Quantitative Biology 2018, 6(4): 307–312
https://doi.org/10.1007/s40484-018-0158-1

RESEARCH ARTICLE

Selecting near-native protein structures from *ab initio* models using ensemble clustering

Li Li, Huanqian Yan and Yonggang Lu*

School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China
* Correspondence: ylu@lzu.edu.cn

Received March 9, 2018; Revised April 23, 2018; Accepted May 5, 2018

Background: Ab initio **protein structure prediction is to predict the tertiary structure of a protein from its amino acid sequence alone. As an important topic in bioinformatics, considerable efforts have been made on designing the *ab initio* methods. Unfortunately, lacking of a perfect energy function, it is a difficult task to select a good near-native structure from the predicted decoy structures in the last step.**
Methods: **Here we propose an ensemble clustering method based on *k*-medoids to deal with this problem. The *k*-medoids method is run many times to generate clustering ensembles, and then a voting method is used to combine the clustering results. A confidence score is defined to select the final near-native model, considering both the cluster size and the cluster similarity.**
Results: **We have applied the method to 54 single-domain targets in CASP-11. For about 70.4% of these targets, the proposed method can select better near-native structures compared to the SPICKER method used by the I-TASSER server.**
Conclusions: **The experiments show that, the proposed method is effective in selecting the near-native structure from decoy sets for different targets in terms of the similarity between the selected structure and the native structure.**

Keywords: near-native structure; protein structure prediction; *ab initio*; decoy; ensemble clustering; *k*-medoids

Author summary: It is a difficult task to select a good near-native structure from the predicted decoy structures produced by *ab initio* structure prediction methods. The *k*-medoids is usually used for the purpose due to its simplicity and efficiency. However, the result of the *k*-medoids method may be affected by its initial centroid selection. The paper proposes a new ensemble clustering method based on *k*-medoids to deal with this problem. The experiments show that the proposed method is effective in selecting the near-native structure from decoy sets for different targets.

INTRODUCTION

Determining the tertiary structure of a protein is a crucial step for understanding its functionality. Currently, X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-EM are the three major methods for determining the protein structures experimentally. Due to the heavy costs and difficulties in specimen preparation, the number of available protein structures still lags far behind the number of available protein sequences. The new release of UniProtKB/TrEMBL protein database in June 2017 contains 87,291,332 sequence entries, but only about 120,000 of them have experimentally solved structures [1]. As a result, protein structure prediction is an important topic in bioinformatics and computational structural biology.

There are three protein structure prediction methods: comparative modeling, fold recognition, and *ab initio* modeling [2]. While comparative modeling and fold recognition depend on the availability of known structure templates, *ab initio* modeling can predict protein structure given only its amino acid sequence. Although considerable efforts have been made, the *ab initio* prediction of protein structure still remains an outstanding unsolved problem. One of the challenges in designing the *ab initio* method is to select the best near-native model from a large number of predicted models in the last step. The best near-native model of a protein can be easily identified if given a perfect energy function according to the basis of minimum energy. But the available energy functions all use some kind of approximations which make them unreliable for selecting the near-native models.

Statistical analysis has been used effectively in protein study. For example, regression analysis has been used for protein sequences analysis [3], and support vector machine (SVM) has been used for analyzing protein-protein interactions [4,5]. Similarly, statistical approaches, such as cluster analysis, have been used for the selection of near-native structure too. Generally, people believe that the native state should be the most populated at low temperatures [6]. So the near-native models are usually the cluster centers densely surrounded by many predicted models.

The *k*-medoids method [7] is a clustering method related to the *k*-means method which is a classical clustering method that clusters the data set of *n* objects into *k* clusters. Many soft subspace clustering algorithms are also based on *k*-means model [8]. In contrast to the *k*-means method, *k*-medoids chooses real data points as cluster centers. So, the *k*-medoids method is often used in the *ab initio* methods for selecting the near-native structure from decoys [9]. However, *k*-medoids method may be affected by its random selection of the cluster centroids during initialization. It may produce different clustering results given different initializations, where each clustering result can be viewed as a possible "look" through the data. The ensemble clustering exploits the complementary nature of different partitions to obtain a good overview [10]. The ensemble approach has also be applied in solving other biological problems [11–13], such as identifying core atom sets, clustering protein-protein interaction networks and processing biological datasets.

