Protein Structure Alignment

Laboratory of Bioinformatics I
Module 2

March 14 and 16, 2017

Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Department of Biological, Geological, and Environmental Sciences (BiGeA)
University of Bologna

Structure Superimposition

Given two sets of points with some the dimension $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots b_n)$ in Cartesian space, find the optimal rigid body transformation G between the two subsets A and B that minimizes a given distance metric D over all possible rigid body transformation G, i.e.

$$\boldsymbol{Y = G(X) = A * X + B}$$

A = 3x3 rotation matrix

B = the translation vector

X = original point

$$\mathbf{RMSD} = \sqrt{\frac{\sum_{i=1}^{n} (\mathbf{a}_i - \mathbf{b}_i)^2}{\mathbf{n}}}$$

$$\mathbf{A} = \begin{bmatrix} \cos\theta\cos\psi & \cos\phi\sin\psi + \sin\phi\sin\theta\cos\psi & \sin\phi\sin\psi - \cos\phi\sin\theta\cos\psi \\ -\cos\theta\sin\psi & \cos\phi\cos\psi - \sin\phi\sin\theta\sin\psi & \sin\phi\cos\psi + \cos\phi\sin\theta\sin\psi \\ \sin\theta & -\sin\phi\cos\theta & \cos\phi\cos\theta \end{bmatrix}$$

Therefore structural superimposition correspond the best rototraslation which computational complexity is O(n).

Structural Alignment

Given two sets of points $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots b_m)$ in Cartesian space, find the optimal subsets A(P) and B(Q) with |A(P)| = |B(Q)|, and 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 G, i.e.

$$\min_{G}\{D[A(P) - G(B(Q))]\}$$

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

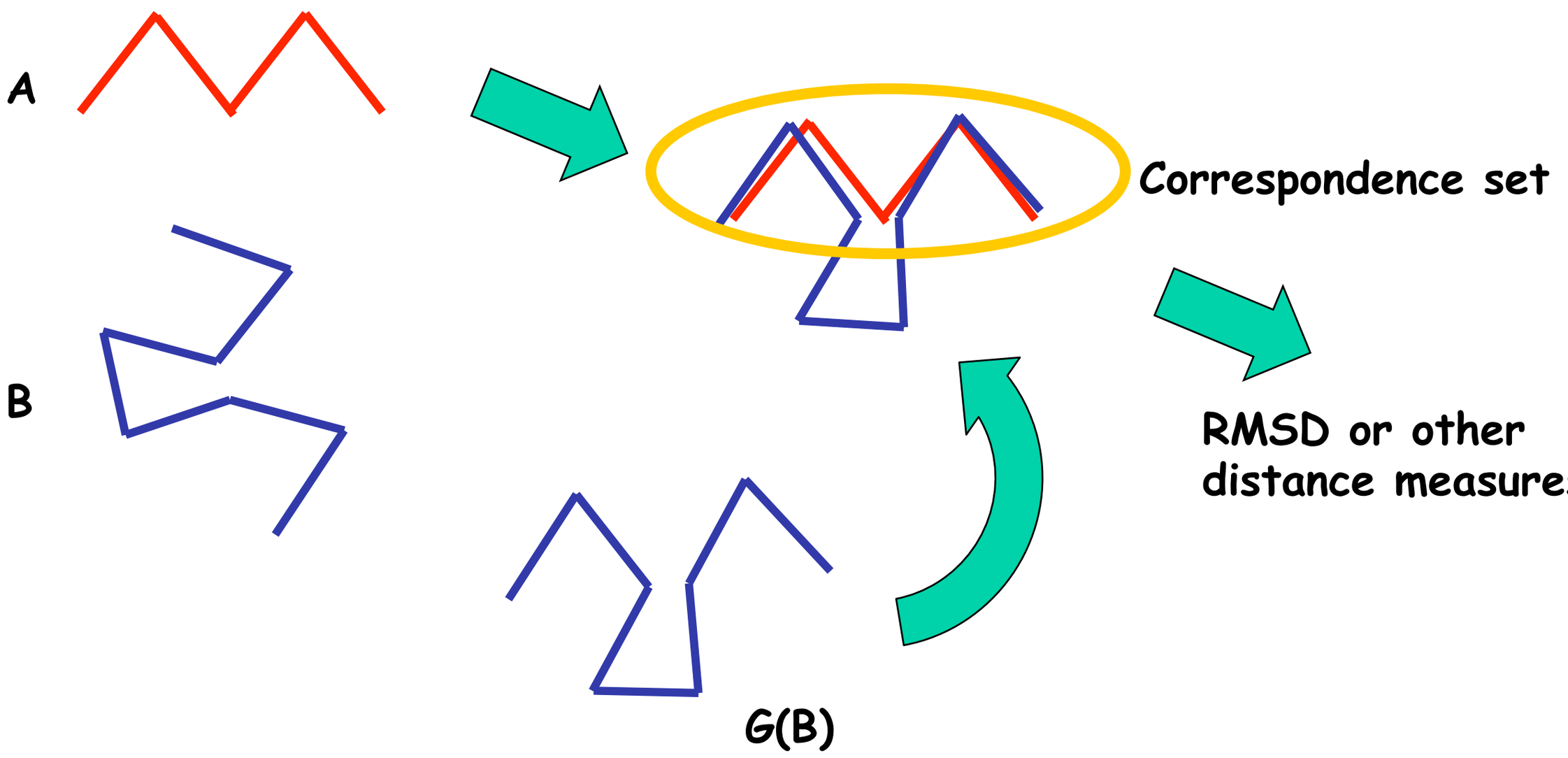
The two subsets A(P) and B(Q) define a “correspondence”, and p = |A(P)| = |B(Q)| is called the correspondence length. Naturally, the correspondence length is maximal when A(P) and B(Q) are similar.

Therefore there are essentially two problems in structure alignment:

- Find the correspondence set (which is NP-hard), and
- Find the alignment transform (which is O(n)).

Bourne P. 2012

Structural Alignment



Correspondence: (A_1,B_1), (A_2,B_2), (A_3,B_6), (A_4,B_7), (A_5,B_8)

Superimposition vs Alignment

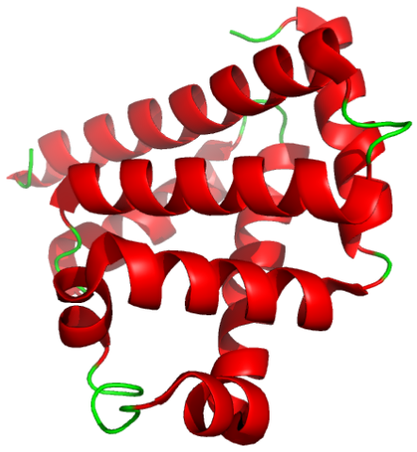
- Structure superposition assumes you already know which atoms to superimpose (correspondence set)

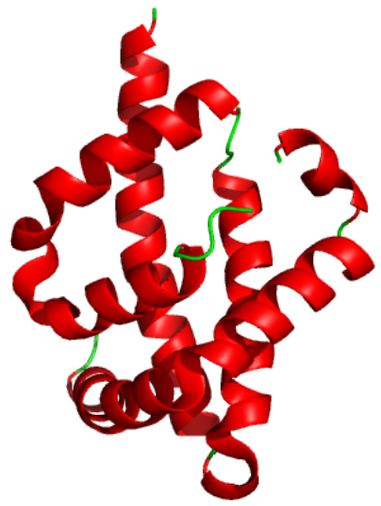
  it merely optimizes the position of the chosen atoms (relatively simple)

- Structure alignment must first determine what atoms to align (difficult).

