Chapter 3 Structural Comparison and Alignment

3.1 Introduction
3.2 Main Methods
   1. Basic algorithms review
      Dynamic programming
      Distance matrix
   2. SARF2, CE, DALI, SSAP

3.3 Recent Methods
   MAMMOTH, RAPIDO, SABERTOOT, TOPOFIT
3. Structural Comparison and Alignment
3.2 Main Methods

2. SARF2, VAST, COMPARER

Components of structure elements to be compared:

- Local geometry ($C_\alpha$, $C_\beta$, 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 Spatial Arrangement of Backbone Fragments** *(Nickolai N Alexandrov, 1998)*

**Based**
comparison of $C_\alpha$ 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 $\alpha$-helices and $\beta$-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 $C_\alpha$ 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
3. Structural Comparison and Alignment
3.2 Main Methods

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
3. Structural Comparison and Alignment

3.2 Main Methods

- **From PDB services**

  - **ENDscript**
    A Web server for searching homologous sequences and giving information on secondary structure elements, accessibility, hydropathy and protein-protein contacts
  - **ESPrint**
    Easy Sequencing in Postscript
  - **Non-covalent bond finder**
    Software for finding non-covalent interactions for use with Chime 2 or higher
  - **Procheck**
    A program that checks the stereochemical quality of a protein structure
  - **ProFit**
    A program for fitting protein structures onto each other
  - **SARF2**
    A program which searches for similar structural motifs (via an analysis of backbone fragments) in protein structures
  - **STRAP**
    Structural Alignment Program for Proteins, interactive extendable and scriptable editor for large protein alignments which integrates amino acid sequence, secondary structure, 3D structure and genomic and mRNA sequence; Windows, Mac OSX, Unix, Linux
  - **Surface Racer**
    A program that calculates exact accessible surface area, molecular surface area and average curvature of molecular surface, and analyzes cavities in the protein interior inaccessible from the outside.
  - **SURFNET**
    A program which generates surfaces and void regions between molecular surfaces
  - **WHAT_CHECK**
    A system for protein structure validation derived from the WHAT IF program
  - **WHAT IF**
    A protein structure analysis program that may be used for mutant prediction, structure verification and molecular graphics
3. Structural Comparison and Alignment

3.2 Main Methods

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 $x_\alpha$ 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. Structural Comparison and Alignment

3.2 Main Methods

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. Structural Comparison and Alignment
3.2 Main Methods

3. CE

Alignment algorithm

Input and 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:

\[ \text{rmsd} = \min_r \sqrt{\frac{\sum_{k=1}^{L} (a_{i_k} - T_{b_{j_k}})^2}{L}} \]

max L

min Gaps:

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

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. Structural Comparison and Alignment
3.2 Main Methods

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: \((n^A - m) \times (n^B - m)\)

**AFPs \( i \) and \( i+1 \) extension if and only if**

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

\[ P^A_{i+1} = P^A_i + m \]
\[ P^B_{i+1} = P^B_i + m \]

Condition (2): Gaps inserted in protein A

\[ P^A_{i+1} > P^A_i + m \]
\[ P^B_{i+1} = P^B_i + m \]

Condition (3): Gaps inserted in protein B

\[ P^A_{i+1} = P^A_i + m \]
\[ P^B_{i+1} > P^B_i + m \]
3. Structural Comparison and Alignment  
3.2 Main Methods

3. CE Method 1. From ONLY structural information

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

\[ P_{i+1}^A \leq P_i^A + m + G \]

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

\[
D_{ij} = \frac{1}{m} \left( \left| d_{P_i^A P_i^A}^A - d_{P_i^B P_j^B}^B \right| + \left| d_{P_i^A+m-1, P_i^A+m-1}^A - d_{P_i^B+m-1, P_j^B+m-1}^B \right| + \sum_{k=1}^{m-2} \left| d_{P_i^A+k, P_j^A+m-1-k}^A - d_{P_i^B+k, P_j^B+m-1-k}^B \right| \right)
\]

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

\[
D_{ij} = \frac{1}{m^2} \left( \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} \left| d_{P_i^A+k, P_j^A+l}^A - d_{P_i^B+k, P_j^B+l}^B \right| \right)
\]
3. Structural Comparison and Alignment
3.2 Main Methods

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.

