Prediction of Mitochondrial Matrix Protein Structures Based on Feature Selection and Fragment Assembly

Gualberto Asencio-Cortés[1], Jesús S. Aguilar-Ruiz[1], Alfonso E. Márquez-Chamorro[1], Roberto Ruiz[1], and Cosme E. Santiesteban-Toca[2]

[1] School of Engineering, Pablo de Olavide University, Seville, Spain
{guaasecor,aguilar,amarcha,robertoruiz}@upo.es
[2] Centro de Bioplantas, University of Ciego de Ávila, Cuba
cosme@bioplantas.cu

Abstract. Protein structure prediction consists in determining the three-dimensional conformation of a protein based only on its amino acid sequence. This is currently a difficult and significant challenge in structural bioinformatics because these structures are necessary for drug designing. This work proposes a method that reconstructs protein structures from protein fragments assembled according to their physico-chemical similarities, using information extracted from known protein structures. Our prediction system produces distance maps to represent protein structures, which provides more information than contact maps, which are predicted by many proposals in the literature. Most commonly used amino acid physico-chemical properties are hydrophobicity, polarity and charge. In our method, we performed a feature selection on the 544 properties of the AAindex repository, resulting in 16 properties which were used to predictions. We tested our proposal on 74 mitochondrial matrix proteins with a maximum sequence identity of 30% obtained from the Protein Data Bank. We achieved a recall of 0.80 and a precision of 0.79 with an 8-angstrom cut-off and a minimum sequence separation of 7 amino acids. Finally, we compared our system with other relevant proposal on the same benchmark and we achieved a recall improvement of 50.82%. Therefore, for the studied proteins, our method provides a notable improvement in terms of recall.

Keywords: Protein structure prediction, physico-chemical amino acid properties, fragment assembly, protein distance map, feature selection.

1 Introduction

Knowing the protein native 3D structures is currently a difficult and significant challenge because these structures determine protein function and they are necessary to design new drugs. Experimental methods to determine protein structures, generally X-ray crystallography and nuclear magnetic resonance, are very expensive and they have limitations with the structures of some proteins. Moreover,

the great number of protein sequences whose three-dimensional structures must be determined, make computational methods of protein structure prediction a very useful tool.

Protein structure prediction (PSP) consists in determining a three-dimensional model based only on the amino acid sequence of a protein and it is currently an issue with great significance in structural bioinformatics [1].

There are currently two main approaches for the PSP problem. The first is the ab initio methods, which find the structure that corresponds to a global minimum of a function, generally a energy function, based in sequence properties. These methods do not use any protein as a template, their computational cost is generally very high and their reliability decreases when the sequence length increases [2].

The second main approach is homology methods, also known as comparative modeling, which try to solve the structure based on protein templates (template-based modeling). This approach is based on the structural conservation of proteins in a protein family, since the 3D structures are more conserved in evolution than sequences. These methods are considered the most currently reliable approach for PSP problem [2].

Template-based modeling 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-modeling is used.

Within free-modeling methods we find the fragment assembly methods that reconstruct the structure of a protein from structural fragments of other proteins. Three of most relevant fragment assembly-based methods are Fragment-HMM [3], FragFold [4] and ROSETTA [5]. 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 minimizing the corresponding energy function.

Since all information used in structure prediction must be inferred from amino acid sequence, there is many useful information derived from sequence used in the literature. Among this information, there are recent methods that use a great set of physico-chemical properties of amino acids [6]. However, the most commonly used properties are hydrophobicity, polarity and charge, which are used, for example, in the models HP and HPNX [7]. There is a database of amino acid properties named AAindex [8] which contains currently 544 properties, from which we selected a subset of 16 in this work by a feature selection process.

The motivation for applying feature selection (FS) techniques has shifted from being optional to becoming a real prerequisite for model building. Specifically, in the PSP problem, the feature selection was also applied and improves the accuracy of predictions [9]. Theoretically, having more features should give us more discriminating power. However, this can cause several problems: increased computational complexity and cost; too many redundant or irrelevant features; and estimation degradation in the classification error.

Based on the generation procedure, FS can be divided into individual feature ranking (FR) and feature subset selection (FSS) [10,11]. FR measures feature-class relevance, then ranks features by their scores and selects the top-ranked ones. These methods are widely used due to their simplicity, scalability, and good empirical success [12]. However, FR is criticized because it can only capture the relevance of features to the target concept, while redundancy and basic interactions between features are not revealed. Furthermore, it is difficult to determine the number of features retained, because a threshold is required. In contrast, FSS attempts to find a set of features that performs well. It integrates the metrics for measuring feature-class relevance and feature-feature interactions.

