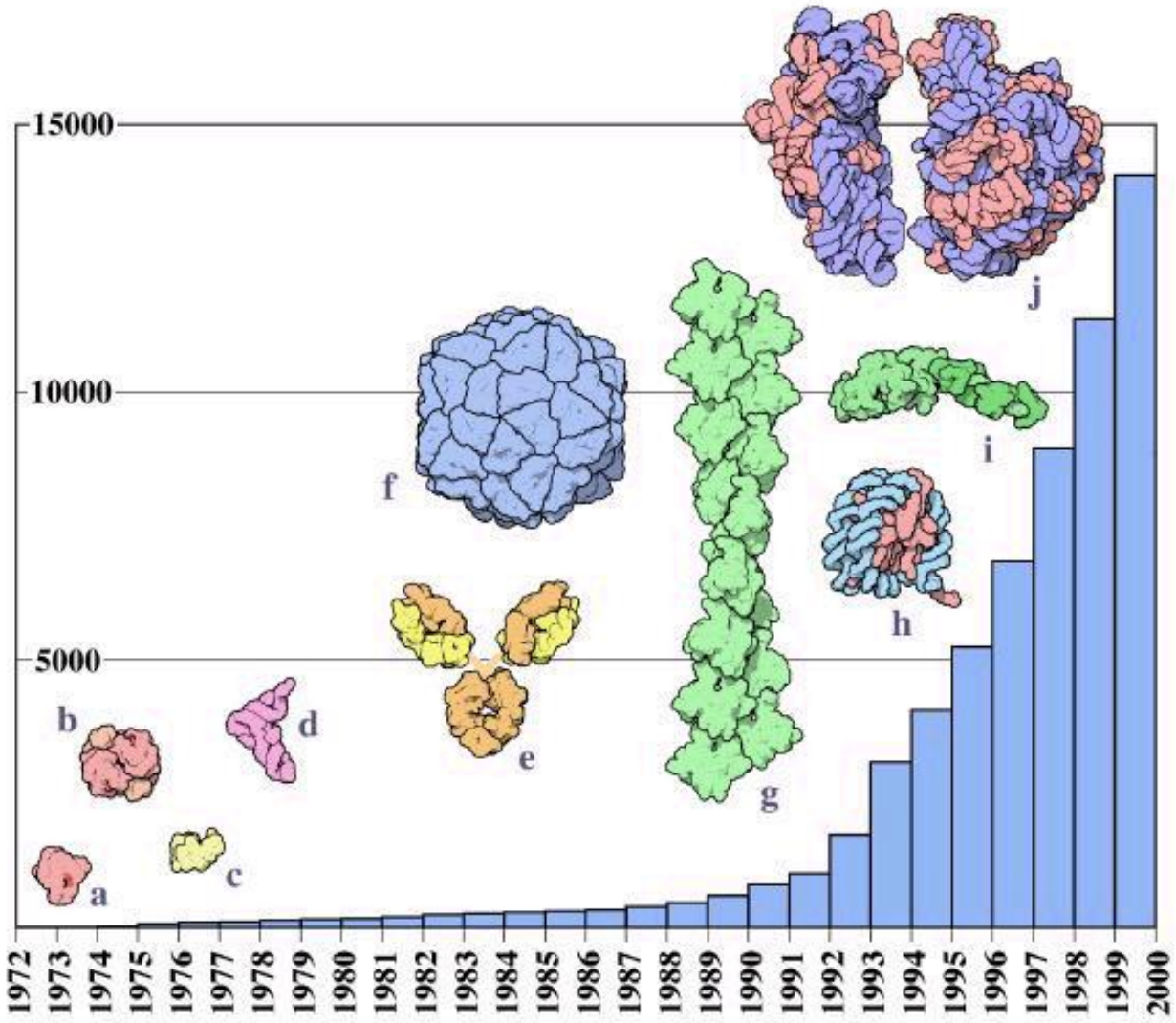
Multiple Structural Alignment

Protein structures



Protein structure databases

- PDB
  - 3D structures
- SCOP
  - Murzin, Brenner, Hubbard, Chothia
  - Classification
    - Class (mostly alpha, mostly beta, alpha/beta (interspersed), alpha+beta (segregated), multi-domain, membrane)
    - Fold (similar structure)
    - Superfamily (homology, distant sequence similarity)
    - Family (homology and close sequence similarity)