In this paper, we have proposed an ensemble clustering method based on *k*-medoids to deal with the near-native structure selection problem. The experiments show the effectiveness of the proposed method.

The rest of paper is organized as follows. First the related works are discussed. Second the experimental results are shown and a discussion about the results is given. At last, each step of our method is described in detail.

RELATED WORK

TM-score

The measure of decoy structure similarities used is TM-score [14], which is a variation of the Levitt-Gerstein (LG) score to assess the quality of protein structure templates and predict full-length models. All the residues of the modeled proteins are evaluated by a protein size dependent scale, rather than using a specific distance cutoff and focusing only on the fractions of structures in the MaxSub or GDT-scoring function. TM-score is more sensitive to the correctness of global topology than the local structural errors, while the RMSD measure is sensitive to local small disorientations which may result in a big overall RMSD change even though the core region of the model may be correct. TM-score is defined as:

$$\mathrm{TM-score} = \mathrm{Max}\left[\frac{1}{L_N}\Sigma_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right], \qquad (1)$$

where L_N is the length of the native structure, L_T is the length of the aligned residues to the template structure, d_i is the distance between the *i*-th pair of aligned residues and d_0 is a scale to normalize the match difference. d_0 can be approximated by the following formula:

$$d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8, \qquad (2)$$

which represents the average structure match difference of random related structures.

The TM-score is between 0 and 1, where 1 indicates a perfect match between two structures. Generally, scores below 0.2 correspond to randomly chosen unrelated proteins, besides the score of structures roughly have the same fold is higher than 0.5.

CASP

Critical assessment of protein structure prediction (CASP) is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994 [15]. CASP provides research groups with an opportunity to objectively test their structure prediction methods and delivers an independent assessment of the state of the art in protein structure modeling to the research community and software users.

I-TASSER server [16,17] is ranked as the top server for protein structure prediction in recent CASP experiments [18]. So the decoy sets, generated by I-TASSER, of single-domain targets in the CASP11 are used in our experiments.

SPICKER

SPICKER [19] is a clustering approach to identify near-native protein folds developed by Zhang and Jeffery. It clusters the protein structures as follows. At first, a self-adjusting cutoff between 7.5 to 12 Å is found in an iterative way to make sure that the largest cluster contains less than 70% and more than 15% of total decoys. Then, another iterative approach is applied to identify the cluster with the most neighbors under the cutoff excluding the members of clusters found in the previous iterations. Finally, an averaging model, called final model, is built from all the decoy members of the cluster in the current iteration.

These results show that SPICKER is an efficient strategy to identify near-native folds, and show significant improvement over their previous clustering algorithms. It is to be noted that the final model is an averaged model, which does not exist in the original decoy set.

Ensemble clustering

There are two main steps in ensemble clustering: generating clustering ensembles and combining multiple clustering results. Three different methods can be used to generate clustering ensembles: using different clustering algorithms, running the same algorithm many times with different parameters or initializations, or using different samples of the dataset. Two approaches are usually used to combine multiple clustering results in the ensemble clustering [20]: median partition based approach, and object co-occurrence based approach. The object co-occurrence based approach includes voting based method, co-association matrix based method and graph based method.

In our method, the *k*-medoids algorithm with different initializations is run many times to generate clustering ensembles, and then a voting based method is used to combine the clustering results. In the voting procedure, good clustering ensembles will be accumulated, leading to a large weight.

RESULTS

Dataset

Fifty-four decoy sets, generated by I-TASSER for CASP11, which are single-domain targets and have experimental native structures, are downloaded from the Zhang Lab website. These decoy sets contain structurally non-redundant set of protein structures from the raw decoy sets. Each decoy set generally contains 750 or less decoys for forty easy targets and 1550 decoys for fourteen hard targets. The native structure, the generated model by SPICKER used in I-TASSER sever, and the best TM-score for the target in the decoy set can also be found on the webpage. These decoy sets can be downloaded from the Zhang Lab website: http://zhanglab.ccmb.med.umich.edu/decoys/casp11.

