BIOINFORMATICS Structures

Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a
Contents: Structures

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - Distance Matrix based methods
  - Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
Molecular Biology Information:
Macromolecular Structure

- DNA/RNA/Protein
  - Almost all protein
  (RNA Adapted From D Soll Web Page,
  Right Hand Top Protein from M Levitt web page)
Molecular Biology Information: Protein Structure Details

- Statistics on Number of XYZ triplets
  ◊ 200 residues/domain → 200 CA atoms, separated by 3.8 Å
  ◊ Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic Å
  ◊ => ~1500 xyz triplets (=8x200) per protein domain
  ◊ 10 K known domain, ~300 folds

ATOM 1  C  ACE  0  9.401  30.166  60.595  1.00  49.88  1GKY 67
ATOM 2  O  ACE  0 10.432  30.832  60.722  1.00  50.35  1GKY 68
ATOM 3  CH3  ACE  0  8.876  29.767  59.226  1.00  50.04  1GKY 69
ATOM 4  N  SER  1  8.753  29.755  61.685  1.00  49.13  1GKY 70
ATOM 5  CA  SER  1  9.242  30.200  62.974  1.00  46.62  1GKY 71
ATOM 6  C  SER  1 10.453  29.500  63.579  1.00  41.99  1GKY 72
ATOM 7  O  SER  1 10.593  29.607  64.814  1.00  43.24  1GKY 73
ATOM 8  CB  SER  1  8.052  30.189  63.974  1.00  53.00  1GKY 74
ATOM 9  OG  SER  1  7.294  31.409  63.930  1.00  57.79  1GKY 75
ATOM 10 N  ARG  2 11.360  28.819  62.827  1.00  36.48  1GKY 76
ATOM 11 CA  ARG  2 12.548  28.316  63.532  1.00  30.20  1GKY 77
ATOM 12 C  ARG  2 13.502  29.501  63.500  1.00  25.54  1GKY 78

ATOM 1444 CB  LYS 186 13.836  22.263  57.567  1.00  55.06  1GKY1510
ATOM 1445 CG  LYS 186 12.422  22.452  58.180  1.00  53.45  1GKY1511
ATOM 1446 CD  LYS 186 11.531  21.198  58.185  1.00  49.88  1GKY1512
ATOM 1447 CE  LYS 186 11.452  20.402  56.860  1.00  48.15  1GKY1513
ATOM 1448 NZ  LYS 186 10.735  21.104  55.811  1.00  48.41  1GKY1514
ATOM 1449 NZT LYS 186 16.887  23.841  56.647  1.00  62.94  1GKY1515
TER 1450 LYS 186

3.8 Å
Other Aspects of Structure, Besides just Comparing Atom Positions

Atom Position, XYZ triplets

Lines, Axes, Angles

Surfaces, Volumes
What is Protein Geometry?

- Coordinates (X, Y, Z’s)
- Derivative Concepts
  - Distance, Surface Area, Volume, Cavity, Groove, Axes, Angle, &c
- Relation to
  - Function, Energies (E(x)), Dynamics (dx/dt)
Depicting Protein Structure: Myoglobin, Whale and Sperm
Structure alignment - Method

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  ◊ Distance Matrix based methods
  ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surface I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
Structural Alignment of Two Globins
Automatic Alignment to Build Fold Library

Alignment of Individual Structures

Fusing into a Single Fold “Template”

Hb  VLSPADKTNKAAGKGAVGAEYGAELERMFLSFTPFTKTYFPF–ARDS-----HGSAQVKGHKGGKVAADALTNAV
     ||| . ||| . ||| . ||| . ||| .. | |.|| | . | . | | | | | | | .| .| || | || .
Mb  VLSEGEWWLHVAKEADVAGHGDILIRLFKSHPELEKDFRFKHLKTEAEMKASEDLKKGVTNLALTGAIL

Hb  AHVD–DMNLSALSDLHAKLRVDPVNFKKLLSCLLVLAAHLPAEFTPADVHASLDKFLASVSTLVTSKYR------
     | | . || . . . | . . . . | . . . . | . . . . | . . . . | . . . . | . . . . | .
Mb  KK–KGHHEAEELKPLAQSHATKHKLPIKYLEFISEAIIHLHSLRHPGFADAGAMNKALEGFRKDLAAAYKELGYQG

Elements: Domain definitions; Aligned structures, collecting together Non-homologous Sequences; Core annotation

Previous work: Remington, Matthews ‘80; Taylor, Orengo ‘89, ‘94; Artymiuk, Rice, Willett ‘89; Sali, Blundell, ‘90; Vriend, Sander ‘91; Russell, Barton ‘92; Holm, Sander ‘93; Godzik, Skolnick ‘94; Gibrat, Madej, Bryant ‘96; Falicov, F Cohen, ‘96; Feng, Sippl ‘96; G Cohen ‘97; Singh & Brutlag, ‘98
Automatically Comparing Protein Structures

- Given
  2 Structures (A & B),
  2 Basic Comparison Operations

  1. Given an alignment optimally **SUPERIMPOSE** A onto B
     - Find Best R & T to move A onto B
  2. **Find an Alignment** between A and B based on their 3D coordinates
RMS Superposition (1)
RMS Superposition (2):
Distance Between an Atom in 2 Structures

\[ d^2_i = (\vec{x}_{A1} - \vec{x}_{B1}) \cdot (\vec{x}_{A1} - \vec{x}_{B1}) \]
RMS Superposition (3):
RMS Distance Between Aligned Atoms in 2 Structures

\[ \text{RMS} = \sqrt{\frac{\sum_{i=1}^{5} (\mathbf{x}_{Ai} - \mathbf{x}_{Bi})^2}{5}} \approx \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5} \]
RMS Superposition (4):
Rigid-Body Rotation and Translation of One Structure (B)

\[ \vec{x}_{B_i} = R(\theta) \vec{x}_{B_i} + \vec{T} \]

\[ \vec{T} = (T_x, T_y, T_z) \quad R(\Theta, \Phi, \Psi) \]

