Chapter 3 Structural Comparison and Alignment

3.1 Introduction

3.2 Main Methods

1. Basic algorithms review

Dynamic programming

Distance matrix

2. SARF2, VAST, COMPARER

3. CE, DALI, SSAP

3.3 Recent Methods

MAMMOTH, RAPIDO, SABERTOOT, TOPOFIT

3. Structural Comparison and Alignment
Introduction

Goal:

Determination of equivalences between amino acid residues by taking into account 3D structures

Relationships between primary protein sequence, 3D structure and biological function

Four steps

- **Structure alignment**: find equivalences of amino acid residues based on known 3D structures of protein folds
- **Structure comparison**: once shared similarities are known the structures are compared
- **Structure superposition**: find the optimal overlap of both proteins (matches information about residues for protein A and B is known)
- **Structure classification**: assign the protein to a certain class

Structural alignments provide information that is unavailable through current sequence alignment methods: distant sequence relationship

3. Structural Comparison and Alignment
Introduction

Motivation:

From Genome sequencing to amino acids/nucleotides primary structure
From amino acids/nucleotides primary structure to 3D Structure Prediction

2008 In PDB data base 49192 Structures structures
Feb 24, 2009 56066 Structures
2008 SWISS PROT 356 194 entries sequence
10-Feb-2009 UniProtKB/Swiss-Prot Release 56.8 of : 410 518 entries

Ratio of 1 structure to 7 sequences

3. Structural Comparison and Alignment
Introduction

Protein structures are more highly conserved than sequences : Evolutionary changes like insertions and deletions take place mainly in loop regions.

No alterations in the final fold and limiting the number of possible folds

Complexity is reduced through evolution maintaining diversity and adaptability

All proteins all species 1000-5000 protein folds (Chotia, 1992)

Similar structures may be formed by alternative folding of the amino acids' Cα backbone: matched regions separated by unmatched segments

Partial local similarities do not automatically transfer to similarities in structure

Same nucleus BUT different end

30% Sequence identity adopt the same folds: homologous folds

5% Similarity can result in the same fold: analogous folds

1 Basic algorithms review

Structures can be compared, assuming they adopt the same fold

Structural comparison and alignment as NP-hard problems: non-deterministic polynomial time problems solved by heuristic approaches

Possible solution as the best analytical answer but NO biological mean

Find the most suitable method to solve the optimization of the alignment and to reduce the computing time consuming problem: 5 of 10 structures inferred without special algorithms

New proteins weekly released in PDB previous all-against-all comparison

Known sequence-structure relationships are used

PDB structures are grouped (only a subset is compared)

1 Basic algorithms review

Steps for algorithm optimization

A. **Structure comparison and alignment**

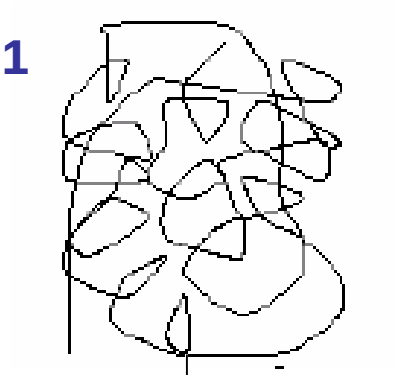
i. Representation of the pair of proteins 1 and 2, domains or fragments to be compared and aligned
ii. Compare 1 and 2
iii. Optimize the alignment between 1 and 2
iv. Statistically significant measurement of the alignment against a random set of structures

B. **Multiple structure alignment**

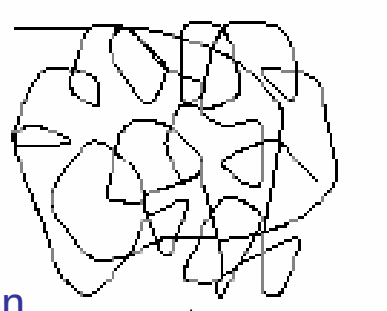
i. Starting from the initial alignment found in Aiii., the next step is running a search within a constraining sequence window to find the optimal alignment against all structures using profiles; HMMs or Monte Carlo approaches.

How to compare structures

Structure 1



Structure 2



Feature extraction

1. Structure description 1 — Structure description 2

2. Comparison algorithm

3. Scores

Statistical Significance (PDB)

Similarity-Classification

Dynamic programming in Structural Bioinformatics

Aligning Sequences :A row of amino acids in one sequence matches a row of identical or substituted positions in the second sequence; insertions or deletions as gaps

Aligning Structures: A scoring matrix is built to compare the positions of the atoms in both 3D structures

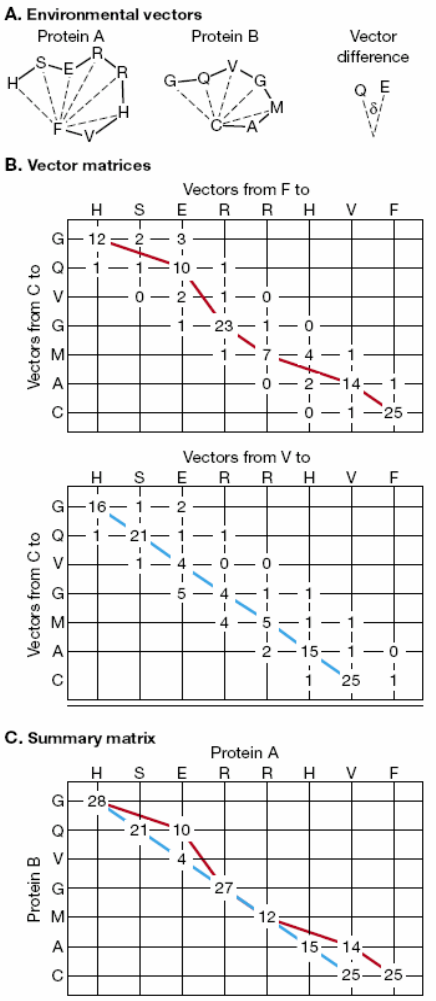
i. Scoring matrix gives scores of how well any of the 20 amino acids fits to a single position in the structure. Calculation of an optimal alignment

ii. Positions of SSEs within a domain: similar types, positions and numbers

iii. Distances between the Ca (NH- Ca- Cb) and Cb (Ca-CbO=NH) atoms within these domains, and later within the whole structure

iv. Determination of the degree of superimposition

Dynamic programming in Structural Bioinformatics



Two steps

1. Atoms or molecules as vectors:

A coded value is given describing the local environment of each amino acid

- Interatomic distances
- Bond angles
- R groups

Cartesian coordinates are assigned to each (X, Y, Z)

Direction of the bond angles is included

2. The alignment of 2D structures:

Determine of the interatomic distances between each amino acid in the polypeptide chain

"The better the arrangement, joining and 2D alignments are, the more significant and convincing is the result"

