Structure Comparison and Alignment

Modified from the slides by Ilya Shindyalov, SDSC/UCSD
(http://www.sdsc.edu/pb/edu/pharm201/11/11.ppt)
Agenda

• Concepts:
  – Similarity vs. Homology
  – Structure vs. Sequence Similarity
  – Alignment: Pairwise vs. Multiple
  – Complexity vs. Performance
  – Advantages vs. Bottlenecks

• Structural similarity measures
• Structure Similarity: “How To” or “One of the problems which would never be solved finally”? 
• In depth analysis of one algorithm (CE)
• How new biological insights come from structure similarity?
Structure Similarity: “How To” or “One of the Problems which would never be solved finally”? 
How to build structural alignment statistics?

**Parameters/factors to consider:**

Baseline: what is random structure?

Distance: RMSD vs. intra-protein distances or contacts/interactions?

Gaps vs. unaccounted unaligned areas?

Similar sequence vs. similar structure?

Known functionally important residues vs. similarity in structure?

How to handle structural movements and flexible areas?
How to evaluate significance of the structural alignment?

For sequence alignment the baseline is random sequence. What is random structure?

Is this random?

Or this?
Approach of Levitt and Gerstein (1998)

1. Use a resource where structure similarity is defined – SCOP.

2. Find a scoring function which allows good separation of true and false positives:

\[
S_{str} = M \left( \sum \frac{1}{1 + \left( \frac{d_{ij}}{d_0} \right)^2} - \frac{N_{gap}}{2} \right)
\]

\(d_{ij}\) - distance between residues \(i\) and \(j\) in alignment;
\(M = 25;\)
\(d_0 = 5\, \text{Å};\)
\(N_{gap}\) – total number of gaps in local alignment (i.e. not counting terminal gaps);
Approach of Levitt and Gerstein (1998)
Biological vs Geometric Alignments
Plastocyanin versus Azurin (from Godzik, 1996)

Maintain 9 of 10 interactions
RMSD 1.5 Å

Maintain 5 of 10 interactions
RMSD 0.5 Å
CDK2 in two functional states

while sequence is identical, there is a drastic movement of catalytic loop

structural alignment doesn’t align the catalytic loop
Contradictions between interactions of sequence with structural alignment?

Existing algorithms do not deliver the answer.
Current State of the Art

• There are many papers published on this, but relatively few have code to download or Web sites from which to perform comparisons
• 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
# Common Methods

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Citations by 05/02</th>
<th>URL and Web Resource Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>Combinatorial Extension of the Optimum Path (Shindyalov and Bourne, 1998)</td>
<td>76</td>
<td><a href="http://cl.sdsc.edu/ce.html">http://cl.sdsc.edu/ce.html</a>&lt;br&gt;Shindyalov and Bourne (2001)</td>
</tr>
<tr>
<td>DALI</td>
<td>Distance Matrix Alignment (Holm and Sander, 1993)</td>
<td>890</td>
<td><a href="http://www.ebi.ac.uk/dali/">http://www.ebi.ac.uk/dali/</a>&lt;br&gt;Deitmann et al., (2001)</td>
</tr>
<tr>
<td>SSAP</td>
<td>Sequential Structure Alignment Program (Taylor and Orengo, 1989)</td>
<td>248</td>
<td><a href="http://www.biochem.ucl.ac.uk/~orengo/~ssap.html">http://www.biochem.ucl.ac.uk/~orengo/~ssap.html</a></td>
</tr>
</tbody>
</table>
**Comparison of some methods** *(Novotny et al, 2004)*

