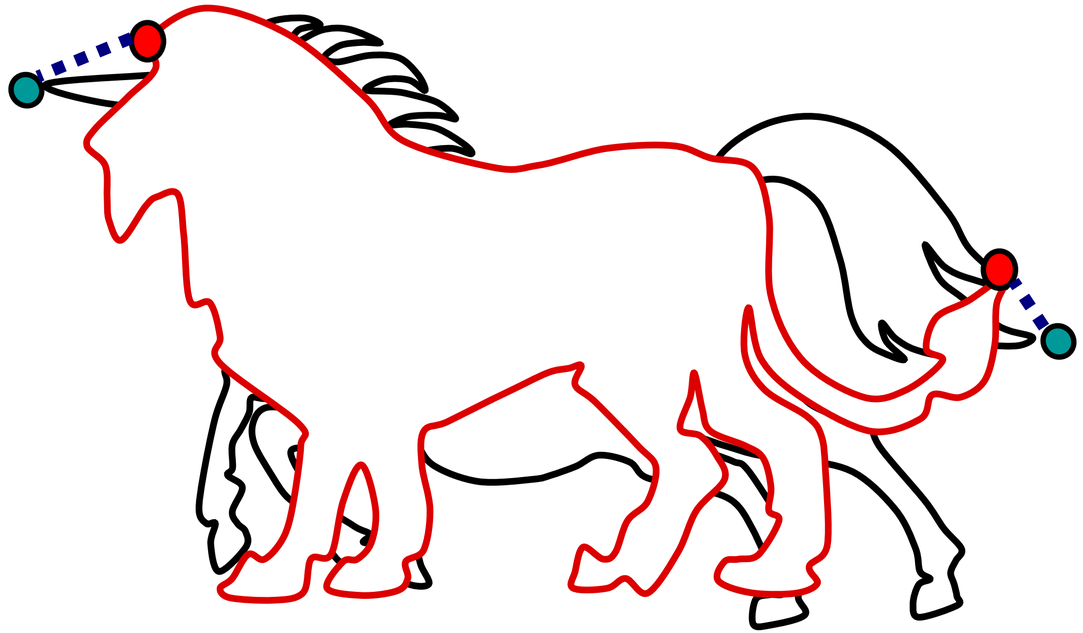
- Solution - Regard the shapes as sets of points and try to “match” these sets using a transformation

TRANSFORMATIONS



- Not a great alignment

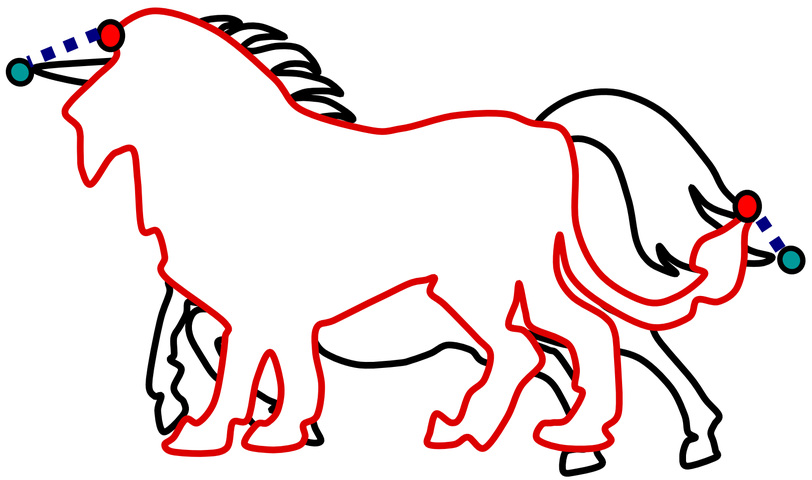
TRANSFORMATIONS



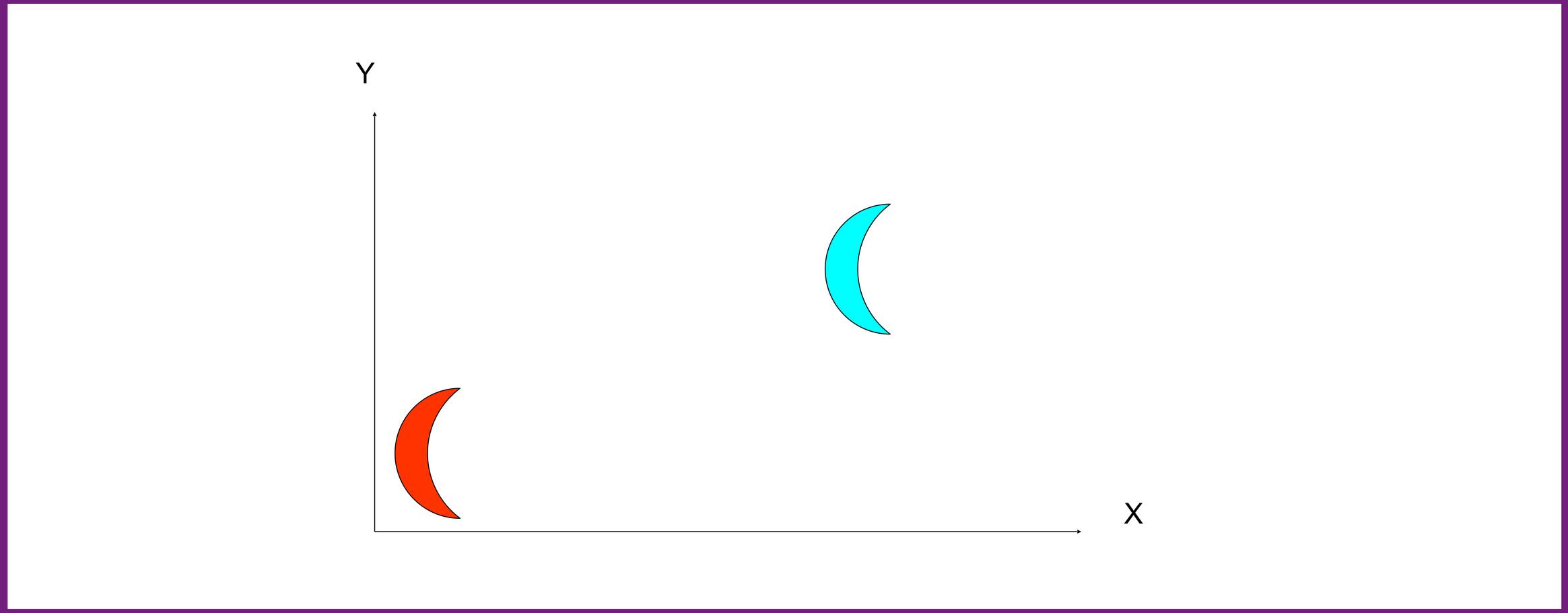
- Good alignment

TRANSFORMATIONS

- Shape transformations to align points in space
  - Rotation
  - Translation
  - Scaling
  - …

