A Nearest Neighbour-Based Approach for Viral Protein Structure Prediction

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

School of Engineering, Pablo de Olavide University
{guaasecor,aguilar,amarcha}@upo.es

Abstract. Protein tertiary structure prediction consists of determining the three-dimensional conformation of a protein 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. Several existing protein tertiary structure prediction methods produce contact maps as their output. Our proposed method produces a distance map, which provides more information about the structure of a protein than a contact map. In addition, many existing approaches use the physicochemical properties of amino acids, generally hydrophobicity, polarity and charge, to predict structure. In our method, we used three different physicochemical properties of amino acids obtained from the literature. Using this method, we performed tertiary structure predictions on 63 viral capsid proteins with a maximum identity of 30% obtained from the Protein Data Bank. We achieved a precision of 0.75 with an 8-angstrom cut-off and a minimum sequence separation of 7 amino acids. Thus, for the studied proteins, our results provide a notable improvement over those of other methods.

Keywords: protein tertiary structure prediction, physicochemical amino acid properties, comparative modeling methods, fragment matching, distance map, nearest neighbors.

1 Introduction

Protein structure prediction is currently an issue of great significance in structural bioinformatics. This significance stems from the fact that the three-dimensional structure of a protein determines its function, which in turn has important repercussions in medicine and biology, particularly in areas such as drug design.

Although experimental procedures exist for determining the structures of proteins, including X-ray crystallography and nuclear magnetic resonance (NMR), these procedures are very expensive. Consequently, there is increasing interest in developing prediction algorithms for protein structure prediction.

Since the experiments of Anfinsen [1], it has been generally accepted that all of the necessary information for determining the structure of a protein is encoded in its sequence of amino acids. Thus, methods for tertiary structure prediction

have been designed. Such methods construct a three-dimensional model based solely on the amino acid sequence of a protein.

There are currently two main approaches for predicting protein structure. On the one hand, ab initio methods try to solve the structure of a protein by optimising an energy function, generally based on physicochemical principles and without using any protein as a template. However, these methods are only adapted for proteins of relatively small size [2]. In contrast, homology-modelling methods try to solve the structure based on protein templates (template modelling). The latter method is currently considered to be the most reliable approach for protein structure prediction [3].

The template-based modelling methods achieve good results when solved structures are available for proteins with sequences similar to the sequence of the target protein. However, when no homologous proteins with solved structures exist, free modelling is used. Within the free-modelling methods are fragment assembly methods that reconstruct the structure of a protein from structural fragments of other proteins; these methods include FragFold [4], Fragment-HMM [5] and ROSETTA [6]. ROSETTA uses a two-stage approach, which begins with a low-resolution model and continues with a representation of all the atoms of the protein, with the goal of minimising the corresponding energy function. In contrast, several methods for protein structure prediction are based on the physicochemical properties of amino acids. Among the most commonly used properties are hydrophobicity, polarity and charge, which are used, for example, in the models HP and HPNX [7].

Many protein structure prediction algorithms produce a contact map to represent the predicted structure. In contrast, our method produces a distance map, which includes 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.

Our method is a free-modelling method based on fragment assembly that selects the best distances between pairs of amino acids using fragments of known structures of proteins. The fragments are chosen through a process of searching for nearest neighbours by similarity in length and three physicochemical properties of amino acids selected from the literature. We tested this model by carrying out predictions on viral capsid proteins from the Protein Data Bank (PDB) [8] with a maximum identity of 30%. We have performed predictions with a minimum sequence separation of 7 amino acids, as proposed in the work of Fariselli et al. 2001 [9]. Finally, we compared our results to those obtained by other methods to determine the quality of the predictions obtained with our method.

In section 2, we describe the elements, procedures and evaluation measures used by our prediction method. In section 3, we detail the protein dataset used, the experimental settings and the obtained results. Finally, in section 4, we describe the main conclusions of the performed study and outline approaches for future studies.

2 Methods

2.1 Definition of Protein Distance Map

The distance map or distance matrix of a protein sequence is a square matrix of order N, where N is the number of amino acids in the sequence. The distance matrix is divided in two parts: observed part (upper triangular) and predicted part (lower triangular). The element (i, j), where $i < j$, of the distance matrix is the observed distance measured in angstroms (Å) between the i^{th} and j^{th} amino acid in the sequence. To measure the distances between amino acids, a reference atom is used. The most commonly used reference atoms are the alpha carbon and the beta carbon of an amino acid [9]. In our method, we used the beta carbon (with the exception of glycine, for which the alpha carbon was used). The distances predicted by the algorithm are stored in the lower triangular of the distance map. Thus, the element (i, j) with $i > j$ of the distance matrix is the predicted distance measured in angstroms between the i^{th} and j^{th} amino acid of the sequence.

