Amino acid torsion angles enable prediction of protein fold classification

Kun Tian
Renmin University of China, School of Mathematics
Xin Zhao
Tsinghua University Department of Mathematical Sciences
Xiaogeng Wan
Tsinghua University Department of Mathematical Sciences
Stephen Yau (yau@uic.edu)
Tsinghua University https://orcid.org/0000-0001-6639-0054

Research article

Keywords: protein structure, torsion angle, prediction, fold classification, distance

Posted Date: January 9th, 2020

DOI: https://doi.org/10.21203/rs.2.20475/v1

Abstract

Background

Protein structure can provide insights that help biologists to predict and understand protein functions and interactions. However, the number of known protein structures has not kept pace with the number of protein sequences determined by high-throughput sequencing. Current techniques used to determine the structure of proteins, such as X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy, are complex and may require a lot of time to analyze the experimental results, especially for large protein molecules. The limitations of these methods have motivated us to create a new approach for protein structure prediction.

Results

Here we describe a new approach that uses integration and analysis of torsion angle information from the Protein Data Bank to enable prediction of protein structures from amino acid sequences. Our prediction model performed well in comparison with previous methods when applied to the structural classification of two CATH datasets with more than 5000 protein domains. This new prediction model performs well with an average of 92.5% accuracy for structure classification, which is higher than the previous research. We also used our model to predict four known protein structures with a single amino acid sequence, while many other existing methods could only obtain one possible structure for a given sequence.

Conclusions

The results show that our method provides a new effective and reliable tool for protein structure prediction research.

Background

Resolution of protein three-dimensional structure is one of the most important research problems in the field of structural biology. The structure of a protein is directly related to its function, and structural prediction is an important goal of bioinformatics and theoretical chemistry, with great potential benefits in the fields of medicine and biotechnology. Hence, how to predict three-dimensional structures from protein sequences has been an unsolved and significant problem. Although amino acid sequences determine protein structures, other factors also contribute to structural modification, which demands us find an efficient technique to delineate the global properties of protein structure space [1–4]. Current techniques for the determination of protein structures include X-ray crystallography, nuclear-magnetic-resonance (NMR) spectroscopy, structure alignment and so on. With modern machine learning methods such as neural networks and support vector machines, some of these new methods also appear in protein structure prediction work [5–18]. For example, Chou develops methods that make allowance for taking into account the coupling effect among different amino acid components of a protein by a covariance

matrix [8, 9]. Brevern defines a structural alphabet, which allows the local approximation of the 3D protein structure by using a Bayesian approach based on the relation protein block amino acid propensity [11]. Wood provides a method called DESTRUCT using a sequence and structure representation and an iterative prediction algorithm [12]. Jung has created a web server providing structural information and analysis based on the backbone torsional representation of a protein structure [13]. More and more structure prediction software tools have appeared recently, including homology modeling, protein threading, ab initio methods, secondary structure prediction, transmembrane helix and signal peptide prediction, such as RaptorX [19], I-TASSER [20], HHpred [21]. However, these methods often require time-consuming analysis of experimental results, especially for large protein molecules which make them unreliable and ineffective for structure prediction. Thus, the speed of computation and accuracy still have room for improvement. As we know, there are many examples of proteins which have the same amino acid sequence but different structures. Beside of this, many existing methods may have limitations and drawbacks for predicting the structures of these kinds of sequences since these tools only obtain the most likely possible structure for each sequence. Therefore, it is necessary to develop a more accurate, fast and effective method to delineate the relationship between sequence code and structure space.

Here, we have therefore attempted to develop a methodology that uses primary amino acid sequence information to make a precise and effective prediction of the possible structures for a particular protein, and to visualize the comparison between the native structure and the predicted structure. Our method is based on the integration and analysis of torsion angle information from the Protein Data Bank (PDB) database, which contains information from over 10 million torsion angles. By taking into account the torsion angles between protein sequences, our algorithm improves secondary structure prediction in general. It not only determines the class of the most likely structure for a given amino acid sequence, but it can also predict and model multiple structures of the same sequence, something many other software tools are not able to achieve this point. We performed our method and compared our results with previously published methods [8, 9, 22] for prediction of protein domain structures in two large CATH protein structure classification datasets [23]. The CATH database contains a hierarchical classification of protein domains on the basis of class (C), architecture (A), topology (T) and homologous superfamily (H). This new prediction method performed well with an average of 92.5% accuracy for structure classification, which is a great improvement than Rackovsky's previous research. The method was also applied to a single amino acid sequence to model four different known protein structures. We also used the RaptorX method to predict the structure of the same sequence and compared the results with our method. The precision and reliability of our results were verified by calculating the dissimilarity of the predicted and actual protein structures. We used both the root-mean-square deviation (RMSD) measure and the Yau-Hausdorff distance to calculate dissimilarity [24, 25]. The Yau-Hausdorff distance is a metric to measure the difference of two proteins of any lengths based on the three-dimensional coordinates of their atoms which does not need aligning and superimposing two structures [24, 25]. Our results demonstrate that this new approach is efficient and reliable on protein structure prediction, and can obtain multiple different structures for a same sequence, improve protein-folding recognition, classification of structural motifs and refinement of sequence alignment.

