Protein Crystallization Prediction with a Combined Feature Set

Hui-Huang Hsu* and Shiang-Ming Wang
Department of Computer Science and Information Engineering
Tamkang University
Taipei, Taiwan
*E-Mail: *h_hsu@mail.tku.edu.tw*

Abstract

Using X-ray crystallography to determine the 3D structure of a protein is a costly and time-consuming process. One of the major reasons is that the protein needs to be purified and crystallized first, and the failure rate of protein crystallization is quite high. Thus it is desired to use a computational method to predict protein crystallizability based on the primary structure information before the whole process starts. This can dramatically lower the average cost for protein structure determination. In this paper, we investigated the feature sets used in previous research. The support vector machine (SVM) was chosen as the predictor. Different weightings are set for the penalty parameters of the two classes to deal with the imbalanced data problem. As a result, a combined set of features is able to produce better results, especially on the specificity.

1. Introduction

The importance of structural biology [1] research has been highlighted in recent years. The two mainstream methods widely used in protein structure determination are the nuclear magnetic resonance (NMR) spectroscopy [2] and the X-ray crystallography [3]. However, these two methods have respective limitations and are not suitable for all proteins. NMR is a physical phenomenon based upon the quantum mechanical magnetic properties of an atom's nucleus. In NMR, the protein also has to be produced in large quantities in highly concentrated solutions and requires weeks of data acquisition, expensive stable isotope labeling, and extensive manual analysis of data. On the other hand, X-ray crystallography is still the most powerful technique for determining the 3D structure of a protein. It can result in a 3D structure with a higher resolution than the NMR. However, X-ray crystallography has one important condition. That is the protein target needs to be crystallized first and the quality of the resulted crystal should be good enough to diffract the X-ray to sufficient resolution.

The process of experimental determination of protein structure has a high ratio of failures at different stages. This increases the average cost for protein structure determination. There are many individual steps in protein structure determination. Two of the key processes are protein production and protein crystallization. The motivation for this work was from the fact that some proteins can not be crystallized. If a protein cannot be crystallized, the X-ray crystallography cannot be utilized.

In this research, the SVM is used as the predictor since it is a powerful binary classifier which can always result in an optimized hyperplane to separate two classes. We use information from a protein's primary structure as the input to the SVM to predict the protein's crystallizability. The main issue here is to decide which features to be used for the classifier. We collect a few feature sets from previous researches and select feature sets from them. A simple feature selection method was used. From the experimental results, a combined feature set can be used to improve the prediction accuracy on protein crystallization. Specificity can be raised and is comparable to sensitivity. Our test result on the PDB dataset reached an overall prediction accuracy of 79.5% in a 5-fold cross-validation experiment where the sensitivity is 80.8% and the specificity is 78.3%.

The rest of this paper is organized as follows. Section 2 reviews related work on protein crystallization prediction. Section 3 presents the materials and methods for this research. Section 4 shows the prediction results of various tests. Finally, Section 5 draws a brief conclusion.

978-1-4244-3397-1/08/$25.00 ©2008 IEEE

2. Related work

In 2006, Smialowski et al. proposed a method – SECRET [4], to predict the protein crystallization. They got the protein sequence information from the Protein Data Bank (PDB) in 2004. They chose amino acid sequences with a length in the range from 30 to 200. They then used the SVM as the primary classifier and the Bayes network as the meta-classifier. The SECRET's prediction accuracy was 66.9% in a 10-fold cross-validation experiment. The result is not satisfied since this is a binary classification problem; a random guess can have 50% accuracy. Nevertheless, it is a good first attempt; valuable insights are provided.

Then in 2007, Chen et al. proposed a new approach for protein crystallization prediction – CRYSTALP [5]. CRYSTALP used only 45 amino acid pairs to represent the protein sequence and used the same dataset as used in SECRET. At last CRYSTALP was shown to predict crystallization with 77.5%, which is much better than the result by SECRET. However, the specificity of CRYSTALP was only 71.3%. The problem might be that the negative examples are much less than the positive examples. It seems that CRYSTALP did not deal with this imbalanced data problem well.

Overton and Barton proposed another criterion – OB-Score to rank potential protein targets by their predicted propensity to produce diffraction-quality crystals in 2006 [6]. The OB-Score summarizes predicted isoelectric point and hydrophobicity (pI and GRAVY) as the protein features. A percentage of 73.4 was reported for PfamA families that contain at least one member with a high OB-Score. The OB-Score thus can be used to pick proteins that are more likely to succeed in the process. However, definite classification criteria were not given.

