Title
CNNcon: Improved Protein Contact Maps Prediction Using Cascaded Neural Networks

Journal
PLoS ONE, 8(4)

ISSN
1932-6203

Authors
Ding, Wang
Xie, Jiang
Dai, Dongbo
et al.

Publication Date
2013-04-23

DOI
10.1371/journal.pone.0061533

OPEN ACCESS Freely available online

PLOS ONE

CNNcon: Improved Protein Contact Maps Prediction Using Cascaded Neural Networks

Wang Ding[1], Jiang Xie[1,2,3], Dongbo Dai[1], Huiran Zhang[1], Hao Xie[4], Wu Zhang[1,2]*

1 School of Computer Engineering and Science, Shanghai University, Shanghai, People's Republic of China, **2** Institute of Systems Biology, Shanghai University, Shanghai, People's Republic of China, **3** Department of Mathematics, University of California Irvine, Irvine, California, United States of America, **4** College of Stomatology, Wuhan University, Wuhan, People's Republic of China

Abstract

Backgrounds: Despite continuing progress in X-ray crystallography and high-field NMR spectroscopy for determination of three-dimensional protein structures, the number of unsolved and newly discovered sequences grows much faster than that of determined structures. Protein modeling methods can possibly bridge this huge sequence-structure gap with the development of computational science. A grand challenging problem is to predict three-dimensional protein structure from its primary structure (residues sequence) alone. However, predicting residue contact maps is a crucial and promising intermediate step towards final three-dimensional structure prediction. Better predictions of local and non-local contacts between residues can transform protein sequence alignment to structure alignment, which can finally improve template based three-dimensional protein structure predictors greatly.

Methods: CNNcon, an improved multiple neural networks based contact map predictor using six sub-networks and one final cascade-network, was developed in this paper. Both the sub-networks and the final cascade-network were trained and tested with their corresponding data sets. While for testing, the target protein was first coded and then input to its corresponding sub-networks for prediction. After that, the intermediate results were input to the cascade-network to finish the final prediction.

Results: The CNNcon can accurately predict 58.86% in average of contacts at a distance cutoff of 8 Å for proteins with lengths ranging from 51 to 450. The comparison results show that the present method performs better than the compared state-of-the-art predictors. Particularly, the prediction accuracy keeps steady with the increase of protein sequence length. It indicates that the CNNcon overcomes the thin density problem, with which other current predictors have trouble. This advantage makes the method valuable to the prediction of long length proteins. As a result, the effective prediction of long length proteins could be possible by the CNNcon.

Citation: Ding W, Xie J, Dai D, Zhang H, Xie H, et al. (2013) CNNcon: Improved Protein Contact Maps Prediction Using Cascaded Neural Networks. PLoS ONE 8(4): e61533. doi:10.1371/journal.pone.0061533

Editor: Bin Xue, Uni. of South Florida, United States of America

Received December 17, 2012; **Accepted** March 11, 2013; **Published** April 23, 2013

Funding: This work was supported by Key Project of Science and Technology Commission of Shanghai Municipality [No.11510500300] (URL: http://www.stcsm.gov.cn), Shanghai Leading Academic Discipline Project [No.J50103] (URL: http://www.shec.edu.cn), Innovation Program of Shanghai Municipal Education Commission [No. 11YZ03] (URL: http://www.shec.edu.cn) and Ph.D. Programs Fund of Ministry of Education of China [No. 20113108120022] (URL: http://www.cutech.edu.cn). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wzhang@shu.edu.cn

Introduction

It is well known that discovering the three-dimensional (3D) structure of a protein can provide important clues to understand of the mechanism of protein functions. Unfortunately, determination of 3D protein structure through experimental methods, such as X-ray crystallography or NMR spectroscopy, are time consuming and not working effectively with all kinds of proteins, especially membrane proteins [1]. Additionally, there are more than 24 million protein sequences in UniPortKB [2] currently, among which only about 84,508 proteins have had their structures solved experimentally [3]. Furthermore, almost 10,000 entries are newly added into Protein Data Bank (PDB) yearly [3]. That means more than 2,400 years are needed to solve the currently existed protein structures through experimental methods, under the situation of current experimental technology and no more newly discovered proteins. In fact, the number of newly discovered sequences grows much faster than the number of structures solved with experimental methods. The computation method is obviously the only way to bridge the huge protein sequence-structure gap.

