Available online at www.sciencedirect.com

ScienceDirect

BBRC

ELSEVIER

Biochemical and Biophysical Research Communications 355 (2007) 764–769

www.elsevier.com/locate/ybbrc

Prediction of protein crystallization using collocation of amino acid pairs

Ke Chen, Lukasz Kurgan [*], Mandana Rahbari

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alta., Canada

Received 30 January 2007
Available online 15 February 2007

Abstract

While above 80% of protein structures in PDB were determined using X-ray crystallography, in some cases only 42% of soluble purified proteins yield crystals. Since experimental verification of protein's ability to crystallize is relatively expensive and time-consuming, we propose a new in silico prediction system, called CRYSTALP, which is based on the protein's sequence. CRYSTALP uses a novel feature-based sequence representation and applies a Naïve Bayes classifier. It was compared with recent, competing in silico method, SECRET [P. Smialowski, T. Schmidt, J. Cox, A. Kirschner, D. Frishman, Will my protein crystallize? A sequence-based predictor, Proteins 62 (2) (2006) 343–355], and other state-of-the-art classifiers. Based on experimental tests, CRYSTALP is shown to predict crystallization with 77.5% accuracy, which is better by over 10% than the SECRET's accuracy, and better than accuracy of the other considered classifiers. CRYSTALP uses different and over 50% less features to represent sequences than SECRET. Additionally, features used by CRYSTALP may help to discover intra-molecular markers that influence protein crystallization.

Keywords: Protein crystallization; X-ray crystallography; Collocated amino acid pairs; Classification; CRYSTALP; Naïve Bayes

Proteins are nano-scale machines that catalyze chemical reactions (enzymes), form the cytoskeleton (tubulin), perform transporting functions (hemoglobin), implement immune responses (antibodies), regulate cell processes (hormones), etc. Their wide range of functionality is due to ability to adopt a huge number of three-dimensional shapes, although they are all assembled using only 20 building blocks, called amino acids (AAs). Four distinct levels of protein structure, i.e., primary (linear) sequence of AAs, secondary, tertiary and quaternary structure, are usually distinguished. Knowledge of the tertiary (three-dimensional) protein structure is of pivotal importance to the understanding and manipulation of protein's biochemical and cellular functions.

The protein sequence can be deduced from known DNA sequence, and can be learned based on Edman degradation and mass spectrometry methods, which are relatively cheap and easy to perform for virtually all proteins. Currently, over 2.8 million of nonredundant protein sequences are known and can be obtained from large repositories like National Center for Biotechnology Information (NCBI). At the same time, the tertiary structure is more difficult to obtain. The two mainstream methods that are widely used to determine the tertiary structure of proteins are the nuclear magnetic resonance (NMR) spectroscopy and the X-ray crystallography [12,13]. However, these methods have some limitations and are not suitable for all proteins. The main advantage of NMR is that the structure is determined in solvent, and therefore the obtained structure is native. The biggest advantage of the X-ray crystallography is that it is easier to perform and takes less time than NMR. As of October 2006, there are over 32,500 of protein structures in the Protein Data Bank (PDB) that are solved by X-ray crystallography, which is over five times more than the number of structures determined by NMR (less than 6000) [2].

The motivation for this work comes from the fact that not all proteins can be crystallized. Some of the X-ray crystallography experiments are unsuccessful, which results in

[*] Corresponding author. Fax: +1 780 492 1811.
E-mail address: lkurgan@ece.ualberta.ca (L. Kurgan).

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.02.040

the wastage of both resources and time. For instance, only 42% of the soluble purified proteins that were used for initial crystallization trials of the nonmembrane proteins of the archeaon *Methanobacterium thermoautotrophicum* yielded crystals [18]. If a computational method could predict, with acceptable accuracy, which sequence could be successfully crystallized, the experimental success rate for the crystallization would be improved. The increase of success rate would save resources that could be devoted to the successful crystallization of additional proteins.

