

Predicting Protein Contact Maps by Bagging Decision Trees

Chuqiao Ren
Department of Computer Science
Bucknell University
Lewisburg, PA 173837 USA
chuqiao.ren@bucknell.edu

Brian R. King^{*}
Department of Computer Science
Bucknell University
Lewisburg, PA, 17837 USA
brian.king@bucknell.edu

ABSTRACT

Substantial progress has been made in the prediction of protein structure. Despite these achievements, the computational complexity of protein folding remains a challenge. Instead, many methods aim to predict a protein contact map from sequence. We introduce an ensemble method for protein contact map prediction based on bagging multiple decision trees. The amino acid alphabet is clustered to improve generality. A random sampling method is used to address the large class imbalance in contact maps. Our results show that our technique performs favorably against existing methods.

Categories and Subject Descriptors

J.3 [Computer Applications]: Life and Medical Sciences—*Biology and Genetics*

General Terms

Algorithms

Keywords

sequence analysis, protein contact map prediction

1. INTRODUCTION

The protein folding problem continues to remain among the grand challenges in science. Progress has been made in the development of methods that can predict various aspects of the tertiary structure of a protein from its sequence. Despite these achievements, the computational complexity of protein folding remains a challenge; not a single prediction method can achieve reliable quality over experimental methods [3]. Instead, many methods aim to predict a *protein contact map* (PCM) from sequence. A PCM is a two-dimensional (2D) binary distance matrix between all pairs

of residues in a protein, where each entry represents binary value depending on whether the actual distance between a specific pair of residues in the 3D structure of the protein falls within a predetermined threshold [5]. A PCM can help recover important elements in the 3D structure of the protein [9]. The majority of existing approaches use machine learning methods, such as artificial neural networks [5][3] and support vector machines [2].

We introduce a prediction method based on the C4.5 decision tree method [6]. An existing approach using this popular classifier has been previously developed, yielding an accuracy of 0.34 from 10-cross validation [7]. We improve the performance of this algorithm and improve its generality on new data by reducing the amino acid alphabet through clustering according to predetermined physicochemical properties. [8] Residue contacts are scarce entities in a PCM. To address the class imbalance problem with contacts, we use a random sampling technique to increase the representation of the minority class. Finally, we employ an ensemble model approach using a variation of bagging, where multiple trees are grown using our random sampling method.

2. METHODS

2.1 Protein Contact Map

The protein contact map (PCM) of a protein of length L is a $L \times L$ binary matrix. An entry e_{ij} in a PCM is 1 if two amino acids a_i and a_j are in contact; otherwise, e_{ij} is 0. The alpha carbon (C_α) of an amino acid is the first carbon that attaches to the functional group. Two amino acids a_i and a_j are in contact if the distance between their alpha carbons is less than 8. All entries that are less than 6 positions in the sequence apart are discarded.

2.2 Amino Acid Clustering

There are 20 amino acids in nature, but many amino acids share similar physicochemical properties. In order to reduce the the attribute space of the data, we used hierarchical clustering method with complete linkage to cluster amino acids over 12 major properties that may have substantial influence on how a protein folds [8].

2.3 Model Architecture

The sparseness of amino acid contacts results in a dataset that is highly imbalanced. To balance data and improve performance, we designed an enhancement of bagging to maximize the predictive power of the minority class. Given a dataset of contacts D , a sample of D , denoted S_k , is gener-

^{*}Corresponding author

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).

BCB'14, September 20–23, 2014, Newport Beach, CA, USA.

ACM 978-1-4503-2894-4/14/09.

<http://dx.doi.org/10.1145/2649387.2660818>.

ated by selecting all contacts, and a random sample of non-contacts without replacement. Sample S_k is formed such that it is 20% of the size of D . A single tree T_k is induced from S_k . This technique is repeated 50 times, resulting 50 different decision trees. A prediction is generated for each entry from each T_k . The average probability of each residue pair being in contact is generated over all 50 trees. All the predictions from each fold are combined and model performance is assessed. We used Quinlan’s C4.5 decision tree method with parameters $C = 0.5$ and minimum nodes = 20.