Results

Prediction of protein structures in the CATH dataset

We used our torsion angle method to predict the most likely structure of each protein domain in two subsets of the CATH dataset. The '59 CAT' group consisted of 59 CAT classes with at least 20 members (a total of 4319 sequences), whereas the '60 CAT' group consisted of 60 CAT classes with 10–19 members (a total of 821 sequences). For each protein domain, we regarded its predicted classification correct if the class of predicted structure was the same as that of the empirically determined one. The accuracy rate of this prediction was defined as the number of correct classifications divided by the total number of proteins that were classified. We compared our results with those of a previous study that used a 10-dimensional vector method to analyze protein secondary structure [22]. We also applied the methods developed by Chou on the same dataset [8, 9]. Complete results are shown in Table 1. From this table, we can find that the accuracies by our method are higher than the other methods, which indicate our torsion angle method performs as well or better than the previous method for prediction of all the domain categories.

Table 1

Comparison about the accuracies of different methods for prediction of protein domain structures.

	60 CAT group accuracies			59 CAT group accuracies		
Class	C = 1	C = 2	C = 3	C = 1	C = 2	C = 3
Protein numbers	195	145	481	762	1220	2337
Torsion angle method	87%	87%	96%	94%	97%	94%
10-dimensional vector method [22]	66%	56%	73%	92%	97%	93%
Method in [8]	50%	77%	90%	47%	76%	60%
Euclidian distance method [9]	74%	59%	61%	67%	69%	60%
Hamming distance method [9]	72%	54%	61%	62%	66%	61%

Each group is divided into alpha structure (C = 1), beta structure (C = 2) and mixed structure (C = 3) classes.

Prediction of multiple protein structures from a single amino acid sequence

Our method was tested by analysis of a 148 amino acid sequence, to predict four known protein structures (1a29, 1cfd, 1cll and 2bcx) based on this sequence. We first checked the locations for each of the 142 heptamers appeared in the 96501 reliable protein structures database and collected the torsion angle points associated with the central amino acid of the heptamer. The torsion angles of the 78th heptamer is shown in Fig. 1 as an example, and detailed steps for constructing the four predicted

structures corresponding the four proteins with the same sequence are explained in the Methods section. The alignments between the known and predicted protein structures using our method are shown in Fig. 2. The Yau–Hausdorff distances and RMSD values for each pair of structures are also calculated in Table 2. We also used the RaptorX method to predict the protein structure for this amino acid sequence. It could only provide one most likely structure which performed not well in predicting multiple structures for specific sequence. The Yau–Hausdorff distances and RMSD values between the constructed structure and the four known ones are listed in Table 2, and the alignments are shown in Fig. 3. In Table 2, both Yau-Hausdorff distance and RMSD measure between each of the constructed structure performed by our method and the empirically determined one is smaller than that of RaptorX method. It also indicates that the predicted structures of our method are more similar than RaptorX by comparing Figs. 2 and 3. Although the predicted and known protein structures do not completely overlap by our method that is probably because the torsion angles of the predicted structure are not the same as the empirically determined one, the distances are small enough (with the diameter of every structure being larger than 50 angstrom) to indicate that each pair of structures is similar, demonstrating that this methodology can predict empirically determined structures from a specific amino acid sequence.

Table 2

Yau–Hausdorff distances and RMSD values between the empirically determined and predicted structures of the four proteins with the same amino acid sequence using our method and RaptorX method.