(a) $D_{ij}$

(b) $D_{ii}$
3. Structural Comparison and Alignment

3.2 Main Methods

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. Structural Comparison and Alignment
3.2 Main Methods

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. Structural Comparison and Alignment

3.2 Main Methods

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 $C_\alpha$

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

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

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

**Secondary structure**: Property 3

**Solvent Exposure**: Property 4

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

**Conservation Index**: Property 5

$$P_{ij} = 20 - |I_i - I_j|$$
3. Structural Comparison and Alignment

3.2 Main Methods

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

\[ P_{ij} = \sum_k w_k \times P_{ij}^k \]

Gap initialization penalty of 10 and gap extension penalty of 1

\[ \alpha^D = \sum_i \alpha_i^D \]

\[ \alpha^D = \begin{cases} 1, & \text{if } \alpha_i^1 \neq -1 \text{ and } \alpha_i^1 \neq \alpha_i^2 \\ 0, & \text{otherwise} \end{cases} \]
3. Structural Comparison and Alignment

3.2 Main Methods

3. CE Method

2. From structural information AND adding composite properties

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

STR: structure based on the rmsd calculated for the superposition of $C_\alpha$ 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. Structural Comparison and Alignment

3.2 Main Methods

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. Structural Comparison and Alignment

3.2 Main Methods

3. DALI Method

Substructures of protein A and B matching by Additive similarity score

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

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

Based on Similarity measure of the Ca-Ca distances

\[ \Phi^R(i, j) = \Phi^R - |d^A_{ij} - d^B_{ij}| \]

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

\[ \Phi^E(i, j) = \begin{cases} \left( \Phi^E - \frac{|d^A_{ij} - d^B_{ij}|}{\Phi^E} \right) w(d^*_{ij}), & i \neq j \\ \Phi^E, & i = j \end{cases} \]

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

\[ w(r) = e^{-\frac{r^2}{\alpha^2}} \]
3. Structural Comparison and Alignment

3.2 Main Methods

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. Structural Comparison and Alignment

3.2 Main Methods

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. Structural Comparison and Alignment

3.2 Main Methods

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. Structural Comparison and Alignment

3.2 Main Methods


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

$$LE = \sum R + \text{bonds angles} + \text{interatomic distances} + \text{degree of burial in hydrophobic core} + \text{type of secondary structure}$$
3. Structural Comparison and Alignment

3.2 Main Methods

3. SSAP

Interactomic vectors between positions $i$ and $j$ of the protein $n$:

Average vector

$$\bar{v}_{ni}$$

Error associated

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

Score

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

$$S_{ij} = A_i - A_j + \bar{r}_{ij}$$

Shift vector to build up a consensus vector

$$A_j = A_j + \frac{S_{ij}}{e_j |j - i|^{\frac{1}{2}}}$$

Additional weight $A$ reflecting the conservation of the error associated
3. Structural Comparison and Alignment
3.2 Main Methods

3. SSAP

<table>
<thead>
<tr>
<th>Structure</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2rhe00</td>
<td>87.7</td>
</tr>
<tr>
<td>1cd800</td>
<td>84.7</td>
</tr>
<tr>
<td>3fabH1</td>
<td>77.0</td>
</tr>
<tr>
<td>3fabH2</td>
<td>74.4</td>
</tr>
<tr>
<td>3fabL1</td>
<td>80.6</td>
</tr>
<tr>
<td>3fabL2</td>
<td>86.4</td>
</tr>
<tr>
<td>1fc1A1</td>
<td>87.7</td>
</tr>
<tr>
<td>1fc1A2</td>
<td>80.9</td>
</tr>
<tr>
<td>2fb4H1</td>
<td>78.2</td>
</tr>
<tr>
<td>2fb4H2</td>
<td>84.8</td>
</tr>
<tr>
<td>3hlaB0</td>
<td></td>
</tr>
</tbody>
</table>