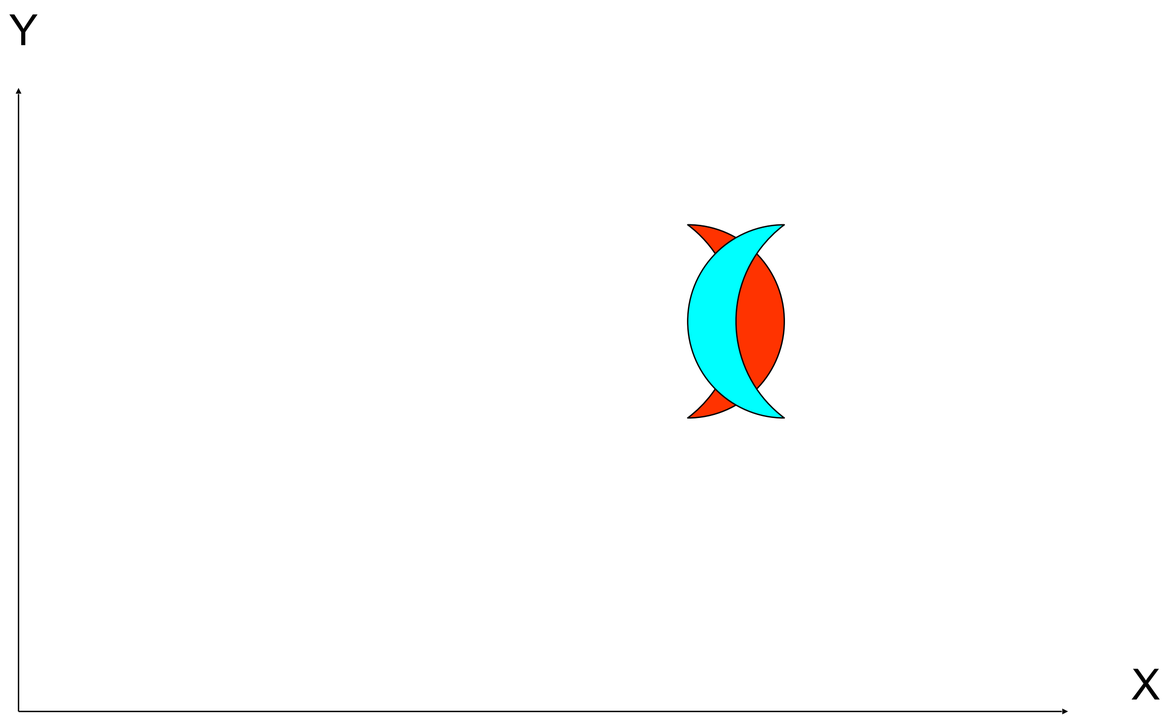


TRANSFORMATIONS



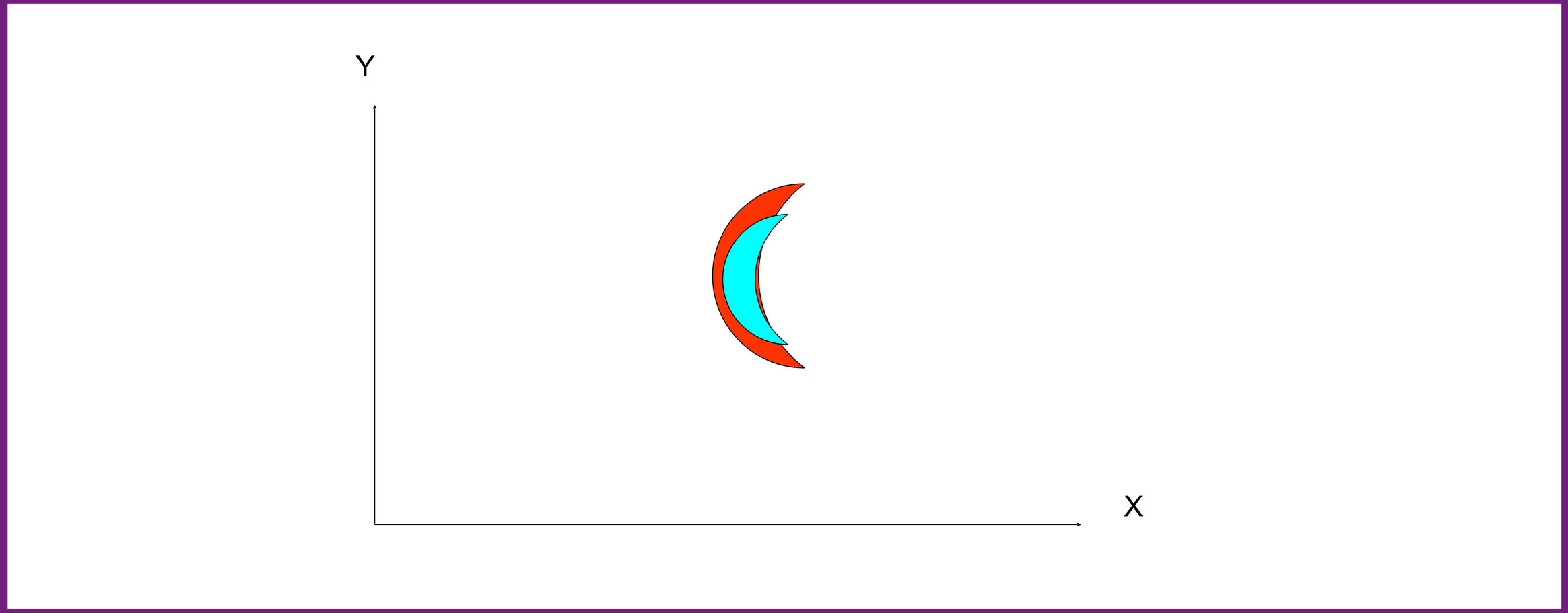
- Translation

TRANSFORMATIONS



- Rotation

TRANSFORMATIONS

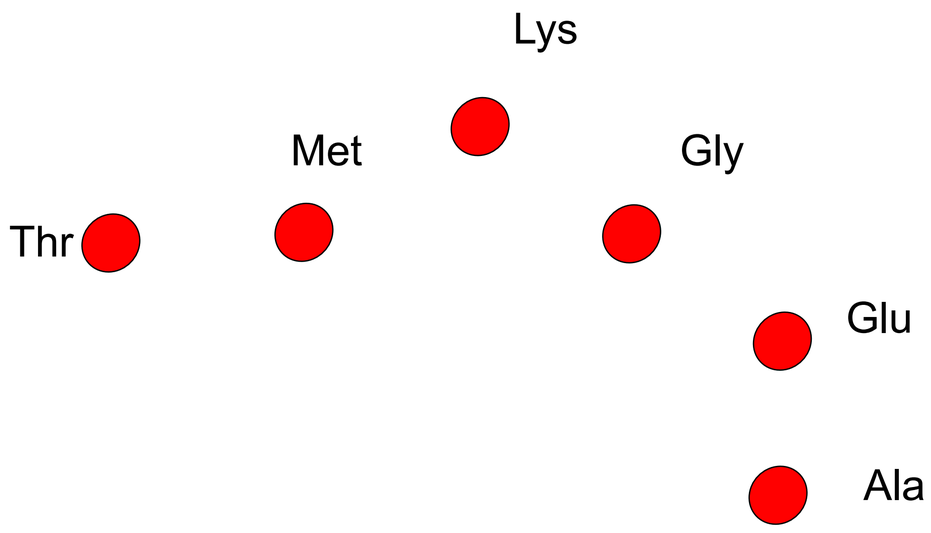


- Scale

Aligning Structures

ALIGNING STRUCTURES

- We represent a protein as a geometric object
  - Object consists of points in space represented by coordinates
    - (x, y, z)



ALIGNING STRUCTURES

- Given two sets of points A = (a1, a2, …, an) and B = (b1,b2,…bm) in Cartesian space
  - Find the optimal subsets A(P) and B(Q) with |A(P)| = |B(Q)|