Structures Comparison

Sperm Whale Myoglobin (1JP6:A)





Bacterial Haemoglobin (1VHB:A)

Feature Extraction

Structure 1

Structure 2

Algorithm

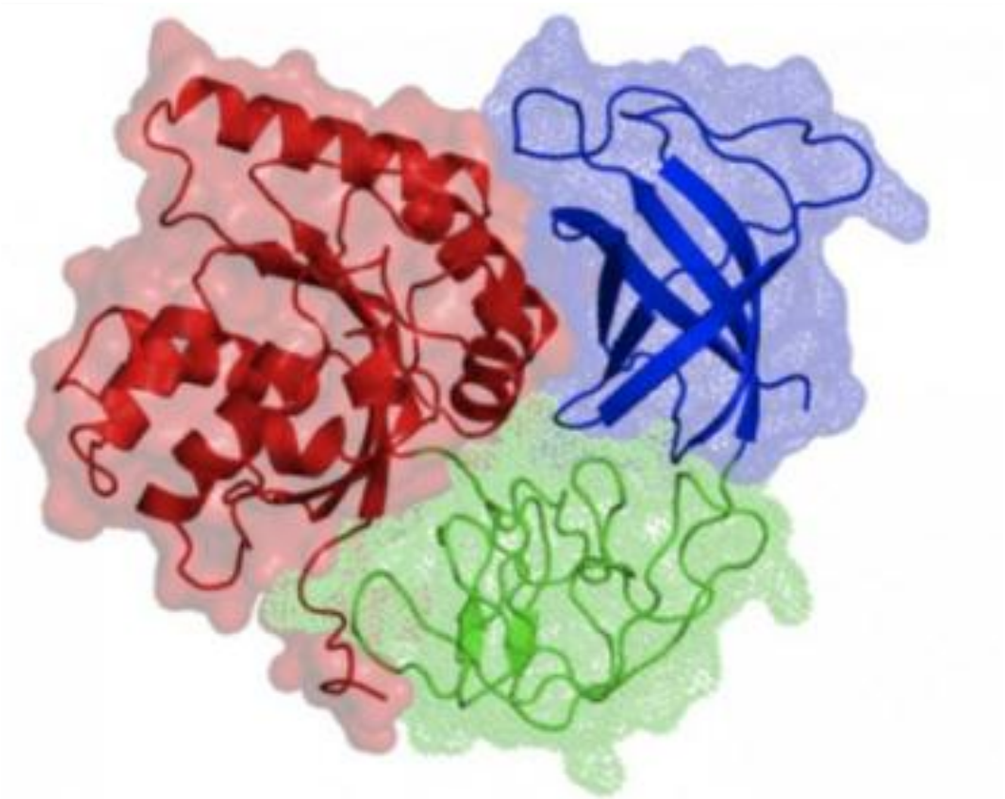
Comparison Algorithm

Statistical Significance

Score

Bourne P. 2012

Level of Comparison

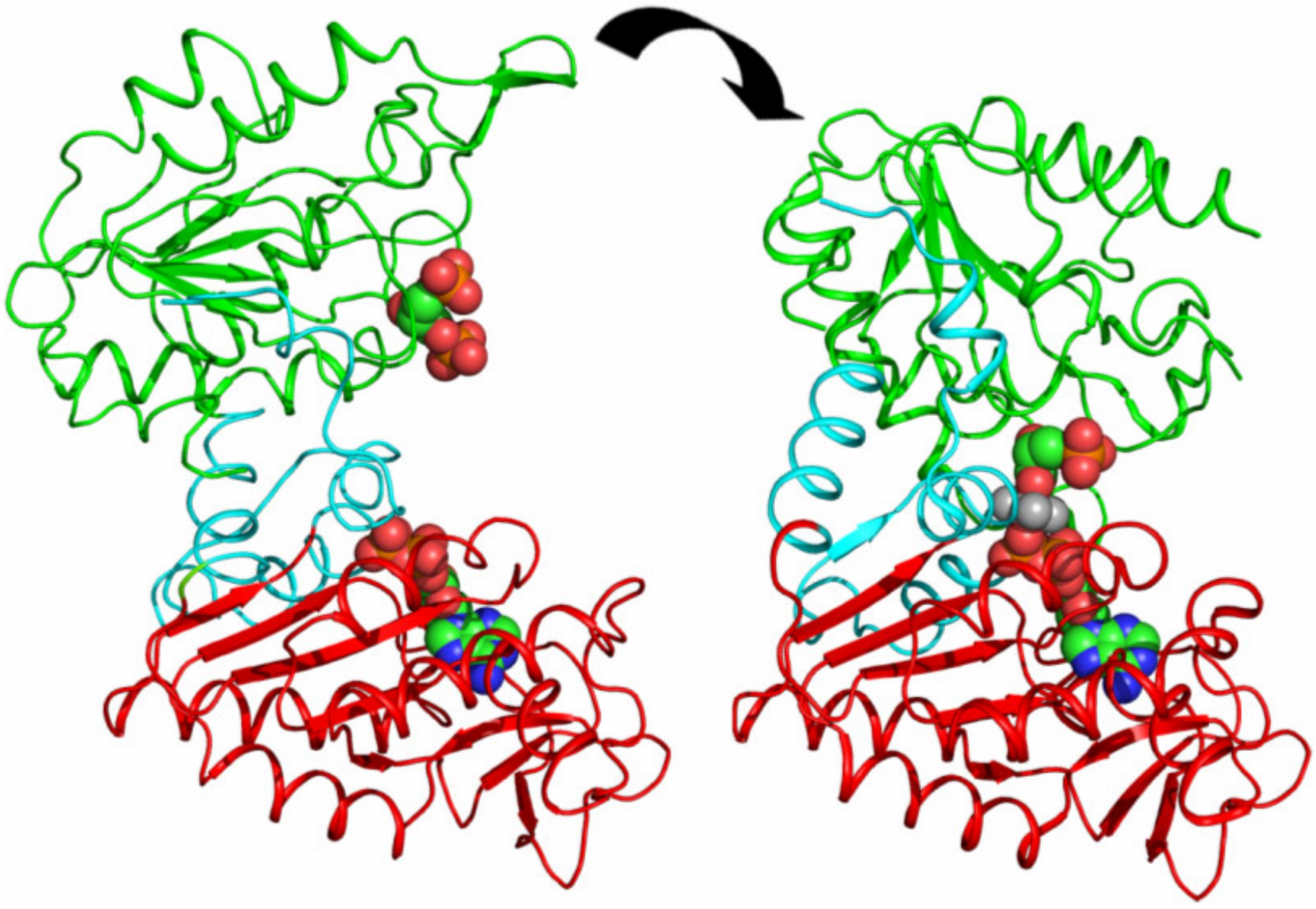


Three domains of Thermus aquaticus elongation factor EF-Tu:
in blue (all-β), red (α/β) and green (all-β).

Structural domains (the units of fold) are independently stable tertiary structures of proteins. They are distinct functional and/or structural units and can evolve, exist and function independently. Therefore, the same domain can be a part of different protein (EBI on-line course)

The definition of domain is often heuristic and questionable (the independent evolution/existence and functionality is rarely experimentally tested

