CRYSpred: accurate sequence-based protein crystallization propensity prediction using sequence-derived structural characteristics

Marcin J Mizianty[1] and Lukasz A Kurgan[1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Canada

[*] Corresponding author
Phone: +1 (780) 492-5488
Fax: +1 (780) 492-1811
Email: lkurgan@ece.ualberta.ca

Abstract

Relatively low success rates of the X-ray crystallography, which is the most popular method for solving proteins structures, motivate development of novel methods that support selection of tractable protein targets. This aspect is particularly important in the context of the current structural genomics efforts that allow for a certain degree of flexibility in the target selection. We propose CRYSpred, a novel in-silico crystallization propensity predictor that uses a set of 15 novel features which utilize a broad range of inputs including charge, hydrophobicity, and amino acid composition derived from the protein chain, and the solvent accessibility and disorder predicted from the protein sequence. Our method outperforms seven modern crystallization propensity predictors on three, independent from training dataset, benchmark test datasets. The strong predictive performance offered by the CRYSpred is attributed to the careful design of the features, utilization of the comprehensive set of inputs, and the usage of the Support Vector Machine classifier. The inputs utilized by CRYSpred are well-aligned with the existing rules-of-thumb that are used in the structural genomics studies.

Key words: crystallization, crystallization propensity prediction, protein structure, structural genomics, target selection, X-ray crystallography.

Introduction

The knowledge of protein structure is essential to determine protein functions and interactions [1, 2], which in turn are used to gain insights into numerous biological processes, and to perform drug design [3]. The current methods that predict the tertiary structure from the protein sequence do not yet offer a sufficient degree of accuracy [4], and thus researchers rely on the availability of the structures that are solved experimentally. The most popular experimental approach is the X-ray crystallography, which according to the Protein Data Bank (PDB) [5] as of January 2011 was used to obtain 86.9% of the deposited protein structures. One of the biggest contributors who solve protein structures is the Structural Genomics (SG) initiative [6]. SG is an international effort that focuses on solving representative structures for unsolved protein families [7]. This approach is characterized by certain flexibility in choosing the representative proteins, concentrating on unsolved proteins that are of potential interest for the structural biology community [8]. Specifically, the Protein Structure Initiative (PSI), which includes a few major SG centers, focuses on selection of target proteins from large, structurally uncharacterized protein domain families and subfamilies in very large and diverse families with incomplete structural coverage [9]. Unfortunately, the X-ray crystallography is characterized by a significant rate of attrition and is among the most complex and least understood problems in structural biology [10]; only about 2-10% of the pursued protein targets yield diffraction-quality crystal structures [11]. A more recent study shows that the success rates at the SG centers are at about 4.6% [12]. Moreover, more than 60% of the costs of structure determination are consumed by the failed attempts [13]. In spite of the advances made in the context of protein crystallization [14], the production of high-quality crystals is one of the major bottlenecks in the X-ray crystallography-based structure determination pipelines [15-17]. Fortunately, the flexibility in the target selection allows the SG centers to concentrate resources on the tractable proteins. This motivates the development of computational methods which predict the propensity of a given protein chain to provide the diffraction-quality crystals. Such methods could save time and resources, since they would potentially reduce the amount of work spent on the failed attempts.

The design of the crystallization propensity predictors requires historical information about both successful and unsuccessful crystallization attempts [18]. This information can be found in TargetDB [19] and PepcDB (Protein Expression Purification and Crystallization DataBase) [20] databases. TargetDB, which concentrates on providing details concerning successful crystallization attempts, was launched July 2001, and it builds upon the work on the PRESAGE database [21]. The PepcDB, which was established around 2004, extends TargetDB and collects more detailed status information and experimental details for each step in the protein structure production pipeline. This database stores a complete history of the status of the experimental steps in each production trial, the current status, and stop conditions. The availability of the databases motivated several studies that investigated relations between different characteristics of proteins chains and the success of the crystallization [22]. In [18], the authors analyzed several protein characteristics, such as the presence of transmembrane helices, low-complexity regions, and coiled-coil regions. In [23], the team from the Joint Center for Structural Genomics discovered a few features which correlate with crystallization output, which include isoelectric point (pI), sequence length, average hydropathy, and the presence of low-complexity regions, signal peptides, and trans-membrane helices. More recent studies [24-26] added several additional factors, such as sequence conservation across organisms, the presence of charged residues, the number of protein binding partners, and the amount of intrinsic disorder. The above studies demonstrate that the propensity of a given protein to crystallize is predictable from the sequence and they motivated the development of several predictors.