Protein ID	Yau–Hausdorff distance by our method	Yau–Hausdorff distance by RaptorX method	RMSD by our method	RMSD by RaptorX method
1a29	1.901	5.830	3.704	14.929
1cfd	2.574	2.654	4.786	6.782
1cll	1.124	4.295	3.330	11.62
2bcx	2.743	2.899	5.821	12.221

Discussion

Structural dynamics of proteins with the same sequence

One most significant potential application of our method is it could be applied to predict the structure of a sequence for which there is no prior structural information. Given a protein sequence without structural information, we can predict the most likely structure for it. Another potential application may be used in structural dynamics. As the four known protein structures all correspond to the same amino acid sequence, it is possible that each structure could transform into one of the other structures. As described above, all possible torsion angles for each heptamer are calculated, enabling construction of all possible structures of the sequence. The dynamic process of transformation between protein structures with the same amino acid sequence is able to be constructed based on these possible structures. The

transformation among the predicted structures can be ordered according to a metric, such as minimize the Yau–Hausdorff distances, beginning with one known structure and finishing with the other. The further in-depth study will discuss the structural dynamics.

Conclusions

With the continuing development of sequencing technologies, methods are required for prediction of protein structures from amino acid sequences. In this study, we have provided an unsupervised method for protein structure prediction and constructing structures using the amino acid sequence via integrating and analyzing large torsion angle information in the Protein Data Bank. We reconstruct the structures of four proteins with the same sequence and compare the result with that obtained by RaptorX method, which is only able to predict one possible structure for a given sequence. One can clearly view the similarity comparison and calculate the value using different kinds of scores, such as the Yau-hausdorff distance [24, 25], RMSD, and TM-score between the native structure and constructed structure, then verify the precision of our method. It is able to generate multiple structures according to the amino acid sequence as well as provide a most likely structure to determine the property of the protein sequence. The new prediction model performs well, with an average of 92.5% accuracy for structure classification on two large CATH datasets, which makes a great improvement than many other methods [8, 9, 22]. This demonstrates our method is efficient and reliable on protein structure prediction study.

Methods

Datasets

To determine the possible torsion angles of the central residues of amino acid heptamers, 96501 reliable protein structures were downloaded from the PDB to provide a structure database (see Additional file 1).

The CATH database contains sequence and structure information for a large number of protein domains, organized hierarchically by class, architecture, topology and homology. Our method was compared with previous methods for its ability to predict the class assignment of two groups of protein domains, as defined previously [8, 9, 22]. We used the same dataset as that of Rackovsky's [22]. The classes are: 'C = 1', α-helical structures; 'C = 2', β-sheet/barrel structures; and 'C = 3', mixed α/β structures. After deleting the sequences with fewer than 60 amino acids from the CathDomainSeqs.S35.ATOM.v3.1.020 database and restricting our attention to the CAT classes, the '59 CAT' group consisted of 59 CAT classes with at least 20 members (a total of 4319 domain sequences), whereas the '60 CAT' group consisted of 60 CAT classes with 10–19 members (a total of 821 domain sequences).

Determination of torsion angle clusters

For each sequence S of length N in the CAT groups, the N − 6 possible amino acid heptamers are determined. For example, the nonameric sequence 'CGDYAHCKS' has three heptamers 'CGDYAHC', 'GDYAHCK' and 'DYAHCKS'. It is a common sense that the first three neighboring amino acids have an

effect on the fourth amino acid torsion angles, therefore pentamers are not enough for determining the amino acid torsion angles. Although the first amino acid has effect on the fifth amino acid, it is weak, so the use of nonamers is not necessary. That is why heptamers are chosen for collecting the torsion angles information of amino acids.

For each heptamer of S, all occurrences in the structure database are identified, along with all pairs of torsion angles associated with the central amino acid of the heptamer. A pair of torsion angles can be treated as coordinates of a point in a plane. All identified torsion angle pairs for a heptamer's central amino acid are plotted in a plane. The most-dense cluster is determined by taking each integer point as a center to draw circles of the same size and choosing the center of the corresponding circle that has the highest number of torsion angles as the cluster. This process is repeated for each of the N − 6 heptamers in S.

Predicting the most likely protein structure

For each sequence S, protein structures are predicted with Pymol software on the basis of the most-dense clusters of torsion angle pairs for the N − 6 heptamers. The first cluster (for the first heptamer) represents the torsion angles between the fourth and fifth amino acids of S. In Pymol, the first cluster is used to set the torsion angles between these two amino acids. The second cluster represents the torsion angles between the fifth and sixth amino acids, and so on. With these torsion angles, the positions of each amino acid are fixed in Pymol, enabling prediction of the most likely structure of S.