Slabinski et al. developed a tool for protein crystallization prediction on the Web in 2007. They named the Web tool XtalPred [7] (http://ffas.burnham.org/XtalPred-cgi/xtal.pl).
XtalPred uses several online bioinformatics tools to compute the protein crystallization feasibility score. The prediction is made by combining individual crystallization probabilities into a single crystallization score. According to XtalPred's analysis, it rates the protein sequence crystallization's possibility into five categories: optimal, suboptimal, average, difficult, and very difficult. But the prediction accuracy by XtalPred is yet to be further verified since some of the used online tools might not give a very accurate prediction on the value of a certain feature.

3. Materials and methods

In this section, we will first explain how the protein sequence information is obtained and preprocessed. Then protein feature sets related to crystallization and wrapper-based feature selection are discussed. The classifier - SVM is then briefly introduced with a focus on the imbalanced data problem (IDP).

3.1. Data screening

Protein data were retrieved from the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do, 2007/04/06). There is a total of 39180 protein sequences. In those 39180 protein sequences, 33866 protein sequences used X-ray crystallography and 5314 protein sequences used NMR spectroscopy. It is assumed that proteins that used only NMR to generate its 3D structure cannot be crystallized. This assumption could be doubtful, but it is reasonable and was used in previous research [4]. Then we chose only the protein sequences with a length of 30 to 200 amino acids since if the protein sequence is too long or too short, the protein might not be crystallizable. This will avoid causing a bias on the length of the protein. This length restriction follows the data selection procedure in [4] and [5]. Then CD-HIT [8] was used to search for homology. Only one protein sequence was chosen from a group of similar ones. Redundancy can be avoided through this filtering process.

At last, 5445 protein sequences were retrieved from the PDB for our research. Among them, 3161 protein sequences are labeled "crystallizable" and 2284 protein sequences are viewed as "uncrystallizable." Although the screening process is similar to the one used in SECRET and CRYSTALP, we have resulted in a much larger dataset.

3.2. Protein features

Other than experimental settings, the most important variable for protein crystallization is still the protein itself. In our research, sequence information is first collected for protein crystallization prediction. The constitution of certain amino acids or amino acid subsequences is a key factor to protein crystallization. The feature sets we selected are the one-word size amino acids from [4] and two-word size subsequences from [5] as shown below.

One-word size: R,N,D,Q,E,H,L,F,S,T,W,V

Two-word size:

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

The two-word size subsequences are collocated amino acid pairs. p = 0, 1,. . . ,4 are considered. When p = 0, the pairs reduce to the dipeptides. When p = 1, there is a one-space gap in the amino acid pairs, the space can be any other amino acid.

Next, each amino acid sequence can be represented be a new alphabet according to the hydrophobicity class of the sequence. There are three different hydrophobicity scales: GES [9], Kyte and Doolittle [10] and Rose [11]. GES hydrophobicity scale was chosen as one of our protein feature sets based on the results from [4].

Dale and his colleagues think protein mutation can greatly influence the protein crystallization [12]. If a part of the protein sequence mutates, the protein might not be able to crystallize. Also if there are disorder regions or unstable regions in the protein sequence, the protein might not be able to crystallize. We can delete or truncate those unstable regions to help the protein crystallize. But this is not in the scope of this research.

There is a matrix defining distances between amino acids based on the amino acid mutation. It is called the mutation cost matrix [13]. It separates the amino acid into 12 groups: C, P, (L, M), (I, V), F, (W, Y), G, A, (T, S, H), (N, D), (K, R), (E, Q). Here we use the mutation cost matrix's information as one of the feature sets. To obtain the disorder region information, details of a protein can be downloaded from the PDB. The information at remark 465 can provide us positions of disorder regions in the sequence. Hence the total length of disorder regions in a protein can be computed.

The isoelectric point (pI) is the pH at which a particular molecule, like protein, has no net charge. It is also an important feature for protein crystallization [14]. The pI of each protein used in this research was calculated by the BioPerl pI calculator module with EMBOSS-defined pKa values.

3.3. Feature selection

There are two kinds of feature selection methods - filters and wrappers. The difference is on the evaluation of feature subsets after they are generated. The filter method uses information gain or mutual information to identify redundancy of certain feature subsets. On the other hand, the wrapper method uses a classifier, like the SVM, the neural network, or the Bayesian network, to check the classification accuracy with a certain feature subset [15]. Thus, the wrapper method is more accurate though it is very time-consuming. In addition, the wrapper method is more suitable for data sets with a large quantity and a smaller dimensionality.

