Statistical Potentials Based on Alpha-Shapes

Leo Guibas  Patrice Koehl  Afra Zomorodian
Stanford University
Focus

- Protein Structure Prediction
  - Intellectual exercise
  - Essential to structural genomics

- Classical scheme:
  - generate models for the protein of interest
  - select “best” and hopefully “native-like” model
## Selection Problem

<table>
<thead>
<tr>
<th>Target</th>
<th>Length</th>
<th>Best Decoy</th>
<th></th>
<th>Best Submitted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSD</td>
<td>Length</td>
<td>RMSD</td>
</tr>
<tr>
<td>T087-A</td>
<td>192</td>
<td>5.3</td>
<td>214</td>
<td>6.5</td>
</tr>
<tr>
<td>T087-B</td>
<td>118</td>
<td>4.8</td>
<td>124</td>
<td>6.5</td>
</tr>
<tr>
<td>T091</td>
<td>109</td>
<td>3.1</td>
<td>90</td>
<td>6.1</td>
</tr>
<tr>
<td>T095</td>
<td>244</td>
<td>3.8</td>
<td>178</td>
<td>5.0, 2.9</td>
</tr>
<tr>
<td>T096-B</td>
<td>160</td>
<td>4.9</td>
<td>123</td>
<td>5.7</td>
</tr>
<tr>
<td>T097</td>
<td>105</td>
<td>3.8</td>
<td>100</td>
<td>4.6</td>
</tr>
<tr>
<td>T098</td>
<td>121</td>
<td>4.2</td>
<td>114</td>
<td>3.9</td>
</tr>
<tr>
<td>T102</td>
<td>70</td>
<td>3.2</td>
<td>70</td>
<td>3.56</td>
</tr>
</tbody>
</table>

---

1CASP 4 Experiment: Best results from D. Baker’s Lab
State of the Art

- Good methods for generating decoys
- Not so good at selection
  - very hard problem
  - not well understood
- We need to predict cRMS(D)
State of the Art

- Good methods for generating decoys
- Not so good at selection
  - very hard problem
  - not well understood
- We need to predict cRMS(D)
- This talk: Database-Derived Potentials
Overview

- Potentials
- Alpha-Complex
- Random Databases
- Results
- Discussion
Physical Potentials

- Idea: native has minimum energy
- Problem: No ideal energy function
- VdW
- Electrostatics
- Hydrophobic effects (?)
Database-Derived Potentials

- Idea: native “looks” like a protein
- Problem: need to quantify “looks”
- Physics (Boltzmann law) or Information theory
- Pairwise potential (1979)
- Residue-based (Sippl 1990)
- Atom-based (Samudrala et al. 1997)
Method

- Potential
  - $E \propto - \ln f(x)$
  - $X$: Pair of certain type, e.g. CYS-CYS
  - $Y$: Pair is at distance $r$
  - $\Pr \{X \mid Y\} = \frac{\Pr \{X\} \cdot \Pr \{Y \mid X\}}{\Pr \{Y\}}$
  - $\Pi = - \sum_{i \neq j} \ln (\Pr \{X_{ij} \mid Y_{ij}\})$
Method

• Potential
  ★ $E \propto - \ln f(x)$
  ★ $X$: Pair of certain type, e.g. CYS-CYS
  ★ $Y$: Pair is at distance $r$
  ★ $Pr\{X \mid Y\} = \frac{Pr\{X\} \cdot Pr\{Y \mid X\}}{Pr\{Y\}}$
  ★ $\Pi = - \sum_{i \neq j} \ln(Pr\{X_{ij} \mid Y_{ij}\})$

• Verification
  1. Generate decoys for a known protein
Method

- Potential
  - $E \propto -\ln f(x)$
  - $X$: Pair of certain type, e.g. CYS-CYS
  - $Y$: Pair is at distance $r$
  - $\Pr \{X \mid Y\} = \frac{\Pr \{X\} \cdot \Pr \{Y \mid X\}}{\Pr \{Y\}}$
  - $\Pi = -\sum_{i \neq j} \ln(\Pr \{X_{ij} \mid Y_{ij}\})$

- Verification

  1. Generate decoys for a known protein
  2. Compute cRMS to native and potential
Method

● Potential

★ $E \propto -\ln f(x)$
★ $X$: Pair of certain type, e.g. CYS-CYS
★ $Y$: Pair is at distance $r$
★ $\Pr \{X \mid Y\} = \frac{\Pr \{X\} \cdot \Pr \{Y \mid X\}}{\Pr \{Y\}}$
★ $\Pi = -\sum_{i \neq j} \ln(\Pr \{X_{ij} \mid Y_{ij}\})$

