Project 1: Angle-Based Pairwise Structural Alignment Using Dynamic Programming

Project Group:
Greg Beckham, Ray Terryn, Richie Hoch
Primary Structure

- Primary Structure: the type and sequence of amino acids along the poly-peptide chain
Secondary Structure

- Torsions mandated by minimization of repulsive forces and maximization of attractive forces, until a balance is reached that provides the lowest energy conformation for the protein as a whole.
Secondary Structure
Secondary Structural Sub-units

- β pleated sheet
- Amino acid subunits
- α heli
Tertiary Structure

- **Tertiary Structure**: the overall 3-D arrangement of all atoms in the protein.

- Formed by rearrangement of the subunits based again on attraction and repulsion.

- Responsible for function.
Tertiary Structure

- beta sheet
- disulfide (S-S) bond
- alpha helix
The Zinc Finger: General conformation of the DNA-binding Domain 20-30 amino acids from the “finger" Schematic view showing how four fingers bind to a specific DNA sequence -form “recognition element"
Quaternary Structure

- Quaternary Structure: the 3-D formed by the combination of multiple, separate polypeptide chains into a single complex.

(b) Hemoglobin
3-D Structure Similarities

- Prediction of function
- Realization of evolutionary details
- Structure better preserved than sequence
3D Structure Analysis

- Large amount of 3D protein structures in the Protein Data Bank (PDB)
- Many different techniques to analyze them
- Generally, given two structures, find a rigid motion of one backbone onto the other such that large, contiguous regions of the backbones are matched.
Existing Techniques

- **DALI** – Uses distance matrices from Cα backbones (align sub-matrices)
- **LOCKE** – Represents secondary structures as vectors (local alignment and DP)
- Chew et al. represent Cα backbone as set of points on unit sphere (shift and combine)
- Others:
  - Combinatorial extension, geometric hashing, double dynamic programming
New Technique

- Focuses on representation of backbone as angles between Ca – Ca bonds
- Alignment is independent of locations in space
- Use dynamic programming to align triplet sequences
- Perfect alignment doesn't exist, so align substructures
Program Flow

Pairwise Alignment Program

- Retrieve Ca Locations from Files
- Compute Ca Distance Vectors
- Compute Angles
- Create Angle Sequences
- Global alignment of Angles
- Output Results

C++ vector of triplets
C++ vector of angle triplets
Percentage match

Data Analysis

Filelist.txt
Output.txt

Greg
Raymond
Richie
Our Work

- **Program - Three main sections:**
  - Input processing (Greg)
  - Angle Calculation (Ray)
  - Dynamic Processing (Richie)

- **Results**
  - Similarity scores of proteins
  - Clustering of scores
Program Input

- PDB Example file
- Process multiple files
- Convert to sequences of locations
Vector Calculation

- Vectors from Ca-Ca needed
- PDB format provides locations
- Subtract one position \( a_{i-1} \) from \( a_i \) to achieve vector \( a_{i-1} \)
- All vectors have a magnitude of \( \approx 3.8\text{Å} \)
Compute Local Geometric Representation

- Define vectors along Ca-Cα backbone
- Compute angles \((α_i, β_i, γ_i)\)
ComputeAngles( CaVectors )

for ( i <= 0 to #CaVectors )

\[ a_i \leftarrow \arccos \left( \frac{-C\alpha_{\text{Vector}_{i-1}} \cdot C\alpha_{\text{Vector}_i}}{3.92^2} \right) \]

\[ \beta_i \leftarrow \arccos \left( \frac{-C\alpha_{\text{Vector}_i} \cdot C\alpha_{\text{Vector}_{i+1}}}{3.92^2} \right) \]

\[ \text{normVector}_1 \leftarrow \left[ -C\alpha_{\text{Vector}_{i-1}} \times C\alpha_{\text{Vector}_i} \right] \]

\[ \text{normVector}_2 \leftarrow \left[ -C\alpha_{\text{Vector}_i} \times C\alpha_{\text{Vector}_{i+1}} \right] \]

\[ \text{normVector}_3 \leftarrow \left[ \text{normVector}_1 \times \text{normVector}_2 \right] \]

\[ \gamma_i \leftarrow \arccos \left( \frac{\text{normVector}_1 \cdot \text{normVector}_2}{3.92^2} \right) \]

if ( \text{normVector}_3 \text{ not in same direction as bondVector}_i )

\[ \gamma_i \leftarrow 2\pi - \gamma_i \]

else

\[ \gamma_i \leftarrow \gamma_i \]

\[ \text{angleTriplet}_i \leftarrow a_i, \beta_i, \gamma_i \]

return 0
Algorithm Details

- Zero end gap penalty
- Internal Gap Penalty: $h(k) = a + k \cdot b$
  - $k$=gap length, $a=0.2$, $b=0.2$ (a,b found experimentally)
- Scoring function: $S(A_i, B_i) = K - d(A_i, B_i)$
  - $K=1.4$ (exp.), $d(A_i, B_i)$ = Euclidean distance of $A_i$ to $B_i$
  - $d(A_i, B_i)$ is slightly different than expected

$$d(A_i, B_j) = \left((\alpha_i^A - \alpha_j^B)^2 + (\beta_i^A - \beta_j^B)^2 + g(|\gamma_i^A - \gamma_j^B|)^2\right)^{1/2}$$

$$g(x) = \min(2\pi - x, x)$$
Cluster Results

- Use score from dynamic programming to determine protein similarity
- Can use percentage out of a “Max Score” to predict similar functionality
## Data Matrix

