Predicting protein distance maps according to physicochemical properties

Gualberto Asencio Cortés*, Jesús S. Aguilar-Ruiz

Bioinformatics Group, School of Engineering, Pablo de Olavide University, Spain

Summary

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 in 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, we used a set of 30 physicochemical amino acid properties selected from the AAindex database. 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. We performed several preliminary analysis of the protein physicochemical data distributions using 3D surfaces. Three main pattern types were found in 3D surfaces, thus it is possible to extract rules in order to predict distances between amino acids according to their physicochemical properties. We performed an experimental validation of our method using five non-homologous protein sets and we showed the generality of this method and its prediction quality using the amino acid properties considered. Finally, we included a study of the algorithm efficiency according to the number of most similar fragments considered and we notably improved the precision with the studied proteins sets.

1 Introduction

There are currently two main approaches to predicting protein structure [1]. 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 [2, 3]. Conversely, the homology modeling methods try to solve the structures based on protein templates [4, 5].

The template-based modeling methods achieve good results when there are proteins with sequences similar to the target protein. When no homologous proteins 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 [6, 7, 8], such as Rosetta [9] and FRAGFOLD [10], has been developed.

The physicochemical properties of amino acids have been used in several protein structure prediction studies [11, 12, 13]. The most commonly used properties have been hydrophobicity, polarity and charge; for example, in the HPNX model [14] for lattice predictions.

*To whom correspondence should be addressed. Email: guaasecor@upo.es

doi:10.2390/biecoll-jib-2011-181

There are numerous protein structure prediction algorithms that produce a contact map to represent the predicted structure [15, 16, 17]. Our method produces a distance map [18, 19] 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 Data Analysis section, we perform several preliminary analysis of the protein physicochemical data distribution. In the Experimental results section, we explain the experiments and performed studies and we analyze results. Finally, in the conclusion section, we summarize the main results of this work.

2 Methods

The prediction system, called ASPPred (Aminoacid Subsequence Properties-based 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.

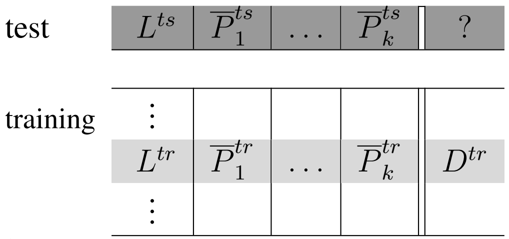
The knowledge-based model consisted of a set of vectors called prediction vectors. Each prediction vector was obtained from a training protein subsequence and contained the length of the subsequence, the average values of the physicochemical properties of its internal amino acids and the actual distance between the ends of the subsequence. In Figure 1, we define formally the prediction vector.

	L	$\overline{P}_1$	$\cdots$	$\overline{P}_k$	D
$a_1a_2...a_m$	$m/lmax$	$\frac{1}{m}\sum_{i=2}^{m-1} P_1(a_i)$	$\cdots$	$\frac{1}{m}\sum_{i=2}^{m-1} P_k(a_i)$	$d(a_1, a_m)$

Figure 1: A prediction vector.

The length L of each subsequence was normalised 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 the prediction vector traits were on the same scale and contributed equally to the prediction. The properties $P_1 \ldots P_k$ of each amino acid within the subsequence, were also normalised, averaged and stored in the prediction vector ($\overline{P}_1 \ldots \overline{P}_k$). Finally, the actual distance D 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. Figure 2 illustrates the search scheme.



test	L^{ts}	$\overline{P}_1^{ts}$	$\ldots$	$\overline{P}_k^{ts}$	?
training	$\vdots$				
	L^{tr}	$\overline{P}_1^{tr}$	$\ldots$	$\overline{P}_k^{tr}$	D^{tr}
	$\vdots$				

Figure 2: Search for the most similar training 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. $\overline{P}_1^{ts} \ldots \overline{P}_k^{ts}$ are the average values of the amino acid properties of the test subsequence and $\overline{P}_1^{tr} \ldots \overline{P}_k^{tr}$ 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.

To compare the prediction vectors, an 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, according to the Equation 1.

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

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 in the 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 ASPPred system is parallelized and it uses 400 execution threads. ASPPred generates the following measures to evaluate the quality of the predictions: accuracy, recall, specificity and precision. To obtain these measures, different cut-off thresholds were established for the distance values, and these were analyzed in the experiments.

3 Data analysis

The set of physicochemical properties of amino acids that was used was obtained by a feature selection from the complete AAindex database [20], which lists 544 properties. Feature selection has been applied previously in protein structure prediction [21]. We explored different feature selection techniques, and we obtained best results with the Relief evaluation algorithm [22] and a Ranker search algorithm over the proteins of experiment 4 (see Experimental results section). This procedure produced a ranking of attributes. We tried with each attribute subset starting at first attribute in ranking (subsets $\{\#1\}, \{\#1, \#2\}, ..., \{\#1, ...\#N\}$). We achieved the best results with the first 30 attributes (subset $\{\#1, ..., \#30\}$), which are showed in the Table 1. Both the set of properties and the set of proteins used can be found at `http://www.upo.es/eps/asencio/asppred`.

Table 1: The 30 physicochemical properties of amino acids considered from AAindex.

AURR980120	Normalized positional residue frequency at helix termini C4' [23]
BUNA790101	alpha-NH chemical shifts [24]
BUNA790103	Spin-spin coupling constants 3JHalpha-NH [24]
CHAM820102	Free energy of solution in water, kcal/mole [25]
DIGM050101	Hydrostatic pressure asymmetry index, PAI [26]
FAUJ880111	Positive charge [27]
FAUJ880112	Negative charge [27]
GARJ730101	Partition coefficient [28]
JOND750102	pK (-COOH) [29]
KARP850103	Flexibility parameter for two rigid neighbors [30]
KHAG800101	The Kerr-constant increments [31]
MAXF760103	Normalized frequency of zeta R [32]
MITS020101	Amphiphilicity index [33]
MONM990201	Averaged turn propensities in a transmembrane helix [34]
NADH010107	Hydropathy scale based on self-information values in the two-state model (50% accessibility) [35]
PRAM820101	Intercept in regression analysis [36]
QIAN880139	Weights for coil at the window position of 6 [37]
RICJ880101	Relative preference value at N" [38]
RICJ880104	Relative preference value at N1 [38]
RICJ880114	Relative preference value at C1 [38]
RICJ880117	Relative preference value at C" [38]
SUEM840102	Zimm-Bragg parameter sigma x 1.0E4 [39]
TANS770102	Normalized frequency of isolated helix [40]
TANS770108	Normalized frequency of zeta R [40]
VASM830101	Relative population of conformational state A [41]
VELV850101	Electron-ion interaction potential [42]
WERD780102	Free energy change of epsilon(i) to epsilon(ex) [43]
WERD780103	Free energy change of alpha(Ri) to alpha(Rh) [43]
WILM950104	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O [44]
YUTK870103	Activation Gibbs energy of unfolding, pH7.0 [45]