To date, the factors that determine the successful crystallization of proteins are not clear. At the same time, others have observed that the propensity of a given protein to yield crystal under a given range of experimental conditions is an individual protein trait [16]. Therefore, we could expect that the ability to form crystals is related to the underlying protein sequence. To investigate this problem we need a dataset that includes sequences which are categorized into those that can and cannot be crystallized. Such dataset would allow finding patterns (regularities) using machine learning methods (classification and feature selection algorithms) that are associated with both crystallizable and noncrystallizable proteins. Smialkowski and the colleagues recently prepared such dataset and proposed a method, called SECRET, for prediction of protein crystallization [16]. The dataset contains 192 sequences that were determined only by NMR (noncrystallizable) and 226 sequences that can be crystallized and were determined by X-ray crystallography (and potentially also by NMR). The two classes of protein are referred to as NMR_ONLY and XRAY_NMR, respectively. Following the authors of [16], the hypothesis that sequences in NMR_ONLY cannot be crystallized is justified by the fact that X-ray crystallography is easier and costs less, and thus this technique is the first choice if the protein can be crystallized. Although this may not be true for all sequences, the NMR_ONLY class was considered acceptable to represent sequences that cannot be crystallized [16]. Despite that this dataset represents the best that can be extracted from PDB, it is characterized by two important limitations:

1. Both crystallizable and noncrystallizable sequences were made nonredundant at the 50% sequence homology using CD-HIT [11] to avoid bias towards a certain set of homologous sequences (family) of proteins. Although this allows for an unbiased prediction, we note that in some cases crystallization of a homologue of a crystallizable protein may be difficult. Therefore, both our method and method proposed in [16], which are based on this dataset, are limited to prediction of crystallization for nonhomologous proteins, and should not be used in prediction aimed at assessing crystallization of a homolog.
2. Since NMR-resolved proteins in PDB include only small and medium proteins, while larger proteins were exclusively determined by X-ray, the dataset includes sequences with 200 or less AAs to avoid the sequence length bias. As a result, both prediction methods are limited to small and medium size proteins. Given that sufficient number of large NMR-only proteins will be deposited to PDB, our future work includes redesigning the prediction method to encompass larger proteins.

The limitations of the SECRET method, which motivate our work, include use of a large number of features to represent protein sequences and low prediction accuracy that equals 67% (50% accuracy can be obtained by a coin flip). Additionally, this method performs prediction in a black-box manner, i.e., the prediction model could not be interpreted and as a results the authors did not describe what factors impact ability/inability of a protein to crystallize [16]. In contrast, our method uses significantly fewer features, predicts with over 77% accuracy, and allows formulating factors that could potentially be associated with protein crystallization. At the same time, both our and the SECRET methods are limited by one more factor, i.e., they consider only intra-molecular interactions, while the impact of the inter-molecular, i.e., protein–protein interaction and/or protein–precipitant interaction, were not included in the prediction model.

In our paper, we use the above dataset that was made available to us by the Frishman's lab [16] to build a classification model that predicts whether a sequence can be crystallized or not. The classification consists of two steps: (1) the protein sequence is converted into representation that consists of a fixed size feature vector, and (2) the feature values are entered into the classification model to predict the protein class (crystallizable/noncrystallizable). We performed the same design and test procedures as in [16]. Namely, we applied 10-fold cross validation to design and select feature representation ([16] used less stringent 5-fold cross-validation) and 10-fold cross-validation to test and compare the classifiers.

Materials and methods

Prediction methods. The SECRET method uses 103 features selected using wrapper based (with Naïve Bayes classifier) forward feature selection from the set of features representing frequencies of mono-, di-, and tripeptides. It applies a collection of Gaussian kernel based Support Vector Machines with that were combined using Naïve Bayes meta-classifier to perform prediction.

In contrast, the proposed method, which we called CRYSTALP, uses a different set of 46 features generated using a novel concept of collocated AA pairs (explained later in the paper), and a simple Naïve Bayes classifier [6]. The proposed method is conceptually simple, easy to implement, uses 50% less features than SECRET, and is shown to outperform the competing method with respect to the predictive accuracy. We also compare our solution, which is based on Naïve Bayes classifiers, to other designs that use the same features and different, state-of-the-art classifiers.

Feature generation. Composition vector is a simple sequence representation that is widely used in prediction of various structural aspects [3,7,9,15,19]. Given 20 alphabetically ordered ($A, C, \ldots, W, Y$) AAs, which are denoted as $AA_1, AA_2, \ldots, AA_{19}$, and AA_{20}, and the number of occurrences of AA_i in the sequence that is denoted as n_i, the composition vector is defined as

$$\left(\frac{n_1}{k}, \frac{n_2}{k}, \ldots\ldots, \frac{n_{19}}{k}, \frac{n_{20}}{k}\right)$$

where k is the length of the sequence.

