LEE/LEER Free Modeling by Global Optimization

InSuk Joung and Jooyoung Lee

http://lee.kias.re.kr
Center for *in-silico* Protein Science
Korea Institute for Advanced Study

Dec. 8, 2014
Iberostar Paraiso, Riviera Maya, Mexico

- Fold Recognition: Sung Jong Lee (U. of Suwon, Korea) & Keehyoung Joo
- Protein 3D Modeling: Keehyoung Joo, InSuk Joung, & Sun Young Lee
- Model Refinement: InSuk Joung & Qianyi Cheng
- Quality Assessment: Sun Young Lee & Balachandran Manavalan
- Database: Jong Yun Kim
- Community Detection and X-ray crystal B-factor: Juyong Lee (NIH, US)
- Others: Jong Young Joung, Seungryong Heo, Mikyung Nam, In-Ho Lee (KRISS, Korea)
1. Weighted network among group leaders
2. Edges are generated according to Google with (# of webpage – 10)
3. Community detection using Mod-CSA (modularity optimization by CSA)
Group leaders’ network of CASP11 from group info web-page

1. Weighted network among group leaders
2. Edges are generated according to Google with (# of webpage – 10)
3. Community detection using Mod-CSA (modularity optimization by CSA)
Protein 3D Modeling

Data-driven modeling (determination):

- Subjective modeling aspects are eliminated as much as possible.
- Only basic stereo-chemical terms such as bond length, bond angle, torsion angle preferences & vdw repulsion are considered.
- Data determine the structure not local properties of proteins.
- Structure determination using X-ray or NMR data
- Contact assisted prediction: T[pscx] targets

Strategy-dependent modeling (prediction)

- Accuracy of Energy/scoring function?
  - Physics / Statistics / Bioinformatics
    - Physical terms: local properties of proteins
    - Statistical terms
    - Template restraints
- Efficiency of conformational search methods?
Data-driven modeling (determination/T[sc])

Sequence

Data & basic local properties → Search Method → 3D model

Strategy-dependent modeling (prediction/T[0px])

Sequence

Energy Function → Search Method → 3D model

*Search Method: Conformational Space Annealing (CSA)

Energy/score function contains:
physics/statistical/bioinformatics terms
Protein Structure Prediction

1. **Physics-based approaches: Principle-based modeling**
   ① **Accurate potential energy function**
   ② **Powerful global optimization method → what we can do better than others**
   ③ **Ab initio, de novo, new fold targets (10-20%)**

2. **Informatics-based approaches: Template-based modeling**
   ① **Map the original problem to a problem with solution → mapping problem (alignment problem)**
   ② **Use templates (problems with solutions) to obtain the solution of the original problem (multiple alignment)**
   ③ **Comparative modeling, fold recognition (80-90%)**
Initially, consider the whole solution space

\[ \downarrow \]

Slowly narrow the search while maintaining diversity of sampling

\[ \downarrow \]

Annealing in the solution space

\[ \downarrow \]

Conformational Space Annealing (CSA)


Flow chart of LEE/LEER

Template Selection

Multiple Alignment

3D Modeling

Side-chain Re-modeling

Model Refinement

Templates are selected from server models.

No MSA is necessary (identical sequences).

Chain is built using CSA and in-house energy function.

RotamerCSA

Restraint-based MD simulation
Template Selection (LEE)

• Network generation:
  – Node: server model
  – Edge: TM-score between 2 nodes
• Each cluster is assessed for ranking
  – Similarity score between nodes
  – QA score of each node
• Model ranking is according to the cluster ranking Community detection score
3D Modeling by Global Optimization

• Energy Function

\[ E = E_{\text{templates}} + E_{\text{stereo-chemistry}} + E_{\text{repulsive}} + E_{\text{DFA}} + E_{\text{dfire}} + E_{\text{goap}} + E_{\text{hbond}} \]