In this work, a hybrid algorithm was used, BARS [13], in order to handle large datasets to take advantage of the above two approaches (FR, FSS) [14]. This method decouple relevance analysis and redundancy analysis, and have proven to be more effective than ranking methods and more efficient than subset evaluation methods in many traditional high-dimensional datasets.

There are many PSP algorithms currently in the literature that produce a contact map to represent the predicted structure [6,15]. 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. There are fewer proposals in the literature that predict distance maps [16], because it is more difficult to perform regression than classification (continuous distances instead binary contacts). Some authors discretize the distances to predict, providing an intermediate representation between contacts and continuous distances, such as the proposal of Walsh et al. 2009 [2] which uses 4-class distance maps. However, unlike 3D models, both distance and contact maps have the desirable property of being insensitive to rotation or translation of the protein molecule.

Our method is a free-modeling approach based on fragment assembly that selects the best distances between pairs of amino acids using fragments of known structures of proteins. These fragments are chosen through a searching process for nearest neighbors by similarity in 16 physico-chemical properties of amino acids selected from the AAindex repository.

We tested our methodology by performing predictions on mitochondrial matrix proteins from the Protein Data Bank (PDB) [17] with a maximum sequence identity of 30%. We have performed predictions with a minimum sequence separation of 7 amino acids, as has been used in the literature [18]. Finally, we compared our system with RBFNN method proposed by Zhang et al. 2005 [19] with the same proteins in the same experimental conditions.

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

2 Methods

2.1 Representation of Protein Structures

The representation of protein structure that we used is the distance map, which is a square matrix of order L, where L is the number of amino acids in the protein 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 actual distance measured in angstroms (Å) between the amino acids i^{th} and j^{th} in the sequence. To measure the distances between amino acids, it is necessary to use a reference atom of each amino acid. The most commonly used reference atoms are the alpha carbon and the beta carbon of amino acids [18]. 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 amino acids i^{th} and j^{th} of the protein sequence.

2.2 Construction of Protein Fragments Knowledge Base

Our prediction system ASPpred (Amino acid Subsequences Properties-based Predictor), works in two phases. In the first phase, it constructs a gallery of protein fragments from all the subsequences of all the proteins in the training set. In the second phase, the target structures of the proteins in test set were predicted using the generated protein fragments model.

The knowledge base consists of a set of vectors called prediction vectors. Each one of these vectors was obtained from one training protein subsequence and contains the physico-chemical properties of the amino acids ends of such fragment. The vector also contains the actual distance between them.

In order to define our prediction vectors formally, it is necessary to define the following elements. In first place, an amino acid sequence of length L is defined by $s_1 \dots s_L$. A fragment or subsequence into a sequence is represented by $s_1 \dots s_b \dots s_e \dots s_L$, where $s_b \dots s_e$ is the fragment, s_b is the beginning amino acid of the fragment, s_e is its ending amino acid and $1 \leq b < e \leq L$.

Moreover, physico-chemical properties are defined by $P_1 \dots P_m$, where m is the number of properties used by the algorithm. The value of the property P_i of an amino acid s_j is defined by $P_i(s_j)$. The prediction vector of a fragment is defined by the tuple showed in Equation 1.

$$\{B_1, E_1, \dots, B_m, E_m, D\} \tag{1}$$

Where D is the distance between amino acids s_b and s_e. B_i and E_i are defined in Equations 2 and 3, respectively. B_i represents the physico-chemical distribution of the entire sequence with decreasing weighting starting at the first amino acid of the fragment. E_i is analogous to B_i starting at the last amino acid of the fragment.

$$B_i = P_i(s_b) + \sum_{\substack{j=1 \\ j \neq b}}^{L} \frac{P_i(s_j)}{L|b-j|}, \forall i \in \{1..m\} \quad (2)$$

$$E_i = P_i(s_e) + \sum_{\substack{j=1 \\ j \neq e}}^{L} \frac{P_i(s_j)}{L|e-j|}, \forall i \in \{1..m\} \quad (3)$$

Note that prediction vectors represent fragments of different lengths, but these lengths is not included in them. The physico-chemical properties included in the prediction vectors are explained in the next subsection. From the point of view of data mining, B_i and E_i are the attributes of training instances and D the class to predict.

2.3 Physico-chemical Feature Selection

To the aim of using the smallest and most effective set of physico-chemical properties, we performed a feature selection from the repository AAindex of physico-chemical properties of amino acids. This repository currently contains 544 amino acid properties.