Although many 3D protein structure predictors (3D-JIGSAW [4], I-TASSER [5], LOMETS [6], MODELLER [7], MODWEB [8], ROBETTA [9], SWISS-MODEL [10] and so on) with different accuracies have been developed in recent years, few predictors can produce desirable resolution structures for applications in medicine, such as drug design. The latest CASP experiment [11] shows that the progress has slowed and even reaches the bottleneck in direct prediction from one-dimensional (sequence) to three-dimensional (structure). With such difficulties, residue contact maps (CM or residue-residue contact) prediction, a matrix representation of protein residue-residue contacts, is the most promising one among recently developed prediction ideas.

Table 1. Performance of sub-networks and final cascade-network.

	Separation[a]	Seq-Len[b]	THR[c]	Chains[d]	Acc[e]	Err_{acc}[f]	Cov[e]	Err_{cov}[f]
Sub-network 1	6	51–70	0.1	30	46.43	9.24	41.91	4.89
Sub-network 2	7	71–90	0.6	40	44.35	9.89	36.43	7.64
Sub-network 3	10	91–130	0.7	199	43.99	8.05	36.33	4.59
Sub-network 4	13	131–190	0.7	246	41.38	8.39	33.95	5.32
Sub-network 5	17	191–290	0.8	201	28.31	7.72	36.57	2.64
Sub-network 6	21	291–450	0.9	87	31.81	9.90	34.47	2.53
Average					**34.01**	8.87	**35.44**	4.60
CNNcon		51–450		803	**57.86**	8.07	**34.28**	4.52

[a]Sequence separation: if value is *s*, then only contacts between pairs *i,j* minimally *s* residues apart are considered, that is $|i-j| \geq s$.
[b]Length Range of protein sequence of corresponding sub-network training and testing data sets.
[c]Minimal prediction value to determine residues contact or not.
[d]Size of test data set for each sub-network.
[e]Acc: prediction accuracy(%), defined in equation (1) and Cov: coverage(%), defined in equation (2).
[f]Standard error.
doi:10.1371/journal.pone.0061533.t001

CM of a protein is a simplified version of the protein structure and provides a new avenue for predicting 3D protein structure [12]. As these two-dimensional representations capture all the important features of a protein fold, the whole complex and difficult 3D structure prediction task can be divided into two steps. That is solving the one-dimensional to two-dimensional prediction firstly and then the final two-dimensional to three-dimensional prediction. This idea of divide and conquer makes the problem much easier and also help reconstruct final 3D structure from predicted contact maps. Protein CM has some advantages below. First, CM conveys strong information about the 3D protein structure. Second, the binary CM nature can be regarded as a classical problem of a two-state classification which has been thoroughly studied. Third, it has been shown that the empirical reconstruction algorithms are quite insensitive to high levels of random noise in CM, so that it is not necessary to predict all contacts correctly for reconstructing the protein 3D structure [12–15]. So far, several contact maps prediction methods, such as NNcon [16], PROFcon [17], SVMcon [18], RECON [19], CMWeb [20] and CMAPpro [21], have been developed successfully.

An improved multiple neural networks based contact map predictor, CNNcon, was proposed in this paper. It's composed of six input sub-networks and one output network, which forms a two-level cascaded network architecture. All the networks used are standard back-propagation neural networks. For network inputs, different sources of information were mixed and most of them had been used separately in some way before.

Results and Conclusion

Assessment of the Prediction Efficiency

To score the efficiency of the CNNcon method, two widely used and accepted statistical indices are introduced. Here, we only sketch these scores that are described in detail in [22–25].