Multi Domain Alignment



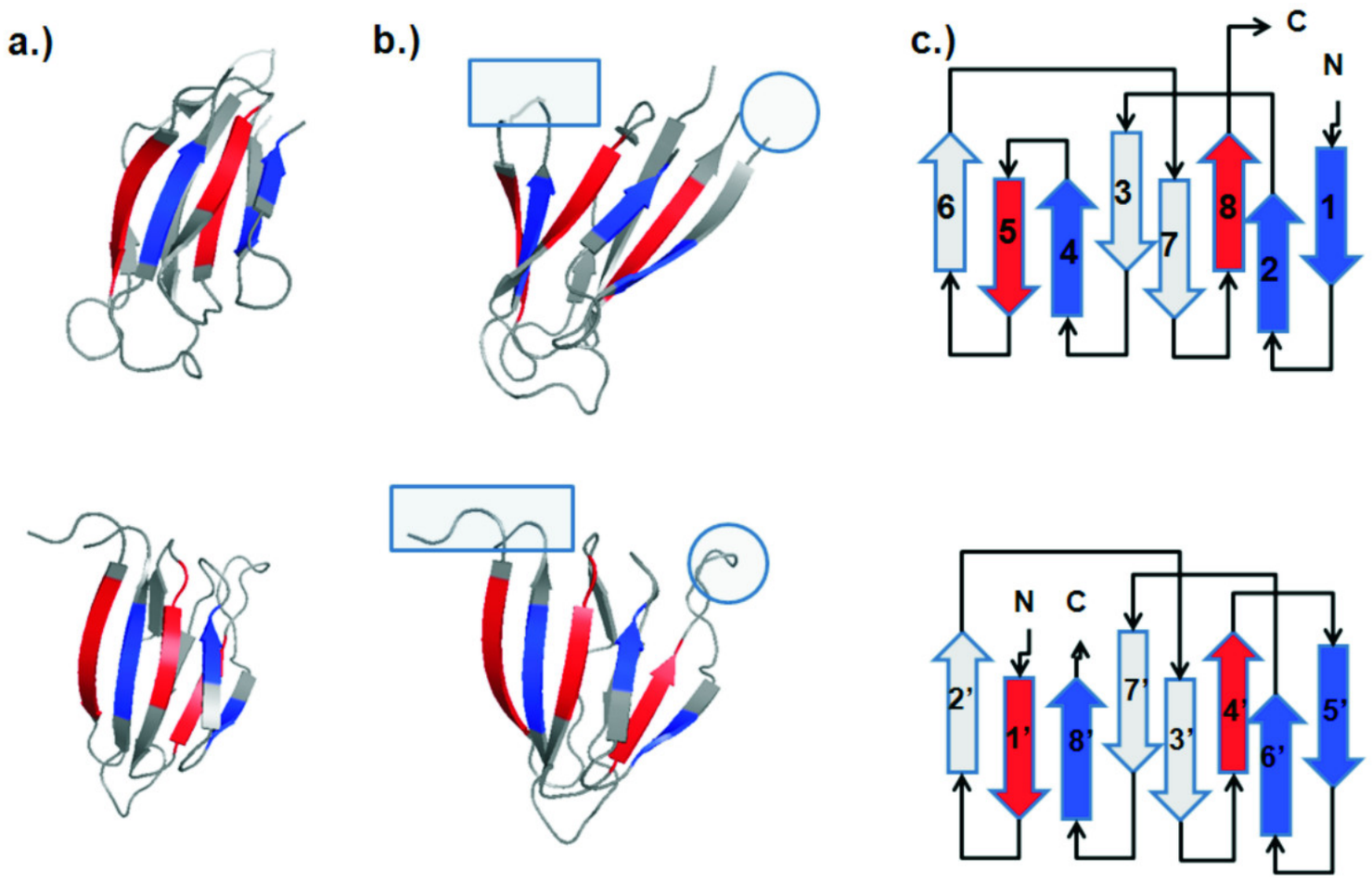
Domain movements in PGK catalysis. The fully-open resting state of the enzyme defined by refinement against SAXS data (left) binds the substrates 13BPG in the N domain (green) and ADP in the C-domain (red).

A rotation of ~56° of the hinge region (blue) brings the substrates together to initialise catalysis and ATP production (right)

Image credit: M.W. Bowler

Topology Independent Alignments

Most protein structural alignment methods can reliably classify proteins into similar folds given the structural units from each protein are in the same sequential order. However, the evolutionary possibility of proteins with different structural topology but with similar spatial arrangement of their secondary structures pose a problem.



Nucleoplasmin-core (1k5j, chain E, top panel), and the fragment of residues 37–127 of auxin binding protein 1 (1lrh, chain A, bottom panel). a) These two proteins superimpose well spatially, with an RMSD value of 1.36Å for an alignment length of 68 residues

Dundas et al. BMC Bioinformatics. 2007

Structural Alignment Tools

There are several well-documented, easy to use software packages for structural alignment. More than 100 are reported on wikipedia.

NAME	Description	Class	Type	Flexible	Link	Author	Year
MAMMOTH	**MA**tching **M**olecular **M**odels **O**btained from **T**heory	Cα	Pair	No	server download	CEM Strauss & AR Ortiz	2002
CE	**C**ombinatorial **E**xtension	Cα	Pair	No	server	I. Shindyalov	2000
CE-MC	**C**ombinatorial **E**xtension-**M**onte **C**arlo	Cα	Multi	No	server	C. Guda	2004
DaliLite	**D**istance Matrix **Ali**gnment	C-Map	Pair	No	server	L. Holm	1993
TM-align	**TM**-score based protein structure **align**ment	Cα	Pair	nil	server and download	Y. Zhang & J. Skolnick	2005
VAST	**V**ector **A**lignment **S**earch **T**ool	SSE	Pair	nil	server	S. Bryant	1996
PrISM	**P**rotein **I**nformatics **S**ystems for **M**odeling	SSE	Multi	nil	server	B. Honig	2000
SSAP	**S**equential **S**tructure **A**lignment **P**rogram	SSE	Multi	No	server	C. Orengo & W. Taylor	1989
SARF2	**S**patial **AR**rangements of Backbone **F**ragments	SSE	Pair	nil	server	N. Alexandrov	1996
KENOBI/K2	NA	SSE	Pair	nil	server	Z. Weng	2000
STAMP	**ST**ructural **A**lignment of **M**ultiple **P**roteins	Cα	Multi	No	site server	R. Russell & G. Barton	1992

https://en.wikipedia.org/wiki/Structural_alignment_software

Method Classification

Type

Pair Pairwise Alignment (2 structures only);
Multi Multiple Structure Alignment;

Class

Cα Backbone Atom (Cα) Alignment;
AllA All Atoms Alignment;
SSE Secondary Structure Elements Alignment;
Seq Sequence-based alignment
C-Map Contact Map
Surf Connolly Molecular Surface Alignment
SASA Solvent Accessible Surface Area
Dihed Dihedral Backbone Angles
PB Protein Blocks

Protein descriptors

Flexible

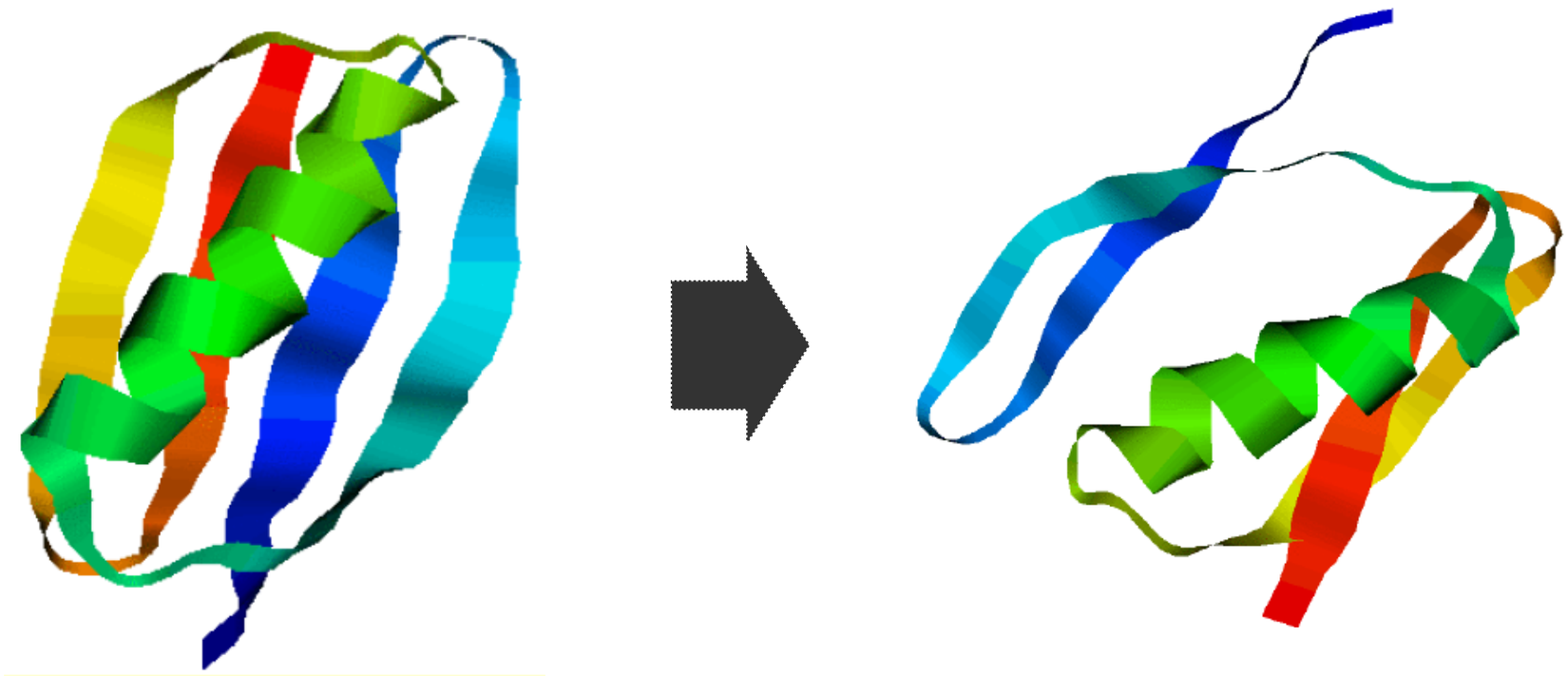
No Only rigid-body transformations are considered between the structures being compared.
Yes The method allows for some flexibility within the structures being compared, such as movements around hinge regions

