Prediction of protein distance maps by assembling fragments according to physicochemical similarities

Gualberto Asencio Cortés, Jesús S. Aguilar-Ruiz and Alfonso E. Márquez Chamorro

Abstract The prediction of protein structures is a current issue of great significance in structural bioinformatics. More specifically, the prediction of the tertiary structure of a protein consists of determining its three-dimensional conformation based solely on its amino acid sequence. This study proposes a method in which protein fragments are assembled according to their physicochemical similarities, using information extracted from known protein structures. Many approaches cited in the literature use the physicochemical properties of amino acids, generally hydrophobicity, polarity and charge, to predict structure. In our method, implemented with parallel multithreading, a set of 30 physicochemical amino acid properties selected from the AAindex database were used. Several protein tertiary structure prediction methods produce a contact map. Our proposed method produces a distance map, which provides more information about the structure of a protein than a contact map. The results of experiments with several non-homologous protein sets demonstrate the generality of this method and its prediction quality using the amino acid properties considered.

1 Introduction

There are currently two main approaches to predicting protein structure. On the one hand, the ab initio and de novo methods try to solve the structure of a protein based on physicochemical principles and without using any protein as a template. Conversely, the homology modeling methods try to solve the structures based on protein templates.

The template-based modeling methods achieve good results when there are proteins with sequences similar to the target protein. When no homologous proteins

Gualberto Asencio Cortés, Jesús S. Aguilar-Ruiz and Alfonso E. Márquez Chamorro
School of Engineering, Pablo de Olavide University, e-mail: {guaasecor,aguilar,amarcha}@upo.es

with solved structures exist, free modeling is used. Within the free modeling methods, fragment assembly methods that reconstruct the structure of a protein from other protein structural fragments, such as Rosetta [1], have been developed.

The physicochemical properties of amino acids have been used in several protein structure prediction studies. The most commonly used properties have been hydrophobicity, polarity and charge; for example, in the HPNX model [2] for lattice predictions.

There are numerous protein structure prediction algorithms that produce a contact map to represent the predicted structure. Our method produces a distance map that incorporates more information than a contact map, because it incorporates the distances between all of the amino acids in the molecule, irrespective of whether they make contact. Unlike 3D models, both contact maps and distance maps have the desirable property of being insensitive to rotation or translation of the molecule.

The proposed method selects the most reliably known distances between amino acid pairs from known protein structural fragments. The fragments are chosen for similarities in length and in 30 physicochemical properties of their amino acids. We evaluated the predictions obtained from several sets of proteins with low sequence identity to determine the generality of the prediction method.

In the methods section, we describe the procedures used in our prediction. In the experimental results section, we explain the data sets used as well as how the results were obtained. Finally, in the conclusion section, we discuss the main results of the study.

2 Methods

The prediction system, called ASPF-PRED (Aminoacid Subsequences Property File Predictor), was divided into two phases. In the first phase, a knowledge-based model was generated from all of the fragments or subsequences from all the proteins in a training set. In the second phase, structures were predicted for all of the proteins in a test set using the knowledge-based model generated in the first phase.

The knowledge-based model consisted of a set of vectors called prediction vectors. Each prediction vector was obtained from a training protein subsequence and contained the length of the subsequence, the average values of the physicochemical properties of its internal amino acids and the actual distance between the ends of the subsequence.

The length of each subsequence was standardized between 0 and 1. For this standardization, the length of each subsequence was divided by the maximum length of all the training proteins. The standardization ensured that all of the prediction vector traits were on the same scale and contributed equally to the prediction. The properties, attributable to each amino acid within the subsequence, were also standardized, averaged and stored in the prediction vector. Finally, the actual distance between the amino acid ends (first and last of the subsequence) was added to each vector.

In the second phase of prediction, all of the test protein prediction vectors were obtained and a full sequential search was conducted, comparing each of them with the training protein prediction vectors. The objective was to find the training protein prediction vector that was the most similar to each test protein prediction vector. For the search process, only the training vectors with the same ends as the test vectors were considered.