We used BARS to perform the feature selection over all the properties in AAindex. BARS is an agglomerative algorithm due to the way it constructs the final subset of selected features. The method begins by generating a ranking. Then, pairs of features are obtained with the ranking's first features, in combination with each one of the remaining features on the list. The pairs of features are ranked according to the value of the evaluation, and the process is repeated, that is, the subsets made up by the first sets on the new list are compared with the rest of the sets. At the end, the algorithm returns the best positioned feature subset of all the subsets evaluated.

BARS can use any measure to evaluate feature subsets. Taken into account this domain with a numeric class attribute where distance between amino acids is represented, we used linear regression as evaluator criteria when the search process is carried out to find a relevant and not redundant subset of features.

The dataset that we used for the feature selection is published by Fariselli et al. 2001 [18], that contains 173 proteins with a sequence identity lower than 25%, without chain breaks and with alignments with more than 15 sequences in the corresponding families. This process results on 16 physico-chemical properties that are showed in Table 1 with the same name and description used in AAindex.

2.4 Structure Reconstruction

The second phase of our system consists in obtaining the prediction vectors of the target proteins and in performing a full sequential search to compare each test prediction vector with the training prediction vectors achieved in the first phase. The objective was to find the training prediction vector most similar to

Table 1. The 16 physico-chemical properties of amino acids considered from AAindex

CHOC760104	Proportion of residues 100% buried
LEVM760104	Side chain torsion angle phi(AAAR)
MEIH800103	Average side chain orientation angle
PALJ810107	Normalized frequency of alpha-helix in all-alpha class
QIAN880112	Weights for alpha-helix at the window position of 5
WOLS870101	Principal property value z1
ONEK900101	Delta G values for the peptides extrapolated to 0 M urea
BLAM930101	Alpha helix propensity of position 44 in T4 lysozyme
PARS000101	p-Values of mesophilic proteins based on the distributions of B values
NADH010102	Hydropathy scale based on self-information values in the two-state model (9% accessibility)
SUYM030101	Linker propensity index
WOLR790101	Hydrophobicity index
JACR890101	Weights from the IFH scale
MIYS990103	Optimized relative partition energies - method B
MIYS990104	Optimized relative partition energies - method C
MIYS990105	Optimized relative partition energies - method D

each test prediction vector. For the search process, we consider only training vectors with the same amino acid ends (first and last of each subsequence) than the test vectors. Figure 1 shows this search scheme.

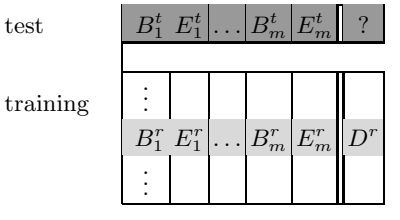


Fig. 1. Search for the most similar training prediction vector

In the search scheme of the Figure 1, $B_1^t \dots B_m^t$ and $E_1^t \dots E_m^t$ are the elements of the test subsequence explained above and $B_1^r \dots B_m^r$ and $E_1^r \dots E_m^r$ are those of the training subsequence with more similarity to the test subsequence. The distance field D^r of the most similar training vector is assigned to the distance field (symbolized with ?) of test vector.

The training vector with the highest similarity to a test vector satisfies the condition showed in Equation 4. As can be seen in that condition, for the comparison of the test and training vectors, an Euclidean distance is used, which includes all the attributes in these vectors with same weigths. All these attributes are normalised previously. The normalization ensured that all of the attributes are on the same scale and contributed equally to the prediction.

$$\min_{r \in TrainingSet} \sqrt{\sum_{i=1}^{m} (B_i^t - B_i^r)^2 + \sum_{i=1}^{m} (E_i^t - E_i^r)^2} \tag{4}$$

Finally, once predicted distances are assigned in test vectors, these distances are stored in the lower triangular of the distance map of the test sequence. Specifically, the distance field of the prediction vector of the subsequence $s_b \ldots s_e$ is assigned to the position (e, b) of the distance map. Thus the structure of each target sequence, by its distance map, is reconstructed.

2.5 Evaluation of Predicted Models

We used several measures to evaluate the quality of the predictions. The first measure is the precision, that is the percentage of predicted contacts that are present in the native structure. This measure is largely used in the literature of protein structure prediction, as in the works of Fariselli et al. [18,20]. The second one is the recall, that is the percentage of native contacts that are predicted to be contacts. Recall has also been widely used in other protein prediction methods [19]. 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 percentages [21]. The following formulas (5,6,7,8,9) define these five measures.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

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 [18,20,19]. 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, as in Fariselli et al. [18].