Figures 3 and 4 shows the distribution of distances between amino acids according to the physicochemical average value of amino acids among them, i.e. the $(\overline{P}, D)$ points from prediction vectors. For this study, we used the proteins of experiment 4. We include only the distance distributions of two properties (WILM9501040 in Figure 3 and GARJ730101 in Figure 4), but the distributions of other properties are similar. The x-axis represents the normalized value of the physicochemical property and the y-axis represents the distance between amino acids.

As can be seen in Figures 3 and 4, 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.

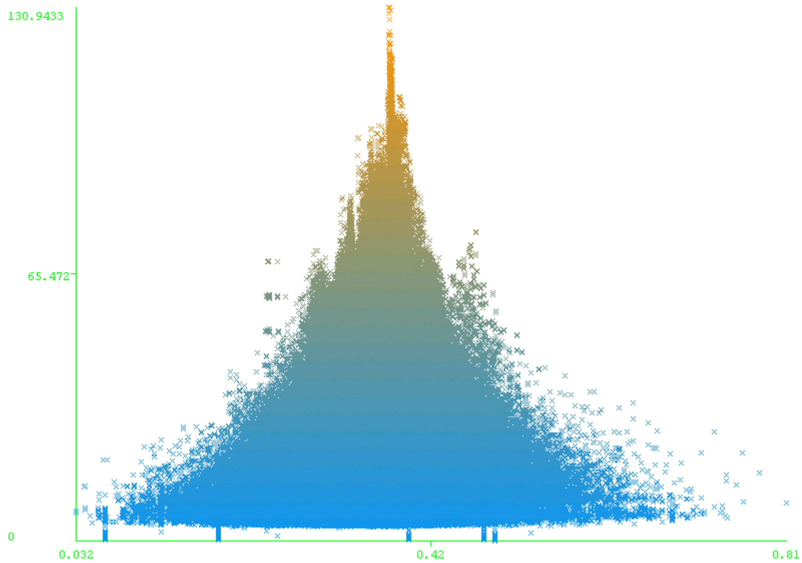


Figure 3: Distance distribution of property WILM950104. The x-axis represents the normalized value of the physicochemical property and the y-axis represents the distance between amino acids.

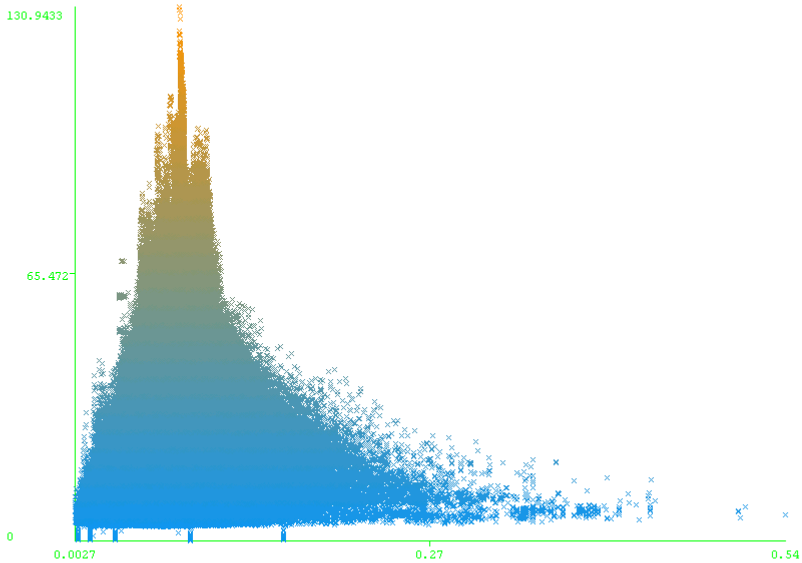
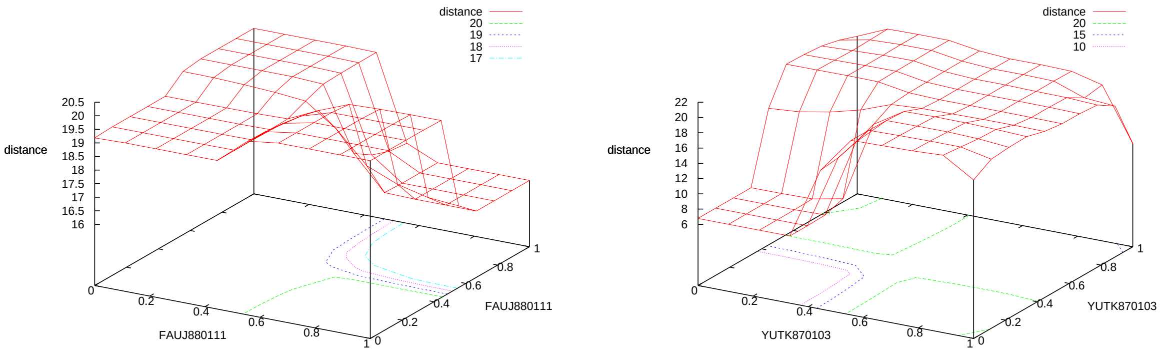


Figure 4: Distance distribution of property GARJ730101. The x-axis represents the normalized value of the physicochemical property and the y-axis represents the distance between amino acids.

We performed another study in order to provide a more detailed representation of the distance distribution according to the nature of the residues. This study shows the mean distances among pairs of amino acids according to their physicochemical values. These distances are represented into a 3D surface, in order to find possible visual patterns. We used harmonic means to represent the distances, in order to mitigate the impact of large distance values. We obtained a 3D surface for each physicochemical property (30 surfaces are generated).

