Lecture 09

Jens Meiler

VANDERBILT UNIVERSITY
Center for Structural Biology
Departments of Chemistry, Pharmacology, and Biomedical Informatics
Sampling and Scoring for Protein Folding Simulation

- Local Sequence Bias
  - Approximate local interactions using the distribution of conformations seen for similar sequences in known protein structures

- Monte Carlo simulations
  - Select broadest minima using cluster analysis

- Energy evaluation of non-local interactions using knowledge-based energy function
  - Steric overlap
  - Residue environment
  - Pair wise interactions
  - Strand pairing
  - Compactness
  - Secondary Structure Packing

Sampling and Scoring for Side Chain Repacking and Design

Local Rotamer Bias
Approximate interactions within sidechain using the distribution of sidechain conformations (rotamers) seen in known protein structures

Energy function
Statistically derived potential function
- VDW interaction
- solvation
- hydrogen bonding potential
- pair wise interactions
- rotamer probability

Simulated Annealing
Monte Carlo energy minimization

Refinement Cycle with Side Chain Repacking and All Atom Minimization

- random perturbation of one or several backbone torsion angles
- fast side-chain optimization using a rotamer representation
- gradient-based minimization with respect to backbone and side chain torsion angles
Combining Strengths: Building Accurate Models from Sparse Data

- Complete Conformational Space
- Local Sequence Bias
- Energy Evaluation of non-Local Interactions

Protein Structures consistent with sparse experimental data
Protein Structures consistent with rich experimental data
Overview of the Rebuilding-and-Refinement Method.

### Improvement of model accuracy and molecular replacement by a rebuilding and refinement protocol

<table>
<thead>
<tr>
<th>X-ray structure</th>
<th>Starting model*</th>
<th>Length (n)†</th>
<th>Sequence identity to best template (%)‡</th>
<th>GDT-HA§</th>
<th>TFZ¶ in molecular replacement</th>
<th>Auto-traced residues (backbone, side chain)¶§</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Best template</td>
<td>Refined model</td>
</tr>
<tr>
<td>NMR</td>
<td>1hb6</td>
<td>2abd</td>
<td>86</td>
<td>N/A/A</td>
<td>0.58</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>1who</td>
<td>1bmw</td>
<td>94</td>
<td>N/A/A</td>
<td>0.59</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>1gnu</td>
<td>1kot</td>
<td>119</td>
<td>N/A/A</td>
<td>0.64</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>1a19</td>
<td>1ab7</td>
<td>89 (2)</td>
<td>N/A/A</td>
<td>0.63</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>1fvk</td>
<td>1a24</td>
<td>189 (2)</td>
<td>N/A/A</td>
<td>0.49</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>1mzl</td>
<td>1afh</td>
<td>93</td>
<td>N/A/A</td>
<td>0.60</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>1tvg</td>
<td>1xpw</td>
<td>143</td>
<td>N/A/A</td>
<td>0.63</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>2snm</td>
<td>2so b</td>
<td>97</td>
<td>N/A/A</td>
<td>0.45</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>1agm</td>
<td>1ezy</td>
<td>129</td>
<td>N/A/A</td>
<td>0.49</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>1abq</td>
<td>1awo</td>
<td>56</td>
<td>N/A/A</td>
<td>0.58</td>
<td>0.83</td>
</tr>
<tr>
<td>CM</td>
<td>2hhz (T0331)</td>
<td>1ty9A</td>
<td>149</td>
<td>14.5</td>
<td>0.49</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>2hr2 (T0368)</td>
<td>2c21C</td>
<td>158 (6)</td>
<td>14.8</td>
<td>0.57</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>2hq7 (T0380)</td>
<td>2fhqA</td>
<td>145 (2)</td>
<td>25.4</td>
<td>0.58</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>2ib0 (T0385)</td>
<td>1jgcB</td>
<td>170 (2)</td>
<td>7.8</td>
<td>0.62</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>2hio (T0329_D2)</td>
<td>1rqlA</td>
<td>92 (2)</td>
<td>8.8</td>
<td>0.52</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>2hc7 (T0330_D2)</td>
<td>1vhb</td>
<td>75</td>
<td>14.1</td>
<td>0.51</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>2hi6 (T0357)**</td>
<td>1aco</td>
<td>132</td>
<td>8.4</td>
<td>0.45</td>
<td>0.52</td>
</tr>
<tr>
<td>DN</td>
<td>2hh6 (T0283)</td>
<td>2b2j</td>
<td>112</td>
<td>3.6</td>
<td>0.22</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Refinement of NMR Structures (a-d) and Comparative Models (e-h)