  - Find the optimal rigid body transformation G between the two subsets A(P) and B(Q) that minimizes a given distance metric D over all possible rigid body transformation
  - $$\min_{G}\{D(A(P) - G(B(Q)))\}$$

ALIGNING STRUCTURES

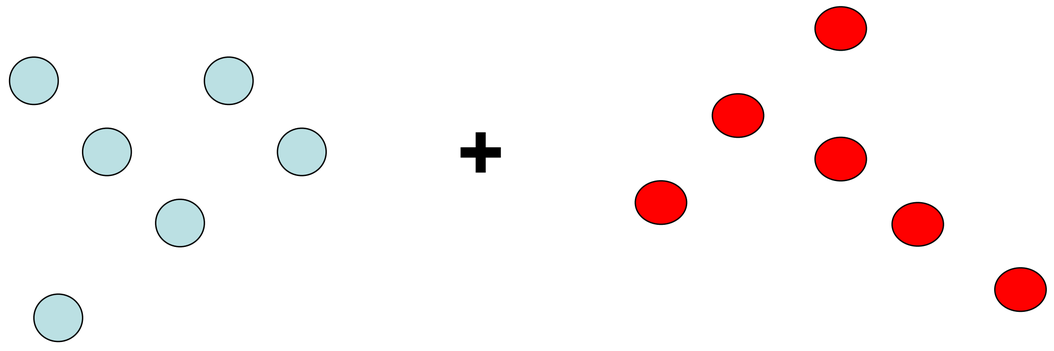
- The two subsets A(P) and B(Q) define a "correspondence", and p = |A(P)| = |B(Q)| is called the correspondence length
  - The correspondence length is maximal when A(P) and B(Q) are similar

$$\min_{G}\left\{D(A(P) - G(B(Q)))\right\}$$

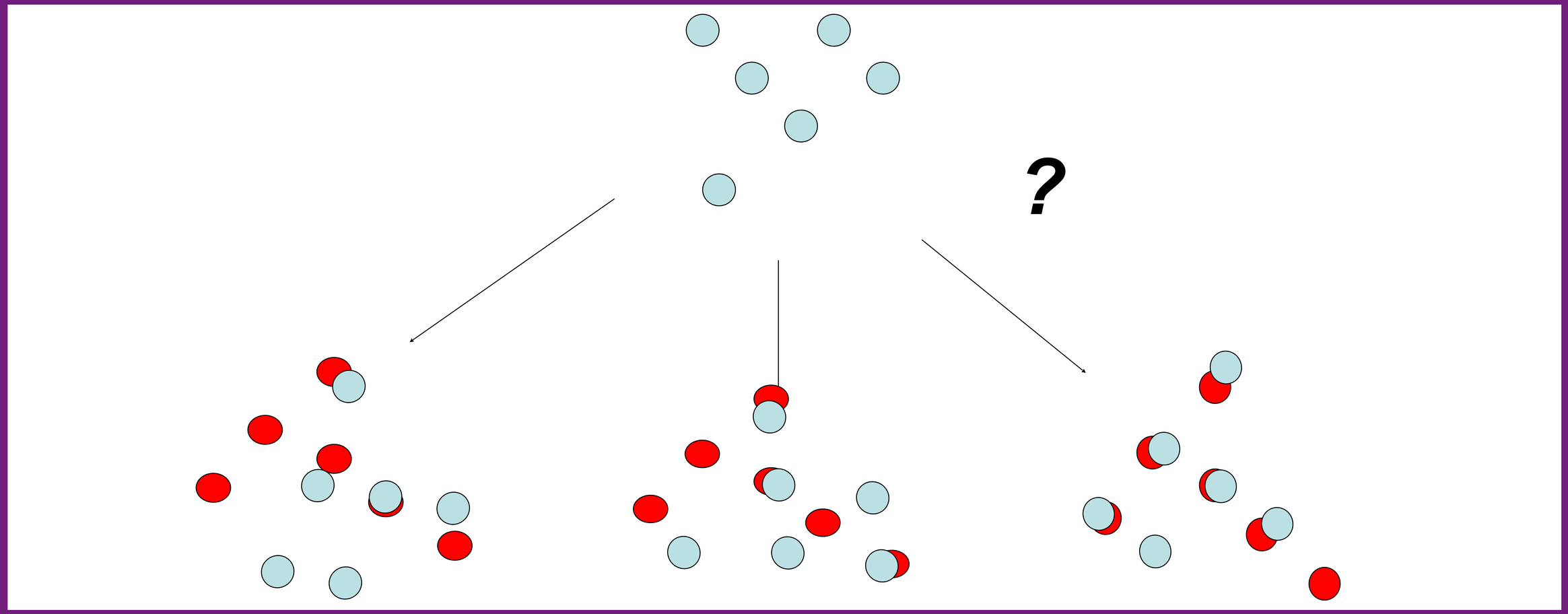
- Therefore there are essentially two problems in structure alignment:
  - Find the correspondence set
  - Find the alignment transform