In 2006, the first machine learning-based predictor, called SECRET, was proposed [27]. In the same year, Overton and Barton developed the OB-Score, which is a normalized scale for SG target ranking that is based on only two features: pI and hydrophobicity [28]. In the following years several other methods have been proposed including CRYSTALP [29], ParCrys [30], XtalPred [13, 31], CRYSTALP2 [32], MetaCrys

[33], P_{XS} [34], SVMCrys [35], and MCSG-Z score [36]. Some of these methods, including OB-Score and XtalPred, were already utilized by the SG centers. We note that while majority of these methods were designed using data generated by multiple SG centers (i.e., data coming from the TargetDB and PepcDB), the P_{XS} and MCSG-Z score were developed based on data coming from one SG center and thus they may not generalize to other centers or to applications by structural biologists. This is also why we do not include the latter two methods in our comparative analysis. Details concerning the above crystallization propensity predictors can be found in a recent review [12].

The current methods are based on a limited set of simple features (including SECRET, CRYSTALP, CRYSTALP2, OB-Score, and ParCrys) which include distribution of amino acid (AA) in the input chain, average hydrophobicity and isoelectric point, or use a relatively simple model (including XtalPred and OB-Score) to generate predictions. The only exception is the SVMCRYS predictor which uses a Support Vector Machines (SVM) [37] classifier and a more comprehensive feature set that includes predicted secondary structure. In this work, we explore a few new aspects which are potentially relevant in the context of the crystallization process to build an improved predictor of the crystallization propensity. First, we perform a comprehensive search of the AA properties from the AAindex database [38] to find which of these indices are relevant for the prediction of the crystallization propensity. Second, as shown in [39, 40], the characteristics of the residues which are located on the surface of the protein are more informative than the characteristics of all constituent residues. Therefore, we hypothesize that the use of the relevant AA properties from the AAindex database, such as charge, hydrophobicity, propensity to form certain secondary structures, for solvent exposed and for buried residues, which are annotated base on the predicted solvent accessibility [41], would improve predictions. Finally, our method effectively combines multiple inputs, which include the sequence, the AA indices, and the predicted disorder and solvent accessibility to provide predictions characterized by success rates that improve over the rates obtained by the existing crystallization propensity predictors.

Materials and Methods

Datasets

We employed three datasets which were recently introduced in [30], and one from [32]. These datasets were produced using the TargetDB and PepcDB using procedure developed in [30]. We designed our method based on five-fold cross validation on the FEAT dataset [30], which is composed of 1456 proteins sequences, with 728 crystallizable (C) and 728 non-crystallizable (NC) chains; this dataset was also used to design several recent crystallization propensity predictors, such as ParCrys, CRYSTALP2, and SVMCrys. Our method was tested on an independent TEST dataset (144 chains with 72 C and 72 NC) which is characterized by low sequence similarity with the chains in the FEAT dataset [30]. Similarly to work in [30], we also performed tests on the TEST-RL dataset (86 chains with 43 C and 43 NC), which is also is characterized by low sequence similarity with the chains in the FEAT dataset and which includes sequences with restricted length. We note that the chains that are included in these two test datasets, TEST and TEST-RL, share similarity to the sequences in the training datasets in the [30], which include the FEAT dataset, that is below the ’similar structure’ thresholds defined in [42]. The values of these thresholds are dependent on the length of the sequence and they are at about 25%. The similarity was measured using PSI-BLAST and a given chain was eliminated if the similarity exceeded the threshold in any of the PSI-BLAST iterations. Chains in the TEST-RL dataset have length between 46 and 200 residues, which enables comparison against SECRET and CRYSTALP as these two methods can only predict proteins with < 200 AAs. We also compare our predictor with the other methods on a more recent and larger test dataset with 2000 proteins (hereafter TEST-NEW with 1000 C and 1000 NC chains), which was introduced in [32] and which does not implement any restrictions on the sequence identity with respect to the training dataset. The TEST-NEW set is used to assess the quality of predictions for newer targets, i.e., it includes targets that were included in the TargetDB and PepcDB after the FEAT, TEST, and TEST-RL datasets were developed.

