

# DeepSF: Deep CNN for mapping protein sequences to folds

Zhiyang's presentation about [1]

Zhiyang Ong

Department of Electrical and Computer Engineering  
Dwight Look College of Engineering,  
Texas A&M University  
College Station, TX

March 27, 2020



① Problem and Knowledge Gap

② Proposed Solution(s)

③ Experimental Results

④ Discussion and Muddy Points

# Table of Contents

① Problem and Knowledge Gap

② Proposed Solution(s)

③ Experimental Results

④ Discussion and Muddy Points

# Context of the Problem

## Background Information

Context of the problem.

- ① The structures of most ( $> 99\%$ ) proteins are unknown
- ② Protein fold recognition enables us to associate a protein sequence to a protein fold
  - Identifying protein homologs that share the same protein fold
  - With this protein fold, determine its protein structure [Tramontano2003]
  - With this protein structure, determine its protein function
    - [Osadchy2011,OConnor2010,Tramontano2003,Hegyi1999]
    - [OConnor2014a, *From Contents: Unit 2, How Do Cells Decode Genetic Information into Functional Proteins?: §2.4 The Functions of Proteins Are Determined by Their Three-Dimensional Structures*]
    - [OConnor2014, *From Contents: Unit 2, How Do Cells Decode Genetic Information into Functional Proteins?: §2.4 The Functions of Proteins Are Determined by Their Three-Dimensional Structures*]

# Problem Definition - 1

What is the problem that [1] is solving?

Problem description...

- ① Problems with sequence-based methods, especially sequence alignment methods (including profile-sequence and profile-profile alignment methods)
  - Methods for mapping protein sequences to protein folds are indirect
  - Can't explain relationship between protein sequences & protein folds, even machine learning (ML) methods
  - traditional ML methods also can't work for classifying data into large number of categories ( $> 1000$ )
    - multi-layer perceptron
    - support-vector machines
    - ensemble classifiers
    - kernel-based learning
- ② Other methods have methodological limitations

## Problem Definition - 2

What is the problem that [1] is solving?

Problem description... Continued.

- ① Protein fold recognition enables us to associate a protein sequence to a protein fold
  - With this protein fold, we can determine its protein structure
  - With this protein structure, we can determine its protein function

Specifically... Solve the protein fold recognition problem.

- ① Given a protein sequence, map it to a protein fold **directly**
- ② Explain relationship between protein sequences & protein folds

# Problem Importance

Why is it important?

Why is this problem important?

- This facilitates protein structure prediction
- Knowing about protein folds help us in protein structure prediction
- Protein structure prediction enables protein function prediction
- Knowing about protein structure and function facilitates:
  - drug/medication design [Nogrady2005, §1.6.4, pp. 54] [Golan2008, Chapter 1, pp. 4]
  - biotechnology [Walsh2014, Chapter 2]
  - synthetic biology [Zhao2013d, Chapter 2]
  - personalized [Cullis2015, Chapter 2, pp. 26] or precision medicine [Mousa2020, §24.3.2, pp. 778]
  - gene therapy [Wecker2010, Chapter 5, pp. 51]

# What are the Knowledge Gaps in [1]?

How would [1] address these knowledge gaps?

List of knowledge gaps:

- Lack of direct methods for mapping protein sequences to protein folds
- Methods can't explain relationship between protein sequences & protein folds

# Table of Contents

① Problem and Knowledge Gap

② Proposed Solution(s)

③ Experimental Results

④ Discussion and Muddy Points

# Proposed Solution(s) - 1

What do they proposed to address the knowledge gap?

Proposed Solution(s).

- Use 1-D deep convolutional neural network (deep CNN)

# Proposed Solution(s) - 2

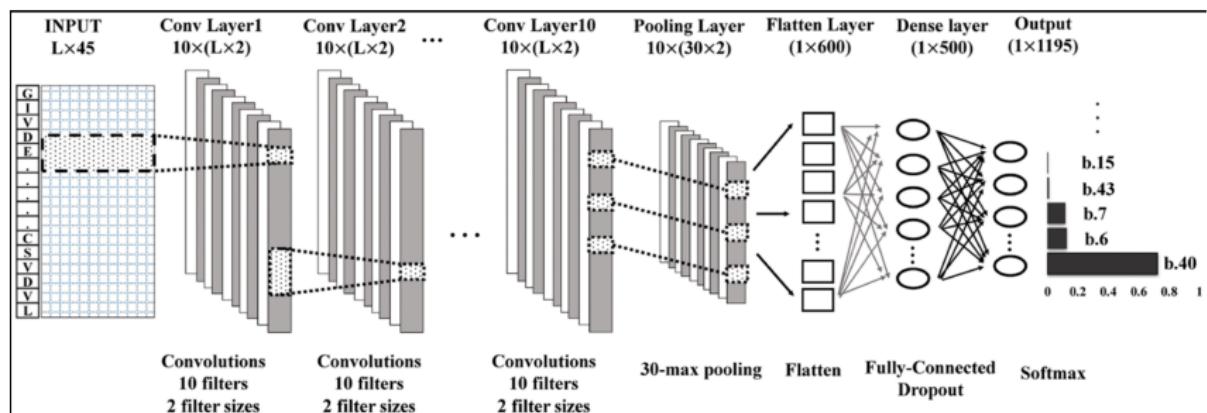
What do they proposed to address the knowledge gap?