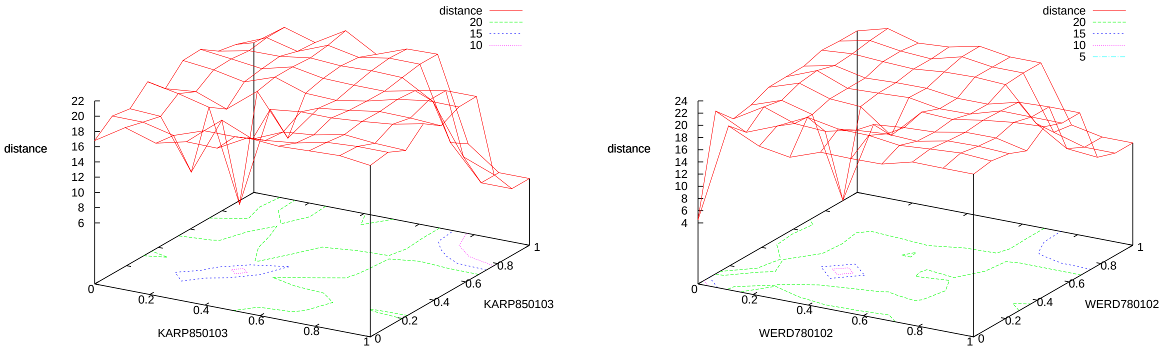
Three main pattern types were found in 3D surfaces. We called "stepped surfaces" to the first pattern type. The surfaces of this type present several areas of different levels of distance. We include two 3D surfaces of this type in Figure 5 (for the properties FAUJ880111 and YUTK870103). The properties FAUJ880112 and MONM990201 present the same behaviour.

The second pattern type has been called "corner-biased surfaces". The surfaces of this type present a low distance values at the same lower or higher physicochemical property values. We



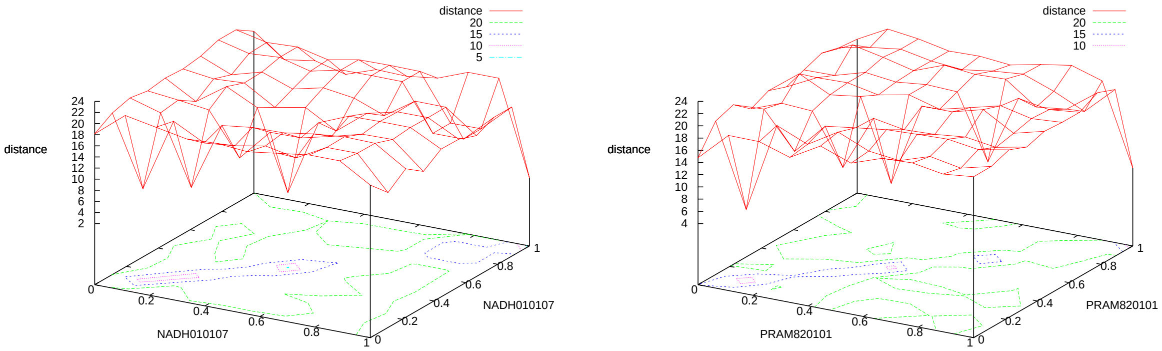
(a) FAUJ880111 (b) YUTK870103

Figure 5: Pattern type: "stepped surfaces". The 3D surface represents the mean distances among pairs of amino acids according to properties (a) FAUJ880111 and (b) YUTK870103.



(a) KARP850103 (b) WERD780102

Figure 6: Pattern type: "corner-biased surfaces". The 3D surface represents the mean distances among pairs of amino acids according to properties (a) KARP850103 and (b) WERD780102.



(a) NADH010107 (b) PRAM820101

Figure 7: Pattern type: "valley surfaces". The 3D surface represents the mean distances among pairs of amino acids according to properties (a) NADH010107 and (b) PRAM820101.

include two 3D surfaces of this type in Figure 6 (properties KARP850103 and WERD780102). There are many other properties that present the same behaviour (QIAN880139, RICJ880101, BUNA790101, BUNA790103, KHAG800101, MAXF760103, RICJ880117 and GARJ730101).

We called “valley surfaces” to the third pattern type. The surfaces of this type present a low distance values at the same physicochemical property values. We include two 3D surfaces of this type in Figure 7 (properties NADH010107 and PRAM820101). There are also many other properties that present the same features (AURR980120, CHAM820102, JOND750102, TANS770108, RICJ880114, VASM830101 and VELV850101).

From the performed study of 3D surfaces, it would be possible to extract rules in order to predict distances between amino acids according to their physicochemical properties. We found that, in most of the studied properties (second and three pattern types), the distances between amino acids with similar properties are low. These properties were AURR980120, CHAM820102, NADH010107, PRAM820101, RICJ880114, SUEM840102, TANS770102, TANS770108, VASM830101, VELV850101 and WILM950104. Therefore, according to the study, our approach seems to be appropriate, since we perform the predictions using fragments with most physicochemical similarities.

4 Experimental results

Five experiments were conducted to test the performance of the ASPPred system. An identical initial configuration was established for all of the experiments, varying only the set of proteins used. In experiments 1, 2, 3 and 4, ten-fold cross validation was used. In experiment 5 we used independent training and test proteins.

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 [46] 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, in order to analyse the predictor behaviour with more protein sequence diversity. We used in experiments 2, 3 and 4 minimum chain lengths (70, 40 and 70 amino acids respectively) in order to avoid small chains which are more easily predicted.

There are other parameters to obtain protein sets: resolution and R-factor. Both parameters measure the quality of the experimental obtaining of proteins. If their values are close to zero, more accurate will be the structure models, but will be less number of available models. With higher quality structure models, less error is introduced in the data. Therefore, the predictor learns knowledge models more adjusted to the real protein structures.

In 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 [47]. 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 [48]. 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.

We used similar configuration in experiments 2 and 4, but in experiment 4 we set sequence identity to 5% in order to analyse results with very different protein sequences. However, we set resolution to 0-1.1 (0.1 more than experiment 2) in order to use more number of proteins. In experiment 3 we decrease the R-factor value to 0-0.12 (0.08 less than experiments 2 and 4) and we increase the resolution to 0-1.4 (0.4 more than experiments 2 and 4), with the aim of analyse the results with high resolution and low R-factor proteins.

