Protein Folds and Protein Structure Superposition

Amit P. Singh
Xiaole Liu
Jeff Chang
Thomas D. Wu
March 3, 1999
- 9105 released atomic coordinate entries
- 7733 proteins, peptides, and viruses
- About 4.5 new entries per day
Structural Challenges

• Compare all known structures to each other
• Classify and organize all structures in a biological way
• Find common folding patterns and structural motifs
• Compute evolutionary distances between protein structures
• Study interactions between structures and other molecules (Protein Docking)
• Use known structures to predict structure from sequence (Protein Threading)
• Many more ...
Overview

• Protein folds
  • Classification of protein structures
  • Current databases

• Structure Superposition
  • Existing methods
  • LOCK - hierarchical protein structure superposition
  • 3dSearch - fast secondary structure superposition

• Demonstrations
Classification of Protein Structures

- **Class**
  - Similar secondary structure content
  - All α; all β; α&β; α/β; etc
- **Fold (Architecture)**
  - Major structural similarity
  - SSE’s in similar arrangement
  - globin-like fold, TIM barrel fold
- **Superfamily (Topology)**
  - Probable common ancestry
  - globins + phycocyanin
- **Family**
  - Clear evolutionary relationship
  - Sequence similarity usually > 25%
Classes of Protein Structures

- Mainly $\alpha$
- Mainly $\beta$

$\alpha/\beta$
- Parallel $\beta$ sheets, $\beta-$\$\alpha$-$\beta$ units

$\alpha+\beta$
- Anti-parallel $\beta$ sheets, segregated $\alpha$ and $\beta$ regions
- helices mostly on one side of sheet
Classes of Protein Structures

- Others
  - Multi-domain, membrane and cell surface, small proteins, peptides and fragments, designed proteins
Folds / Architectures

- Mainly $\alpha$
  - Bundle
  - Non-Bundle
- Mainly $\beta$
  - Single sheet
  - Roll
  - Barrel
  - Clam
  - Sandwich
  - Prism
  - 4/6/7/8 Propeller
  - Solenoid

$\alpha/\beta$ and $\alpha+\beta$
- Closed
  - Barrel
- Open
  - Sandwich
  - Clam, ...
eg. The TIM Barrel Fold
A Conceptual Problem ...

Another example: Globin vs Colicin
A Difficult Alignment
Databases of Folds

• **SCOP**
  - Murzin AG, Brenner SE, Hubbard T, Chothia C
  - Structural Classification of Protein Structures
  - Manual assembly by inspection
  - All nodes are annotated (eg. All-alpha, alpha/beta)
  - Structural similarity search using 3dSearch (Singh and Brutlag)

• **CATH**
  - Dr. C.A. Orengo, Dr. A.D. Michie, Dr. S. Jones, Dr. M.B. Swindells, Dr. G. Hutchinson, Dr. A. Martin, Dr. D.T. Jones, Prof. J.M. Thornton
  - Class - Architecture - Topology - Homologous Superfamily
  - Manual classification at Architecture level
  - Automated topology classification using the SSAP algorithm No structural similarity search
Databases of Folds

• FSSP
  • L. L. Holm and C. Sander
  • Fully automated using the DALI algorithm (Holm and Sander)
  • No internal node annotations
  • Structural similarity search using DALI

• Pclass
  • A. Singh, X. Liu, J. Chang, D. Brutlag
  • Fully automated using the LOCK and 3dSearch algorithms
  • All internal nodes automatically annotated with common terms
  • JAVA based classification browser
  • Structural similarity search using 3dSearch
Automating Structure Classification and Fold Detection

• Growth of PDB demands automated techniques for classification and fold detection

• Protein Structure Comparison
  • computing structure based evolutionary distances
  • identifying protein function
  • understanding functional mechanism
  • identifying structurally conserved regions in the protein
  • finding binding sites or other functionally important regions of the protein
Structure Superposition

- Find the transformation matrix that best overlaps the table and the chair
- i.e. Find the transformation matrix that minimizes the root mean square deviation between corresponding points of the table and the chair
- Correspondences:
  - Top of chair to top of table
  - Front of chair to front of table, etc.
Absolute Orientation Algorithm

The key is finding corresponding points between the two structures

Closed-form solution of absolute orientation using unit quaternions

Algorithms for Structure Superposition

- **Distance based methods:**
  - DALI (Holm and Sander): Aligning scalar distance plots
  - STRUCTAL (Gerstein and Levitt): Dynamic programming using pair wise inter-molecular distances
  - SSAP (Orengo and Taylor): Dynamic programming using intra-molecular vector distances
  - MINAREA (Falicov and Cohen): Minimizing soap-bubble surface area