2.2 Training Phase

The proposed prediction system, which we have named ASPF-PRED (Amino acid 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 of 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. These vectors represent physicochemical properties of training protein fragments. Each prediction vector was obtained from a training protein subsequence. The vector contains the length of the subsequence, the average values of the physicochemical properties of its amino acids and the actual distance between the ends of the subsequence. In Figure 1, the content of prediction vectors from all subsequences $S_1 \dots S_z$ of a protein sequence is formally defined.

The length L of each subsequence was normalised to fall between 0 and 1. For this normalization, the length of each subsequence was divided by the maximum length $lmax$ of all the training proteins. The normalization ensured that all of

	L	$\overline{P}_1$	$\cdots$	$\overline{P}_k$	D
$S_1 : a_1 a_2 ... a_m$	$m/lmax$	$\frac{1}{m}\sum_{i=2}^{m-1} P_1(a_i)$	$\cdots$	$\frac{1}{m}\sum_{i=2}^{m-1} P_k(a_i)$	$d(a_1, a_m)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$S_z : z_1 z_2 ... z_n$	$n/lmax$	$\frac{1}{n}\sum_{i=2}^{n-1} P_1(z_i)$	$\cdots$	$\frac{1}{n}\sum_{i=2}^{n-1} P_k(z_i)$	$d(z_1, z_n)$

Fig. 1. Prediction vector definition

the prediction vector traits were on the same scale and contributed equally to the prediction. The properties $P_1 \dots P_k$ of each amino acid within the subsequence, were also normalised, averaged and stored in the prediction vector ($\overline{P}_1 \dots \overline{P}_k$). Finally, the actual distance D between the amino acid ends (first and last of the subsequence) was added to each vector.

Our model used the following three physicochemical properties of amino acids: accessible surface area of residues in a folded protein [10], average relative fractional occurrence in ER [11] and RF value in high-salt chromatography [12].

2.3 Prediction of Protein Distance Maps

In the second phase of our method, we obtain the test prediction vectors of the test proteins and we perform a full sequential search to compare each test prediction vector with the training prediction vectors. The objective was to find the training prediction vector that was the most similar to each test prediction vector. For the search process, only training vectors with the same ends (first and last of the subsequence) as the test vectors were considered. Figure 2 illustrates the search scheme.

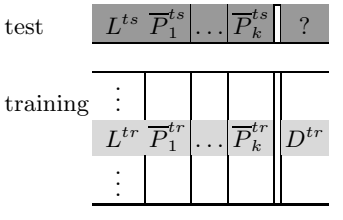

Fig. 2. Nearest neighbor search for each test prediction vector

In the search scheme of the Figure 2, L^{ts} is the length of the test subsequence. L^{tr} is the length of the training subsequence with more similarity to the test subsequence. $\overline{P}_1^{ts} \dots \overline{P}_k^{ts}$ are the average values of the amino acid properties of the test subsequence and $\overline{P}_1^{tr} \dots \overline{P}_k^{tr}$ are those of the nearest training subsequence. The distance to be predicted is symbolised with ? and is assigned the same value as the distance D^{tr} of the most similar training vector.

The training vector with the greatest similarity to the test vector satisfies the condition showed in formula 1. As can be seen in that condition, for the comparison of the test and training vectors, a Euclidean distance is used, which includes the lengths of the subsequences and the average values of the properties of their amino acids.

$$min\sqrt{(L^{ts} - L^{tr})^2 + (\overline{P}_1^{ts} - \overline{P}_1^{tr})^2 + \ldots + (\overline{P}_k^{ts} - \overline{P}_k^{tr})^2} \quad (1)$$

The distance field for each test vector was assigned the value of the distance field of the nearest training vector. The distance assigned to each test vector represents the predicted distance between the amino acid ends of the subsequence

to which the vector refers. Finally, the predicted distances are stored in the lower triangular of the distance map of the test sequence.