Classification of protein structures

Two methods are used for determination of the classification which each constructed most likely protein domain structure belongs to. One approach uses the Definition of Secondary Structure of Proteins (DSSP) tool for standardization of secondary structure assignment [26]. DSSP is a software of secondary structure assignments for all protein domain structures entries. It is used for determining the classification of our prediction of the structure of the most likely protein domain by putting the predicted structure into the software and running the program directly.

A second approach uses the Ramachandran plot method to visualize energetically allowed regions for backbone dihedral angles ψ against φ of amino acid residues in protein structures [27]. Because dihedral-angle values are circular and − 180° is equal to 180°, the edges of the Ramachandran plot 'wrap' right-to-left and bottom-to-top. The regions where the majority of the torsion angles lie are different for each of the protein domain structure classes 'C = 1', 'C = 2' and 'C = 3'. For example, most of the torsion angles of protein domain structures in class 'C = 1' lie in the upper left side of the Ramachandran plot. Based on this location feature of the three classes, classifications of our predictions of the most likely protein domain structures are determined by identification of the regions in which most of the torsion angles are located in the Ramachandran plot.

Constructing multiple protein structures for a given sequence

Given an amino acid sequence S of length N, we can predict all possible structures for it. As described above, all occurrences of the torsion angles associated with the central amino acid of the N − 6 heptamers in sequence S are determined from the structure database at first. Not only the most-dense cluster is determined for predicting the structure, but also the second most-dense cluster is used as another choice for some heptamers with large number of appearance times in the structure database when constructing multiple structures for the sequence S. Among the whole possible structures constructed by these cluster points, the ones which have the minimum Yau-Hausdorff distance with the known structures are chosen as the multiple predicted structures for sequence S.

Yau-Hausdorff distance between protein structures

The Yau-Hausdorff distance is used to calculate the dissimilarity between protein structures here [24, 25]. Each protein structure is regarded as a three-dimensional point set consisting of all the atom coordinates.

Define the minimum one-dimensional Hausdorff distance of two finite point sets A_1 and B_1 in $\mathbb{R}$ as

$$H^1(A_1, B_1) = \min_{t\in\mathbb{R}} h(A_1 + t, B_1),$$

where h is the Hausdorff distance

$$h(A_1, B_1) = \max\left\{\max_{a\in A_1}\min_{b\in B_1} d(a,b), \max_{b\in B_1}\min_{a\in A_1} d(b,a)\right\},$$

here $d(a,b)$ is the Euclidean distance between two points a and b, and $h(A_1 + t, B_1)$ stands for the Hausdorff distance between A_1 and B_1 after shifting A_1 by t. The Yau-Hausdorff distance $D(A,B)$ of two point sets A and B in $\mathbb{R}^3$ is then defined in terms of H^1:

$$D(A,B) = \max\left\{\max_{\theta^2}\min_{\varphi^2} H^1\left(P_x\left(A^{\theta^2}\right), P_x\left(B^{\varphi^2}\right)\right), \max_{\varphi^2}\min_{\theta^2} H^1\left(P_x\left(A^{\theta^2}\right), P_x\left(B^{\varphi^2}\right)\right)\right\},$$

where $P_x\left(A^{\theta^2}\right)$ is a one-dimensional point set representing the projection of A on the x-axis after being rotated by three-dimensional rotation angle θ^2.

The Yau-Hausdorff distance is a natural metric which takes all possible translation and rotation into consideration for calculating the dissimilarity between protein structures. Comparing with aligning methods, the computational complexity has been reduced by projecting three-dimensional point sets into one-dimensional space in calculation without losing any information.

Abbreviations

NMR: nuclear magnetic resonance; PDB: protein data bank; CATH: class, architecture, topology and homologous superfamily; RMSD: root mean square deviation; DSSP: definition of secondary structure of proteins.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets supporting the conclusions of this article are included within the article and its additional file.

Competing interests

The authors declare that they have no competing interests.

Funding

This study is supported by the National Natural Sciences Foundation of China (91746119), Tsinghua University start up fund. The funders did not take part in study design; in collection and analysis of data; in the writing of the manuscript; in the decision to publish this manuscript.

Authors' contributions

SSTY conceived the ideas. XZ, KT and SSTY designed the methodology used; XZ and KT collected and analyzed the data; XZ, KT and SSTY led the writing of the manuscript. All authors contributed critically to the draft and gave final approval for publication.