Parameter selection

One important problem for clustering is to select the number of clusters. This parameter is necessary for many clustering methods, for example, *k*-medoids. Generally, it is hard to tell the optimal number of clusters for a target protein. One of the advantages of ensemble clustering is that it can even improve the clustering result generated from arbitrary number of clusters. When the selected number of clusters is closer to the optimal number of clusters, the result of the ensemble clustering will be more accurate.

Different values from 2 to 10 are used for the parameter *k* in *k*-medoids. Two models are used to evaluate the experimental results, which are the model found by our method and the best model in the decoy set. For each *k* value, the difference of TM-scores between two models and the corresponding native structure is calculated. The boxplot in Figure 1 shows the distribution of the TM-score differences of all targets for different value of *k*. As we can see from Figure 1, the result becomes better when *k* is larger than 2, while it is getting worse when *k* is larger than 5. So the parameter *k* is set to be 5.

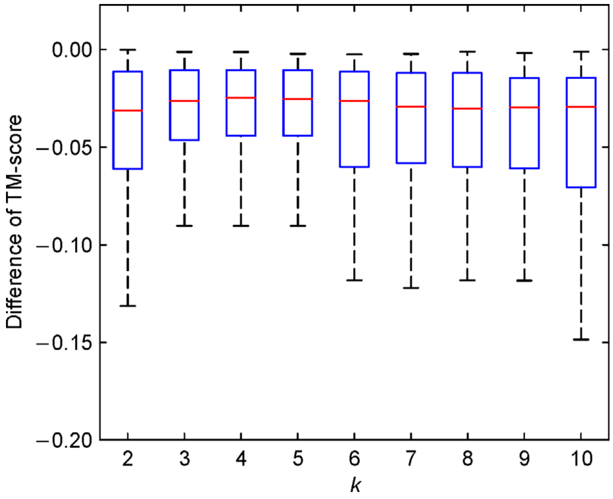

Figure 1. Parameter selection.

Experimental results

To evaluate the proposed method, the near-native structures selected by our method and the near-native structure generated by the SPICKER method used in I-TASSER server are compared. The TM-score between two near-native structures and the native structure are computed. The result is shown in a scatter plot in Figure 2, in which each target protein is represented as one point. The *x*-axis represents the TM-score produced by our method comparing to the native structure and the *y*-axis represents the SPICKER's. The red diagonal line in Figure 2 represents $y = x$. It is to be noticed that the same TM-score does not mean the same model. The model selected by our method is the cluster center which does exist in the decoy set for the target, while the model generated by SPICKER is an averaging model.

As shown in Figure 2, for about 70.4% of all targets the points fall below the diagonal line, which shows that our method produces better results than SPICKER in terms of the TM-score. From the aspect of the difficulty of a target according to the CSAP classification, 72.5% of our results are better for these easy targets, while 64.3% of our results are better for the difficult targets.

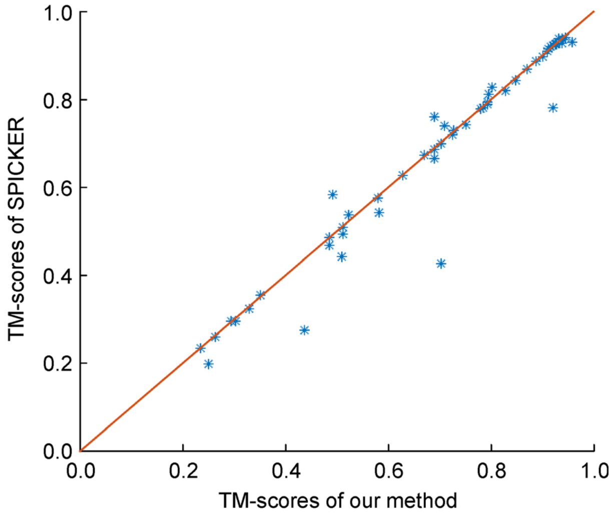

Figure 2. The comparison of TM-scores produced by two methods.