[Diagram showing rotation and translation of points]
RMS Superposition (5):
Optimal Movement of One Structure to Minimize the RMS

Methods of Solution:
- springs ($F \sim kx$)
- SVD
- Kabsch

\[
\text{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_{Ai} - x_{Bi})^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_{Ai} - R(\Theta)x_{Bi} - T)^2}
\]
**Alignment (1)**

**Make a Similarity Matrix**

**(Like Dot Plot)**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>N</th>
<th>Y</th>
<th>R</th>
<th>Q</th>
<th>C</th>
<th>L</th>
<th>C</th>
<th>R</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
**Structural Alignment (1b)**

**Make a Similarity Matrix**

(Generalized Similarity Matrix)

- **PAM(A,V) = 0.5**
  - Applies at every position

- **S(aa @ i, aa @ J)**
  - Specific Matrix for each pair of residues
    - i in protein 1 and J in protein 2
  - Example is Y near N-term. matches any C-term. residue (Y at J=2)

- **S(i,J)**
  - Doesn’t need to depend on a.a. identities at all!
  - Just need to make up a score for matching residue i in protein 1 with residue J in protein 2
Structural Alignment (1c*)

Similarity Matrix

for Structural Alignment

• Structural Alignment
  ◊ Similarity Matrix $S(i, J)$ depends on the 3D coordinates of residues $i$ and $J$
  ◊ Distance between CA of $i$ and $J$
    \[
    d = \sqrt{(x_i - x_J)^2 + (y_i - y_J)^2 + (z_i - z_J)^2}
    \]
  ◊ $M(i,j) = 100 / (5 + d^2)$

• Threading
  ◊ $S(i, J)$ depends on the how well the amino acid at position $i$ in protein 1 fits into the 3D structural environment at position $J$ of protein 2

\[
\alpha \quad \beta \quad \gamma
\]

\[
\begin{array}{cccccc}
100 & & & & & 100 \\
& & & & & 5 + d^2 \\
& & & & & 5 + d^2 \\
\end{array}
\]
Alignment (2): Dynamic Programming, Start Computing the Sum Matrix

\[
\text{new\_value\_cell}(R,C) \leq \\
\text{cell}(R,C) \quad \{\text{Old value, either 1 or 0}\} \\
+ \text{Max}[
\text{cell}(R+1, C+1), \quad \{\text{Diagonally Down, no gaps}\} \\
\text{cells}(R+1, C+2 \text{ to } C_{\text{max}}), \quad \{\text{Down a row, making col. gap}\} \\
\text{cells}(R+2 \text{ to } R_{\text{max}}, C+2) \quad \{\text{Down a col., making row gap}\}
]\]

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>N</th>
<th>Y</th>
<th>R</th>
<th>Q</th>
<th>C</th>
<th>L</th>
<th>C</th>
<th>R</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>N</th>
<th>Y</th>
<th>R</th>
<th>Q</th>
<th>C</th>
<th>L</th>
<th>C</th>
<th>R</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Alignment (3): Dynamic Programming, Keep Going
Alignment (4): Dynamic Programming, Sum Matrix All Done
Alignment (5): Traceback

Find Best Score (8) and Trace Back

A B C N Y - R Q C L C R - P M
A Y C - Y N R - C K C R B P

```
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>N</th>
<th>Y</th>
<th>R</th>
<th>Q</th>
<th>C</th>
<th>L</th>
<th>C</th>
<th>R</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Y</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Y</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>K</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
```
In Structural Alignment, Not Yet Done (Step 6*)

- Use Alignment to LSQ Fit Structure B onto Structure A
  - However, movement of B will now change the Similarity Matrix

- This Violates Fundamental Premise of Dynamic Programming
  - Way Residue at i is aligned can now affect previously optimal alignment of residues (from 1 to i-1)
Structural Alignment (7*),
Iterate Until Convergence

1. Compute Sim. Matrix
3. RMS Fit Based on Alignment
4. Move Structure B
5. Re-compute Sim. Matrix
6. If changed from #1, GOTO #2
Structure alignment - Scoring

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  ◦ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  ◦ Distance Matrix based methods
  ◦ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surface I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
Score \( S \) at End Just Like SW Score, but also have final RMS

\[
S = \text{Total Score} \\
S(i,j) = \text{similarity matrix score for aligning } i \text{ and } j \\
\text{Sum is carried out over all aligned } i \text{ and } j \\
n = \text{number of gaps (assuming no gap ext. penalty)} \\
G = \text{gap penalty}
\]

\[
S = \sum_{i,j} S(i, j) - nG
\]
Some Similarities are Readily Apparent others are more Subtle

Easy: Globins
125 res., ~1.5 Å

Tricky: Ig C & V
85 res., ~3 Å

Very Subtle: G3P-dehydrogenase, C-term. Domain
>5 Å
Some Similarities are Readily Apparent others are more Subtle

**Easy:**
Globins

125 res.,
~1.5 Å

**Tricky:**
Ig C & V

85 res.,
~3 Å

**Very Subtle:** G3P-dehydrogenase, C-term. Domain

>5 Å
• Significance Statistics
  ◊ For sequences, originally used in Blast (Karlin-Altschul). Then in FASTA, &c.
  ◊ Extrapolated Percentile Rank: How does a Score Rank Relative to all Other Scores?

