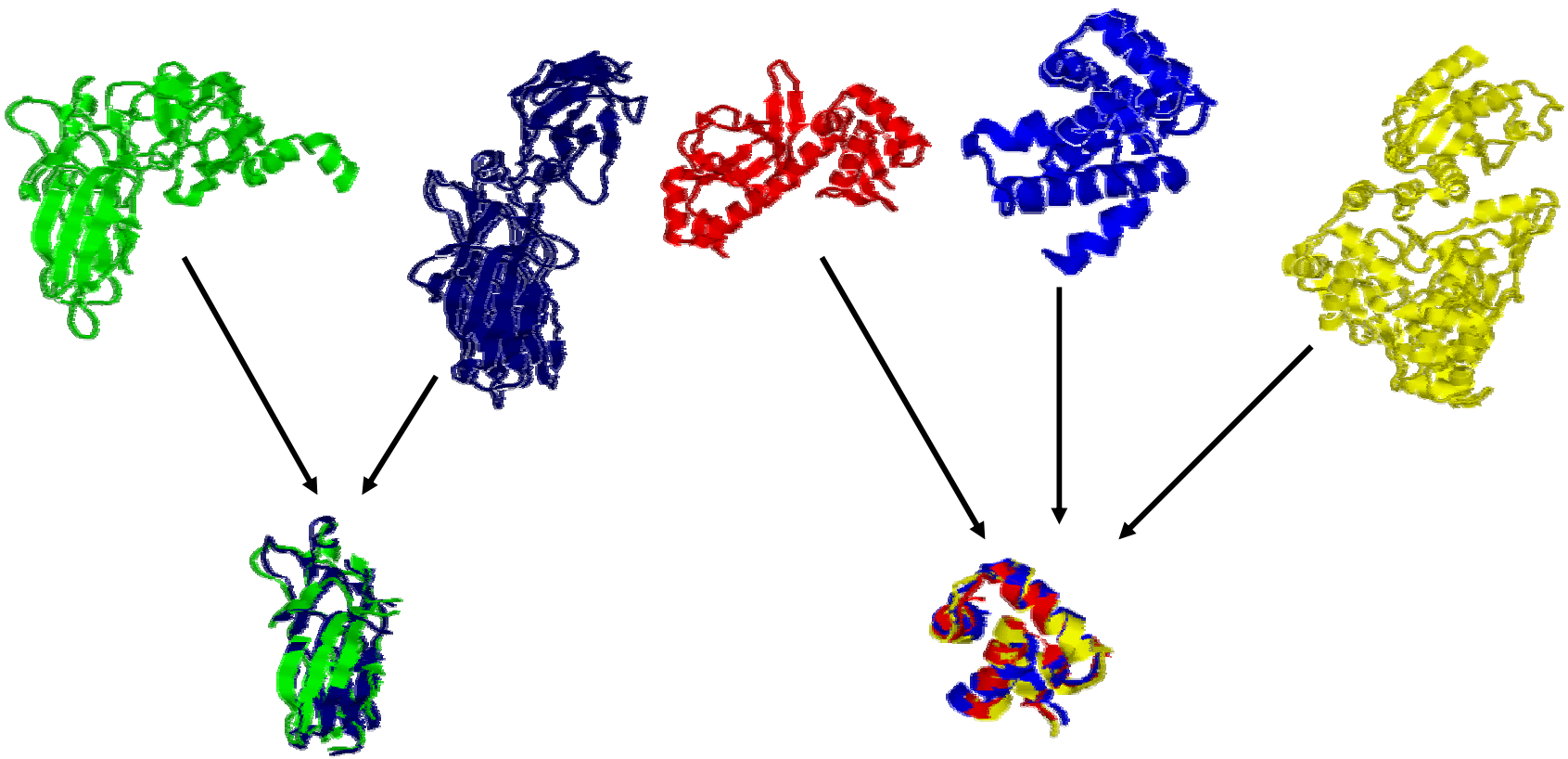
Multiple Structural Alignment

Problem Definition

Input: a collection of 3D protein structures

Goal: find substructures common to two or more proteins



Motivation

- *"Two structures whisper, a full multiple structure alignment shouts out load."* (A. Lesk)

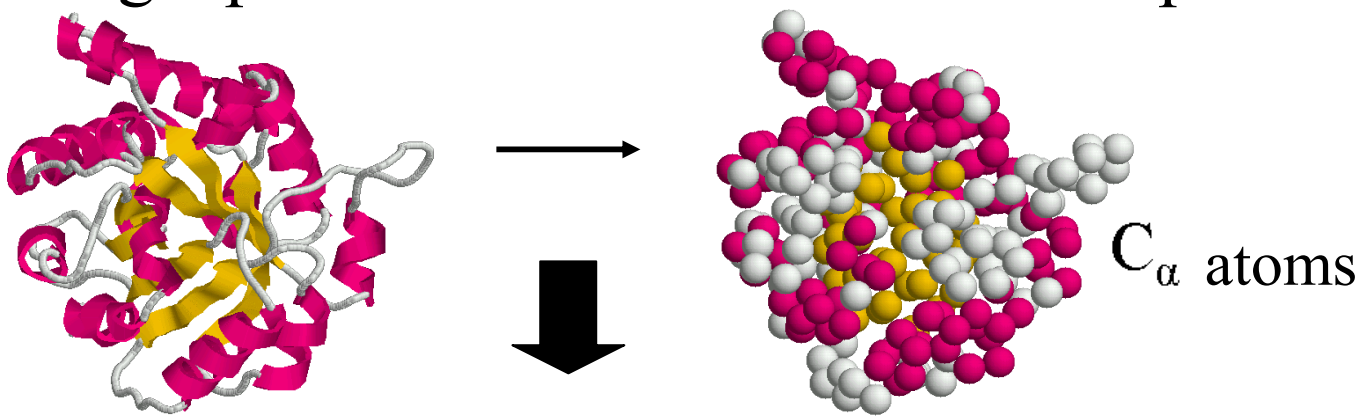
- Multiple alignment vs. pairwise alignment
  - carries significantly more information
  - filers out the noisy and insignificant alignments

Applications

- Protein Structural classification
- Evolutionary analysis
- Functional and structural motif detection
- Structure prediction algorithms
  - Homology modeling
  - Threading

The problem is a hard one

Representing a protein structure as a set of 3D points



Multiple Structural Alignment

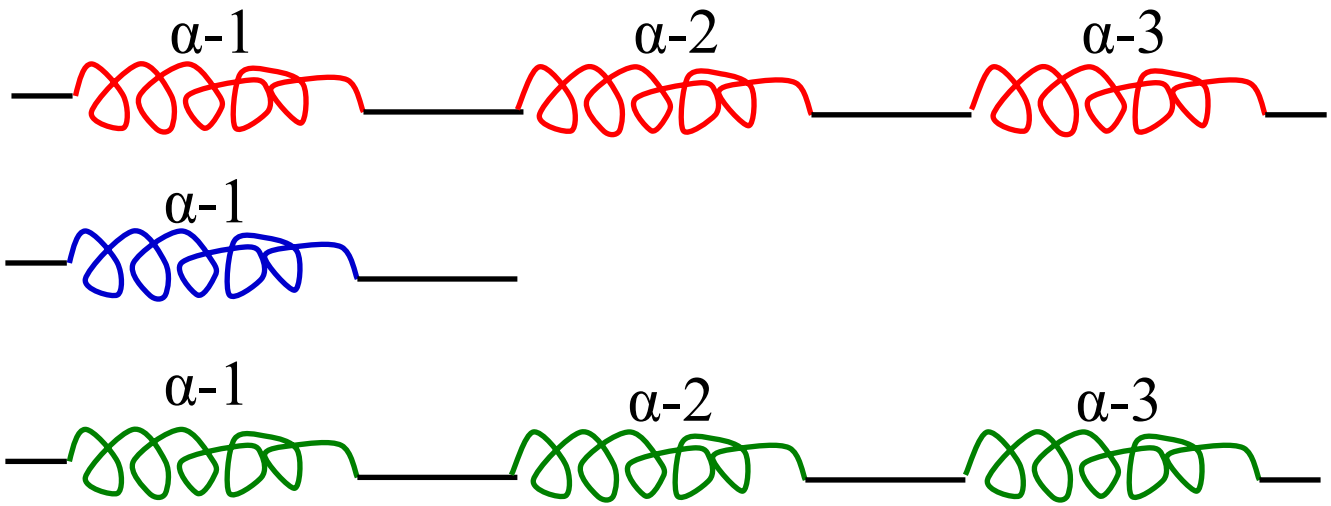
≡

Largest Common Point set (LCP)

NP-hard (Akutsu, 2000)

Solution Space

The number of solutions might be exponential.

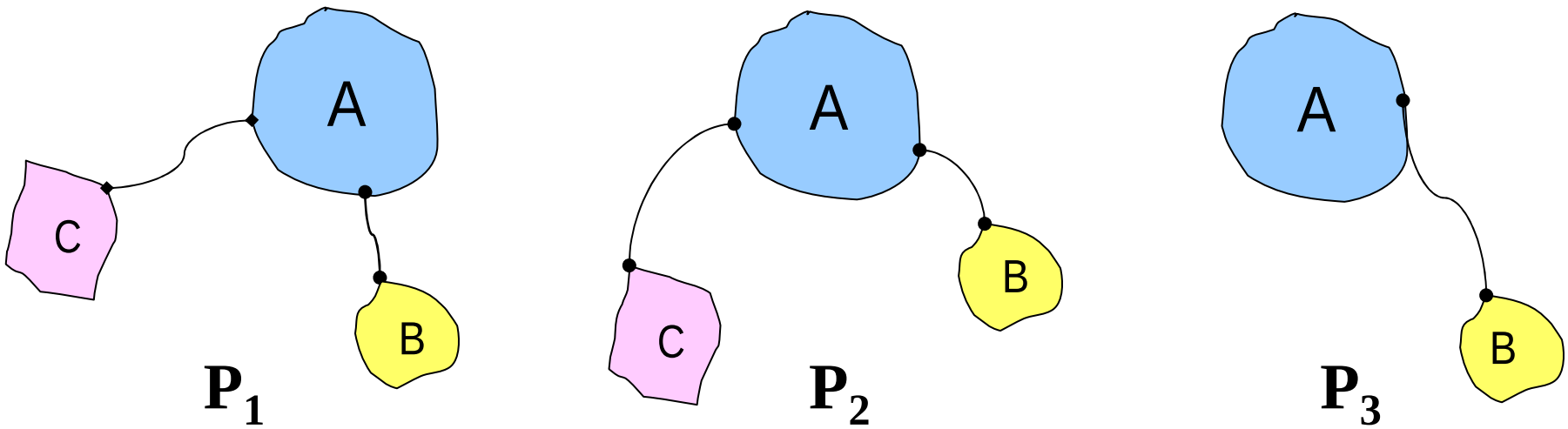


3•1•3 possible solutions

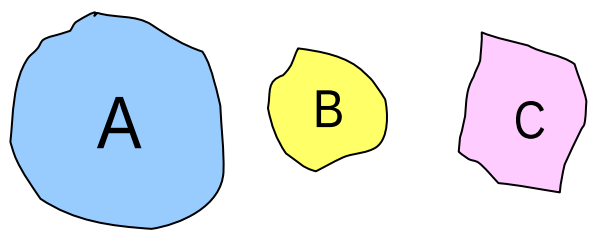
k^m possible solutions

(k – maximum number of helices per molecule, m – number of molecules)

- The problem is more complicated due to:
  - Similar substructures instead of identical
  - Partial alignments (smaller common substructures)
  - Subset alignments



Common substructures:

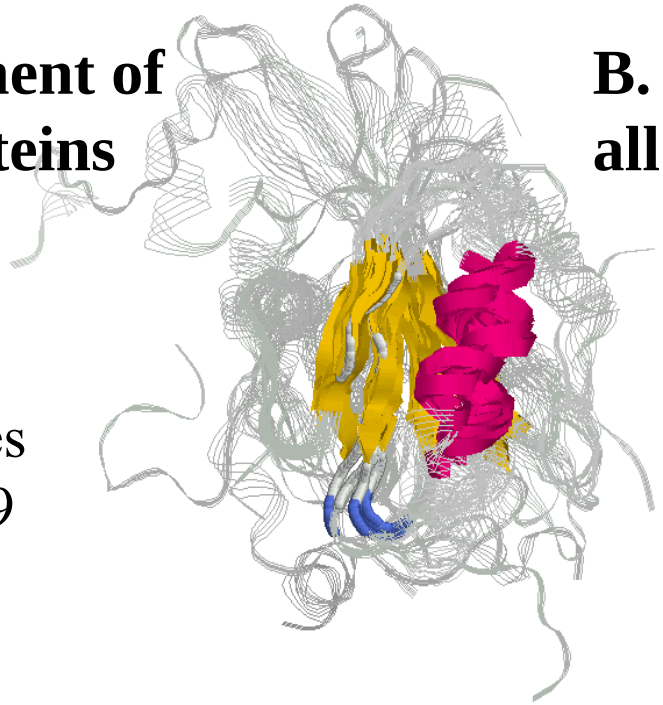


Subset Alignments
Cofilin-like and Gelsolin-like Families

- The ensemble contains 12 sequentially non-redundant structures taken from the two families of the *Actin depolymerizing proteins* fold:
  - Cofilin-like (CL) family (4 molecules)
  - Gelsolin-like (GL) family (8 molecules)

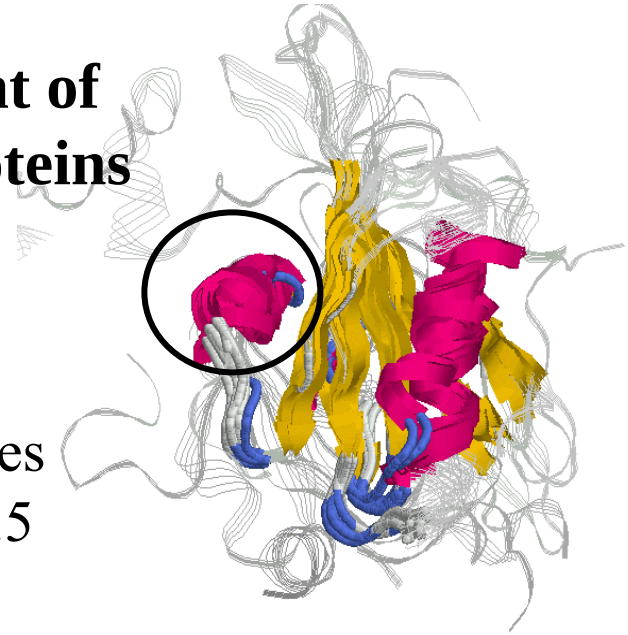
A. Alignment of all 12 proteins

28 residues
RMSD 1.9



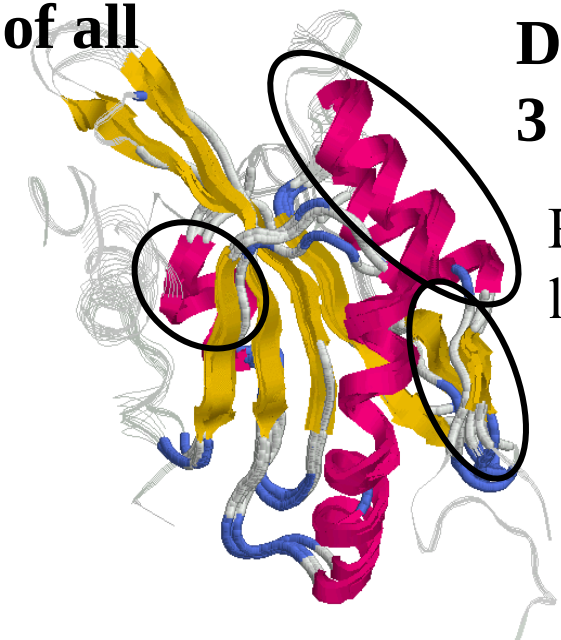
B. Alignment of all 8 GL proteins

63 residues
RMSD 1.5



C. Alignment of all 4 CL proteins

104 residues
RMSD 1.2



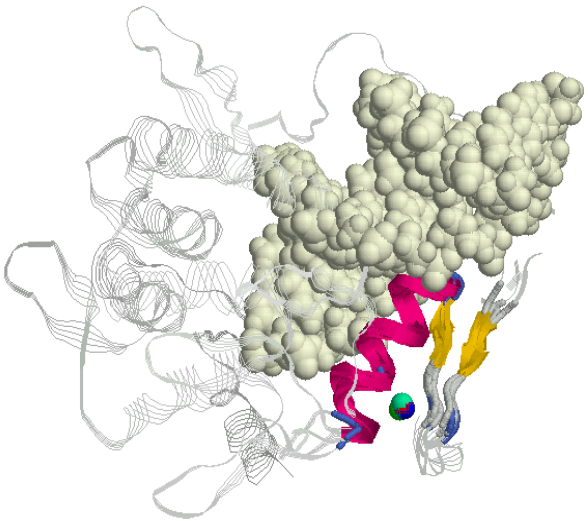
D. Alignment of 3 CL proteins

PDB:1f7s → lacks this helix

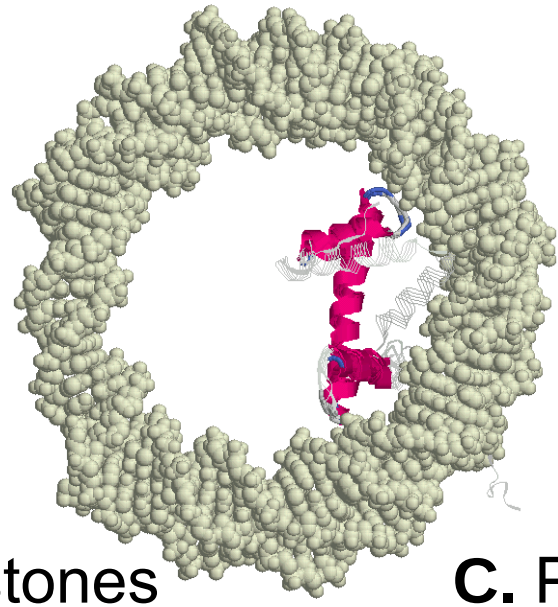
120 residues
RMSD 1.3

Classification of DNA-Binding Proteins

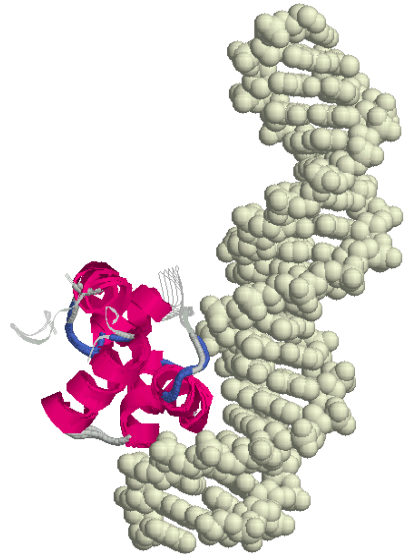
- The ensemble contains 18 DNA-binding proteins that can be classified into 5 structural classes:
  - Classic zinc finger (7 molecules)
  - Histones (3 molecules)
  - Phage repressors (3 molecules)
  - Restriction endonuclease-like (3 molecules)
  - Winged helix (3 molecules).