<table>
<thead>
<tr>
<th>Program</th>
<th>Mainly α (19)a</th>
<th>Mainly β (19)a</th>
<th>Mixed α-β (15)a</th>
<th>Few SSEs (8)a</th>
<th>Overall (%)b</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>17</td>
<td>19</td>
<td>13</td>
<td>8</td>
<td>93</td>
</tr>
<tr>
<td>DALI</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>8</td>
<td>90</td>
</tr>
<tr>
<td>DEJAVU</td>
<td>14</td>
<td>19</td>
<td>9</td>
<td>4</td>
<td>75</td>
</tr>
<tr>
<td>LOCK</td>
<td>0</td>
<td>14</td>
<td>11</td>
<td>8</td>
<td>54</td>
</tr>
<tr>
<td>MATRAS</td>
<td>11</td>
<td>19</td>
<td>14</td>
<td>8</td>
<td>85</td>
</tr>
<tr>
<td>PRIDE</td>
<td>14</td>
<td>14</td>
<td>7</td>
<td>3</td>
<td>62</td>
</tr>
<tr>
<td>SSM</td>
<td>5</td>
<td>13</td>
<td>10</td>
<td>5</td>
<td>54</td>
</tr>
<tr>
<td>TOP</td>
<td>14</td>
<td>18</td>
<td>12</td>
<td>7</td>
<td>84</td>
</tr>
<tr>
<td>TOPS</td>
<td>2</td>
<td>15</td>
<td>14</td>
<td>7</td>
<td>62</td>
</tr>
<tr>
<td>TOPSCAN</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>70</td>
</tr>
<tr>
<td>VAST</td>
<td>12</td>
<td>17</td>
<td>15</td>
<td>7</td>
<td>87</td>
</tr>
</tbody>
</table>

*a* The number of cases for which at least one true positive (different from the query structure itself) was retrieved are listed. At most, this is equal to the number of members of the structural family (this number is listed in brackets in the column headers).

*b* The overall success rate is defined as the percentage of cases of all four classes in which a true positive (apart from the query structure itself) was retrieved. For instance, for TOPS this is \((100\% \times (2 + 15 + 14 + 7)/(19 + 19 + 15 + 8)) = 62\%\).

### Fold Comparison Programs Tested

<table>
<thead>
<tr>
<th>Program</th>
<th>URL</th>
<th>Database used</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td><a href="http://cl.sdsc.edu/">http://cl.sdsc.edu/</a></td>
<td>Structure representatives</td>
</tr>
<tr>
<td>DALI</td>
<td><a href="http://www2.ebi.ac.uk/dali">http://www2.ebi.ac.uk/dali</a></td>
<td>Default</td>
</tr>
<tr>
<td>DEJAVU</td>
<td><a href="http://portray.bmc.uu.se/dejavu">http://portray.bmc.uu.se/dejavu</a></td>
<td>&lt;100% Sequence identity</td>
</tr>
<tr>
<td>LOCK</td>
<td><a href="http://gene.stanford.edu/LOCK">http://gene.stanford.edu/LOCK</a></td>
<td>Largest database (582 proteins)</td>
</tr>
<tr>
<td>MATRAS</td>
<td><a href="http://bongo.lab.nig.ac.jp/~takawaba/cgi-bin/Matras/LibMatForm.pl.cgi">http://bongo.lab.nig.ac.jp/~takawaba/cgi-bin/Matras/LibMatForm.pl.cgi</a></td>
<td>&lt;40% Sequence identity</td>
</tr>
<tr>
<td>PRIDE</td>
<td><a href="http://hydra.icgeb.trieste.it/pride">http://hydra.icgeb.trieste.it/pride</a></td>
<td>CATH database</td>
</tr>
<tr>
<td>SSM</td>
<td><a href="http://www.ebi.ac.uk/msd-srv/ssm">http://www.ebi.ac.uk/msd-srv/ssm</a></td>
<td>PDB or SCOP</td>
</tr>
<tr>
<td>TOP</td>
<td><a href="http://bioinf1.mbfys.lu.se/TOP">http://bioinf1.mbfys.lu.se/TOP</a></td>
<td>4220 structures in SCOP</td>
</tr>
<tr>
<td>TOPS</td>
<td><a href="http://tops.ebi.ac.uk/tops/compare1.html">http://tops.ebi.ac.uk/tops/compare1.html</a></td>
<td>CATH and SCOP databases</td>
</tr>
<tr>
<td>TOPSCAN</td>
<td><a href="http://www.rubric.rdg.ac.uk/~andrew/bioinf.org/topscan">http://www.rubric.rdg.ac.uk/~andrew/bioinf.org/topscan</a></td>
<td>Probe database</td>
</tr>
</tbody>
</table>
How to Compare Structures?