ALIGNING STRUCTURES

- Correspondence unknown:
  - Given two configurations of points in the three dimensional space

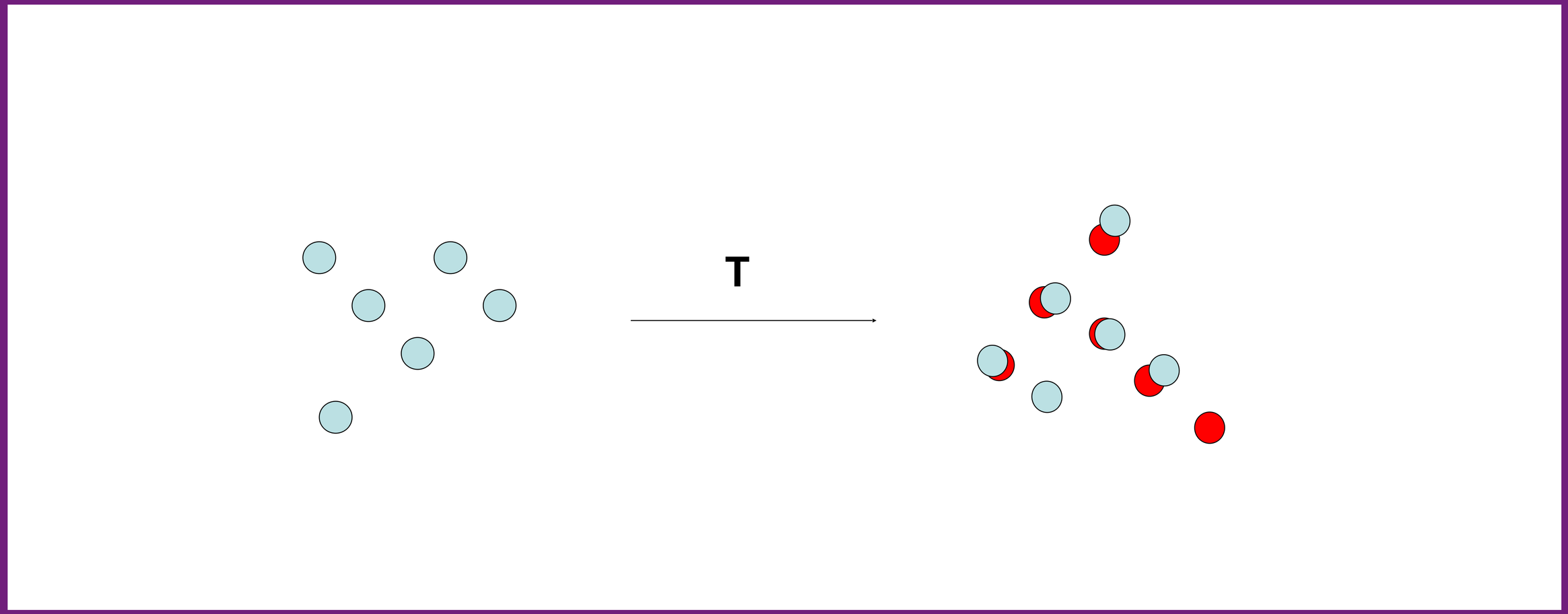


ALIGNING STRUCTURES



- Find rotations and translations of one of the point sets which produce “large” superimpositions of corresponding 3-D points

