CONFOLD: Residue-Residue Contact-guided
ab initio Protein Folding

Badri Adhikari
Student of Dr. Jianlin Cheng
Department of Computer Science
University of Missouri
Columbia MO 65211

3/13/2015

will present a method we developed, called CONFOLD, which can build 3D protein
models using residue contacts.
In the field of protein structure prediction, since last 5 years, there has been a resurgence of interest in the development of inter-residue contact-prediction methods.

- The excitement is not just because of breakthroughs but also because of having abundance of protein sequence data to analyze.
- A lot is being done on left side and much less on right (building 3D models)
amount of effort currently being spent to improve contact prediction accuracy
In the field of protein structure prediction, since last 5 years, there has been a resurgence of interest in the development of inter-residue contact-prediction methods.

The excitement is not just because of breakthroughs but also because of having abundance of protein sequence data to analyze.

A lot is being done on left side and much less on right (building 3D models)
residue-residue contacts in proteins
a contact in a protein

Residues $i$ and $j$ are in contact if
\[
\text{distance}(\text{C}_\beta_i, \text{C}_\beta_j) \leq 8 \text{ Å}
\]

Variations:
- use of C\alpha instead of C\beta
- other thresholds like 7 Å or 12 Å
- minimum sequence separation of 5 or 6 residues
a helix circle (not a real protein) to demonstrate how distance threshold influences # of contacts
[The audience will think how easy/difficult it can be to use contacts to reconstruct different shapes of proteins]
building 3D models using predicted contacts
- available benchmarks
- widely used for evaluating contact predictions
- popularly known as EVFOLD (2011) and FRAGFOLD (2014)
- building models from the scratch vs building models using fragments
- EVFOLD is very fast, no need of big sample space
- Fragment based folding cannot fold complicated structures (need a big sample space)
ab initio - Challenges

1. Reconstruct secondary structures
   reconstruct Helices and beta-sheets

2. Handle noise in predicted contacts
   If we try to satisfy all contacts,
   the model can easily jumble up
CONFOLD method

- convert secondary structures into distance, angle and h-bond restraints
- combine contact restraints and secondary structure restraints using weighting schemes
- use of customized CNS suite v1.3 for modelling
- two stage model-building method
  - improves beta-sheet quality
  - improves accuracy of models

The two novel ideas I introduce
In the second stage, CONFOLD tries to filter out noisy contacts to improve the quality of models.
Results
First verification of the CONFOLD method
- Superposition of crystal structure (green) and reconstructed top model (orange)
- beta-alpha-beta barrel protein 1YPI and antiparallel beta barrel protein 2QOM
[Primary reason for a lower reconstruction rate of β-sheets than helices is the presence of proline in strands. Since proline acts as hydrogen-bond acceptor only, when it appears in strands, the hydrogen-bonding pattern is broken.]
- Predicted contacts for 15 proteins in the EVFOLD data set
- Improvement in 12 out of 15 proteins
- Best of 400 models for each protein
only keep one image, roll it, color code it
- EVFOLD and stage1 are similar
- RR filter and CONFOLD are more similar
- Contact filtering is the major contributor of the improvement
- protein = 1NRV(A), L = 100, # of contacts = top 0.6L (sixty)
- green is native, left is stage 1 best model, right is stage 2 best model
- 8 out of 60 contacts are removed in stage 2
- 5 out of these 8 contacts in the native structure are separated by large distances, which certainly were hindering in stage 1
- for example, the contact 30-36 (in blue) at the two ends of a strand; they cannot be in contact
- 22% improvement ins TM-score
Conclusion

• We can reconstruct secondary structures, helices and beta-sheets, in proteins from scratch, very well

• Existing *ab initio* methods for building 3D models using contacts can be improved

• Two-stage model building approach can improve models’ accuracy by removing noisy contacts
Acknowledgements

to my advisor Dr. Cheng and
to my friends Deb, Renzhi, Jilong and Jie in our research lab.
blank - on purpose
Additional slides on CONFOLD method
# of secondary structure restraints in CONFOLD

For a helix of 10 residues 107 restraints in total were derived, including 20 dihedral angle restraints, 7 hydrogen bond restraints, and 80 backbone atom restraints. For a pair of strands, each 10 residues long, connected as antiparallel, 108 restraints were derived, including 20 dihedral restraints and 9 O-O backbone distance restraints for each strand, 10 hydrogen bond restraints, and 40 backbone atom restraints.
Upper bounds and lower bounds of hydrogen bond and oxygen-oxygen distance, dihedral angle and backbone atom-backbone atom distance measurements derived from the SABmark database with $\lambda = 0.5$ for reconstructing alpha helices, strands and $\beta$-sheets. In all sub-tables, the first column defines secondary structure type: parallel (P) or anti-parallel (A), generic strand (U), and helix (H). Measurements of upper and lower bounds of hydrogen bond distances for anti-parallel and parallel $\beta$-sheets and helices (sub-Table A), adjacent oxygen-oxygen atom distances in strands (sub-Table B), dihedral angles (sub-Table C). Distance restraints for reconstructing helices and $\beta$-sheets are presented in sub-Table D. In sub-Table D, second column defines atom pair (atom of residue 1 – atom of residue 2), third column is the hydrogen bond reference atom (oxygen or hydrogen), and fourth column is the neighbor distance of the second residue. If strands a-b and c-d (a, b, c and d being residue numbers) are antiparallel and have a hydrogen bond between residues b and c, with oxygen atom of b connected to hydrogen atom of c, then, referring to the first row from sub-Table D, we apply distance restraint of [7.4Å, 8.0Å] between oxygen of residue b and oxygen of residue (c+1).
Limitations of EVFOLD

1. Too many distance restraints for secondary structures
   eg. for a helix of 10 residues around 300 restraints
   because of only distance restraints

2. Low quality of secondary structures
   Confidence of contact prediction vs secondary structure prediction

3. β-strand pairing
   If there are only two strands, and some contacts,
   why don’t we directly pair them?

all limitations are secondary structures related
EVFOLD vs CONFOLD

1. secondary structure (already discussed)
2. using contacts
   all contacts have same weight
1. selecting contacts
   instead of $N_c = 30$ up to $L$ in steps of 10, I choose $0.4L$ to $2.2L$
1. weighting between contacts and secondary structure
   I have two weighting schemes: one with high weight to contacts and vice versa.
Additional slides on CONFOLD results
- Left: second verification of CONFOLD method
- Right: best models reconstructed for the protein 5p21
Improvement in the accuracy of best models (left) and all 400 models (right) in the second stage of CONFOLD over the first stage for 150 proteins in FRAGFOLD dataset.
Number of best models and the number of contacts used to build the best models for 150 proteins in FRAGFOLD dataset.
Comparison of accuracy and secondary structure quality of best models built by CONFOLD and EVFOLD. Columns H and E are number of helix and β-sheet residues assigned by DSSP. RMSD values are in Å.

<table>
<thead>
<tr>
<th>UNIPROT-NAME</th>
<th>Native</th>
<th>EVFOLD</th>
<th>CONFOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>H</td>
<td>E</td>
</tr>
<tr>
<td>YES_HUMAN</td>
<td>48</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>CHEY_ECOLI</td>
<td>114</td>
<td>47</td>
<td>20</td>
</tr>
<tr>
<td>SPTB2_HUMAN</td>
<td>106</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>OMPR_ECOLI</td>
<td>77</td>
<td>31</td>
<td>6</td>
</tr>
<tr>
<td>OPSD_BOVIN</td>
<td>248</td>
<td>165</td>
<td>8</td>
</tr>
<tr>
<td>O45418_CAEEL</td>
<td>95</td>
<td>11</td>
<td>31</td>
</tr>
<tr>
<td>RNH_ECOLI</td>
<td>140</td>
<td>53</td>
<td>44</td>
</tr>
<tr>
<td>PCBP1_HUMAN</td>
<td>63</td>
<td>25</td>
<td>17</td>
</tr>
<tr>
<td>ELAV4_HUMAN</td>
<td>71</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>THIO_ALIAC</td>
<td>103</td>
<td>30</td>
<td>25</td>
</tr>
<tr>
<td>CADH1_HUMAN</td>
<td>100</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>BPT1_BOVIN</td>
<td>53</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>RASH_HUMAN</td>
<td>161</td>
<td>57</td>
<td>40</td>
</tr>
<tr>
<td>A8MVQ9_HUMAN</td>
<td>107</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>TRY2_RAT</td>
<td>216</td>
<td>7</td>
<td>72</td>
</tr>
<tr>
<td>Average</td>
<td>113</td>
<td>36</td>
<td>25</td>
</tr>
</tbody>
</table>
CONFOUND vs FRAGFOLD

The best models predicted by FRAGFOLD have TM-score of 0.54, and those by CONFOLD have TM-score of 0.55, on average.

For the 150 proteins, we calculated the Pearson’s correlation between the precision of top-L/2 predicted contacts and the TM-scores of the best models for both FRAGFOLD and CONFOLD in order to find which method is more contact driven.

The correlations between the contact precision and FRAGFOLD models and CONFOLD models are 0.53 and 0.70 respectively.

This suggests that contacts played a more important role in the modeling process of CONFOLD than in FRAGFOLD.

However, the comparison here should be only considered a qualitative understanding of the performance of CONFOLD because the models of the two methods were not generated in the exactly same conditions. The caveats are that: (a) FRAGFOLD’s best models are best of 5 whereas CONFOLD’s best models are best of 400 models, (b) FRAGFOLD used fragment information and CONFOLD did not, and (c) the secondary structures used by CONFOLD may not be same as the one used by FRAGFOLD.
Additional slides on introduction
What are proteins?
Problem is more computational than biological, it is a 3D shape like many 3D shapes
<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCA</td>
<td>Inverse Potts Model: maximum entropy</td>
</tr>
<tr>
<td>EV couplings/fmDCA</td>
<td>Inverse Potts Model: maximum entropy</td>
</tr>
<tr>
<td>plmDCA</td>
<td>Inverse Potts Model: maximum likelihood</td>
</tr>
<tr>
<td>GREMLIN</td>
<td>Inverse Potts Model: maximum likelihood</td>
</tr>
<tr>
<td>PSICOV</td>
<td>Sparse inverse covariance estimation</td>
</tr>
<tr>
<td>RLS</td>
<td>Regularized Least Squares inverse covariance</td>
</tr>
<tr>
<td>mdMI</td>
<td>Multi-dimensional Mutual Information</td>
</tr>
<tr>
<td>gaussDCA</td>
<td>Continuous maximum entropy inverse Potts model</td>
</tr>
</tbody>
</table>

**Opportunities and limitations in applying coevolution-derived contacts to protein structure prediction**

Stuart Tetchman, Tomasz Kosciolek, David T. Jones

Department of Computer Science, University College London, London, UK

*Bio-Algorithms and Med-Systems*
Precision %

DNcon

AUC_PR

ConSIP2 (021)
UCI-I6B-Cmpor (398)
Pcons-net (410)
MULTICOM-cluster (420)
MLID (105)
RBO_Aleph (479)
Shen-Group (124)
SAM-T06-server (086)
MULTICOM-novel (041)
RaportX-Contact (057)
MULTICOM-construct (008)
Icos (455)
SAM-T06-server (073)
CNIO (067)
PLCT (231)
FALCON_Contract (262)
Distill (349)
IASL-COPE (402)
FLOUDAS_A1 (157)
FLOUDAS_A2 (326)
mns (038)
LEE (169)
FLOUDAS_A3 (235)
eThread (454)
Raghavagps-painint (047)

Proteins: Structure, Function, and Bioinformatics
Evaluation of residue-residue contact prediction in CASP10


First published: 31 August 2013
- test set of 100 proteins
- the predictions are performed with CORNET
Additional slides on reconstruction using true contacts
- All the 100000 proteins are grouped according to their shape into 6 classes x families and xyz.
- All protein reconstruction tasks until now use just contact. Adding secondary structures I show that we have a much stronger yes.
All the 100000 proteins are grouped according to their shape into 6 classes x families and xyz.
All protein reconstruction tasks until now use just contact. Adding secondary structures I show that we have a much stronger yes.