native crystal structure (blue), template/NMR structure (red), and the refined model (green)
Phasing by refined NMR structures, comparative and de novo models

- NMR (a\(\rightarrow\)b) / CM (c\(\rightarrow\)d)
- de novo model (c\(\rightarrow\)d)
RosettaNMR: Usage of CSs, NOEs, and RDCs

- NMR data are used in addition to the Local Sequence Bias
  - CS derived dihedral angle restrictions (via TALOS)
  - Local NOE distance restraints
  - RDC orientation restraints

- NMR data are used in addition to energy evaluation of non-local Interactions
  - Long-range NOE distance restraints
  - RDC orientation restraints

RosettaNMR: High-Resolution from “one” Restraint per Amino Acid

- Profilin I with CSs and 1.01 NOEs/AA
- Ubiquitin with 0.89 RDCs/AA
- Profilin I with CSs and 1.25 RDCs/AA

NMR Structure Determination for Proteins Using Backbone-Only Data

Blind Structure Predictions with CS-RDC-Rosetta.

Fig. 3. Blind predictions with the CS-RDC-Rosetta and iterative CS-RDC-Rosetta protocols. (Left side of each panel) Superposition of the 10 lowest-energy predicted structures (red) over the experimentally solved ensemble of NMR structures (blue). (Right side of each panel) Magnified view of the core side chains. Rosetta models in (A) to (D) were determined with CS-RDC-Rosetta, and in (E) with iterative CS-RDC-Rosetta. (A) Bcr268F, (B) Dnu115G, (C) MaR214A, (D) Sr115C, and (E) At7.
Fold Determination and High-Resolution Model Refinement

Sequence
SKLIVPPDEQFTR

ROSETTA de novo fold prediction

NMR

Automated spectra interpretation

Lists with CS-, NOE-, and RDC-information

MONTE CARLO optimization

Optimized assignments

Best Model

Ranking

Improved partial CS-atom assignment

ROSETTA NMR refinement

Refined Model

1ubi_ Refinement 1st Step
1ubi_ Refinement 2\textsuperscript{nd} Step
1ubi_ Refinement 3rd Step

RMSD to native

number of correctly assigned backbone atoms
1ubi_ Refinement 4th Step
1ubi_ Refinement 5th Step

RMSD to native

number of correctly assigned backbone atoms
Backbone RMSD is 0.6Å

All Core Amino Acids have Correct Side Chain Conformation

Structure Elucidation of Lysozyme from 25 Experimental EPR Distances

- 25 distance restraints
- 57 surface exposure restraints

Thanks to Hassane Mchaourab and his lab for experimental data

25 Experimental Distance Restraints

- Restraint: \( (d_{SL-SL} - \sigma_{SL-SL} - 10\text{Å}) \leq d_{CB-CB} \leq (d_{SL-SL} + \sigma_{SL-SL}) \)
- Harmonic penalty function
Influence of Experimental Data on Sampling and Model Quality

- RMSD histogram
- C

![RMSD histogram graph](image)

![Contact Map](image)

6 December 2011 © Jens Meiler
Influence of Experimental Data on Sampling and Model Quality

- RMSD histogram
- C

![RMSD histogram graph](image)

- Model Count
- RMSD in Å

- Plain Rosetta
- With EPR restraints
High Resolution Energy Refinement

- Lowest scoring ~11,000 models out of 500,000 were refined
Backbone RMSD is 0.96Å

All but Two Core Amino Acids have Correct Side Chain Conformation

Maximal sequence separation and unique SSE connections!
Optimized labeling patterns improve de novo protein structure prediction

A

Percentage of Models

25
20
15
10
5
0

RMSD (Å)

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

B

RMSD < 7.5Å

70
60
50
40
30
20
10
0


c

RMSD < 3.5Å

3
2
1
0

Optimized Restraints
Randomized Restraints
No Restraints

6 December 2011 © Jens Meiler
Improvement in model quality requires a limited number of distance restraints.
Adenovirus Protein IIIa – A Novel Topology?

