Prediction of protein 3D structure

Nikos C. Papandreou
Section of Cell Biology and Biophysics, Department of Biology, National and Kapodistrian University of Athens
1. Introduction

Anfinsen's dogma: At least for a small globular protein in its standard physiological environment, the native structure is determined only by the protein's amino acid sequence.

Primary Structure (Aminoacid Sequence) → Secondary Structure → Tertiary Structure → Function

Knowledge of the 3D structure of proteins provides valuable information regarding their function.

**Experiments related to:**
- Site directed mutagenesis
- Detection of mutations related to diseases
- Drug Design
- Biomolecular Interactions
- Protein Folding
- Protein engineering
- etc..

**Are facilitated:**
- Atomic coordinates of amino acid residues in 3D space
In recent years great progress has been achieved in the experimental determination of 3D structures of biological macromolecules by

X-Ray Crystallography
Cryo-Electron Microscopy
NMR Spectrometry

(-)
High Cost
Time consuming
The experimental determination of protein structures is not always possible
# PDB File

<table>
<thead>
<tr>
<th>ATOM</th>
<th>1</th>
<th>N</th>
<th>SER</th>
<th>2</th>
<th>54.667</th>
<th>35.581</th>
<th>53.555</th>
<th>1.00</th>
<th>86.83</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>2</td>
<td>CA</td>
<td>SER</td>
<td>2</td>
<td>53.586</td>
<td>35.391</td>
<td>54.515</td>
<td>1.00</td>
<td>91.77</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>3</td>
<td>C</td>
<td>SER</td>
<td>2</td>
<td>53.292</td>
<td>33.910</td>
<td>54.722</td>
<td>1.00</td>
<td>93.28</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>4</td>
<td>O</td>
<td>SER</td>
<td>2</td>
<td>52.974</td>
<td>33.477</td>
<td>55.830</td>
<td>1.00</td>
<td>95.38</td>
<td>O</td>
</tr>
<tr>
<td>ATOM</td>
<td>5</td>
<td>CB</td>
<td>SER</td>
<td>2</td>
<td>52.321</td>
<td>36.116</td>
<td>54.050</td>
<td>1.00</td>
<td>89.80</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>6</td>
<td>OG</td>
<td>SER</td>
<td>2</td>
<td>52.556</td>
<td>37.503</td>
<td>53.885</td>
<td>1.00</td>
<td>92.50</td>
<td>O</td>
</tr>
<tr>
<td>ATOM</td>
<td>7</td>
<td>N</td>
<td>GLY</td>
<td>3</td>
<td>53.403</td>
<td>33.139</td>
<td>53.646</td>
<td>1.00</td>
<td>88.76</td>
<td>N</td>
</tr>
<tr>
<td>ATOM</td>
<td>8</td>
<td>CA</td>
<td>GLY</td>
<td>3</td>
<td>53.143</td>
<td>31.713</td>
<td>53.695</td>
<td>1.00</td>
<td>82.53</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>9</td>
<td>C</td>
<td>GLY</td>
<td>3</td>
<td>51.754</td>
<td>31.387</td>
<td>53.184</td>
<td>1.00</td>
<td>78.25</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>10</td>
<td>O</td>
<td>GLY</td>
<td>3</td>
<td>50.871</td>
<td>32.245</td>
<td>53.190</td>
<td>1.00</td>
<td>80.98</td>
<td>O</td>
</tr>
<tr>
<td>ATOM</td>
<td>11</td>
<td>N</td>
<td>PRO</td>
<td>4</td>
<td>51.550</td>
<td>30.139</td>
<td>52.739</td>
<td>1.00</td>
<td>67.80</td>
<td>N</td>
</tr>
<tr>
<td>ATOM</td>
<td>12</td>
<td>CA</td>
<td>PRO</td>
<td>4</td>
<td>50.259</td>
<td>29.714</td>
<td>52.189</td>
<td>1.00</td>
<td>58.44</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>13</td>
<td>C</td>
<td>PRO</td>
<td>4</td>
<td>49.177</td>
<td>29.566</td>
<td>53.256</td>
<td>1.00</td>
<td>62.47</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>14</td>
<td>O</td>
<td>PRO</td>
<td>4</td>
<td>49.465</td>
<td>29.170</td>
<td>54.386</td>
<td>1.00</td>
<td>59.81</td>
<td>O</td>
</tr>
<tr>
<td>ATOM</td>
<td>15</td>
<td>CB</td>
<td>PRO</td>
<td>4</td>
<td>50.581</td>
<td>28.349</td>
<td>51.578</td>
<td>1.00</td>
<td>46.57</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>16</td>
<td>CG</td>
<td>PRO</td>
<td>4</td>
<td>51.726</td>
<td>27.844</td>
<td>52.378</td>
<td>1.00</td>
<td>46.15</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>17</td>
<td>CD</td>
<td>PRO</td>
<td>4</td>
<td>52.546</td>
<td>29.054</td>
<td>52.719</td>
<td>1.00</td>
<td>63.15</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>18</td>
<td>N</td>
<td>VAL</td>
<td>5</td>
<td>47.943</td>
<td>29.889</td>
<td>52.886</td>
<td>1.00</td>
<td>63.19</td>
<td>N</td>
</tr>
</tbody>
</table>