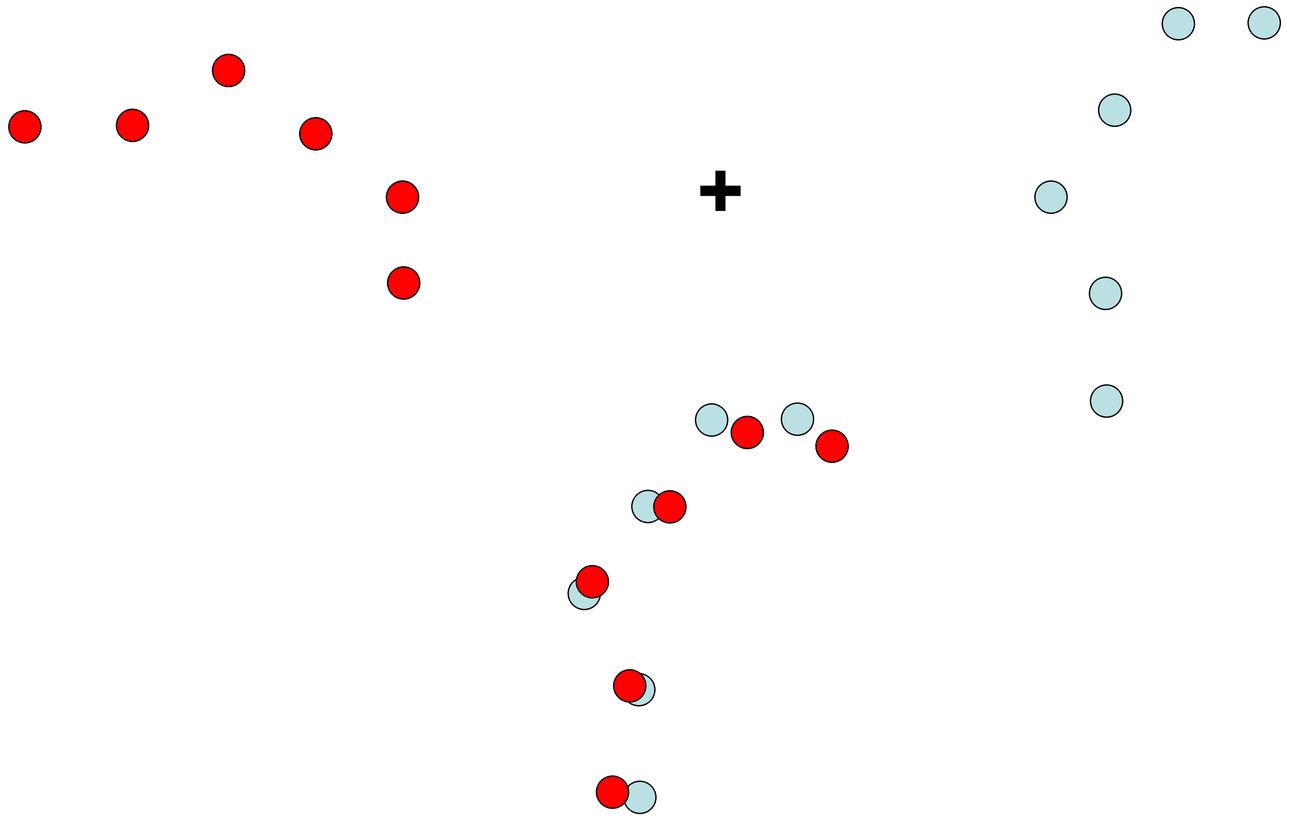
ALIGNING STRUCTURES



- The best transformation

ALIGNING STRUCTURES

- Simple case of two closely related proteins with the same number of amino acids
- How do we assess the quality of alignment?



ALIGNING STRUCTURES

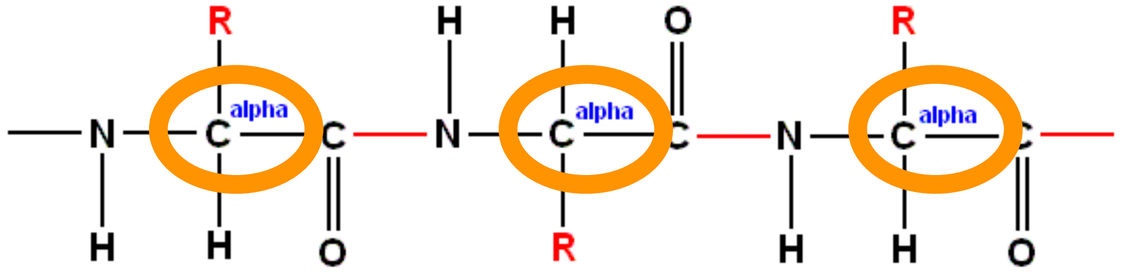
- Scoring the alignments
  - Two point sets:
    - A={ai} i=1…n
    - B={bj} j=1…m
  - Pairwise Correspondence:
    - (ak1,bt1) (ak2,bt2)… (atN,btN)
  - Bottleneck at max
    - max ||ak1 – bt1||

SCORING THE ALIGNMENT

- RMSD – Root Mean Square Deviation
  - Given two sets of 3-D points :
    - A={pi}, B={qi} , i=1,…,n;

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(a_i - b_i)^2}{n}}$$

$$\text{cRMSD} = \underset{B}{\text{MIN}}\sqrt{\frac{\sum_{i=1}^{n}(a_i - b_i)^2}{n}}$$



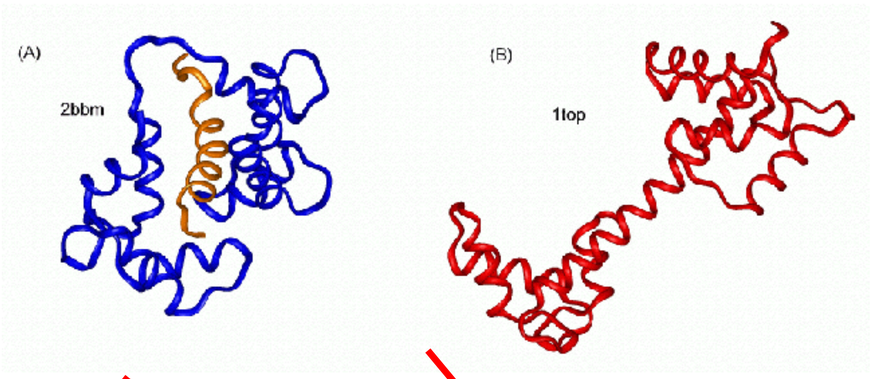
SCORING THE ALIGNMENT

- Alignment algorithm aims
  - Find a 3-D transformation T such minimized RMSD
  - Find the highest number of atoms aligned with the lowest RMSD
- RMSD is biased
  - All atoms are treated equally
  - Residues on the surface have a higher degree of freedom than those in the core
  - Best alignment does not always mean minimal RMSD
  - Does not take into account the attributes of the amino acids