A new representation, which is based on frequency of *collocation* of AA pairs in the sequence, was developed for the proposed prediction method. Our motivation was that the *composition vector* is insufficient to represent a sequence, since it only counts the frequencies of individual AAs. At the same time, frequencies of AA pairs (dipeptides) provide more information since they reflect interaction between local (with respect to the sequence) AA pairs. Based on this argument, we would count all dipeptides in the sequence. Since there are 400 possible AA pairs (AA, AC, $AD, \ldots, YY$), a feature vector of that size is used to represent occurrence of these pairs in the sequence. For instance, if AG pair occurs four times in a sequence, the corresponding value in the vector is set to 4, while if KN pair would not occur in the sequence, the corresponding values would be set to 0. Since short-range interactions between AAs, rather than only interactions between immediately adjacent AAs, have impact of folding [4], the proposed representation also considers collocated pairs of AAs, i.e. pairs that are separated by p other AAs. Collocated pairs for $p = 0, 1, \ldots, 4$ are considered, where for $p = 0$ the pairs reduce to the dipeptides. These pairs can be understood as the dipeptides with gaps. For each value of p, there are 400 corresponding feature values. Table 1 summarizes both composition vector and p-collocated AA pairs with respect to their corresponding number of features. As a result, we propose representation that includes total of $400(4+1)+20 = 2020$ features.

Feature selection. As the abovementioned proposed representation includes relatively large number of features, a feature selection method was used to reduce the dimensionality and potentially improve the prediction accuracy. A correlation-based feature subset selection method (CFSS) was used [5]. CFSS evaluates the value of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The strategy for searching subsets was the best-first-search method. Best-first-search explores the space of attribute subsets by using the greedy hill-climbing augmented with the backtracking. We also tested several other feature selection methods, and concluded that the CFSS method provides the best results. The feature selection was performed using 10-fold cross-validation to avoid overfitting, and features that were found significant by the feature selection method in at least 5 folds were selected. Among the original set of 2020 features, 46 features, which include 45 collocated AA pairs (see Table 2), and a composition vector value for AA Tyrosine (Y) were selected.

Results and discussion

Experimental setup

The classification systems used to develop and compare the proposed method were implemented in Weka, which is a comprehensive open-source library of machine learning methods [17]. The proposed CRYSTALP method was compared with several state-of-the-art classifiers such as Support Vector Machine (SVM) [8], Multiple Logistic Regression [10], instance learning based IBK algorithm [1] and C4.5 decision tree [14] using the same 46 features to represent sequences. It was also compared with the competing SECRET method, which utilizes a different sequence representation. The experimental evaluation was performed using 10-fold cross-validation to avoid overfitting and assure statistical validity of the results. The reported results include the following quality indices: accuracy, sensitivity, specificity, and Matthews's correlation coefficient (MCC).

Table 2
Features selected using CFSS/best-first-search feature selection method from the set of 2000 p-collocated AA pairs; the feature selection method also included one more feature, which is the composition of AA Tyrosine

$p = 0$	$p = 1$	$p = 2$	$p = 3$	$p = 4$
DL	HH	EC	AG	CS
EH	IC	FQ	CL	DN
LR	LE	IP	EL	FT
PD	QL	LE	EQ	GR
RI	TE	QS	HS	IG
RT	TT	SL	LD	MA
SS	YF	TG	MA	MY
WC		WV	NI	NH
YT		YN	NQ	TG
				TY
				VT

Results and comparison with competing methods

Each of the abovementioned classifiers was optimized (by adjusting internal parameters) with respect to accuracy. The optimization was performed using 10-fold cross-validation and the proposed set of 46 features. We first compare the CRYSTALP with the other classifiers (these methods use the same feature based sequence representation), and next with the competing SECRET method.

The best, optimized results for the CRYSTALP and the other four classifiers are shown in Table 3. The proposed CRYSTALP, which uses simple Naïve Bayes, provides the best accuracy that equals 77.51%. The CRYSTALP applies Naïve Bayes that uses kernel estimator for numeric attributes [6]. The second best SVM achieves 76.08% accuracy when using polynomial kernel (we also considered Gaussian kernel). The other three methods provide lower accuracies.