Proposed Solution(s).

- Use input feature generation and label assignment from PSI-BLAST.
- Specifically, use position-specific scoring matrix
- Train all mini-batches for 100 epochs
- Distance metrics
  - ① Euclid-D, Euclidean distance
  - ② Manh-D, Manhattan distance
  - ③ Corr-D, Pearson's Correlation score
  - ④ KL-D, KL-Divergence

# Proposed Solution(s) - 3

Architecture for a 1-D deep CNN



# Dataset and Methods

Which datasets were used?

- SCOP 1.75 was used for training and validation (§3.1)
- SCOP 2.06 (§3.2) and CASP (§3.3) was used for test

# Table of Contents

① Problem and Knowledge Gap

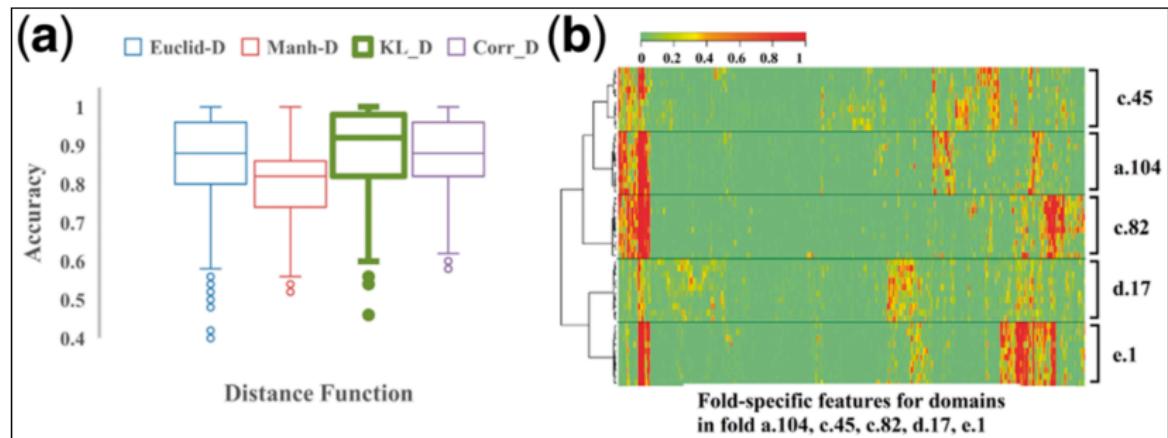
② Proposed Solution(s)

③ Experimental Results

④ Discussion and Muddy Points

# Experimental Results - 1

Evaluation of distance metrics based on clustering accuracy



**Figure:** The prediction accuracy at family/superfamily/fold levels for top 1, top 5 and top 10 predictions of DeepSF and PSI-BLAST, on SCOP 1.75 test dataset

# Experimental Results - 1

How do their proposed solutions compare with existing solutions?

**Table 1.**

The prediction accuracy at family/superfamily/fold levels for top 1, top 5 and top 10 predictions of DeepSF and PSI-BLAST, on SCOP 1.75 test dataset

| Level                       | Methods   | Top1   | Top5   | Top10  |
|-----------------------------|-----------|--------|--------|--------|
| Family (1272 proteins)      | DeepSF    | 76.18% | 94.50% | 97.56% |
|                             | PSI-BLAST | 96.80% | 97.40% | 97.60% |
| Superfamily (1254 proteins) | DeepSF    | 50.71% | 77.67% | 77.67% |
|                             | PSI-BLAST | 42.20% | 51.40% | 54.60% |
| Fold (718 proteins)         | DeepSF    | 40.95% | 70.47% | 82.45% |
|                             | PSI-BLAST | 5.60%  | 11.60% | 16.20% |

# Experimental Results - 2

How do their proposed solutions compare with existing solutions?

**Table 2.**

The accuracy of DeepSF on SCOP 2.06 dataset and its subsets

| DeepSF           | Top1   | Top5   | Top10  |
|------------------|--------|--------|--------|
| SCOP2.06 dataset | 73.00% | 90.25% | 94.51% |
| 'Large' folds    | 79.64% | 94.87% | 97.81% |
| 'Medium' folds   | 74.16% | 75.61% | 76.06% |
| 'Small' folds    | 67.93% | 86.86% | 94.74% |

# Experimental Results - 3

How do their proposed solutions compare with existing solutions?