For compare the prediction vectors, a Euclidean distance between the test and training vectors was used. This distance was calculated from the lengths of the subsequences and the average values of the properties of their internal amino acids.

After the predictions were made, a distance map was generated for each of the test protein sequences. The distance map of a sequence is a square matrix of order N, where N is the number of amino acids possessing this sequence. The factor (i, j) with $i < j$ of the matrix is the distance, measured in Angstroms, observed between the ith and the jth amino acids of the sequence. To measure the distances, the beta carbons were used (except for glycine, for which the alpha carbon was used). The predicted distances are finally stored in the lower triangle of each distance map.

The ASPF-PRED system generated the following measures to evaluate the quality of the prediction: accuracy, recall, specificity and precision. To obtain these measures, different cut-off thresholds were established for the actual distance values, and these were analyzed in the experiment.

3 Experimental results

Four experiments were conducted to test the performance of the ASPF-PRED system. An identical initial configuration was established for all of the experiments, varying only the set of proteins used. For all of the experiments, ten-fold cross validation was used.

The set of physicochemical properties of amino acids that was used was obtained by a selection of traits from the complete AAindex database [3], which lists 544 properties. The selection of traits that produced the best results has 30 traits, showed in the Table 1, and was obtained by the Relief evaluation algorithm with the 10 nearest neighbors and a Ranker search algorithm. Both the set of properties and the set of proteins used can be found at http://www.upo.es/eps/asencio/aspfpred30.

Table 1 Physicochemical properties of amino acids considered from AAindex

UTK870103	MONM990201	VELV850101	KHAG800101	BUNA790103
MITS020101	TANS770108	WERD780103	NADH010107	MAXF760103
CHAM820102	TANS770102	RICJ880104	FAUJ880111	RICJ880117
KARP850103	VASM830101	JOND750102	QIAN880139	RICJ880101
GARJ730101	BUNA790101	WERD780102	WILM950104	RICJ880114
FAUJ880112	AURR980120	DIGM050101	SUEM840102	PRAM820101

The objective followed in the selection of the protein sets was to use non-homologous proteins (identity less or equal to 30%). Therefore, it was possible to ascertain whether the prediction method is general enough and assert that it does not work only for specific families of proteins.

In the first experiment, 20 proteins that were randomly selected from the PDB Web [4] in April 2010 and had less than or equal to 30% identity to each other were used. In this experiment we used a small set of proteins to test the behavior offered by the predictor with a poor training information.

In the following experiments we used a larger number of proteins to see if it increases the quality of the predictions with increasing training information. In addition, we have used identity values lower than that of experiment 1. Resolution values used in obtaining experimental proteins was less than 1.4, with the aim of providing accurate training.

In the second experiment, proteins with more than 70 amino acids with a resolution between 0-1.0, an R-factor between 0-0.2 and a maximum of 10% identity (118 proteins) were obtained from CullPDB [5]. In the third experiment, proteins with more than 40 amino acids with a resolution between 0-1.4, an R-factor between 0-0.12 and a maximum of 25% identity (170 proteins) were obtained from PDBselect [6]. In the fourth experiment, proteins with more than 70 amino acids with a resolution between 0-1.1, an R-factor between 0-0.2 and a maximum of 5% identity (221 proteins) were obtained from CullPDB.