Acknowledgements

The authors wish to thank Dr. Benson from Department of Computer Science, Seattle Pacific University for help with revising the manuscript, and the Department of Mathematical Science at Tsinghua University for providing the work space and library facilities.

References

1. Bu Z, Callaway D. Proteins move! Protein dynamics and long-range allostery in cell signaling. Advances in Protein Chemistry and Structural Biology. 2011;83:163-221. doi: 10.1016/B978-0-12-381262-9.00005-7.
2. Tokuriki N, Tawfik D. Protein dynamism and evolvability. Science. 2009;324:203-207. doi: 10.1126/science.1169375.
3. Frauenfelder H, Sligar S, Wolynes P. The energy landscapes and motions of proteins. Science. 1991;254:1598-1603.

4. Akinori K, Nobuhiro G. Refinement of protein dynamic structure: Normal mode refinement. Proceedings of the National Academy of Sciences. 1990;87:3718-3722. doi: 10.1073/pnas.87.10.3718.
5. Guzzo A. Influence of Amino-Acid Sequence on Protein Structure. Biophysical Journal. 1965;5:809-822. doi: 10.1016/S0006-3495(65)86753-4.
6. Prothero J. Correlation between Distribution of Amino Acids and Alpha Helices. Biophysical Journal. 1966;6:367-370. doi: 10.1016/S0006-3495(66)86662-6.
7. Schiffer M, Edmundson A. Use of Helical Wheels to Represent Structures of Proteins and to Identify Segments with Helical Potential. Biophysical Journal. 1967;7:121-135. doi: 10.1016/S0006-3495(67)86579-2.
8. Chou K. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function, and Genetics. 1995;21:319-344. doi: 10.1002/prot.340210406.
9. Chou K, Liu W, Maggiora G, Zhang C. Prediction and classification of domain structural classes. Proteins: Structure, Function, and Genetics. 1998;31:97-103. doi: 10.1002/(SICI)1097-0134(19980401)31:1<97::AID- PROT8>3.0.CO;2-E.
10. Dor O, Zhou Y. Achieving 80% Ten-fold Cross-validated Accuracy for Secondary Structure Prediction by Large-scale Training. Proteins: Structure, Function and Bioinformatics. 2007;66:838-845. doi: 10.1002/prot.21298.
11. Brevern A, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins: Structure, Function and Bioinformatics. 2000;41:271-287. doi: 10.1002/1097-0134(20001115) 41:3<271::AID-PROT10>3.0.CO;2-Z.
12. Wood M, Hirst J. Protein secondary structure prediction with dihedral angles. Proteins: Structure, Function and Bioinformatics. 2005;59:476-481. doi: 10.1002/prot.20435.
13. Jung S, Bae S, Ahn I, Son H. Protein Backbone Torsion Angle-Based Structure Comparison and Secondary Structure Database Web Server. Genomics and informatics. 2013;11:155-160. doi: 10.5808/GI.2013.11.3.155.
14. MacCallum J, Perez A, Dill K. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. Proceedings of the National Academy of Sciences. 2015;112:6985-6990. doi: 10.1073/pnas.1506788112.
15. Huang P, Boyken S, Baker D. The coming of age of de novo protein design. Nature. 2016;537:320-327. doi: 10.1038/nature19946.
16. Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. Scientific reports. 2016;6:1-11. doi: 10.1038/srep18962.
17. Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K, Zhou Y. Sixty-five years of the long march in protein secondary structure prediction: the final stretch. Briefings in bioinformatics. 2018;1:1-13. doi: 10.1093/bib/bbw129.

18. Söding J. Big-data approaches to protein structure prediction. Science. 2017;355:248-249. doi: 10.1126/science.aal4512.
19. Wang S, Li W, Liu S, Xu J. RaptorX-Property: a web server for protein structure property prediction. Nucleic Acids Research. 2016;44:W430-W435. doi: 10.1093/nar/gkw306.
20. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9:1-8. doi: 10.1186/1471-2105-9-40.
21. Söding J, Biegert A, Lupas A. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research. 2005;33:W244-W248. doi: 10.1093/nar/gki408.
22. Rackovsky S. Sequence physical properties encode the global organization of protein structure space. Proceedings of the National Academy of Sciences. 2009;106:14345-14348. doi: 10.1073/pnas.0903433106.
23. Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J. CATH — a hierarchic classification of protein domain structures. Structure. 1997;5:1093-1108. doi: 10.1016/S0969-2126(97)00260-8.
24. Tian K, Yang X, Kong Q, Yin C, He R, Yau S. Two dimensional Yau-Hausdorff distance with applications on comparison of DNA and protein sequences. PLoS ONE. 2015;10:e0136577. doi: 10.1371/journal.pone.0136577.
25. Tian K, Zhao X, Zhang Y, Yau S. Comparing protein structures and inferring functions with a novel three-dimensional Yau-Hausdorff method. Journal of Biomolecular Structure and Dynamics. 2019;37:4151-4160. doi: 10.1080/07391102.2018.1540359.
26. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577-2637. doi: 10.1002/bip.360221211.
27. Ramachandran G, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology. 1963;7:95-99. doi: 10.1016/S0022-2836(63)80023-6.