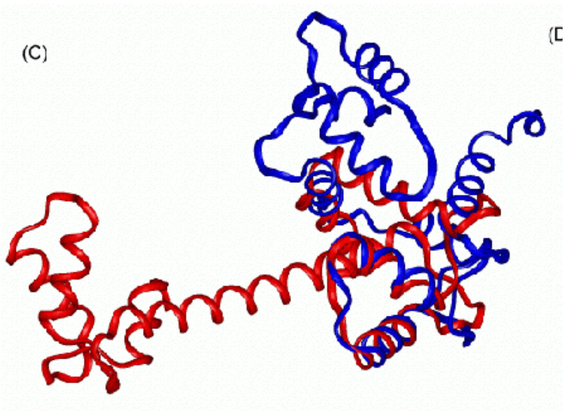
DIFFERENT APPROACHES AND ALGORITHMS

TRANSFORMATIONS

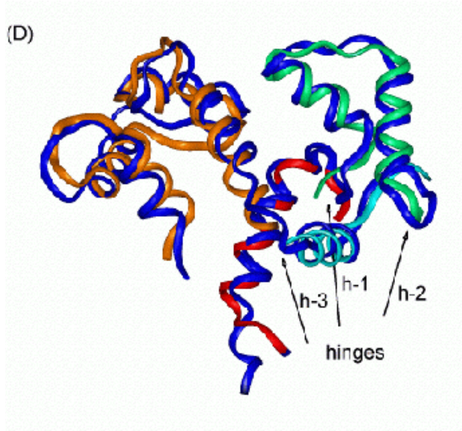
- The best transformation



Rigid alignment



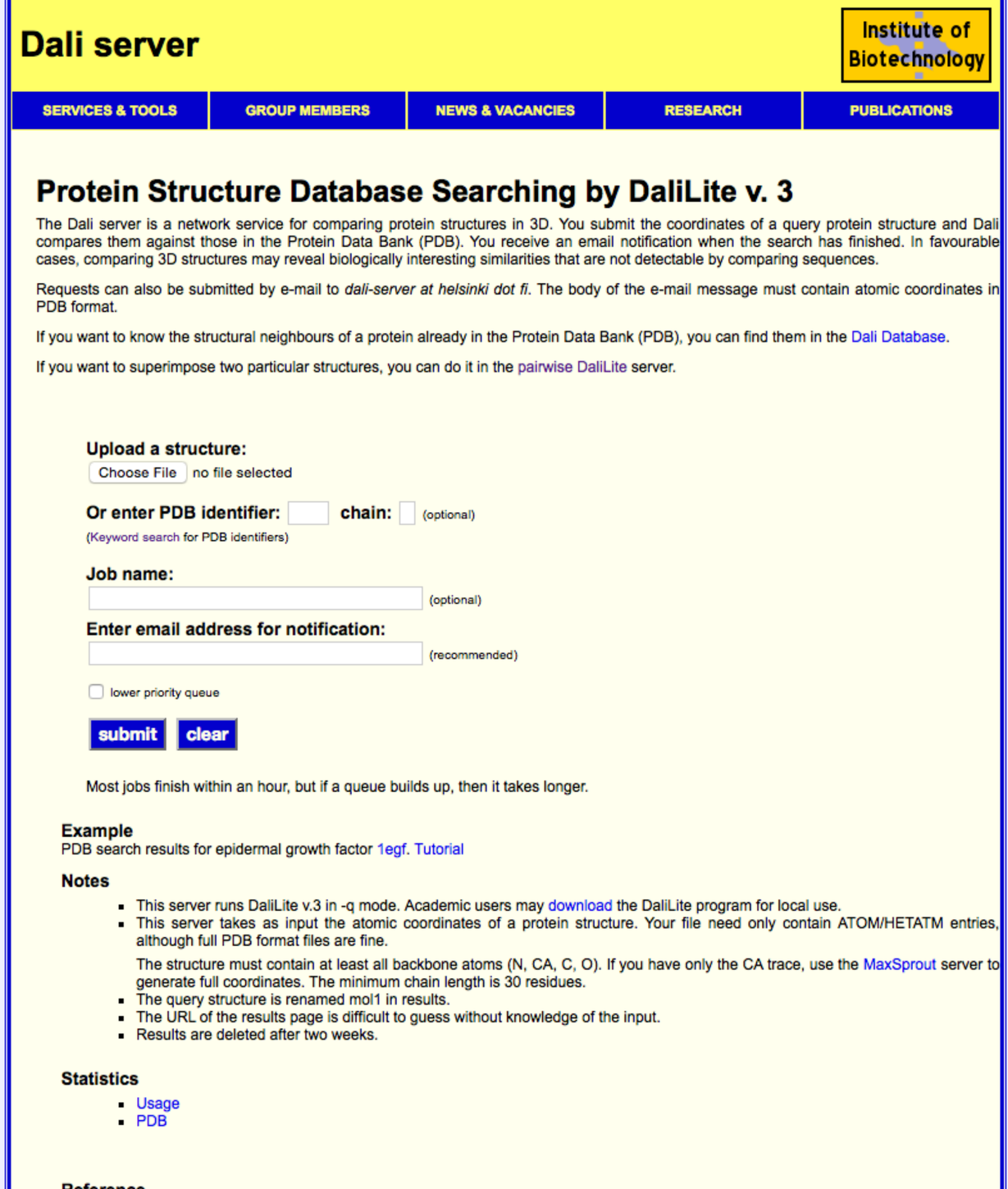
Flexible alignment



Structural Alignment Software

ALIGNMENT SOFTWARE

- DALI
  - Holm and Sander. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993, 233:123-128
  - Uses 2D distance matrices between CA atoms to represent each structure
  - Maximally overlay the matrices



Dali server

Institute of Biotechnology

SERVICES & TOOLS | GROUP MEMBERS | NEWS & VACANCIES | RESEARCH | PUBLICATIONS

Protein Structure Database Searching by DaliLite v. 3

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). You receive an email notification when the search has finished. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

Requests can also be submitted by e-mail to *dali-server at helsinki dot fi*. The body of the e-mail message must contain atomic coordinates in PDB format.

If you want to know the structural neighbours of a protein already in the Protein Data Bank (PDB), you can find them in the Dali Database.

If you want to superimpose two particular structures, you can do it in the pairwise DaliLite server.

Upload a structure:
Choose File no file selected

Or enter PDB identifier: **chain:** (optional)
(Keyword search for PDB identifiers)

Job name:
(optional)

Enter email address for notification:
(recommended)

lower priority queue

submit clear

Most jobs finish within an hour, but if a queue builds up, then it takes longer.

Example
PDB search results for epidermal growth factor 1egf. Tutorial

Notes

- This server runs DaliLite v.3 in -q mode. Academic users may download the DaliLite program for local use.
- This server takes as input the atomic coordinates of a protein structure. Your file need only contain ATOM/HETATM entries, although full PDB format files are fine.