3 Experimentation and Results

3.1 Prediction of Mitochondrial Matrix Proteins

We performed an experimental validation of our predictor using all mitochondrial matrix proteins (GO ID: 5759) published in the PDB with a maximum identity of 30% (non-homologous proteins) at October 2011 (74 proteins with a maximum length of 1094 amino acids). In Table 2, we show the PDB codes of the proteins used in this study. We classified proteins in three groups of sequence length L ($L \leq 300$, $300 < L \leq 450$ and $L > 450$) in order to show the prediction behavior for each sequence length interval.

Table 2. Mitochondrial matrix proteins used to train and test our predictor

$L \leq 300$	2CW6	1CSH	2DFD	3CMQ	1PJ3	3C5E
1BWY	2GRA	1D2E	2E0A	3EXE	1WDK	3DLX
1EFV	2HDH	1F0Y	2IB8	3GH0	1WLE	3E04
1KKC	2O23	1GKZ	2IZZ	3KGW	1ZMD	3IHJ
1MJ3	2WYT	1HW4	2OAT	7AAT	2FGE	3IKL
1QQ2	3ED7	1I4W	2QB7	$L > 450$	2J6L	3IKM
1R4W	3EMN	1OTH	2QFY	1A4E	2JDI	3MW9
1RHS	3QUW	1RX0	2R2N	1CJC	2UXW	3N9Y
1TG6	3ULL	1W6U	3AF0	1G5H	2WYA	3OEE
1XX4	5CYT	2A1H	3BLX	1HR6	2XIJ	3OU5
1ZD8	$L300 - 450$	2BFD	3BPT	1OHV	2ZT5	3SPA

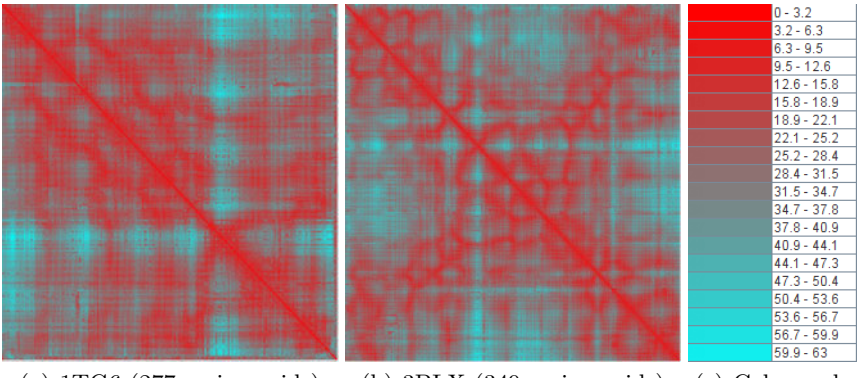
A cross-validation was performed over each group of proteins and over the all 74 proteins. We used a leave-one-out scheme in order to avoid the effect of fold choice in a cross-validation with folds. Table 3 shows the five evaluation measures obtained in this experiment.

We achieved a recall value of 0.80 and a precision of 0.79 for the complete group of proteins, as shown in Table 3. We obtain best predictions, in terms of recall and precision, with proteins of length between 300 and 450 amino acids. In this group of proteins, we achieved recall of 0.84 and precision of 0.83.

In most cases the precision obtained in predicting proteins with long sequences (more than 300 amino acids) is lower than with proteins of short sequences. For example, in the work of Fariselli et al. 2001 [18], which also uses a cross validation, cut-off of 8 angstroms and minimum sequence separation of 7 amino acids, achieved a precision value of 0.11 for proteins of more than 300 amino acids.

Table 3. Efficiency of our method predicting mitochondrial matrix proteins

Protein set	Recall	Precision	Accuracy	Specificity	MCC
All proteins (74)	0.80	0.79	0.97	0.97	0.82
$L \leq 300$ (20)	0.77	0.76	0.98	0.98	0.75
$300 < L \leq 450$ (27)	0.84	0.83	0.99	0.99	0.83
$L > 450$ (27)	0.77	0.76	0.95	0.95	0.82



(a) 1TG6 (277 amino acids) (b) 3BLX (349 amino acids) (c) Color scale

Fig. 2. Predicted distance maps for the mitochondrial matrix proteins 1TG6 (a) and 3BLX (b) with their color scale (c)

Figure 2 shows the predicted distance maps for protein 1TG6 (277 amino acids) and 3BLX (349 amino acids) from the study set. We show a color scale to represent the distances, ranging from the minimum (red) to the maximum (blue) distance. We can appreciate in these distance matrices that the lower triangular (predicted distances) is largely similar to the upper triangular (real distances).

