Improving *De novo* Protein Structure Prediction using Contact Maps Information

Karina Baptista dos Santos
PhD student in Computational Modeling

National Laboratory for Scientific Computing (LNCC)
Petrópolis, RJ - Brazil
Introduction

- Protein Structure Prediction (PSP) have a large potential for biotechnological applications

  - Can provide essential clues to understanding diseases
  - Allows the creation of new proteins and the development of new drugs

MQTVLKKRKKSGGYIPD IADIRDFSYTEKSVIAA LPPKVDLTSPFQVDQGR IGSCTANALAAAIFERI HDKQSPEFIPSLFI

Protein Function
Introduction

- CASP (Critical Assessment of Techniques for Protein Structure Prediction)

  http://predictioncenter.org/

  - 12th CASP (1994-2016)
  - Assess the state of the art of protein structure prediction methods
    - Evaluation via process of **blind prediction**
  - Progress in PSP methodologies
    - Use of protein **contact maps** in the search of native conformation
  - Highlight where the future effort may be focused

Proceedings of the National Academy of Sciences
pages 7298-7303, 30 NOV 2006
http://www.pnas.org/content/103/19/7298/F1.expansion.html
Protein Contact Maps → CASP10, CASP11 and CASP12

- A minimalist representation of the 3D protein structure.

- Two residues in a protein are “in contact” if the Euclidean distance between their $C\beta$ atoms ($C\alpha$ for Glycine) is at most some threshold value, usually 8Å.

$$C\beta_i - C\beta_j \leq 8.0\text{Å}$$
Introduction

- Protein Contact Maps prediction → Residue covariation analysis
  - Sufficiently large number of homologous proteins (MSA)
    - Which residue pairs are most likely to be in contact?
    - Which residue pairs are important to be maintained by evolution?
  - These contacts are not template!
  - Precision threshold $[0, 1] \rightarrow$ Probability of such residues being in contact

Plos One, v.6, n.12
page e28766, 7 DEC 2011 DOI: 10.1002/prot.24367
https://doi.org/10.1371/journal.pone.0028766
- Protein Contact Maps prediction
  - MetaPSICOV\(^1\) (the CASP11 winner)
    - Stage 1
      - Classic contact prediction features
      - 3 three coevolution-based methods
        - PSICOV, mfDCA-Free Contact and CCMpred
    - Stage 2
      - Fill in the gaps in contact map and remove outliers (Filter)
      - Hydrogen bonds patterns between residues in contact

Objectives

Develop a strategy to use the information of protein contact maps in the GAPF PSP program

- Contact Map Term → a new term of the GAPF Fitness Function

Assess the predictive information capacity of the contact maps provided by MetaPSICOV Stage 1
Methods

- GAPF\textsuperscript{2, 3} - Genetic Algorithm for Protein Folding
  - Phenotypic \textbf{crowding-based} steady-state genetic algorithm
  - Candidate solutions are encoded by the \textbf{backbone dihedral angles}
  - \textbf{Coarse-Grained} representation for the side chain atoms

Methods

- GAPF - *Genetic Algorithm for Protein Folding*

  - Protein Fragment Libraries $\rightarrow$ Profrager$^4$

  ![Fragment Structure](image1)

  ![Protein Model](image2)

- Secondary Structure Prediction $\rightarrow$ PSIPRED$^5$

  ![Secondary Structure Prediction](image3)

---


Methods

• **GAPF - Fitness Function**\(^7\)

  - Based on the energy from the interaction between the atoms of the protein
    - Dihedral potential (from Gromos96 force field);
    - Atomic repulsive term;
    - Hydrophobic compaction term;
    - 4 Hydrogen bond terms

---

\(^7\)Rocha, Gregório Kappaun. Desenvolvimento de Metodologias Para Predição de Estruturas de Proteínas Independente de Moldes. Laboratório Nacional de Computação Científica. 2015.
Contact Map Term → Distance constraints

\[ \lambda(i, j) = \begin{cases} \gamma \times 1000 & \text{if } 2 \text{ Å} \leq \sigma(i, j) \leq 8 \text{ Å} \\ \gamma \times 500 & \text{if } 8 \text{ Å} < \sigma(i, j) \leq 10 \text{ Å} \end{cases} \]

- \( \lambda \) = Contribution value add to the predicted model energy
- \( \gamma \) = Precision threshold associated with the probability of such residues being in contact
- \( \sigma(i, j) \) = Euclidean distance between C\( \beta \) atoms of residues \( i \) and \( j \)

\[ CM_{rr} = - \sum_{i,j}^{rr\_pairs} \lambda(i, j) \]

- The more residues are respecting the distance constraintsts, the greater the contribution to the fitness function
• Test Set – Experimentally determined 3D structures

3FIL – 56 residues

T0773-D1 – 67 residues
(PDB 2N2U)

T0820-D1 – 90 residues

T0769-D1 – 97 residues
(PDB 2MQ8)

T0766-D1 – 108 residues
(PDB 4Q53)
Contact Maps

- **Native Map** → All residues pairs found in native structure (PDB file)
  - Precision threshold (PT) = 1.0

<table>
<thead>
<tr>
<th>R1</th>
<th>R2</th>
<th>d_{min}</th>
<th>d_{max}</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>115</td>
<td>0</td>
<td>8</td>
<td>1.0</td>
</tr>
<tr>
<td>64</td>
<td>74</td>
<td>0</td>
<td>8</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
<td>0</td>
<td>8</td>
<td>1.0</td>
</tr>
<tr>
<td>33</td>
<td>40</td>
<td>0</td>
<td>8</td>
<td>1.0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methods

• Contact Maps

  ○ **Filtered Map** → Only the residues pairs predicted by MetaPSICOV Stage1 that are present in the Native Map.