The first and most frequently used one is accuracy, also referred to as 'Specificity', defined as follows:

$$Acc = \frac{N_{cp}*}{N_{cp}} = \frac{TP}{TP+FP} \quad (1)$$

where $N_{cp}*$ and N_{cp} are the number of correctly assigned contacts and that of total predicted contacts respectively. They also correspond to the sum of true positives (TP) and the sum of both TPs and false positives (FT) respectively. Routinely the accuracy is evaluated for each test protein and then averaged over the protein set.

We also evaluate the performance on the coverage of correct predicted contacts, also referred to as 'Sensitivity', defined as:

$$Cov = \frac{N_{cp}*}{N_{obs}} = \frac{TP}{TP+FN} \quad (2)$$

where $N_{cp}*$ is the same in equation (1) and N_{obs} is the number of observed contacts, which corresponds to the sum of TPs and false negatives (FN).

Results

Table 1 gives the prediction results of sub-networks and the final cascade-network, respectively. Two conclusions follow from these results. First, the prediction accuracy of each sub-network alone is comparable to other neural network based methods [16–18], whose performance is showed in Table 2. It indicates that our idea of assigning different prediction tasks to specific sub-networks corresponding to the protein length is practicable. Second, the remarkable improvement of accuracy from final cascade-network with little coverage loss proves that the CNNcon method is extremely effective and valuable.

In general, it is neither straightforward nor completely fair to compare the performance of different contact map predictors. First, different predictors are usually suitable for different length range proteins. Second, there also not existed a benchmark data set big enough and accepted widely. Therefore, the comparisons with other current contact map predictors in Table 2 are used for reference. The results show that the CNNcon method achieves the best accuracy and the coverage is the second best, which is almost as good as the best one. Moreover, the largest test data set is used in order to make the present results reliable.

To further verify the performance of the CNNcon method, we applied all the compared methods on the same test data set, 64 CASP10 targets. This test data set contains all the targets with length from 51 to 450 and valid PDB codes. Since different methods predict different number of contacts, in order to correctly

Table 2. Comparison results with other current methods.

Predictor	Acc[e]	Cov[e]	Targets[g]	Method
CNNcon[h]	57.86	34.28	803	Neural network based; Using optimized thresholds.
NNcon[i]	54.50	35.00	116	Neural network based; Top *L*/5 predicted.
PROFcon[j]	32.40	19.60	633	Neural network based; Top *L*/2 predicted.
SVMcon[k]	37.00	21.00	48	Support vector machine based; Top *L*/5 predicted.

[e]As in Table 1.
[g]Size of test data sets.
[h]This work.
[i,k]Results are summarized from previous works [16–18], respectively.
doi:10.1371/journal.pone.0061533.t002

compare them, n predicted contacts with the highest probabilities are selected. To increase the comparison preciseness, instead of being assigned one value, n was assigned to $T/2$, $2T/3$ and T, respectively, where T was the total true contacts of the whole test data set. Then the final compared statistical indices take the average values. The details of the compared results are given in Table 3. Both accuracy and coverage of the present method are better than others.

The prediction accuracies upon all proteins in the six test sets by corresponding sub-networks are shown in Figure 1. Clearly, accuracies decrease sharply while protein sequence length increases owing to the density of contacts decreasing greatly as the inverse of the protein length [12,26]. This also troubles most other current contact predictors. However, the prediction accuracies from the present method almost keep the same with the increase of protein sequence length in Figure 2. That means the CNNcon method overcomes the thin density problem [12,26], which suggests that it might be a valuable candidate for long length protein prediction.

Conclusion

An improved neural network based approach for protein contact map prediction, called CNNcon, was developed in this paper. The method performs better on prediction accuracy than other compared state-of-the-art methods. Further, the CNNcon method has better consistency and stability on prediction accuracy as protein length increases. Although training the six sub-networks and one cascade network costs computationally more than single-network predictors, it is one-time work. While in testing, the CNNcon method can divide the contact map prediction task naturally and run in parallel, based on the specially designed architecture. This advantage makes the method almost as fast as other single network based methods. It is expected that the CNNcon will be used to enhance parallel performance with longer protein length. As the neural network can be improved by adding more input information and training with a larger training data set, next work will be focus on combining more input information (e.g. correlated mutation information) and adding more protein chains to training data set. Parallel version of the CNNcon algorithm will also be implemented and worked on super-computers in the future.

