Monte Carlo and Structure Optimization Methods for Biology, Chemistry and Physics

28–30 March, 1999
Supercomputer Computations Research Institute
Florida State University

Contributed talk presented by H. H. Gan
Lattice protein folding with two and four–body statistical potentials

Hin Hark Gan                        Alex Tropsha
New York University    U. North Carolina, Chapel Hill

Tamar Schlick
New York University & Howard Hughes Medical Institute

1. Structure–derived statistical potentials
   – 2–body potentials
     Miyazawa–Jernigan (shifted)
     Hinds–Levitt
   – 4–body potential
     Delaunay tessellation

2. Generating conformational ensembles on lattice
   – chain–growth or scanning method

3. Assessment of predictions
   – RMS, RMS–energy plots,
     distribution functions
   – 4–body score of 2–body ensembles

Supported by NSF, NIH, and HHMI
Monte Carlo

Chain Growth Method


A protein chain is grown segment–by–segment on a lattice. The m–th segment is chosen with a probability

\[ P_m = \exp\left[-\frac{u_{m,k}}{k_BT}\right]/w_m \]

in direction k, where \( u_{m,k} \) is the potential of residue m+1 and

\[ w_m = \sum_k \exp\left(-\frac{u_{m,k}}{k_BT}\right) \]

\( c(m) \) is the number of vacant sites at step m.
**Statistical Average** (canonical ensemble)

\[
<A> = \frac{\sum_{\Lambda} A(\Lambda) W(\Lambda, T)}{\sum_{\Lambda} W(\Lambda, T)}
\]

(configuration \(\Lambda\))

**Statistical weight** (for a protein with \(N+1\) residues)

\[
W(\Lambda, T) = \prod_{m=1}^{N} \left\{ \sum_k \exp(-u_{m,k}/k_B T) \right\}
\]

**Generalized Rosenbluth weight**

\[
W_R(\Lambda, T) = \prod_{m=1}^{N} \left[ 1/P_m(T) \right]
\]

for SAW,

\[
W_R(\Lambda) = \prod_{m=1}^{N} c(m)
\]
Model

A. Lattice

- Cubic lattice (311) with 24 basis vectors (moves):
  \[
  \{(3,1,1),(3,1,-1),\ldots,(-3,-1,-1)\}
  \]
- Lattice spacing, \( L = 1.146 \, \textrm{Å} \)

B. Chain Geometry

allowed bond angle range: \( 68 < \theta < 143 \)

Excluded volume per site = \( (2L)^3 = 12 \, \textrm{Å}^3 \)
Geometry of 311 model

excluded site

zone with repulsive energy
(radius = 4 Å)

$L = 1.146 \text{ Å}$

$b = 3.8 \text{ Å}$
Energy Functions

Miyazawa–Jernigan (MJ) interaction matrix $e(ij)$


- use (shifted) MJ matrix to parameterize the square–well potential $u(ij,R)$

![Diagram](image-url)
D. Four–body statistical potential

A. Tropsha, R.K. Singh, I.I. Vaisman, and W. Zheng, 

Background

1. Voronoi and Delaunay tessellations of disordered structures

2. Delaunay tessellation of a 3D protein structure defines the irregular tetrahedra of four–residue clusters in the structure.
Definition of four-body statistical energy function

\[ e_{ijkl}^{(\alpha)} = -k_B T \ln \left[ \frac{f_{ijkl}^{(\alpha)}}{p_{ijkl}} \right] \]

- \( f_{ijkl}^{(\alpha)} \) – observed frequency of \((ijkl)\) in a set of 309 native proteins
- \( p_{ijkl} \) – expected frequency of \((ijkl)\) i.e. the random, compact reference state

- 5 types of quadruplets: \( \alpha = 0, 1, 2, 3, 4 \)
- use only 6 classes of amino acid residues; all residues in the same class are equivalent
- the potential function has \( 5 \times 126 = 630 \) entries
- potential range, \( 11^\circ \), is chosen to be large in order to avoid low counts
Calbindin 4icb, 76 residues

MJ energy/residue ($k_B T$)

RMSD (Angstrom)
Calbindin D9K (4icb), 76 residues

Native

Lowest RMSD = 5.92 Å

Lowest MJ energy, RMSD = 8.08 Å
Native 4icb
Calbindin 4icb, 76 residues

M J

Hinds-Levitt

4-body
due peptide

Daura et al. J. Mol. Model. (1998) 280 925-

MD simulation with GROMOS96

\[ T = 360 \text{ K} \]
Native 1r69

434 Repressor 1r69, 63 residues

Hinds-Levitt

4-body
434 REPRESSOR (AMINO–TERMINAL DOMAIN) (R1–69)

1r69, 63 residues

Native

Lowest RMSD=5.96 Å

Lowest MJ energy, RMSD=7.53 Å
# RESULTS

RMSD values (Å) of predicted protein structures

<table>
<thead>
<tr>
<th>Protein</th>
<th>Size Class</th>
<th>S–S bonds</th>
<th>RMS for lowest E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MJ</td>
</tr>
<tr>
<td>sini</td>
<td>31</td>
<td>α</td>
<td>7.62</td>
</tr>
<tr>
<td>1fct</td>
<td>32</td>
<td>α</td>
<td>7.18</td>
</tr>
<tr>
<td>1pnt</td>
<td>36</td>
<td>α</td>
<td>6.31</td>
</tr>
<tr>
<td>1r69</td>
<td>63</td>
<td>α</td>
<td>7.53</td>
</tr>
<tr>
<td>2cro</td>
<td>65</td>
<td>α</td>
<td>8.68</td>
</tr>
<tr>
<td>4iecb</td>
<td>76</td>
<td>α</td>
<td>8.08</td>
</tr>
<tr>
<td>1bg8</td>
<td>76</td>
<td>α</td>
<td>10.91</td>
</tr>
<tr>
<td>1apo</td>
<td>42</td>
<td>β</td>
<td>10.03</td>
</tr>
<tr>
<td>1atx</td>
<td>46</td>
<td>β</td>
<td>8.14</td>
</tr>
<tr>
<td>1oma</td>
<td>48</td>
<td>β</td>
<td>9.14</td>
</tr>
<tr>
<td>8rxna</td>
<td>52</td>
<td>β</td>
<td>7.77</td>
</tr>
<tr>
<td>1pn HD</td>
<td>31</td>
<td>α,β</td>
<td>7.36</td>
</tr>
<tr>
<td>7pti</td>
<td>58</td>
<td>α+β</td>
<td>9.72</td>
</tr>
<tr>
<td>2sn3</td>
<td>65</td>
<td>α+β</td>
<td>12.22</td>
</tr>
<tr>
<td>1ctf</td>
<td>68</td>
<td>α/β</td>
<td>9.76</td>
</tr>
<tr>
<td>1ubq</td>
<td>76</td>
<td>α+β</td>
<td>11.81</td>
</tr>
</tbody>
</table>

SW – highest statistical weight, W(Λ,T)
Limitations of statistical potentials


2. Lack of sufficient statistics for multibody potentials.

Concluding remarks

1. A modest goal: to reproduce the path of the native chain (with RMSD ~ 6 Å).

2. Generality of a prediction scheme is demonstrated by consistency of predictions for different protein classes (all α, all β, α+β, and α/β).

3. RMS–energy correlation plot reveals the quality of potential energy function used.