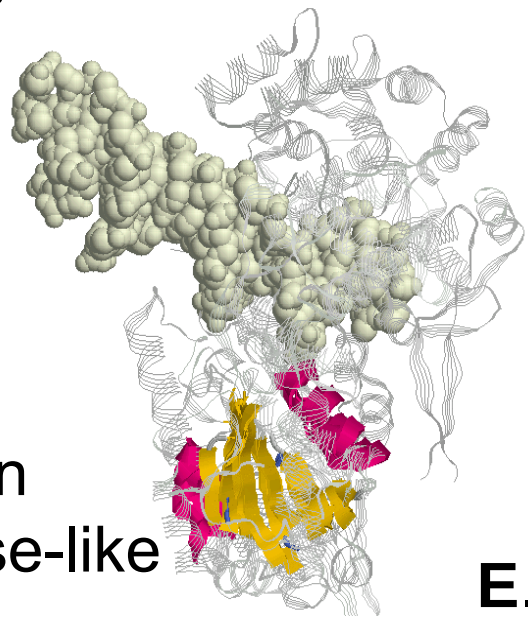
A. Zinc Finger



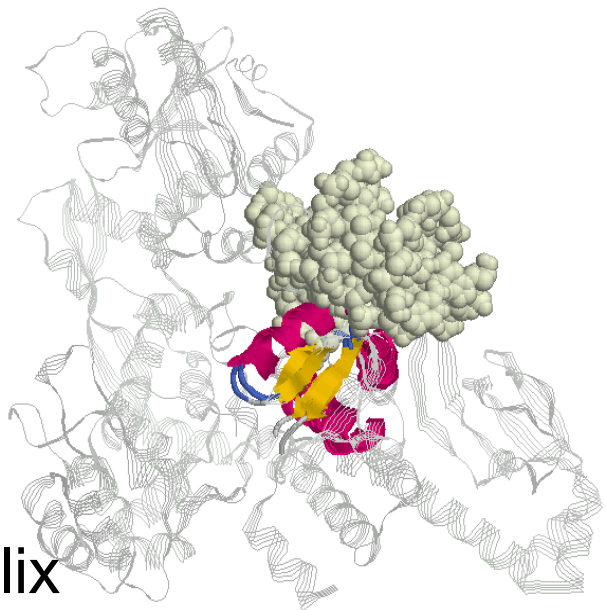
B. Histones



C. Phage repressors



D. Restriction endonuclease-like



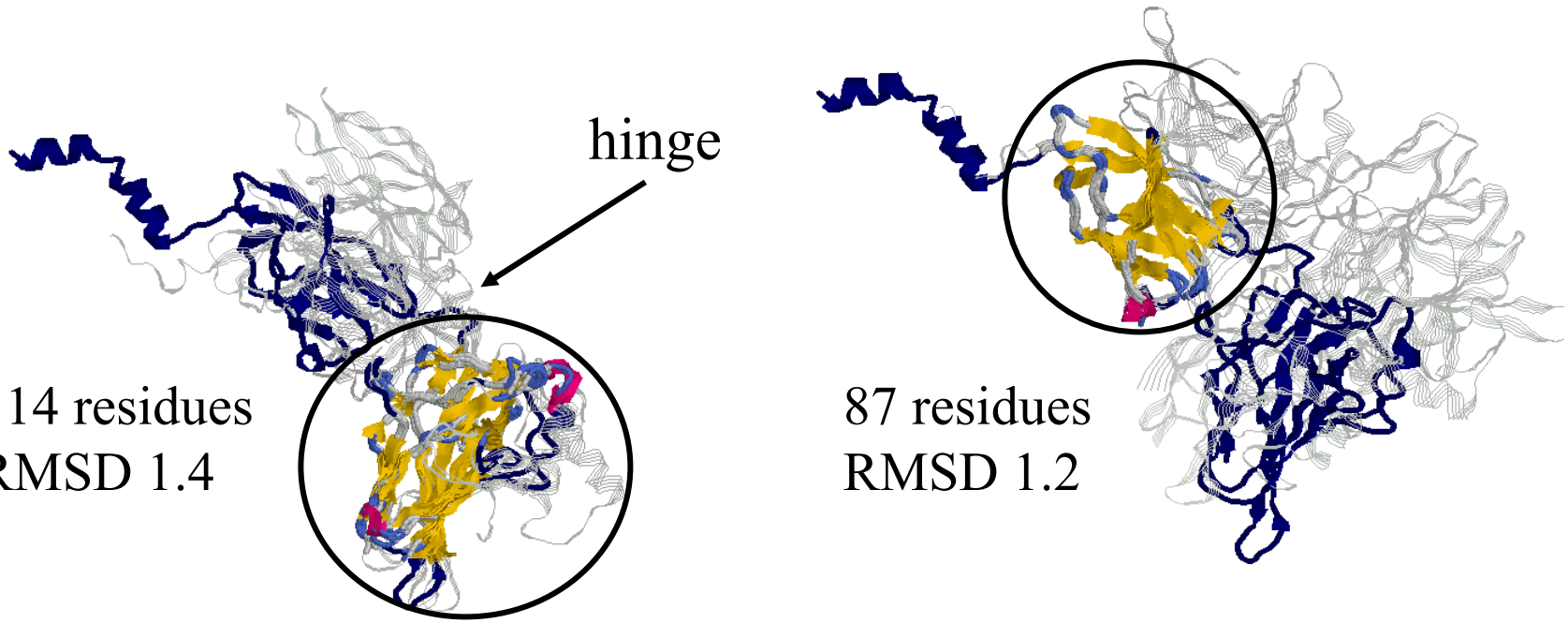
E. Winged Helix

- DNA

Partial Alignments

Detection of Two Common Domains

- Ensemble: 5 proteins that share two domains



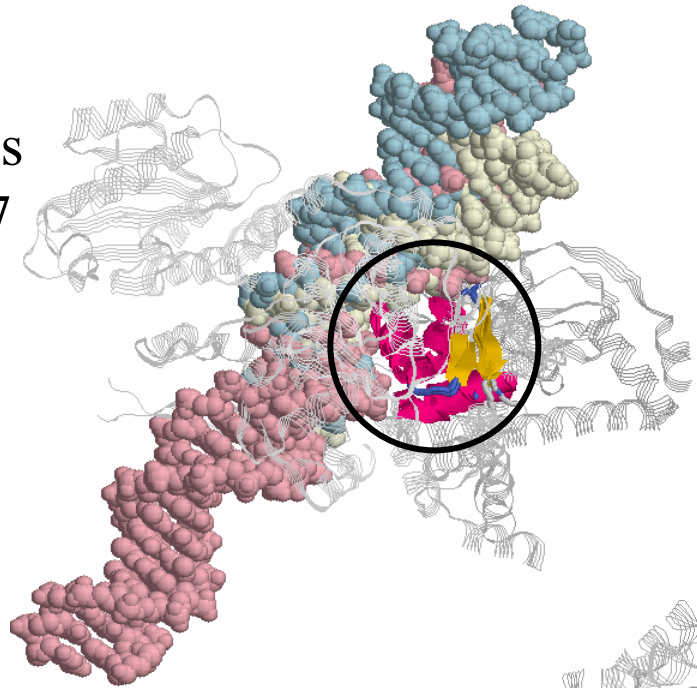
p53-like transcription factors **E set domains**

■ - PDB:1nfiA

Detection of Two Common Motifs

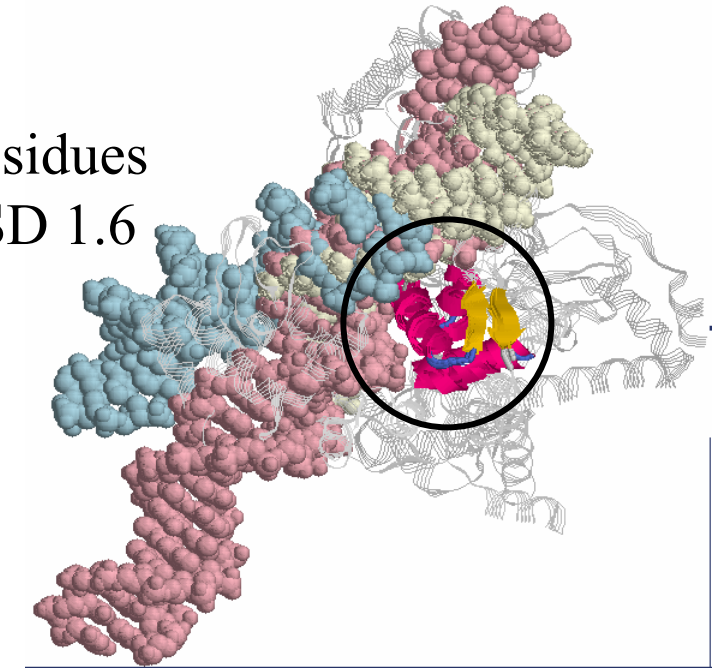
A.

46 residues
RMSD 1.7

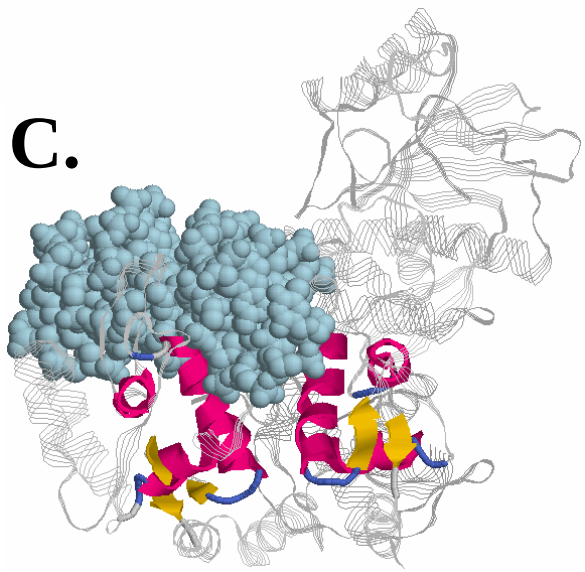


B.

45 residues
RMSD 1.6



C.



Winged-helix proteins

- DNA of PDB:1cgpA
- DNA of PDB:1ddnA
-DNA of PDB:1fokA

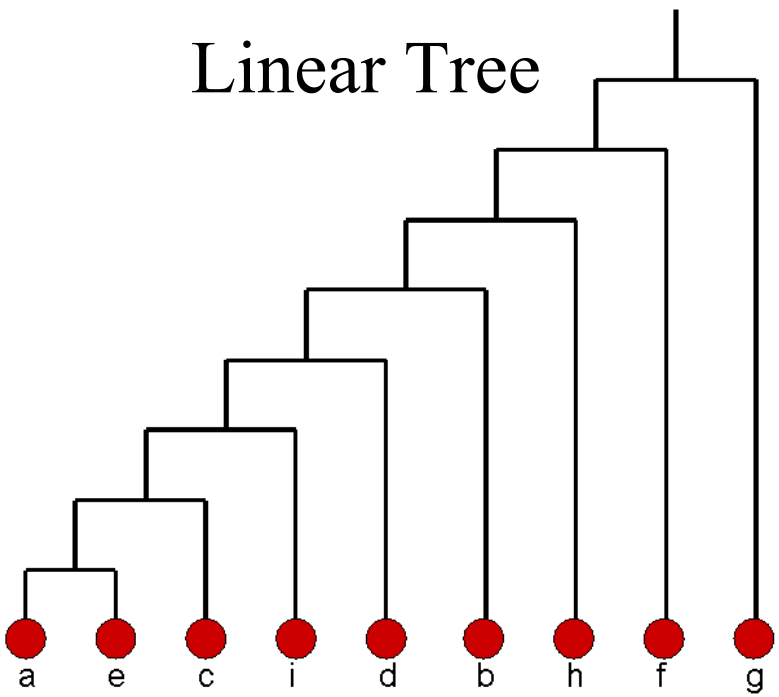
Pairwise-Based Methods

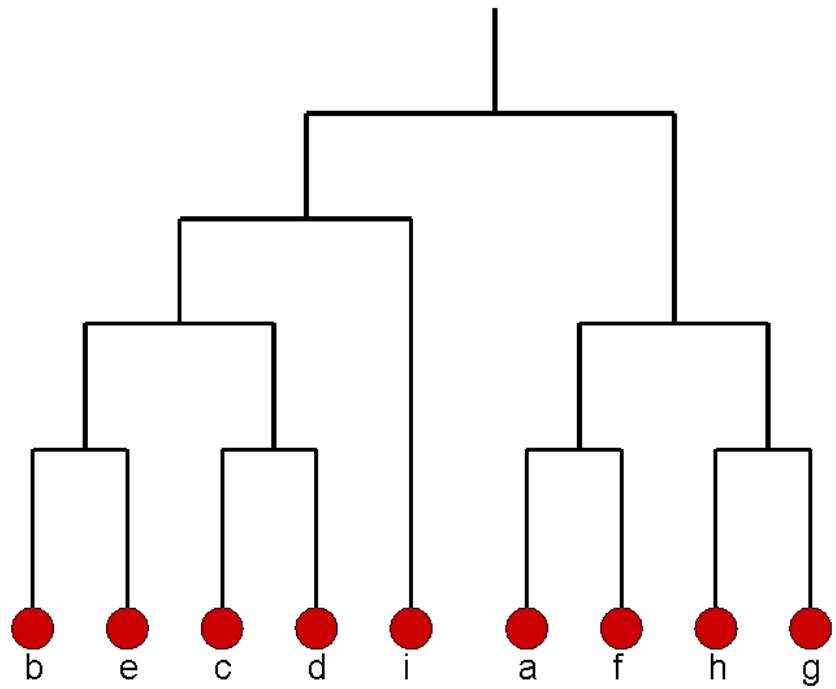
Two heuristics approaches

Center Star

Linear Tree

Progressive Tree





Pairwise-Based Methods

Two heuristics approaches

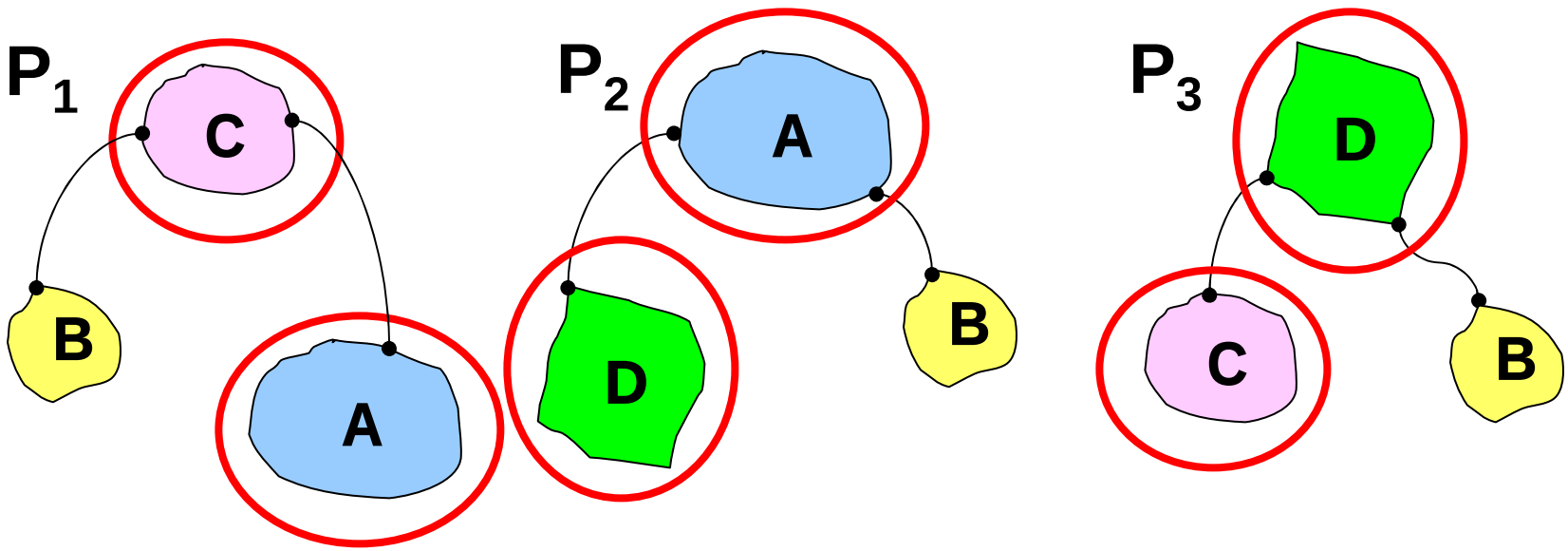
Center Star

- Gerstein & Levitt, 1996
- Akutsu & Sim, 1999
- Guda *et al.,* 2001

Progressive Tree

- COMPARER (Sali & Blundell, 1990)
- STAMP (Russell & Barton, 1992)
- Ding *et al.*, 1994
- (May & Johnson, 1995)
- SSAPm (Taylor *et al*., 1994)
- PrISM (Yang & Honig, 2000)

- Pros: provide a hierarchical clustering
- Cons: might miss significant alignments