Distance matrix

No alignments help is needed

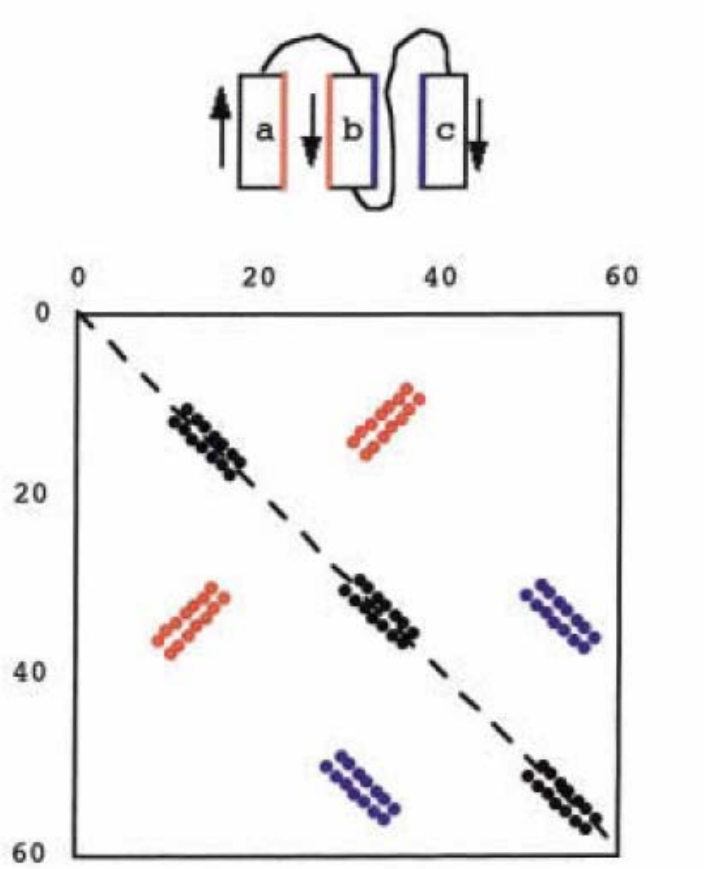
Each position in the 2D matrix represents the distance between corresponding Cα atoms in the 3D structure

i. Distances between Cα atoms along the polypeptide chain and between Cα atoms within the protein structure are compared

ii. Similar groups of 2D structural elements are superimposed (sum distance minimization in the aligned Cα atoms resulting in a common core)

"The smallest the distance, the most closely packed atoms within SSEs and regions of the 3D structures "

Distance matrix



Bases of Distance method

Degree to which all of the matched elements can be superimposed

Protein A ------- helices a and b interacting
Protein B ------- helices a' and b'

Helices superimposition-set of Cα

i^A and i^B in helix a and a'
j^A and j^B in helix b and b'

For matching pairs

dijA = distance between i^A and j^A

dijB = distance between i^B and j^B

SS = |dijA – dijB| / dij*

dij* = average of dijA and dijB

0.2< SS =0.2

SS 1Å β strands SS 2-3Å α and β

2. SARF2, VAST, COMPARER

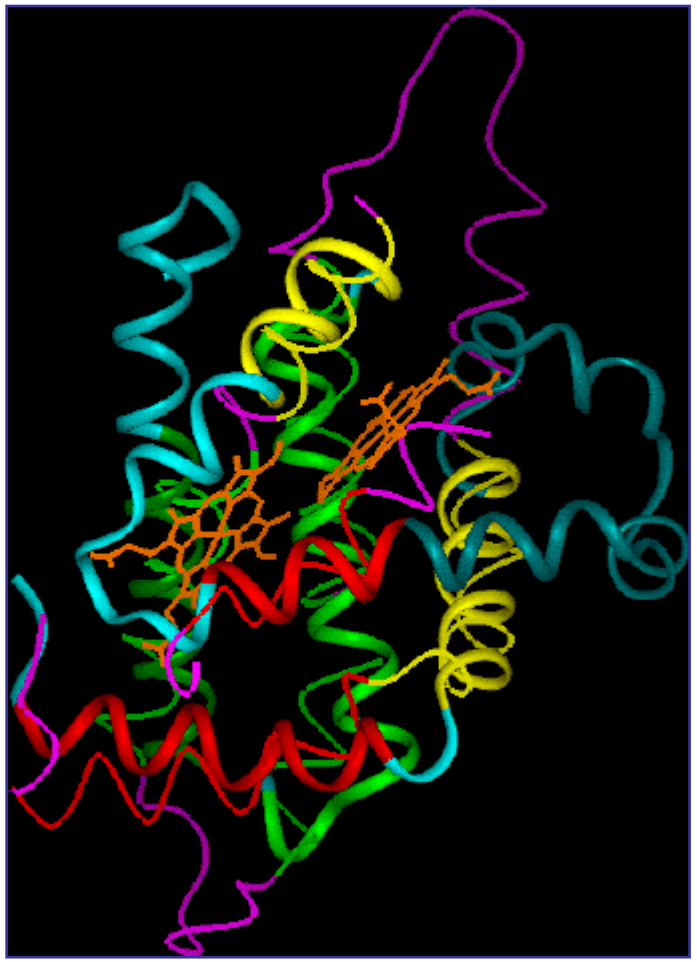
Components of structure elements to be compared:

- Local geometry (Cα, Cβ, Torsion angles)
- Side chain contacts
- Distance matrix
- Distances of inter and intra aligned fragment pairs
- Properties as SSs, hydrophobic clusters

SARF2 and VAST : predictions based on vector comparisons by converting

- Position
- Direction
- Length

Used to compare new structures to the existing DB or to view structural similarities already in the DB



http://123d.ncifcrf.gov/sarf2.html

2. **SARF2** **S**patial **A**rrangement of **B**ackbone **F**ragments (Nickolai N Alexandrov, 1998)

Based: comparison of Cα of each residue in the SSEs of each protein

Goal:to find those SSEs which can form similar spatial arrangements but have different topological connections

How: SSEs detected through comparison with common templates for α-helices and β- strands, then larger assemblies of SSEs are constructed from the compatible pairs found

Similarity Score: Calculated as a function of rmsd and the number of matched Cα atoms

RMSD: Measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimate→ Measure of accuracy (wikipedia)

The significance of the comparison is considered contrasting this score with the one built up once a protein is compared with a non redundant set of structures

2. SARF2

1st step: pairs of SSEs are matched up

Shortest distance between their axes

Closest point on the axes

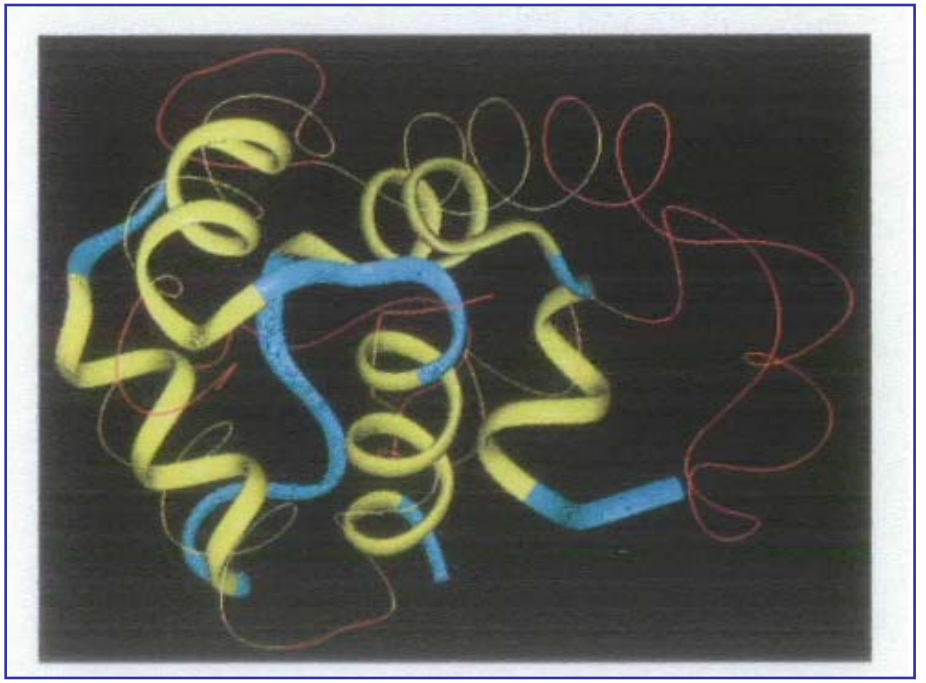
Minimum and maximum distances from each SSE

2nd Step: Largest ensembles are formed

Graph theory and maximum clique problem approximation

3rd Step: Extension of the alignment

Additional residues included



Blue ribbon shown as repressor 434 and recovering as red line.