- **Vector based methods:**
  - VAST (Bryant): Graph theory based secondary structure alignment
  - 3dSearch (Singh and Brutlag): Fast secondary structure index lookup

- **Both vector and distance based:**
  - LOCK (Singh and Brutlag): Hierarchically uses both secondary structure vectors and atomic distances
An intra-molecular distance plot for myoglobin
DALI

- Based on aligning 2-D intra-molecular distance matrices
- Computes the best subset of corresponding residues from the two proteins such that the similarity between the 2-D distance matrices is maximized
- Searches through all possible alignments of residues using Monte-Carlo and branch-and-bound algorithms

\[
\text{Score}(i, j) = 1.5 - |\text{distance}^A(i, j) - \text{distance}^B(i, j)|
\]
**STRUCTAL**

- Based on Iterative Dynamic Programming to align inter-molecular distances
- Pair-wise alignment score in each square of the matrix is inversely proportional to distance between the two atoms
VAST - Vector Alignment Search Tool

- Aligns only secondary structure elements (SSE)
- Represents each SSE as a vector
- Finds all possible pairs of vectors from the two structures that are similar
- Uses a graph theory algorithm to find maximal subset of similar vectors
- Overall alignment score is based on the number of similar pairs of vectors between the two structures
Algorithms for Structure Superposition

• **Distance based methods:**
  - DALI (Holm and Sander): Aligning scalar distance plots
  - STRUCTAL (Gerstein and Levitt): Dynamic programming using pair wise inter-molecular distances
  - SSAP (Orengo and Taylor): Dynamic programming using intra-molecular vector distances
  - MINAREA (Falicov and Cohen): Minimizing soap-bubble surface area

• **Vector based methods:**
  - VAST (Bryant): Graph theory based secondary structure alignment
  - 3dSearch (Singh and Brutlag): Fast secondary structure index lookup

• **Both**
  - LOCK (Singh and Brutlag): Hierarchically uses both secondary structure vectors and atomic distances
Three Step Algorithm

• Local Secondary Structure Superposition
  • Find an initial superposition of the two proteins by using dynamic programming to align the secondary structure vectors

• Atomic Superposition
  • Apply a greedy nearest neighbor method to minimize the RMSD between the C-α atoms from query and the target (i.e. find the nearest local minimum in the alignment space)

• Core Superposition
  • Find the best sequential core of aligned C-α atoms and minimize the RMSD between them
Step 1: Local Secondary Structure Superposition
Step 1: Local Secondary Structure Superposition
Step 2: Atomic Superposition
Step 3: Core Superposition
Features