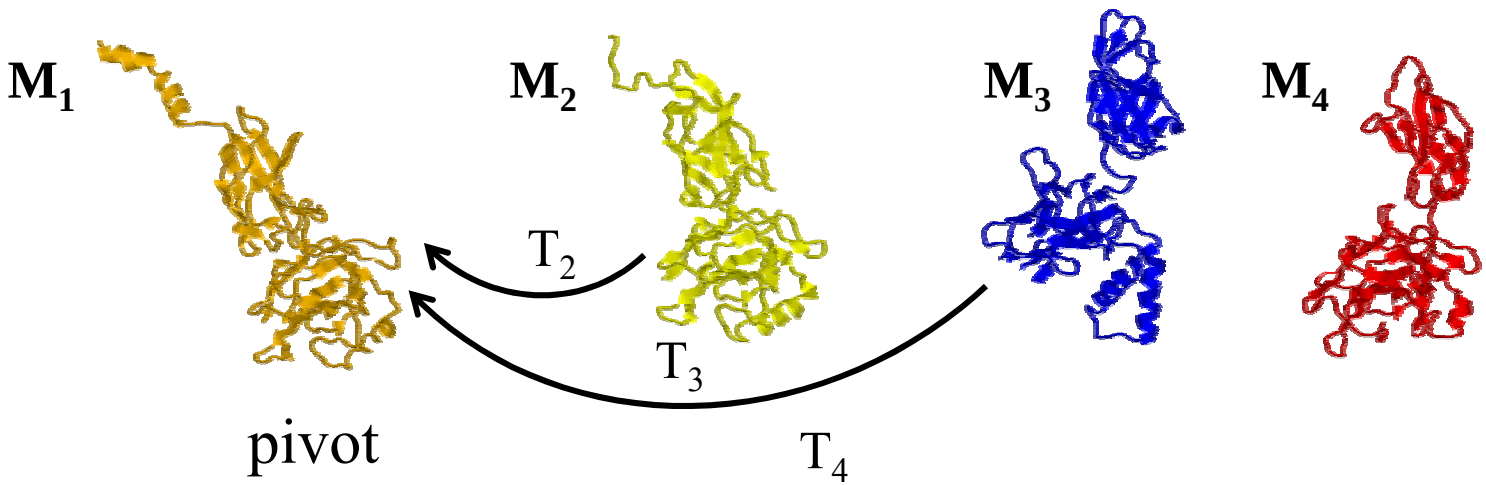
Substructure common to all structures: none instead of 

Simultaneous Methods

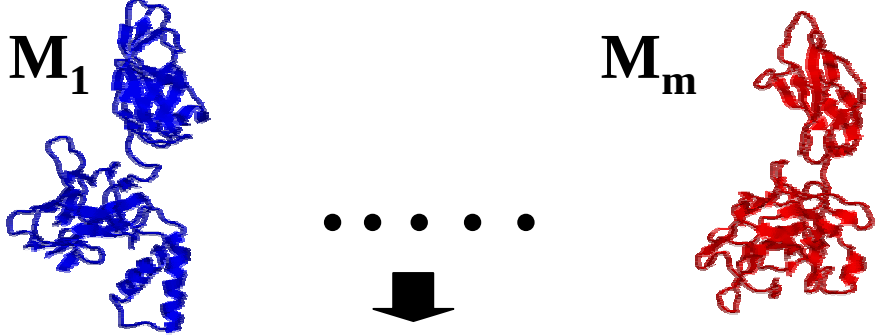
- Combinatorial Assembly (Escalier *et al.*, 1998)
- MUSTA (Leibowitz *et al.*, 2001)
- MultiProt (Shatsky *et al.*, 2002)
- MASS (Dror *et al.*, 2003)

MUSTA, MultiProt & MASS

- Using a pivot to define a multiple alignment as a vector of 3D rigid transformations:



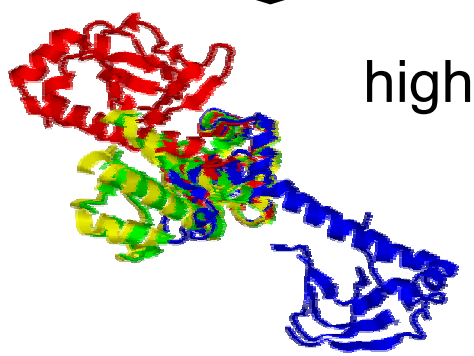
A multiple alignment between M_1,M_2,M_3 & M_4 is defined by (T_2, T_3, T_4)



Local Alignment

Clustering
and
Global Extension

Scoring



highest scoring multiple alignments

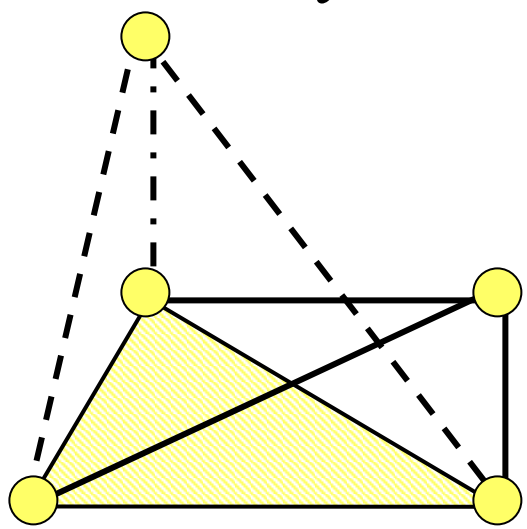
MUSTA:
MUltiple **ST**ructural **A**lignment

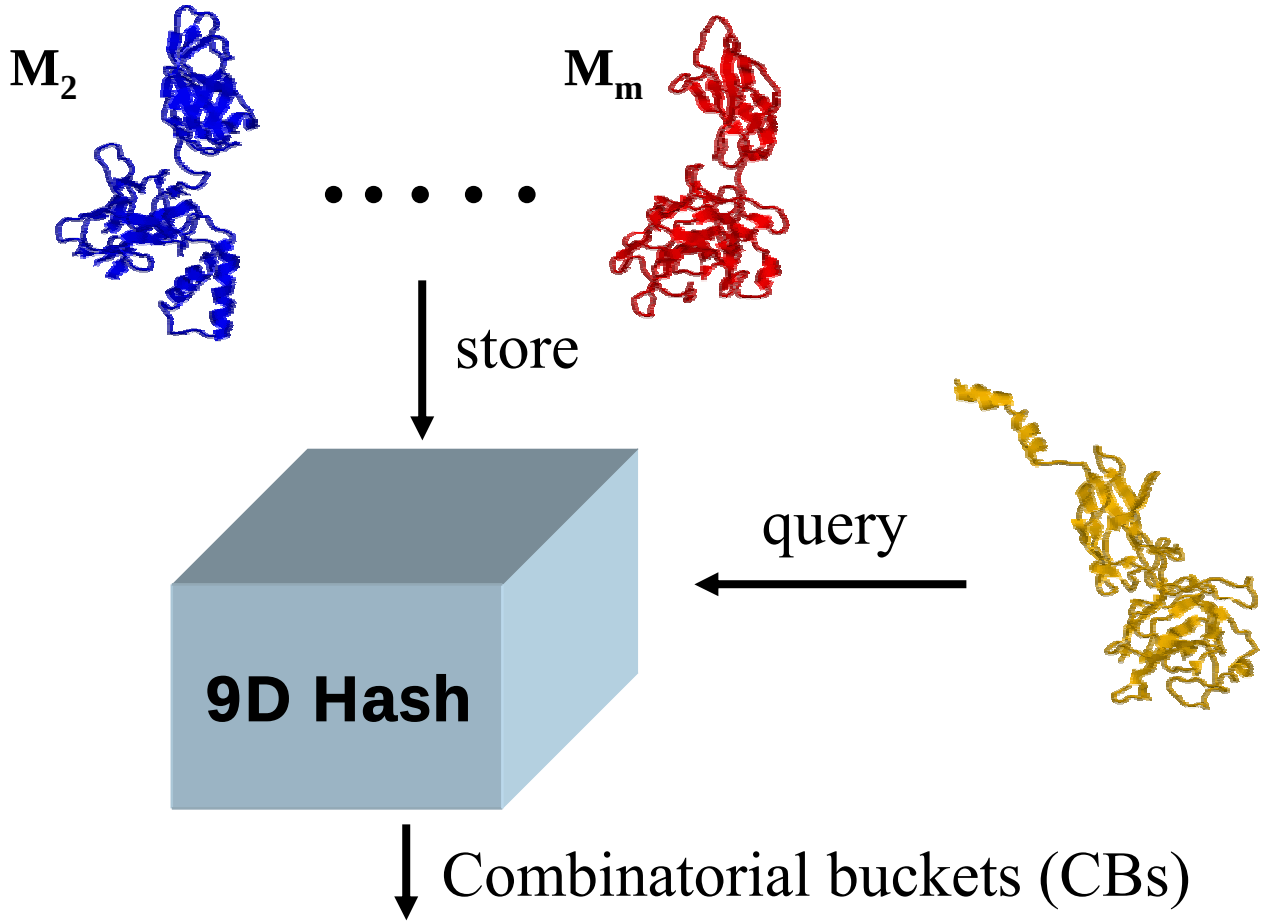
- Aligns **all** *m* input molecules
  - does not detect subset alignments
- The input is sets of 3D points
  - applicable to comparisons of drugs, protein active sites/surfaces/interfaces, etc.

- **Detection of Local Alignments:**

  Geometric Hashing:

  - Detects *k*-tuples of atoms whose configuration appears in all *m* molecules.
  - In practice k=5.
  - A 5-tuple is represented by a 9D vector of inter-distances.

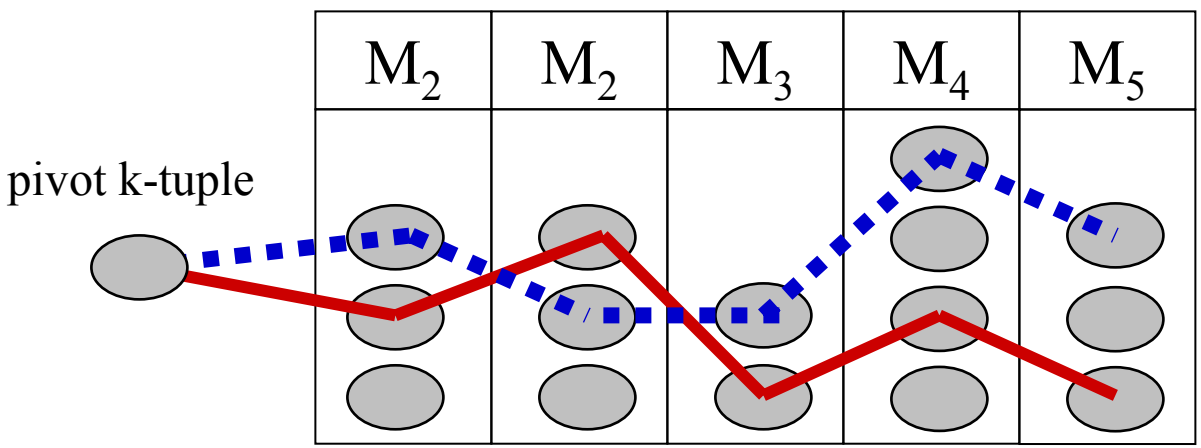



M2
Mm
store
query
9D Hash
Combinatorial buckets (CBs)
M2
M2
M3
M4
M5
pivot k-tuple

Combinatorial Buckets (CBs):

- Only buckets with k-tuples from all molecules are considered.
- A path of *k*-tuples, one from each column, defines a local multiple alignment.
- Worst Case: Exponential number of alignments.

(3•3•2•4•3) local alignments

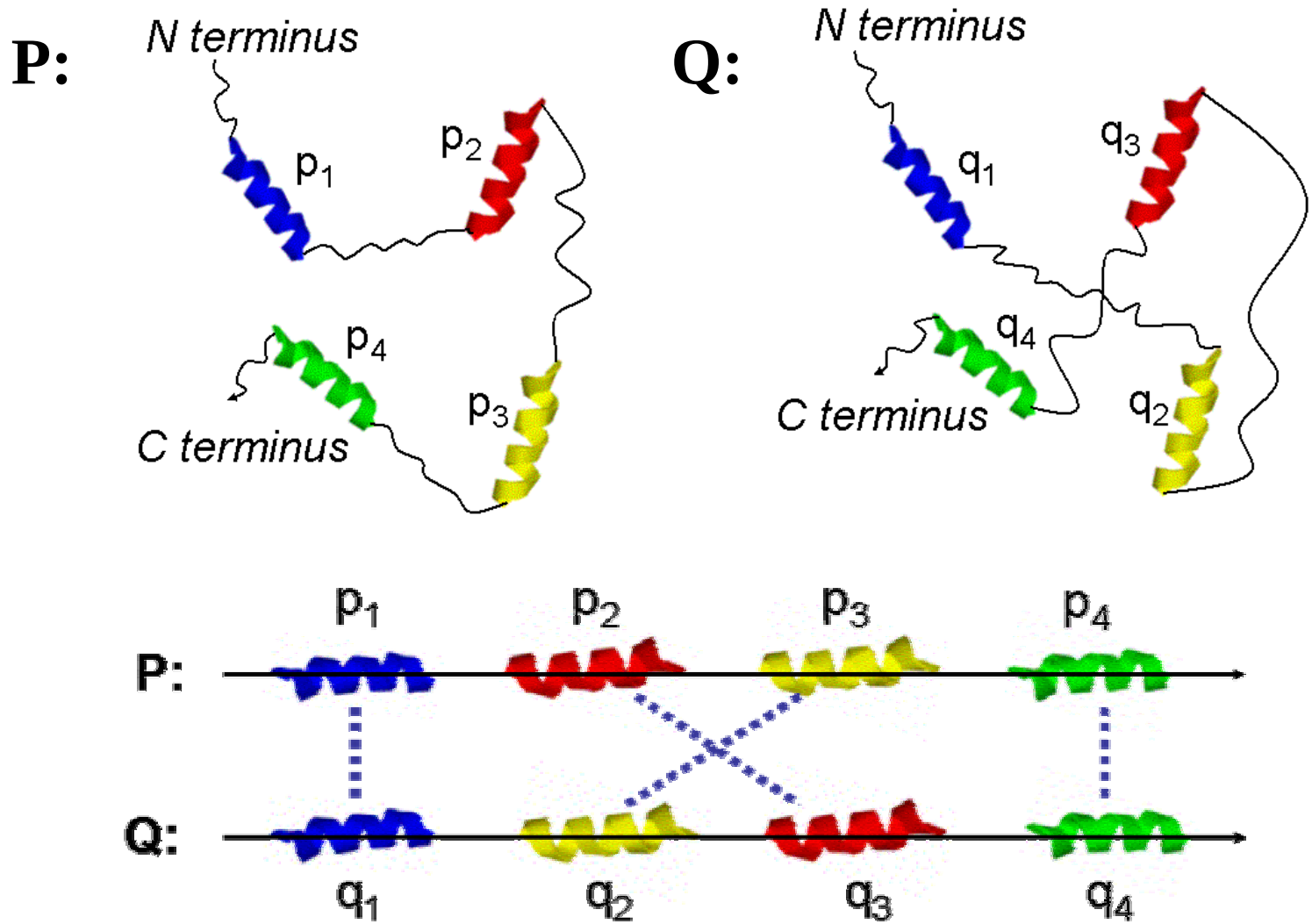


- Computes the pairwise transformations of the CBs
- **Clustering**
  - Criteria: transformations of the same cluster map relevant atoms to almost the same location
  - Distance between transformations = number of atoms mapped to different locations
  - Reduces the complexity of the CBs
- **Scoring**
  - Alignment's score = the size of the match list

MASS: **M**ultiple 3D **A**lignment by **S**econdary **S**tructures

- Considers all structures **simultaneously**
- Capable of detecting **subset alignments**
- Exploits secondary structure information
  - **Stability:** proteins are inherently composed of secondary structure elements (SSEs)
  - **Efficiency**: introduces great savings in structural description
  - **Accuracy:** filters noisy results