Yellow fragments can be superimposed with rmsd = 2.61

52 Cα matched found

No evolutionary relationship but structural stability is apparent

2. VAST **V**ector **A**lignment **S**earch **T**ool (Gibrat et al,1996)

Based: SSEs-pair alignment

How: Structures as a set of vectors of secondary structural elements whose direction, type and connectivity infer the topology of the structure

Once the alignment is achieved, it uses Gibbs sampling algorithm to examine alternative alignments

Gibbs sampling: an algorithm to generate a sequence of samples from the joint probability distribution of two or more random variables (wikipedia)

The statistical theory similar to BLAST

BLAST: probability to get the same score when aligning a test sequence against a DB sequence would be found by comparing random sequences

VAST: SS is the likelihood that the score would be the result of a random alignment of unrelated structures

Score: number of superimposed SSEs

2. VAST

SS = N1 x N2

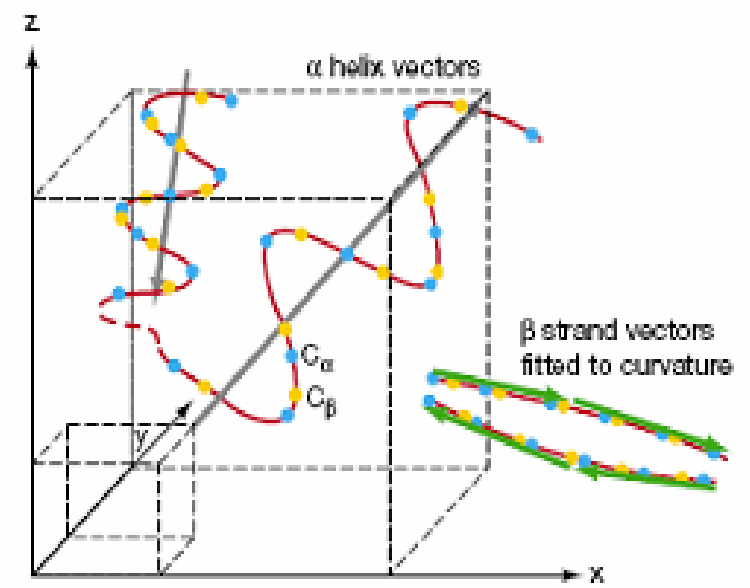
N1 = probability of picking up elements randomly and get the same score

N2 = number of alternative element pair combinations

Optimal alignment: the one with highest relation to the background distribution of Cα in the superimposed amino acid residues

Elements in two structures are similarly arranged so expected similarity on their corresponding 3D structures

Larger alignment groups by clustering the SSEs similarities found



3D structures of proteins are predicted to be similar if, once the representations of their vectors were compared, the type and arrangement are alike within a rational range

2. COMPARER (Sali and Blundell, 1990)

Based: on the sequence alignment algorithm of Needleman and Wunsch by using equivalences between protein structures to define general topologies

How: both comparison of properties and relationships through simulated annealing and dynamic programming

Properties: dynamic programming algorithm used to find an optimal alignment

Residues like identity and local conformation
Segments like SSs type and orientation relative to the center of gravity

Relationships: combinatorial simulated annealing technique

Relations between residues like hydrogen bonds and hydrophobic clusters
Relations between segments like distance to one or more closer neighbors and the relative orientation of two or more segments.

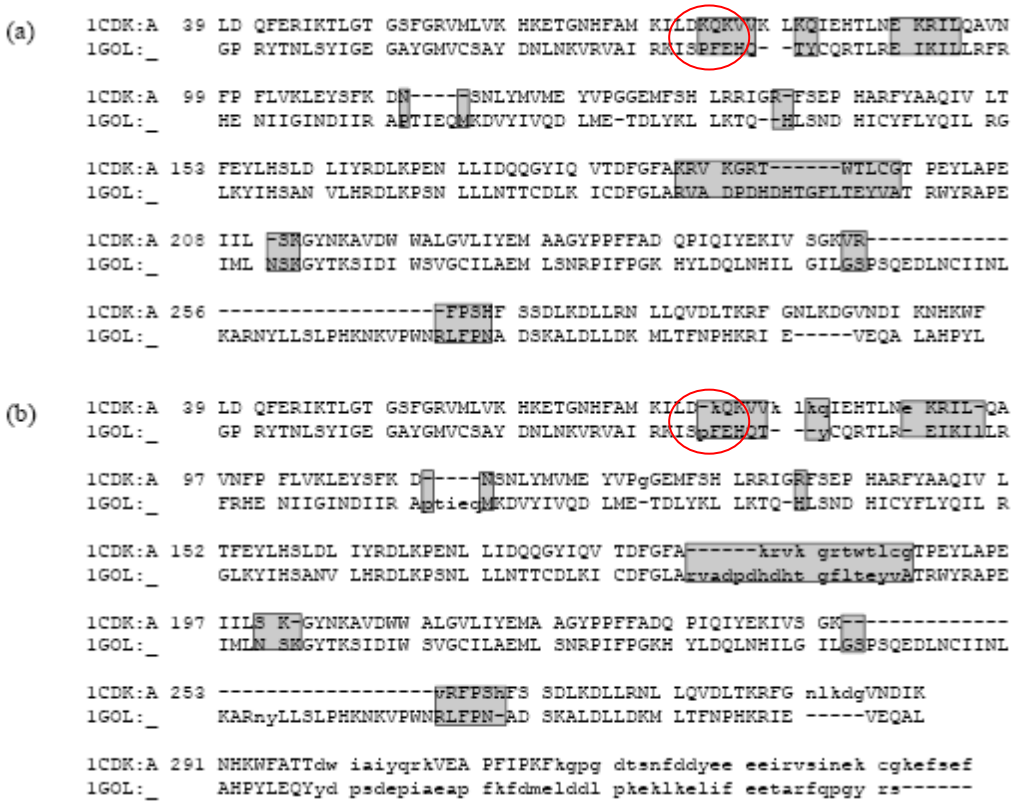
Up to 14 Properties and relationships are compared

For statistical analysis: E (residue equivalences) and A (gap penalties) comparing to the values obtained by using two unrelated proteins and two random sequence relationships

3. CE Combinatorial Extension of the Optimum Path (Shindyalov and Bourse, 1998)

Former heuristic methods whose results were contradictory:

Finding the best rmsd is not enough to match biologically meaningful features (the most significant structure alignment)
Different alignments produced by different methods
Close scores got from same methodology but residues in the alignments are far positioned



New algorithm that uses a combinatorial extension of the optimal path; the path is defined by the use of protein properties relevant to structural and functional features

3. CE

Based

Target function: heuristics assumes continuity and optimal path existence

Compare octameric fragments – an aligned fragment pair (AFP)

Distance matrices: distances between each Ca of each octamer fragment combination from both proteins is plotted and represented

Combinations of AFP "representing" possible continuous alignment path are selected and extended

Find the optimal path through the AFPs

Optimize the alignment through dynamic programming

Measure the statistical significance of the alignment

3. CE

Assumed rules

Remove highly homologous chains

The rmsd between two chains < 2Å

The length difference between two chains < 10%

The number of gap positions in alignment between two chains < 20% of aligned residue positions

At least 2/3 of the residue positions in the represented chain are aligned

3. CE

Alignment algorithm

Input and output of alignment algorithm

Input: two proteins:

$$A = \{a_1, \cdots, a_m\} \qquad B = \{b_1, \cdots, b_n\}$$

Output: An alignment

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

and scores

$$i_1 < i_2 < \cdots < i_L, \; j_1 < j_2 < \cdots < j_L$$

Constraints:

min rmsd:

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

max L

min Gaps:

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

Penalization gaps: Computational speed lost of non topological alignments and insertions of more than 30 residues

3. CE

Two methods for detecting structural homology

1. From ONLY structural information

Alignment Path

Distance Measure for Similarity Evaluation

2. From structural information AND adding composite properties

(i) Octamer A and Octamer B satisfy a similarity criterion: AFP

(ii)Three threshold

1st detecting AFP

2nd detecting the correctness of a next candidate AF relative to the current one

3rd evaluating all alignments to find the optimal ones

(iii) Statistical significance

Numerical table

Two distributions corresponding to both proteins rmsd and Gaps values for the non redundant set

Assuming normality the final z-score is calculated by combining both z-scores

3. CE Method 1. From ONLY structural information

Alignment Path

Selection of starting point by the ones leading the longest alignment found

Longest continuous path P of AFPs in a similarity matrix S

Protein A length: n^A
Protein B length: n^B
Similarity matrix size: $(n^A - m)$ $(n^B - m)$

AFPs *i* and *i*+1 extension if and only if

Condition (1): No Gaps between AFPs *i* and *i*+1

$$P^A_{i+1} = P^A_i + m \qquad P^B_{i+1} = P^B_i + m$$

Condition (2): Gaps inserted in protein A

$$P^A_{i+1} > P^A_i + m \qquad P^B_{i+1} = P^B_i + m$$

Condition (3): Gaps inserted in protein B

$$P^A_{i+1} = P^A_i + m \qquad P^B_{i+1} > P^B_i + m$$

3. CE Method 1. From ONLY structural information

Condition (4): Gaps on protein A ; Condition (5): Gaps on Protein B

$$P^A_{i+1} \leq P^A_i + m + G \qquad P^B_{i+1} \leq P^B_i + m + G$$

Distance Measure for Similarity Evaluation: 2 distances are measured and the rmsd

i. Using an independent set of inter-residue distances: to evaluate combination of two AFPs

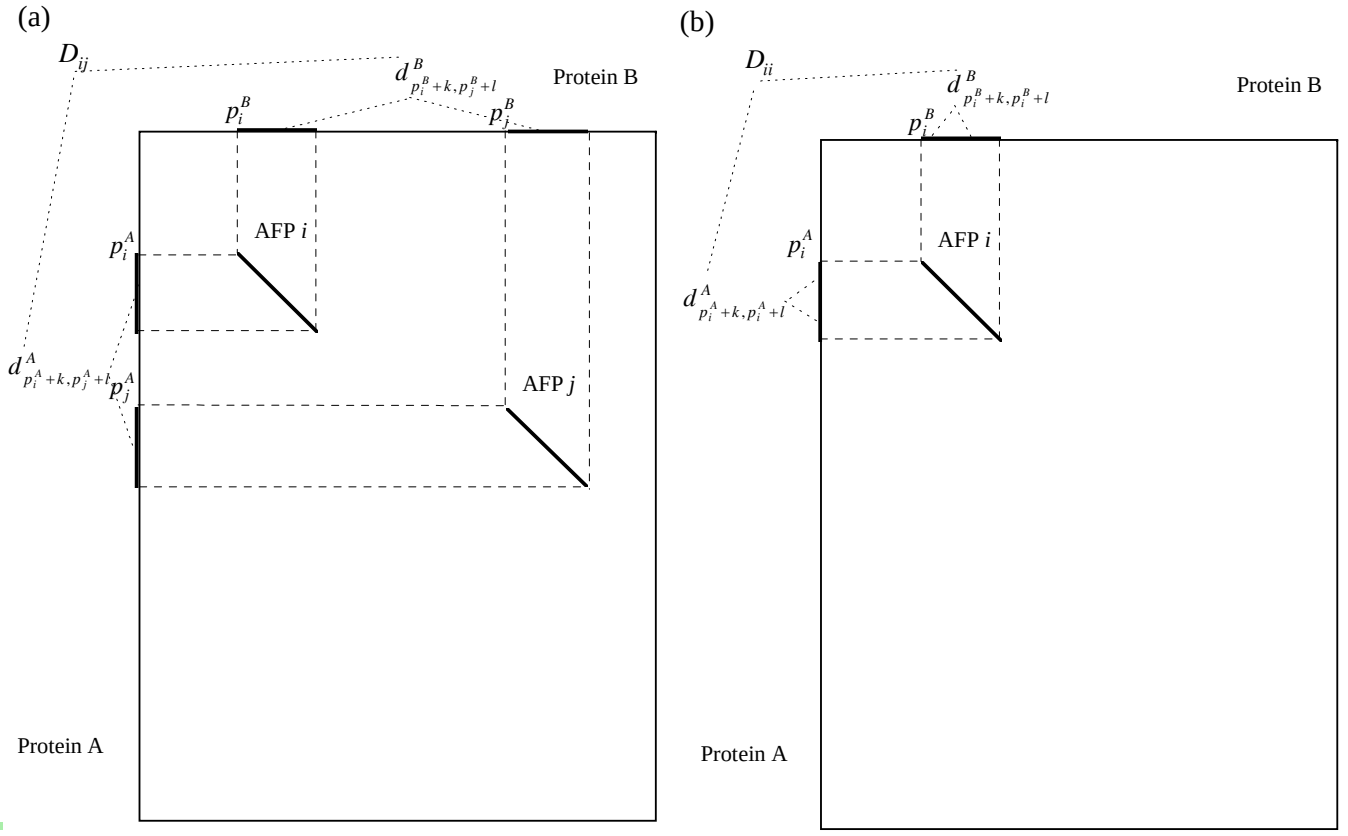
$$D_{ij} = \frac{1}{m}\left(\left| d^A_{P^A_i P^A_j} - d^A_{P^B_i P^B_j} \right| + \left| d^A_{P^A_i+m-1,\ P^A_j+m-1} - d^B_{P^B_i+m-1,\ P^B_j+m-1} \right| + \sum_{k=1}^{m-2} \left| d^A_{PA_i+k,\ P^A_j+m-1-k} - d^B_{P^B_i+k,\ P^B_j+m-1-k} \right| \right)$$

ii. Using a full set of inter-residue distances: to evaluate a single AFP

$$D_{ij} = \frac{1}{m^2}\left(\sum_{k=0}^{m-1} \sum_{l=0}^{m-1} \left| d^A_{P^A_i+k,\ P^A_j+l} - d^B_{P^B_i+k,\ P^B_j+l} \right| \right)$$

3. CE

Calculation of distance: a)D_{ij} for alignment represented by two AFPs i and j from the path, b) D_{ii} for a single AFP i from the path



3. CE

iii. RMSD obtained from structures optimally superimposed: to select the best alignments and for the optimization of gaps in the final alignment