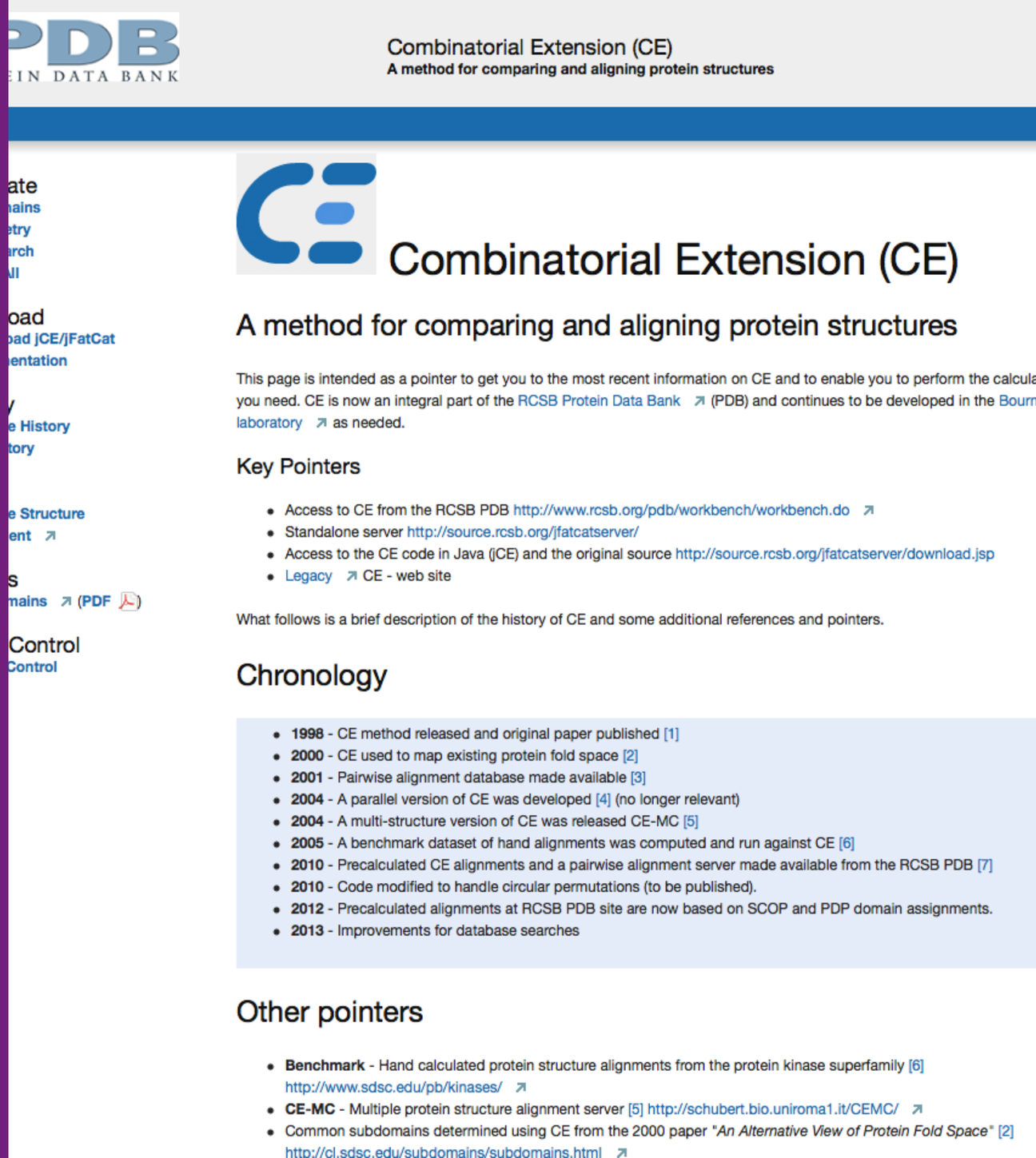
  The structure must contain at least all backbone atoms (N, CA, C, O). If you have only the CA trace, use the MaxSprout server to generate full coordinates. The minimum chain length is 30 residues.
- The query structure is renamed mol1 in results.
- The URL of the results page is difficult to guess without knowledge of the input.
- Results are deleted after two weeks.

Statistics

- Usage
- PDB

ALIGNMENT SOFTWARE

- CE (Combinatorial extension)
  - Shindyalov and Bourne, Protein structure alignment by incremental combinatorial extension (CE) of optimal path. Prot Eng, 1998, 11:739-747
  - Uses characteristics of local geometry to seed structural alignments
  - Joins these regions of local similarity into an "optimal" path for the full alignment
  - Bottom-up approach



PDB
EIN DATA BANK

Combinatorial Extension (CE)
A method for comparing and aligning protein structures

ate
ains
etry
arch
ll

oad
oad jCE/jFatCat
entation

y
e History
tory

e Structure
ent

s
mains (PDF)

Control
Control

Combinatorial Extension (CE)

A method for comparing and aligning protein structures

This page is intended as a pointer to get you to the most recent information on CE and to enable you to perform the calcula
you need. CE is now an integral part of the RCSB Protein Data Bank (PDB) and continues to be developed in the Bourn
laboratory as needed.

Key Pointers

- Access to CE from the RCSB PDB http://www.rcsb.org/pdb/workbench/workbench.do
- Standalone server http://source.rcsb.org/jfatcatserver/
- Access to the CE code in Java (jCE) and the original source http://source.rcsb.org/jfatcatserver/download.jsp
- Legacy CE - web site

What follows is a brief description of the history of CE and some additional references and pointers.