In experiment 5 we used as training the proteins of the Experiment 2. As test, in experiment 5 we used all the proteins currently available in PDBselect (most recently in February 2011) with a maximum identity of 25% (5130 proteins).

In Tables 2 and 3 we show the results obtained in experiments. We indicate the mean and standard deviation 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.

Table 2: Efficiency of our method at 4 Å of distance threshold ($\mu \pm \sigma$ values).

Experiment	Recall	Precision	Accuracy	Specificity
1	0.10±0.05	0.08±0.10	0.99±0.00	0.99±0.00
2	0.31±0.10	0.39±0.11	0.99±0.00	0.99±0.00
3	0.48±0.04	0.43±0.05	0.99±0.01	0.99±0.01
4	0.40±0.05	0.41±0.05	0.99±0.01	0.99±0.01
5	0.14±0.08	0.14±0.08	0.99±0.05	0.99±0.05

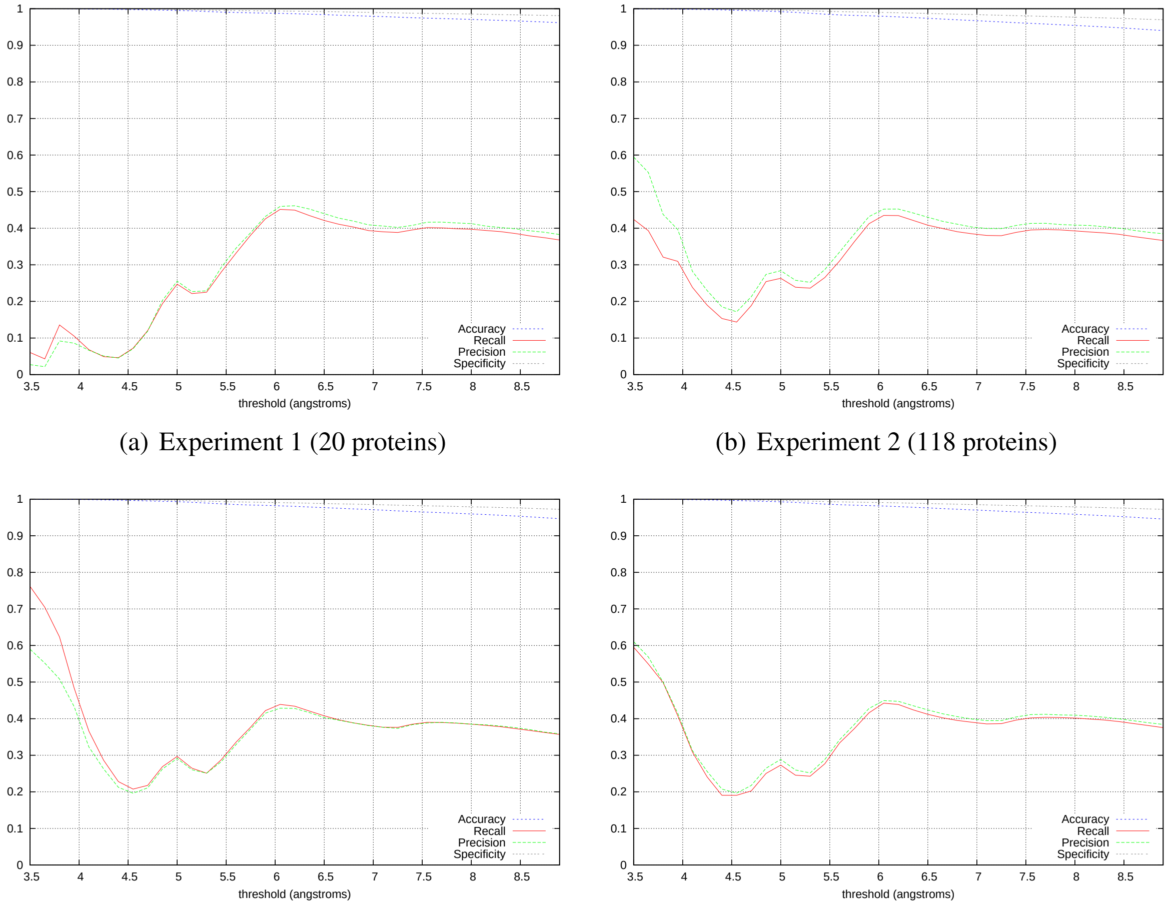
Table 3: Efficiency of our method at 8 Å of distance threshold ($\mu \pm \sigma$ values).

Experiment	Recall	Precision	Accuracy	Specificity
1	0.39±0.06	0.41±0.08	0.97±0.03	0.98±0.01
2	0.39±0.07	0.40±0.07	0.95±0.01	0.97±0.02
3	0.38±0.02	0.38±0.02	0.95±0.02	0.97±0.01
4	0.40±0.03	0.41±0.03	0.95±0.01	0.97±0.01
5	0.51±0.11	0.51±0.11	0.92±0.06	0.95±0.07

For 8 Å of threshold, recall and precision are basically the same in experiments 1 to 4. We found that our predictor no needs many proteins as training. Seems to it find good similar fragments in poor trainings. In experiment 5 we achieved better recall and precision than other experiments; however, standard deviation values are higher. This may be due to the great number of different types of proteins (structural classes or number of domains, for instance) in all the PDBselect.

Thus, we included detailed information about each protein in five experiments as supplemental material at `http://www.upo.es/eps/asencio/asppred`. We indicate for each protein its CATH [49] structural class, CATH ID, CATH description, number of PFAM [50] domains, resolution, chain lengths and, finally, recall and precision values achieved by our predictor. We also included five tables (one for each experiment) that contain means and standard deviations of recall and precision for proteins of each CATH protein structural class.

In order to show the complete results of experiments and to test the efficiency of our algorithm at different thresholds, one graph has been included for experiments 1, 2, 3 and 4 (Figure 8). 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.