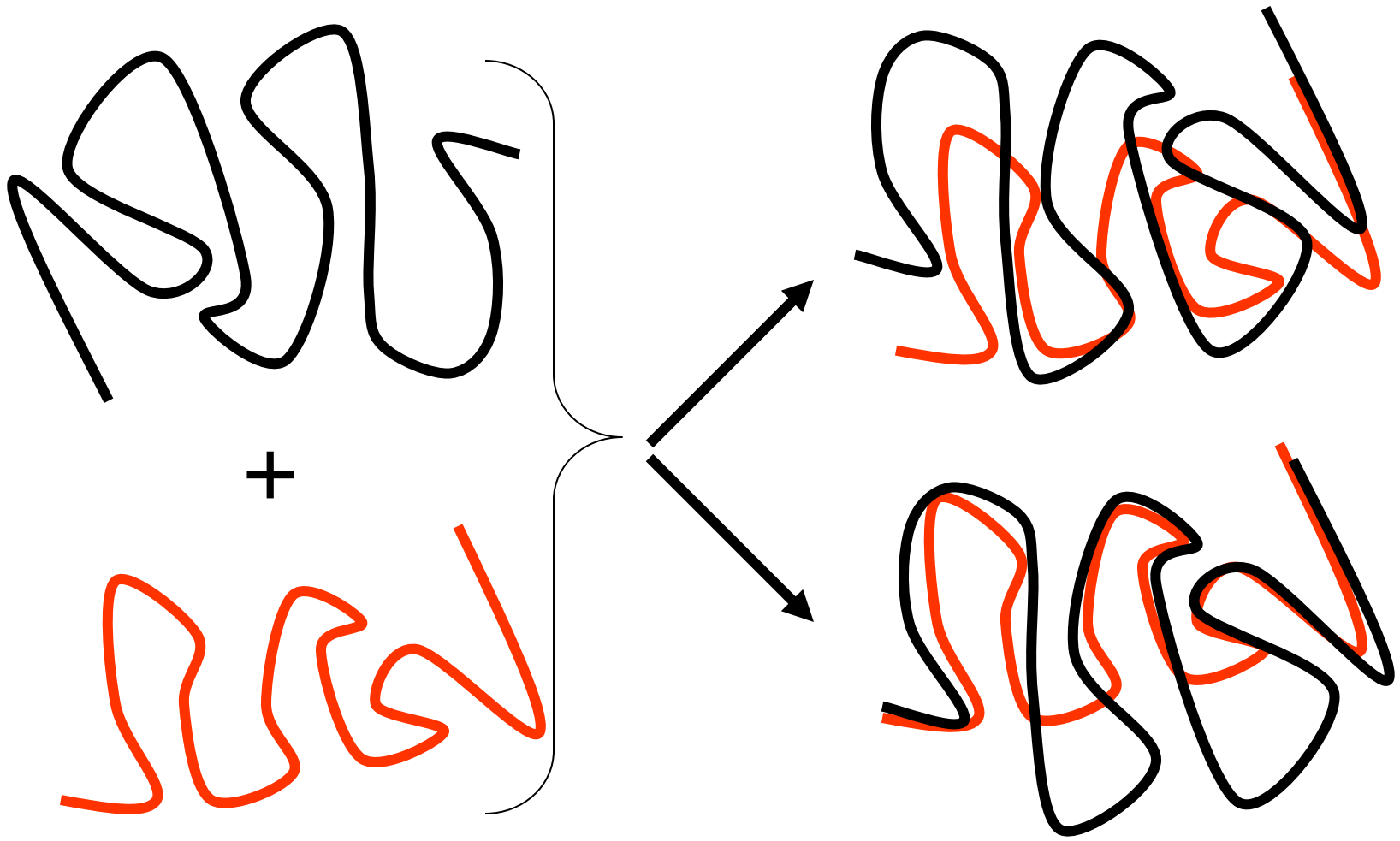
Protein structure comparison

- Levels of structure description
  - Atom/atom group
  - Residue
  - Fragment
  - Secondary structure element (SSE)
- Basis of comparison
  - Geometry/architecture of coordinates/relative positions
  - sequential order of residues along backbone, ...
  - physio-chemical properties of residues, …

How to compare?