When adding the next AFP three strategies can be followed

All possible AFPs which extend the path and satisfy the similarity criteria

Only the best AFP which extend the path and satisfy the similarity criteria

Intermediate criteria

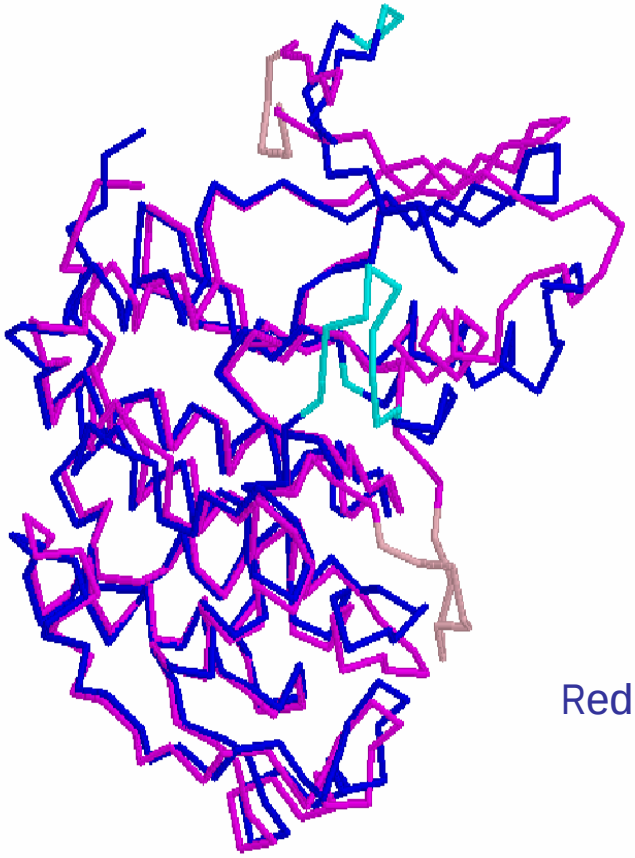
Three heuristic and three conditions to decide

Condition (6): Single AFP < 3Å $D_{nn} < D_0$

Condition (7): AFP against the path < 4Å $\frac{1}{n-1} \sum_{i=0}^{n-1} D_{in} < D_1$

Condition (8): Whole path $\frac{1}{i^2} \sum_{i=0}^{n} \sum_{j=0}^{n} D_{ij} < D_1$

3.CE Optimization of the Final Path



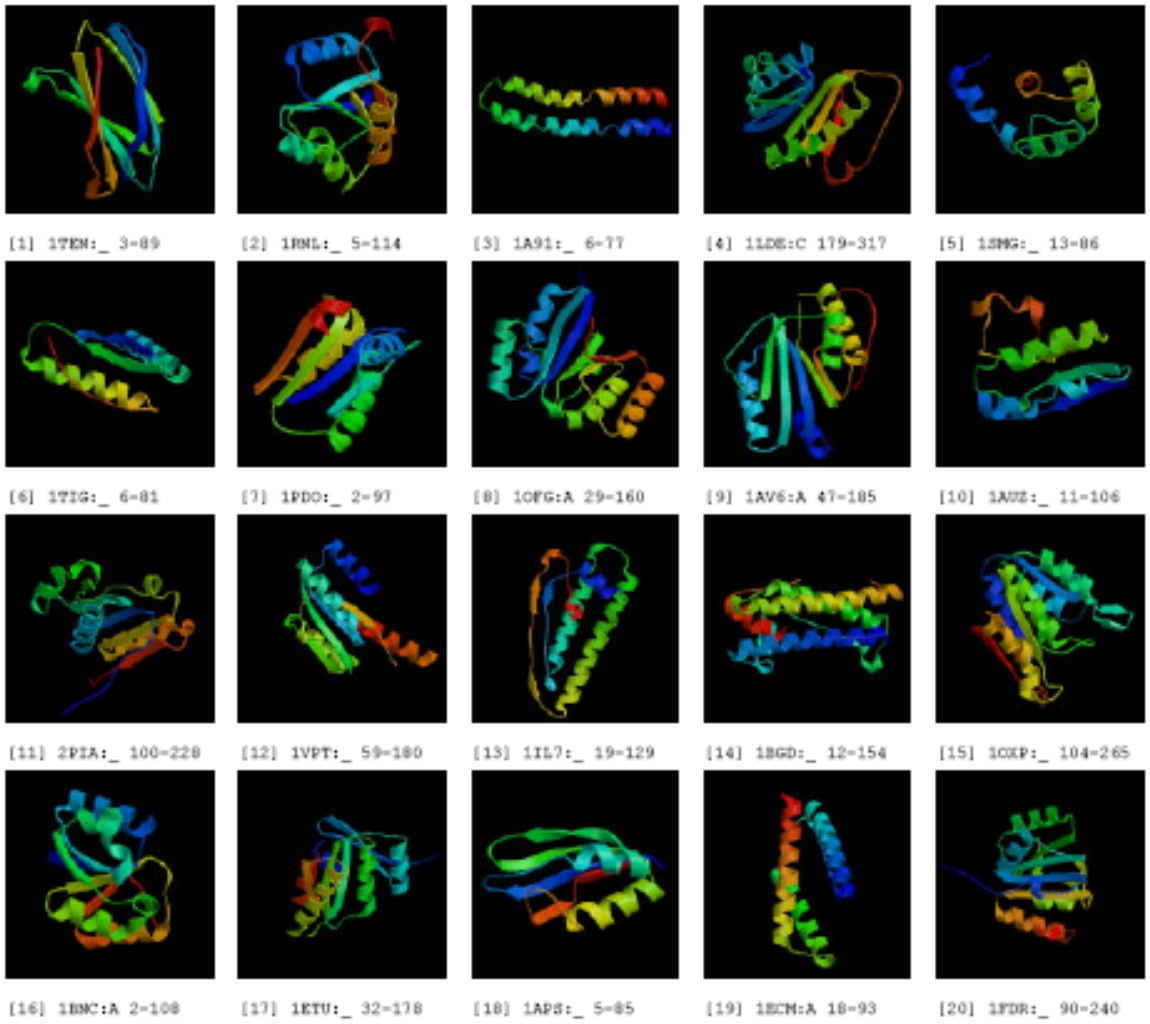
The 20 best alignments with a Z score > 3.5 are assessed
based on RMSD and the best kept: approx. one error in 1000 structures

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

Red-brown and light blue : Insertions

Will not find non-topological alignments (outside the bounds of the dotted lines)
CE works on chains and not in domains

3.CE Optimization of the Final Path



Gaps included and analyzed for relocation in both directions m/2

RMSD improvements in superimposed structures

New boundaries adopted

Dynamic programming on the distance matrix using residues from the 2 superimposed structures

3. CE Method 2. From structural information AND adding composite properties

Similarity is calculated by adding the following properties represented as scores

P_{ij} measures the match between residues i and j from two proteins A and B

d_{ij} distance between residues i and j in proteins A and B after CE superimposition

Structure: Property 1, defined by coordinates of Cα

$$P_{ij} = \begin{cases} c_1 - d_{ij}, & \text{if } c_1 - d_{ij} > c_2 \\ c_2, & \text{otherwise} \end{cases}$$

Sequence: Property 2, value of PET91 matrix for amino acids at positions i and j

Secondary structure: Property 3

$$P_{ij} = \begin{cases} 1, & \text{if } s_i = s_j \\ 0, & \text{otherwise} \end{cases}$$

Solvent Exposure: Property 4

$$P_{ij} = E_0 - |E_i - E_j|$$

Conservation Index: Property 5

$$P_{ij} = 20 - |I_i - I_j|$$

3. CE Method 2. From structural information AND adding composite properties

The calculus is done residue by residue

Dynamic programming to find the optimal alignment for the whole polypeptide chain

The composite property that measure structural similarity at residue level is defined

$$\tilde{P}_{ij} = \sum_k w_k * P_{ij}^k$$

Gap initialization penalty of 10 and gap extension penalty of 1

$$a^D = \sum_i a_i^D$$

$$a^D = \begin{cases} 1, & \text{if } a_i^1 \neq -1 \text{ and } a_i^1 \neq a_i^2 \\ 0, & \text{otherwise} \end{cases}$$

3. CE Method 2. From structural information AND adding composite properties

Method	PKA (1CDK:A) vs MAPK (1GOL:_) length of alignment = 248	PKA (1CDK:A) vs CDK2 (1FIN:A) length of alignment = 251
Dali	34 (13.7%)	30 (12.0%).
STR	8 (3.2%)	8 (3.2%)
STR+SEQ+CONS	3 (1.2%)	5 (2.0%)
SEQ	98 (39.5%)	76 (30.3%)
SS	76 (30.6%)	77 (30.3%).
CONS	84 (33.9%)	107 (42.6%)
EXP	45 (18.1%)	62 (24.7%)
STR+SEQ	4 (1.6%)	6 (2.4%)