(a) Experiment 1 (20 proteins)

(b) Experiment 2 (118 proteins)

(c) Experiment 3 (170 proteins)

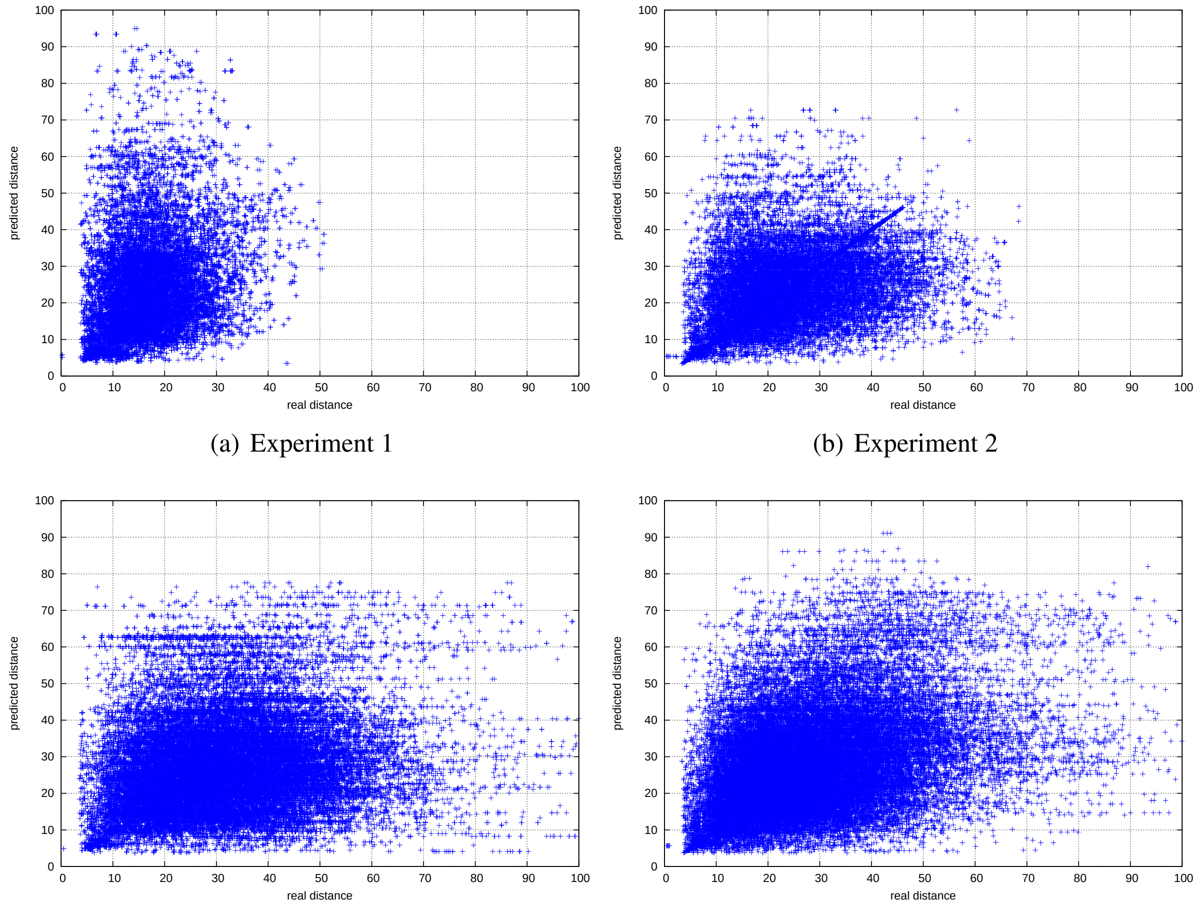
(d) Experiment 4 (221 proteins)

Figure 8: Accuracy, recall, precision and specificity values of the four experiments.

As we can see in Figure 8, with a poor knowledge (experiment 1 with 20 proteins), the quality of prediction for thresholds between 3.5 and 4.8 Å, in terms of recall and precision, is lower than other experiments. 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). However, the behaviour of the measures for higher thresholds to 4.8 Å is similar in all experiments. The main conclusion we extract from Figure 8 is the same prediction ability of our method independent to the number and diversity of proteins.

In order to analyse predictions deeply, we include a plot for experiments 1, 2, 3 and 4 in Figure 9 that shows real and predicted distances for each pair of amino acids.

From Figure 9 we can observe the effect of data volume and predictions trend, from experiment 1 to experiment 4. In experiment 1 there are more predictions such that predicted distance is higher than its real value. Increasing the number of proteins, there are more predictions such that predicted distance is lower than its real value, especially for distances higher than 40 Å.



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

(d) Experiment 4

Figure 9: Real and predicted distances in experiments 1, 2, 3 and 4.

Figure 10 shows the distance map obtained for the protein 3CCD (85 residues) from experiment 2. We used a colour scale to represent the distances, ranging from the minimum distance (red) to the maximum (blue). As shown in the figure 10, the lower triangular of the matrix (prediction) is largely similar to the upper triangular (observation).

Figure 11 shows the contact map of the same protein 3CCD, obtained using the distance map in Figure 10 with a cut-off of 8 angstroms. As distance map, there is great similarity between the real and predicted parts of the contact map.

A very well-known proteins characteristic is that their core contains hydrophobic amino acids that are generally located close among them. In addition, the protein surface consists of non-hydrophobic amino acids that are usually in contact. Therefore, in both cases it is easy to predict their relative distances. In order to analyse if the best predicted distances by our method corresponded to the hydrophobic core of proteins or to their surface, we include the Table 5.

The idea is to check if the error is lower (best predictions) for predicted distances between amino acids with same hydrophobicity (amino acids whose difference of hydrophobicity is close to 0). In this table we show the mean absolute error between real and predicted distances achieved for each range of hydrophobicity difference between all pairs of amino acids. The hydrophobicity corresponds to the property WILM950104. We have taken the proteins of Experiment 4 for this analysis, because it contains the largest number of proteins.

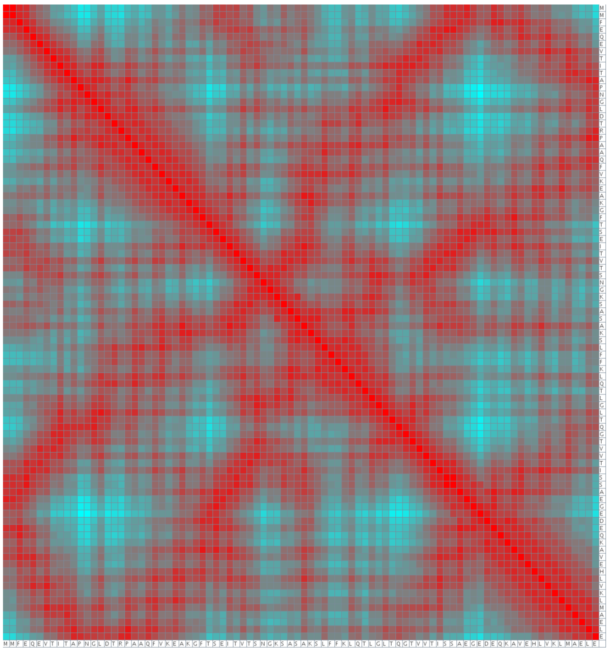


Figure 10: Predicted distance map for the protein 3CCD.