**Table 3.**

The prediction accuracy at family/superfamily/fold level for top 1, top 5 and top 10 predictions, on SCOP 2.06 test dataset

| Type                        | Methods   | Top1   | Top5   | Top10  |
|-----------------------------|-----------|--------|--------|--------|
| Family (742 proteins)       | DeepSF    | 75.87% | 91.77% | 95.14% |
|                             | PSI-BLAST | 82.20% | 84.50% | 85.30% |
| Superfamily (1754 proteins) | DeepSF    | 72.23% | 90.08% | 94.70% |
|                             | PSI-BLAST | 86.90% | 88.40% | 89.30% |
| Fold (37 proteins)          | DeepSF    | 51.35% | 67.57% | 72.97% |
|                             | PSI-BLAST | 18.90% | 35.10% | 35.10% |

# Experimental Results - 4

How do their proposed solutions compare with existing solutions?

**Table 4.**

The performance of the methods on 88 template-based proteins in the CASP dataset

| Method         | Top1   | Top5   | Top10  |
|----------------|--------|--------|--------|
| DeepSF         | 46.59% | 73.86% | 84.09% |
| HHSearch       | 43.20% | 61.40% | 67.00% |
| Cons_HH_DeepSF | 59.10% | 77.30% | 85.20% |
| PSI-BLAST      | 15.90% | 31.80% | 47.70% |

# Experimental Results - 5

How do their proposed solutions compare with existing solutions?

**Table 5.**

The performance of the methods on 95 template-free proteins in the CASP dataset

| Method         | Top1   | Top5   | Top10  |
|----------------|--------|--------|--------|
| DeepSF         | 24.21% | 51.58% | 70.53% |
| HHSearch       | 11.58% | 34.74% | 44.21% |
| Cons_HH_DeepSF | 23.16% | 56.84% | 70.53% |
| PSI-BLAST      | 8.42%  | 15.79% | 32.63% |

# Experimental Results - 6

How do their proposed solutions compare with existing solutions?

**Table 6.**

Accuracy of protein structure predictions on 95 template-free targets

| Methods  | TM-score |      |      |      |
|----------|----------|------|------|------|
|          | Min      | Max  | Mean | Std  |
| DeepSF   | 0.15     | 0.54 | 0.27 | 0.07 |
| HHSearch | 0.11     | 0.52 | 0.25 | 0.08 |

# Experimental Results - 2

How do their proposed solutions compare with existing solutions?

## Experimental Results - 2.

-

# Experimental Results - 3

## Benchmarking

Benchmarking of results with PSI-BLAST.

- Solution is 12.63-26.32% better than HHSearch on template-free modeling targets.
- Solution is 3.39-17.09% better on hard template-based modeling targets.

# Table of Contents

① Problem and Knowledge Gap

② Proposed Solution(s)

③ Experimental Results

④ Discussion and Muddy Points

# Discussion of Experimental Results

What do the experimental results tell us?

Discuss the experimental results.

- Method is robust against:
  - ① sequence mutation
  - ② insertion
  - ③ deletion
  - ④ truncation
- Can solve other protein pattern recognition problems:
  - ① protein clustering
  - ② protein comparison
  - ③ protein ranking

# Weaknesses - 1

What are the weaknesses of this paper?

- Poor benchmarking methodology:
  - ① Benchmarked solution (especially in Table 1) against PSI-BLAST (Altschul S.F. et al. , 1997)
  - ② Did not benchmark with other modern solutions for protein fold prediction
    - ① PFPA (2015, IEEE Transactions on NanoBioscience); DOI: 10.1109/TNB.2015.2450233
    - ② random forest (2014, BMC Bioinformatics); DOI:10.1186/1471-2105-15-S11-S14
    - ③ PFP-RFSM (2013, J. Biomedical Science and Engineering); DOI:10.4236/jbise.2013.612145
  - Did not use box plots or bar charts to compare methods for each protein sequence in the datasets:

## Weaknesses - 2

What are the weaknesses of this paper?

- Inadequate **cross-referencing of figures and tables** in the text:
  - ① Benchmarked solution (especially in Table 1) against PSI-BLAST (Altschul S.F. et al. , 1997)
  - ② Did not benchmark with other modern solutions for protein fold prediction
    - ① PFPA (2015, IEEE Transactions on NanoBioscience); DOI: 10.1109/TNB.2015.2450233
    - ② random forest (2014, BMC Bioinformatics); DOI:10.1186/1471-2105-15-S11-S14
    - ③ PFP-RFSM (2013, J. Biomedical Science and Engineering); DOI:10.4236/jbise.2013.612145
- Did not use box plots or bar charts to compare methods for each protein sequence in the datasets:

# Muddy Points

What do I not understand about [1]?

What do I not understand about [1]?

- Why is it bad for top 1 predicted fold, but better for top 5 and top 10 predicted folds?

## References

-  Jie Hou, Badri Adhikari, and Jianlin Cheng.  
DeepSF: deep convolutional neural network for mapping  
protein sequences to folds.  
*Bioinformatics*, 34(8):1295–1303, April 15 2018.