Discussion

Optimized Thresholds were Crucial for Performance

Table 1 (Column 'THR') gives the optimized thresholds for all sub-networks. They are minimal prediction values to determine

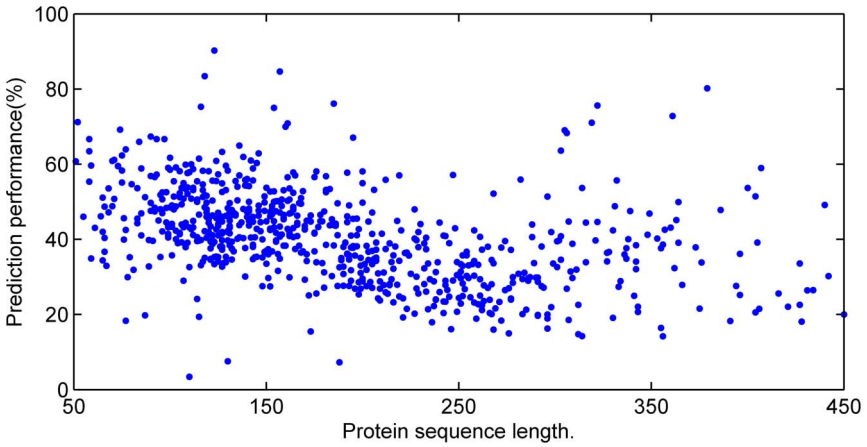

Figure 1. Prediction results upon all test proteins by corresponding sub-networks. The X axis is length range of tested proteins. The Y axis is prediction accuracy (%). Each point represents the predicted accuracy of a protein by its belonged sub-network. The average accuracy is as high as 34.01%. However, the accuracies decrease while the length of proteins increases.
doi:10.1371/journal.pone.0061533.g001

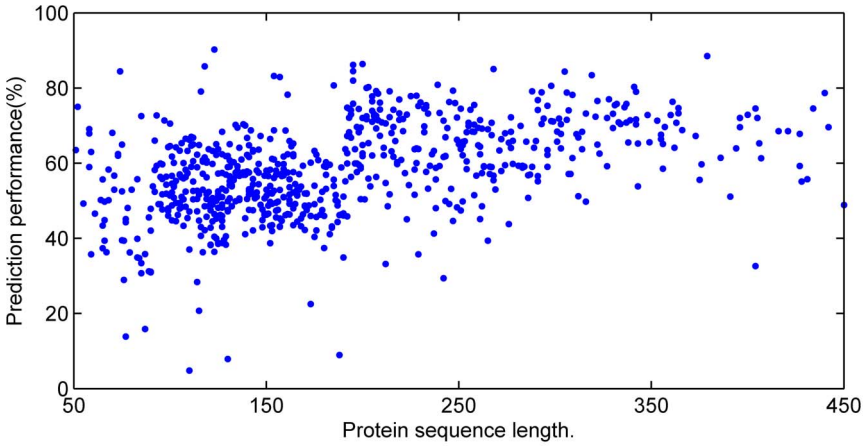

Figure 2. Prediction results upon all test proteins by the final cascade-network. The X axis and Y axis are the same in Figure 1. Each point represents the predicted accuracy of a protein by the final cascade-network. The average accuracy is as high as 57.86%. Moreover, the accuracies keep steady while the length of proteins increases.
doi:10.1371/journal.pone.0061533.g002

residues contact or not for corresponding sub-networks. Different thresholds resulting in both different accuracies and different coverages are found. And different sub-networks have their own optimized thresholds. This was probably related to the different contact densities of different protein length ranges, according to which the sub-networks were introduced. Further, it is discovered that the coverage score dropped sharply while the threshold was once greater than a specific value. These specific values were used as our final optimized thresholds for the corresponding sub-networks.

