Machine learning algorithms for structural and functional genomics

Jian Peng
Department of Computer Science
College of Medicine, Institute of Genomic Biology
University of Illinois at Urbana-Champaign
The protein folding problem

Initial structure

Intermediate states

Native structure

Energy
Structure provides insights on function
Protein structure prediction

Amino acid sequence → Successful prediction algorithms → 3D structure

Successful prediction algorithms:
- Rosetta
- RaptorX
- I-TASSER
- ALPHAFOLD

Progress in recent CASPs:
- Machine Learning for molecular modeling
- Coevolution-based contact prediction methods in literature

From CASP13 evaluation by Abriata and Dal Peraro
Structure prediction
Current status: Protein structure prediction

Sequence → Alignment of homologs

Template-based modeling

Identify templates → PDB

Contact-assisted modeling

Co-evolutionary analysis → Deep learning models

Template 1

Template 2

Optimization → Backbone building

Backbone Optimization

Fragment assembly

Side-chain construction

Residue-residue constraints
Exploiting co-evolution for contact prediction

Multiple sequence alignment

Local residue preference

Strength of co-evolution

Evolutionary and structural constraints

Residue contact
Learning couplings from protein alignment

Capture independent sites

<table>
<thead>
<tr>
<th>Amino acid i</th>
<th>$X_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>K</td>
</tr>
<tr>
<td>D</td>
<td>T</td>
</tr>
<tr>
<td>D</td>
<td>R</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>T</td>
<td>D</td>
</tr>
<tr>
<td>D</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>R</td>
</tr>
<tr>
<td>T</td>
<td>S</td>
</tr>
<tr>
<td>D</td>
<td>R</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
</tr>
</tbody>
</table>

Multiple sequence alignment

$P(x) = p_1(x_1)p_2(x_2) \cdots p_L(x_L)$

or equivalently

$P(x) = \frac{1}{Z} \exp \left( \sum_i e_i(x_i) \right)$

Single potentials

Local preference
Learning couplings from protein alignment

Capture pairwise interactions

Multiple sequence alignment

Amino acid $i$ ($X_i$)

Amino acid $j$ ($X_j$)

\[
P(x) = \frac{1}{Z} \exp \left( \sum_i e_i(x_i) + \sum_{i \neq j} e_{ij}(x_i, x_j) \right)
\]

**Single potentials**
**Pairwise potentials**

Local preference
Co-evolution strength

Markov random field
Ising (Potts) model
Undirected graphical model
Learning with Markov Random Fields

\[ L(e) = \prod_{n=1}^{N} \frac{1}{Z_e^{(n)}} \prod_{i}^{L} \exp \left[ e_i(x_i^n) + \sum_{j \neq i} e_{i,j}(x_i^n, x_j^n) \right] \]

- **Partition function**
- **Singleton potentials**
- **Pairwise potentials**

**Local AA preference**
**Pairwise AA couplings**

**Learning algorithms:**
- Mean fields approximation: EVFold, DirectInfo
- Gaussian approximation: PSICOV
- Pseudolikelihood: GREMLIN, CCMpred

Deep convolutional NNs recognize image patterns

Deep convolutional NNs recognize coevolutionary patterns

DeepContact for contact prediction

Train on set of inferred couplings and known structures:

Co-evolution + PDB structure

... Co-evolution + PDB structure

Learn parameters

DeepContact convolutional neural network (CNN):

Novel prediction from sequence:

Co-evolution map from sequence

DeepContact contact map

Deep learning improves coevolution-based contact prediction

Ranked at the top in CASP12 in 2016 in Z-score ranking on par with two other deep-learning based methods (RaptorX-Contact and (Deep) MetaPSICOV) on other metrics

<table>
<thead>
<tr>
<th>#</th>
<th>GR code</th>
<th>GR name</th>
<th>Domains Count</th>
<th>SUM Zscore (&gt; -2.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>079</td>
<td>iFold_1</td>
<td>37</td>
<td>40.1644</td>
</tr>
<tr>
<td>2.</td>
<td>219</td>
<td>Deepfold-Contact</td>
<td>36</td>
<td>34.5989</td>
</tr>
<tr>
<td>3.</td>
<td>451</td>
<td>RaptorX-Contact</td>
<td>38</td>
<td>34.4778</td>
</tr>
<tr>
<td>4.</td>
<td>109</td>
<td>naive</td>
<td>36</td>
<td>27.8373</td>
</tr>
<tr>
<td>5.</td>
<td>431</td>
<td>Shen-Group</td>
<td>38</td>
<td>21.3752</td>
</tr>
<tr>
<td>6.</td>
<td>013</td>
<td>MetaPSICOV</td>
<td>38</td>
<td>17.6201</td>
</tr>
<tr>
<td>7.</td>
<td>287</td>
<td>MULTICOM-CLUSTER</td>
<td>38</td>
<td>15.1225</td>
</tr>
<tr>
<td>8.</td>
<td>345</td>
<td>MULTICOM-NOVEL</td>
<td>38</td>
<td>14.1530</td>
</tr>
<tr>
<td>9.</td>
<td>236</td>
<td>MULTICOM-CONSTRUCT</td>
<td>38</td>
<td>12.3499</td>
</tr>
<tr>
<td>10.</td>
<td>320</td>
<td>raghavagps</td>
<td>38</td>
<td>11.4462</td>
</tr>
</tbody>
</table>
Why is deep learning effective?

2D projection of 1st layer filters using tSNE

Deep neural network learns contact patterns
Recent developments go beyond contact prediction

CASP14: DeepMind’s AlphaFold 2

Blue: Predicted
Green: Actual

ORF8

ORF3a
Function Prediction
Optimization of protein function

Sequence \rightarrow Function

Antibody

Binding affinity

Fluorescent protein

Fluorescence

CRISPR/Cas9

Specificity
Sequence-to-function modeling

Need to differentiate function levels of closely related sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>...DNGVDGEWTYDDATKTFTVTE</td>
<td>1.0</td>
</tr>
<tr>
<td>...DNGCDGEWTYDDATKTFTVTE</td>
<td>0.2</td>
</tr>
<tr>
<td>...DNGVGGEWTYDDATKTFTVTE</td>
<td>3.9</td>
</tr>
<tr>
<td>...DNGVWGGEWTYDDATKTFTFTE</td>
<td>5.4</td>
</tr>
<tr>
<td>...DNGVMGGEWTYDDATKTFTDTE</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Need to model non-additive effect (epistasis)

\[ f_{ij} = f_i + f_j + \epsilon_{ij} \]

Need to generalize to unseen sequences/mutations

Image from Schmiedel et al., 2019
Successful sequence-to-function models

- Function / Fitness:
  - FC layers
  - Attention
  - LSTM

- Protein sequence

- Convolution
  - Global max pool
  - Fully connected

- Kinase-peptide binding (protein phosphorylation)
  - [ Luo*, Ma* et al., RECOMB, 2018 ]
  - [ Luo*, Ma* et al., RECOMB, 2019 ]

- Protein-RNA binding
  - [ Su*, Luo* et al., PLOS CB, 2019 ]

- Kinase-drug binding
  - [ Winner of IDG-DREAM Challenge, 2019 ]
Challenge: labeled data are expensive to get

Idea: unsupervised representation learning using language models using unlabeled data

Trained on Pfam / UniProt database with unlabeled data

Too general, not specific/sensitive to a single or few changes in the sequence

Another idea: Learning from natural evolution

Not enough for fitting language models but enough for getting good features

Co-evolution correlates with function

Evolutionary couplings

ECNet: integrating evolutionary contexts for protein function prediction

Specific and sensitive to mutations
Capturing global semantics

Luo, et al. RECOMB 2020
Evaluation on single-mutation datasets
Generalization to high-order mutations

Test on high-order (4~11) GFP variants

High-order (3~15) resistant TEM

Box plots showing predicted fitness for random variants and inhibitor resistant TEM alleles.

Spearman $\rho$ values for different training data configurations:
- single
- double
- triple
- single + double + triple

Training data comparison.
Engineer inhibition-resistant TEM1 beta-lactamase

**Experimental validation**

Variants construction & fitness measurement (ampicillin resistance)

ECNet

Fraction of improved predictions
Thank you & acknowledgements

**Students:** Sheng Wang, Yunan Luo, Hoon Cho, Nate Russell, Wesley Qian, Yang Liu, Yufeng Su, Palash Sashittal,

**Collaborators:** Bonnie Berger, Vik Khurana, Trey Ideker, Mikko Taipale, Jianzhu Ma, Jianyang Zeng, Qiang Liu, Mohammed El-Kebir, Huimin Zhao, Martin Burke, David Heckerman, ChengXiang Zhai, Jiawei Han, Aditya Parameswaran, Jadwiga Bienkowska, Lenore Cowen, Alex Schwing, Jonathan Carlson, Sumaiya Nazeen, Dina Zielinski