2.4 Datasets

We assessed datasets from Griep and Hobohm’s PDBselect list (25% threshold, nsigma = 3.0, method = X-ray, resolution ≤ 1.5) [4], which has 4350 proteins. We collected proteins that have a length of 50 to 150 residues. Data is partitioned into four groups determined by protein length (see **Table1**) A classifier is built for each group independently. Data instances are collected from contact maps by sliding a 5×5 window across the contact map. Residues labels along the sliding window represent attributes, and the cell closest to the diagonal represents the target class.

3. RESULTS

We used 10-fold cross validation to evaluate our method. For a protein of length L , only the top $L/5$ and $L/10$ predictions are assessed for accuracy [1]. For each set of proteins, we estimate $L/5$ and $L/10$ by calculating the average protein length, denoted L_{ave} . Let N represent the number of proteins in the data. Then, our estimate for $L/5$ is computed by $\frac{L_{ave}}{5} \times N$ and $L/10$ is $\frac{L_{ave}}{10} \times N$. The accuracy (or precision) is then computed for the top $L/5$ and $L/10$ predictions as $Acc = \frac{TP}{(TP+FP)}$. A ROC curve is generated for each data set with different protein length (**Figure1**). Accuracy results are reported in **Table1**.

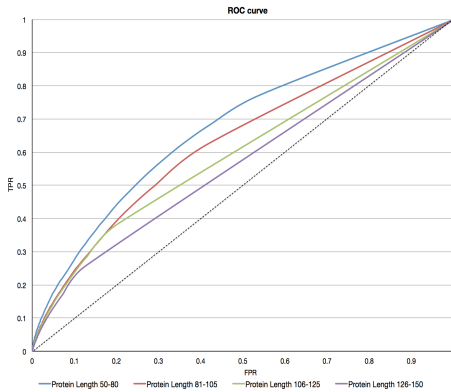


Figure 1: ROC curve for each protein data set with different length

4. DISCUSSION

Our results compare favorably with existing methods for contact map prediction, including DTP [7]. We are currently working to improve the performance of our predictor.

Table 1: Performance results for proteins with different length of 10-fold cross validation.

Protein Sequence Length	L/10	L/5
50 - 80	0.40	0.36
81 - 105	0.28	0.23
106 - 125	0.26	0.25
126 - 150	0.46	0.35

To this end, we found that proteins that are relatively abundant with alpha helices are harder to predict using a distance threshold of 8. Many helices do not appear properly in a contact map unless the minimum distance threshold is increased to 10 or 11. Our preliminary work has found that the performance of these alpha-rich proteins can be substantially improved by simply increasing the distance threshold. Our current work is focused on improving the overall performance for proteins with all types of structures.

5. REFERENCES

- [1] P. Chen and J. Li. Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. *BMC structural biology*, 10 Suppl 1(Suppl 1):S2, Jan. 2010.
- [2] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC bioinformatics*, 8(1):113, Jan. 2007.
- [3] W. Ding, J. Xie, D. Dai, H. Zhang, H. Xie, and W. Zhang. CNNcon: improved protein contact maps prediction using cascaded neural networks. *PloS one*, 8(4):e61533, Jan. 2013.
- [4] S. Griep and U. Hobohm. PDBselect 1992-2009 and PDBfilter-select. *Nucleic acids research*, 38(Database issue):D318–9, Jan. 2010.
- [5] M. Punta and B. Rost. PROFcon: novel prediction of long-range contacts. *Bioinformatics (Oxford, England)*, 21(13):2960–8, July 2005.
- [6] J. Quinlan. Induction of decision trees. *Machine learning*, pages 81–106, 1986.
- [7] C. Santiesteban-Toca, G. Asencio-Cortes, A. Marquez-Chamorro, and J. Aguilar-Ruiz. Short-Range Interactions and Decision Tree-Based Protein Contact Map Predictor. In M. Giacobini, L. Vanneschi, and W. S. Bush, editors, *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics 10th European Conference, EvoBIO 2012 Proceedings*, volume 7246 of *Lecture Notes in Computer Science*, pages 224–233, Málaga, Spain, 2012. Springer Berlin Heidelberg.
- [8] A. Sharma, K. K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, and S. Miyano. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. *BMC bioinformatics*, 14(1):233, Jan. 2013.
- [9] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding & design*, 2(5):295–306, Jan. 1997.