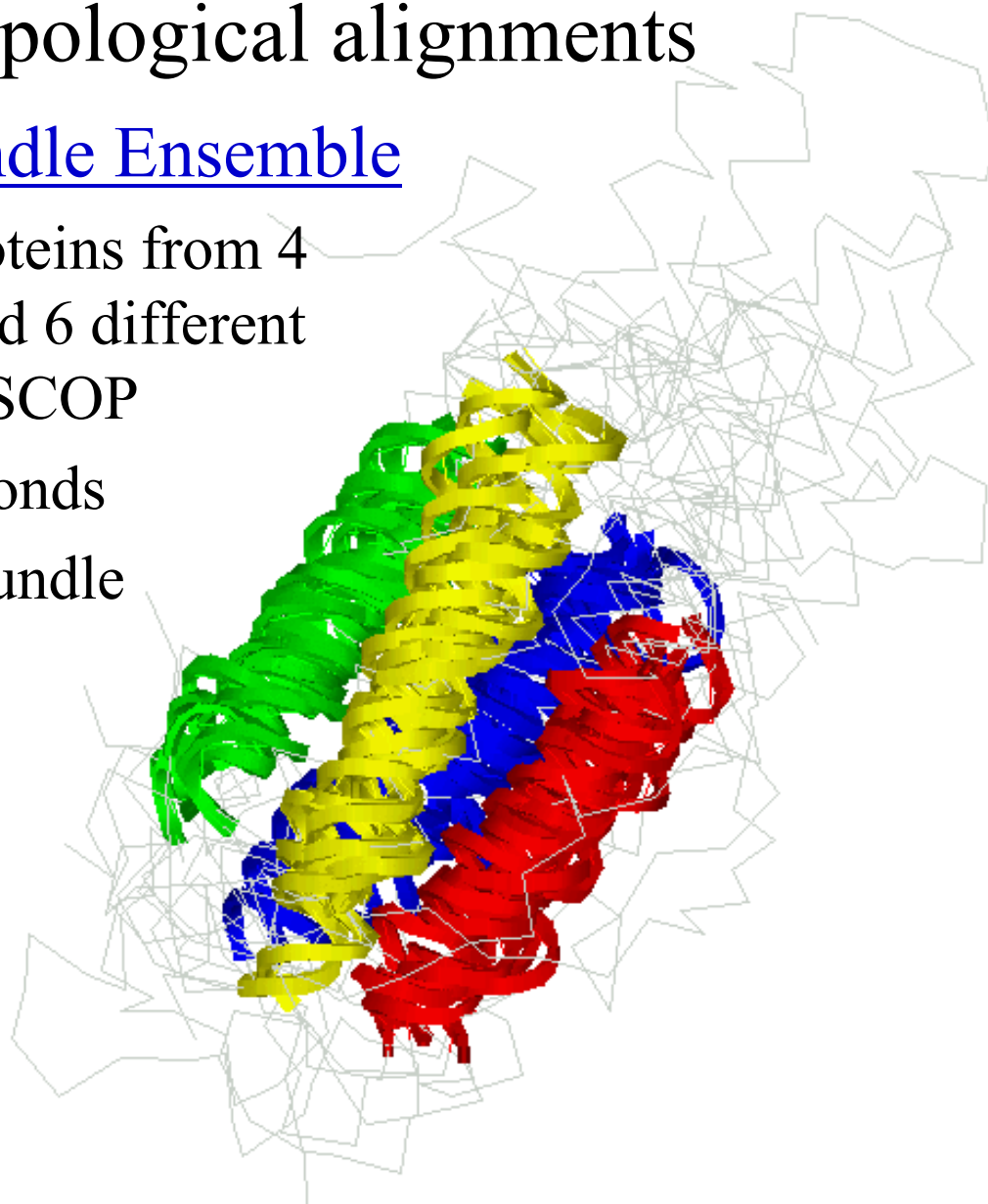
- Sequence-order independent



- Can detect non-topological alignments

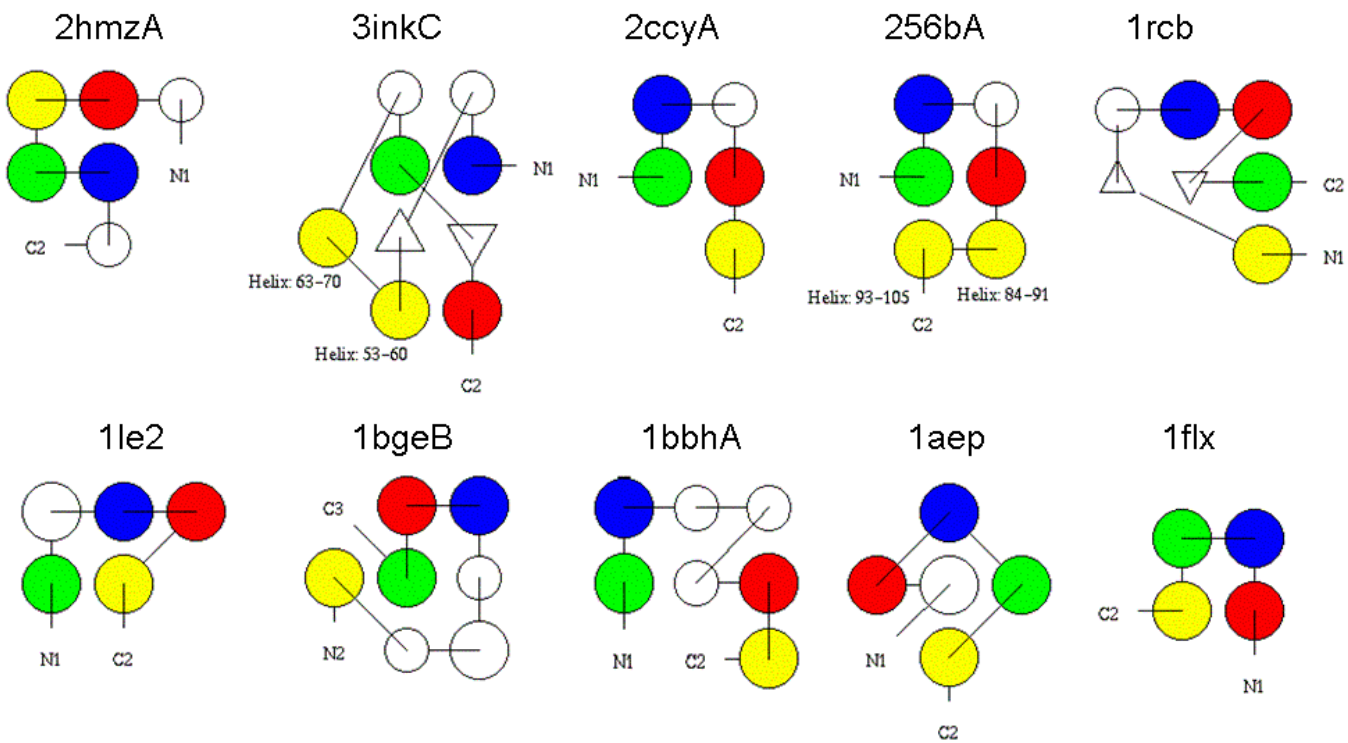
  Example: Helix-bundle Ensemble

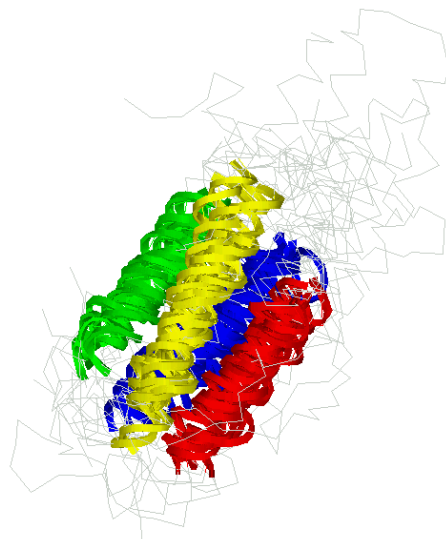
  - Ensemble: 10 proteins from 4 different folds and 6 different superfamilies in SCOP
  - Runtime: 48 seconds
  - Core: 4-helical bundle



2hmzA	3inkC	2ccyA	256bA	1rcb	1le2	1bgeB	1bbhA	1aep	1flx
H1	H6	H3	H3	H3	H3	H5	H5	H1	**H0**
H2	H2, H3	H4	H4, H5	H0	H4	H0	H6	H4	**H3**
H3	H5	H0	H0	H4	H0	H6	H0	H3	**H2**
H4	H0	H1	H1	H2	H2	H4	H1	H2	**H1**

Match List: H - helix





TOPS Diagrams: ○ - helices ; △ - strands

SSE Assignment

SSE Representation

Local Basis Alignment

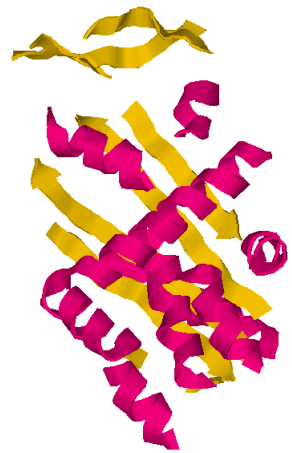
Clustering

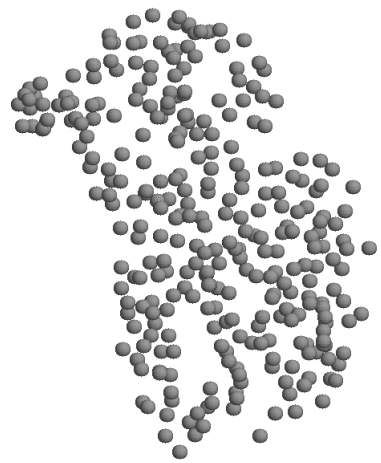
Global Extension

Filtering and Scoring

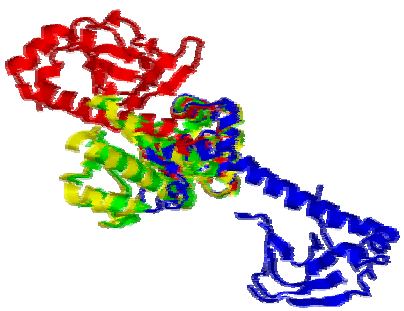
SSE Level

C_α Atomic Level

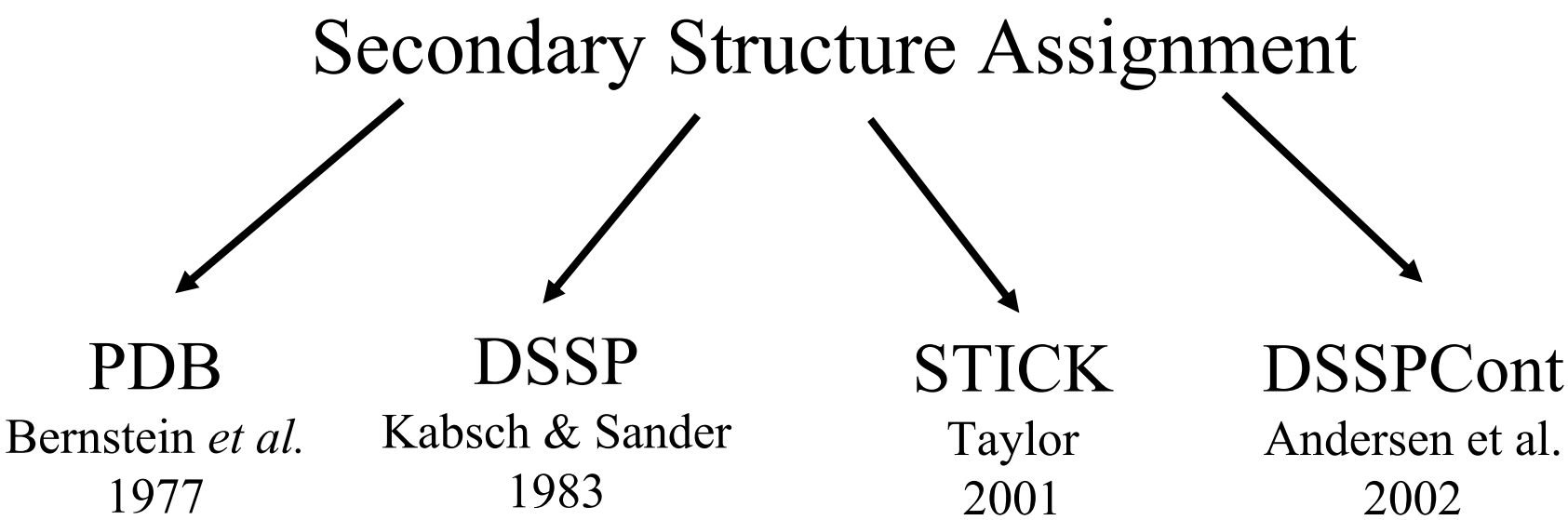




The highest scoring multiple alignments

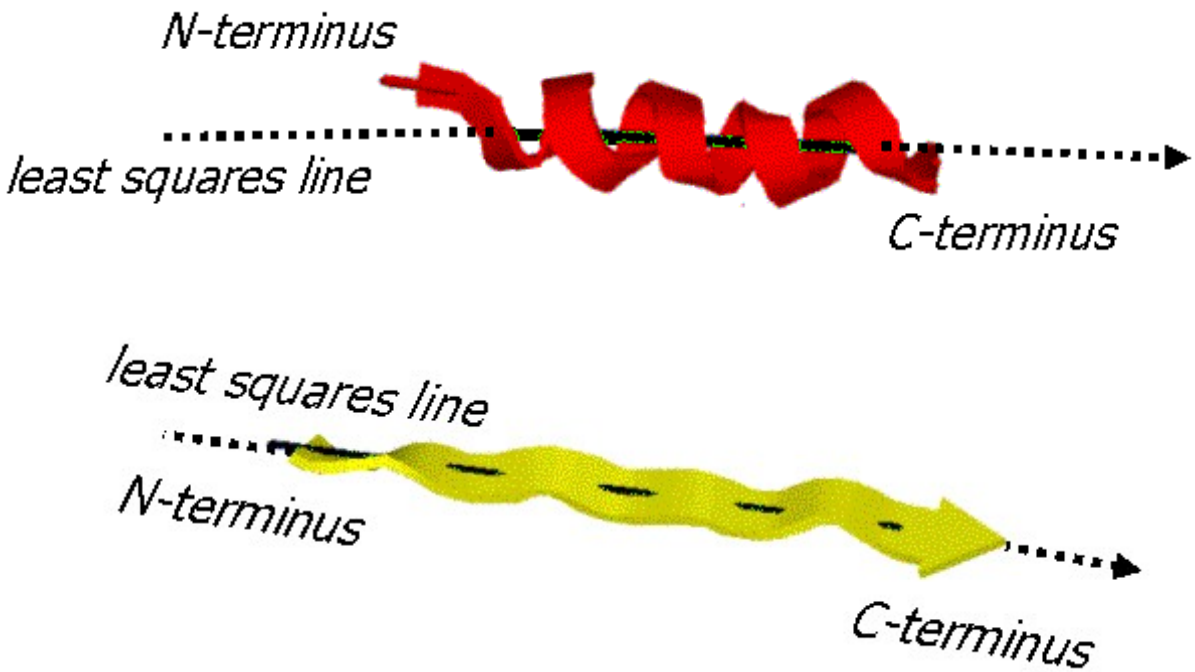


SSE Assignment

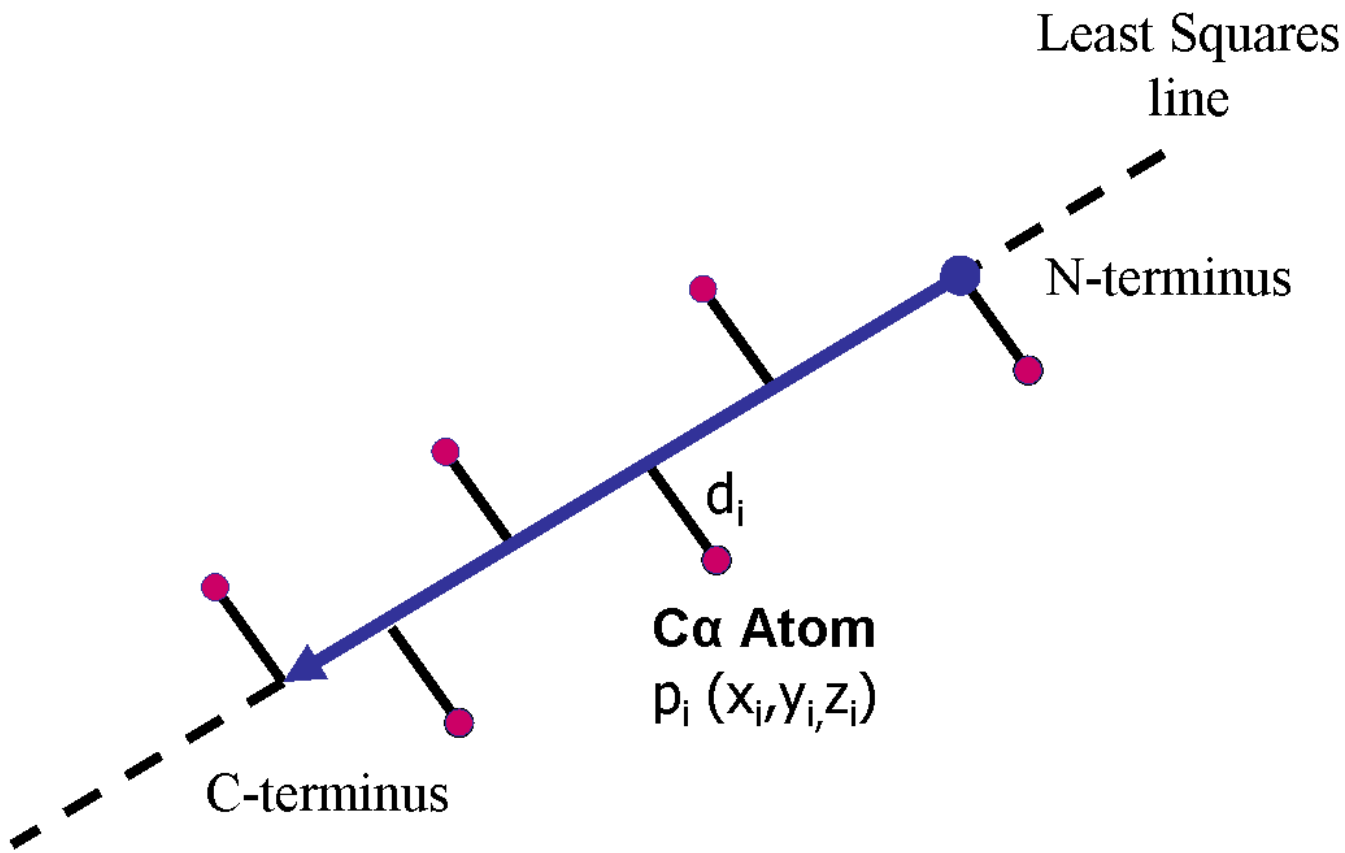


SSE Representation

- An SSE is represented by a 3D line segment with fuzzy endpoints



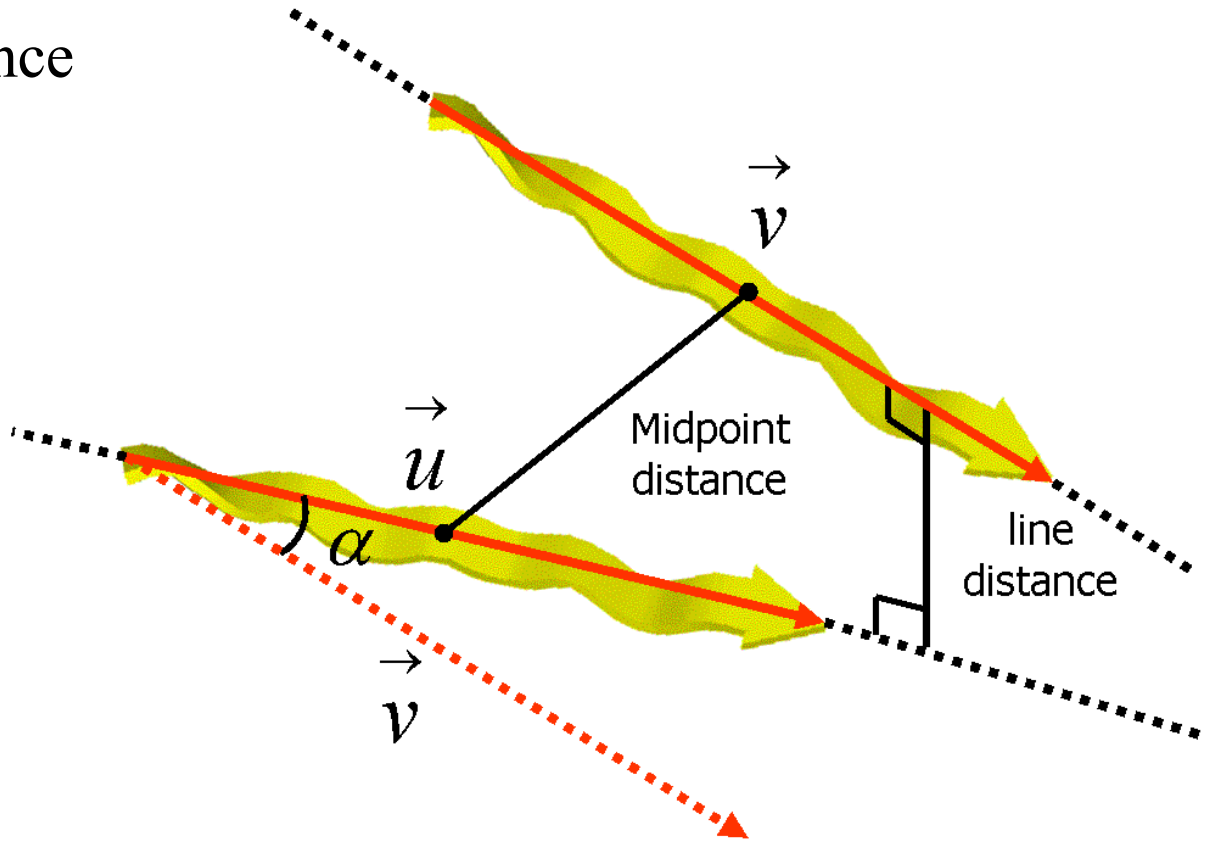
- The SSE *least-square line* minimizes $\sum_i d_i^2$