1. Structure description 1
2. Comparison algorithm
3. Scores

- Feature extraction
- Statistical significance
- Similarity, classification
Components of Structure Alignment

1. Structure Description

- Local geometry
- Side chain contacts
- Geometric hashing
- Distance matrix (Dali, 1993)
- Properties (secondary structure, hydrophobic clusters (Comparer, 1990)
- Secondary structure elements (VAST, 1996)
- Distances of inter & intra aligned fragment pairs (CE, 1998)
- Contact map (Celera, 2004)
- Geometry invariants (Jia et al, 2004)
Components of Structure Alignment

2. Alignment algorithms
   - Monte Carlo (Dali, VAST)
   - Heuristics (CE)
   - Dynamic Programming (CE)
   - Probabilistic

3. Statistical significance
Components of Structure Alignment

2. Alignment algorithms

- Input & output of alignment algorithm

  **Input:** two proteins: \( A = \{a_1, \ldots, a_m\} \) and \( B = \{b_1, \ldots, b_n\} \)

  **Output:** An alignment \( L(A, B) = \{(a_{i_1}, b_{j_1}), \ldots, (a_{i_L}, b_{j_L})\} \), and scores \( i_1 < i_2 < \cdots < i_L, j_1 < j_2 < \cdots < j_L \)

  **Constraints:**

  \[
  \min \text{rmsd: } \quad \text{rmsd} = \min_T \sqrt{T \sum_{k=1}^{L} (a_{i_k} - T b_{j_k})^2 / L} \\
  \max L \\
  \min \text{Gaps: } \quad \text{Gaps} = \sum_{t=1}^{L-1} [(i_{t+1} - i_t - 1) + (j_{t+1} - j_t - 1)]
  \]

- Dynamic programming, Integer programming, Monte Carlo...

3. Statistical significance

- Levitt and Gerstein, PNAS, 1998

- Random Model and CE scoring function (Jia et al, 2004)
Understand one method (CE) in more detail

Basic Approach

• Compare octameric fragments – an aligned fragment pair (AFP)

• Stitch together AFPs

• Find the optimal path through the AFPs

• Optimize the alignment through dynamic programming

• Measure the statistical significance of the alignment
Calculation of distance: (a) $D_{ij}$ for alignment represented by two AFPs $i$ and $j$ from the path; (b) $D_{ii}$ for single AFP $i$ from the path.
Definition of the Alignment Path

\[ p_{i+1}^A = p_i^A + m \] \text{ and } \[ p_{i+1}^B = p_i^B + m \] \hspace{1cm} (1)

or

\[ p_{i+1}^A > p_i^A + m \] \text{ and } \[ p_{i+1}^B = p_i^B + m \] \hspace{1cm} (2)

or

\[ p_{i+1}^A = p_i^A + m \] \text{ and } \[ p_{i+1}^B > p_i^B + m \] \hspace{1cm} (3)

\[ p_{i+1}^A \leq p_i^A + m + G \] \hspace{1cm} (4)

and

\[ p_{i+1}^B \leq p_i^B + m + G \] \hspace{1cm} (5)
Evaluation based upon the following three distance similarity measures

1. Distance calculated from independent set of inter-residue distances where each distance is used only once

2. Full set of inter-residue distances

3. Rmsd from least squares superposition
Evaluation based upon the following three distance similarity measures

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

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

3. Rmsd from least squares superposition - used to select few best fragments
How to Extend the Path?

1. Consider all possible AFPs that extend the path

2. Consider only the best AFP

3. Use some intermediate strategy
How to Extend the Path?

1. Consider all possible AFPs that extend the path
   Computationally expensive

2. Consider only the best AFP
   Works well with the right heuristics

3. Use some intermediate strategy
What Heuristics?

\[ D_{nn} < D_0 \] (8)

\[ \frac{1}{n-1} \sum_{i=0}^{n-1} D_{in} < D_1 \] (9)

\[ \frac{1}{n^2} \sum_{i=0}^{n} \sum_{j=0}^{n} D_{ij} < D_1 \] (10)

Candidate AFPs are based upon 8
The best AFP is based upon 9
The decision to extend or terminate the path is based upon 10

\[ D_0 = 3\text{Å}, \quad D_1 = 4\text{Å} \]
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.
Test Case: Phycocyanin versus Colicin A
Test Case: Phycocyanin versus Colicin A
Cyclin-dependent kinases
Open (purple) Closed (blue)
Pavelitch et al. (1997)
Limitations