• Our Strategy: Fit to Observed Distribution
  1) All-vs-All comparison
  2) Graph Distribution of Scores in 2D (N dependence); 1K x 1K families -> ~1M scores; ~2K included TPs
  3) Fit a function $\rho(S)$ to TN distribution (TNs from scop); Integrating $\rho$ gives $P(s>S)$, the CDF, chance of getting a score better than threshold $S$ randomly
  4) Use same formalism for sequence & structure

[ e.g. $P(score\ s>392) = 1\% \ chance]$
Statistics on Range of Similarities

For 2107 pairs, only 2% Outliers (with subtle similarity)

Histogram of RMS or RMS' values

For 2107 scop superfamily pairs

RMS or RMS'

Frequency

RMS

For 2107 pairs, only 2% Outliers (with subtle similarity)

Num. Aligned
Scores from Structural Alignment Distributed Just Like Ones from Sequence Alignment (E.V.D.)
Same Results for Sequence & Structure

3 Free Parm. fit to EVD involving: a, b, σ.
These are the only difference betw. sequence and structure.

\[ Z = \frac{S - (a \ln N + b)}{c} \]

\[ S = \sum_{i,j} M(i, j) - G \]

\[ \rho(z) = \exp(-z - e^{-z}) \]

N, G, M also defined differently for sequence and structure.
N = number of residues matched.
G = total gap penalty.
M(i,j) = similarity matrix
(Blossum for seq. or M_{str}(i,j), struc.)
Score Significance (P-value) derived from Extreme Value Distribution (just like BLAST, FASTA)

F(s) = E.V.D of scores
F(s) = exp(-Z(s) - exp(-Z(s)))

Z(s) = As + ln(N) + B
s = Score from random alignment
N length of sequence matched
A & B are fit parameters

P(s>S) = CDF = integral[ F(s) ]
P(s>S) = 1 - exp(-exp(-Z(s)))
Given Score S (1%), P (s > S) is the chance that a given random score s is greater than the threshold
i.e. P-value gives chance score would occur randomly

Exactly like Sequence Matching Statistics (BLAST and FASTA)
RMS is a similarity Score

- Also, RMS doesn’t work instead of structural alignment (no EVD fit)
  - RMS penalizes worst fitting atoms, easily skewed

\[ S_{str} = \sum \frac{100}{5 + d_i^2} \]
\[ \text{RMS} = \sqrt{\sum d_i^2} \]
Structure alignment - Other methods

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - Distance Matrix based methods
  - Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surface I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
Significance Ignoring Crucial Features in Structural Similarity
Other Methods of Structural Alignment

- RMS fitting used universally, but other alignment methods

- Comparison of Distance Matrices
  - Holm & Sander, DALI
  - Taylor & Orengo

Structure Hashing
- Bryant, VAST
- Rice, Artymiuk

Others
- Cohen (Soap)
- Sippl
- Godzik (Lattice)
Fold Library vs. Other Fundamental Data structures

Parts List Database; Statistical, rather than mathematical relationships and conclusions

Folds in Molecular Biology 1000-10000

| Part | 1 | 2 | 3 | 4 | 5 | 6 | ...
|------|---|---|---|---|---|---|---|

- Physics
- Chemistry
- Finance
- Politics

(Large than physics and chemistry, Similar to Finance (Exact Finite Number of Objects (3,056 on NYSE by 1/98), descrip. by Standardized Statistics (even abbrevs, INTC) and groups (sectors)) Smaller than Social Surveys, Indefinite Number of People, Not Well Defined Vocabulary and statistics.
Fold Classifications

- **Scop**
  - Chothia, Murzin (Cambridge)
  - Manual classification, auto-alignments available
  - Evolutionary clusters

- **Cath**
  - Thornton (London)
  - Semi-automatic classification with alignments
  - Class, arch, topo., homol.

- **FSSP**
  - Sander, Holm (Cambridge)
  - Totally automatic with DALI
  - Objective but not always interpretable clusters

- **VAST**
Sequence-structure Relationships

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - Distance Matrix based methods
  - Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surface I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
Adding Structure to Functional Genomics, Function to Structural Genomics

Why Structure? Do we really need it?

1. Most Highly Conserved
2. Precisely Defined Modules
4. Link to Chemistry, Drugs

Purely Seq. Based Analysis -- e.g. EcoCyc, ENZYME, GenProtEC, COGs, MIPS

%ID v. RMS

%ID v. Genomes

Sequence

Genome

Folds v. Genomes

Structure

Drug
Chothia & Lesk, 1986 -- 32 points

EMBO J 4: 823 (1986)
“The relation between the divergence of sequence and structure in proteins”
32 pairs of homologous proteins
RMS, percent identity
Δ = 0.40 e^{1.87H}

Now redo with >16,000 pairs in scop + auto-alignments (pdb95d)....
Chothia and Lesk, revisited
16K points

C&L ‘86:
\[ \Delta = 0.4 \exp(1.9 \ H) \]
Here:
\[ \Delta = 0.2 \exp(1.3 \ H) \]
\[ \Delta = 0.2 \exp(1.9 \ H) \]
Problems with RMS