Additional File

Additional file 1.docx-This additional file contains the IDs of 96501 reliable protein structures used in our study.

Figures

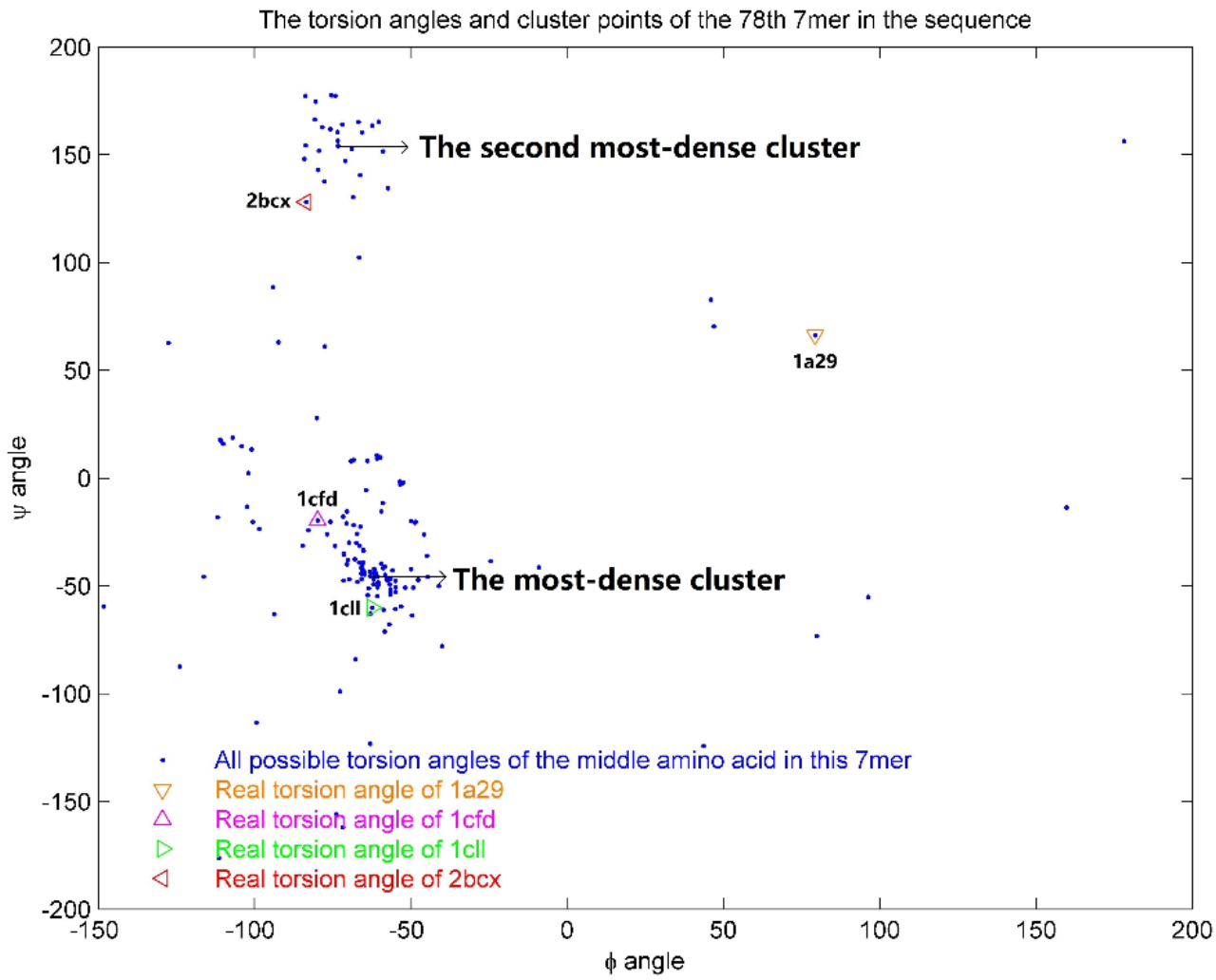

Figure 1

The torsion angles of the 78th amino acid heptamer in a sequence that results in four protein structures (1a29, 1cfd, 1cll and 2bcx). The blue points represent all possible torsion angles and the torsion angles corresponding to each of the four protein structures are indicated. The most-dense cluster and second most-dense cluster used for constructing the predicted structures are pointed out.