Quality Measures and Evaluation Protocols

We performed the evaluation at the protein level for binarized predictions (predicted class crystallizable vs. non-crystallizable) and predicted propensity of crystallization. For the binarized prediction we report the accuracy and Matthews's Correlation Coefficient (MCC), whereas for propensity predictions we provide ROC curves and the corresponding Areas Under the Curve (AUC). These measures are consistent with measures used in prior studies [30, 32, 33, 35]. The accuracy is the number of correct predictions divided by the total number of the test sequences. Given that the TP is the number of true positives (crystallizable protein predicted as being crystallizable), FP is the number of false positives (non-crystallizable predicted to be crystallizable), TN is the number of true negatives (non-crystallizable, predicted as non-crystallizable), and FN is the number of false negatives (crystallizable predicted as non-crystallizable), the accuracy is defined as:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100$$

The MCC values range between -1 and 1, where 0 represents random correlation, and bigger positive (negative) values indicate better (lower) prediction quality. This measure is defined as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(\mathrm{TP} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})}}$$

The Receiver Operating Characteristic ROC curve represents the relationship between the true positive (TP) and false positive (FP) rates when the confidence scores from the predictors are thresholded and the threshold values are varied. This allows comparing prediction qualities under different TP and FP rates. Besides visualizing the ROC curves, we also compute the AUC to quantify the predictive quality.

The design of our predictor was performed using five-fold cross validation on the FEAT dataset; this helps to obtain a robust method which is not over-trained (over-fitted) to the training dataset. In the five-fold cross validation we divide the dataset at random into five equal size subsets (folds), and use one fold as a test set and the others four folds as a training set; the experiment is repeated five times, each time a different fold is chosen to be the test fold. This methodology is consistent with the design protocols used in this area. In addition, for each assessment we repeated the cross validation (each time at random selecting different five folds) several times to compute mean MCC values; the mean is first computed over each set of five test folds and then averaged over the repetitions. The cross validations were repeated for as long as the coefficient of variation (the ratio of the standard deviation to the mean) was below 0.02, or for a maximum of five times. Once the design process is completed, we computed our prediction model using the FEAT dataset, and we tested our model and compared it against the existing predictors using the TEST, TEST-RL, and TEST-NEW dataset. This is consistent with the test protocols used in prior studies [30, 32, 33, 35].

Architecture of the Predictor

The input protein sequence is first converted into a numerical feature vector. These features are based on the information extracted from the protein chain and the solvent accessibility and disorder predicted from the sequence. The feature vector is inputted into an SVM model that outputs the predicted class (crystallizable vs. non-crystallizable) and the predicted crystallization propensity. The design includes formulation of the input features, selection of the relevant features, and parameterization of the SVM classifier.

Features

We explore a wide range of AA indices from the AAIndex1 database [38] and we combine them with the solvent accessibility predicted using SPINE [43]. We define a given residue as solvent exposed if its predicted relative solvent accessibility > 0.25; otherwise we assume that the residue is buried. We investigate total of 531 AA indices; we exclude the indices that have missing values for any of the 20 AAs. We also include features which are based on the disorder predicted with DISOPRED2 [44].

We used each of the 531 AA indices to compute the following three values (which results in 1593 features):

- {AAIndex} – sum of the index value for each residue divided by the sequence length (531 features)
- {AAIndex}_exp – sum of the index value for residues predicted as solvent exposed divided by the number of the exposed residues (531 features)
- {AAIndex}_bur – sum of the index value for residues predicted as buried divided by the number of the buried residues (531 features)

The 7 features generated using the predicted disorder include:

- DIS_RES – number of the predicted disordered residues divided by the sequence length (1 feature)
- DIS_MAX{_norm} – the length of the longest predicted disordered region (either normalized with respect to the sequence length or not) (2 features)
- DIS_AVG{_norm} – the average length of the predicted disordered regions (either normalized with respect to the sequence length or not) (2 features)
- DIS_REAL – sum of predicted disordered scores for each residue, divided by the sequence length (1 feature)
- DIS_SEG – number of predicted disordered regions (1 feature).