Number of Possible Placements of y Helices in x Density Rods

- Secondary structure prediction with y helices
- EM map with x density rods

\[ n = \max(x, y) \quad \| \quad k = \min(x, y) \]

- Permutations = \( k! \)
- Orientations = \( 2^k \)
- Combinations = \( n! / k! (n - k)! \)
- Total = \( 2^k n! / (n - k)! \) \( \| \) 3 helices and 2 densities: \( 2^2 \cdot 3! / (3 - 2)! = 24 \)
Size of Search Space Grows Exponentially

- About $10^{17}$ possible placements for protein IIIa
- Sample one per second and you are done in $4 \cdot 10^9$ years
- This is about the age of the earth
EM-Fold: “Fold into Density” Protocol

Knowledge-based energy potentials

Primary sequence
MQIFVKLTGKTITLEVEPSDTIENV

Pool of helices and strands

Monte Carlo assembly

Monte Carlo refinement

ROSETTA:
loop and side chain building, atomic resolution refinement

cryoEM density

restraints for moves and scores

Final model

A Density Specific MC Move Set Assembles Protein Fold

- **Model before move**
- **Add**
- **Delete**
- **Flip**
- **Swap**
- **Swap with pool**
- **Move**
A Density-Driven Scoring Function Evaluates Plausibility of Models

- loop score favors models that have native-like loop length
- occupancy score favors models that have all the density rods filled
- connectivity score favors models that connect density rods with strong density in between them
True topology is found in 70% of the benchmark cases

- **Realistic SSE Pool**

<table>
<thead>
<tr>
<th>protein</th>
<th>assembly</th>
<th>refinement</th>
<th>loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1IE9</td>
<td>3.7 Å (1)</td>
<td>3.7 Å (5)</td>
<td>5.9 Å (1)</td>
</tr>
<tr>
<td>1N83</td>
<td>6.2 Å (1)</td>
<td>5.9 Å (2)</td>
<td>7.1 Å (1)</td>
</tr>
<tr>
<td>1OUV</td>
<td>3.0 Å (6)</td>
<td>2.9 Å (4)</td>
<td>4.3 Å (1)</td>
</tr>
<tr>
<td>1QKM</td>
<td>3.6 Å (16)</td>
<td>2.7 Å (2)</td>
<td>3.9 Å (2)</td>
</tr>
<tr>
<td>1TBF</td>
<td>3.1 Å (100)</td>
<td>2.8 Å (20)</td>
<td>4.1 Å (1)</td>
</tr>
<tr>
<td>1V9M</td>
<td>- (-)</td>
<td>- (-)</td>
<td>- (-)</td>
</tr>
<tr>
<td>1XQO</td>
<td>- (-)</td>
<td>- (-)</td>
<td>- (-)</td>
</tr>
<tr>
<td>1Z1L</td>
<td>3.1 Å (150)</td>
<td>3.2 Å (72)</td>
<td>5.9 Å (1)</td>
</tr>
<tr>
<td>2AX6</td>
<td>4.0 Å (1)</td>
<td>3.2 Å (5)</td>
<td>6.6 Å (3)</td>
</tr>
<tr>
<td>2CWC</td>
<td>- (-)</td>
<td>- (-)</td>
<td>- (-)</td>
</tr>
</tbody>
</table>
True topology is found in 70% of the benchmark cases