In this work, the wrapper method with the SVM is used. There are several searching algorithms for the wrapper to decide the order of selecting feature subsets. A simple forward feature selection algorithm is used to seek for a better feature set combination. It is picked because of its simplicity. So all feature subsets are tested by the SVM first and the feature subsets are ordered according to the resulted prediction accuracies. Then the feature subsets are added to the combined feature set one by one following the order. The combined feature set with the highest prediction accuracy is the one we are looking for.

3.4. Support vector machine

The support vector machine (SVM) is a supervised learning method for classification [16]. The main concept of the SVM is to determine a hyperplane which separates binary class samples. The hyperplane maximizes the margin between the two classes of data samples. The hyperplane in SVM is defined by support vectors which are the data samples near the decision boundary of their corresponding class. (The margin is the distance of the two boundaries.)

In practical problems, the data samples might not be linearly separable by a hyperplane. In such cases, a nonlinear kernel is needed to transfer the data samples into a new feature space. In the new feature space, the SVM then can be trained as in a linear feature space. The most used nonlinear kernel in SVM is the radial basis function (RBF) kernel which is also used in this research.

3.5 Imbalanced data problem

The imbalanced data problem often exists in the real world data set. Imbalanced data classification refers to a two-class learning problem when the number of samples in one class is much smaller than that in the

other class. Sometimes the small data set's information is more important. While the majority of learning methods are designed for well-balanced training data, data imbalance presents a unique challenging problem to classifier design.

There is an algorithmic approach to improve the SVM for imbalanced training. We can use different weightings for the majority class and the minority class to overcome the imbalanced data problem [17]. Considering a two-class classification problem, we can give the two classes different penalty parameters for misclassification. The majority class is given a smaller penalty parameter and the minority class is given a larger penalty parameter. This can suppress the misclassification rate of the minority class comparing to the majority class. The proportion between the two penalty parameters is roughly the same as the rate between the numbers of samples of the two classes.

4. Experimental results

The RBF kernel based SVM is used to predict the protein crystallization. We have the one-word size amino acids, two-word size amino acid pairs, hydrophobicity/hydorphilicity, mutation cost index, disorder region length, and pI to represent the protein. Totally there are six feature subsets with 74 features.

The six protein feature subsets are ranked in Table 1 according to their accuracies. The accuracies were produced by 5-fold cross-validation. The feature subsets are then added one by one following their ranks in accuracy. The results are shown in Table 2. In this procedure, the best feature set combination is 1+2+3+4+5. If the sixth feature subset pI is added, prediction accuracy drops. This means pI might not be helpful in our cases. This result contradicts the conclusion from previous research in [14].

Table 1. Test results on feature sets for wrapper-based feature selection

No.	Features (dimensionality)	Accuracy (5-fold)
1	2-word size amino acid pairs (45)	75.3%
2	1-word size amino acids (12)	73.1%
3	Hydrophobicity/hydrophilicity (3)	72.8%
4	Mutation cost matrix (12)	70.9%
5	Disorder region length (1)	66.7%
6	pI (1)	61.9%

Table 2. Results from the wrapper procedure

Feature Sets	Accuracy (5-fold)
1+2	76.1%
1+2+3	76.4%
1+2+3+4	77.1%
1+2+3+4+5	**79.5%**
1+2+3+4+5+6	79.4%

From the results in Table 2, the feature set combination 1+2+3+4+5 is chosen as our combined feature set with a total number of features at 73. The prediction accuracy is 79.5%. The sensitivity of this is 80.8%; the specificity is 78.3%. The numbers are pretty good with the strict test of 5-fold cross-validation. For comparison, 10-fold cross-validation is also used to test the combined feature set. The result is compared with the results reported in SECRET [4] and CRYSTALP [5]. Please refer to Table 3. SEN means the sensitivity and SPE means the specificity. Our combined feature set with the SVM is able to achieve a higher accuracy. Also, the specificity is dramatically improved. This is very important since if specificity is too low, a large portion of proteins might be wrongly classified as uncrystallizable.

To do more test on the generated model, crystallization data from TargetDB (http://targetdb.pdb.org/, 2008/05) are extracted. TargetDB is a database created for collection of crystallization data. It asks the researchers to report both positive examples and negative examples. It is expected that the data samples are more balanced. The positive subset of the testing set was extracted from the protein sequences which are labeled "<status>Crystal Structure</status>" where "Crystal Structure" means that the protein structure was decided by X-ray crystallography. We include the protein sequences which are labeled "<status>NMR Assigned</status>" in our negative testing subset. The TargetDB test set includes totally 5574 protein sequences. 3831 protein sequences are positive examples and 1743 protein sequences are negative ones. So the PDB data set is used as the training set to construct a SVM model. The model is then tested on the TargetDB testing set. The result is shown in Table 4. The accuracy reaches 80.1% ((3211+1254)/5574) which is pretty good. The specificity is a bit low (71.9%, 1254/1743), but still acceptable. Please be aware that the two data sets were collected with different criteria.