- \( E_{\text{templates}} \) : restraints from templates (Lorentzian shape)
- \( E_{\text{DFA}} \) : dynamic fragment assembly term
- \( E_{\text{dfire}} \) : dfire statistical potential term
- \( E_{\text{goap}} \) : orientation-dependent statistical potential term
- \( E_{\text{hbond}} \) : local hydrogen bonding term

• Global Optimization by Conformational Space Annealing (CSA)
Harmonic vs Lorentzian

\[ E_{\text{harmonic}} = \frac{1}{2} \left( \frac{(r-r_0)^2}{\sigma^2} \right) \]

\[ E_{\text{Lorentzian}} = \frac{1}{\sigma} \frac{(r-r_0)^2}{(r-r_0)^2 + \sigma^2} \]

- \( \sigma \) controls the strength of individual distance restraint.

- Harmonic terms for sure/measured restraint terms
- Lorentzian terms for prediction-based restraint terms
Improved $\sigma$ Prediction

Variability prediction of distance information

$$\sigma_{\text{native}} = \frac{(r_{\text{native}} - r_{\text{template}})}{r_{\text{native}}}$$

Comparison of old and new $\sigma$ of T0536

<table>
<thead>
<tr>
<th>Modeller</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0517</td>
<td>0.291</td>
</tr>
<tr>
<td>T0523</td>
<td>0.020</td>
</tr>
<tr>
<td>T0527</td>
<td>0.182</td>
</tr>
<tr>
<td>T0536</td>
<td>0.197</td>
</tr>
<tr>
<td>T0538</td>
<td>0.136</td>
</tr>
<tr>
<td>T0539</td>
<td>0.261</td>
</tr>
<tr>
<td>T0545</td>
<td>0.200</td>
</tr>
<tr>
<td>T0552</td>
<td>0.105</td>
</tr>
<tr>
<td>T0557</td>
<td>0.298</td>
</tr>
<tr>
<td>T0559</td>
<td>0.147</td>
</tr>
<tr>
<td>T0560</td>
<td>-0.138</td>
</tr>
<tr>
<td>T0566</td>
<td>0.319</td>
</tr>
<tr>
<td>T0567</td>
<td>0.271</td>
</tr>
<tr>
<td>T0580</td>
<td>0.071</td>
</tr>
<tr>
<td>T0586</td>
<td>0.128</td>
</tr>
<tr>
<td>T0590</td>
<td>0.139</td>
</tr>
<tr>
<td>T0592</td>
<td>-0.291</td>
</tr>
<tr>
<td>T0594</td>
<td>0.189</td>
</tr>
<tr>
<td>T0598</td>
<td>0.263</td>
</tr>
<tr>
<td>T0602</td>
<td>0.065</td>
</tr>
<tr>
<td>T0610</td>
<td>0.242</td>
</tr>
<tr>
<td>T0614</td>
<td>-0.083</td>
</tr>
<tr>
<td>T0615</td>
<td>0.329</td>
</tr>
<tr>
<td>T0619</td>
<td>0.061</td>
</tr>
<tr>
<td>T0622</td>
<td>0.021</td>
</tr>
<tr>
<td>T0632</td>
<td>-0.078</td>
</tr>
<tr>
<td>T0634</td>
<td>0.101</td>
</tr>
</tbody>
</table>

Average 0.113 0.352
Side-chain Re-modeling

• Statistics from 3D modeling results:
  – If only one cluster of X1/X2 is identified at the residue level, 10 sets of rotamers are added to the SCWRL4 library.
• Using the modified SCWRL4 library, E is optimized by CSA
  
  \[ E = E_{\text{prob}} + E_{\text{sb}} + E_{\text{ss}} + E_{\text{dfire}} + E_{\text{disulfide}} \]

  – prob: rotamer probability
  – sb: sidechain-backbone clash
  – ss: sidechain-sidechain clash
  – disulfide: disulfide reward
• For surface residues, the regular Scwrl4 is used.