Since many of the above 1600 features are likely irrelevant or weakly-relevant to the prediction of the crystallization propensity, we used a simple wrapper-based filter to remove these features. We ranked each feature according to its MCC value, based on the five-fold cross fold validation on FEAT dataset with the Flexible Naïve Bayes classifier [45] using this feature as the only input. We selected this classifier since it allows for quick computation of the MCC value, while we had to compute 1600*5 = 8000 experiments. We use the MCC value for the isoelectric point, which equals 0.286 and which is one of the features that are known to affect the crystallization [23], as a cut-off threshold, i.e., the features with the MCC values < 0.286.were filtered out. Consequently, we selected 161 features that have the MCC values between 0.286 and 0.415.

Parameterization and Feature Selection

We used WEKA workbench [46] to design the proposed predictor. We use the SVM, which is among the top 10 data mining algorithms [47] and which was previously shown to provide accurate predictions for the crystallization propensity [35], outer membrane proteins [48], disordered residues [49], catalytic residues [50], and RNA-binding residues, to name just a few. The SVM requires parameterization, which includes selection of a suitable kernel function (including setting parameters of this function), and the selection of the value of the complexity constant *C*. We tried 3 popular kernel functions including Radial Basis Function (RBF), Polynomial kernel (POLY), and normalized polynomial kernel (NPOLY). For each kernel, we tuned its parameter including *gamma* for RBF and *exponent* for the latter two kernels. Consequently, we tuned two parameters for each kernel type, *C* and *gamma* for RBF, and *C* and *exponent* for both polynomial kernels. The parameterization was performed using grid search based on the five-fold cross validation on the FEAT dataset to maximize the MCC values, where the considered values of the SVM parameter were: $C \in \{0.5, 1, 2.5, 5, 7.5, 10, 15 ,20, 25, 30\}$, $gamma = g*10^{e}$, where $g \in \{0.01, 0.025,$

0.05, 0.075} and $e \in \{0, 1, 2\}$, and *exponent* = 0.5+ 0.25**i*, where $i \in \{0, 1, .., 10\}$ (except *exponent* = 1 for NPOLY kernel since this value is prohibited). We first parameterized each of the 3 kernels using the 161 features; these parameters were used to perform feature selection. Upon completion of the features selection, we parameterized the 3 kernels again using the corresponding selected feature sets.

Table **1** Summary of results for the wrapper-based feature selection, which are based on the five-fold cross validation on the FEAT datasets, for the three considered classifiers, i.e., SVM models that use three kernel types including Radial Basis Function (RBF), Polynomial kernel (POLY), and normalized polynomial kernel (NPOLY), and two search methods, the best first and the greedy stepwise. The results are sorted in the descending order according to the MCC values.

Feature selection method		# of selected features	Results on the FEAT dataset	
Classifier	Search method		Accuracy	MCC
SVM_NPOLY	Best First	15	78.5	0.572
SVM_NPOLY	Greedy stepwise	15	77.8	0.558
SVM_RBF	Best First	17	77.6	0.554
SVM_POLY	Best First	21	77.3	0.548
SVM_RBF	Greedy stepwise	15	77.2	0.546
SVM_POLY	Greedy stepwise	15	76.9	0.538

Table **2** Comparison of predictive quality between the proposed CRYSpred method and the existing crystallization propensity predictors including ParCrys, OB-Score, XtalPred, CRYSTALP, CRYSTALP2, SECRET, and SVMCrys. The evaluation is computed based on three independent test datasets, TEST, TEST-RL and TEST-NEW. The CRYSpred was trained on the FEAT dataset, whereas the results for the ParCrys, OB-Score, XtalPred, CRYSTALP, CRYSTALP2 and SECRET were obtained from the web servers, and for the SVM-CRYS using the author-provided standalone application. The SVM-CRYS and CRYSTALP do not produce crystallization propensity scores (they only generate binary predictions) and thus we could not compute their AUC values. The SECRET and CRYSTALP could be tested only on the TEST-RL dataset since they predict only for sequence < 200 residues. The results on each dataset are sorted in the descending order according to the MCC values and the best results for each quality index and each dataset are shown in bold font.