Chronology

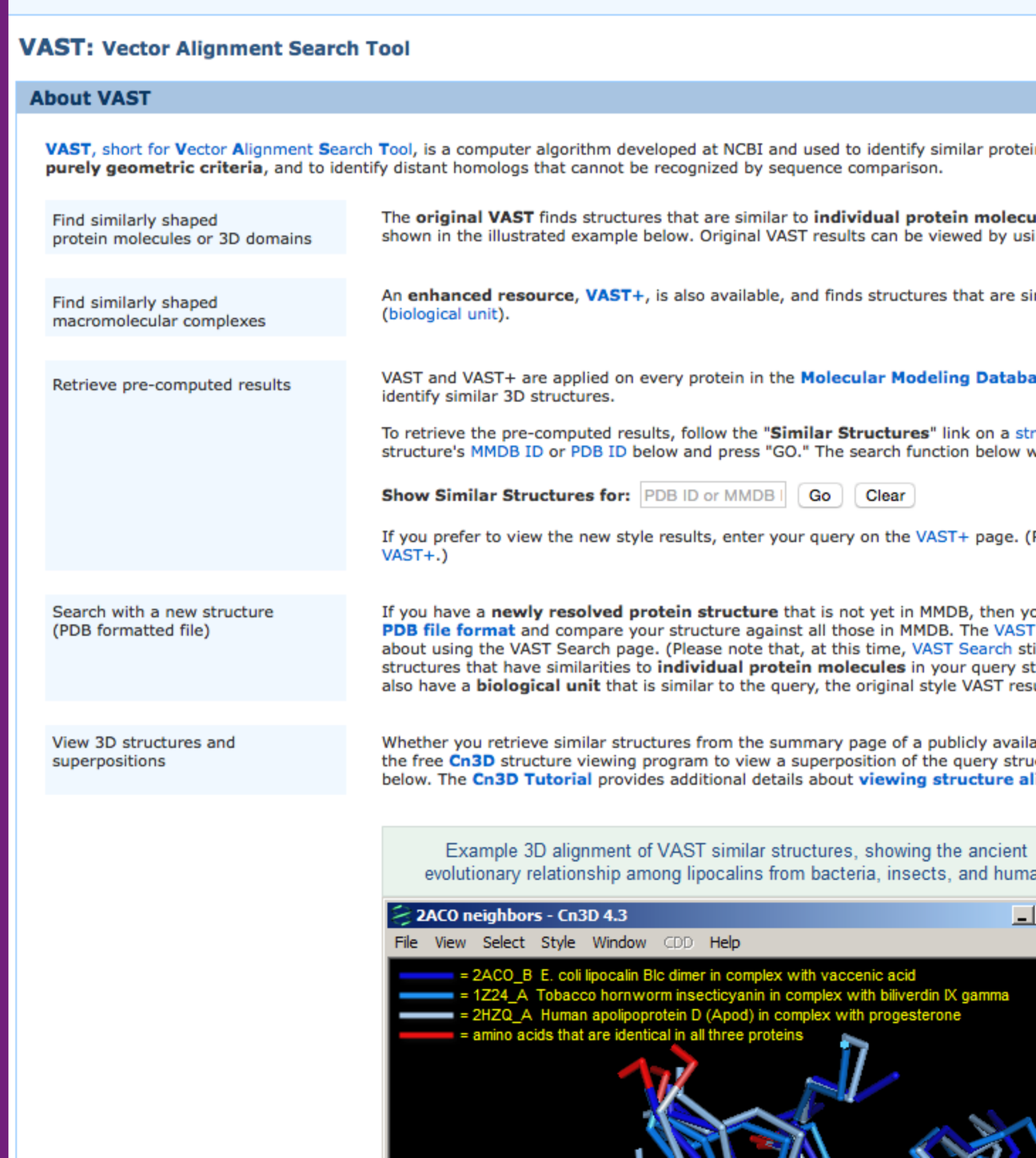
- **1998** - CE method released and original paper published [1]
- **2000** - CE used to map existing protein fold space [2]
- **2001** - Pairwise alignment database made available [3]
- **2004** - A parallel version of CE was developed [4] (no longer relevant)
- **2004** - A multi-structure version of CE was released CE-MC [5]
- **2005** - A benchmark dataset of hand alignments was computed and run against CE [6]
- **2010** - Precalculated CE alignments and a pairwise alignment server made available from the RCSB PDB [7]
- **2010** - Code modified to handle circular permutations (to be published).
- **2012** - Precalculated alignments at RCSB PDB site are now based on SCOP and PDP domain assignments.
- **2013** - Improvements for database searches

Other pointers

- **Benchmark** - Hand calculated protein structure alignments from the protein kinase superfamily [6] http://www.sdsc.edu/pb/kinases/
- **CE-MC** - Multiple protein structure alignment server [5] http://schubert.bio.uniroma1.it/CEMC/
- Common subdomains determined using CE from the 2000 paper "*An Alternative View of Protein Fold Space*" [2] http://cl.sdsc.edu/subdomains/subdomains.html

ALIGNMENT SOFTWARE

- VAST (Vector Alignment Search Tool)
  - Treats secondary structure elements as vector
  - Purely geometrical approach
  - Fast, but looses information



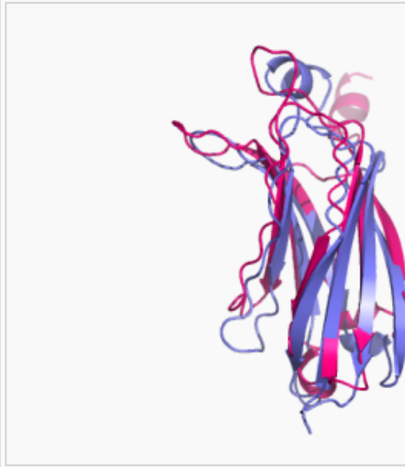
ALIGNMENT SOFTWARE

- PyMOL has built in structural alignment algorithms
  - align
  - fit
  - super
  - cealign
  - pair-fit
  - …

Page Discussion Read View source View history

Align

align performs a sequence alignment followed by a structural superposition, and then carries out zero or more cycles of refinement in order to reject structural outliers found during the fit. **align** does a good job on proteins with decent sequence similarity (identity >30%). For comparing proteins with lower sequence identity, the super and cealign commands perform better.



Two proteins after structure alignment

Contents [hide]

Usage

```
align mobile, target [, cutoff [, cycles
    [, gap [, extend [, max_gap [, object
    [, matrix [, mobile_state [, target_state
    [, quiet [, max_skip [, transform [, reset ]]]]]]]]]]]]]]]]
```

Arguments

- **mobile** = string: atom selection of mobile object
- **target** = string: atom selection of target object
- **cutoff** = float: outlier rejection cutoff in Angstrom {default: 2.0}
- **cycles** = int: maximum number of outlier rejection cycles {default: 5}
- **gap, extend, max_gap**: sequence alignment parameters
- **object** = string: name of alignment object to create {default: (no alignment object)}
- **matrix** = string: file name of substitution matrix for sequence alignment {default: BLOSUM62}
- **mobile_state** = int: object state of mobile selection {default: 0 = all states}
- **target_state** = int: object state of target selection {default: 0 = all states}
- **quiet** = 0/1: suppress output {default: 0 in command mode, 1 in API}
- **max_skip** = ?
- **transform** = 0/1: do superposition {default: 1}
- **reset** = ?

Alignment Objects