Combining and Balancing Multiple Predictions Improves Accuracies

As expected, the prediction accuracies of sub-networks are at the same level of most single neural network based methods. However, the final prediction accuracy is improved greatly by our cascade-network because of the following two advantages of our model. First, instead of being processed by a single network, each test protein was input to its corresponding sub-network, left-next sub-network and right-next sub-network for prediction in parallel. This increases the opportunity of contacted amino acids to be found. Second, three optimized balancing weights were introduced to balance the predicted results of sub-networks during final cascade-network prediction.

Table 3. Comparison results on 64 CASP10 targets.

Predictor	Acc[e]	Err_{acc}[f]	Cov[e]	Err_{cov}[f]
CNNcon	55.48	17.13	36.89	4.79
NNcon	46.39	11.79	31.70	9.49
PROFcon	39.90	7.02	25.55	9.87
SVMcon	38.15	9.02	25.62	10.93

[e,f]As in Table 1.
doi:10.1371/journal.pone.0061533.t003

Materials and Methods

Contact Map Definition

The contact map of a protein with N amino acids is an $N \times N$ binary symmetric matrix $C_{N\times N}$. The components $C_{ij}(i<j)$ are defined as follows:

$$C_{ij} = \begin{cases} 1 & \text{if amino acid } i \text{ and } j \text{ are in contact} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

We define two amino acids as being in contact if the distance between their C_β atoms (C_α for glycines which having a hydrogen substituent as its side-chain) is less than 8 Å, a standard threshold widely used [12,22,27–29].

Neural Network Architecture

The finding that number of contacts in a protein is proportional to the protein length N, while the number of possible contacts increases with $N(N-1)/2$ [12], implies the contact densities in the map decrease as the inverse of the protein length. In other words, long proteins have lower contact densities than short ones [26]. This makes the contact maps of long proteins more difficult to predict and the prediction accuracy is affected by the protein length greatly. Six specific sub-networks for different protein length range respectively and one cascade-network are introduced in order to solve this problem. They are all classical feed-forward 3-layer neural networks trained with the same standard back-propagation algorithm [30]. Architectures of all the six sub-networks are the same and composed of 1747 input nodes, 5 hidden nodes and 1 output node. The cascade-network contains 9 input nodes, 6 hidden nodes and 1 output node. The numbers of middle nodes are actually decided by repeated trials in experiment depending on the balance of computation time and prediction accuracy. Same values are assigned to the number of middle nodes of all sub-networks. In fact, it might be more suitable to assign the parameter of each sub-network with its different and specific values, since each sub-network is designed for proteins with different lengths. In next improved version of CNNcon (v2.0, also parallel and super computer version), this work will be considered and these optimal values of this parameter will be picked out through experiments performed on each sub-network. Among