3.2 Comparison with RBFNN on the Same Benchmark

In order to assess the quality of the predictions obtained with our method and to validate our predictor, we compared our proposal with RBFNN method proposed by Zhang et al. 2005 [19]. We predicted the same test proteins with the same training sets in the same conditions.

Zhang et al. 2005 used recall (namely accuracy (A_p) by the authors), predicted and desired numbers to evaluate the performance. Predicted numbers N_p is the count of the predicted contacts by the algorithm and desired numbers N_d is the total number of contacts. The contact threshold was set at 8 Å.

In Table 4 we show the results of this experimentation. As we can see in this table, the average recall (A_p) of ASPpred is 50.82% higher than RBFNN.

Table 4. Comparison at 8 Å with RBFNN on the same benchmark

PDB code (length)	RBFNN			ASPpred		
	N_p	N_d	A_p	N_p	N_d	A_p
1TTF (94)	376	1421	26.46	1307	1421	91.96
1E88 (160)	1006	3352	30.01	3075	3352	91.73
1NAR (290)	3346	10524	31.79	1797	10524	17.07
1BTJ_B (337)	3796	14283	26.58	14026	14283	98.20
1J7E (458)	6589	25026	26.33	23407	25026	93.53
Average			27.67			78.49

N_p: predicted numbers; N_d: desired numbers; A_p: prediction accuracy (%).

Only the protein 1NAR is poorly predicted because there is only one protein as training in the benchmark and it seems to be insufficient to build an effective knowledge base of protein fragments. Thus on the same benchmark dataset, ASPpred yields a sizable improvement.

4 Conclusions and Future Work

In this work we have proposed a method in which protein fragments are assembled according to their physico-chemical similarities, using 16 physico-chemical properties of amino acids selected from AAindex by the BARS feature selection algorithm. We have predicted distance maps, which provide more information about the structure of a protein than contact maps. We have performed an experimental validation of the method on all non-homologous mitochondrial matrix proteins currently available in PDB. We have achieved a recall of 0.80 and a precision of 0.79 with an 8-angstrom cut-off and a minimum sequence separation of 7 amino acids. Finally, we have compared our system with RBFNN method proposed by Zhang et al. 2005 on the same benchmark and we have achieved a recall improvement of 50.82%. Therefore we achieved a significant improvement over previous algorithms.

As future work, we propose to use other prediction vector definitions, including more specific descriptors of the fragment that represent, as amino acid windows. We will also include in these vectors information of the secondary structure of the fragment and its solvent accessibility. We are designing feasibility measures for the geometry derived from predicted distance maps and adjustment algorithms in order to improve our results.

References

1. Zhou, Y., Duan, Y., Yang, Y., Faraggi, E., Lei, H.: Trends in template/fragment-free protein structure prediction. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 128, 3–16 (2011)

2. Walsh, I., Bau, D., Martin, A., Mooney, C., Vullo, A., Pollastri, G.: Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Structural Biology 9(1), 5 (2009)
3. 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)
4. Jones, D.T.: Predicting novel protein folds by using fragfold. Proteins (suppl.5), 127–132 (2001)
5. 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 (2004)
6. Li, Y., Fang, Y., Fang, J.: Predicting residue-residue contacts using random forest models. Bioinformatics (2011)
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. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.: Aaindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36(Database issue), D202–D205 (2008)
9. Lin, K.-L., Lin, C.-Y., Huang, C.-D., Chang, H.-M., Yang, C.-Y., Lin, C.-T., Tang, C.Y., Hsu, D.F.: Feature selection and combination criteria for improving accuracy in protein structure prediction. IEEE Transactions on NanoBioscience 6(2), 186–196 (2007)
10. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245–271 (1997)
11. Guyon, I.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
12. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
13. Ruiz, R., Riquelme, J.C., Aguilar-Ruiz, J.S.: Best agglomerative ranked subset for feature selection. Journal of Machine Learning Research - Proceedings Track 4, 148–162 (2008)
14. Yu, L., Liu, H., Guyon, I.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224 (2004)
15. Wu, S., Szilagyi, A., Zhang, Y.: Improving protein structure prediction using multiple sequence-based contact predictions. Structure 19(8), 1182–1191 (2011)
16. Kloczkowski, A., Jernigan, R., Wu, Z., Song, G., Yang, L., Kolinski, A., Pokarowski, P.: Distance matrix-based approach to protein structure prediction. Journal of Structural and Functional Genomics 10, 67–81 (2009)
17. 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)
18. 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)

19. 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)
20. Fariselli, P., Casadio, R.: A neural network based predictor of residue contacts in proteins. Protein Engineering 12(1), 15–21 (1999)
21. 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)