The details of the comparison can also be found in Table 1, in which the first column is the ID of the target protein, the second column is the TM-score of the selected near-native model by our method comparing to the corresponding native structure, the third column is the TM-score of the SPICKER model, and the last column shows the best TM-score for each target in the decoy set.

There are 9 special targets (T0763, T0768, T0797, T0816, T0817, T0836, T0837, T0851 and T0855), whose best model in the decoy set is pretty better than the SPICKER model. More specifically, the difference of the TM-score between the best decoy and the SPICKER's model is more than 0.1 for the 9 special targets, while the score difference is very small for the other targets. It is found that, for these special targets, our method produces better results in 8 of 9 targets, in which 4 of 8 have obvious improvements.

Taking target T0851 as an example, Figure 3 shows the superposition between the native structure and the near-native structure found by our method or SPICKER. The red model is the native structure and the blue is the structure selected by our method in Figure 3A, the other blue structure is generated by SPICKER in Figure 3B. It can be seen from Figure 3 that the SPICKER model has an obvious mismatch in the right half part of the protein, leading to a TM-score of 0.782, while the model selected by our method matches better with the native structure and has a TM-score of 0.919.

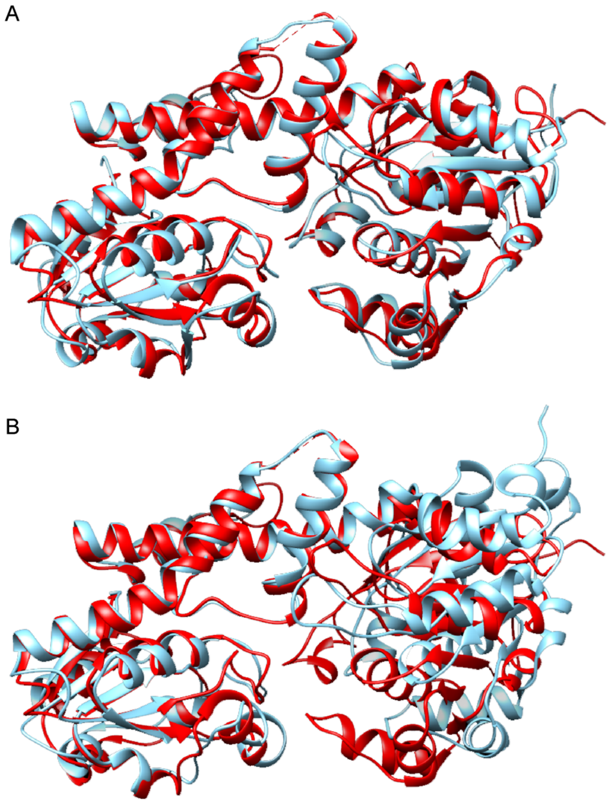

Figure 3. Comparison of the selected near-native structures by our method and SPICKER. (A) The superposition of T0851 native structure and the near-native structure found by our method. (B) The superposition of T0851 native structure and the model generated by the SPICKER method.

CONCLUSION

The paper proposes a new ensemble clustering method based on *k*-medoids for selecting the near-native model from a decoy set produced by *ab initio* structure prediction method. The experiments show that the proposed method is effective in selecting the near-native structure from decoy sets for different targets. The most time-consuming part of the method is to calculate the similarity matrix using the TM-score, which is quite slow for a large decoy set. Our future work is intended to design a fast and effective protein structure comparison method for constructing the similarity matrix.