Comparing Torsion Angles

Torsion Angles (Φ,Ψ) are:

- local by nature
- invariant upon rotation and translation of the molecule
- compact - complexity o(n)

Good for alignment of local region but
possible problems on the alignment of the whole structure



Credit: Predrag Radivojac

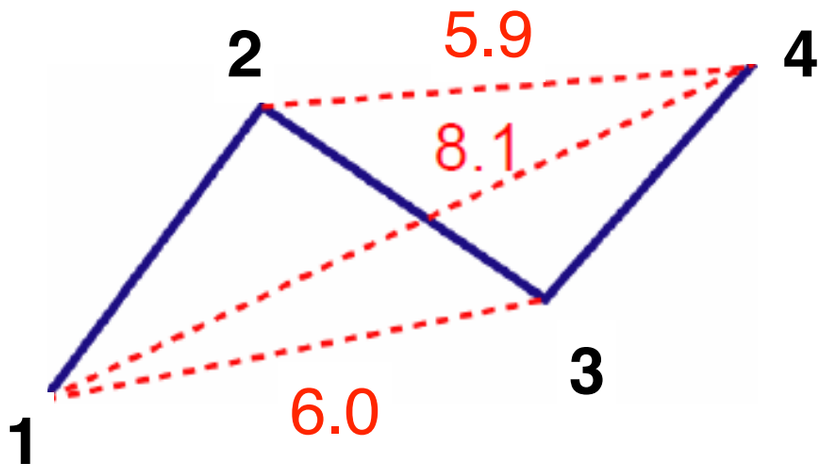
Distance Matrix

Advanctage:

- invariant upon rotation and translation of the molecule
- can be used for protein comparison

Disadvantages

- Comparing matrices is an hard computational problem
- Complexity is o(n^2) where n represents the number of residues
- Insensitive to chirality



	1	2	3	4
1	0.0	3.8	6.0	8.1
2	3.8	0.0	3.8	5.9
3	6.0	3.8	0.0	3.8
4	8.1	5.9	3.8	0.0

Credit: Predrag Radivojac

Structural Alignment Components

Input & output of alignment algorithm

Input: two proteins: $A = \{a_1, \cdots, a_m\}$ $B = \{b_1, \cdots, b_n\}$

Output: An alignment and scores

$$L(A,B) = \{(a_{i_1}, b_{j_1}), \cdots, (a_{i_L}, b_{j_L})\},$$
$$i_1 < i_2 < \cdots < i_L, j_1 < j_2 < \cdots < j_L$$

Constraints:

min rmsd:

$$rmsd = \min_T \sqrt{\frac{\sum_{k=1}^{L} (a_{i_k} - Tb_{j_k})^2}{L}}$$

max L

min Gaps

$$Gaps = \sum_{t=1}^{L-1} \left[(i_{t+1} - i_t - 1) + (j_{t+1} - j_t - 1) \right]$$

Dynamic programming, Integer programming, Monte Carlo…

Statistical Significance

Phil Bourne 2012

State Of The Art

- All methods can identify obvious similarities between two structures
- Remote similarities are detected by a subset of methods – different remote similarities are recognized by different methods
- Good alignments are much harder to come by
- Speed is a serious issue with some algorithms

Desirable Method Features

- Biologically meaningful alignments not just geometrically meaningful

- Complete database of all alignments

- Ability to apply to structures not in the PDB

Phil Bourne 2012

CE Algorithm

- Compare octameric fragments – an aligned fragment pair (AFP) (local alignments)

- Stitch together AFPs

- Find the optimal path through the AFPs

- Optimize the alignment through dynamic programming

- Measure the statistical significance of the alignment

Shindyalov and P.E. Bourne Protein Engineering 1998, 11(9) 739-747

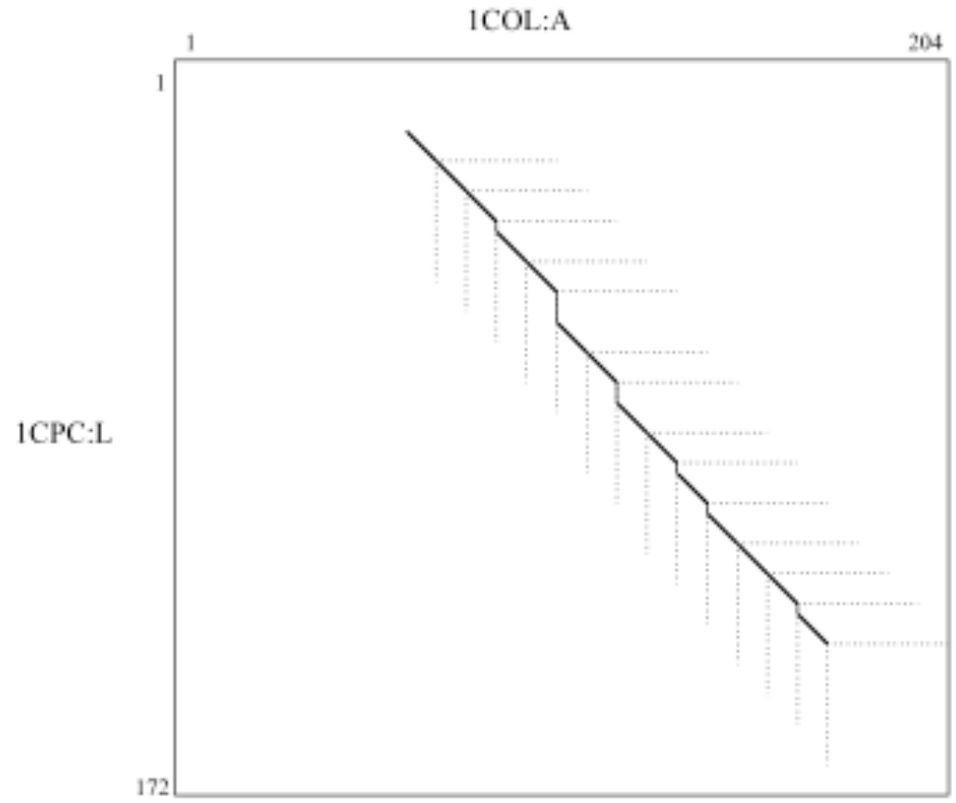
Constrain The Search

The alignment between two proteins A and B is the longest continuous path P of AFPs of size m in a similarity matrix

Similarity Matrix S represents all AFPs conforming to some similarity criterion (e.g., low RMSD):

$$S=(n_A-m).(n_B-m)$$

m = Length of AFP
n_A = Length of protein A



This is very large to compute – constraints are needed

Shindyalov and P.E. Bourne Protein Engineering 1998, 11(9) 739-747

Path Definition

p^A_i = AFPs starting residue position in protein A at the i-th position of the alignment path

m = longest continual path – set as 8

One of the conditions (1)-(3) should be satisfied for 2 consecutive AFPs i and i+1 in the path

(1) = 2 consecutive AFPs aligned without gaps
(2) = Two consecutive AFPs with a gap in protein A
(3) = Two consecutive AFPs with a gap in protein B

$$p^A_{i+1} = p^A_i + m \text{ and } p^B_{i+1} = p^B_i + m \quad (1)$$

or

$$p^A_{i+1} > p^A_i + m \text{ and } p^B_{i+1} = p^B_i + m \quad (2)$$

or

$$p^A_{i+1} = p^A_i + m \text{ and } p^B_{i+1} > p^B_i + m \quad (3)$$

Extension of the Path

Gap sizes are limited to G – heuristically set as 30 residues

$$p_{i+1}^{A} \leqslant p_i^{A} + m + G \qquad (4)$$

$$p_{i+1}^{B} \leqslant p_i^{B} + m + G \qquad (5)$$

Similarity Measures

1. RMSD from least squares superposition
used to select few best fragments

2. Full set of inter-residue distances
used for a scoring single AFP

$$D_{ij} = \frac{1}{m^2}\left(\sum_{k=0}^{m-1} \sum_{l=0}^{m-1} \left| d^{A}_{p_i^A+k, p_j^A+l} - d^{B}_{p_i^B+k, p_j^B+l} \right| \right) \qquad (7)$$

3. Distance calculated from independent set of inter-residue distances where each distance is used only once
used for combinations of 2 AFPs

$$D_{ij} = \frac{1}{m}\left(\left| d^{A}_{p_i^A, p_i^A} - d^{B}_{p_i^B, p_i^B} \right| + \left| d^{A}_{p_i^A+m-1, p_j^A+m-1} - d^{B}_{p_i^B+m-1, p_j^B+m-1} \right| + \sum_{k=1}^{m-2} \left| d^{A}_{p_i^A+k, p_j^A+m-l-k} - d^{B}_{p_i^B+k, p_j^B+m-l-k} \right| \right) \qquad (6)$$

Phil Bourne 2012

Statistical Evaluation

Evaluate the probability of finding an alignment path of the same length or smaller gaps and distance from a random set of non-redundant structures.

Optimization:

The 20 best alignments with a Z score above 3.5 are assessed based on RMSD and the best kept. This produces approx. one error in 1000 structures

Each gap in this alignment is assessed for relocation up to m/2

Iterative optimization using dynamic programming is performed using residues for the superimposed structures

Limitations

- Will not find non-topological alignments (outside the bounds of the dotted lines)
- What are the correct “units” to be comparing?
- CE initially worked on chains – as we shall see in future weeks domains are the correct units, but definition of the domains is not straightforward

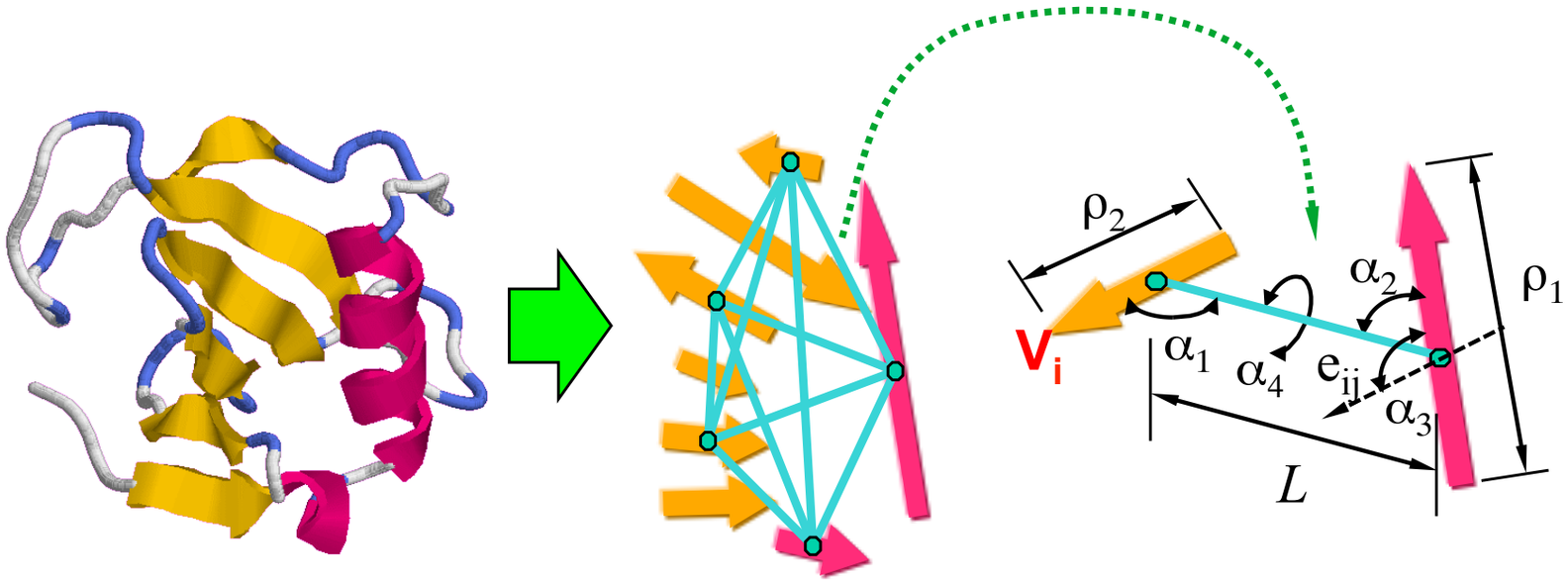
PDBe Fold

- Protein secondary structure elements (SSE) – natural and convenient objects for building three dimensional graphs.

- Secondary structures provide most functionality and is conserved through evolution

- Details of protein fold – expressed in terms of two SSE – helices and strands

Graph Representation (I)



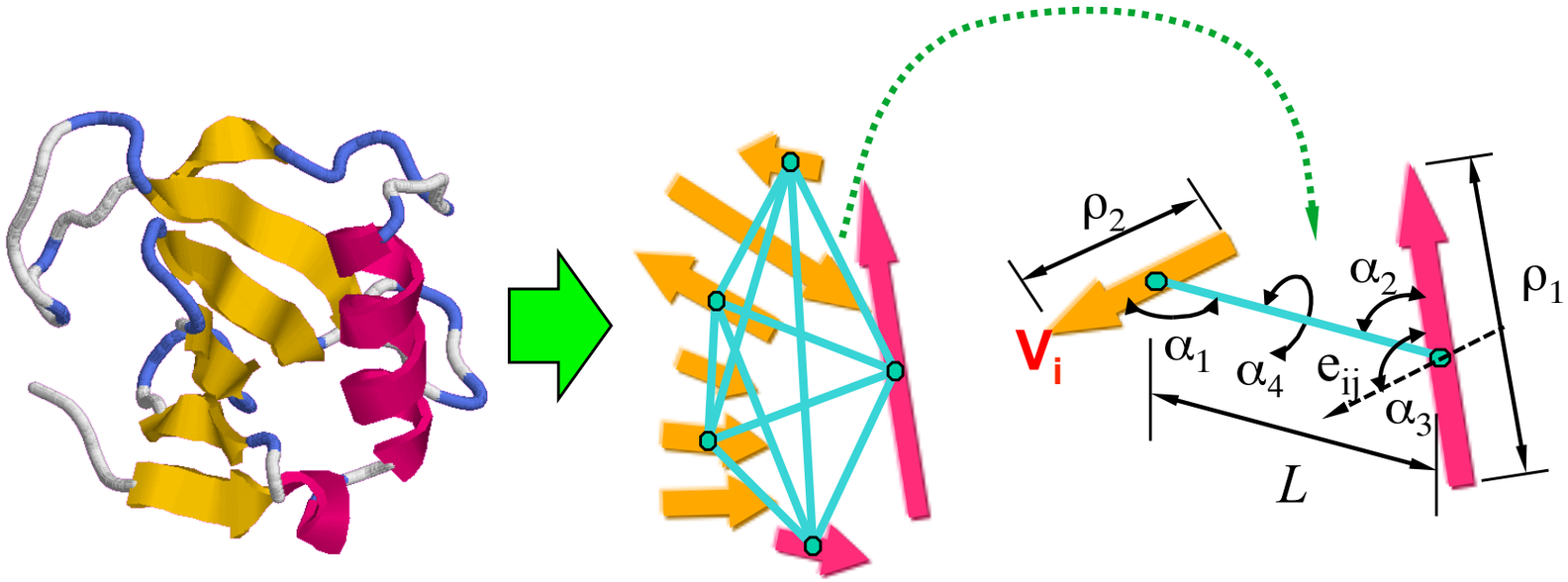
SSE graphs- represented by vectors

Each SSE can be used as graph vertices (T_i, ρ_i)

Any 2 vertices are connected by an edge label L – describes position and orientation of the connected SSEs

Each edge labelled with a property vector – $\alpha_{1/2}$ angle between edge and vertices, torsion angle between vertices, length of the edge L

Graph Representation (II)