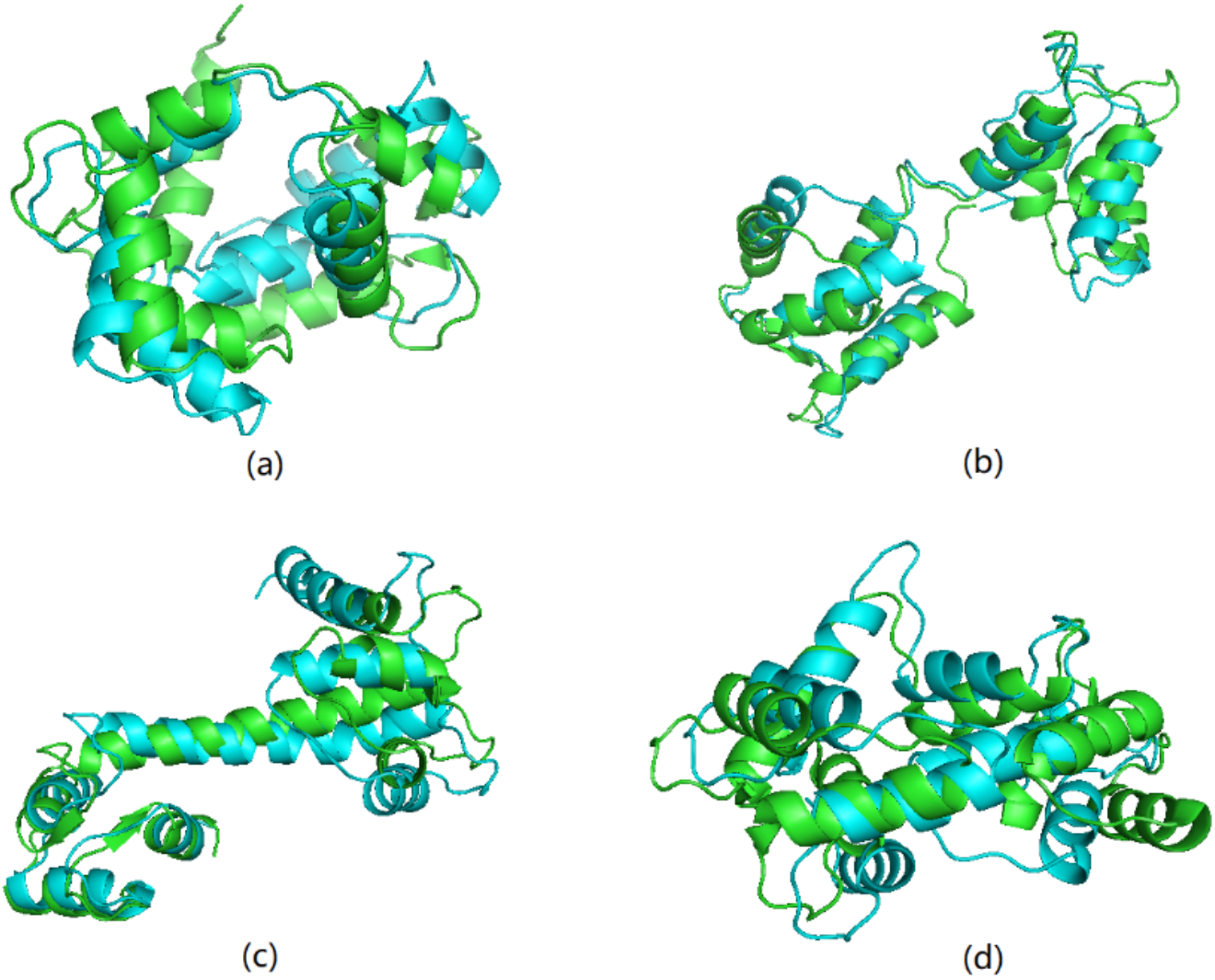

Figure 2

Alignment of empirically determined and predicted structures corresponding to a single amino acid sequence using our method. Known structures are shown in green, and predicted structures in blue, for (a) protein 1a29, (b) protein 1cfd, (c) protein 1cll and (d) protein 2bcx.