  ✓ Precision threshold (PT) = MetaPSICOV Predicted value

<table>
<thead>
<tr>
<th>R1</th>
<th>R2</th>
<th>d_{min}</th>
<th>d_{max}</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>115</td>
<td>0</td>
<td>8</td>
<td>0.71</td>
</tr>
<tr>
<td>64</td>
<td>74</td>
<td>0</td>
<td>8</td>
<td>0.56</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
<td>0</td>
<td>8</td>
<td>0.44</td>
</tr>
<tr>
<td>33</td>
<td>40</td>
<td>0</td>
<td>8</td>
<td>0.34</td>
</tr>
</tbody>
</table>

......
Methods

• Contact Maps
  
  ○ **Filtered Map** → Only the residues pairs predicted by MetaPSICOV Stage1 that are present in the Native Map.
  
  ✓ Precision threshold (PT) = MetaPSICOV Predicted value

<table>
<thead>
<tr>
<th>R1</th>
<th>R2</th>
<th>d_{min}</th>
<th>d_{max}</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>115</td>
<td></td>
<td></td>
<td>0.71</td>
</tr>
<tr>
<td>64</td>
<td>74</td>
<td></td>
<td></td>
<td>0.56</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
<td></td>
<td></td>
<td>0.44</td>
</tr>
<tr>
<td>33</td>
<td>40</td>
<td></td>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Native X Filtered

Same Residues / Different Precision Threshold
Filtered Map - Precision threshold profile

![Graph 1: Precision Threshold Ranges](image1)

![Graph 2: Precision Threshold Ranges](image2)

![Graph 3: Precision Threshold Ranges](image3)

![Graph 4: Precision Threshold Ranges](image4)

Contact Pairs

Precision Threshold Ranges

1.00

0.88

0.66

0.50

0.38

0.63
Native x False contacts in the MetaPSICOV predicted map

- Filter native contacts from predicted map → Great challenge

No method described is able to do this selection efficiently

- Precision threshold above some value (e.g. 0.5, 0.4 …)
- Number of contacts associated with the sequence length (e.g., L/5, L/2…)

<table>
<thead>
<tr>
<th>Contact Pairs</th>
<th>All MetaPSICOV Stage1 contacts</th>
<th>Native contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>3filA</td>
<td>1275</td>
<td>107</td>
</tr>
<tr>
<td>T0773-D1</td>
<td>1953</td>
<td>132</td>
</tr>
<tr>
<td>T0820-D1</td>
<td>3655</td>
<td>78</td>
</tr>
<tr>
<td>T0769-D1</td>
<td>4278</td>
<td>204</td>
</tr>
<tr>
<td>T0766-D1</td>
<td>5356</td>
<td>256</td>
</tr>
</tbody>
</table>

17
Results

- Impact of the use of Contact Maps
  - 4000 models predicted by GAPF

**Best of All** Models Predicted

**Top5** Best Energy Models

- RMSD $\leq 4.0\text{Å}$ -> Good predictions
Results

• Impact of the use of Contact Maps
  ○ 4000 models predicted by GAPF

**Best of All Models Predicted**

**Top5 Best Energy Models**

○ RMSD $\leq 4.0\text{Å} \rightarrow$ Good predictions
• Satisfied contact **does not** mean correct contact!

<table>
<thead>
<tr>
<th></th>
<th>Best of All</th>
<th>Top 5 Best Energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native Map</td>
<td>60%</td>
<td>65%</td>
</tr>
<tr>
<td>Filtered Map</td>
<td>48%</td>
<td>44%</td>
</tr>
</tbody>
</table>

\[ \beta_i - \beta_j \leq 8.0 \text{Å} \]

Native Structure
\[ \beta - \beta = 7.71 \text{Å} \]

Model
\[ \beta - \beta = 4.23 \text{Å} \]
Results

- RMSD x Energy (Native Map) → The greatest contribution to GAPF
Results

- RMSD x Energy (Standard Procotol)
  - Difficult to select the best structure model → methodology to select decoys
Results

- **T0820-D1 - Summary of improvements**

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSD</th>
<th>GDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAFP_St</td>
<td>9.60 Å</td>
<td>40.28 %</td>
</tr>
<tr>
<td>GAFP_Cm Native map</td>
<td>3.19 Å</td>
<td>68.89 %</td>
</tr>
<tr>
<td>GAFP_Cm Filtered map</td>
<td>3.73 Å</td>
<td>66.11 %</td>
</tr>
</tbody>
</table>

Best of All Models

- RMSD = 9.60 Å
- GDT = 40.28 %

Top 5 Best Energy Models

- RMSD = 13.79 Å
- GDT = 26.94 %

Experimental Structure

Model predicted by GAPF program
Conclusions

• Contact Map in the form of Distance Constrains → Useful Strategy for PSP

• **Naive Potential** → Properly combined with GAPF fitness function

• Predictors give probability for all possible residue-residue contacts
  - **Lack of full confidence** in the prediction of the contacts

• Development of strategies to **filter and enhance** contact maps
  - Which contacts are **most valuable** for PSP
Thank you!

karinabs@lncc.br