● Verification

1. Generate decoys for a known protein
2. Compute cRMS to native and potential
3. Compare cRMS with potential
- 653 decoys, correlation 0.66²

²[Park & Levitt 1996]
677 decoys, correlation: -0.08
Signal and Noise

CYS - CYS

Pair Distance (Angstroms)

Number of Pairs

0 2 4 6 8 10 12 14 16 18 20

Number of Pairs

0 50 100 150 200 250 300 350
Signal and Noise
Alpha Complex

- van der Waals model
- Captures the topology of the set of ball set
- Subcomplex of the Delaunay complex
Pruning
Pruning
Database

- 2,145 domains from SCOP
- 29,654,812 $C^\alpha$ pairs (edges) with distance $\leq 20\text{Å}$
- $\alpha = 10$ gives 3,643,018 $C^\alpha$ edges
- $\approx 12.3\%$ of possible non-backbone pairs
- 211 types
• 653 decoys, correlation 0.66
653 decoys, correlation 0.79
- 630 decoys, correlation 0.70
630 decoys, correlation 0.56
1ctf – Random 12.3% Database

- 630 decoys, correlation 0.16
Random Trials

- 10 12.3% databases, 10 trials each: average correlation 0.23
## Correlation

<table>
<thead>
<tr>
<th>protein</th>
<th>type</th>
<th># residues</th>
<th># decoys</th>
<th>ap</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ctf</td>
<td>74</td>
<td>a+b</td>
<td>630</td>
<td>0.70</td>
<td>0.56</td>
</tr>
<tr>
<td>1r69</td>
<td>69</td>
<td>a</td>
<td>675</td>
<td>0.31</td>
<td>0.43</td>
</tr>
<tr>
<td>1sn3</td>
<td>65</td>
<td>a</td>
<td>660</td>
<td>-0.04</td>
<td>0.002</td>
</tr>
<tr>
<td>2cro</td>
<td>75</td>
<td>a</td>
<td>674</td>
<td>0.32</td>
<td>0.52</td>
</tr>
<tr>
<td>3icb</td>
<td>75</td>
<td>a</td>
<td>653</td>
<td>0.66</td>
<td>0.79</td>
</tr>
<tr>
<td>4pti</td>
<td>58</td>
<td>small</td>
<td>687</td>
<td>0.18</td>
<td>0.04</td>
</tr>
<tr>
<td>4rxn</td>
<td>54</td>
<td>small</td>
<td>677</td>
<td>-0.08</td>
<td>-0.22</td>
</tr>
</tbody>
</table>
# Improved Selection

<table>
<thead>
<tr>
<th>protein</th>
<th>best cRMS</th>
<th>selected (rank)</th>
<th>native</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ap</td>
<td>α</td>
</tr>
<tr>
<td>1ctf</td>
<td>1.32</td>
<td>3.97 (156)</td>
<td>2.32 (57)</td>
</tr>
<tr>
<td>1r69</td>
<td>0.88</td>
<td>5.35 (410)</td>
<td>4.47 (268)</td>
</tr>
<tr>
<td>1sn3</td>
<td>1.31</td>
<td>6.56 (428)</td>
<td>6.36 (383)</td>
</tr>
<tr>
<td>2cro</td>
<td>0.81</td>
<td>4.19 (246)</td>
<td>1.87 (24)</td>
</tr>
<tr>
<td>3icb</td>
<td>0.94</td>
<td>6.32 (387)</td>
<td>1.45 (16)</td>
</tr>
<tr>
<td>4pti</td>
<td>1.41</td>
<td>6.28 (424)</td>
<td>4.75 (169)</td>
</tr>
<tr>
<td>4rxn</td>
<td>1.36</td>
<td>3.91 (140)</td>
<td>7.01 (606)</td>
</tr>
</tbody>
</table>
Experiments

- Decoy ’R’ Us datasets
- All atom databases
- Backbone databases
- 3 body, 4 body potentials
- CHARMM classification
- Significance measures
Discussion

• No significant progress in 20 years: Geometry?

• Future work:
  ★ Restrict to domains
  ★ Alternate functions and Multivariate regression
  ★ No distance dependence

• Why cRMS?

• Induced Questions: rigidity, local computation