The proposed CRYSTALP method gives the best overall accuracy, highest sensitivity for crystallizable class, highest specificity for noncrystallizable class and the best MCC value. IBK with the optimal number of neighbors set at 13 gives the highest specificity for crystallizable proteins and highest sensitivity for noncrystallizable proteins. In short, the results demonstrate superiority of the proposed CRYSTALP method over the other commonly used classifiers.

Table 1
Size of feature sets for the proposed sequence representation

Feature representation	Composition vector	Collocated AA pairs				
		Adjacent pairs (dipeptides)	1-Collocated pairs	…	p-Collocated pairs	Total
Number of features	20	400	400	…	400	$400(p+1)+20$

Table 3
Prediction quality of the proposed CRYSLAP method and the other four classifiers

Method	Accuracy (%)	Crystallizable		Noncrystallizable		MCC	Confusion matrix			
		Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)		TP	FP	FN	TN
CRYSTALP	77.51	82.74	77.27	71.35	77.84	0.55	187	55	39	137
SVM	76.08	78.32	77.63	73.44	74.21	0.52	177	51	49	141
Logistic regression	71.53	73.01	73.99	69.79	68.72	0.43	165	58	61	134
IBK	72.49	68.58	77.89	77.08	67.58	0.46	155	44	71	148
C4.5	61.96	62.83	65.44	60.94	58.21	0.24	142	75	84	117

The results were based on the proposed feature set that includes 46 features.

The results of CRYSTALP were compared with the recently proposed competing SECRET method [16]. Since the authors of SECRET used the same data and experimental setup (10-fold cross-validation) we can directly compare the results, see Fig. 1.

CRYSTALP achieved 10.5% better accuracy than SECRET (77.51% vs. 66.99%) and significantly higher MCC (0.53 vs. 0.34) when compared with SECRET. The sensitivity and specificity for both protein classes obtained by CRYSTALP are also higher than the same values achieved by SECRET. Direct comparison of the confusion matrixes that were achieved by both methods shows that CRYSTALP predicts significantly more true positives (187 vs. 147) and similar (slightly larger) number of true negatives (137 vs. 133) when compared with SECRET. This illustrates that CRYSTALP provides better quality in predicting crystallizable proteins. Additionally, SECRET uses 103 features to represent the sequence and a complex classifier that applies multiple SVMs and a Naïve Bayes as the meta-classifier. In contrast, the proposed CRYSTALP uses about 50% less features and simple (and very fast to learn) Naïve Bayes classifier.

In the nutshell, the CRYSTALP method is shown to outperform, on virtually all aspects, the competing SECRET method.

Discussion of the proposed sequence representation

Although features selected for the CRYSTALP method may seem random (see Table 2), below we discuss several interesting patterns that concern the proposed sequence representation.

One of the patterns associated with the noncrystallizable proteins is the LE/EL pairs. LE is selected in 1-collocated and 2-collocated AA pairs, and EL is selected in 3-collocated pairs. The DL dipeptide and 3-collocated LD pair are also associated with the noncrystallizable proteins. We note that the higher the occurrence frequency of these collocated pairs in the sequence, the lower the probability that this sequence can be crystallized. We show detailed data for the 1-collocated LE pair and the DL dipeptide in Fig. 2A, i.e., the numbers inside the bars show how many proteins that contain a given number of collocated pairs are in fact noncrystallizable. One explanation for such consistency is that aspartate (D) and glutamate (E) have similar side chains. Moreover, substitution matrices (such as BLOSUM), have a relatively high value for the DE pair, which means that these AAs are characterized by a high exchange rate.

On the other hand, proteins that contain more of the MA, TG, and TY/YT pairs have higher probability to crystallize. Example data for the 4-collocated TG pair and 3-collocated MA pair are shown in Fig. 2B.