Dataset	Method	Accuracy	MCC	AUC
TEST	**CRYSpred**	**79.9**	**0.60**	**0.85**
	XtalPred	79.2	0.58	0.83
	CRYSTALP2	75.7	0.52	0.79
	ParCrys	71.5	0.45	0.75
	OB-Score	64.6	0.32	0.68
TEST-RL	**CRYSpred**	**80.2**	**0.60**	**0.86**
	ParCrys	79.1	0.58	0.84
	XtalPred	76.7	0.54	0.82
	CRYSTALP2	69.8	0.40	0.72
	OB-Score	69.8	0.40	0.71
	SECRET	58.1	0.16	0.58
	CRYSTALP	46.5	-0.07	N/A
TEST-NEW	**CRYSpred**	**73.4**	**0.47**	**0.78**
	ParCrys	70.6	0.43	0.75
	SVMCrys	70.4	0.43	N/A
	XtalPred	70.0	0.40	0.76
	CRYSTALP2	69.3	0.39	0.74

The feature selection was performed based on the wrapper approach, with the 3 SVMs that are based on the 3 kernels parameterized using the 161 features, in which we utilized two search techniques: the best first search and the greedy stepwise search. The objective of both search types was to select a set of features that maximizes the MCC value, which is computed based on five-fold cross validation on the FEAT dataset. We generated total of 6 features sets (2 search procedures for each of the 3 kernel types). The results from the five-fold cross validation on the FEAT dataset for the six selected feature sets are summarized in Table 1. We observe that the six considered configurations provide relatively similar predictive performance. The MCC values range between 0.54 and 0.57 and the number of selected features is between 15 and 21. The best performing model utilizes SVM with the NPOLY kernel along and 15 features selected using the best first search-based feature selection. This configuration, for which $C = 25$ and *exponent* = 2.75, was selected to implement the proposed CRYSpred method.

Results and Discussion

Comparative Study

Our CRYSpred was trained on the FEAT dataset and tested on the three independent (from the FEAT dataset) tests datasets: TEST, TEST-RL, and TEST-NEW. We compare our model with the OB-Score, ParCry, CRYSTALP2, and XtalPred methods on the TEST, TEST-RL and TEST-NEW dataset. We also compare with SECRET and CRYSTALP on the TEST-RL dataset (these two latter methods predict only for sequence with less than 200 residues and thus they could not be tested on the other two test sets) and with the SVMCrys on the largest TEST-NEW dataset. We do not include the results from the SVMCrys on the TEST and TEST-RL datasets since we believe that these predictions were overfitted into these datasets, i.e., the authors report the accuracy of 89.5% on the TEST-RL dataset and 86.8% on the TEST dataset which are substantially higher that the 77.4% accuracy that this method obtained on the training FEAT dataset [35]. We note that the XtalPred generates one of five crystallization propensity classes, including optimal, suboptimal, average, difficult, and very difficult, whereas our test datasets classify each chain into one of the two classes, crystallizable and non-crystallizable. Therefore, we mapped the output of XtalPred into the two classes as follows: the optimal, suboptimal, and average classes are assumed to be predicted as crystallizable, while the difficult and very difficult classes as the non-crystallizable. This mapping was previously shown to result in a favorable prediction quality when considering the two classes [32]. The results are presented in Table 2. The corresponding ROC curves obtained for the top four predictors, ParCrys, XtalPred, CRYSTALP2, and CRYSpred, are shown in Figure 1. We note that the SVM-CRYS and CRYSTALP do not generate the crystallization propensity values, and thus we could not generate the ROC curves and compute AUC for these two methods. We also did not include the ROC curves for the OB-Score and SECRET since these methods obtain lower AUC values, see Table 1.