- **Key problem**: find an optimal correspondence between the arrangements of atoms in two molecular structures (say A and B) in order to align them in 3D
- Optimality of the alignment is determined using a root mean square measure of the distances between corresponding atoms in the two molecules
- **Complication**: It is not known a priori which atom in molecule B corresponds to a given atom in molecule A (the two molecules may not even have the same number of atoms)

Find the optimal alignment



RMSD

Root Mean Square Deviation (*RMSD*)

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{n}}$$

n = number of atoms

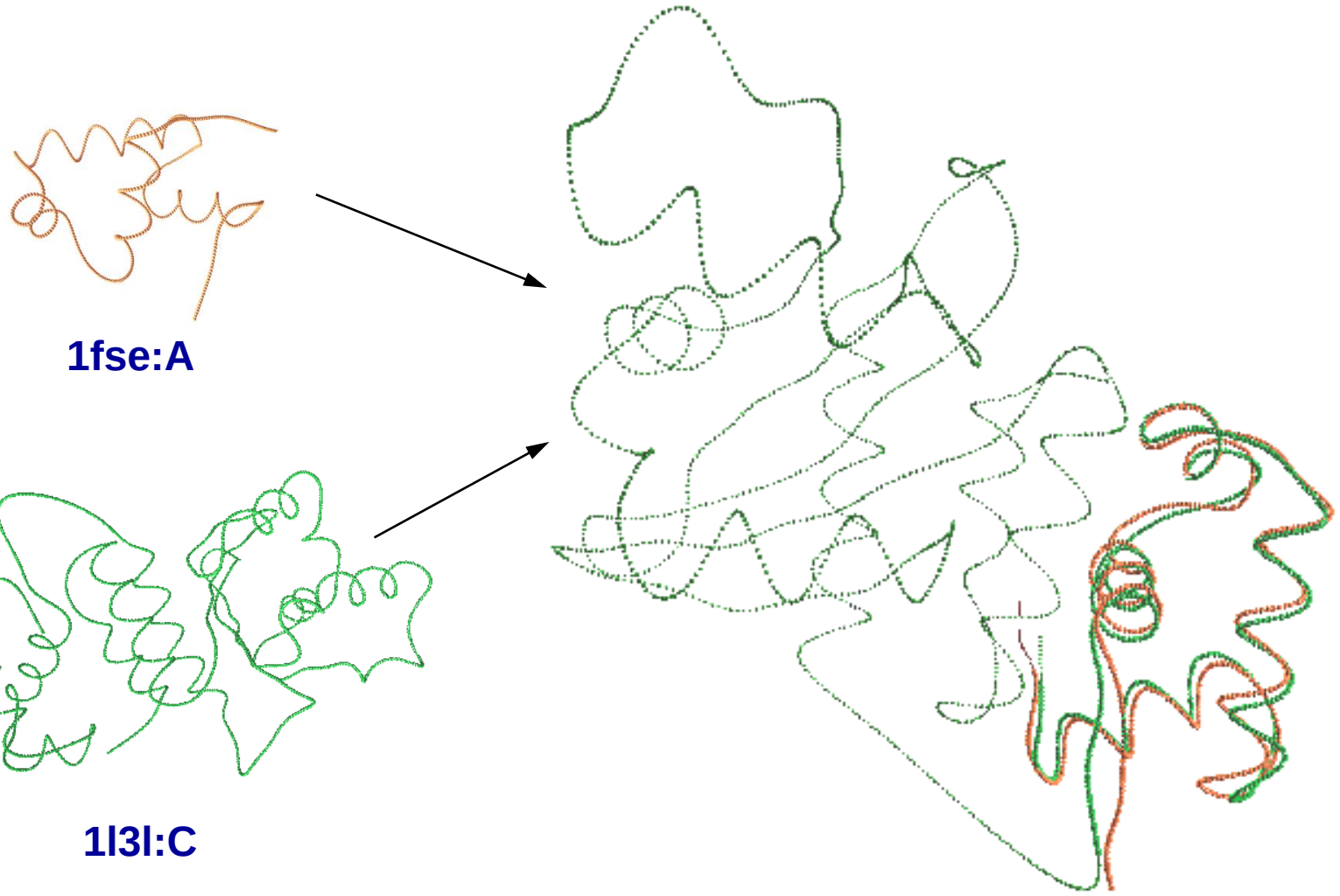
d_i = distance between 2 corresponding atoms i in 2 structures

RMSD

Unit of RMSD => e.g. Ångstroms

- identical structures => *RMSD* = “0”
- similar structures => *RMSD* is small (1 – 3 Å)
- distant structures => *RMSD* > 3 Å