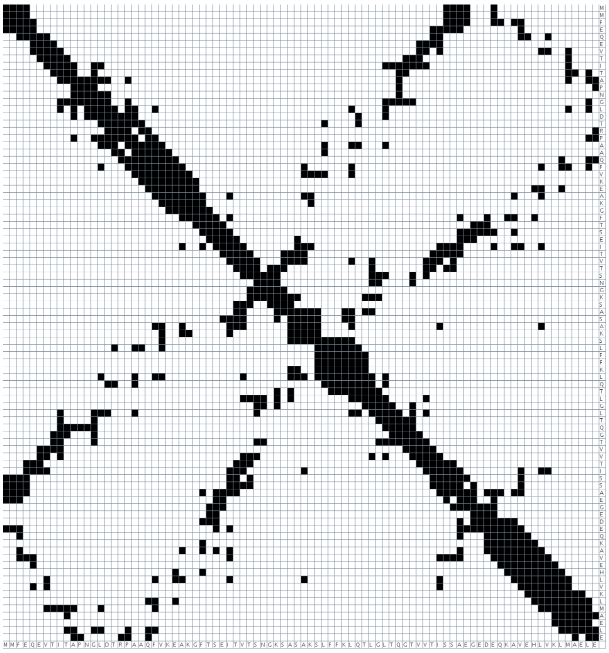


Figure 11: Predicted contact map for the protein 3CCD with a cut-off of 8Å.

As we can see in Table 5, our proposal achieves approximately the same error and deviation regardless of the hydrophobicity of amino acids whose distance is predicted. Furthermore, the error and its deviation decrease when amino acids have more different hydrophobicities. Therefore, our method predicts properly not only the amino acids in the hydrophobic core of proteins or to their surface.

We performed a study of the algorithm efficiency according to the number (K) of most similar training prediction vectors used in the search scheme (see Figure 2). When several prediction vectors are used, we calculate the predicted distance by an arithmetic mean of their distance

Table 4: Study of the number (K) of prediction vectors at 8 Å of distance threshold ($\mu \pm \sigma$ values).

Experiment	K	Recall	Precision	Accuracy	Specificity
1	1	0.39±0.06	0.41±0.08	0.97±0.03	0.98±0.01
	3	0.40±0.05	0.63±0.07	0.97±0.02	0.98±0.00
	5	0.39±0.05	0.73±0.05	0.98±0.02	0.98±0.00
	7	0.38±0.05	0.78±0.05	0.98±0.01	0.98±0.00
	9	0.37±0.04	0.80±0.04	0.98±0.00	0.99±0.00
	11	0.36±0.04	0.83±0.04	0.99±0.00	0.99±0.00
	13	0.35±0.04	0.84±0.04	0.99±0.00	0.99±0.00
	15	0.35±0.04	0.86±0.04	0.97±0.00	0.98±0.00
2	1	0.39±0.07	0.40±0.07	0.95±0.01	0.97±0.02
	3	0.40±0.04	0.73±0.02	0.98±0.01	0.98±0.00
	5	0.39±0.03	0.81±0.01	0.98±0.00	0.99±0.00
	7	0.38±0.03	0.85±0.01	0.99±0.00	0.99±0.00
	9	0.38±0.03	0.86±0.01	0.99±0.00	0.99±0.00
	11	0.37±0.03	0.87±0.01	0.99±0.00	0.99±0.00
	13	0.37±0.03	0.88±0.01	0.99±0.00	0.99±0.00
	15	0.37±0.03	0.88±0.01	0.99±0.00	0.99±0.00
3	1	0.38±0.02	0.38±0.02	0.95±0.02	0.97±0.01
	3	0.40±0.02	0.72±0.01	0.97±0.01	0.98±0.00
	5	0.39±0.01	0.81±0.01	0.98±0.00	0.99±0.00
	7	0.38±0.01	0.84±0.01	0.99±0.00	0.99±0.00
	9	0.38±0.01	0.86±0.01	0.99±0.00	0.99±0.00
	11	0.37±0.01	0.87±0.01	0.99±0.00	0.99±0.00
	13	0.37±0.01	0.88±0.01	0.99±0.00	0.99±0.00
	15	0.37±0.01	0.88±0.01	0.99±0.00	0.99±0.00
4	1	0.40±0.03	0.41±0.03	0.95±0.01	0.97±0.01
	3	0.40±0.04	0.72±0.01	0.96±0.01	0.97±0.00
	5	0.39±0.04	0.81±0.01	0.97±0.00	0.98±0.00
	7	0.39±0.04	0.84±0.00	0.99±0.00	0.99±0.00
	9	0.38±0.04	0.86±0.00	0.99±0.00	0.99±0.00
	11	0.38±0.04	0.87±0.00	0.99±0.00	0.99±0.00
	13	0.37±0.04	0.88±0.00	0.99±0.00	0.99±0.00
	15	0.37±0.04	0.88±0.00	0.99±0.00	0.99±0.00

values. In this study we tested with 1, 3, 5, 7, 9, 11, 13 and 15 prediction vectors. We show in Table 4 the results of this study for the protein data sets of experiments 1, 2, 3 and 4 using 8 Å of distance threshold.

When $K = 1$, the results are naturally the same as in Table 3. As we can observe in Table 4, we notably improved the precision when the parameter K increases, up to 0.88 in experiments 2, 3, 4 and $K \geq 13$. However, recall values decreases when K increases, up to 0.37 in experiments 2, 3, 4 and $K \geq 13$. However, the precision increment is very higher than the recall decrement. Therefore, it seems to be useful to use $K = 15$. For values of K higher than 15, the results are stabilized.

Table 5: Analysis of error according to the difference of hydrophobicities of amino acids whose distance is predicted, using proteins of Experiment 4.