The CRYSpred outperforms the other solutions in both the binary predictions (based on the MCC and ACC scores) and the real-valued crystallization propensities (based on the AUC values) on the three test datasets. The MCC and AUC values offered by our method are around 0.6 and 0.85, respectively, for the older two datasets (TEST and TEST-RL), and they equal 0.47 and 0.78 for the newer and larger TEST-NEW dataset. The AUC scores and the corresponding ROC curves show that CRYSpred works on average better than the ParCrys and XtalPred methods, and that these improvements hold for the majority of the range of the TP- and FP-rates, see Figure 1. We note that our method improves over the second-best ParCrys on the newer TEST-NEW dataset by 2.8% in accuracy, 0.04 in MCC, and 0.03 in AUC. The accuracies on the TEST-NEW dataset are lower by around 6% for the CRYSpred and 9% for second-best ParCrys, when compared with the results on the TEST and TEST-RL datasets. This drop in the quality of the predictions could be explained by the fact that the TEST-NEW set includes newer data. Consequently, recent advances in the crystallization protocols [52, 53], which potentially enable crystallization of previously non-crystallizable proteins, would confuse the results generated by the prediction models that were established using older data. This motivates development of new crystallization propensity predictors which would utilize newer data in order to keep up with the dynamic nature of the modern crystallization pipelines.