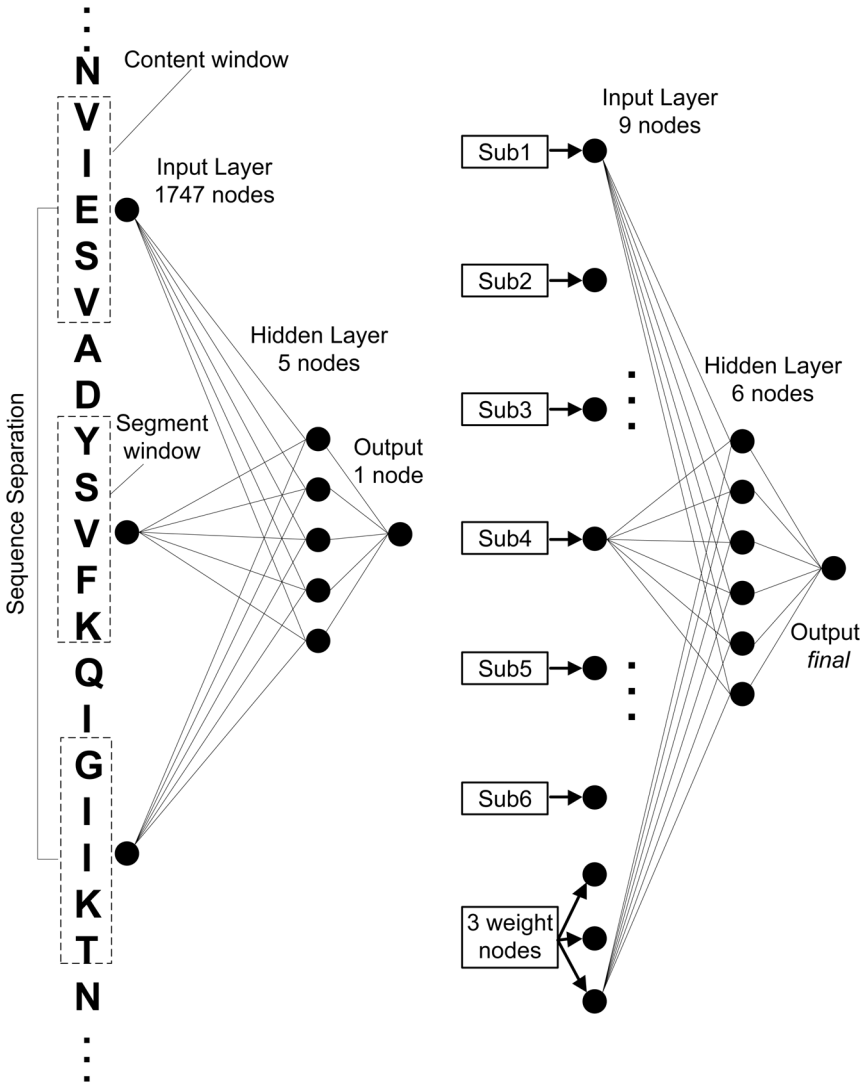

Figure 3. Architectures of sub-neural network (left) and cascade-neural network (right). Since architectures of all the six sub-networks are the same, only one of them is shown here (left).
doi:10.1371/journal.pone.0061533.g003

these input nodes, six are coded by prediction results from sub-networks and the remaining three are coded by balanced weights. The whole architecture of the CNNcon method is shown in Figure 3.

Each sub-network was trained and tested with its corresponding data set. The data sets and length range divisions are mentioned in section of data sets below. While for testing, the target protein was coded first and input to its corresponding sub-networks for prediction. We defined the sub-network id as 1 to 6 as increase of its length coverage and the particular sub-network with length range covering the target protein length defined as i. Thus for each target protein, its corresponding sub-networks were $i-1$, i and $i+1$, that is just local communication needed while in parallel. After prediction by its corresponding sub-networks, the intermediate results along with three optimized balance factors were input to the cascade-network to finish the final prediction.

Input Codings

The basic input coding method used here is the same as previously introduced in [31]. Each residue pair is characterized by an vector containing 210 elements $(20\times(20+1)/2)$, representing all the possible ordered couples of residues. The input coding vectors of each residue couple and its symmetric ones are the same.

In the method, multiple sequence information instead of single sequence was used, since evolutionary information had been proved to improve prediction performance greatly [17]. Multiple sequence alignment information of each protein sequence was gained from its corresponding HSSP file [32]. Considering the prediction performance and our computing resource, we chose as most as 100 multiple sequence alignment sequences (including the target one) with the identity of each aligned sequence less than 80%.

For each sequence in the alignment, a pair of residues in position i and j were counted. The final input coding, representing the frequency of each pair in the alignment, was normalized to the number of the aligned sequences [31].

Conservation weights and secondary structures [33] information from HSSP file were also coded with one and three elements

respectively. Thus the length of the input coding vector becomes 218 $(210+(1+3)\times 2)$.