Table 3. Comparison with SECRET and CRYSTALP

Method	SEN	SPE	No. of features	Accuracy (10-fold)
SECRET	65.0%	69.3%	103	66.9%
CRYSTALP	82.7%	71.3%	45	77.5%
This Work	80.8%	79.2%	73	80.2%

Table 4. Confusion matrix with tests on TargetDB data

Class \ Classified as	Positive	Negative
Positive	3211	620
Negative	489	1254

5. Conclusion

In this paper, we re-investigate a few feature sets used in previous researches for the protein crystallization prediction problem. We are able to find a combined feature set that has a reasonable dimensionality and can achieve higher overall accuracy with the SVM as the classifier. The specificity of our result is especially better than the ones reported by SECRET and CRYSTALP.

As abovementioned, accurate prediction of protein crystallization can certainly lower the cost and shorten the whole process for 3D structure determination. It would be interesting to do further research to find an optimal feature set for this problem. Only a very simple wrapper-based feature selection method was used in this research. More advanced feature selection can be applied to this problem.

References

[1] J.M. Chandonia and S.E. Brenner, "The Impact of Structural Genomics: Expectations and Outcomes," *Science*, 311: 347-351, 2006.

[2] J.M. Tyszka, S.E. Fraser and R.E. Jacobs, "Magnetic Resonance Microscopy: Recent Advances and Applications," *Current Opinion in Biotechnology*, 16(1): 93-99, 2006.

[3] W.L. Bragg, "The Structure of Some Crystals as Indicated by Their Diffraction of X-rays," *Proceedings of the Royal Society* (London), A89:248–277, 1914.

[4] P. Smialowski, T. Schmidt, J. Cox, A. Kirschner, and D. Frishman, "Will My Protein Crystallize? A Sequence-Based Predictor," *PROTEINS: Structure, Function, and Bioinformatics*, 62:343–355, 2006.

[5] K. Chen, L. Kurgan and M. Rahbari, "Prediction of Protein Crystallization Using Collocation of Amino Acid Pairs," *Biochemical and Biophysical Research Communications*, 355:764-769, 2007.

[6] I.M. Overton and G.J. Barton, "A Normalised Scale for Structural Genomics Target Ranking: The OB-Score," *FEBS Letters*, 580:4005–4009, 2006.

[7] L. Slabinski1, L. Jaroszewski, L. Rychlewski, I.A. Wilson, S.A. Lesley and A. Godzik, "XtalPred: a Web Server for Prediction of Protein Crystallizability," *Bioinformatics Applications Note*, 23:3403–3405, 2007.

[8] W. Li and A. Godzik, "CD-HIT: a Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics*, 22:1658-9, 2006.

[9] D.M. Engelman, T. A. Steitz and A. Goldman, "Identifying Nonpolar Transbilayer Helices in Amino Acid Sequences of Membrane Proteins," *Annu Rev Biophys Biophys Chem*, 15:321–353, 1986.

[10] J. Kyte and R. F. Doolittle, "A Simple Method for Displaying the Hydropathic Character of a Protein," *J Mol Biol*, 157:105–132, 1982.

[11] G. D. Rose, A. R. Geselowitz, G.J. Lesser, R.H. Lee and M.H. Zehfus, "Hydrophobicity of Amino Acid Residues in Globular Proteins," *Science*, 229:834–838, 1985.

[12] G.E. Dale, C. Oefner and A. D'Arcy, "The Protein as a Variable in Protein Crystallization," *Journal of Structural Biology*, 142:88–97, 2003.

[13] D. Gilis, S. Massar, N.J. Cerf and M. Rooman, "Optimality of the Genetic Code with Respect to Protein Stability and Amino-Acid Frequencies," *Genome Biol*, 2:RESEARCH0049, 2001.

[14] K.A. Kantardjieff and B. Rupp, "Protein Isoelectric Point as a Predictor for Increased Crystallization Screening Efficiency," *Bioinformatics Applications Note*, 20:2162–2168, 2004.

[15] R. Kohav and G.H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, 97:273-324, 1997.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM : a Library for Support Vector Machines," 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[17] X. Chen, B. Gerlach and D. Casasent, "Pruning Support Vectors for Imbalanced Data Classification," *Proceedings of 2005 International Joint Conference on Neural Networks*, Montreal, Canada, 2005.