Pairwise Alignment



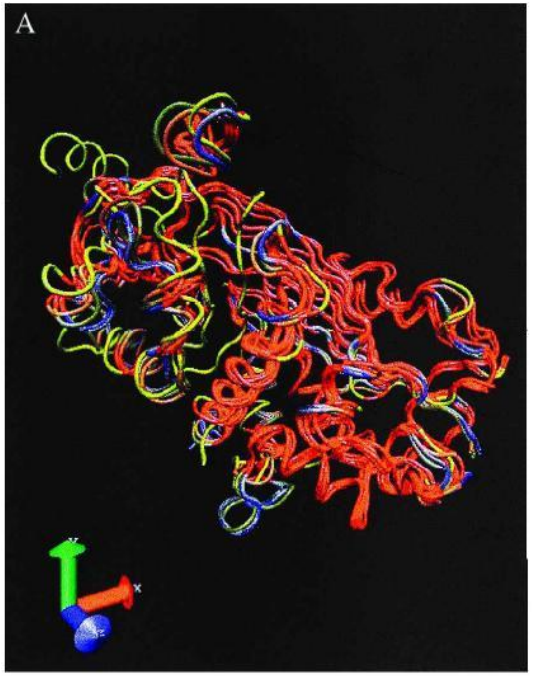
Multiple Structure Alignment

- The idea is similar to Multiple Sequence Alignment:
  - Find regions that are conserved among a set of input proteins
- The difference:
  - We do not use sequence information but atomic coordinate positions (3D structures of proteins) to determine conserved regions

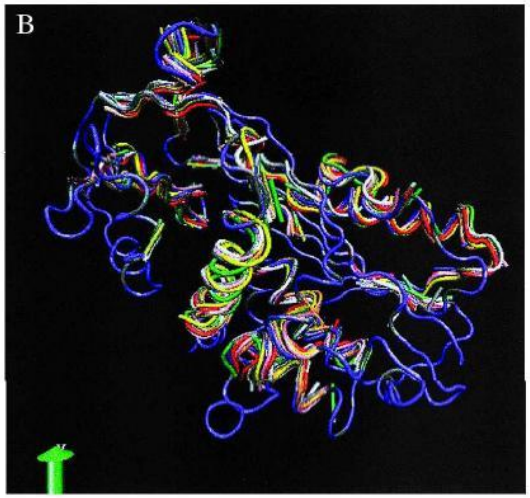
Pairwise vs. Multiple and Sequence vs. Structure

- Optimum pairwise sequence alignment can be found in $O(n^2)$ time.
- Multiple sequence alignment is exponential time.
- Pairwise structure alignment problem is NP-complete (Lathrop, 1994)
- So, Multiple Structure Alignment is a difficult problem

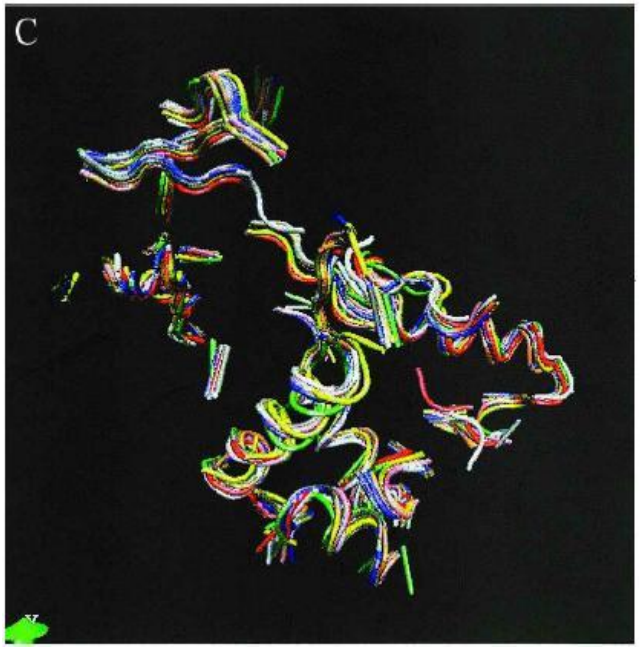
Serpins



First 6 molecules
(core highlighted in red)