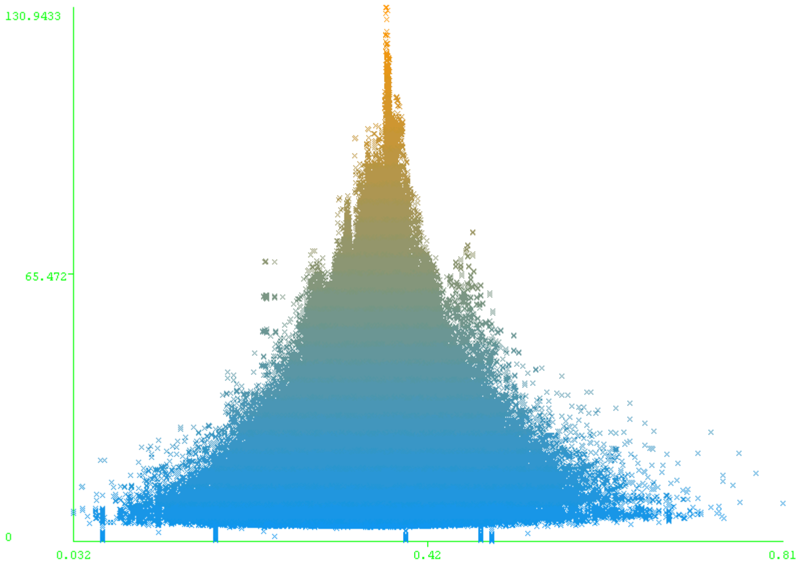

Fig. 1 Distance distribution for property WILM9501040. The x-axis represents the normalized value of the physicochemical property and the y-axis represents the distance between amino acids that have the value of the property

Figures 1 and 2 shows the distribution of distances between amino acids according to two physicochemical properties used (WILM9501040 and GARJ730101). For this distribution of distances, have been referred to all the 221 amino acids of all proteins of experiment 4. They include only the distributions of distances for two physicochemical properties of amino acids, but the distribution of other properties is similar. The x-axis of Figures 1 and 2 represents the normalized value of the physicochemical property and the y-axis represents the distance between amino acids that have the value of the property.

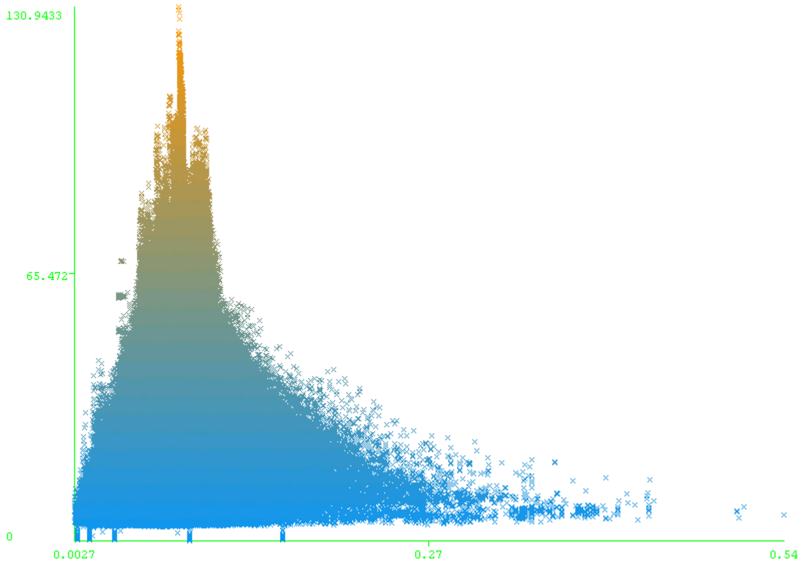

Fig. 2 Distance distribution for property GARJ730101. The x-axis represents the normalized value of the physicochemical property and the y-axis represents the distance between amino acids that have the value of the property

As can be seen in Figures 1 and 2, the distances between amino acids seem to follow a normal distribution with mean 0.402 and deviation 0.31 in the case of WILM9501040 property, and with mean 0.047 and deviation 0.059 in the case of property GARJ730101.

In Tables 2 and 3 we show the results obtained in protein structure prediction of the four experiments. We indicate the values of accuracy, recall, specificity and precision. In Table 2 we used a cut-off of 4 Å and in Table 3 a cut-off of 8 Å.

To show the complete results of the experiments and facilitate their analysis, one graph has been included for each experiment (Figure 3). In each graph, the distance threshold values (in Angstroms) are shown on the x-axis, and the accuracy, recall, specificity and precision values are shown on the y-axis.

Table 2 Efficiency of our method at 4 Å of distance threshold.