- Will not find non-topological alignments (outside the bounds of the dotted lines)
- What are the correct “units” to be compared?
- CE works on chains – domains are the correct units, but definition of the domains is not straightforward
Computation of All x All

- Took 11,748 chain in the PDB (1/98)
- Computed for 1868 representatives
- 24,000 Cray T3E processor hours
- Loaded pairwise alignments into database
2004

- 27,000 proteins ~ 45,000 chains
- $45,000^2/2 \times 30$ seconds = 963 yrs
- Options:
  - Use a redundant set of chains
  - Use parallel architectures

One Criterion for Redundancy

- Remove highly homologous chains;
- The RMSD between two chains is less than 2Å;
- The length difference between two chains is less than 10%;
- The number of gap positions in alignment between two chains is less than 20% of aligned residue positions;
- At least 2/3 of the residue positions in the represented chain are aligned with the representing chain.
Review examples where structure comparison has revealed new biological insights
Example 1

- CE revealed putative Ca++ binding domain in acetylcholine esterase
- Sequence similarity to neuroligins predicts Ca++ binding too – confirmed experimentally
- Members of the a/b hydrolase family bind Ca++ which may be important for heterologous cell associations

Structural similarity between Acetylcholinesterase and Calmodulin found using CE (Tsigelny et al, Prot Sci, 2000, 9:180)
Example 1- cont.

2ACE vs. 1TN4: RMSD = 4.6Å  Z-score = 4.6  LALI = 86 LGAP = 8 Seq. Identity = 3.5%

Other algorithms to author’s best knowledge cannot find this similarity - possibly because of very different size (537 vs. 159) of chains and high RMSD of the match, despite low number of gaps)

Full view – nonaligned parts are in grey
Example 2

Structural and functional analysis of ataxin-2 and ataxin-3.

Albrecht M, Golatta M, Wullner U, Lengauer T.

Max-Planck-Institute for Informatics, Saarbrucken, Germany. maro.albrecht@mpi-sb.mpg.de

Spinocerebellar ataxia types 2 (SCA2) and 3 (SCA3) are autosomal-dominantly inherited, neurodegenerative diseases caused by CAG repeat expansions in the coding regions of the genes encoding ataxin-2 and ataxin-3, respectively. To provide a rationale for further functional experiments, we explored the protein architectures of ataxin-2 and ataxin-3. Using structure-based multiple sequence alignments of homologous proteins, we investigated domains, sequence motifs, and interaction partners. Our analyses focused on presumably functional amino acids and the construction of tertiary structure models of the RNA-binding Lsm domain of ataxin-2 and the deubiquitinating Josephin domain of ataxin-3. We also speculate about distant evolutionary relationships of ubiquitin-binding UIM, GAT, UBA and CUE domains and helical ANTH and UBX domain extensions.

PMID: 15265035 [PubMed - indexed for MEDLINE]
Example 3 - 20 most frequent common subdomains
Example 3 – cont.

Detailed view of alignments for substructure [1], spoiaa, a phosphorylatable component of the system that regulates transcription factor sigm of bacillus subtilis (1AU2: ... residues 11–106) showing specific region of the alignment. The columns represented are: arbitrary reference index; start and end residue number of subdomain structure; start and end residue number of neighbor; index for neighbor in the database; overall length of neighbor chain; and resulting sequence alignment. The sequence is color-coded such that red is alpha helix, blue is beta strand, and yellow is unassigned.
Example 3 – cont.

Distribution of structural similarities in PDB detected by CE

![Graph showing distribution of structural similarities in PDB detected by CE.](image)
Example 4

ERABUTOXIN B (NMR, MINIMIZED AVERAGE STRUCTURE)
GLUTAMINE PHOSPHORIBOSYLPYROPHOSPHATE AMIDOTRANSFERASE

This is a similarity the biological meaning of which still remains undefined.
The Future (also a general rule)

- Gold standards are important
- For structure comparison a human generated alignment standard is important
- Algorithms are then challenged to meet the standard
- Eventually those algorithms highlight problems with the standard
- The cycle continues