- Dominated by worst-fitting atoms
- Trimming is arbitrary (50%)
- “Bunching up” between 20% and 0% identity
Structural Comp. Score vs. Smith-Waterman Score

overcomes zero bunching, trimming problem

\[ S_{str} = 100(21 - 11 \exp(-0.0054 \text{ SWS}) \]
Problems with Structural Alignment Score

Different Lengths give different scores. Scores follow equation of the form:

\[ y = An + Mx + B \]
Modern statistical language

\[ P_{str} = 10^{-10} P_{seq}^{0.05} \]
\[ P_{str} = 10^{-6} P_{seq}^{0.274} \]

overcomes length dependency
Focus on Twilight Zone

- **Sequence Sig. without structure signif.**
  - Protein motions
  - Small proteins
  - Low-res, NMR

- **Struc. Sig. without Seq. signif.**
  - More in bottom-right than top-left
Relationship of Similarity in Sequence & Structure to that in Function

![Graphs showing the relationship between sequence similarity and structural similarity.](image)

- Fractional percentage with functional similarity vs. RMS Cα separation (50% trim).
- Fractional percentage with functional similarity vs. log($P_{seq}$).
- Percentage of pairs with same class or function vs. % sequence identity.
### Relationship of Similarity in Sequence & Structure, & Function - Summary

<table>
<thead>
<tr>
<th></th>
<th>Sequence Similarity</th>
<th>Structural Similarity</th>
<th>Features</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional Scores</td>
<td>Percent sequence identity</td>
<td>RMS C$\alpha$ separation</td>
<td>Well understood, in use</td>
<td>RMS depends most highly on worst matches, requiring arbitrary trimming</td>
</tr>
<tr>
<td>Alignment Similarity Scores</td>
<td>$S_{seq}$</td>
<td>$S_{str}$</td>
<td>Analogous similarity scores, $S_{str}$ depends most highly on best matches</td>
<td>Dependence on alignment length</td>
</tr>
<tr>
<td>Modern Probabilistic Scores</td>
<td>$P_{seq}$</td>
<td>$P_{str}$</td>
<td>Statistical significance, unified framework for different comparisons</td>
<td>Not as familiar as RMS and percent identity, some residual length-dependency</td>
</tr>
</tbody>
</table>
Surfaces I

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - Distance Matrix based methods
  - Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
• Why calculate?
  ◊ Protein is solid object. Surface is where action takes place.
  ◊ Surface useful for docking and drug-design
  ◊ Hydrophobic energy proportional to surface area

• Various Types of Protein Surfaces
  ◊ Accessible Surface
  ◊ Molecular Surface
  ◊ Hydration Surface

• Accessible Surface
  ◊ Roll sphere (water) on surface and look at locus of sphere centers.
  ◊ Usually represented as a dot surface
  ◊ Not smooth and continuously differentiable (relevant for energy calculations). It has sharp cusps.
Molecular Surface

- Cusps in the Accessible Surface
- Solution: the smooth molecular surface.
  - M.S. = contact surface + re-entrant surface
  - C.S. = points of tangency between probe sphere and protein when probe sphere is only touching one atom
  - R.S. = solid angle of probe sphere when tangent to two protein atoms
  - First proposed by Richards, but hard to calculate. First numeric calc. by Connely. Later analytic calculation by Connelly.
  - Analytic version is continuously differentiable.
Richards’ Molecular and Accessible Surfaces

<table>
<thead>
<tr>
<th>Probe Radius</th>
<th>Part of Probe Sphere</th>
<th>Type of Surface</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Center (or Tangent)</td>
<td>Van der Waals Surface (vdWS)</td>
</tr>
<tr>
<td>1.4 Å</td>
<td>Center</td>
<td>Solvent Accessible Surface (SAS)</td>
</tr>
<tr>
<td>***</td>
<td>Tangent (1 atom)</td>
<td>Contact Surface (CS, from parts of atoms)</td>
</tr>
<tr>
<td>***</td>
<td>Tangent (2 or 3 atoms)</td>
<td>Reentrant Surface (RS, from parts of Probe)</td>
</tr>
<tr>
<td>***</td>
<td>Tangent (1, 2, or 3 atoms)</td>
<td>Molecular Surface (MS = CS + RS)</td>
</tr>
<tr>
<td>10 Å</td>
<td>Center</td>
<td>A Ligand or Reagent Accessible Surface</td>
</tr>
<tr>
<td>∞</td>
<td>Tangent</td>
<td>Minimum limit of MS (related to convex hull)</td>
</tr>
<tr>
<td>***</td>
<td>Center</td>
<td>Undefined</td>
</tr>
</tbody>
</table>
How to Calculate Accessible Surface Area

- Lee & Richards algorithm (first method, 1970)
  - Pick an arbitrary direction from which to view the protein. Slice it into many sections perpendicular to this direction.
  - In each section, cycle over all the atoms. Each atom is represented as a sphere with a radius that is the sum of its VDW radius plus that of a probe solvent -- i.e. 1.4 for water.

For each atom determine the circle corresponding to the intersection of this sphere with the sectioning plane. Remove all parts (i.e. arcs) of this circle occluded by the circles of other atoms.