• Able to detect both global structural similarity as well as sub-domain similarity
• Fast: $O(n^3)$, where $n =$ number of secondary structure elements
• ~4 min to compare myoglobin (153 residues) to 800 representative proteins from the PDB (on a MIPS R10000, 195 MHz processor)
• Can align flexible proteins that contain multiple sub-domains
<table>
<thead>
<tr>
<th>Code</th>
<th>RMSD</th>
<th>Chain Length</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>myh-A</td>
<td>0.56</td>
<td>153</td>
<td>MYOGLOBIN (AQUOMET, PH 7.1) MUTANT</td>
</tr>
<tr>
<td>dHB-A</td>
<td>1.33</td>
<td>132</td>
<td>HEMOGLOBIN (HORSE, DEOXY)</td>
</tr>
<tr>
<td>eca</td>
<td>1.49</td>
<td>130</td>
<td>HEMOGLOBIN (ERYTHROCRUORIN, AQUOMET)</td>
</tr>
<tr>
<td>flp</td>
<td>1.34</td>
<td>129</td>
<td>HEMOGLOBIN I (MONOMERIC) (FERRIC)</td>
</tr>
<tr>
<td>lhb</td>
<td>1.08</td>
<td>128</td>
<td>HEMOGLOBIN V (CYANO, MET)</td>
</tr>
<tr>
<td>hds-B</td>
<td>1.23</td>
<td>128</td>
<td>HEMOGLOBIN (SICKLE CELL)</td>
</tr>
<tr>
<td>ith-A</td>
<td>1.30</td>
<td>126</td>
<td>HEMOGLOBIN (CYANOMET)</td>
</tr>
<tr>
<td>mba</td>
<td>1.50</td>
<td>126</td>
<td>HEMOGLOBIN (MET) ($PH 7.0)</td>
</tr>
<tr>
<td>hbg</td>
<td>1.41</td>
<td>125</td>
<td>HEMOGLOBIN (CARBON MONOXY)</td>
</tr>
<tr>
<td>ash</td>
<td>1.43</td>
<td>122</td>
<td>HEMOGLOBIN (DOMAIN ONE)</td>
</tr>
<tr>
<td>hlB</td>
<td>1.47</td>
<td>121</td>
<td>HEMOGLOBIN (SEA CUCUMBER)</td>
</tr>
<tr>
<td>hbi-A</td>
<td>1.40</td>
<td>118</td>
<td>HEMOGLOBIN I (OXYGENATED, HOMODIMER)</td>
</tr>
<tr>
<td>gdi</td>
<td>1.47</td>
<td>118</td>
<td>LEGHEMOGLOBIN (CARBON MONOXY)</td>
</tr>
<tr>
<td>cpc-A</td>
<td>1.78</td>
<td>80</td>
<td>C-PHYCOCYANIN</td>
</tr>
<tr>
<td>cpc-L</td>
<td>1.68</td>
<td>74</td>
<td>C-PHYCOCYANIN</td>
</tr>
<tr>
<td>tox-A</td>
<td>1.64</td>
<td>63</td>
<td>DIPTHERIA TOXIN DIMER COMPLEXED WITH NAD</td>
</tr>
<tr>
<td>pbg-A</td>
<td>1.70</td>
<td>60</td>
<td>MOL_ID: 1;</td>
</tr>
<tr>
<td>col-A</td>
<td>1.79</td>
<td>58</td>
<td>COLICIN *A (C-TERMINAL DOMAIN)</td>
</tr>
<tr>
<td>sbL-B</td>
<td>1.51</td>
<td>53</td>
<td>LIPOTROXOGENASE-1 (SOYBEAN)</td>
</tr>
<tr>
<td>oxa</td>
<td>1.85</td>
<td>52</td>
<td>CYTOCHROME P450 (DONOR:O2 OXIDOREDUCTASE)</td>
</tr>
<tr>
<td>krb-C</td>
<td>1.68</td>
<td>49</td>
<td>MOL_ID: 1;</td>
</tr>
<tr>
<td>irk</td>
<td>1.40</td>
<td>48</td>
<td>INSULIN RECEPTOR (TYROSINE KINASE DOMAIN)</td>
</tr>
<tr>
<td>hpd-A</td>
<td>1.61</td>
<td>48</td>
<td>CYTOCHROME P450 (BM-3)</td>
</tr>
<tr>
<td>le2</td>
<td>1.62</td>
<td>48</td>
<td>APOLIPOPROTEIN-*E2</td>
</tr>
<tr>
<td>cp4</td>
<td>1.90</td>
<td>47</td>
<td>CYTOCHROME P450CAM (CAMPHOR MONOOXYGENASE)</td>
</tr>
</tbody>
</table>
## PDB Search: Helix-Turn-Helix