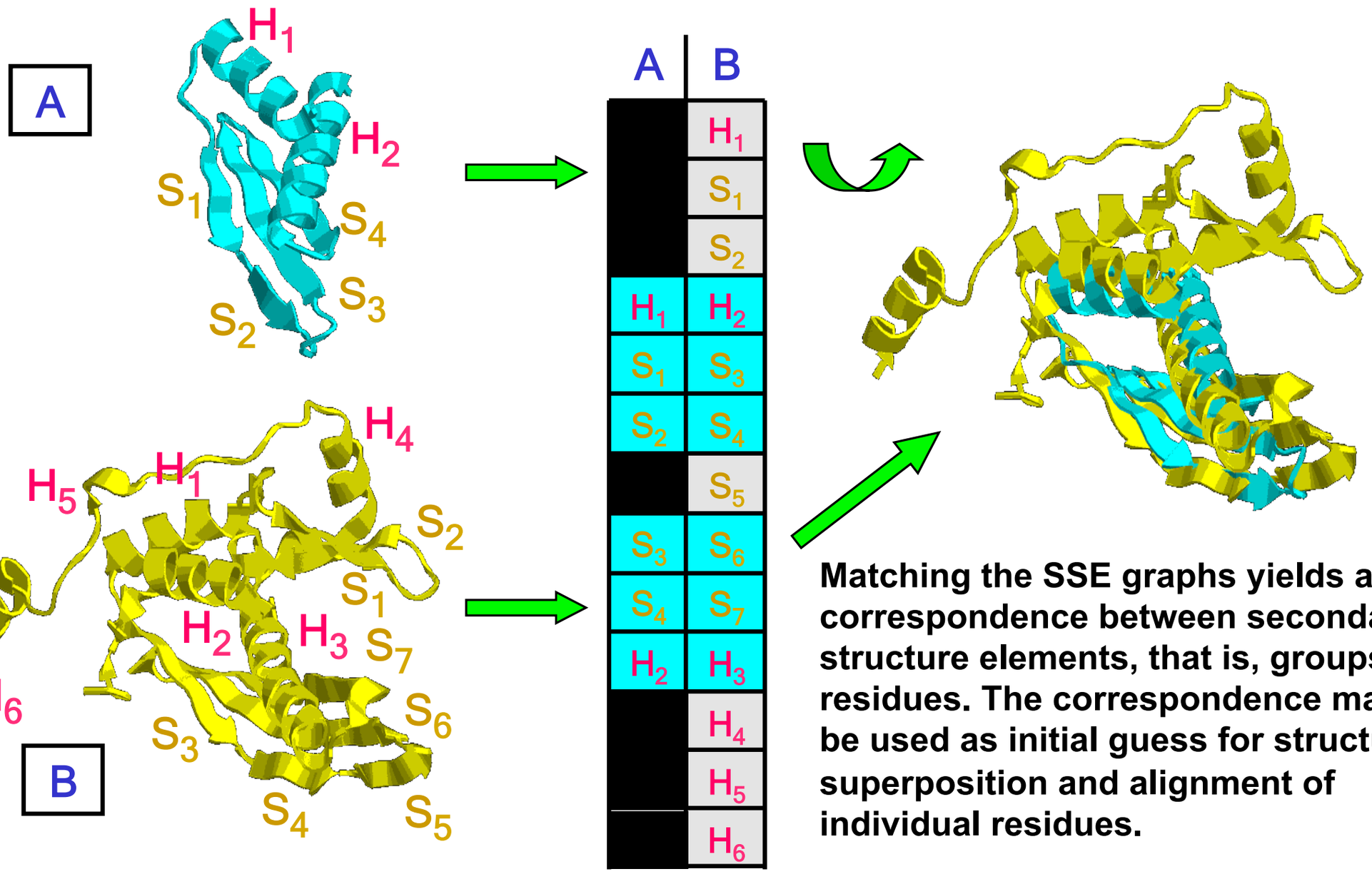
Sets of vertices, edges and their labels provides full definition of the graph.

Graph matching algorithm is required – set of rules for comparing individual vertices and edges – tolerances chosen empirically

Relative and absolute vertex and edge lengths are used for comparison – allows larger absolute differences for longer vertices and edges

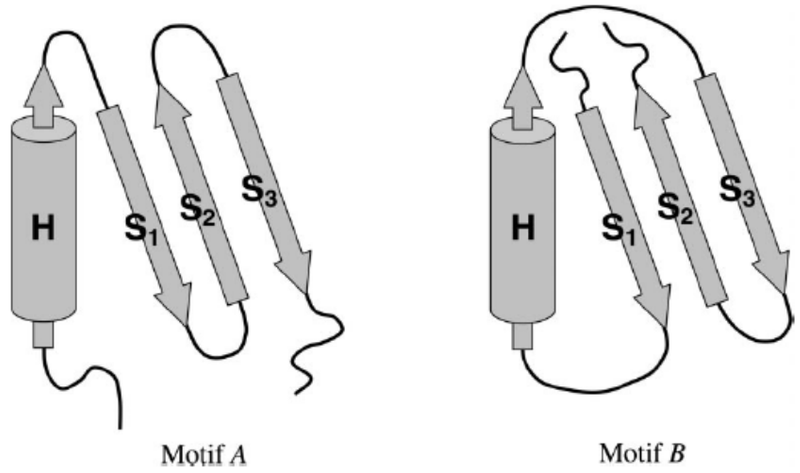
Torsion angle comparison – distinguish mirror symmetry mates

Graph Matching



PDBe Fold Approaches

1) Connectivity of SSE Neglected



2) Soft connectivity – general order of SSEs along their protein chains are same in both structures BUT any number of missing/unmatched SSE between matched ones allowed

3) Strict connectivity – matched SSEs follow same order along their protein chains – separated only by equal number of matched/ unmatched SSE in both structures

To obtain 3D alignment of individual residues – represent them by their C-alpha atoms – use results of graph matching as a starting point

MAMMOTH Algorithm

The MAMMOTH (MAtching Molecular Models Obtained from Theory) algorithm is one of the fastest methods for structural alignment .

The method represents a protein structure as a set of unit vectors build using the vectors between C-α atoms.

MAMMOTH uses a dynamic programming algorithm to find the bast alignment between two protein structure.

MAMMOTH-Mult

- MAMMOTH-mult is a multiple alignment version of MAMMOTH. It multiply aligns protein structures, providing a common 3D superimposition, a corresponding structure-based sequence alignment and a dendrogram for the set of structures aligned.
- Version: 1.0
- Free use for Educational and Research Purposes.
- Contact
- Reference: *Lupyan D, Leo-Macias A, Ortiz AR (2005) Bioinformatics (2005) 21, 3255-63*

Align your protein against one SCOP family.

Upload the **pdb file** containing the coordinates of your protein: Choose File No file chosen

Type the SCOP tag of the family you want to align your protein against (is five numbers code, e.g.: 50045)

Your **e-mail** for results to be sent back:

*some calculations may take upto few minutes, it is recommended that you include your email!

Align! Reset

Align your own proteins.

Upload your **MAMMOTH-mult** input file (See example): Choose File No file chosen