- Multiply the total amount of non-occluded arc length by the sectioning width to get the surface area for atom. Sum over all atoms and all sections to get total area.
Shrake & Rupley algorithm (easier)

- Surround each atom with sphere of uniformly spaced dots (e.g. 92).
- Remove dots contained in other atoms spheres. Total number of remaining dots is accessible surface.
Calculation of Volumes

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  ◊ Distance Matrix based methods
  ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
Voronoi Volumes

- Each atom surrounded by a single convex polyhedron and allocated space within it
  - Allocation of all space (large V implies cavities)
- 2 methods of determination
  - Find planes separating atoms, intersection of these is polyhedron
  - Locate vertices, which are equidistant from 4 atoms
Classic Papers


Calculating Volumes with Voronoi polyhedra

- In 1908 Voronoi found a way of partitioning all space amongst a collection of points using specially constructed polyhedra. Here we refer to a collection of "atom centers" rather than "points."

- In 3D, each atom is surrounded by a unique limiting polyhedron such that all points within an atom's polyhedron are closer to this atom than all other atoms.

- Likewise, points equidistant from 2 atoms form planes (lines in 2D). Those equidistant from 3 atoms form lines, and those equidistant form 4 centers form vertices.
Determining Voronoi Volumes

- Integrating on a Grid
  ◊ The simplest method for calculating volumes with Voronoi polyhedra is to put all atoms in the system on a fine grid. Then go to each grid-point (i.e., voxel) and add its infinitesimal volume to the atom center closest to it. This is prohibitively slow for a real protein structure, but it can be made somewhat faster by randomly sampling grid-points. It is, furthermore, a useful approach for high-dimensional integration.

- Solving for the Vertices
  ◊ In the basic Voronoi construction, each atom is surrounded by a unique limiting polyhedron such that all points within an atom’s polyhedron are closer to this atom than all other atoms. Points equidistant from 2 atoms lie on a dividing plane; those equidistant from 3 atoms are on a line, and those equidistant from 4 centers form a vertex.
  ◊ It is straightforward to solve for possible vertex coordinates using the equation of a sphere. (That is, one uses four sets of coordinates \((x,y,z)\) and the equation \((x-a)^2 + (y-b)^2 + (z-c)^2 = r^2\) to solve for the center \((a,b,c)\) and radius \((r)\) of the sphere.) One then checks whether this putative vertex is closer to these four atoms than any other atom; if so, it is a real vertex.
Collecting Vertices and Calculating Volumes

- To systematically collect the vertices associated with an atom, label each one by the indices of the four atoms with which it is associated. To traverse the vertices on one face of a polyhedron, find all vertices that share two indices and thus have two atoms in common — e.g., a central atom (atom 0) and another atom (atom 1). Arbitrarily pick a vertex to start and walk around the perimeter of the face. One can tell which vertices are connected by edges because they will have a third atom in common (in addition to atom 0 and atom 1). This sequential walking procedure also provides a way to draw polyhedra on a graphics device. More importantly, with reference to the starting vertex, the face can be divided into triangles, for which it is trivial to calculate areas and volumes.
Atoms have different sizes

- Difficulty with Voronoi Meth.
  Not all atoms created equal

- Solutions
  - Bisection -- plane midway between atoms
  - Method B (Richards)
    Positions the dividing plane according to ratio
  - Radical Plane

- VDW Radii Set
Complexity from different atom sizes requires new ways to calculate polyhedra.

**Vertex Error**

**Chopping Down Method of Calculating Polyhedra**
Representing Lines, Planes, and Their Intersection

Intersection of Planes and Lines

\[ \begin{align*}
\hat{z} & = \frac{\pi - w}{\sqrt{w^2 + z^2}} \\
\hat{y} & = \frac{\theta z}{\sqrt{z^2 + \theta^2}} \\
\hat{x} & = \frac{-\theta y}{\sqrt{z^2 + \theta^2}} \\
\end{align*} \]
Calculating Areas and Volumes from Vectors
Delauney Triangulation, the Natural Way to Define Packing Neighbors

- Related to Voronoi polyhedra (dual)
- What “coordination number” does an atom have? Doesn’t depend on distance
- alpha shape
- threading
Properties of Voronoi Polyhedra

- If Voronoi polyhedra are constructed around atoms in a periodic system, such as in a crystal, all the volume in the unit cell will be apportioned to the atoms. There will be no gaps or cavities as there would be if one, for instance, simply drew spheres around the atoms.
- Voronoi volume of an atom is a weighted average of distances to all its neighbors, where the weighting factor is the contact area with the neighbor.
Voronoi diagrams are generally useful, beyond proteins

- Border of D.T. is Convex Hull
- D.T. produces "fatest" possible triangles which makes it convenient for things such as finite element analysis.
- Nearest neighbor problems. The nearest neighbor of a query point in center of the Voronoi diagram in which it resides
- Largest empty circle in a collection of points has center at a Voronoi vertex
- Voronoi volume of "something" often is a useful weighting factor. This fact can be used, for instance, to weight sequences in alignment to correct for over or under-representation
Voronoi Volumes & Packing

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  ◊ Distance Matrix based methods
  ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
Voronoi Volumes, the Natural Way to Measure Packing

Packing Efficiency

\[
\frac{\text{Volume-of-Object}}{\text{Space-it-occupies}} = \frac{V(\text{VDW})}{V(\text{Voronoi})}
\]

- Absolute v relative eff.
  \[
  \frac{V_1}{V_2}
  \]
- Other methods
  ◇ Measure Cavity Volume
    (grids, constructions, &c)
Close-Packing of Spheres

- Efficiency
  ◊ Volume Spheres / Volume of space

- Close packed spheres
  ◊ 74% volume filled
  ◊ Coordination of 12
  ◊ Two Ways of laying out

- Fcc
  ◊ cubic close packing
  ◊ ABC layers

- hcp
  ◊ Hexagonally close packed
  ◊ ABABAB
Other Well Known Sphere Arrangements

- Simple cubic packing
  - 8 nbrs
  - 52% efficiency
- bcc cubic packing
  - one sphere sits in middle of 8 others (body-centered)
  - 8 nbrs
  - 68% efficiency
- fcc -> bcc -> simple
  - apx 3/4, 2/3, 1/2
Optimal Packing Finally Proved

After Four Centuries, an Answer

What's the best way to stack a bunch of round objects? The answer, whether they are cannonballs or oranges, seems to be an extension of the familiar pyramid-shaped stack seen in grocery stores everywhere.