<table>
<thead>
<tr>
<th>2VM6.pdb</th>
<th>2GM.pdb</th>
<th>2DR5.pdb</th>
<th>1HA6.pdb</th>
<th>1PXX.pdb</th>
<th>2D0O.pdb</th>
<th>1ZMO.pdb</th>
<th>2G6Z.pdb</th>
<th>3VVB.pdb</th>
<th>1ITH.pdb</th>
<th>2CPE.pdb</th>
<th>2OXY.pdb</th>
<th>3CXE.pdb</th>
<th>1P2U.pdb</th>
<th>2D0L.pdb</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.36114</td>
<td>0.37559</td>
<td>0.37559</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
<td>0.37643</td>
</tr>
<tr>
<td>0.21819</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
</tr>
<tr>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
<td>0.34217</td>
</tr>
<tr>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
<td>0.33625</td>
</tr>
<tr>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
<td>0.32774</td>
</tr>
<tr>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
<td>0.34964</td>
</tr>
<tr>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
<td>0.31812</td>
</tr>
<tr>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
<td>0.25019</td>
</tr>
<tr>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
<td>0.24611</td>
</tr>
<tr>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
<td>0.21863</td>
</tr>
</tbody>
</table>

### References
- [Data Matrix](https://example.com/data_matrix)
Results Graphed

Z – axis
Percent Match

X and Y axis
Protein Files
2VM6 vs 2Q81

Percentage Match: .56

Green: 2VM6
Red: 2Q81
2VM6 vs 2Q81

**2VM6**

**Molecular Function**
- protein binding

**Biological Process**
- apoptosis
- anti-apoptosis
- regulation of apoptosis
- protein binding
- microtubule binding
- in utero embryonic development
- kidney development
- induction of apoptosis
- cell-matrix adhesion
- activation of pro-apoptotic gene products
- post-embryonic development

**2Q81**

**Molecular Function**
- nucleic acid binding
- DNA binding
- transcription factor activity
- protein binding
- zinc ion binding
- metal ion binding

**Biological Process**
- gastrulation with mouth forming second
- transcription
- regulation of transcription, DNA-dependent
- multicellular organismal development
- ectoderm development
- negative regulation of cell cycle
2E19 vs 3DWB

Percentage Match: .49

Green: 2E19
Red: 3DWB
<table>
<thead>
<tr>
<th>2E19</th>
<th>3DWB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Molecular Function</strong></td>
<td><strong>Molecular Function</strong></td>
</tr>
<tr>
<td>nucleic acid binding</td>
<td>endopeptidase activity</td>
</tr>
<tr>
<td>DNA binding</td>
<td>metalloendopeptidase activity</td>
</tr>
<tr>
<td>transcription factor activity</td>
<td>peptidase activity</td>
</tr>
<tr>
<td>transcription coactivator activity</td>
<td>metallopeptidase activity</td>
</tr>
<tr>
<td>transcription corepressor activity</td>
<td>zinc ion binding</td>
</tr>
<tr>
<td>zinc ion binding</td>
<td>hydrolase activity</td>
</tr>
<tr>
<td><strong>Biological Process</strong></td>
<td><strong>Biological Process</strong></td>
</tr>
<tr>
<td>negative regulation of transcription from RNA polymerase II promoter</td>
<td>response to hypoxia</td>
</tr>
<tr>
<td>transcription</td>
<td>positive regulation of receptor recycling</td>
</tr>
<tr>
<td>regulation of transcription, DNA-dependent</td>
<td>regulation of systemic arterial blood pressure by endothelin</td>
</tr>
<tr>
<td>regulation of transcription from RNA polymerase II promoter</td>
<td>proteolysis</td>
</tr>
<tr>
<td>immune response</td>
<td>apoptosis</td>
</tr>
<tr>
<td></td>
<td>heart development</td>
</tr>
</tbody>
</table>
2D5R vs 1T7H

Percentage Match: 0.01
2D5R vs 1T7H

2D5R
Molecular Function
- nucleic acid binding
- transcription factor activity
- signal transducer activity
- protein binding
- transcription activator activity

Biological Process
- carbohydrate metabolic process
- transcription
- regulation of transcription, DNA-dependent
- signal transduction
- positive regulation of transcription from RNA polymerase II promoter

1T7H
Molecular Function
- receptor binding
- hormone activity
- protein binding
- endothelin A receptor binding
- endothelin B receptor binding

Biological Process
- skeletal system development
- patterning of blood vessels
- response to hypoxia
- in utero embryonic development
- regulation of systemic arterial blood pressure by endothelin
- regulation of pH
- cell surface receptor linked signal transduction