Range of hydrophobicity difference	Mean absolute error	Error standard deviation
$[0, 0.2)$	13.327	9.472
$[0.2, 0.4)$	13.029	9.240
$[0.4, 0.6)$	12.873	9.295
$[0.6, 0.8)$	12.637	9.222
$[0.8, 1]$	12.379	9.131

To statistically validate this study, the results obtained in Table 4 were subjected to a test of statistical significance. In order to use a enough number of examples to perform the statistical test, we used the recall and precision values achieved for each protein in each experiment. Through a D'Agostino-Pearson test [51], it could be confirmed that all the data obtained for this study did not meet the criteria of normality. For this reason, a non-parametric test was selected to statistically validate the differences among the efficiency for each K value in experiments. The non-parametric process followed is described in [52] and involves the use of the Friedman test. Thus, we performed eight Friedman tests, two tests for each experiment: one for recall values and other one for precision values.

After executing the Friedman tests, all the obtained p-values were lower than $1.21 * 10^{-7}$, such that the null hypothesis was rejected. Therefore, the efficiency improvements achieved with different values of K are statistically significant.

In Table 6 we included the execution times (in minutes) of experiments 1, 2, 3 and 4 and K value from the previous study. We indicate in parentheses the number of proteins of each experiment. Preprocess time is necessary at the beginning of each experiment and it is used in all K values. We used a 2 Intel Xeon X5482 3.2GHz with 32GB RAM machine for this study.

Table 6: Execution times (in minutes) of each experiment and K value.

K	Exp 1 (20)	Exp 2 (118)	Exp 3 (170)	Exp 4 (221)
Preprocess (one time)	2.07	3.52	7.56	8.40
1	0.40	15.57	71.12	93.81
3	0.40	15.59	73.94	94.12
5	0.39	15.58	74.23	95.03
7	0.38	15.61	74.55	95.43
9	0.37	15.60	75.14	96.77
11	0.36	15.62	75.53	97.01
13	0.35	15.64	75.62	97.69
15	0.35	15.65	80.43	99.86

5 Conclusions

In this work we have proposed a method in which protein fragments are assembled according to their physicochemical similarities using 30 physicochemical properties of amino acids. Then we predict distance maps, which provide more information about the structure of a protein than contact maps.

As we have tested in the protein data analysis, the distances between amino acids seem to follow a normal distribution, with different means and deviations, according to all the studied physicochemical properties of amino acids among them. We also studied the mean distances among pairs of amino acids according to their physicochemical values. We have represented these distances into 3D surfaces.

We find three main pattern types in 3D surfaces. Therefore, it is possible to extract rules in order to predict distances between amino acids according to their physicochemical properties. Also, we found that, in most of the studied properties (second and three pattern types), the distances between amino acids with similar properties are lower. This is desirable for our approach, since we perform the predictions using the fragments with most physicochemical similarities.

We performed an experimental validation of our method on five non-homologous protein sets obtained from different repositories. We have trained our predictor with a poor training knowledge (experiment 1) and with a higher and sequence diverse (up to 5% of sequence identity) training knowledge (experiments 2, 3 and 4). In addition, we tested our algorithm with all the currently available proteins in PDBselect. In this last experiment we obtained precision and recall of 0.51 and 0.11 as standard deviation with a cut-off of 8 angstroms.

We found that, with a poor knowledge (experiment 1 with 20 proteins), the quality of prediction is low, in terms of recall and precision, for lower thresholds. This behaviour may have been due to the low number of training proteins and, consequently, to the low achieved knowledge of the search space (native protein structures).

We have performed a study of the algorithm efficiency according to the number (K) of most similar training prediction vectors used in the search scheme. We notably improved the precision when the parameter K increases, from 0.38 to 0.88 in experiment 3. However, recall values decreases when K increases, from 0.40 to 0.37 in experiment 4. These improvements are statistically significant. Therefore, it is useful to use $K = 15$ in order to predict distances between amino acids.

Finally, we found empirically that the response of our method over large 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 had very low identity. These results are desirable, in theory, since this study sought generality of the method.

References

[1] C.A. Floudas. Computational methods in protein structure prediction. *Biotechnol Bioeng*, 97(2):207–213, 2007.

[2] Ciro Pierri, Anna De Grassi, and Antonio Turi. Lattices for ab initio protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 73(2):351–361, 2008.

[3] C.A. Floudas, H.K. Fung, S.R. Mcallister, M. Monnigmann, and R. Rajgaria. Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, 61(3):966–988, 2006.

[4] F. Melo and A. Sali. Fold assessment for comparative protein structure modeling. *Protein science : a publication of the Protein Society*, 16(11):2412–2426, 2007.

[5] Krzysztof Ginalski. Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology*, 16(2):172 – 177, 2006.

[6] Shuai Cheng Li, Dongbo Bu, Jinbo Xu, and Ming Li. Fragment-hmm: a new approach to protein structure prediction. *Protein science : a publication of the Protein Society*, 17(11):1925–1934, 2008.

[7] J.M. Bujnicki. Protein-structure prediction by recombination of fragments. *Chembiochem*, 7(1):19–27, 2006.

[8] D. T. Jones, K. Bryson, A. Coleman, L. J. McGuffin, M. I. Sadowski, J. S. Sodhi, and J. J. Ward. Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins: Structure, Function, and Bioinformatics*, 61(S7):143–151, 2005.

[9] 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.

[10] D.T. Jones. Predicting novel protein folds by using fragfold. *Proteins*, Suppl 5:127–132, 2001.

[11] T. Herges and W. Wenzel. An all-atom force field for tertiary structure prediction of helical proteins. *Biophysical Journal*, 87(5):3100 – 3109, 2004.

[12] M. Scott Shell, S. Banu Ozkan, Vincent Voelz, Guohong Albert Wu, and Ken A. Dill. Blind test of physics-based prediction of protein structures. *Biophysical Journal*, 96(3):917 – 924, 2009.

[13] Takeshi N. Sasaki, Hikmet Cetin, and Masaki Sasai. A coarse-grained langevin molecular dynamics approach to de novo protein structure prediction. *Biochemical and Biophysical Research Communications*, 369(2):500 – 506, 2008.

[14] 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.

[15] Narjes Khatoon Habibi and Mohammad Hossein Saraee. Protein contact map prediction based on an ensemble learning method. *Computer Engineering and Technology, International Conference on*, 2:205–209, 2009.