2.4 Evaluation of the Efficiency

We used several measures to evaluate the quality of the predictions. The first measure was precision, which is used in the works of Fariselli et al. [13,9]. The second was a measure of recall, which has been used in other protein prediction methods [14]. Finally, we have obtained measures of accuracy, specificity and Matthews Correlation Coefficient, that may often provide a much more balanced evaluation of the prediction than, for instance, the percentages [15]. The following formulas (2,3,4,5,6) define these five measures.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

These measures are used to evaluate the quality of a classification: i.e., each predicted value is assigned a value of 0 or 1. Thus, there are four possible outcomes depending on the quality of the predictions: a) both the real and predicted values are 1 (true positive, TP), b) both the real and predicted values are 0 (true negative, TN), c) the real value is 1 and the predicted value is 0 (false negative, FN) and d) the real value is 0 and the predicted value is 1 (false positive, FP). Because in this case, the class to predict is a real value (a distance), to obtain these measures it is necessary to binarise the class using a distance threshold or cut-off.

In this work, we used a cut-off value of 8 angstroms, which is commonly used in the literature [13,9,14]. In the evaluation of the measures, we omitted predictions of amino acid pairs with a minimum separation in the protein sequence of 7 amino acids.

3 Experimentation and Results

To verify the validity of the method, we performed a test on all viral capsid proteins (viral capsid, GO ID: 19028) published in the Protein Data Bank with a maximum identity of 30% (non-homologous proteins), as of November 2010

Table 1. The database of proteins used to train and test the predictor APSF-PRED

$L < 150$	1TD4	1CD3	2IZW	1C8D	1MUK	3IYH
1C5E	1VDO	1EI7	2VTU	1DZL	1OPO	3IYK
1GFF	1W8X	1F15	2VVF	1EJ6	1P2Z	3IYL
1HGZ	2C0W	1F2N	2WLP	1FN9	1QHD	3JYR
1IFK	2KX4	1JS9	2ZL7	1HX6	1SVA	3KIC
1IFL	2QUD	1STM	3FMG	1IHM	1YUE	3KZ4
1IFP	2VF9	1VPS	3KML	1KVP	2BBD	
1JMU	$L150 - 300$	1X36	$L > 300$	1LP3	2JHP	
1MSC	1AUY	1ZA7	1A6C	1M1C	2TBV	
1QBE	1C8N	2BUK	1BVP	1M3Y	2XVR	

Table 2. Efficiency of our method predicting distance maps of viral capsid proteins

Protein set	Recall	Precision	Accuracy	Specificity	MCC
All proteins (63)	0.77	0.75	0.99	0.99	0.75
$L < 150$ (16)	0.85	0.83	0.99	0.99	0.84
$150 \leq L < 300$ (19)	0.80	0.75	0.99	0.99	0.77
$L \geq 300$ (28)	0.75	0.73	0.99	0.99	0.73

(63 proteins, maximum length of 1284 amino acids). In Table 1, we show the PDB codes of the proteins used in the study. A leaving-one-out cross-validation was used to avoid the effect of choice of folds in a fold cross-validation. Table 2 shows the evaluation measures obtained in the experiment.

As shown in Table 2, we obtained a precision value of 0.75 and a recall value of 0.77 for the complete group of study proteins. To assess the quality of the predictions obtained with our method and to have reference values, we indicate the results obtained with other protein structure prediction approaches. In particular, in the work of Zhang et al. 2005 [14], a recall value of 0.27 was obtained with a cut-off of 8 angstroms for 5 test proteins. Fariselli et al. 2001 [9] achieved by cross validation a precision value of 0.21 for a cut-off of 8 angstroms and a minimum separation of 7 amino acids in the sequence.

Generally, the precision of the prediction of structures of proteins with long sequences (more than 300 amino acids) is lower than those proteins with short sequences. For example, in the work of Fariselli et al. 2001 [9], a precision value of 0.11 was obtained for proteins of 300 amino acids or more. With our method, a precision of 0.73 was obtained for proteins with lengths in this range.

Figure 3 shows the distance map obtained for the protein 1M3Y (413 amino acids) from the study set. We used a colour scale to represent the distances, ranging from the minimum distance (red) to the maximum (blue). As shown in the figure 3, the lower triangular of the matrix (prediction) is largely similar to the upper triangular (observation).