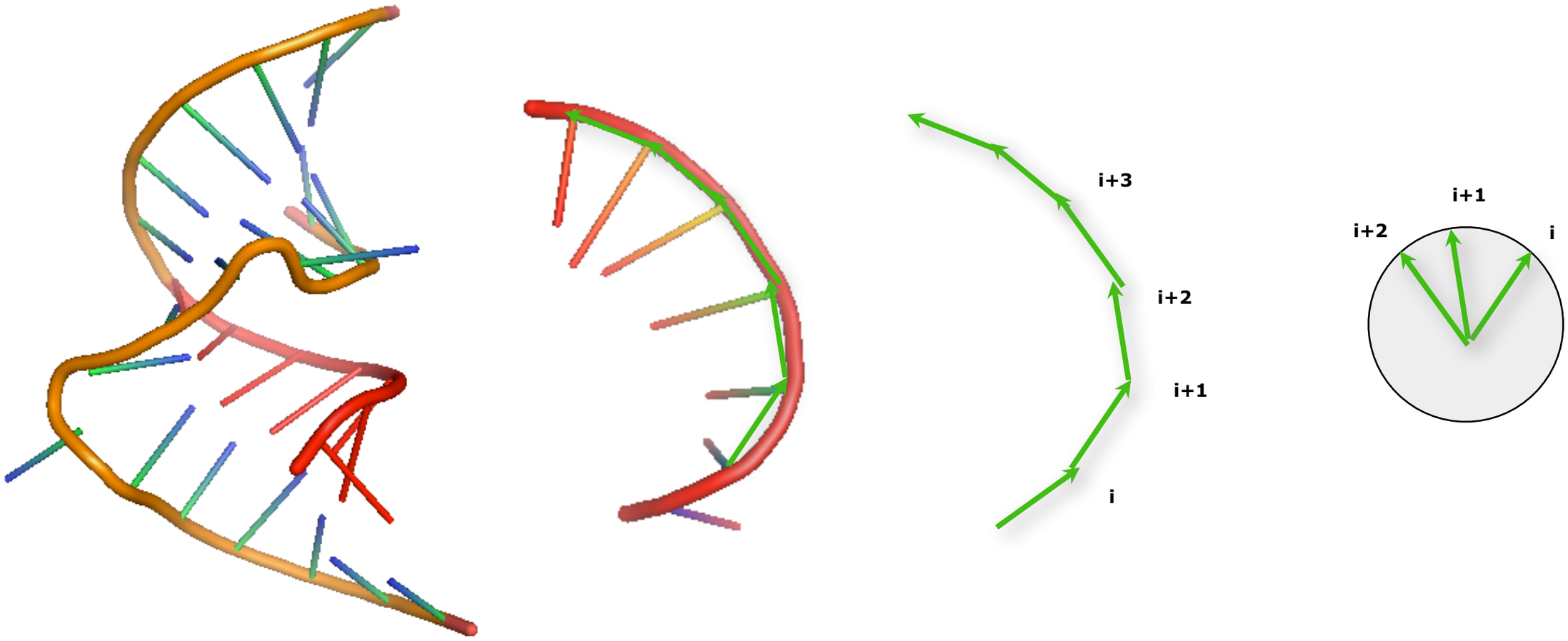
Your **e-mail** for results to be sent back:

*some calculations may take upto few minutes, it is recommended that you include your email!s

Align! Reset

https://ub.cbm.uam.es/software/online/mamothmult.php

Unit Vector Representation

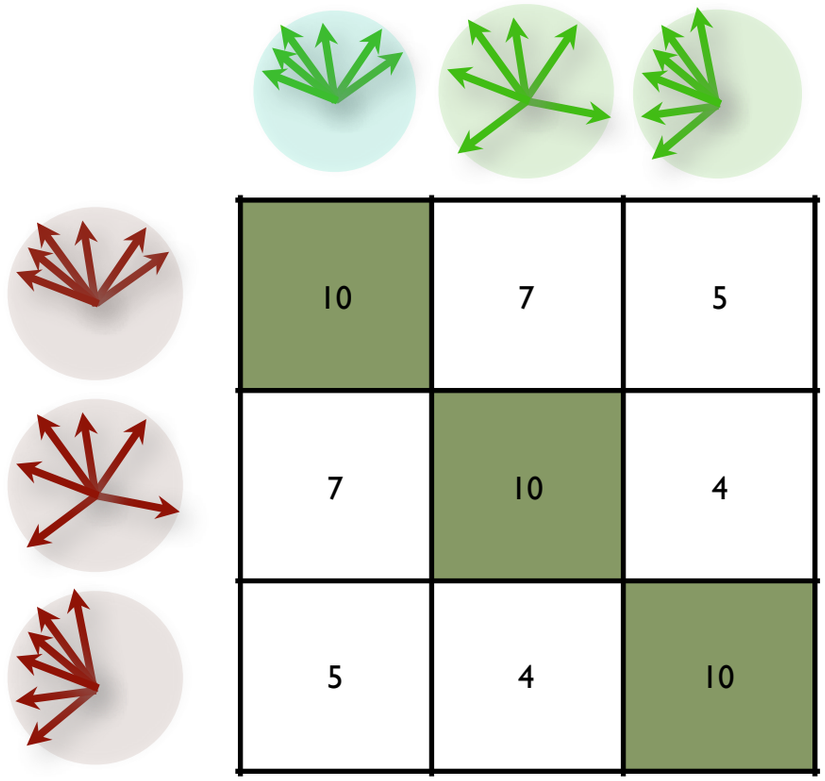


A Unit Vector is the normalized vector between two successive Cα atoms.

For each position *i* consider the *k* consecutive vectors, which will be mapped into a unit sphere representing the local structure of k residues.

Ortiz et al. *Proteins* 2002

Unit Vector Scoring



$$URMS^{R} = \sqrt{2.0 - \frac{2.84}{\sqrt{k}}}$$

$$S_{ij} = \frac{(URMS^{R} - URMS^{ij})}{URMS^{R}} \Delta(URMS^{R}, URMS^{ij})$$

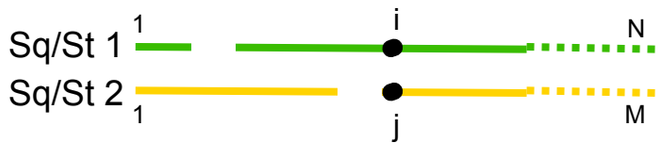
$$\Delta(URMS^{R}, URMS^{ij}) = 10 \Rightarrow URMS^{R} > URMS^{ij}$$
$$\Delta(URMS^{R}, URMS^{ij}) = 0 \Rightarrow URMS^{R} \leq URMS^{ij}$$

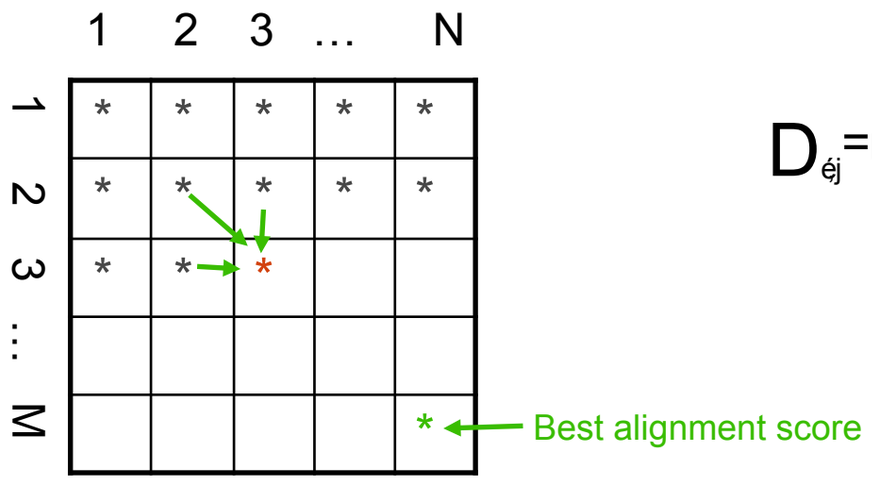
For each position i, the k consecutive unit vectors (k=6) are grouped and aligned to the j set of unit vectors. Each pair of aligned unit vectors will be evaluated by calculating Unit Root Mean Square distance ($URMS^{ij}$).

The obtained URMS values are compared the minimum expected URMS distance between two random set of k unit vectors ($URMS^{R}$).

The alignment score is than calculated normalizing $URMS^{ij}$ to the $URMS^{R}$ value.

Alignment





$$D_{ij} = \min \begin{cases} D_{i,j-1} + Score_{(-,rj)} \\ D_{i-1,j-1} + Score_{(ri,rj)} \\ D_{i-1,j} + Score_{(ri,-)} \end{cases}$$

Backtracking to get the best alignment

A Dynamic Programming procedure is then applied to search for the optimal structural alignment using a global alignment with zero end gap penalties.