[16] Allison Tegge, Zheng Wang, Jesse Eickholt, and Jianlin Cheng. Nncon: improved protein contact map prediction using 2d-recursive neural networks. *Nucleic acids research*, 37(Web Server issue), 2009.

[17] G. Tradigo. On the integration of protein contact map predictions. In *22nd IEEE International Symposium on Computer-Based Medical Systems, 2009. CBMS 2009.*, pages 1 –5, 2-5 2009.

[18] I Walsh, D Baù, AJ Martin, C Mooney, A Vullo, and G Pollastri. Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. *BMC structural biology*, 9:5, 2009.

[19] Andrzej Kloczkowski, Robert Jernigan, Zhijun Wu, Guang Song, Lei Yang, Andrzej Kolinski, and Piotr Pokarowski. Distance matrix-based approach to protein structure prediction. *Journal of structural and functional genomics*, 10(1):67–81, 2009.

[20] 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.

[21] K.-L. Lin, Chun-Yuan Lin, Chuen-Der Huang, Hsiu-Ming Chang, Chiao-Yun Yang, Chin-Teng Lin, Chuan Yi Tang, and D.F. Hsu. Feature selection and combination criteria for improving accuracy in protein structure prediction. *NanoBioscience, IEEE Transactions on*, 6(2):186 –196, june 2007.

[22] Marko Robnik-Sikonja and Igor Kononenko. An adaptation of relief for attribute estimation in regression. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 296–304, 1997.

[23] R. Aurora and G. Rose. Helix capping. *Protein Science*, 7:21–38, 1998.

[24] Arno Bundi and Kurt Wuthrich. 1h-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides h-gly-gly-x-l-ala-oh. *Biopolymers*, 18(2):285–297, 1979.

[25] Marvin Charton and Barbara I. Charton. The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology*, 99(4):629 – 644, 1982.

[26] Di Giulio M. A comparison of proteins from pyrococcus furiosus and pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code. *Gene*, 346:1–6, 2005.

[27] Jean-Luc Fauchare, Marvin Charton, Lemont B. Kier, Arie Verloop, and Vladimir Pliska. Amino acid side chain parameters for correlation studies in biology and pharmacology. *International Journal of Peptide and Protein Research*, 32(4):269–278, 1988.

[28] Filliol D. Garel, J.P. and P. Mandel. Coefficients de partage d'aminoacides, nucleobases, nucleosides et nucleotides dans un systeme solvant salin. *J. Chromatogr.*, 78:381–391, 1973.

[29] Daniel D. Jones. Amino acid properties and side-chain orientation in proteins: A cross correlation approach. *Journal of Theoretical Biology*, 50(1):167 – 183, 1975.

[30] P.A. Karplus and G.E. Schulz. Prediction of chain flexibility in proteins. *Naturwiss*, 72:212–213, 1985.

[31] G. Khanarian and W.J. Moore. The kerr effect of amino acids in water. *Aust. J. Chem.*, 33:1727–1741, 1980.

[32] F.R. Maxfield and H.A. Scheraga. Status of empirical methods for the prediction of protein backbone topography. *Biochemistry*, 15:5138–5153, 1976.

[33] Hirokawa T. Mitaku, S. and T. Tsuji. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics*, 18:608–616, 2002.

[34] Nilsson I. Elofsson A. Monne, M. and G. von Heijne. Turns in transmembrane helices: determination of the minimal length of a ”helical hairpin” and derivation of a fine-grained turn propensity scale. *J. Mol. Biol.*, 293:807–814, 1999.

[35] Sadeghi M. Arab S. Naderi-Manesh, H. and A.A. Moosavi Movahedi. Prediction of protein surface accessibility with information theory. *Proteins*, 42:452–459, 2001.

[36] M. Prabhakaran and P.K. Ponnuswamy. Shape and surface features of globular proteins. *Macromolecules*, 15:314–320, 1982.

[37] N. Qian and T.J. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, 202:865–884, 1988.

[38] J.S. Richardson and D.C. Richardson. Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240:1648–1652, 1988.

[39] Lee S. Powers S.P. Denton J.B. Konishi Y. Sueki, M. and H.A. Scheraga. Helix-coil stability constants for the naturally occurring amino acids in water. *Macromolecules*, 17:148–155, 1984.

[40] S. Tanaka and H.A. Scheraga. Statistical mechanical treatment of protein conformation. 5. a multiphasic model for specific-sequence copolymers of amino acids. *Macromolecules*, 10:9–20, 1977.

[41] Nemethy G. Vasquez, M. and H.A. Scheraga. Computed conformational states of the 20 naturally occurring amino acid residues and of the prototype residue alpha-aminobutyric acid. *Macromolecules*, 16:1043–1049, 1983.

[42] Cosic I. Dimitrijevic B. Veljkovic, V. and D. Lalovic. Is it possible to analyze dna and protein sequences by the method of digital signal processing? *IEEE Trans. Biomed. Eng.*, 32:337–341, 1985.

[43] D.H. Wertz and H.A. Scheraga. Influence of water on protein structure. an analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules*, 11:9–15, 1978.

[44] Aguilar M.I. Wilce, M.C. and M.T. Hearn. Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from rp-hplc of peptides. *Anal Chem.*, 67:1210–1219, 1995.

[45] Ogasahara K. Tsujita T. Yutani, K. and Y. Sugino. Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit. *Proc. Natl. Acad. Sci. USA*, 84:4441–4444, 1987.

[46] 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.

[47] Guoli Wang and Roland Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics (Oxford, England)*, 19(12):1589–1591, 2003.

[48] Sven Griep and Uwe Hobohm. Pdbselect 1992-2009 and pdbfilter-select. *Nucl. Acids Res.*, 38(suppl1):D318–319, 2010.

[49] Christine A Orengo, James E Bray, Daniel W A Buchan, Andrew Harrison, David Lee, Frances M G Pearl, Ian Sillitoe, Annabel E Todd, and Janet M Thornton. The cath protein family database: a resource for structural and functional annotation of genomes. *Proteomics*, 2:11–21, 2002.

[50] Robert D. Finn, John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. The pfam protein families database. *Nucleic Acids Research*, 36(suppl 1):D281–D288, 2008.

[51] Ralph B. D'Agostino, Albert Belanger, and Ralph B. D'Agostino Jr. A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4):316–321, November 1990.

[52] S. Garcia and F. Herrera. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2008.