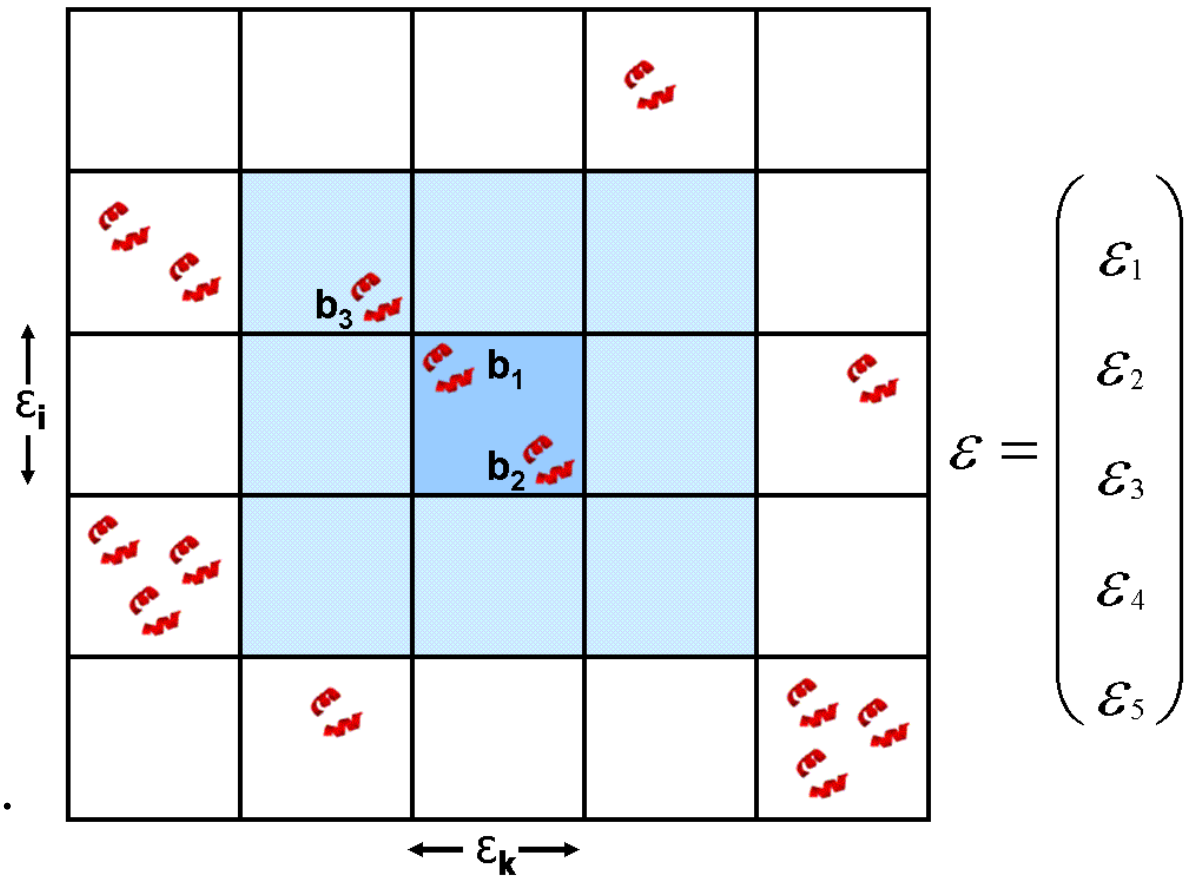
Detection of Local Basis Alignments

- A ***basis*** – an ordered pair of SSEs.
- Aim: Detection of bases with similar 3D configuration that appear in several proteins
- Geometric Hashing:
  - Each basis is represented by a ***fingerprint*** which is invariant to 3D rotations and translations.

- A basis fingerprint is a 5D vector composed of:
  - SSE types: helix, strand
  - Midpoint distance
  - Line distance
  - Angle

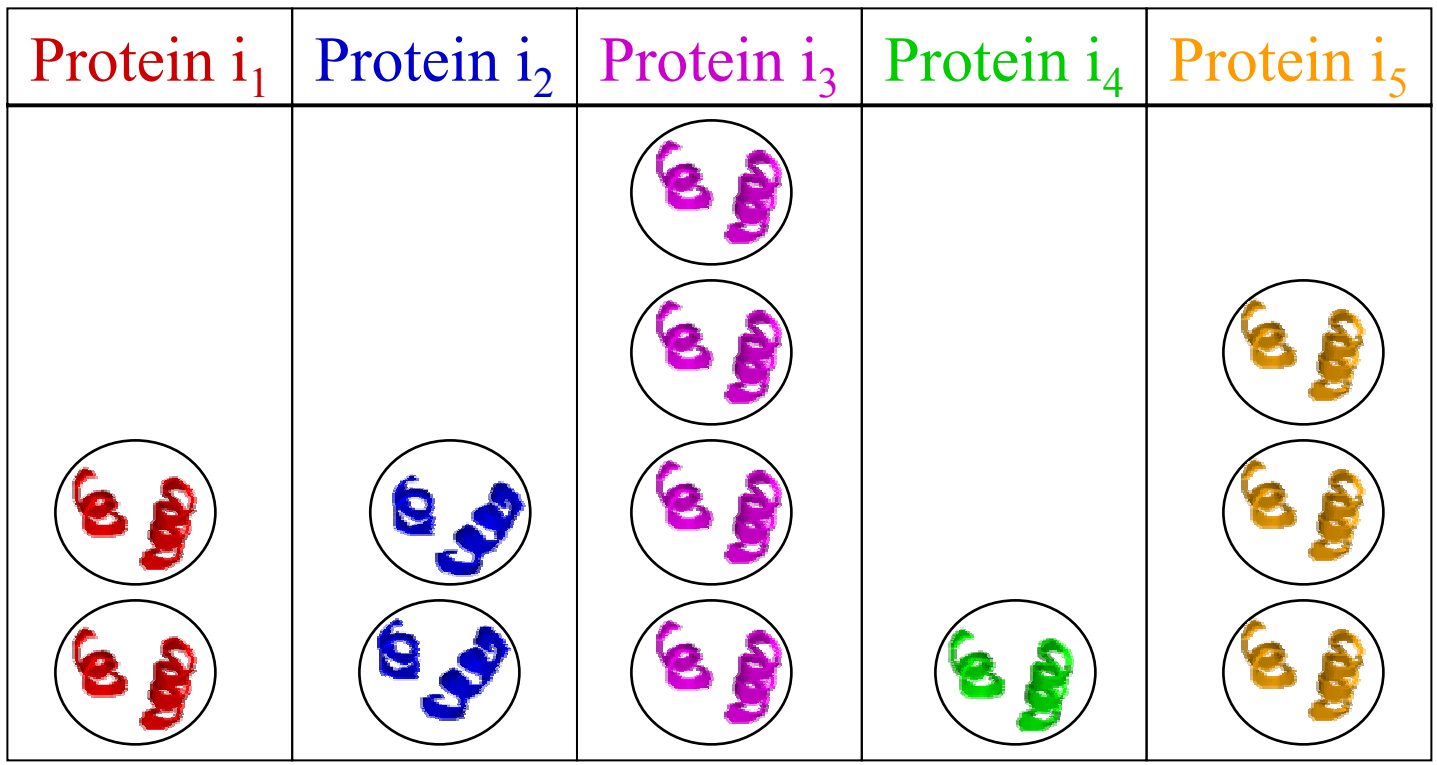


- The bases of all proteins are stored in a 5D grid, addressed by their fingerprints.

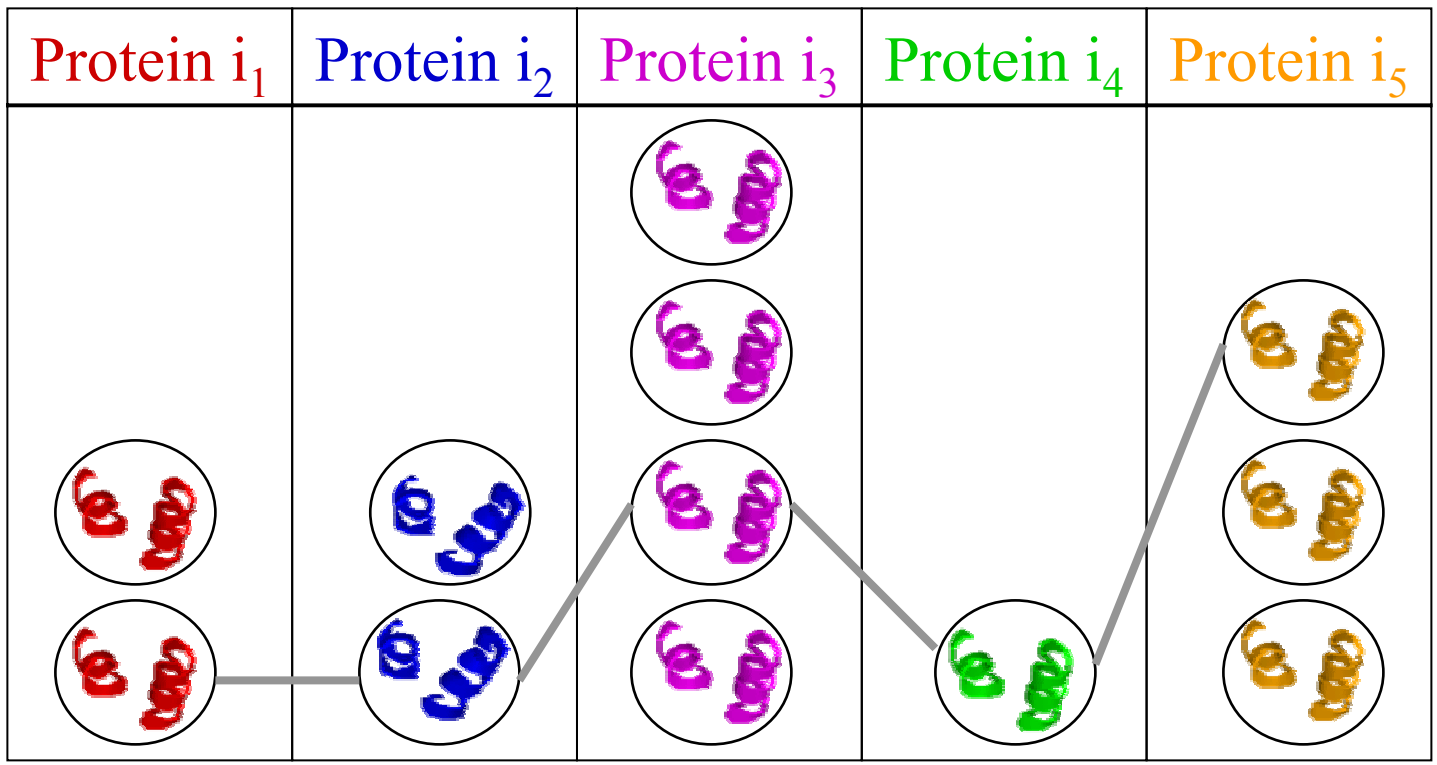


2D Cut.

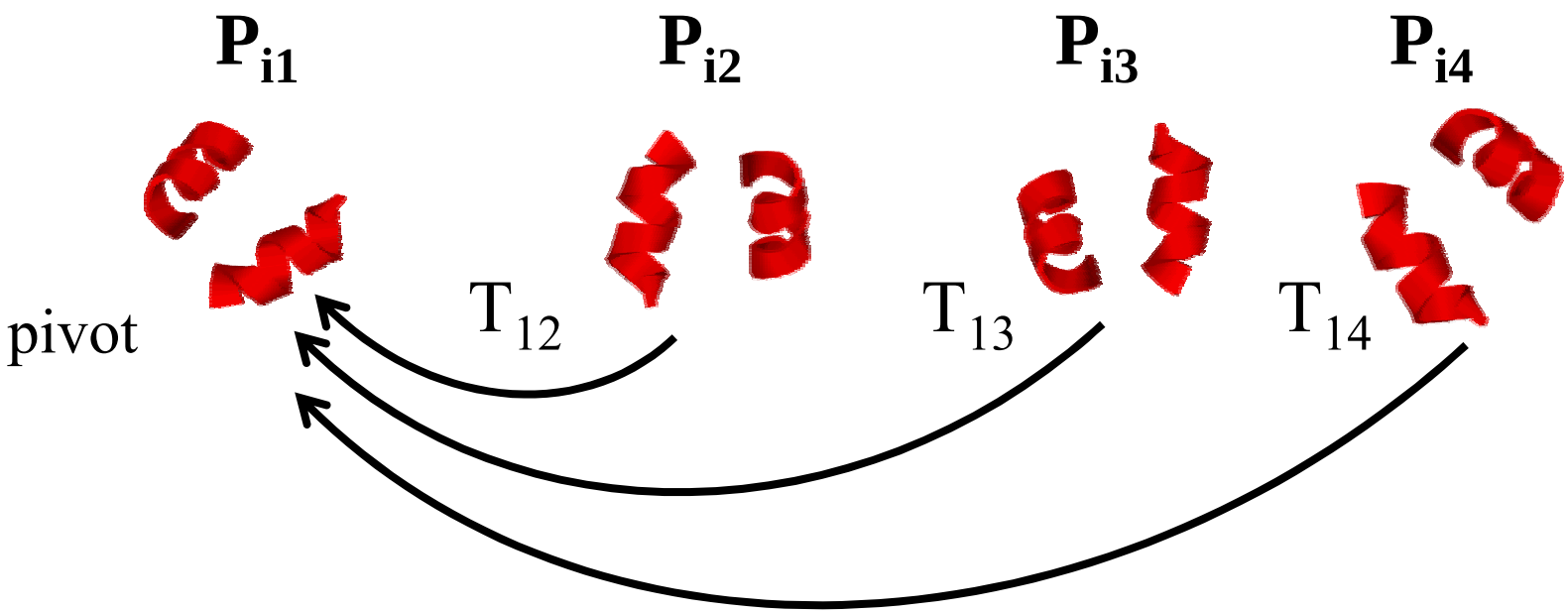
- For each bin in the grid:
  - Retrieve all the bases in that bin and in the adjacent bins and store them in a ***Basis Bucket***



- Bases from different columns define a multiple local alignment between the respective proteins.
- Exponential number of local basis alignments



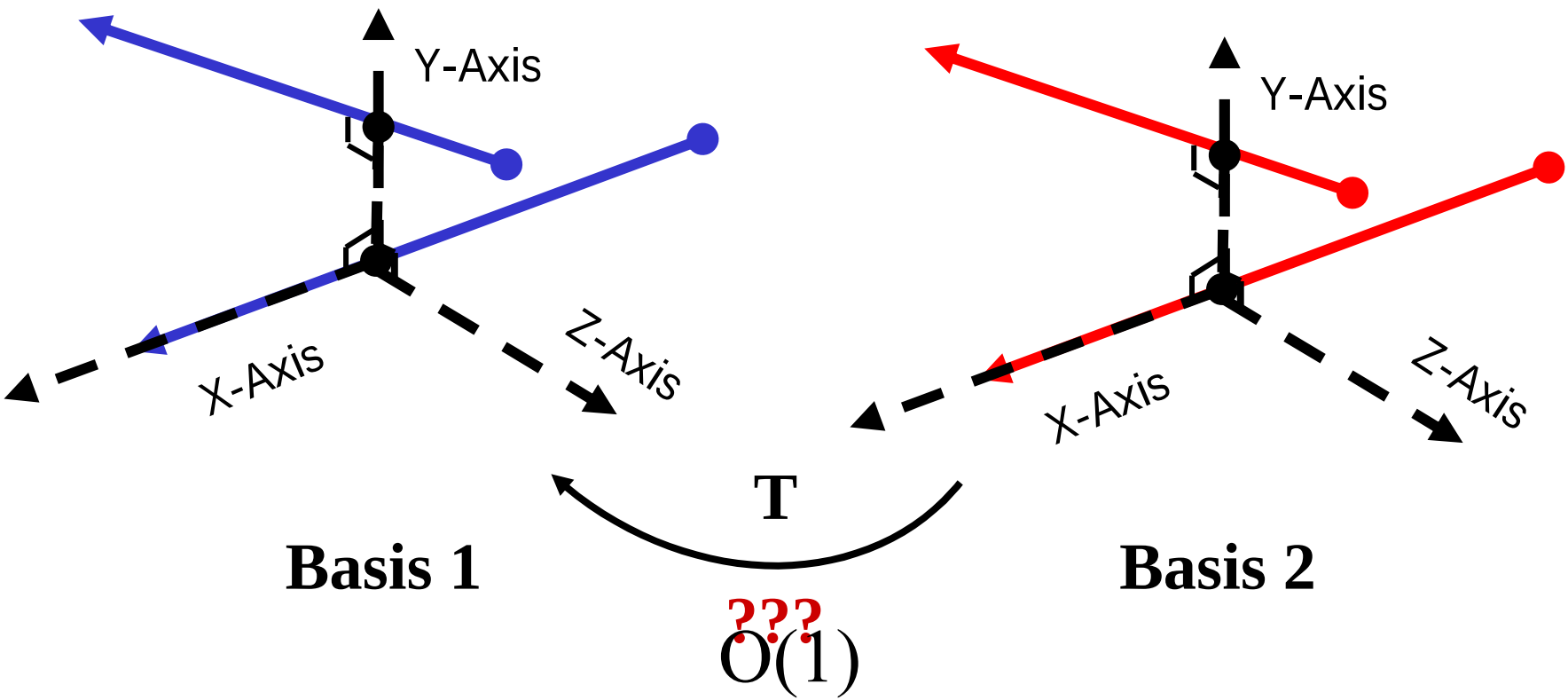
- How do we compute a multiple basis alignment?