Experiment	Recall	Precision	Accuracy	Specificity
1	0.10	0.08	0.99	0.99
2	0.31	0.39	0.99	0.99
3	0.48	0.43	0.99	0.99
4	0.40	0.41	0.99	0.99

Table 3 Efficiency of our method at 8 Å of distance threshold.

Experiment	Recall	Precision	Accuracy	Specificity
1	0.39	0.41	0.97	0.98
2	0.39	0.40	0.95	0.97
3	0.38	0.38	0.95	0.97
4	0.40	0.41	0.95	0.97

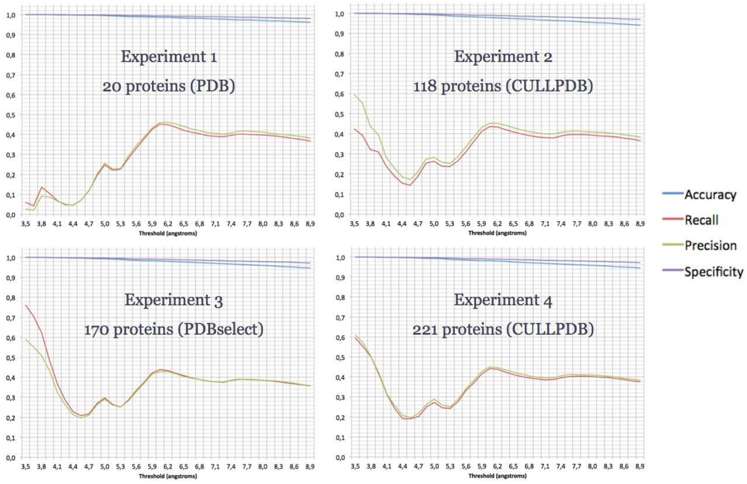

Fig. 3 Accuracy, recall, specificity and precision values of the four experiments

4 Conclusions

We performed four experiments to test the efficiency of our predictor with a poor training knowledge (experiment 1) and with a higher and diverse training knowledge (experiments 2, 3 and 4).

We found that, with a poor knowledge (experiment 1 with 20 proteins), the quality of prediction, in terms of recall and precision, is low for thresholds between 3.5 and 4.8 Å. In particular, we obtain a recall of 0.10 and a precision of 0.08 for 4 Å of cut-off. This difference may have been due to the lower number of training proteins and, consequently, to the lower knowledge of the search space (protein structures).

We tested our predictor with greater number of proteins and with great diversity in their sequences (identities of 25%, 10% and up to 5% in experiment 4). The quality of the predictions in terms of recall and precision for low thresholds (between 3.5 and 4.8 Å) is higher than in experiment 1. However, the behavior of the measures for higher thresholds to 4.8 Å is similar to experiment 1.

Finally, we found empirically that the response of our method over protein sets with great diversity in their sequences seems to be the same irrespective of the type of protein to be predicted. In fact, the protein sets of these experiments have very low identity. This result is desirable, in theory, since this study sought generality of the method.

References

1. Carol A. Rohl, Charlie E. M. Strauss, Kira M. S. Misura, and David Baker. Protein structure prediction using rosetta. In Ludwig Brand and Michael L. Johnson, editors, *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pages 66 – 93. Academic Press, 2004.
2. Tamjidul Hoque, Madhu Chetty, and Abdul Sattar. Extended hp model for protein structure prediction. *Journal of computational biology : a journal of computational molecular cell biology*, 16(1):85–103, 2009.
3. Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. Aaindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D202–D205, Jan 2008.
4. Helen Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya Shindyalov, and Philip Bourne. The protein data bank. *Nucl. Acids Res.*, 28(1):235–242, 2000.
5. Guoli Wang and Roland Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics (Oxford, England)*, 19(12):1589–1591, 2003.
6. Sven Griep and Uwe Hobohm. Pdbselect 1992-2009 and pdbfilter-select. *Nucl. Acids Res.*, 38(suppl1):D318–319, 2010.