- **Realistic SSE Pool**

<table>
<thead>
<tr>
<th>protein</th>
<th>assembly</th>
<th>refinement</th>
<th>loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1IE9</td>
<td>3.7 Å (1)</td>
<td>3.7 Å (5)</td>
<td>5.9 Å (1)</td>
</tr>
<tr>
<td>1N83</td>
<td>6.2 Å (1)</td>
<td>5.9 Å (2)</td>
<td>7.1 Å (1)</td>
</tr>
<tr>
<td>1OUV</td>
<td>3.0 Å (6)</td>
<td>2.9 Å (4)</td>
<td>4.3 Å (1)</td>
</tr>
<tr>
<td>1QKM</td>
<td>3.6 Å (16)</td>
<td>2.7 Å (2)</td>
<td>3.9 Å (2)</td>
</tr>
<tr>
<td>1TBF</td>
<td>3.1 Å (100)</td>
<td>2.8 Å (20)</td>
<td>4.1 Å (1)</td>
</tr>
<tr>
<td>1V9M</td>
<td>- (-)</td>
<td>- (-)</td>
<td>- (-)</td>
</tr>
<tr>
<td>1XQO</td>
<td>- (-)</td>
<td>- (-)</td>
<td>- (-)</td>
</tr>
<tr>
<td>1Z1L</td>
<td>3.1 Å (150)</td>
<td>3.2 Å (72)</td>
<td>5.9 Å (1)</td>
</tr>
<tr>
<td>2AX6</td>
<td>4.0 Å (1)</td>
<td>3.2 Å (5)</td>
<td>6.6 Å (3)</td>
</tr>
<tr>
<td>2CWC</td>
<td>- (-)</td>
<td>- (-)</td>
<td>- (-)</td>
</tr>
</tbody>
</table>

- **Perfect SSE Pool**

<table>
<thead>
<tr>
<th>protein</th>
<th>assembly</th>
<th>refinement</th>
<th>loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1IE9</td>
<td>3.3 Å (1)</td>
<td>2.6 Å (1)</td>
<td>7.8 Å (1)</td>
</tr>
<tr>
<td>1N83</td>
<td>3.2 Å (1)</td>
<td>2.4 Å (1)</td>
<td>3.7 Å (7)</td>
</tr>
<tr>
<td>1OUV</td>
<td>3.1 Å (10)</td>
<td>2.3 Å (6)</td>
<td>4.8 Å (1)</td>
</tr>
<tr>
<td>1QKM</td>
<td>3.1 Å (1)</td>
<td>3.3 Å (1)</td>
<td>4.2 Å (7)</td>
</tr>
<tr>
<td>1TBF</td>
<td>3.2 Å (8)</td>
<td>2.7 Å (17)</td>
<td>4.2 Å (3)</td>
</tr>
<tr>
<td>1V9M</td>
<td>3.3 Å (1)</td>
<td>2.0 Å (1)</td>
<td>6.7 Å (2)</td>
</tr>
<tr>
<td>1XQO</td>
<td>3.3 Å (2)</td>
<td>2.1 Å (7)</td>
<td>5.0 Å (1)</td>
</tr>
<tr>
<td>1Z1L</td>
<td>3.4 Å (3)</td>
<td>2.5 Å (13)</td>
<td>5.5 Å (1)</td>
</tr>
<tr>
<td>2AX6</td>
<td>3.4 Å (1)</td>
<td>3.4 Å (1)</td>
<td>9.2 Å (8)</td>
</tr>
<tr>
<td>2CWC</td>
<td>2.9 Å (2)</td>
<td>2.4 Å (8)</td>
<td>7.1 Å (2)</td>
</tr>
</tbody>
</table>
Structure Determination of Adenovirus Protein IIIa

- One capsid protein of adenovirus with unknown structure
- Determine electron density by exclusion of density of known proteins
Adenovirus Protein IIIa Density has Fourteen α-Helical Rods
Eleven of Fourteen α-Helices are Predicted with High Confidence
Cryo-EM PIIIa Model Superimposed with 3.5 Å X-ray density


Refinement to Models Accurate at Atomic-Detail

- Blue – round 1
- Green – round 2
- Red - native

RMSD to native structure

1X91
Blue – round 1
Green – round 2
Red - native

6 December 2011
© Jens Meiler
EM-Fold is Applicable to Proteins that Contain β-Strands