To obtain local information of each residue, similar to [17], we used two content windows of size 2 centered around i and j (window of i: $\{i-2,i-1,i,i+1,i+2\}$, window of j: $\{j-2,j-1,j,j+1,j+2\}$) respectively. That means that, for each residue pair $\{i,j\}$, we incorporated information from all residues in those two windows of five consecutive residues. Thus, the length of the input coding vector was increased to 1090 (218×5).

Further, we introduced a segment window with size of 2 to code information from the segment connecting i and j. For each residue pair $\{i,j\}$, we incorporated information from all residues in the window centered around k $(k=i/2+j/2)$, which was the middle position of i and j. Thus, the segment window spanned the interval $\{k-2,k-1,k,k+1,k+2\}$ and the length of our input coding vector again was added to 1744 $(1090+218\times 3)$.

Finally, we used sequence separation, sequence length and segment separation length to represent the global information from the entire protein. The size of our input coding vector was lastly set to 1747 $(1744+3)$.

Data Sets

Data set used here for training and testing was extracted from the March 2012 25% pdb_select list [34–37] with 5,300 chains and 788,447 residues.

For the goal of algorithm design, we removed all protein chains of non-X-ray determined structures, all chains with resolution greater than 1.5 Å, all backbone broken chains (contain missing backbone atoms in the PDB files), all chains containing non-standard residues in its corresponding PDB files and all chains with obsolete PDB ID (e.g. 3G62 is obsolete and replaced by 4F1U). We reduced the data set further by excluding all protein chains longer than 450 residues. Without loss of generality, all chains shorter than 51 residues were removed as well. After above processing, our final data set contains 1,103 chains (1,082 proteins) and 192,640 residues.

As prediction performances greatly depend on protein length distribution, here we give the protein length distribution of data set to make assessment more reasonable. 7.25% of the proteins have a length from 51 to 70 residues (sub-network 1), 8.16% comprise from 71 to 90 residues (sub-network 2); 22.57% from 91 to 130 residues (sub-network 3); 26.84% from 131 to 190 residues (sub-network 4); 22.76% from 191 to 290 residues (sub-network 5); 12.42% from 291 to 450 residues (sub-network 6). These distributions are also the partitions of length range coverage of sub-networks. That's also why six sub-networks are needed in the CNNcon method. The data set was split into six subsets according to the above length range distributions. Each sub-network was trained with 50 samples randomly selected from its corresponding data set and tested by the remaining. We used all the six test subsets (803 protein chains in total) to test the final cascaded network.

Balanced Training

To address the extreme disproportion distribution of true (contacts) and false (non-contacts) samples during the training phase, we used balanced training to reduce back-propagation learning cycles [38]. A balancing probability factor was also introduced to further reduce the false samples and the whole training data set size in a random way.

Acknowledgments

The authors are grateful to CMBuilder, a small tool calculating contact maps from PDB files. Thanks to all the bioinformatics group members at Shanghai University for their useful discussion and previous significant research work [39–42].

Author Contributions

Conceived and designed the experiments: WD JX DD HZ HX WZ. Performed the experiments: WD DD. Analyzed the data: WD JX DD HZ HX WZ. Contributed reagents/materials/analysis tools: WD JX DD HZ HX WZ. Wrote the paper: WD DD WZ.