The occurrence of the abovementioned pairs may give insights to discover factors that either enable or prevent the protein crystallization. Although we could not find direct and consistent interactions between the side chains of the residues in these pairs, we note that, for instance, the 1-collocated LE and DL dipeptide are characterized by high probability to form a helix, see Fig. 3. This agrees

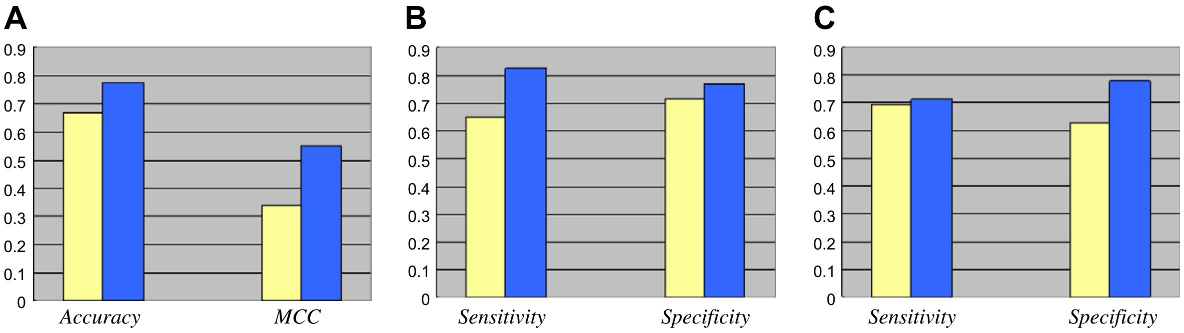

Fig. 1. Comparison between CRYSTALP and SECRET methods: (A) accuracy and MCC, (B) sensitivity and specificity for the class of crystallizable proteins, and (C) sensitivity and specificity for the class of noncrystallizable proteins. The light gray (yellow/left) bar corresponds to performance of SECRET, while the dark gray (blue/right) bar corresponds to the CRYSTALP method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

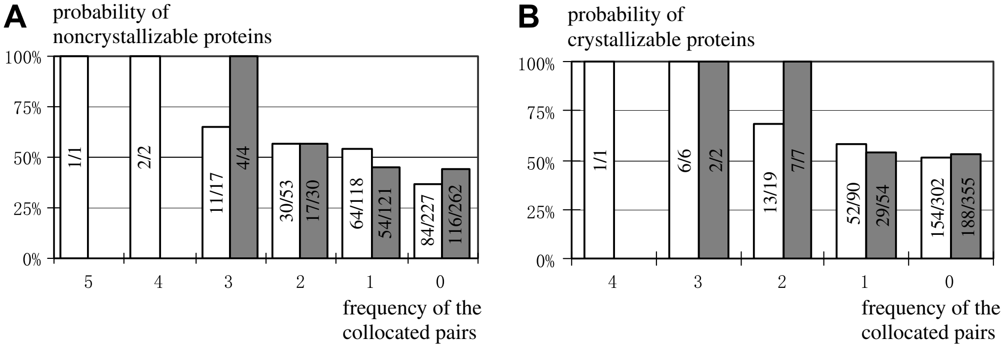

Fig. 2. (A) Probability of noncrystallizable proteins in function of the frequency of the 1-collocated LE pair (white bar) and the DL dipeptide (gray bar), (B) probability of crystallizable proteins in function of the frequency of the 4-collocated TG pair (white bar) 3-collocated MA pair (grey bar). The numbers inside the bars show how many proteins that contain a given number of collocated pairs are in fact noncrystallizable/crystallizable.

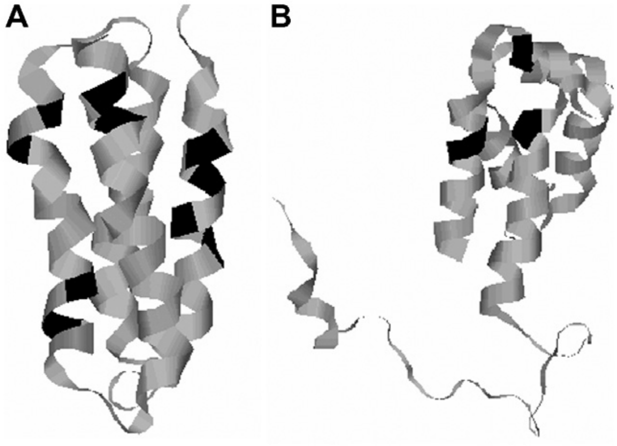

Fig. 3. Examples tertiary structure of proteins that contain significant number of the collocated pairs: (A) structure of chain A of 1p68, which contains five 1-collocated LE pairs; (B) structure of chain A of 1n3k, which contains three DL dipeptides; all pairs (shown in black) are inside helices.

with results of our previous study that shows that Leucine is strongly associated with formation of helices [4]. Statistical analysis in [4] shows that 1-collocated LE pair and the DL dipeptide have 0.71 and 0.50 probability of forming helix, respectively. Such intra-molecular interactions, which are related to formation of secondary structures, may have influence on the protein crystallization.