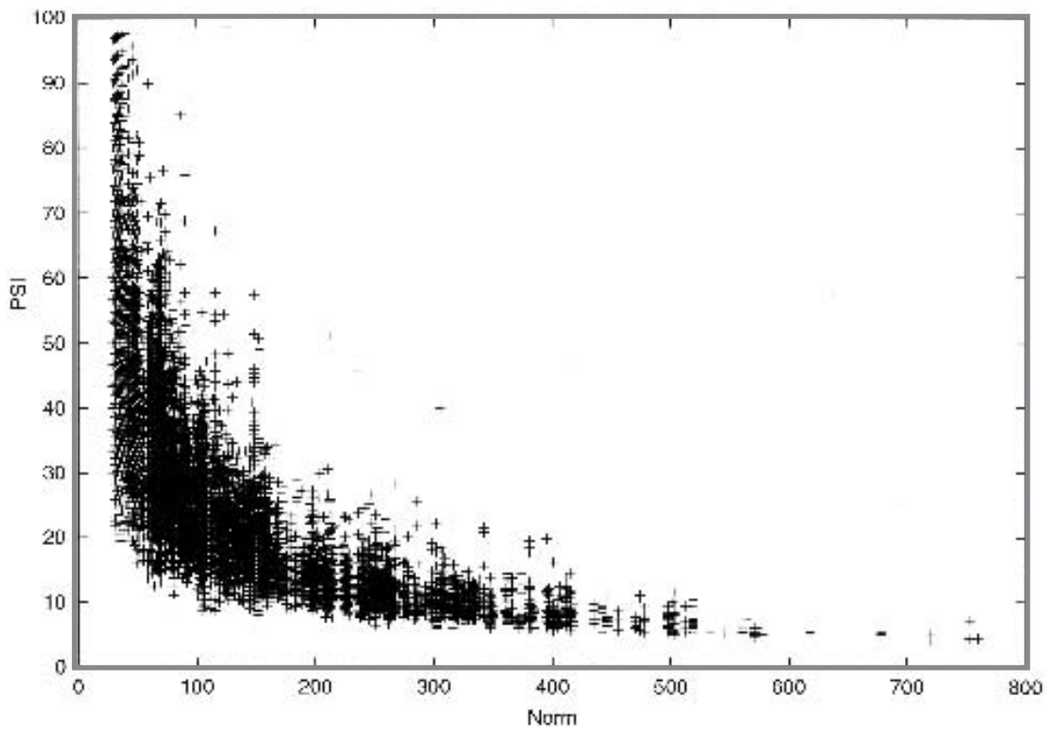
The maximum subset of local structures that have their corresponding Cα within 4.0 Å in the space are evaluated. The number of close atoms is used to evaluate the percentage of structural identity (PSI) using a variant of the MaxSub algorithm.

Needleman and Wunsch J. Mol.Biol 1970
Siew et al. Bioinformatics 2000

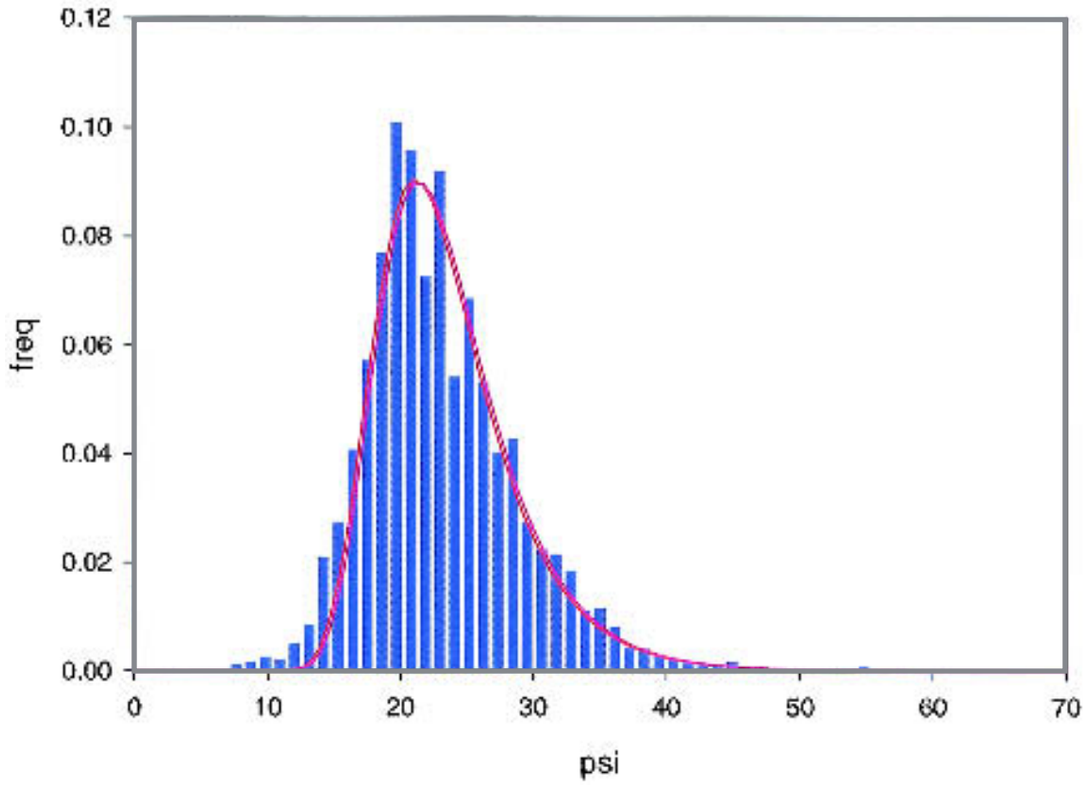
Background Distribution

Considering a dataset of random structures, it is possible to produce pairwise alignments that resulted in a empirical distribution of scores (s). From such distribution we can then evaluate μ and σ needed to calculated the p-value for P(s>x).

Empirical



Analytic



$$P(t > x) = \int_t^{\infty} f(x)dx = 1 - e^{-e^{\frac{-(x-a)}{b}}}$$

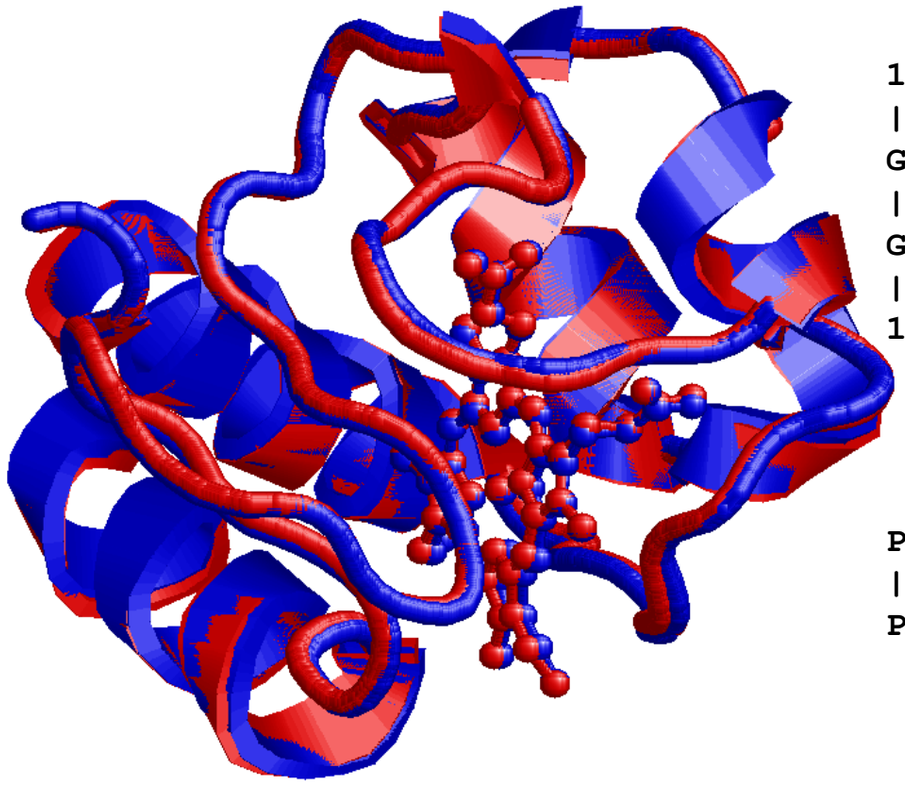
Ortiz et al. *Proteins* 2002

Exercise

Build a Python script for structure superimposition using the class SVDSuperimposer from the biopython libraries.

Test the script on a group of atoms from the following structures

Human Cytochrome C – Uniprot:P99999. PDB: 3ZCF:A
Equine Cytochrome C – Uniprot: P00004. PDB 3O20:A



```
1:A                  20:A                  40:A                  60:A
|        |    .    |    .    |    .    |    .    |    .    |    .    |
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLEN
||||||||||:.||.||||||||||||||||||||||||||||:.||.|||||||.|.|:|||||||||
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMEYLEN
|        |    .    |    .    |    .    |    .    |    .    |    .    |
1:A                  20:A                  40:A                  60:A

          80:A                  100:A
     .    |    .    |    .    |
PKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
||||||||||||.|||||.||.||||||||||||
PKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE
     .    |    .    |    .    |
          80:A                  100:A
```