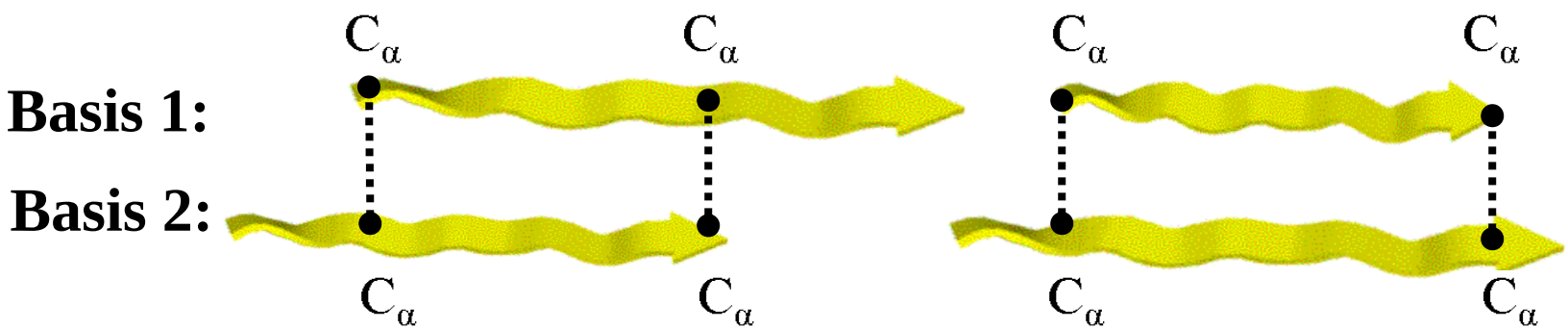
(T_{12}, T_{13}, T_{14})

- How do we align two bases?
  - **Reference Frame Superposition**

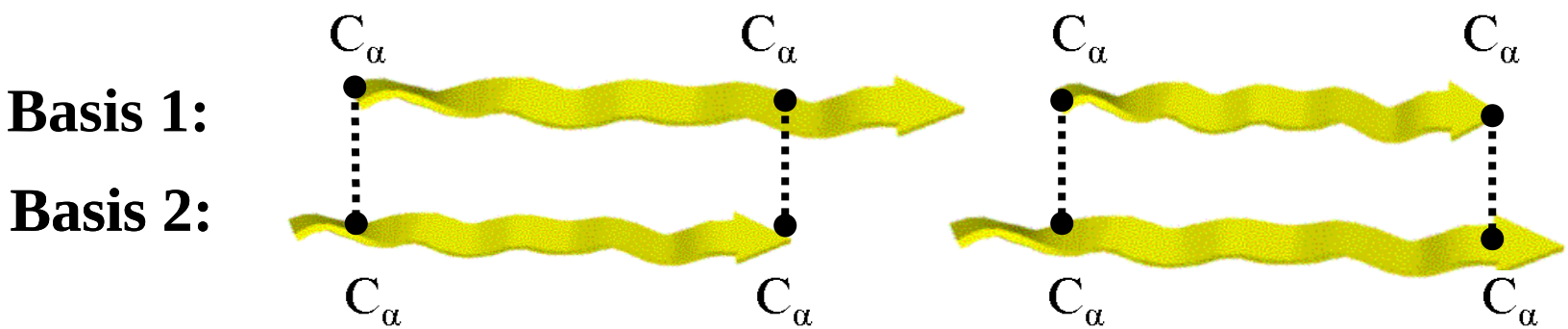
    For each basis we define a Cartesian reference frame



- **Atomic Superposition:**
  - Each SSE is represented by the list of its C_α atoms
  - We iterate over all possibilities of simultaneously aligning the two pairs of atom lists.
  - Atoms of matched SSEs are aligned consecutively
  - Transformations are computed by a Least-Squares Fitting technique.
  - The alignment with the largest match list and with RMSD < ε is selected.

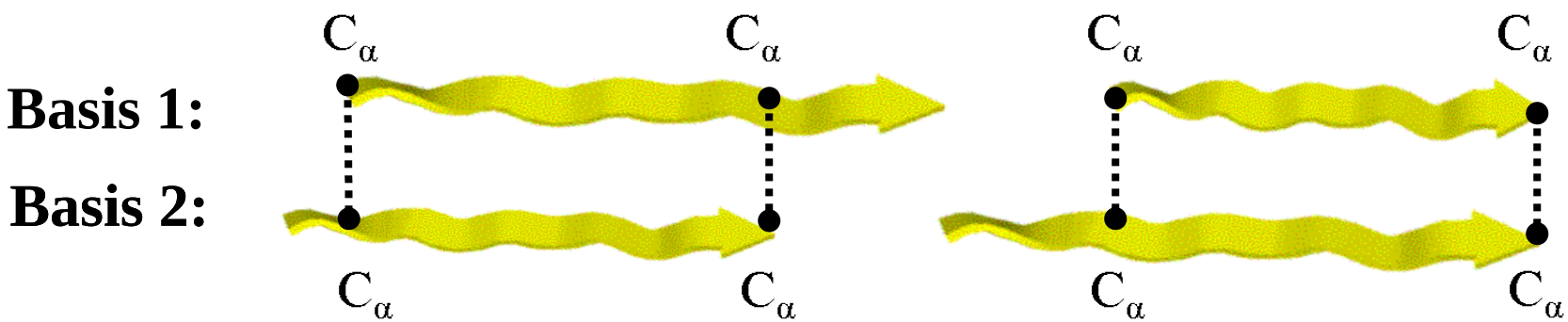


- **RMSD Minimization**
  - Each SSE is represented by the list of its C_α atoms
  - We iterate over all possibilities of simultaneously aligning the two pairs of atom lists.
  - Atoms of matched SSEs are aligned consecutively
  - Transformations are computed by a Least-Squares Fitting technique.
  - The alignment with the largest match list and with RMSD < ε is selected.

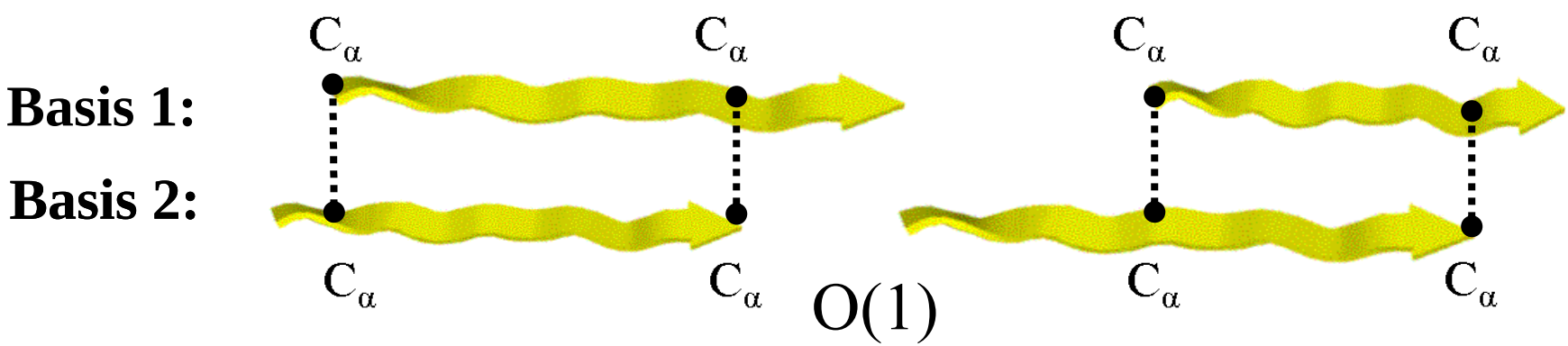


RMSD Minimization

- Each SSE is represented by the list of its C_α atoms
- We iterate over all possibilities of simultaneously aligning the two pairs of atom lists.
- Atoms of matched SSEs are aligned consecutively
- Transformations are computed by a Least-Squares Fitting technique.
- The alignment with the largest match list and with RMSD < ε is selected.

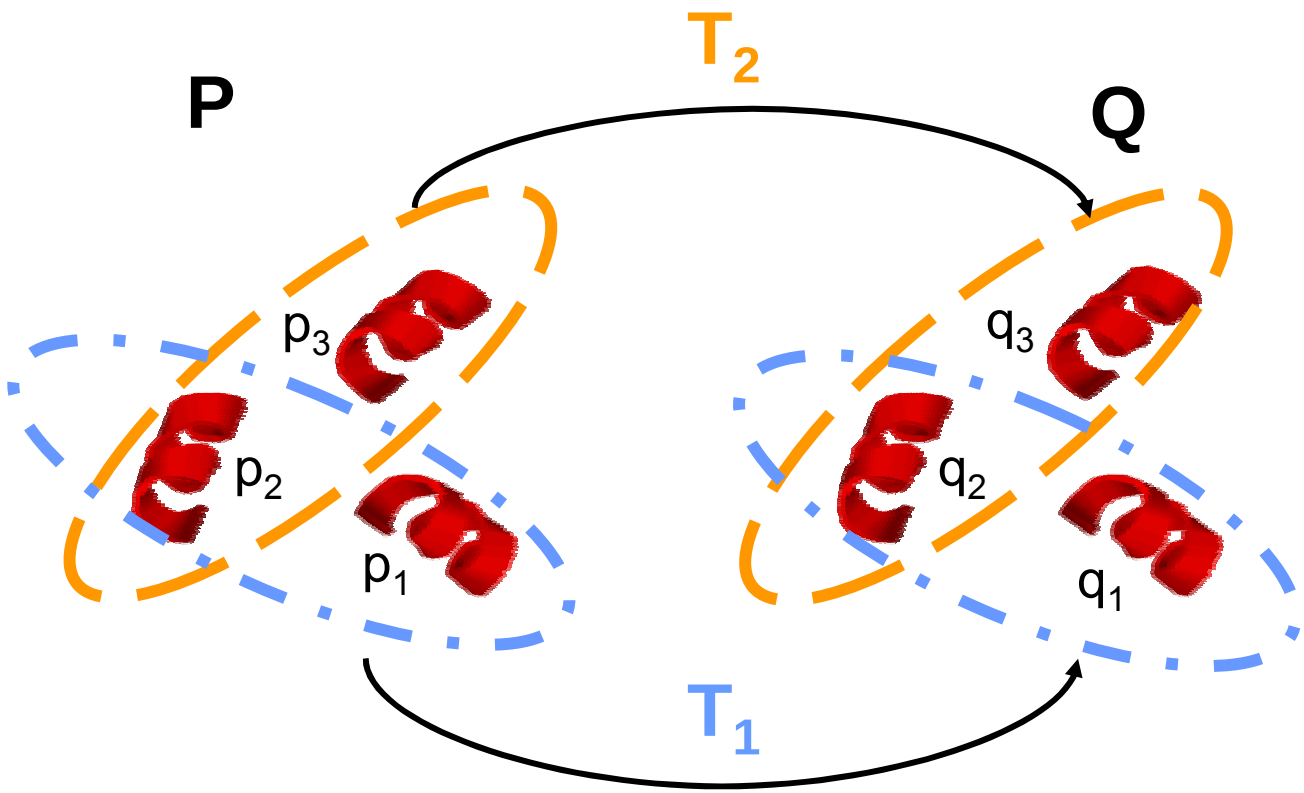


- **RMSD Minimization**
  - Each SSE is represented by the list of its C_α atoms
  - We iterate over all possibilities of simultaneously aligning the two pairs of atom lists.
  - Atoms of matched SSEs are aligned consecutively
  - Transformations are computed by a Least-Squares Fitting technique.
  - The alignment with the largest match list and with RMSD < ε is selected.



Clustering

- RMSD Clustering: Similar to (Rarey et al., 1996)
- $O(p^2 \log p)$ where p = number of alignments



- **RMSD Clustering**: Similar to (Rarey et al., 1996)

$$G = (V, E)$$

$$V = \{T_i\}$$

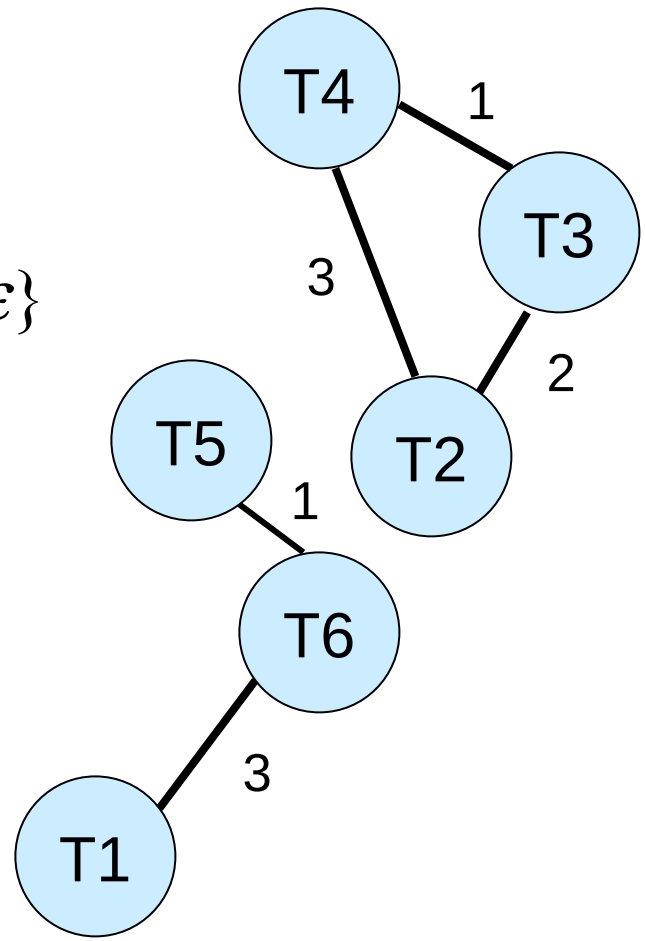
$$E = \{e = edge(T_i, T_j) : Dist(T_i, T_j) \leq \varepsilon\}$$

$$\omega(e) = \omega(edge(T_i, T_j)) = Dist(T_i, T_j)$$

$$where:$$

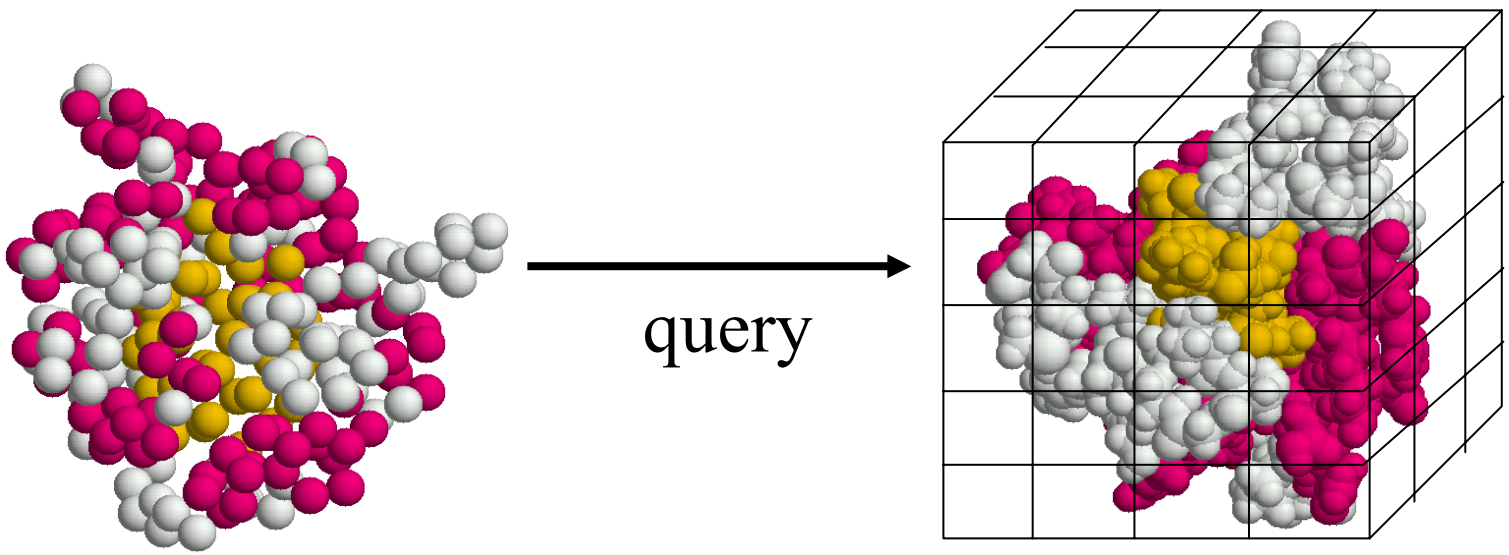
$$Dist(T_i, T_j) = RMSD(T_i(S), T_j(S))$$

$$S \subseteq Pivot$$



Residue Extension

- Extending the core of the alignments by detecting corresponding C_α atoms.