References

1. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL (1994) Knowledge-based protein modeling. Crit Rev Biochem Mol Biol 29: 1–68.
2. UniProtKB Protein Database. Available: http://www.uniprot.org/. Accessed 2012 Sep 11.
3. Protein Data Bank. Available: http://www.rcsb.org/. Accessed 2012 Sep 11.
4. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins: Structure Function and Genetics 455: 39–46.
5. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5: 725–738.
6. Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 35: 3375–3382.
7. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci Chapter 2: Unit 29.
8. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, et al. (2003) Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 31: 3375–3380.
9. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32: W526–W531.
10. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22: 195–201.
11. Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)-round IX. Proteins: Structure, Function, and Bioinformatics 79: 1–5.
12. Bartoli L, Capriotti E, Fariselli P, Martelli PL, Casadio R (2008) The pros and cons of predicting protein contact maps. Methods Mol Biol 413: 199–217.
13. Vendruscolo M, Domany E (2000) Protein folding using contact maps. Vitam Horm 58: 171–212.
14. Fariselli P, Olmea O, Valencia A, Casadio R (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins: Structure Function and Genetics 45: 157–162.
15. Vassura M, Di Lena P, Margara L, Mirto M, Aloisio G, et al. (2011) Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3D structure. BioData Min 4: 1.
16. Tegge AN, Wang Z, Eickholt J, Cheng J (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Research 37: W515–W518.
17. Punta M, Rost B (2005) PROFcon: novel prediction of long-range contacts. Bioinformatics 21: 2960–2968.
18. Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 8: 113.
19. Kundrotas P, Alexov E (2006) Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics 7: 503.
20. Kozma D, Simon I, Tusnady GE (2012) CMWeb: an interactive on-line tool for analysing residueresidue contacts and contact prediction methods. Nucleic Acids Res 40: W329–W333.
21. Di Lena P, Nagata K, Baldi P (2012) Deep architectures for protein contact map prediction. Bioinformatics 28: 2449–2457.
22. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins: Structure Function and Genetics 18: 309–317.
23. Olmea O, Rost B, Valencia A (1999) Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 293: 1221–1239.
24. Eyrich VA, Przybylski D, Koh IY, Grana O, Pazos F, et al. (2003) CAFASP3 in the spotlight of EVA. Proteins: Structure Function and Genetics 53: 548–560.
25. Monastyrskyy B, Fidelis K, Tramontano A, Kryshtafovych A (2011) Evaluation of residue-residue contact predictions in CASP9. Proteins: Structure, Function, and Bioinformatics 79: 119–125.
26. Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. Protein Eng 14: 835–843.

27. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18: 534–552.
28. Lund O, Frimand K, Gorodkin J, Bohr H, Bohr J, et al. (1997) Protein distance constraints predicted by neural networks and probability density functions. Protein Eng 10: 1241–1248.
29. Galaktionov S, Nikiforovich GV, Marshall GR (2001) Ab initio modeling of small, medium, and large loops in proteins. Biopolymers 60: 153–168.
30. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323: 533–536.
31. Fariselli P, Casadio R (1999) A neural network based predictor of residue contacts in proteins. Protein Eng 12: 15–21.
32. Dodge C, Schneider R, Sander C (1998) The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res 26: 313–315.
33. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
34. PDBselect-selection of a representative set of PDB chains. Available: http://bioinfo.mni.th-mh.de/pdbselect/.
35. Hobohm U, Scharf M, Schneider R, Sander C (1992) Selection of representative protein data sets. Protein Sci 1: 409–417.
36. Hobohm U, Sander C (1994) Enlarged representative set of protein structures. Protein Sci 3: 522–524.
37. Hobohm U, Griep S (2010) PDBselect 1992–2009 and PDBfilter-select. Nucleic Acids Research 38: D318–D319.
38. Rost B, Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proceedings of the National Academy of Sciences 90: 7558–7562.
39. Ding W, Dai D, Xie J, Zhang H, Zhang W, et al. (2012) PRT-HMM: A novel hidden Markov model for protein secondary structure prediction. In: Proceeding of 11th International Conference on Computer and Information Science. IEEE Computer Society, 207–212.
40. Xie J, Wang M, Dai D, Zhang H, Zhang W (2012) A network clustering algorithm for detection of protein families. In: Proceeding of the International Conference on Engineering in Medicine and Biology Society. 6329–6332.
41. Xie J, Yi R, Tan J, Cheng X, Dai D, et al. (2011) Multi-database retrieval technology on CPSE-Bio. In: Proceeding of International Conference on Computer Sciences and Convergence Information Technology. IEEE Computer Society, 380–384.
42. Chen J, Song A, Zhang W (2012) Hybrid k-harmonic clustering approach for high dimensional gene expression data. Journal of Convergence Information Technology 7: 39–49.