**SIMPLE CUBIC LATTICE**

STACKING EFFICIENCY 52%

In this arrangement, the spheres sit directly on top of one another, leaving a space between the spheres that is almost equal to the sphere itself.

Stacking efficiency = volume of the spheres / (volume of the spheres + the space between the spheres)

**FACE-CENTERED CUBIC LATTICE**

STACKING EFFICIENCY 74%

In this more efficient arrangement, the spheres sit off-center, resting within the pocket created by the spheres sitting side-by-side below.

Illustration Credits: Singh, New York Times
More Complex Systems -- what to do?

Water v. Argon
Small Packing Changes Significant

- Exponential dependence
- Bounded within a range of 0.5 (.8 and .3)
- Many observations in standard volumes gives small error about the mean (SD/sqrt(N))
Packing ~ VDW force

- Longer-range isotropic attractive tail provides general cohesion
- Shorter-ranged repulsion determines detailed geometry of interaction
- Billiard Ball model, WCA Theory

Electron Overlap Repulsion
\[ U = \varepsilon \left( \frac{r_0}{r} \right)^{12} \]

Dispersion Attraction
\[ U = -4\varepsilon \left( \frac{r_0}{r} \right)^6 \]
Close-packing is Default

- No tight packing when highly directional interactions (such as H-bonds) need to be satisfied
- Packing spheres (.74), hexagonal
- Water (~.35), “Open” tetrahedral, H-bonds
Standard Radii & Volumes

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  ◊ Distance Matrix based methods
  ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
## Different Sets of Radii

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-CH₃</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>1.87</td>
<td>1.90</td>
<td>1.95</td>
<td>2.13</td>
<td>1.82</td>
<td>1.88</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>-CH₂⁻</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>1.87</td>
<td>1.90</td>
<td>1.90</td>
<td>2.23</td>
<td>1.82</td>
<td>1.88</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>&gt;CH⁻</td>
<td>1.70</td>
<td>2.00</td>
<td>2.00</td>
<td>1.87</td>
<td>1.90</td>
<td>1.85</td>
<td>2.38</td>
<td>1.82</td>
<td>1.88</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>=CH</td>
<td>1.80</td>
<td>1.85</td>
<td>*</td>
<td>1.76</td>
<td>1.70</td>
<td>1.90</td>
<td>2.10</td>
<td>1.74</td>
<td>1.80</td>
<td>1.76</td>
<td></td>
</tr>
<tr>
<td>&gt;C=</td>
<td>1.74</td>
<td>1.80</td>
<td>*</td>
<td>1.70</td>
<td>1.76</td>
<td>1.70</td>
<td>1.80</td>
<td>1.74</td>
<td>1.80</td>
<td>1.61</td>
<td></td>
</tr>
<tr>
<td>-NH₂⁺</td>
<td>-</td>
<td>1.80</td>
<td>2.00</td>
<td>1.50</td>
<td>0.70</td>
<td>1.75</td>
<td>1.68</td>
<td>1.40</td>
<td>1.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-NH₂</td>
<td>1.75</td>
<td>1.80</td>
<td>1.50</td>
<td>-</td>
<td>1.65</td>
<td>1.70</td>
<td>1.68</td>
<td>1.40</td>
<td>1.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;NH</td>
<td>1.65</td>
<td>1.52</td>
<td>1.40</td>
<td>1.70</td>
<td>1.65</td>
<td>1.65</td>
<td>1.75</td>
<td>1.68</td>
<td>1.40</td>
<td>1.64</td>
<td></td>
</tr>
<tr>
<td>=O</td>
<td>1.50</td>
<td>1.80</td>
<td>1.40</td>
<td>1.40</td>
<td>1.60</td>
<td>1.56</td>
<td>1.34</td>
<td>1.38</td>
<td>1.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-OH</td>
<td>-</td>
<td>1.80</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.70</td>
<td>1.54</td>
<td>1.53</td>
<td>1.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-OM</td>
<td>1.80</td>
<td>1.89</td>
<td>1.50</td>
<td>1.40</td>
<td>1.60</td>
<td>1.62</td>
<td>1.34</td>
<td>1.41</td>
<td>1.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-SH</td>
<td>-</td>
<td>1.80</td>
<td>1.85</td>
<td>-</td>
<td>1.85</td>
<td>1.80</td>
<td>1.82</td>
<td>1.56</td>
<td>1.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-S⁻</td>
<td>1.80</td>
<td>-</td>
<td>-</td>
<td>1.80</td>
<td>1.85</td>
<td>1.80</td>
<td>1.90</td>
<td>2.08</td>
<td>1.82</td>
<td>1.56</td>
<td>1.77</td>
</tr>
</tbody>
</table>
## ProtOr Parameter Set