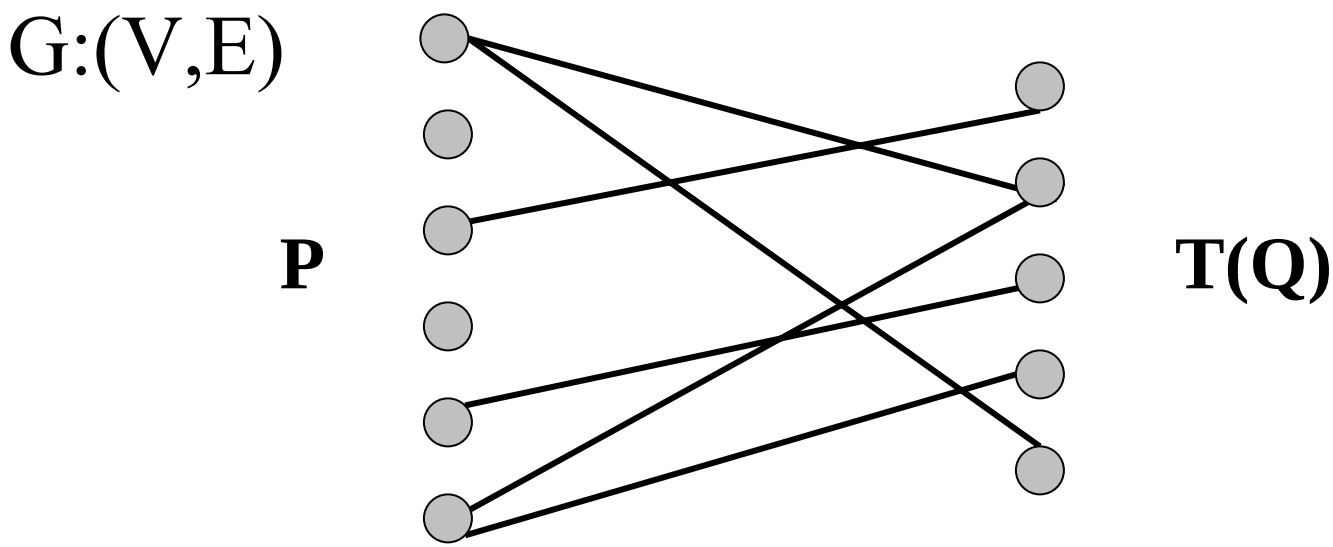
- Maximum matching in a bipartite graph

$$O(|G| + \sqrt{|V|}\,|E|) = O(n + n\sqrt{n}) = O(n\sqrt{n})$$

for ε =3Å

- Heuristic $O(n)$

n -num of residues in a protein



Computing the Best Global Multiple Alignments

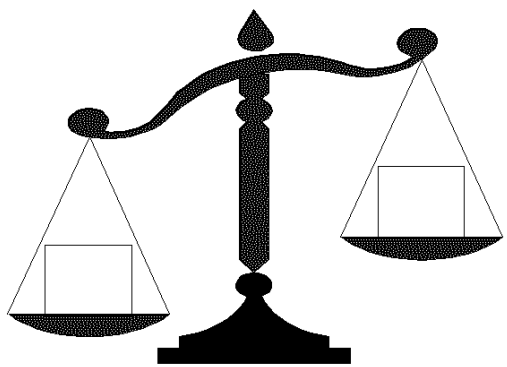
- What are the best global alignments?

  No absolute answer.

- Tradeoff:

  Number of aligned molecules vs. core size

molecules



core size

- Two approaches to address the trade-off:
  - The score of an alignment is defined as:

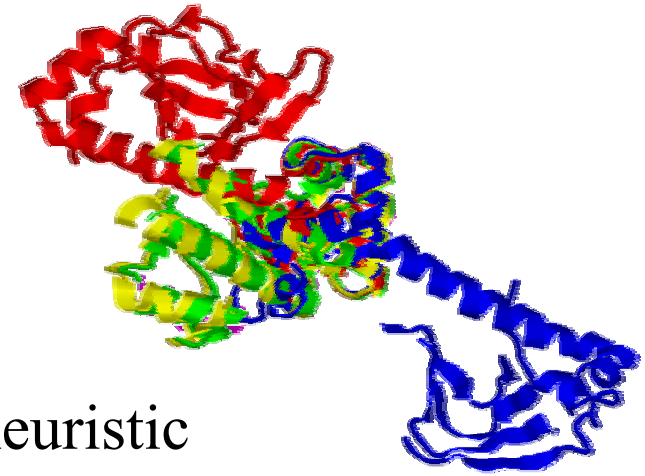
$$F(k,l) = k \cdot \binom{l}{2}$$

l - # molecules
k – core size

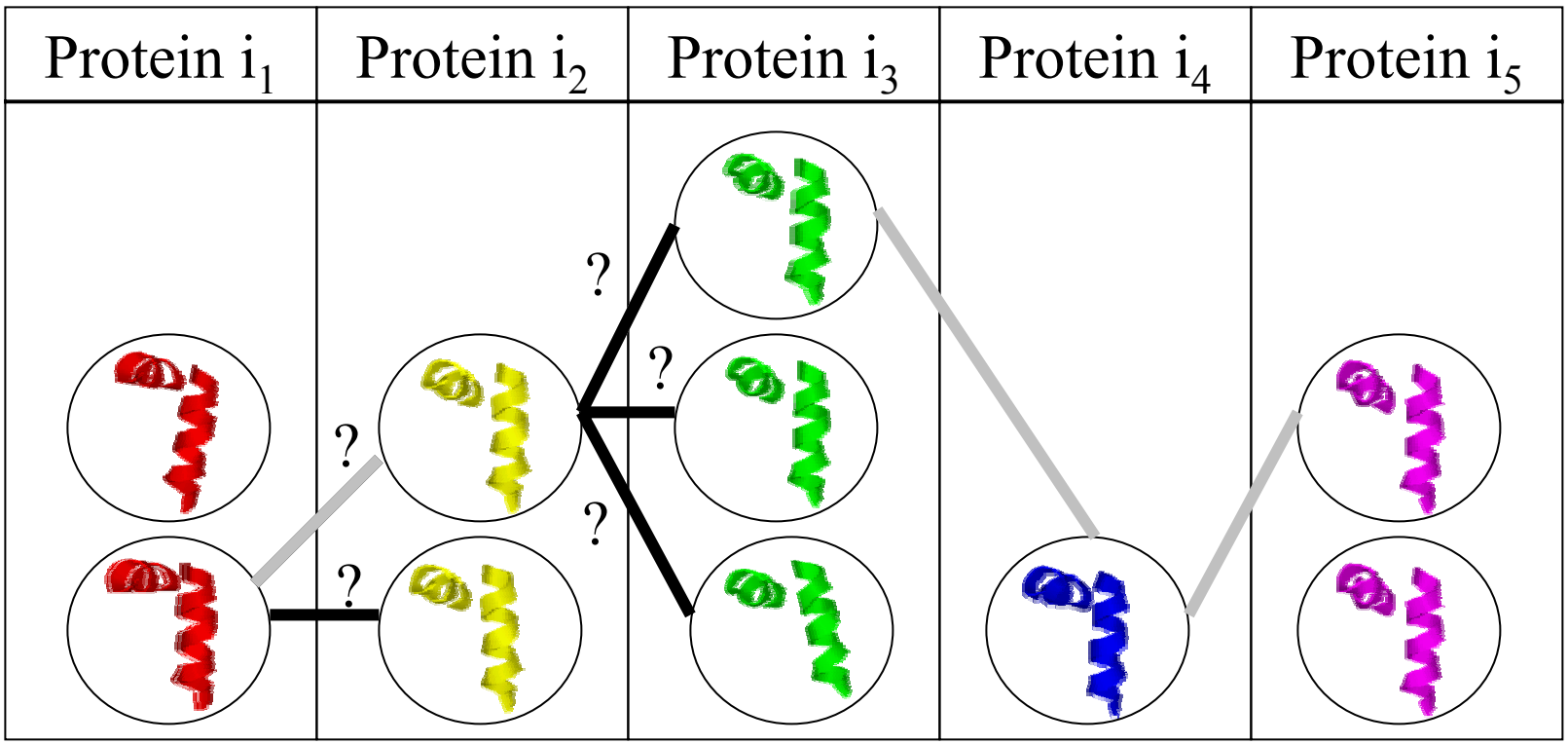
  - Providing the alignments with the largest cores for each possible number of aligned molecules.

# molecules	Alignments with largest core
2	
3	
4	

- **Evaluation**



Polynomial-time heuristic



Complexity

- **Theoretical Worst Case Complexity**

Construction of local pairwise alignments: $O(m^2 s^4) \cdot O(1)$

Clustering: $O(m^2) \cdot O(s^8 \log s)$

Extension: $O(m^2 s^4) \cdot O(n)$

Construction of multiple alignments: $O(ms^2) \cdot O(ms^2 n)$

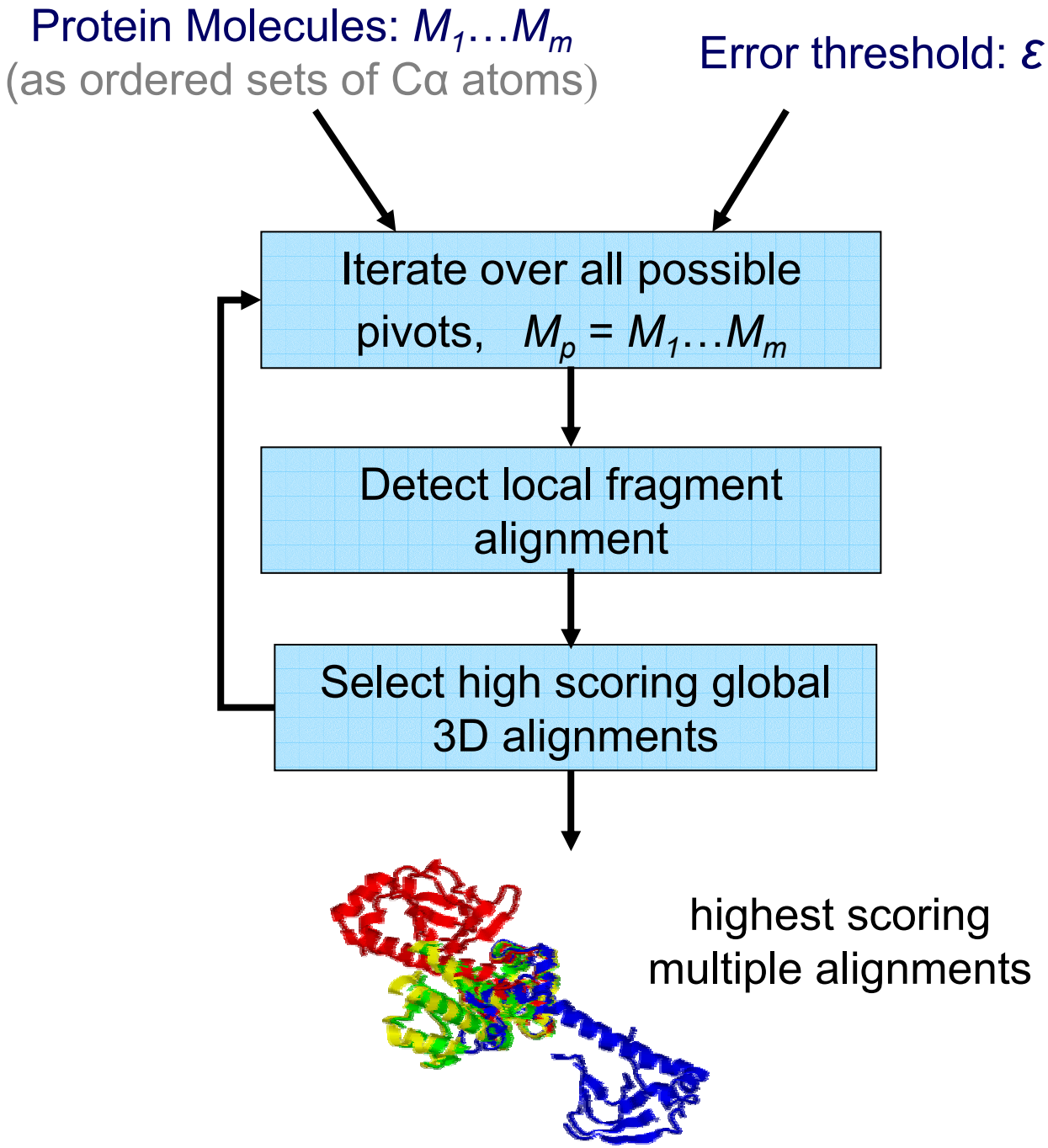
$$O(m^2 s^4 (s^4 \log s + n))$$

m – Number of input molecules
s – Maximum number of SSEs in a protein
n – Maximum number of residues in a protein

- **Practical Complexity**
  - The runtime is influenced by the number of bases in a basis bucket
  - The number of bases in a basis bucket depends on two factors:
    - the number of recurring motifs in a protein
    - the structural variance among the input proteins

MultiProt

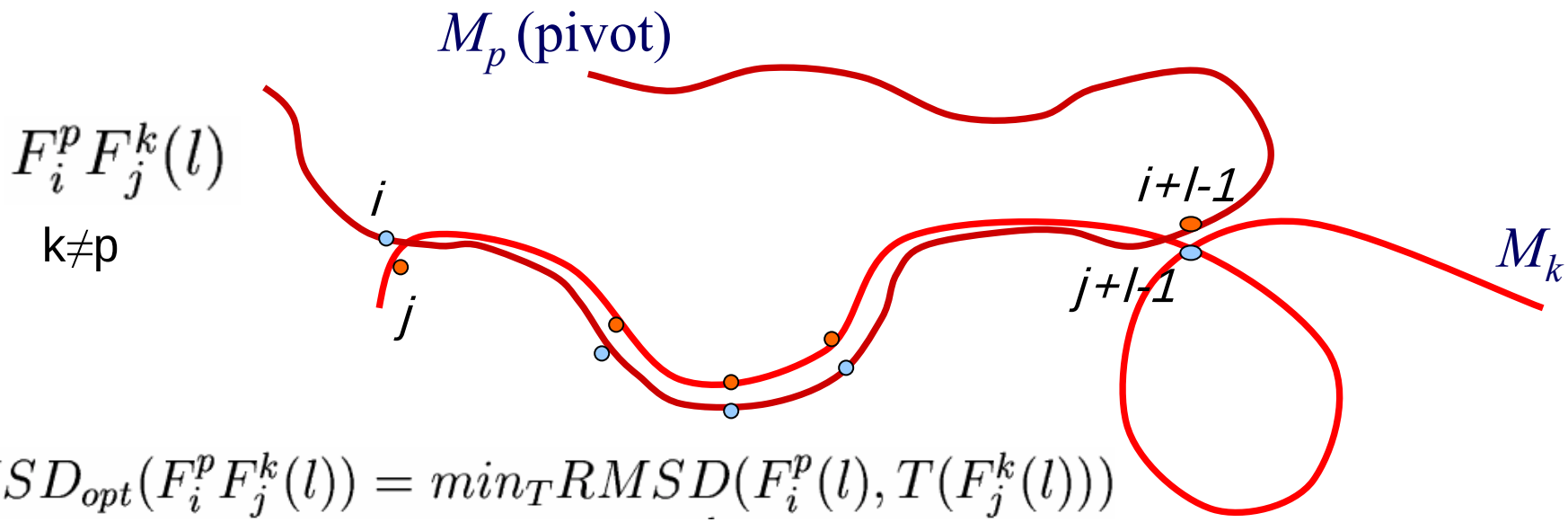
- Considers all structures **simultaneously**
- Capable of detecting **subset alignments**
- **Assumption**:
  A multiple alignment of proteins consists of at least short contiguous fragments of input points (>= 3 points).


Protein Molecules: $M_1 \ldots M_m$
(as ordered sets of Cα atoms)
Error threshold: ε
Iterate over all possible pivots, $M_p = M_1 \ldots M_m$
Detect local fragment alignment
Select high scoring global 3D alignments
highest scoring multiple alignments