<table>
<thead>
<tr>
<th>Chain</th>
<th>Score</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1lmb-4</td>
<td>0.207748</td>
<td>22</td>
<td>DNA-BINDING REGULATORY PROTEIN</td>
</tr>
<tr>
<td>1pra</td>
<td>0.451552</td>
<td>22</td>
<td>GENE REGULATING PROTEIN</td>
</tr>
<tr>
<td>1adr</td>
<td>0.472226</td>
<td>22</td>
<td>TRANSCRIPTION REGULATION</td>
</tr>
<tr>
<td>1yrn-A</td>
<td>0.492206</td>
<td>22</td>
<td>COMPLEX (TWO DNA-BINDING PROTEINS/DNA)</td>
</tr>
<tr>
<td>1dik</td>
<td>0.517499</td>
<td>22</td>
<td>PHOSPHOTRANSFERASE</td>
</tr>
<tr>
<td>1pnrc-A</td>
<td>0.613499</td>
<td>22</td>
<td>COMPLEX (DNA-BINDING REGULATION/DNA)</td>
</tr>
<tr>
<td>4fis-A</td>
<td>0.677586</td>
<td>22</td>
<td>DNA-BINDING PROTEIN</td>
</tr>
<tr>
<td>1oct-C</td>
<td>0.697683</td>
<td>22</td>
<td>DNA-BINDING PROTEIN</td>
</tr>
<tr>
<td>1ftt</td>
<td>0.742871</td>
<td>22</td>
<td>DNA-BINDING PROTEIN</td>
</tr>
<tr>
<td>1cop-E</td>
<td>0.768256</td>
<td>22</td>
<td>GENE REGULATING PROTEIN</td>
</tr>
<tr>
<td>1dtr</td>
<td>0.780090</td>
<td>22</td>
<td>DNA-BINDING PROTEIN</td>
</tr>
<tr>
<td>1gdt-A</td>
<td>0.795171</td>
<td>22</td>
<td>COMPLEX (SITE-SPECIFIC RECOMBINASE/DNA)</td>
</tr>
<tr>
<td>3gap-B</td>
<td>0.817812</td>
<td>22</td>
<td>GENE REGULATORY PROTEIN</td>
</tr>
<tr>
<td>1lfb</td>
<td>0.820489</td>
<td>22</td>
<td>TRANSCRIPTION REGULATION</td>
</tr>
<tr>
<td>1ads</td>
<td>0.920977</td>
<td>22</td>
<td>OXIDOREDUCTASE</td>
</tr>
<tr>
<td>1pdsn-C</td>
<td>0.949104</td>
<td>22</td>
<td>COMPLEX (GENE REGULATING PROTEIN/DNA)</td>
</tr>
<tr>
<td>1hom</td>
<td>0.958631</td>
<td>22</td>
<td>DNA-BINDING PROTEIN</td>
</tr>
<tr>
<td>1mse-C</td>
<td>0.965817</td>
<td>22</td>
<td>COMPLEX (BINDING PROTEIN/DNA)</td>
</tr>
<tr>
<td>1cma-A</td>
<td>1.023169</td>
<td>22</td>
<td>DNA-BINDING REGULATORY PROTEIN</td>
</tr>
<tr>
<td>1trr-A</td>
<td>1.048990</td>
<td>22</td>
<td>DNA-BINDING REGULATORY PROTEIN</td>
</tr>
<tr>
<td>1tpl-B</td>
<td>1.467803</td>
<td>22</td>
<td>LYASE (CARBON-CARBON)</td>
</tr>
<tr>
<td>1hcr-A</td>
<td>0.930009</td>
<td>21</td>
<td>DNA-BINDING</td>
</tr>
<tr>
<td>1neq</td>
<td>0.949148</td>
<td>21</td>
<td>DNA-BINDING PROTEIN</td>
</tr>
<tr>
<td>1ade-A</td>
<td>0.979226</td>
<td>21</td>
<td>LIGASE (SYNTHETASE)</td>
</tr>
<tr>
<td>153l</td>
<td>1.093427</td>
<td>21</td>
<td>HYDROLASE (O-GLYCOSYL)</td>
</tr>
</tbody>
</table>
Algorithms for Structure Superposition

• **Distance based methods:**
  - DALI (Holm and Sander): Aligning scalar distance plots
  - STRUCTAL (Gerstein and Levitt): Dynamic programming using pair wise inter-molecular distances
  - SSAP (Orengo and Taylor): Dynamic programming using intra-molecular vector distances
  - MINAREA (Falicov and Cohen): Minimizing soap-bubble surface area

• **Vector based methods:**
  - VAST (Bryant): Graph theory based secondary structure alignment
  - 3dSearch (Singh and Brutlag): Fast secondary structure index lookup

• **Both**
  - LOCK (Singh and Brutlag): Hierarchically uses both secondary structure vectors and atomic distances
3dSearch

• Place vectors of all target proteins into an index table
• Structure comparison = Index lookup
• Building the table:
  • For each pair of vectors in the target protein, generate a unique coordinate system
  • Record the position and orientation of all remaining vectors in this coordinate system in an index table
• Searching the table:
  • For each pair of vectors in the query protein, generate a unique coordinate system
  • Search the table of target vectors to find the query coordinate system that results in the maximum number of aligned vectors
Generating a Coordinate System

- For each pair of vectors \((V_i, V_j)\)
  - Transform the coordinates such that \(V_i\) lies on the Z-axis and the projection of \(V_j\) onto the X-Y plane is parallel to the Y-axis
Building the Index Table

• For each coordinate system
  • Compute the position and orientation of the remaining vectors in this new frame of reference
## 3dSearch vs. LOCK