Table 1 Comparison of the TM-scores

Target ID	Our method	SPICKER	Best decoy
T0759	0.352	0.356	0.427
T0762	**0.926**	0.925	0.931
T0763	**0.251**	0.198	0.311
T0764	**0.886**	0.885	0.888
T0765	0.689	0.761	0.785
T0766	**0.940**	0.935	0.954
T0768	**0.629**	0.626	0.865
T0769	0.710	0.741	0.782
T0773	0.796	0.812	0.828
T0778	**0.958**	0.929	0.965
T0782	**0.689**	0.687	0.729
T0784	**0.938**	0.937	0.949
T0785	**0.265**	0.261	0.309
T0786	**0.783**	0.782	0.800
T0787	**0.235**	**0.235**	0.257
T0788	**0.901**	0.897	0.903
T0792	0.670	0.672	0.704
T0796	**0.689**	0.666	0.715
T0797	0.801	0.826	0.948
T0798	0.931	0.937	0.955
T0800	**0.579**	0.575	0.602
T0801	**0.937**	0.926	0.943
T0803	**0.486**	0.467	0.524
T0805	**0.848**	0.843	0.858
T0807	0.911	0.913	0.922
T0811	0.944	0.941	0.956
T0812	0.522	0.536	0.612
T0813	0.919	0.922	0.925
T0815	**0.886**	0.885	0.922
T0816	**0.296**	**0.296**	0.721
T0817	**0.724**	0.718	0.917
T0819	0.916	0.920	0.923
T0820	**0.329**	0.324	0.395
T0821	**0.828**	0.818	0.868
T0822	**0.510**	0.442	0.539
T0823	**0.780**	0.779	0.792
T0824	**0.303**	0.296	0.336
T0825	**0.511**	0.509	0.513
T0829	0.493	0.584	0.649
T0833	**0.750**	0.743	0.776
T0835	**0.703**	0.700	0.726
T0836	**0.438**	0.276	0.448
T0837	**0.702**	0.427	0.736
T0838	**0.582**	0.543	0.591
T0841	**0.927**	0.926	0.935
T0843	**0.925**	0.924	0.938
T0847	**0.793**	0.788	0.803
T0849	0.728	0.730	0.769
T0851	**0.919**	0.782	0.928
T0854	0.793	0.794	0.829
T0855	**0.511**	0.494	0.629
T0856	**0.869**	0.869	0.880
T0857	0.485	0.487	0.548
T0858	0.909	0.910	0.924

METHOD

Our ensemble clustering method can be divided into three major steps: constructing a distance matrix for the decoy set using TM-score; finding the most possible cluster center using an ensemble *k*-medoids; selecting the cluster center with the maximum score as the result, *i.e.*, the near-native structure found by our method.

Construct the distance matrix

To produce the distance matrix for the clustering method, a similarity matrix for the decoys needs to be constructed, and then we can get the distance matrix by defining distance = 1–similarity. The distance matrix is a symmetric matrix whose diagonal elements are all 0. The element in *i*-th row and *j*-th column represents the dissimilarity between two decoys *i* and *j*.

Find the most possible cluster center using ensemble clustering

K-medoids is run $m = 500$ times, which is enough to ensure the statistical stability. Running with different random initialization, the number of times a decoy becomes the center of the largest cluster is counted. By combining multiple clustering results, a more general overview of the data can be obtained. It is found that a reasonable value for the parameter *k* used in *k*-medoids is 5, which has already been discussed in Parameter Selection of Results Section.

Select the near-native structure

To consider both the size and the internal similarity of a cluster in selecting the near-native structure, a confidence score is defined for each cluster center, which is:

$$cs = \frac{\sum_{i \in C} \sum_{j \in C} sim(i,j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} sim(i,j)}, \tag{3}$$

where C is the target cluster containing the cluster center, n is the total number of decoys in the decoy set, sim is the similarity matrix between a pair of decoys.

The center with the maximum confidence score within the cluster centers whose count is more than 70% of the maximum count is selected as the near-native structure, where the count is the number of times a decoy becomes the center of the largest cluster as described earlier.

