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 26, 2020

Table of Contents

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
     - 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

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

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, Euclid, Manh-D, Corr-D, and KL-D

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

Experimental Results - 1

How do their proposed solutions compare with existing solutions?

Experimental Results - 1.

-

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

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

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 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.