| 6 | lmba | lcpc-B | 152 | 0.445 | MYOGLOBIN (CARBONMONOXY) |
| 6 | leca | leca   | 131 | 1.512 | OXYGEN TRANSPORT         |
| 5 | lmcy | lflp   | 128 | 1.349 | OXYGEN TRANSPORT         |
| 5 | lhb-A | lbab-A | 127 | 1.276 | OXYGEN TRANSPORT         |
| 5 | lgdi | lmba   | 124 | 1.537 | OXYGEN STORAGE           |
| 5 | lflp | lhb   | 124 | 1.418 | OXYGEN TRANSPORT         |
| 5 | lbab-A | lhb-A | 116 | 1.422 | OXYGEN TRANSPORT         |
| 4 | lhb | lgdi   | 110 | 1.644 | OXYGEN TRANSPORT         |
| 2 | lezm | lcpc-B | 74  | 1.709 | LIGHT HARVESTING PROTEIN |
| 2 | ldm | lpha   | 46  | 1.902 | OXIDOREDUCTASE (OXYGENASE) |
| 2 | lcpc-B | lcs | 46  | 1.856 | LYASE (OXO-ACID)         |
| 2 | 2pia | ilk    | 43  | 1.721 | CYTOKINE                 |
| 2 | 2baa | lhp    | 43  | 1.623 | HYDROLASE                |
| 2 | 2abk | lhyp   | 42  | 1.767 | HYDROPHOBIC SEED PROTEIN |
| 2 | ludg | lasu   | 42  | 1.438 | DNA INTEGRATION          |
| 2 | 1tpf-A | ltml | 41  | 1.538 | BETA-AMYLASE             |
| 2 | ltag | lgs    | 41  | 1.714 | HYDROLASE (O-GLYCOSYL)   |
| 2 | lptf | 2baa   | 40  | 1.651 | HYDROLASE (O-GLYCOSYL)   |
| 2 | 126l | 2cyp   | 39  | 1.775 | OXIDOREDUCTASE (H2O2(A)) |
| 1 | lnr | ltag   | 39  | 1.689 | GTP-BINDING PROTEIN      |
| 1 | lmrj | lbbh-A | 38  | 1.511 | ELECTRON TRANSPORT       |
| 1 | lmm | lezm   | 37  | 1.913 | HYDROLASE                |
| 1 | lmla | 2mhr   | 36  | 1.594 | OXYGEN BINDING           |
| 1 | lst | lrpo   | 36  | 1.438 | TRANSCRIPTION REGULATION |
| 1 | 1mb-4 | lpoa | 36  | 1.711 | HYDROLASE                |
Demonstrations

• Fold Databases
  • SCOP (http://scop.stanford.edu/scop)
  • FSSP (http://www2.ebi.ac.uk/dali/fssp/fssp.html)
  • PClass (http://gene.stanford.edu/PClass)

• Structural Alignment Tools
  • LOCK (http://gene.stanford.edu/lock)
  • 3dSearch (http://gene.stanford.edu/3dSearch/)
  • DALI (http://www2.ebi.ac.uk/dali)
Evaluation

- LOCK and 3dSearch compared to
  - DALI, Structal, VAST, MinArea
- 3 distinct query structures
  - Globin, TIM barrel, Immunoglobulin fragment
- Targets obtained from a representative set of the PDB
  - 685 structures (using PDB-Select [Hobohm and Sander])
- Gold standard was the manually created SCOP database
  - True positives were those that belonged to the same SCOP FOLD as the query structure
- Results plotted on ROC like curves by counting number of TP and FP above each position in sorted list
Evaluation

ROC curve for myoglobin (1mbd)

Number of True Positives

Number of False Positives

DALI
STRUCTAL
VAST
MINAREA
LOCK
3dSEARCH

ROC curve for myoglobin (1mbd)
Evaluation

ROC curve for TIM (1tph-2)

Number of False Positives vs Number of True Positives

- DALI
- STRUCTAL
- VAST
- MINAREA
- LOCK
- 3dSEARCH
Evaluation

ROC curve for immunoglobulin fragment (8fabA)
### Evaluation

Execution times (hrs:min:sec) for aligning each query structure to 685 target structures

<table>
<thead>
<tr>
<th>Query</th>
<th>Number of Residues</th>
<th>DALI</th>
<th>STRUCTAL</th>
<th>MINAREA</th>
<th>LOCK</th>
<th>3dSEARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1mbd</td>
<td>153</td>
<td>1:0:0</td>
<td>2:38:55</td>
<td>0:57:09</td>
<td>0:06:28</td>
<td>0:0:04</td>
</tr>
<tr>
<td>1tph-2</td>
<td>245</td>
<td>4:0:0</td>
<td>3:18:05</td>
<td>1:17:07</td>
<td>0:13:05</td>
<td>0:0:27</td>
</tr>
<tr>
<td>8fab-A</td>
<td>103</td>
<td>1:0:0</td>
<td>1:50:16</td>
<td>0:45:44</td>
<td>0:06:54</td>
<td>0:0:07</td>
</tr>
</tbody>
</table>

- LOCK and DALI demonstrate best sensitivity and specificity
- LOCK is at least an order of magnitude faster than DALI