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. Structural Comparison and Alignment

3.2 Main Methods

### 3. SSAP

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

Two levels of dynamic programming

1. Comparing residues environment 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 immunoglobulins (A, B, C, D, E, F, G, H)

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

Helices

Strands
## 3. Structural Comparison and Alignment
### 3.2 Main Methods

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Description</th>
<th>Focus</th>
<th>Statistical Analysis</th>
<th>Advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>DALI</td>
<td>Distance Matrix Alignment</td>
<td>Complete sequence, Distances between all Cα atoms</td>
<td>Score derived from all against all comparisons. Z-score as the number of standard deviations from the average score derived from the DB distribution.</td>
<td>One single frame of representation. Speed of execution. Ability to recognize distant relationships.</td>
<td>Not algorithm for direct alignment. Statistical significance based on rmsd value which is considered suboptimal. Non topological regions are not detected.</td>
</tr>
<tr>
<td>CE</td>
<td>Combinatorial Extension of the Optimum Path</td>
<td>Distance between Cα of octameric fragments (combinational properties)</td>
<td>Tabulation of rmsd of the distributions of both proteins Z-score as result of combination of both z-scores</td>
<td>Computational speed. High percentage of homology detection.</td>
<td>Reduction in the accuracy. Domains miss recognition. &quot;Non topological&quot; recognition or detection.</td>
</tr>
<tr>
<td>SSAP</td>
<td>Sequential Structure Alignment Program</td>
<td>Domain level Intraprotein CB-Cα vectors comparison</td>
<td>Scores compared against CAIIN database.</td>
<td>Dealing with internal domains. Generation of multiple alignment by pairs alignments concatenation.</td>
<td>Global alignment is missed since the whole structure is broken down into small SSEs. Lot of details when just CB-Cα are compared.</td>
</tr>
<tr>
<td>VAST</td>
<td>Vector Alignment Search Tool</td>
<td>SSEs Vectors comparisons</td>
<td>P-value calculated for the best substructure superposition as if randomly obtained multiplied of alternative substructure alignments possible.</td>
<td>Computational time saved. SSEs converted into vectors.</td>
<td>The whole 3D structure cannot be used but just the predefined SSEs. Not complete SSEs Cα coordinates represented but just the beginning and the ends.</td>
</tr>
<tr>
<td>SARF2</td>
<td>Spatial Arrangement of Backbone Fragments</td>
<td>Superim posables SSEs comparing typical helix and βstrands templates</td>
<td>Score as a function of rmsd and the number of matched Cα atoms. Comparison of scores obtained from non-redundant set of structures.</td>
<td>Computational time saved. SSEs converted into vectors. Difficult cases detection.</td>
<td>The whole 3D structure cannot be used but just the predefined SSEs. Not complete SSEs Cα coordinates represented but just the beginning and the ends.</td>
</tr>
<tr>
<td>COMPARER</td>
<td>Comparer</td>
<td>Comparison of residues properties and relationships</td>
<td>Two scores E and A are contracted, residues equivalences and GAPS penalties respectively</td>
<td>Residues properties and relationships and segments relationships are studied at once.</td>
<td>DP NOT applicable to relationships due to the dependence of the scores for a given relationship on the assignment of other relationships.</td>
</tr>
</tbody>
</table>
3. Structural Comparison and Alignment

3.3 Recent Methods

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
Non-sequential alignment: ignoring the order of the considered protein pair

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. Structural Comparison and Alignment

3.3 Recent Methods

3. MAMMOTH, MAtrching Molecular Models Obtained from THeory

Based:
Developed for comparing models coming from structure prediction (THeory)
Tolerant of large unalienable 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)
3. Structural Comparison and Alignment

3.3 Recent Methods

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,
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
3. Structural Comparison and Alignment
3.3 Recent Methods

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.