Summary and conclusions

A high quality predictor for protein crystallization would improve the success rate for X-ray crystallography, and as a result researchers could reroute resources to structure discovery for other proteins. Therefore, we propose CRYSTALP method that uses a novel protein sequence representation, which includes only 45 features, and applies a Naïve Bayes classifier. This method can be used to predict if small and medium size (<200 AAs), nonhomologous, proteins can be crystallized. CRYSTALP takes into account only intra-molecular factors, which are encoded in the protein's chain, while it may not provide reliable predictions when inter-molecular factors must be considered.

Based on 10-fold cross-validation tests, the proposed CRYSTALP is shown to predict the protein crystallization with 77.5% accuracy, which is 10% higher than the analogous results (for the same data and tests) achieved by the state-of-the-art, recently proposed competing SECRET method, and is also better when compared with four other machine learning methods. CRYSTALP also uses less then half of the number of features than the SECRET method and provides several interesting hypotheses with respect to intra-molecular factors that are associated with protein crystallization. We show that repeated occurrence of certain collocated AA pairs is correlated with crystallization or inability of a sequence to crystallize, which may provide a basis for discovery of crystallization markers. Our future work will include extending the method to predict crystallization for larger proteins and to incorporate inter-molecular interactions.

Acknowledgment

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under the Discovery program.

References

[1] D. Aha, D. Kibler, Instance-based learning algorithms, Machine Learning 6 (1991) 37–66.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Research 28 (2000) 235–242, based on data obtained in Oct 2006 at the <http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html>.

[3] C. Chen, Y.X. Tian, X.Y. Zou, P.X. Cai, J.Y. Mo, Using pseudo-amino acid composition and support vector machine to predict protein structural class, Journal of Theoretical Biology 243 (3) (2006) 444–448.

[4] K. Chen, L. Kurgan, J. Ruan, Optimization of the sliding window size for protein structure prediction, in: Proceedings of the 2006 International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2006), 2006, pp. 366–372.

[5] M. Hall, Correlation based feature selection for machine learning, Ph.D. dissertation, University of Waikato, Dept of Computer Science, 1999.

[6] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995, pp. 338–345.

[7] K.D. Kedarisetti, L. Kurgan, S. Dick, Classifier ensembles for protein structural class prediction with varying homology, Biochemical and Biophysical Research Communications 348 (2006) 981–988.

[8] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murphy, Improvements to Platt's SMO algorithm for SVM classifier design, Neural Computation 13 (2001) 637–649.

[9] L. Kurgan, L. Homaeian, Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, Pattern Recognition, special issue on Bioinformatics 39 (12) (2006) 2323–2343.

[10] C.S. Le, J.C. Houwelingen, Ridge estimators in logistic regression, Applied Statistics 41 (1992) 191–201.

[11] W. Li, L. Jaroszewski, A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases, Bioinformatics 18 (2002) 77–82.

[12] A.G. Palmer, D.J. Patel, Kurt Wuthrich and NMR of biological macromolecules, Structure 10 (12) (2002) 1603–1604.

[13] B. Perman, S. Anderson, M. Schmidt, K. Moffat, New techniques in fast time-resolved structure determination, Cellular and Molecular Biology 46 (5) (2000) 895–913.

[14] Q. Ross, C.45: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[15] H. Shen, K.-C. Chou, Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types, Biochemical and Biophysical Research Communications 334 (1) (2005) 288–292.

[16] P. Smialowski, T. Schmidt, J. Cox, A. Kirschner, D. Frishman, Will my protein crystallize? A sequence-based predictor, Proteins 62 (2) (2006) 343–355.

[17] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, San Francisco, 2005.

[18] A. Yee, K. Pardee, D. Christendat, A. Savchenko, A.M. Edwards, C.H. Arrowsmith, Structural proteomics: toward high-throughput structural biology as a tool in functional genomics, Accounts of Chemical Research 36 (2003) 183–189.

[19] Z. Yuan, Better prediction of protein contact number using a support vector regression analysis of amino acid sequence, BMC Bioinformatics 6 (2005) 248.