Local Fragment Alignment

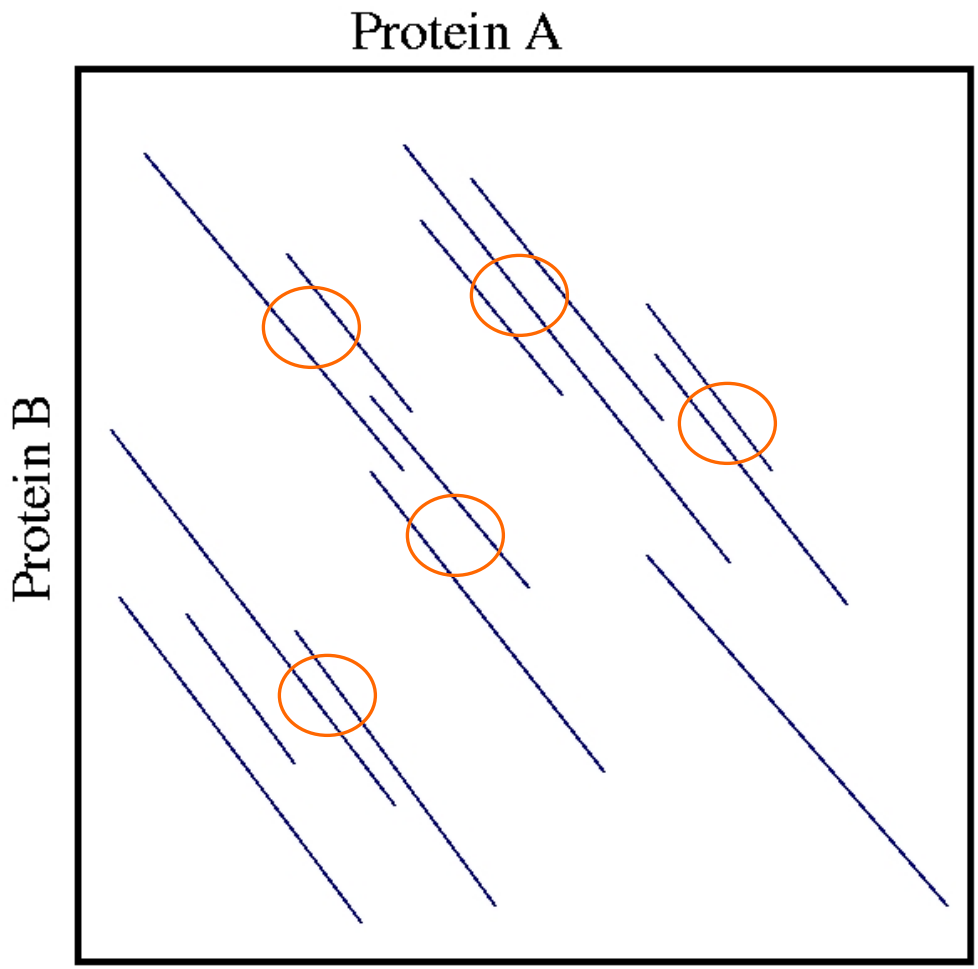
- Detects all ε-congruent fragments pairs between the pivot and the other molecules (similar to *FlexProt*)

$$S^{FP} = \{F_i^p F_j^k(l) : k \neq p,\ RMSD_{opt}(F_i^p F_j^k(l)) \leq \epsilon\}$$



$$RMSD_{opt}(F_i^p F_j^k(l)) = min_T RMSD(F_i^p(l), T(F_j^k(l)))$$

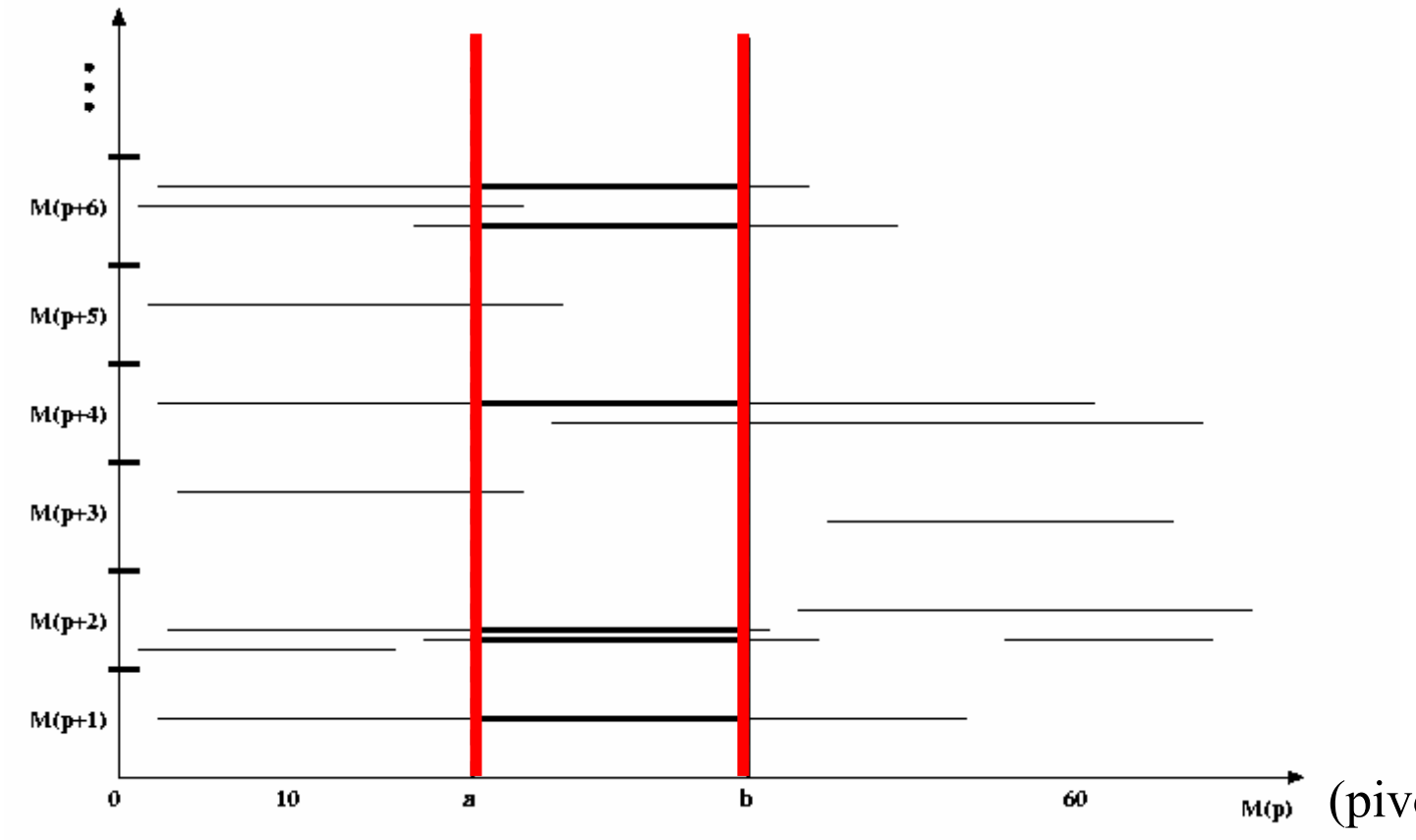
Fragment Clustering



30-40% reduction

- 2D plot of all the ε-congruent fragments pairs:

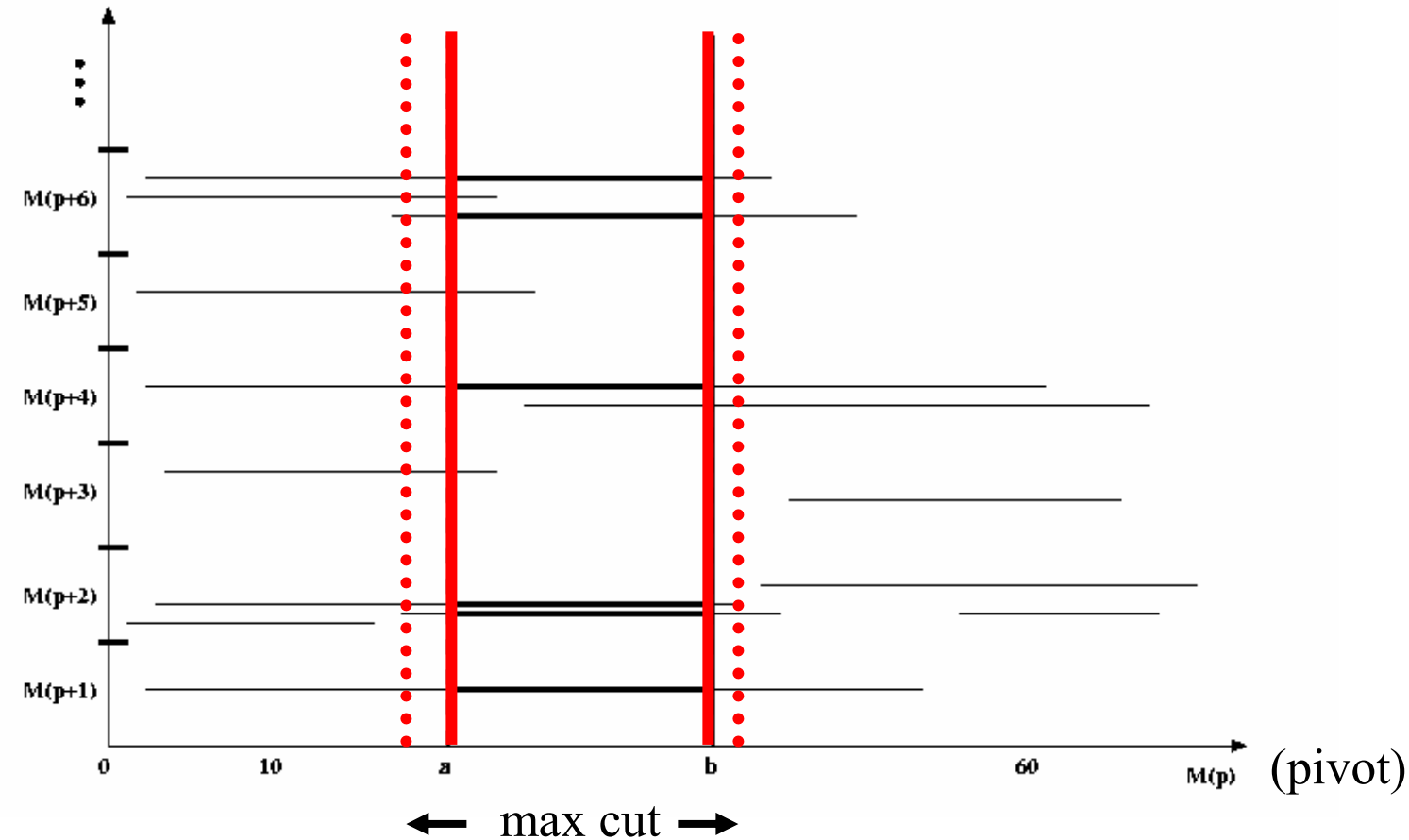
Cut[α,β] : $\{F_i^p F_j^k(l) : i \leq \alpha, (i+l-1) \geq \beta\}$



Number of multiple alignment for ***Cut[α,β]*** = $\prod_i k_{M_i}$ (2•1•2•1)

(exponential number in worst case)

- Detect all max cuts (using the Sweeping method)
- Worst case: $O(n^2)$ max cuts where n is the number of the pivot's points.



Computing the Best Global Multiple Alignments

- Given a ***Cut[α,β]*** : $\{F_i^p F_j^k(l) : i \leq \alpha, (i+l-1) \geq \beta\}$
  Select the high scoring fragment alignments
- Exact Solution – hard
- Heuristic: For each molecule, M_k , select a transformation that has the largest pairwise alignment with the pivot molecule M_p.
- Rank the alignments with the largest cores for each possible number of aligned molecules.

Iterative Extension

Given:

set of molecules $(M_{i1},\ldots,M_{ik})$ and transformations $(T_{i1},\ldots,T_{ik})$

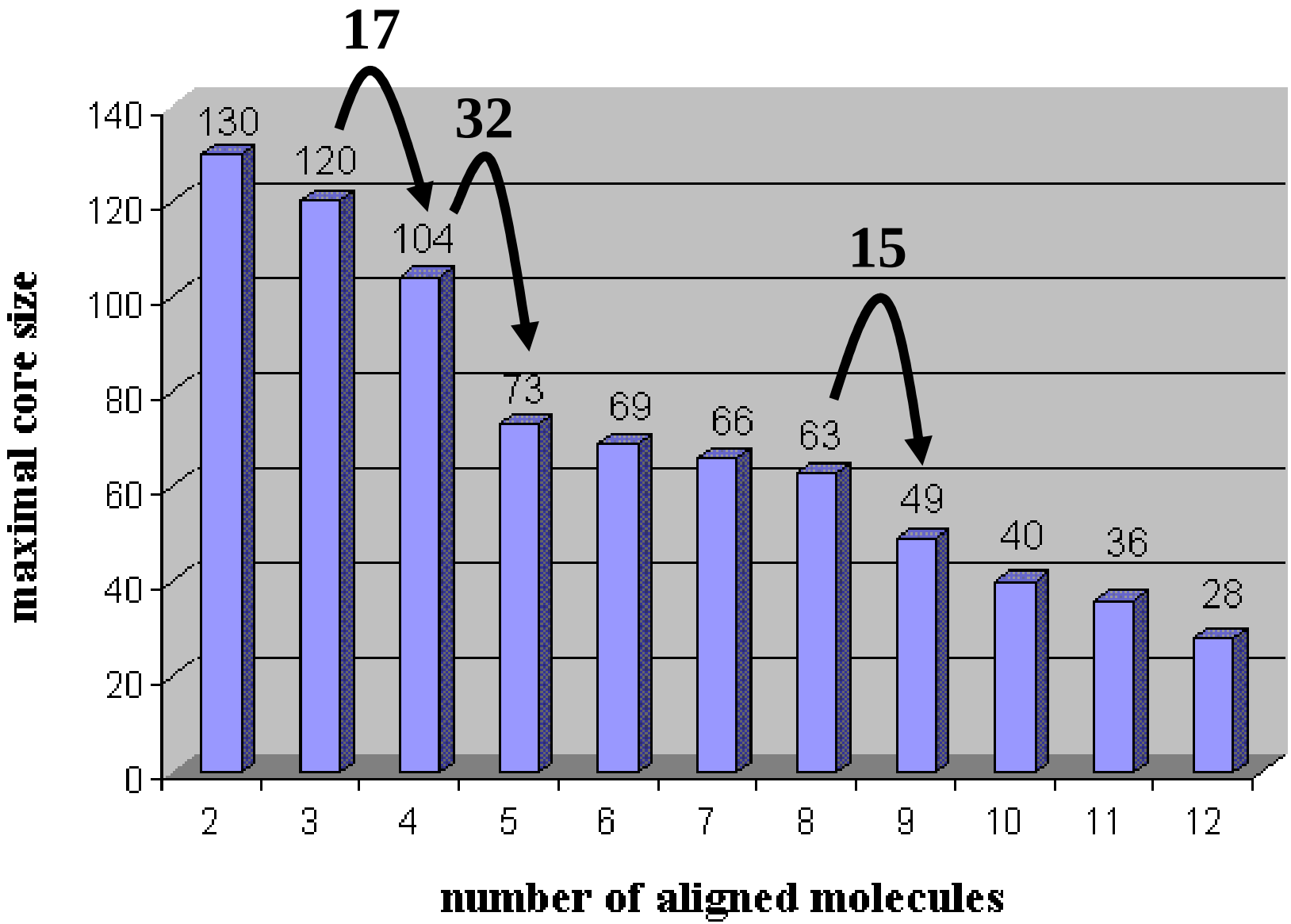
(1) Apply transformations : $M_{ij} = T_{ij}(M_{ij})$

(2) Compute optimal alignment with the pivot molecule. Recompute $(T_{i1},\ldots,T_{ik})$.

(3) Go to (1)

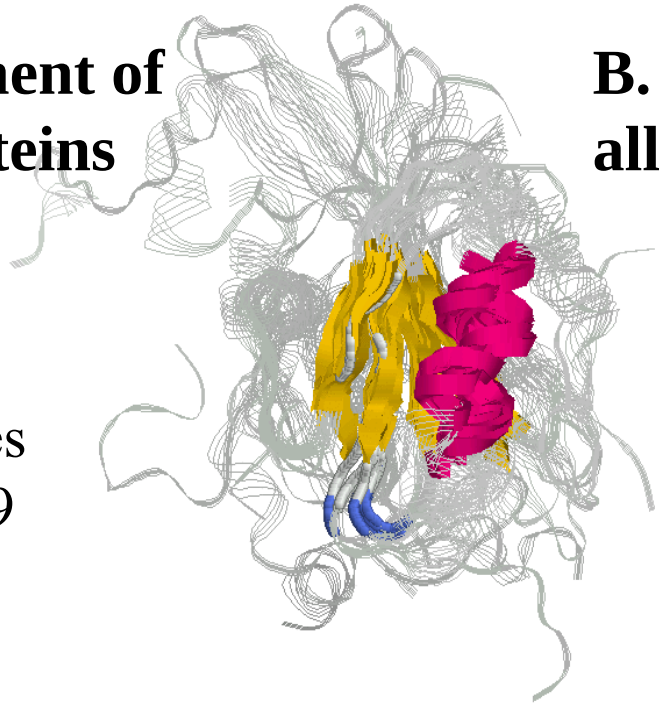
Ranking Example:
Cofilin-like and Gelsolin-like Families

- The ensemble contains 12 sequentially non-redundant structures taken from the two families of the *Actin depolymerizing proteins* fold:
  - Cofilin-like (CL) family (4 molecules)
  - Gelsolin-like (GL) family (8 molecules)


17
32
15
maximal core size
number of aligned molecules
0
20
40
60
80
100
120
140
130
120
104
73
69
66
63
49
40
36
28
2
3
4
5
6
7
8
9
10
11
12

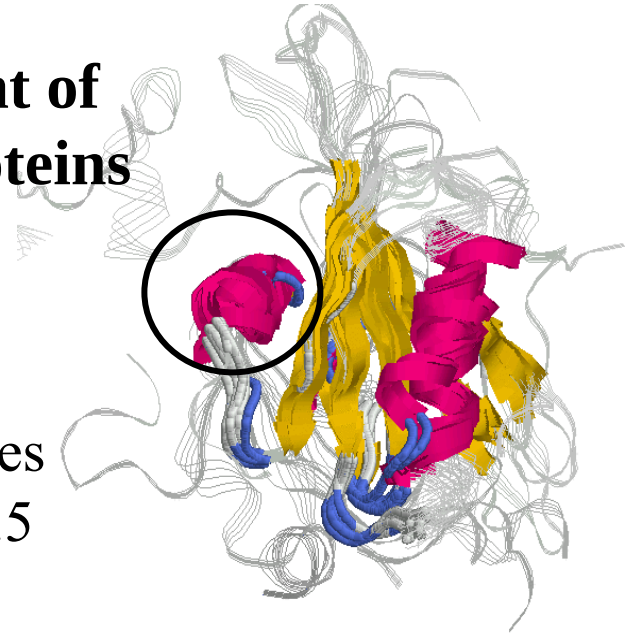
A. Alignment of all 12 proteins

28 residues
RMSD 1.9



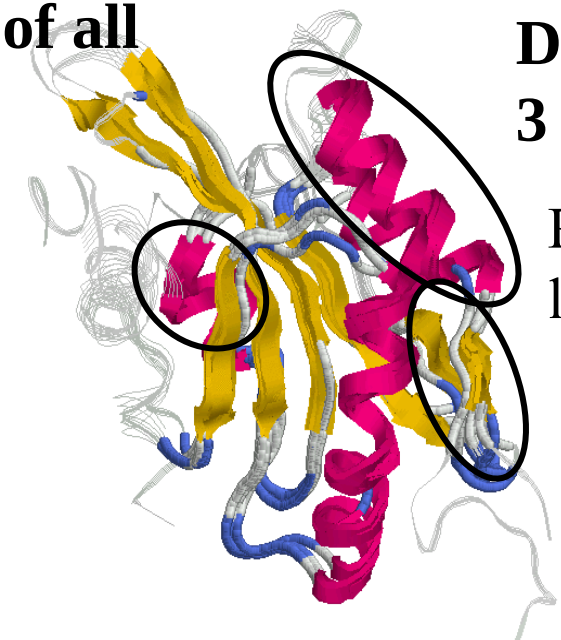
B. Alignment of all 8 GL proteins

63 residues
RMSD 1.5



C. Alignment of all 4 CL proteins

104 residues
RMSD 1.2



D. Alignment of 3 CL proteins

PDB:1f7s → lacks this helix

120 residues
RMSD 1.3