- Consistent Radii, Typing, and Volumes for Packing Calculations

<table>
<thead>
<tr>
<th></th>
<th>Unified Atoms</th>
<th>Residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>atom</td>
<td>radii</td>
<td>volume</td>
</tr>
<tr>
<td>C3H0b</td>
<td>1.61</td>
<td>9.70</td>
</tr>
<tr>
<td>C3H0s</td>
<td>1.61</td>
<td>8.72</td>
</tr>
<tr>
<td>C3H1b</td>
<td>1.76</td>
<td>21.28</td>
</tr>
<tr>
<td>C3H1s</td>
<td>1.76</td>
<td>20.44</td>
</tr>
<tr>
<td>C4H1b</td>
<td>1.88</td>
<td>14.35</td>
</tr>
<tr>
<td>C4H1s</td>
<td>1.88</td>
<td>13.17</td>
</tr>
<tr>
<td>C4H2b</td>
<td>1.88</td>
<td>24.26</td>
</tr>
<tr>
<td>C4H2s</td>
<td>1.88</td>
<td>23.19</td>
</tr>
<tr>
<td>C4H3u</td>
<td>1.88</td>
<td>36.73</td>
</tr>
<tr>
<td>N3H0u</td>
<td>1.64</td>
<td>8.65</td>
</tr>
<tr>
<td>N3H1b</td>
<td>1.64</td>
<td>15.72</td>
</tr>
<tr>
<td>N3H1s</td>
<td>1.64</td>
<td>13.62</td>
</tr>
<tr>
<td>N3H2s</td>
<td>1.64</td>
<td>22.69</td>
</tr>
<tr>
<td>N4H3u</td>
<td>1.64</td>
<td>21.41</td>
</tr>
<tr>
<td>O1H0u</td>
<td>1.42</td>
<td>15.91</td>
</tr>
<tr>
<td>O2H1u</td>
<td>1.46</td>
<td>17.98</td>
</tr>
<tr>
<td>S2H0u</td>
<td>1.77</td>
<td>29.17</td>
</tr>
<tr>
<td>S2H1u</td>
<td>1.77</td>
<td>36.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Calculations

**Volume**

**Affecting Factors**

<table>
<thead>
<tr>
<th>Parameters used in Protein Volume Derivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCOP (87 structures)</td>
</tr>
<tr>
<td>BL+</td>
</tr>
<tr>
<td>BL+ Ratio</td>
</tr>
<tr>
<td>Protein Radii, Tsai et al. (1999)</td>
</tr>
<tr>
<td>Typing with 16 basic types</td>
</tr>
<tr>
<td>Hybrid chemical and numerical</td>
</tr>
</tbody>
</table>

#### Post Calculation Processing

1. Radii Set
2. Atomic Group Typing
3. Atomic Group Selection
   - Plane Positioning Method

#### Derived Volume Types

- Raw 173 Protein Volumes
- Predicted 173 Protein Volumes

#### Set of Protein Structures

- Input Data:
Set of VDW Radii

- Great differences in a sensitive parameter (Radii for carbon 1.87 vs 2.00)
- Complex calculation: minimizing SD, iterative procedure, from protein structures
- Look for common distances in CCD
- **Preliminary Solution**

<table>
<thead>
<tr>
<th>Atom</th>
<th>Bondi</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4___</td>
<td>1.87</td>
<td>1.88</td>
</tr>
<tr>
<td>C3H1</td>
<td>1.76</td>
<td>1.76</td>
</tr>
<tr>
<td>C3H0</td>
<td>1.76</td>
<td>1.61</td>
</tr>
<tr>
<td>O1HO</td>
<td>1.40</td>
<td>1.42</td>
</tr>
<tr>
<td>O2H1</td>
<td>1.40</td>
<td>1.46</td>
</tr>
<tr>
<td>N___</td>
<td>1.65</td>
<td>1.64</td>
</tr>
<tr>
<td>S___</td>
<td>1.85</td>
<td>1.77</td>
</tr>
</tbody>
</table>
Standard Residue Volumes

- Database of many hi-res structures (~100, 2 Å)
- Volumes statistics for buried residues (various selections, resample, &c)
- Standard atomic volumes harder…
  parameter set development...

```
G 64  c 105  T 120  V 139  H 159  M 168  R 194
A  90  C 113  P 124  E 140  L 165  K 170  Y 198
S  94  D 117  N 128  N 150  I 165  F 193  W 233
```
### Standard Core Volumes (Prelim.)

