Bioinformatics Methods

Iosif Vaisman

Email: ivaisman@gmu.edu

Protein Modeling Methods

- *Ab initio methods*: solution of a protein folding problem
  search in conformational space
- *Energy-based methods*: energy minimization
  molecular simulation
- *Knowledge-based methods*: homology modeling
  fold recognition

Molecular structure representation

- Elementary particles
- Atoms
- Groups of atoms

Potential Energy Function

\[
\text{PEF}(R) = \sum_{\text{bonds}} K_b (b(R) - b_0)^2 + \sum_{\text{angles}} K_\theta \theta(R) - \theta_0\right)^2 + \\
\sum_{\text{otherbond}} \frac{K_\chi}{2} (1 + \cos[\eta\psi(R) - \gamma]) + \\
\sum_{\text{nonbonded}} \left[ \frac{A_{ij}}{r_{ij}^n} - \frac{B_{ij}}{r_{ij}^m} + \frac{q_i q_j}{\varepsilon_i \varepsilon_j r_{ij}} \right] \quad (1)
\]

Forcefields: AMBER, CHARMM, CVF, ECEPP, GROMOS

Non-Bonded Interactions

- Torsion
- Bond stretching
- Angle bending

Bond length

\[
E = \sum_{\text{bonds}} k_b (r - r_0)^2
\]
**Bond length**

\[ k_b (r - r_0)^2 \]

**Bond angle**

\[ E = \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \]

**Bond angle**

\[ k_\theta (\theta - \theta_0)^2 \]

**Bond length and angle (parameters)**

\[ k = 0.5, 1.0, 2.0 \]

**Torsional angle**

\[ E = \sum_{\text{tortions}} A [1 + \cos(nt - \phi)] \]

**Torsional angle (parameters)**

\[ A = 2.0, n = 2.0, \phi = 0.0^\circ \]
\[ A = 1.0, n = 1.0, \phi = 30.0^\circ \]
\[ A = 1.0, n = 2.0, \phi = 60.0^\circ \]
Non-bonded terms

\[ \varepsilon = \sum \frac{\sigma_{ij}}{\varepsilon_{ij}} + \sum \eta_{ij} \]

van der Waals term

Electrostatic term

Non-bonded terms (parameters)

\[ -\frac{A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} \]

Potential Energy Function

\[
\text{PEF}(R) = \sum_{\text{bonds}} K \{(b(R)-b_0)^2 + \sum_{\text{angles}} K \{(\theta(R)-\theta_0)^2 + \\
\sum_{\text{dihedrals}} K \{(1 + \cos[N(R)-\gamma]) + \\
\sum_{\text{non-bonded}} \left[ \frac{A_{ij}}{r_{ij}^6} - \frac{B_{ij}}{r_{ij}^{12}} + \frac{q_i q_j}{\varepsilon_{ij} r_{ij}(R)} \right] \}
\]

Forcefields: AMBER, CHARMM, CVF, ECEPP, GROMOS

Energy Minimization

Molecular Dynamics

- Model system
- Initial conditions
- Boundary conditions
- Integration algorithm
- Constraints
- Ensemble
- Results
Molecular Dynamics

\[ \vec{F}_i = m_i \vec{a}_i \]

\[ a_i = \frac{d\vec{v}_i}{dt} \]

\[ \vec{v}_i = \frac{d\vec{r}_i}{dt} \]

\[ -\frac{dE}{dr_i} = F_i \]

\[ -\frac{dE}{dr_i} = m_i \frac{d^2r_i}{dt^2} \]

Periodic Boundary Conditions

MD cycle and integration algorithm

1. solve for \( a_i \) at \( t \) using:
   \[ -\frac{dE}{dr_i} = F_i = m_i \dot{a}_i(t) \]
2. update \( \vec{v}_i \) at \( t + \Delta t \) using:
   \[ \vec{v}_i(t + \Delta t) = \vec{v}_i(t - \Delta t) + a_i(t) \Delta t \]
3. update \( \vec{r}_i \) at \( t + \Delta t \) using:
   \[ r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t) \Delta t \]

Time scales

<table>
<thead>
<tr>
<th>Motion</th>
<th>Characteristic time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative vibration of bonded atoms</td>
<td>( 10^{-14} )</td>
</tr>
<tr>
<td>Rotation of side chains at protein surface</td>
<td>( 10^{-11} - 10^{-10} )</td>
</tr>
<tr>
<td>Torsional libration of buried groups</td>
<td>( 10^{-11} - 10^{-9} )</td>
</tr>
<tr>
<td>Relative motion of different globular regions</td>
<td>( 10^{-11} - 10^{-7} )</td>
</tr>
<tr>
<td>Rotation of medium-sized side chains in protein interior</td>
<td>( 10^{-4} - 1 )</td>
</tr>
<tr>
<td>Local denaturation</td>
<td>( 10^5 - 10 )</td>
</tr>
</tbody>
</table>

Temperature in molecular dynamics

\[ U_{kin} = \frac{1}{2} \sum m_i v_i^2 = \frac{3}{2} N k T \]

- \( N \) – number of atoms
- \( k \) – Boltzmann constant
- \( T \) – absolute temperature

MD of proteins: Solvent model

Adopted from YDiggan (1999)
MD of proteins: mobile regions

Snapshots of V_h domain simulation at 300 and 340 K

Adapted from W.F.V an Gunsteren (2001)

MD of proteins: long runs

1 microsecond simulation of villin

Adapted from I.D. Kuntz and P. Kollman (2001)

MD: Reversible folding of peptides

Adapted from W.F.V an Gunsteren (2001)

Structure verification and validation

Bond lengths (Procheck)

<table>
<thead>
<tr>
<th>Bond</th>
<th>labeling</th>
<th>Value</th>
<th>sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-Calpha</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calpha-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calpha-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-Calpha</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calpha-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calpha-C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bond angles (Procheck)

<table>
<thead>
<tr>
<th>Angle</th>
<th>labeling</th>
<th>Value</th>
<th>sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=N-Calpha</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C=O-Calpha</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C=O-NC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C=O-N-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C=O-N-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C=O-N-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C=O-N-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C=O-N-C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Procheck output

a. Ramachandran plot quality - percentage of the protein's residues that are in the core regions of the Ramachandran plot.

b. Peptide bond planarity - standard deviation of the protein structure's omega torsion angles.

c. Bad non-bonded interactions - number of bad contacts per 100 residues.

d. C4 torsional distortion - standard deviation of the C torsion angle (C=O, N-C, and C=O).

e. Main-chain hydrogen bond energy - standard deviation of the hydrogen bond energies for main-chain hydrogen bonds.

f. Overall G-factor - average of different G-factors for each residue in the structure.