The pseudocode for the selection of the near-native structure is shown in Figure 4.

```
Algorithm SelectNearNativeDecoy
Input: Decoys[1…n], k, MaxIter
Output: The near-native structure
  TM[1…n, 1…n] ← Compute TM-score between each pair
                  of Decoys[1…n]
  Distance[1…n, 1…n] ← 1 − TM
  Count[1…n] ← 0
  Sum[1…n] ← 0
  for iter ← 1 to MaxIter
    Cluster_result ← k-medoids(Distance, k) with random
                     initialization
    LC ← the largest cluster in the Cluster_result
    idx ← the index of the center of LC in the Decoys
    Count[idx] ← Count[idx] + 1
    CS ← compute the confidence score using equation (3)
    Sum[idx] ← Sum[idx] + CS
  end
  Max_count ← Max(Count)
  Score[1…n] ← 0
  for idx ← 1 to n
    if  Count[ idx ]>0.7* Max_count then
      Score[ idx ] ← Score[ idx ] / Count[ dix ]
    end
  end
  Max_index ← arg_max(Score)
  return Decoys [Max_index]
```

Figure 4. The pseudocode for selecting the near-native decoy.

ACKNOWLEDGEMENTS

This work is supported by the National Key R&D Program of China (No. 2017YFE0111900), and the Lanzhou Talents Program for Innovation and Entrepreneurship (No. 2016-RC-93).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Li Li, Huanqian Yan and Yonggang Lu declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. UniProtKB/TrEMBL Protein Database Release Statistics. https://www.ebi.ac.uk/uniprot/TrEMBLstats (Accessed Jun 30, 2017)
2. Zhang, Y. and Skolnick, J. (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. USA, 101, 7594–7599
3. Huang, D. S., Zhao, X. M., Huang, G. B. and Cheung, Y. M. (2006) Classifying protein sequences using hydropathy blocks. Pattern Recognit., 39, 2293–2300
4. Xia, J. F., Zhao, X. M., Song, J. and Huang, D. S. (2010) APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics, 11, 174
5. Huang, D. S., Zhang, L., Han, K., Deng, S., Yang, K. and Zhang, H. (2014) Prediction of protein-protein interactions based on protein-protein correlation using least squares regression. Curr. Protein Pept. Sci., 15, 553–560
6. Shortle, D., Simons, K. T. and Baker, D. (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc. Natl. Acad. Sci. USA, 95, 11158–11162
7. Kaufman, L. and Rousseeuw, P. J. (1987) Clustering by means of medoids. In Statistical Data Analysis Based on The Ll-Norm and Related Methods, Dodge , Y. (ed.). Basel: Birkhäuser Basel
8. Deng, Z., Choi, K. S., Jiang, Y., Wang, J. and Wang, S. (2016) A survey on soft subspace clustering. Inf. Sci., 348, 84–106
9. Bradley, P., Malmström, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., Meiler, J., Misura, K. M. and Baker, D. (2005) Free modeling with Rosetta in CASP6. Proteins, 61, 128–134
10. Jain, A. K. (2010) Data clustering: 50 years beyond *K*-means. Pattern Recognit. Lett., 31, 651–666
11. Snyder, D. A. and Montelione, G. T. (2005) Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. Proteins, 59, 673–686
12. Asur, S., Ucar, D., and Parthasarathy, S. (2006) An ensemble approach for clustering protein-protein interaction networks. Bioinfomatics, 23, i29-i40
13. Pirim, H. and Seker, S. E. (2012) Ensemble clustering for biological datasets. In Bioinformatics, Pérez-Sánchez, H., (Ed.). IntechOpen
14. Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins, 57, 702–710
15. Moult, J., Pedersen, J. T., Judson, R. and Fidelis, K. (1995) A large-scale experiment to assess protein structure prediction methods. Proteins, 23, ii–v
16. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. and Zhang, Y. (2015) The I-TASSER Suite: protein structure and function prediction. Nat. Methods, 12, 7–8
17. Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, 9, 40
18. The 11th Critical Assessment of Techniques for Protein Structure Prediction. predictioncenter.org/casp11/zscores_final.cgi (Accessed Jun 30, 2017)
19. Zhang, Y. and Skolnick, J. (2004) SPICKER: a clustering approach to identify near-native protein folds. J. Comput. Chem., 25, 865–871
20. Vega-Pons, S. and Ruiz-Shulcloper, J. (2011) A survey of clustering ensemble algorithms. Int. J. Pattern Recognit. Artif. Intell., 25, 337–372