STR: structure based on the rmsd calculated for the superposition of Cα atoms after optimal alignment found using the CE algorithm

SEQ: sequence based on PET91 amino-acid similarity measure by Jones and Thornton (1992)

SS: secondary structure based on the SSEs by Kabsch and Sander (1983)

EXP: solvent exposure based on the definition of Lee and Richards (1971)

CONS: conservation index based on sequences compiled for proteins with known structure

()Absolute difference between alignments

3. DALI, Distance Alignment Matrix (Holm and Sander, 1993a)

Based: use of distance matrices to represent each structure as a 2D array for aligning protein structures. Monte Carlo Simulation

Allowance: gaps of any length

reversal of chains in any direction

free topological connectivity

Two categories of searches

Finding predefined structural patterns in a database

Finding the largest common structure between two proteins

How:

- Submatrices of hexapeptides-hexapaptides contact patterns and their distances between Ca-Ca in the 3D are plotted
- Similarities in both matrices, for protein A and B, are paired and combined into larger combined sets of pairs (overlapping)

Similarity score optimized by Monte Carlo simulation and defined as equivalent intramolecular distances

3. DALI Method

Substructures of protein A and B matching by Additive similarity score

$$S = \sum_{i=1}^{L} \sum_{j=1} \Phi\ (i,j)$$

The larger the value of S, better set of residue equivalences

Based on Similarity measure of the Ca-Ca distances

$$\Phi^R(i,j) = \Phi^R - \left| d_{ij}^A - d_{ij}^B \right|$$

Geometrical distortions effects are reduced by including the elastic similarity of the residue-pairs score

$$\Phi^E\ (i,j) = \begin{cases} \left(\Phi^E - \frac{\left| d_{ij}^A - d_{ij}^B \right|}{d_{ij}^*} \right) w\ (d_{ij}^*), & i \neq j \\ \Phi^E, & i = j \end{cases}$$

Envelope function to weight the contribution of pairs in the long distance range

$$w(r) = e^{\frac{r^2}{\alpha^2}}$$

3. DALI

Summarized Method

Hexapeptides-hexapeptides contact patterns: equivalents fragments

Identification of new matching contact patterns sharing the previous equivalent fragment: (a,b)-(b,c)-(c,d)….

Iterative improvement to maximize the similarity of the alignment built up

Outcomes visualized

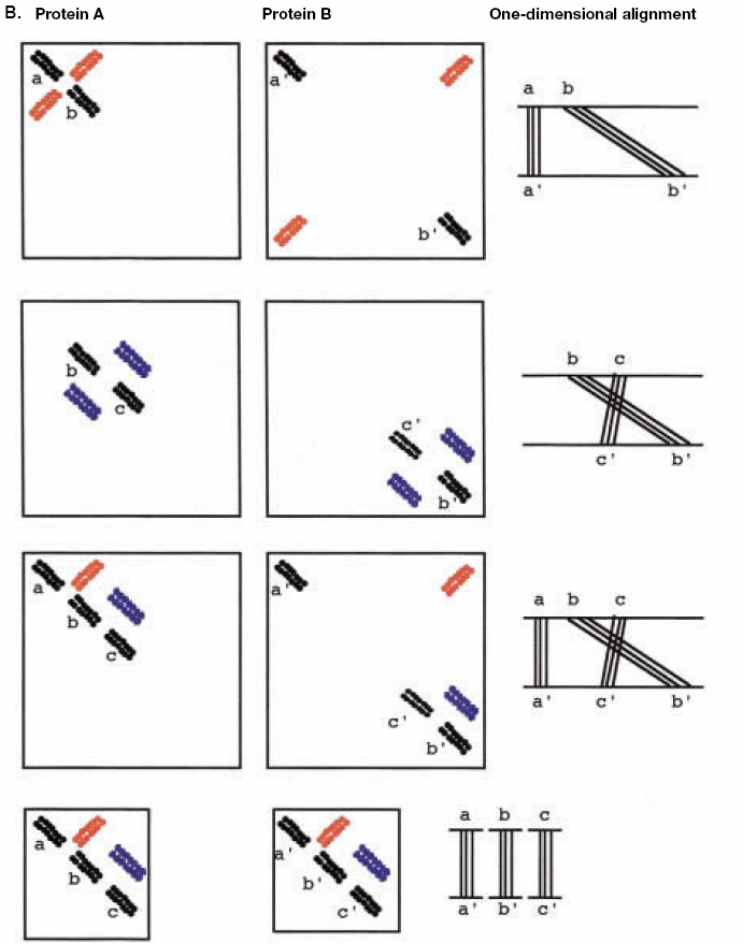
Matches substructures are patches

Main diagonal is formed when overlapping and centered

Locally similar backbone conformations: SSEs
Out of the diagonal: Tertiary structure similarities

Common motifs and structural motifs are represented as disjoint regions of the backbone

3. DALI



Protein A: Helices a, b

Protein B: Helices a' , b'

For each protein: sets of submatrices (6x6) overlapped from the whole matrix

Comparisons and combination: building up the complete alignment

Parallel alignment: insertions and deletions are removed

SS: all against all with < 30% sequence identity

Expressed as the number of standard deviations from the average score derived from the DB distribution

3. DALI Method

Adjacent strands in a b-sheet (distance is 4-5 Å) match within 1 Å

Strands-helix or helix-helix (distance is 8-15 Å) match within 2 Å

Aligns related proteins pairs and detects common 3D folding motifs in database search

Fast enough to scan the entire PDB looking for protein similar to a probe structure

FSSP (Fold Classification based on Structure-Structure alignment of Proteins) and DALI domain dictionary

Drawback: there is not an algorithm for direct alignment because it should find the closest alignment of 2 sets of points in 3D space and that is computationally a difficult problem

3.SSAP, Sequential Structure Alignment Program/Secondary Structure Alignment Program (Taylor and Orengo,1989)

Based:

Double dynamic programming (DDP) to obtain the optimal alignment in terms of matrices of:

A first matrix to get the **selected matches**. Distances between Cb- Cb of positions *i* and *j* of the proteins A and B to all the other proteins positions

A second matrix to get the **scores S_{ik}**. For every pair of positions *i* and *k* of proteins A and B, vectors between Cb at positions I and j are compared based on the first matrix (directionality)

Method

Each amino acid in each sequence is given a local environment

LE = ΣR + bonds angles + interatomic distances + degree of burial in hydrophobic core + type of secondary structure

3.SSAP

Interactomic vectors between positions *i* and *j* of the protein n: $\vec{v}_{nij}$

Average vector

$$\vec{r}_i = \frac{1}{n} \sum_{n=1}^{N} \vec{v}_{nij}$$

Error associated

$$e_{ij} = \frac{1}{N} \sum_{n=1}^{N} (\vec{r}_{ij} - \vec{x}_{ij})^2$$

Score

$$S_{ijmn} = (\vec{r}_{ij} - \vec{r}_{mn})^2$$

$$S_{ij} = \vec{A}_i - \vec{A}_j + \vec{r}_{ij}$$

Difference between the overage vectors of the two pairs residues in the two proteins

Shift vector to build up a consensus vector

$$\vec{A}_j^i = A_j + \frac{\vec{s}_{ij}}{e_j \mid j - i \mid^{\frac{1}{2}}}$$

Additional weight A reflecting the conservation of the error associated

3.SSAP

2rhe00	87.7	86.9	83.9	90.7	79.8	80.4	80.0	86.9	78.5	78.5
1cd800		84.7	76.0	87.4	80.1	79.4	80.2	87.5	78.6	78.4
3fabH1			77.0	85.7	78.2	79.8	79.2	88.6	73.0	79.3
3fabH2				74.4	85.5	84.1	86.8	77.7	91.0	84.8
3fabL1					80.6	76.5	80.0	86.9	79.1	76.6
3fabL2						86.4	88.4	80.3	86.5	86.0
1fc1A1							87.7	80.4	85.0	86.2
1fc1A2								80.9	88.2	89.1
2fb4H1									78.2	79.7
2fb4H2										84.8
3hlaB0										

Domains structures within these family

Local environments of given amino acid of both proteins are compared to find out the match residues

A scoring matrix is derived and the highest scoring region is chosen as the one that defines the optimal structural alignment

Those residues must be the ones having similar buried areas and torsion angles

3.SSAP

PDB code	Title	SSAP	Equivalent
1p11 452	Anthranilate isomerase	86.48 (76.9)	157
5timA 249	Triosephosphate isomerase	85.74 (100)	157
1wsyA 246	Tryptophan synthase	84.58 (68.8)	157
1ald 363	Aldolase A	84.25 (77.8)	157
5rubA 434	Rubisco	77.36 (68.5)	155
4enl 436	Enolase	75.75 (65.9)	141
2taaA 478	Taka-amylase	74.35 (62.9)	128
1ximA 392	Xylose isomerase	73.78 (70.8)	122
1dri 271	D-Ribose binding protein	69.76 (62.1)	139
1cseE 274	Subtilisin Carlsberg	69.23 (61.5)	122
2cmd 312	Malate dehydrogenase	68.78 (58.7)	133
2liv 344	Leucine binding protein	68.12 (60.6)	148
3grs 461	Glutathione reductase	66.53 (59.6)	133
1ldb 291	Lactate dehydrogenase	66.51 (59.9)	120
5p21 166	Ras p21 protein	65.85 (68.3)	122

Two levels of dynamic programming

1. Comparing residues environmnet between pairs of residues

2. Obtaining an alignment from accumulated data on residues pairs

Related folds have more variation on the loops and orientation of Secondary Structures

The SSAP cut off

Similarity of 70%: fold families 150

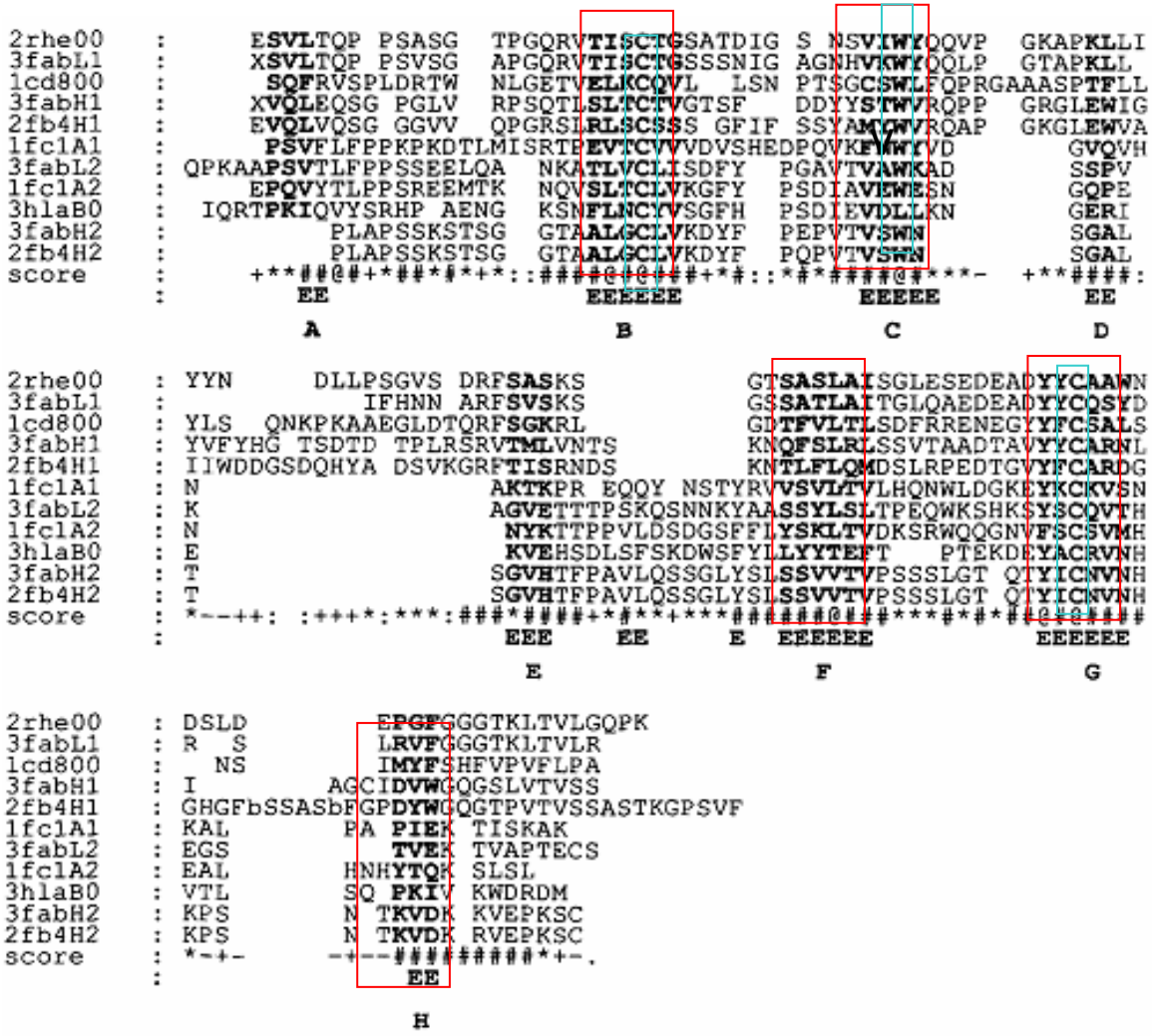
Similarity of 70-80%: analogous folds (variations in loops and orientation of secondary structure)

Similarity of 80%: fold families 200

Similarity of >80: homologous fold (divergence from a common ancestor)

3. Structural Comparison and Alignment
3.2 Main Methods

3.SSAP

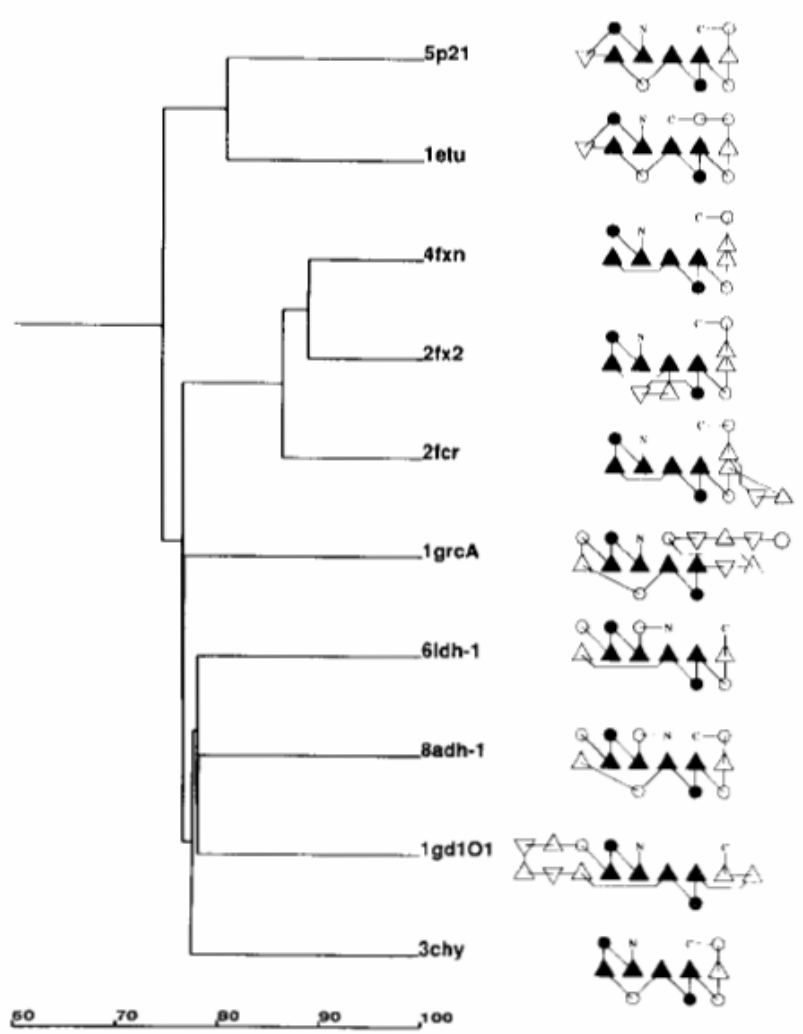


Multiple protein structure alignment derived from pairwise simple alignment concatenation

β Strands constant domains of inmunoglobulins (A, B, C, D, E, F, G, H)

Residues in equivalent SSs regions derived from hydrogen bonds patterns are highlighted

3.SSAP



The comparison of relationship improves the alignment of distantly related structures

The structural relationships fold shape are more easily recognized although secondary structure changes reduce the number of superimposable residues between not close related proteins

SSAP Dendrogram : Structural relationship of inmunoglobulins domains

Helices ●

Strands ▲

3. Structural Comparison and Alignment
3.2 Main Methods

ALGORITHM	DESCRIPTION	FOCCUS	STATISTICAL ANALYSIS	ADVANTAGES	DRAWBACKS
DALI	[a]Distance Matrix Alignment	Complete sequence. Distances between all Cα atoms	Score derived from all against all comparisons. Z-score as the number of standard deviations from the average score derived from the DB distribution.	One single frame of representation. Speed of execution Ability to recognize distant relationships	Not algorithm for direct alignment. Statistical significance based on rmsd value which is considered suboptimal Non topological regions are not detected
CE	[b]Combinatorial Extension of the Optimum Path	Distance between Cα of octameric fragments (combinatorial properties)	Tabulation of rmsd of the distributions of both proteins Z-score as result of combination of both z-scores	Computational Speed High percentage of homology detection	Reduction in the accuracy Domains miss recognition "Non topological" recognition or detection
SSAP	[c]Sequential Structure Alignment Program	Domain level Intraprotein Cβ-Cβ vectors comparison	Scores compared against CATH database.	Dealing with internal domains Generation of multiple alignment by pairs alignments concatenation	Global alignment is missed once the whole structure is broken down into small SSEs Lost of details when just Cβ-Cβ are compared
VAST	[d]Vector Alignment Search Tool	SSEs Vectors comparisons	P-value calculated for the best substructure superposition as if randomly obtained multiplied of alternative substructure alignments possible.	Computational time saved: SSEs converted into vectors	The whole 3D structure can not be used but just the predefined SSEs. Not complete SSEs Cα coordinates represented but just the beginning and the ends
SARF2	[e]Spatial Arrangement of Backbone Fragments	Superimposables SSEs comparing typical αhelix and βstrands templates	Score as a function of rmsd and the number of matched Cα atoms Comparison of scores obtained from non redundant set of structures	Computational time saved: SSEs converted into vectors Difficult cases detection	The whole 3D structure can not be used but just the predefined SSEs. Not complete SSEs Cα coordinates represented but just the beginning and the ends
COMPARER	[f]Comparer	Comparison of residues properties and relationships	Two scores E and A are contrasted, residues equivalences and gaps penalties respectively	Residues properties and relationships and segments relationships are studied at once	DP NOT applicable to relationships due to the dependence of the scores for a given relationship on the assignment of other relationships

3. GANGSTA, Genetic Algorithm for Non-sequential, Gapped protein STructure Alignment (Kolbeck B et al 2006 **)**

Based:

Non-sequential protein structure alignment using a two-level hierarchical approach

Sequential alignment: respecting the sequential order of the SSEs in the polypeptide chains of the considered protein pair non-sequential alignment: ignoring the order

Method:

First level, pairwise contacts and relative orientations between SSEs are maximized using a genetic algorithm (GA) and protein graph representation

Second level, pairwise residue contact maps resulting from the best SSE alignments are optimized

GANGSTA+

Combinatorial algorithm for non-sequential structural alignment of proteins and similarity search in database SSE pairs can optionally be aligned in reverse orientation

3. MAMMOTH, **MA**tching **M**olecular **M**odels **O**btained from **TH**eory

Based:

Developed for comparing models coming from structure prediction (**TH**eory)

Tolerant of large unalignable regions

To work well with experimental models (especially when looking for remote homology)

Genomic scale normalization: is being facilitated by a highly complete database of mammoth-based structure annotation for the predicted structures of unknown proteins covering 150 genomes

Method:

Heptapeptides ´from protein A and B are compared

Similarity score between two heptapeptides is calculated using a unit-vector RMS (URMS) method (molecular dynamics trajectories)

Scores stored in a similarity matrix, and with dynamic programming the optimal residue alignment is calculated

Similarity scores are derived from the likelihood of obtaining a given structural alignment by chance

3. RAPIDO, Rapid Alignment of Proteins In terms of Domains, (Mosca, Schneider TR 2008)

Based:

web server for the 3D alignment of crystal structures of different protein molecules (taking into account conformational changes)

Method

Identifies similar fragments in the two proteins using difference distance matrices

Matching Fragment Pairs (MFPs) are represented as nodes in a graph for the identification of the longest path on a DAG (Directed Acyclic Graph)

Method

Final step of refinement to improve the quality of the alignment

After aligning a genetic algorithm is applied for the identification of conformationally invariant regions (groups of atoms whose interatomic distances are constant)

IRs represent reliable sets of atoms for the superposition of the two structures that can be used for a detailed analysis of changes in the conformation

RAPIDO can identify structurally equivalent regions on fragments that are distant in terms of sequence and separated by other movable domains

3. SABERTOOTH

Based

Vectorial representation

Structural profiles to perform structural alignments

The underlying structural profiles expresses the global connectivity of each residue

Method

Recognizes structural similarities with accuracy comparable to SARF2, VAST

Algorithm has favorable scaling of computation time with chain length

Algorithm is independent of the details of the structural representation

The framework can be generalized to sequence-to-sequence and sequence-to-structure comparison within the same setup

3. TOPOFIT, novel common volume superimposition (Valentin A and col.2004)

Based

Model based common sub-groups to produce structural alignment

Structurally related proteins have a common spatial invariant part (set of tetrahedrons or common spatial sub-graph volume

Identifies common, invariant structural parts between proteins

Method

Similarity of protein structures is analyzed using three-dimensional Delaunay triangulation patterns derived from backbone representation

Method

The superimposition of those groups patterns allows to identify a common number of equivalent residues in the structural alignment

Identifies a feature point on the RMSD/Ne curve (structures correspond to each other including backbone and inter-residue contacts)

Larger RMSD corresponds to a growing number of mismatches between the patterns

The topomax point is present in all alignments from different protein structural classes

Understanding the molecular principles of 3D structure organization and functionality

Helps to detect conformational changes, topological differences in variable parts