Lamia, 14-16/12/2018 Prediction of Structure and Function of Proteins
Entries in Databases

**Structures**
- PDB  ~150000 entries

**Sequences**
- Uniprot/Swissprot  ~560.000 entries
- Uniprot/TrEMBL  ~115.000.000 entries

And the difference increases....
Attempt to fill the gap by application of Computational Methods

The prediction of the 3D structure of a protein is one of the biggest challenges in Computational Biology
Approaches (1)

A. Comparative modeling / Homology modeling

1. Provides the most accurate results
2. Applicable only when sequence similarity between the target protein and experimentally determined protein structures is detected

B. Fold Recognition / Threading

Applied in cases when sequence similarity between the target protein and experimentally determined protein structures is not detected
Approaches (2)

C. De novo (ab initio) prediction of 3D protein structure from its sequence

1. Based on Physical Principles that define Protein Folding
2. They do not rely on evolutionary relationships

D. Integrative or hybrid methods

1. Combination of Computational and Experimental Data
2. May include all the previously mentioned approaches
2. Protein Modelling Methods

A. Comparative modeling/Homology modeling

Exploits the evolutionary relationship of a protein sequence of unknown structure (Target) with a sequence of a protein that is experimentally determined (Template)

It is based on the observation that two sequences that have evolutionary relationship they share similar 3D conformation
Structure is more conserved than Sequence
General Steps:

1. Search for a proper Template and Sequence Alignment

2. Model Building of structurally conserved parts of the protein and Prediction of the structure for the rest

3. Model Optimization

4. Model Evaluation (Quality Accuracy)
1. Search for a proper Template and Sequence Alignment

The proper choice of a template and sequence alignment are crucial

Model Accuracy

- >40% sequence identity -> High

- <40% sequence identity -> Low
Sequence similarity implies structural similarity?

Percentage sequence identity/similarity vs. Number of residues aligned.

Sequence identity implies structural similarity

Safe zone
Template search (1)
Template Quality

• Resolution

• B-factor

• Ramachandran Plot

• NMR structure as a Template
Sequence Alignment Tools

• Clustal Omega
  https://www.ebi.ac.uk/Tools/msa/clustalo/

T-Coffee
  https://www.ebi.ac.uk/Tools/msa/tcoffee/

A Collection of Sequence Alignment Tools is available at:
  http://toolkit.tuebingen.mpg.de/
Pairwise sequence alignment results
CLUSTALW format

**Pairwise sequence alignment results**

**CLUSTALW format**

<table>
<thead>
<tr>
<th>tr</th>
<th>G9BX57</th>
<th>G9BX57_CERCA</th>
<th>3U9C_A</th>
<th>PDBID</th>
<th>CHAIN</th>
<th>SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>--TLPSA ARYTDVNAHKPD EYWDENYVVDWAN QDDYQLVRKLGRGKYS</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGPVPSARVYTDVNTHRPREYWDY ESHVEWGNQDDYQLVRKLGRGKYS</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>..:: ********::<em>::</em> ****:::::--------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVFEA INITTTEKCVVKIL PVKKKIKREIKILENLRGGTNIITLLAVV</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVFEA INTNNEKVVKILPVKKKIKREIKILENLRGPNIIITLADIV</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*****<strong><strong><em>:</em>:::</strong></strong>************************<em>:</em>::</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KDPVSRTPALIFEHVNNNTDFKQLYQTLDYEIRYYLFELLKALDYCHSMG</td>
<td>148</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KDPVSRTPALIFEHVNNNTDFKQLYQTLDYDIRFMYEILKALDYCHSMG</td>
<td>150</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>***************************<em>:::</em>:::--------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMHRDVKPHNVIMIDHNRKLRLIDWGLAEFYHPOEQYNVRVASRYF KGE</td>
<td>198</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMHRDVKPHNVIMIDHNRKLRLIDWGLAEFYHPOEQYNVRVASRYF KGE</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>***************************<em>:::</em>:::--------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLVDYQMYDSLIA WGLCMSMI FRKPE FFGHDN YDQLVRIA KVLG T</td>
<td>248</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLVDYQMYDSLIA WGLCMSMI FRKPE FFGHDN YDQLVRIA KVLG T</td>
<td>250</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>***************************<em>:::</em>:::--------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EELYA YLDK YNIELDPRFH DILQRHSRK RWERFVHS NQQLVSP EALDFL</td>
<td>298</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELYDYOYLDK YNIELDPRFH DILQRHSRK RWERFVHS NQQLVSP EALDFL</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>::** :::********::<em>::</em> :::::*:::--------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DKLLRY DHVERLTAREAMGHYPYFLP IVNGQIKS NQ</td>
<td>335</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DKLLRY DHQSR LT ARE AMGHYPYF T V K D QA------</td>
<td>331</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>***************************<em>:::</em>:::--------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lamia, 14-16/12/2018