<table>
<thead>
<tr>
<th>Atom Types</th>
<th>Num.</th>
<th>Volume $\text{Å}^3$</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mainchain Atoms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>carbonyl carbon (except G)</td>
<td>C</td>
<td>8361</td>
<td>9.2</td>
</tr>
<tr>
<td>alpha carbon (except G)</td>
<td>CA</td>
<td>7686</td>
<td>13.4</td>
</tr>
<tr>
<td>nitrogen (except P)</td>
<td>N</td>
<td>9042</td>
<td>13.9</td>
</tr>
<tr>
<td>carbonyl oxygen</td>
<td>O</td>
<td>7831</td>
<td>15.8</td>
</tr>
<tr>
<td>Gly C</td>
<td></td>
<td>811</td>
<td>10.2</td>
</tr>
<tr>
<td>Gly CA</td>
<td></td>
<td>522</td>
<td>23.5</td>
</tr>
<tr>
<td>Pro N</td>
<td></td>
<td>334</td>
<td>8.6</td>
</tr>
<tr>
<td><strong>Sidechain atoms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trigonal or aromatic carbon</td>
<td>&gt;C=</td>
<td>3026</td>
<td>10.3</td>
</tr>
<tr>
<td>aromatic CH (H,F,W,Y)</td>
<td>-CH=</td>
<td>4333</td>
<td>21.1</td>
</tr>
<tr>
<td>aliphatic CH</td>
<td>&gt;CH-</td>
<td>3411</td>
<td>14.6</td>
</tr>
<tr>
<td>methylene group</td>
<td>-CH2-</td>
<td>5427</td>
<td>23.7</td>
</tr>
<tr>
<td>methyl group (A,V,L,I)</td>
<td>-CH3</td>
<td>5273</td>
<td>36.7</td>
</tr>
<tr>
<td>hydroxyl oxygen (S,T)</td>
<td>-OH</td>
<td>851</td>
<td>17.2</td>
</tr>
<tr>
<td>carbonyl oxygen (N,Q)</td>
<td>=O</td>
<td>272</td>
<td>16.8</td>
</tr>
<tr>
<td>carboxyl oxygen (D,E)</td>
<td>-O</td>
<td>517</td>
<td>16.0</td>
</tr>
<tr>
<td>2° amine (R,H,W)</td>
<td>-NH-</td>
<td>530</td>
<td>15.6</td>
</tr>
<tr>
<td>1° amine or amide (R,N,Q)</td>
<td>-NH2</td>
<td>355</td>
<td>23.4</td>
</tr>
<tr>
<td>tetrahedral nitrogen (K)</td>
<td>-NH3</td>
<td>31</td>
<td>20.0</td>
</tr>
<tr>
<td>thioether or disulfide (C,M)</td>
<td>-S-</td>
<td>1242</td>
<td>19.3</td>
</tr>
<tr>
<td>sulfhydryl (C)</td>
<td>-SH</td>
<td>67</td>
<td>37.8</td>
</tr>
</tbody>
</table>
Clustering into a set of Atom Types I

- Which atoms are equivalent? How many types valid?
- 18 types, [CNOS][34]H[123][bsu]
Clustering into a set of Atom Types II

- Which atoms are equivalent? How many types valid?
- 18 types, [CNOS][34]H[123][bsu]
- E statistic to tell apart
# Compare Different Structure Sets

<table>
<thead>
<tr>
<th>Structure Set</th>
<th>Number</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>100</td>
<td>1.21</td>
</tr>
<tr>
<td>b</td>
<td>90</td>
<td>1.15</td>
</tr>
<tr>
<td>c</td>
<td>110</td>
<td>1.32</td>
</tr>
<tr>
<td>d</td>
<td>80</td>
<td>1.10</td>
</tr>
</tbody>
</table>

The table above compares different structure sets based on their scores. The higher the score, the more similar the structures. The table includes the number of structures in each set and their respective scores.
Overlap of Volumes of Aromatic C3H1 and Aliphatic C4H2
Surfaces II

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  ◊ Distance Matrix based methods
  ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
Packing at Interfaces

- Voronoi volumes (and D. triangulation) to measure packing
- Tight core packing v. Loose surface packing
- Grooves & ridges: close-packing v. H-bonding
- How packing defines a surface (hydration surface)
- Implications for Motions
Packing defines the “Correct Definition” of the Protein Surface

• Voronoi polyhedra are the *Natural* way to study packing!

• How reasonable is a geometric definition of the surface in light of what we know about packing

• The relationship between
  ◊ accessible surface
  ◊ molecular surface
  ◊ Delauney Triangulation (Convex Hull)
  ◊ polyhedra faces
  ◊ hydration surface
Surface and Volume Definitions Linked
Problem of Protein Surface for Voronoi Construction
Sensitivity of Voronoi Construction to Surface Structure

[Diagram showing different structures and graphs comparing volumes and counts for various molecules (A, C3H1; B, C4H2; C, N3H1; D, O1H0).]
Hydration Surface

- Bring together two helices
  - Unusually low water density in grooves and crevices — especially, as compared to uncharged water
  - Fit line through second shell
Defining Surfaces from Packing: Convex Hull and Layers of Waters
Defining a Surface from the Faces of Voronoi Polyhedra
Accessible Surface as a Time-averaged Water Layer
The Hydration Surface: Trying to Model Real Water
Other Applications of Volumes -- Motions, Docking

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  ◊ Distance Matrix based methods
  ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking
• Intercalating Interface, Knobs into Holes

• Packing is a strong constraint on motions
  ◇ Domain or loop motions have to be fast (~10 ps – 100 ns)
  ◇ Can’t cross big energy barriers involved in repacking an interface

• Not applicable to allosteric motions, which are much slower (~1 ms) and do involve repacking interfaces

Interface Packing and Motions
Packing Based Classification: \textbf{Hinge} \textit{v} Shear

\textbf{Hinge} Mechanism involves absence of steric constraints (continuously maintained interface), esp. at hinge.
Absence of Tight Packing at Hinge

Chain Topology is not important
Docking

- The active site of an enzyme is constituted from a relatively small part of the total volume of an enzyme.
- The active site is three-dimensional and formed from distant parts of the linear amino-acid or nucleic acid sequence.
- Substrates are bound to enzymes by multiple weak interactions.
- Active sites are usually clefts or crevices in the enzyme that maximize interaction with the substrate and exclude water.
- The active site creates an unusual microenvironment that specifically stabilizes the chemical transition state.
- The specificity of substrate binding depends upon the precise arrangements of atoms within the active site.
- The active site can be prearranged (rigid lock and key mechanism) or have a dynamic interaction with the substrate (induced fit mechanism).