• This finishes the LEE procedure.
LEER

• The input is a LEE model
• Overall process is similar to PRINCETON_TIGRESS
• Prepare restraints:
  • Positional restraints from a given structure (backbone atoms)
  • distance restraints for β-containing parts (CA atoms)
• Run MD simulations while gradually reducing the strength of the positional restraint.
• Final stage trajectories are averaged and further energy-minimized to generate the LEER model.

Force field: AMBER14SB
Solvation: AMBER GB8
LEE vs LEER (FM, all models)

GDT-TS

MolProbity

LEER wins

LEER wins
LEE vs LEER (FM, model1)

GDT-TS

MolProbity

LEER wins

LEER wins
Assessor’s requests

• Two FM targets
  – T0827-D2 (212-369, 158 residues)
  – T0808-D2 (150-418, 269 residues)
T0827-D2 (212-369, 158 residues)

**green:** native  
**blue:** initial  
**red:** model

**TM-score**  
<table>
<thead>
<tr>
<th>Model</th>
<th>TM-score</th>
<th>Molprobity</th>
</tr>
</thead>
<tbody>
<tr>
<td>starting</td>
<td>31.9</td>
<td>3.54</td>
</tr>
<tr>
<td>LEE</td>
<td>51.8</td>
<td>2.09</td>
</tr>
<tr>
<td>LEER</td>
<td>51.8</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Best: MODEL2 by LEE/LEER

Final model (LEE/LEER): 7 helices aligned
T0827-D2 (212-369, 158 residues)

<table>
<thead>
<tr>
<th>Server Models</th>
<th>TM-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang-Server_TS1</td>
<td>19.0</td>
</tr>
<tr>
<td>QUARK_TS1</td>
<td>38.5</td>
</tr>
<tr>
<td>nns_TS1</td>
<td>27.8</td>
</tr>
<tr>
<td>HHPredA_TS1</td>
<td>35.2</td>
</tr>
<tr>
<td>myprotein-me_TS1</td>
<td>29.0</td>
</tr>
<tr>
<td>LEE2/LEER2</td>
<td>51.8</td>
</tr>
</tbody>
</table>
The challenge is to extract max information from multiple templates without the native structure information.
T0808-D2 (150-418, 269 residues)

Best: MODEL4 by LEE/LEER

<table>
<thead>
<tr>
<th>Server Models</th>
<th>TM-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>slbio_TS1</td>
<td>11.0</td>
</tr>
<tr>
<td>slbio_TS3</td>
<td>10.6</td>
</tr>
<tr>
<td>slbio_TS2</td>
<td>10.8</td>
</tr>
<tr>
<td>BAKER-ROSETTASERVER_TS5</td>
<td>11.3</td>
</tr>
<tr>
<td>slbio_TS5</td>
<td>11.8</td>
</tr>
<tr>
<td>slbio_TS4</td>
<td>11.8</td>
</tr>
<tr>
<td>TASSER-VMT_TS2</td>
<td>12.7</td>
</tr>
<tr>
<td>MULTICOM-NOVEL_TS1</td>
<td>9.8</td>
</tr>
<tr>
<td>MULTICOM-NOVEL_TS2</td>
<td>10.6</td>
</tr>
<tr>
<td>nns_TS1</td>
<td>14.0</td>
</tr>
<tr>
<td>HHPredA_TS1</td>
<td>14.3</td>
</tr>
<tr>
<td>LEE4/LEER4</td>
<td>29.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TM-score</th>
<th>Molprobity</th>
</tr>
</thead>
<tbody>
<tr>
<td>starting</td>
<td>19.3</td>
<td>4.20</td>
</tr>
<tr>
<td>LEE</td>
<td>29.6</td>
<td>1.14</td>
</tr>
<tr>
<td>LEER</td>
<td>29.7</td>
<td>1.12</td>
</tr>
</tbody>
</table>
Conclusions

• Energy balance between template terms and ab initio terms is important.
• Proper use of multiple templates of low quality can lead to good modeling.
• Efficient global optimization can help you.
• MD refinement improved physical reality in terms of MolProbity.