Prediction of Structure and Function of Proteins 19
Pairwise sequence alignment results
PIR format

>P1;3U9C_A|PDBID|CHAIN|SEQUENCE

SGPVPSRARVYTDVNHPRPREYWDYESHVVEWGNQDDYQLVRKLGGRKYS
EVFEANITNNEKSVVKKKKREIKILENLRGGPNHIVLADIV
KDPVSRTPALVFEVNDTFQLYQTLTDYDIRFMYEILKALDYCHSMG
IMHRDVKHPHNVIDHREKLRQIDWGLAEFYHQPGEYNVRVASRYFKGPE
LLVDYQMYDYLDMWSLGCMLASIMIFRKEPPFFHGHDNYQLVRIAKVLGT
EDLYDYIDKNYELDPFNDILGHRKSRKHERFVHSNQHLSPEALDFL
DKLRLYDHQSRLTAREAMEHPYFVTVKDQA------

*
2. Model Building

A. Modeling by Assembly of Rigid Bodies

A comparative model can be assembled from a framework of small number of rigid bodies obtained from the aligned template protein structures.

B. Modeling by Segment Matching or Coordinate Reconstruction

Comparative models can be constructed by using a subset of atomic positions from template structures as “guiding” positions, such as the C\textsubscript{a} atoms, and by identifying and assembling short, all-atom segments that fit these guiding positions.

C. Modeling by Satisfaction of Spatial Restraints

Generation of constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. Homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom–atom contacts that are obtained from a molecular mechanics force field. The model is then derived by minimizing the violations of all the restraints.

D. Combining Alignments, Combining Structures

Modeling Loops, Insertions

1. Database search approaches (Fragment-Based Approach to Loop Modeling)

2. Conformational search approaches (Ab Initio Modeling of Loops)
Side Chains

Side Chains play a key role in molecular recognition and the packing of hydrophobic cores of globular proteins.

Side chain conformations are called rotamers.
3. Model optimization
   Necessary step for geometry optimization and minimization of errors due to steric hindrance

by

Energy minimization and Molecular Dynamics

4. Model Evaluation (Quality Estimation)
   • Verify 3D  http://servicesn.mbi.ucla.edu/Verify3D/
   • The ModFOLD Model Quality Assessment Server
     http://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD_form_2_0.html
   • WHATCHECK
     https://swift.cmbi.umcn.nl/gv/whatcheck/
     etc.....
Accuracy of Modeling Methods

- Critical Assessment of protein Structure Prediction (CASP)
  
  [http://predictioncenter.org/](http://predictioncenter.org/)

- CAMEO continuously evaluate the accuracy and reliability of predictions
  
  [https://www.cameo3d.org/](https://www.cameo3d.org/)
Limitations in the application of Comparative Modeling

1. No template structures available, alignment errors

2. Intrinsically disordered proteins
   (Cell Signaling, Transcription regulation .....)

3. Transmembrane proteins
   e.g. GPCRs
B. Fold Recognition / Threading

Method for modeling of proteins that have the same fold as proteins of known structures, but do not have homologous proteins with known structure

- **I-TASSER** (Iterative Threading ASSEmbly Refinement)
  [https://zhanglab.ccmb.med.umich.edu/I-TASSER/](https://zhanglab.ccmb.med.umich.edu/I-TASSER/)

- **RaptorX**: a Web Portal for Protein Structure and Function Prediction
  [http://raptorx.uchicago.edu/](http://raptorx.uchicago.edu/)

A list of Protein Structure Prediction Software can be found at:
PRACTICAL

During the practical we will use the following programs and discuss their results

• Modeller (standalone application available at https://salilab.org/modeller/ for download)
• Swiss Model (https://swissmodel.expasy.org/)
• I-TASSER (https://zhanglab.ccmb.med.umich.edu/I-TASSER/)
• RaptorX (http://raptorx.uchicago.edu/)
Please do not forget to fill out the questionnaire regarding our workshop!!!

Follow the link: https://goo.gl/forms/KMu6HLeQ3YjwUb6z1

Thank you!!!!!