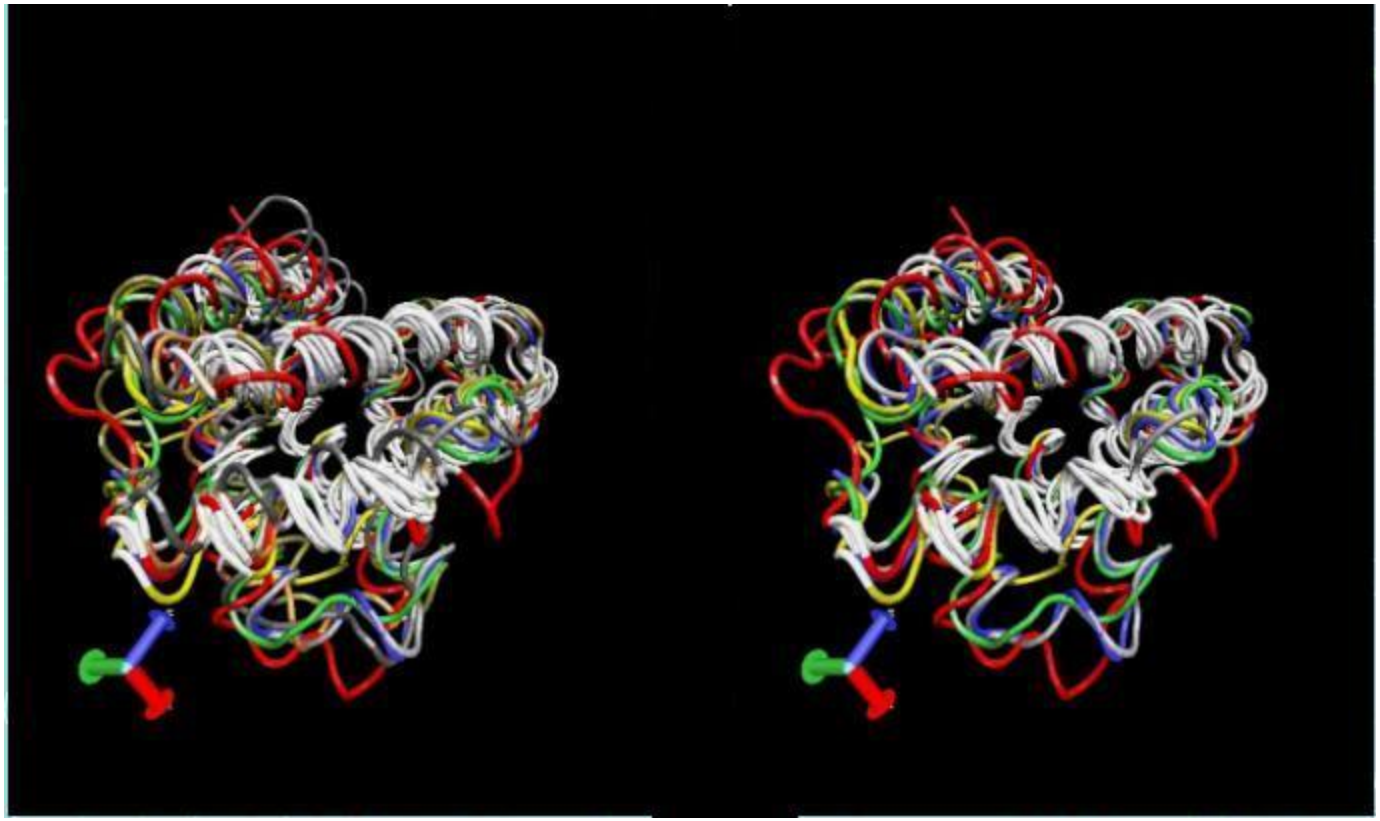
11 molecules



Core alone of 11 molecules

Multiple structure alignment result of MUSTA algorithm

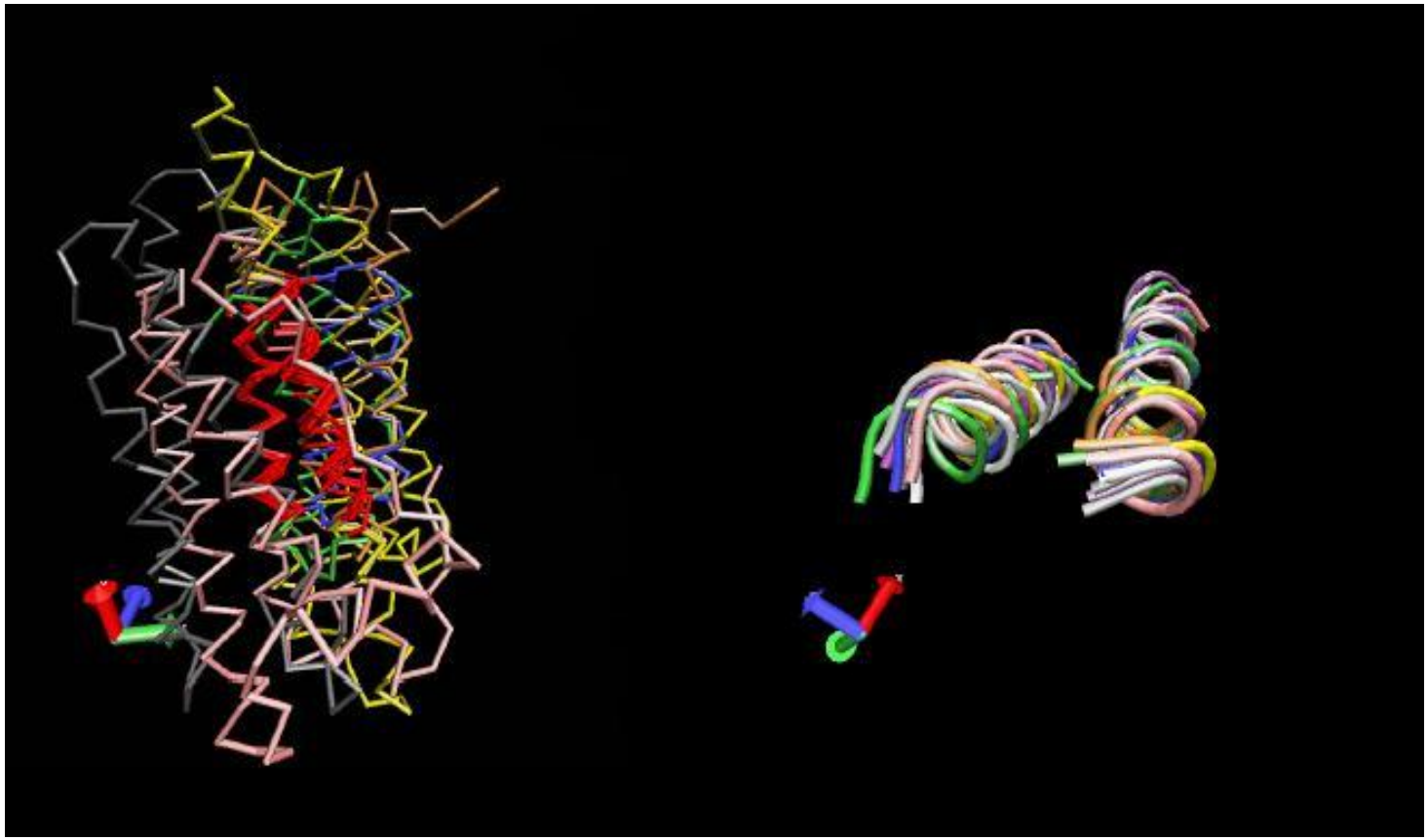
Globins



Running time = 1min (average)

Multiple structure alignment result of MUSTA algorithm

Cal-binding

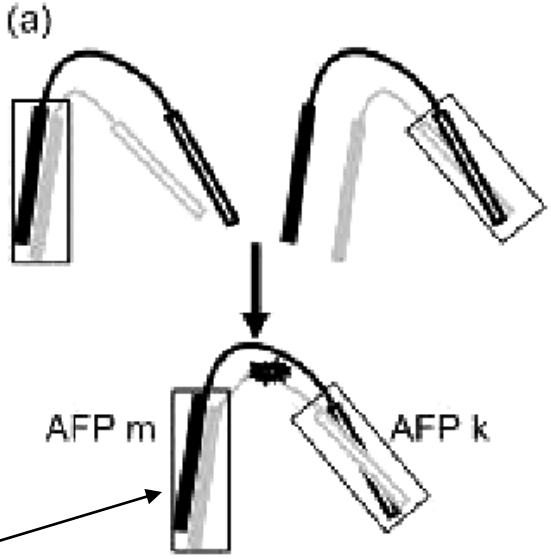


Running time = 8 sec

Multiple structure alignment result of MUSTA algorithm

Flexible vs. rigid body alignment

- Rigid body alignment:
  - Proteins treated as static rigid 3D objects
  - If we are to report a final single alignment, combining two local alignments may produce inconsistencies when two proteins are superpositioned using a single rotation/translation matrix
- Flexible: allow certain parts of the protein structure to twist (rotate and translate) so that we get a better alignment of matched parts



Progressive alignment