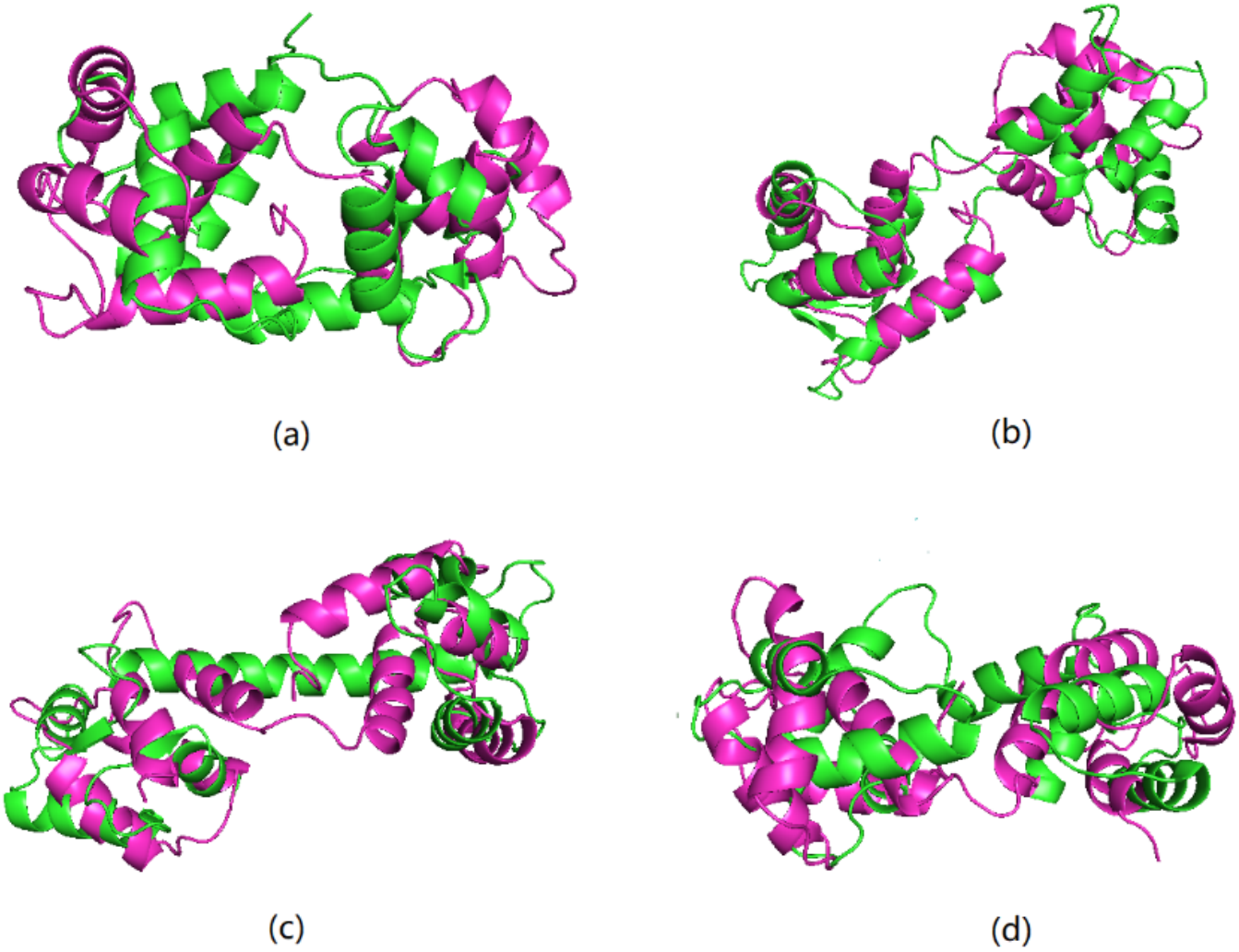

Figure 3

Alignment of empirically determined and predicted structures corresponding to a single amino acid sequence using RaptorX method. Known structures are shown in green, and predicted structure in purple, for (a) protein 1a29, (b) protein 1cfd, (c) protein 1cll and (d) protein 2bcx.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx