scoreD: Deep discriminative ensemble classifiers for protein scoring

Debswapna Bhattacharya
Auburn University

3DSIG @ ISMB 2018
What is protein scoring?

- **Quantitate prediction accuracy** of computational protein models
- Compare with experimental (native)
- What happens *when native is absent*?
  - Predict accuracy of predicted models
- Goes by other names
  - Quality Assessment (QA)
  - Estimate of Model Accuracy (EMA)
  - Quality Estimation (QE)
Why protein scoring is important?

– Distinguish folded structure from alternatives (a.k.a. decoys)

– Rank decoy pool for selecting the best decoy

– Guide conformational search during 3D protein modeling
How protein scoring is done?

Physics-based MM force fields

- AMBER, CHARMM, ...

Knowledge-based statistical potentials

- DFIRE, DOPE, ...

Unsupervised clustering or structural consensus

- ModFOLD, APOLLO, ...

Promising in CASP

Supervised machine learning

- ProQ, SVMQA, ...
Supervised learning for protein scoring

- Support Vector Machine (ProQ2, ProQ3, SVMQA)
- Random Forest (RFMQA)
- Deep Learning (ProQ3D, DeepQA)

Machine learning regression
Global Distance Test (GDT)

- Widely used structural superposition based accuracy metric

\[
GDT - TS = \frac{(N_1 + N_2 + N_3 + N_4)}{4L}
\]

- \(N_1, N_2, N_3, N_4\) = \# of aligned residues below 1, 2, 4, 8Å \text{ thresholds}
- \(L\) = length of the protein
- GDT-TS \(\in [0, 1)\)

- Based on the \textbf{principle of categorization}

\textbf{Classification} is a natural choice
scoreD: What’s conceptually new?

**binary classification** for modeling GDT-TS
Features and class labels

**features**

- Sequence profile conservation
  - PSI-BLAST against NR
  - Information per position
- Predicted vs. true features
  - SPIDER2 and DSSP
  - SS, SA
- Rosetta centroid terms
  - Twelve terms
  - sigmoidal transformation

**labels**

++ Positive class

if \( d_{CA} \leq r \ Å; r \in \{1,2,4,8\} \)

- Negative class

Otherwise

ensemble classifiers

---
Training data and proportion of classes

- **3DRobot set**
  - Evenly distributed
  - 300 proteins x 200 decoys
  - 6,000 decoys

- **GDT-TS binning**
  - 10 bins with width 0.1 GDT-TS
  - Max 1 decoy per bin / target
  - 1,628 decoys

Class imbalance problem exists

\[
\frac{\text{+ve}}{\text{-ve}}
\]

<table>
<thead>
<tr>
<th>Bin Width (Å)</th>
<th>+ve</th>
<th>-ve</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Å</td>
<td>44,405</td>
<td>173,869</td>
<td>0.255</td>
</tr>
<tr>
<td>2 Å</td>
<td>80,147</td>
<td>138,127</td>
<td>0.58</td>
</tr>
<tr>
<td>4 Å</td>
<td>125,642</td>
<td>92,632</td>
<td>1.356</td>
</tr>
<tr>
<td>8 Å</td>
<td>176,221</td>
<td>42,053</td>
<td>4.19</td>
</tr>
</tbody>
</table>
AUC-Maximized Deep Convolutional Neural Fields

- **DeepCNF**
  - Integration of DCNN and CRF
  - Better for sequence labeling task with imbalanced label distribution

- **Maximum-AUC algorithm**
  - Train by maximizing empirical Area Under the ROC Curve (AUC)
  - Shown to greatly outperform maximum-likelihood and maximum labelwise accuracy

Suitable for **class imbalance** problem

---

Xu and coworkers
Overall architecture

- Input features
  - Sequence profile
  - Predicted vs. True SS, SA
  - Rosetta centroid terms

- Ensemble classifiers
  - Ensemble of four DeepCNF
  - Labels with $r = 1, 2, 4, 8\text{Å}$

- Classify each residue
  - Probability to be $\leq r\text{Å}$
  - Four sets of classifications

- Weighted combination
  - Probabilistic GDT-TS
  - $scoreD \in [0, 1)$

---

- 5 hidden layers
- Each having 50 neurons
- Window size of 21
- Sigmoid activation function
scoreD: Results

**benchmark on 258 protein targets**
## Results 1/2: Native recognition

<table>
<thead>
<tr>
<th>datasets</th>
<th>comparisons with statistical potentials</th>
<th>measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Modeller set (20 proteins)</td>
<td>- RW</td>
<td>- % native recognized ↑</td>
</tr>
<tr>
<td>- Rosetta set (58 proteins)</td>
<td>- RAPDF</td>
<td>- Z-score of the native ↑</td>
</tr>
<tr>
<td>- Original + 3DRobot</td>
<td>- KBP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- DFIRE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- DOPE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- SRS</td>
<td></td>
</tr>
</tbody>
</table>
Modeller and Rosetta sets

- better native recognition
- ‘less fooled’ by 3DRobot
Results 2/2. Decoy scoring

datasets
- CASP11 (80 proteins)
- CASP12 (40 proteins)
- CAMEO (60 proteins)

comparisons with top single QA
- ProQ2
- ProQ3
- ProQ3D
- Qprob
- DeepQA

measures
- GDT-HA correlation ↑
- GDT-HA loss ↓
CASP11 set

- highest correlation of 0.65
- comparable loss of 0.079
  - all others < 0.6
  - best is 0.078
CASP12 set

- highest correlation of 0.708
- relatively higher loss of 0.106

all others < 0.68
best is 0.095
CAMEO set

- highest correlation of 0.904 all others < 0.9
- relatively higher loss of 0.043 best is 0.039
Summary

- **Pioneers** binary classification for predicting GDT-TS
- **Improved native recognition**; less ‘fooled’ by challenging sets
- **Best correlation** in CASP, CAMEO; room for improvement in loss
scoreD: **new direction** in protein scoring