Figure 4 shows the contact map of the same protein 1M3Y, obtained using the distance map in Figure 3 and with a cut-off of 8 angstroms. As with the

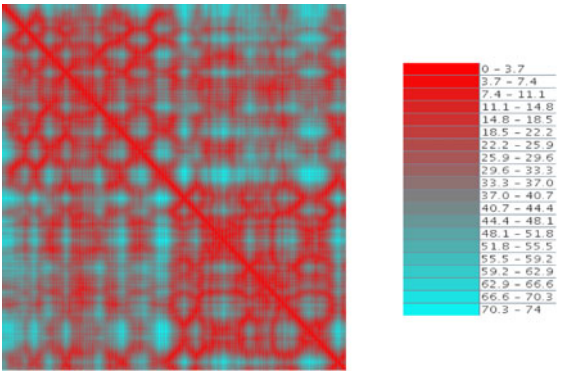

Fig. 3. Predicted distance map for the protein 1M3Y with color scale

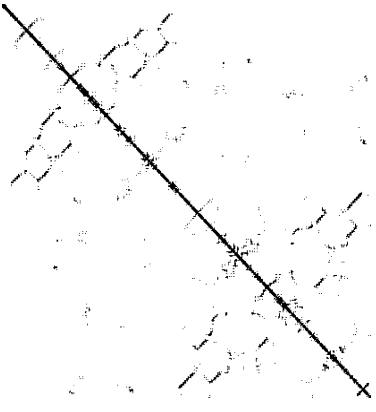

Fig. 4. Predicted contact map for the protein 1M3Y with a cut-off of 8Å

distance map, there is great similarity between the real and predicted parts of the contact map.

4 Conclusions and Future Work

Protein tertiary structure prediction problem consists of determining protein three-dimensional conformation based solely on its amino acid sequence. In this work, we have proposed a method in which protein fragments are assembled according to their physicochemical similarities, using three physicochemical properties of amino acids. We then predict distance maps, which provide more information about the structure of a protein than contact maps. We performed an experimental validation of the method on all non-homologous viral capsid proteins currently available in PDB. We obtained a precision of 0.75 with a cut-off of 8 angstroms and with a minimum sequence separation of 7 amino acids. Our

results are a significant improvement, for the studied proteins, on the results of previous studies.

In future work, we will refine the generated distance maps a posteriori, checking if they satisfy certain geometric and chemical restrictions for distance maps. Additionally, we intend to study other physicochemical properties of amino acids and check their utility for the protein structure prediction problem.

References

1. Anfinsen, C.B.: The formation and stabilization of protein structure. The Biochemical Journal 128(4), 737–749 (1972)
2. Wu, S., Skolnick, J., Zhang, Y.: Ab initio modeling of small proteins by iterative tasser simulations. BMC Biology 5(1), 17 (2007)
3. Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P., Boniecki, M.: Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. Proteins 45(S5), 149–156 (2001)
4. Jones, D.T.: Predicting novel protein folds by using fragfold. Proteins 5(suppl.), 127–132 (2001)
5. Li, S.C., Bu, D., Xu, J., Li, M.: Fragment-hmm: a new approach to protein structure prediction. Protein science: a publication of the Protein Society 17(11), 1925–1934 (2008)
6. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., Baker, D.: Protein structure prediction using rosetta. In: Brand, L., Johnson, M.L. (eds.) Numerical Computer Methods, Part D. Methods in Enzymology, vol. 383, pp. 66–93. Academic Press, London (2004)
7. Hoque, T., Chetty, M., Sattar, A.: Extended hp model for protein structure prediction. Journal of Computational Biology: a Journal of Computational Molecular Cell Biology 16(1), 85–103 (2009)
8. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I., Bourne, P.: The protein data bank. Nucl. Acids Res. 28(1), 235–242 (2000)
9. Fariselli, P., Olmea, O., Valencia, A., Casadio, R.: Prediction of contact maps with neural networks and correlated mutations. Protein Engineering 14(11), 835–843 (2001)
10. Chothia, C.: The nature of the accessible and buried surfaces in proteins. Journal of Molecular Biology 105(1), 1–12 (1976)
11. Rackovsky, S., Scheraga, H.A.: Differential geometry and polymer conformation. 4. conformational and nucleation properties of individual amino acids. Macromolecules 15(5), 1340–1346 (1982)
12. Weber, A.L., Lacey, J.C.: Genetic code correlations: Amino acids and their anticodon nucleotides. Journal of Molecular Evolution 11, 199–210 (1978), 10.1007/BF01734481
13. Fariselli, P., Casadio, R.: A neural network based predictor of residue contacts in proteins. Protein Engineering 12(1), 15–21 (1999)
14. Zhang, G.-Z., Huang, D.S., Quan, Z.H.: Combining a binary input encoding scheme with rbfnn for globulin protein inter-residue contact map prediction. Pattern Recogn. Lett. 26, 1543–1553 (2005)
15. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5), 412–424 (2000)