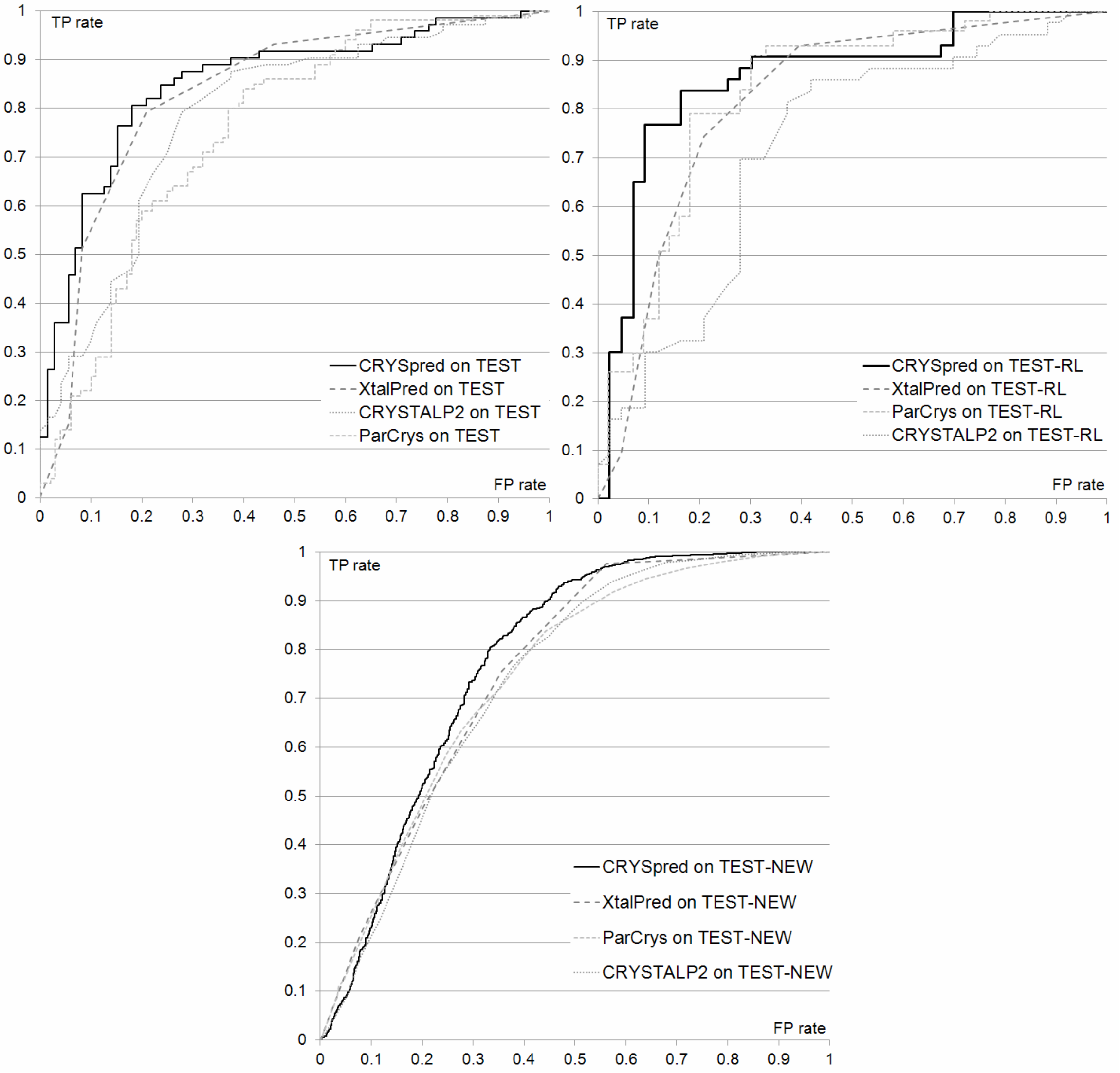

Figure **1** The ROC curves for the ParCrys, XtalPred, CRYSTALP2, and CRYSpred computed on the TEST (top left panel), TEST-RL (top right panel), and TEST-NEW (bottom panel) datasets.

Factors Related to Prediction of Crystallization Propensity

CRYSpred uses 15 features which are summarized in Table 3 and which quantify several structural characteristics of proteins. The table includes the biserial correlation coefficients between the values of our features and the outcomes (annotations of crystallizable and non-crystallizable chains) on the FEAT dataset. The sign and the magnitude of these correlations indicate the direction and the strength of the relation between a given features and the outcomes. We observe a strong presence of information derived from the charge-based AA indices, which agrees with previous observation made in [24], and from the hydrophobicity-based AA indices, which concurs with the observations in several related studies [24, 28-30, 32, 36]. The feature set uses two AA indices that describe AA composition, which was also used in several prior methods that predict crystallization propensity [13, 30, 32, 34]. We hypothesize that the reason why we use three AA indices that quantify propensity for the alpha helix conformation is that membrane spanning regions in protein structure are often implemented with alpha helices, and this information was previously found useful for the crystallization prediction [23, 18, 26]. Our method also utilizes features derived from the predicted disorder, which agrees with the findings in [13, 25], and information concerning the predicted solvent accessibility, which was shown to be important in [39, 40]. Overall, the features utilized by CRYSpred are intuitive, physically reasonable, and their selection is supported by the fact that they were investigated in relevant prior studies. Our main contribution is to efficiently combine these relevant factors to generate high quality predictions of the crystallization propensity.

Table **3** Summary of the selected 15 features used in the CRYSpred method along with their MCC values obtained by the Flexible Naïve Bayes classifier and biserial correlation on the FEAT dataset (see the "Features" section for more details). The features are sorted in the descending order according to the MCC values. The "Feature name" uses identifiers of the corresponding AA indices from the AAindex database.

MCC	Biserial correlation	Feature name	Description	Feature type
0.397	0.349	NAKH900113	Average value of AA index describing AA composition	AA composition
0.368	0.386	KUMS000103	Average value of AA index describing distribution of AAs in the alpha-helices in thermophilic proteins	Secondary structure
0.367	0.390	KUMS000104	Average value of AA index describing distribution of AAs in the alpha-helices in mesophilic proteins	Secondary structure
0.360	-0.417	GRAR740101	Average value of AA index describing AA composition	AA composition
0.347	-0.336	DIS_MAX_norm	The length of the longest predicted disordered region divided by the sequence length	Disorder
0.343	0.361	QIAN880103_exp	Average value of AA index describing weights for alpha-helices	Secondary structure
0.325	-0.232	PARJ860101	Average value of AA index describing HPLC parameter	Hydrophobicity
0.315	0.129	WERD780101	Average value of AA index describing propensity of AAs to be buried	Solvent accessibility
0.312	-0.272	DIS_REAL	Sum of predicted disorder scores for each residue divided by the sequence length	Disorder
0.309	0.183	BIOV880101	Average value of AA index describing solvent accessibility of AAs	Solvent accessibility
0.307	0.116	BAEK050101	Average value of linker index	Disorder
0.307	0.289	COWR900101	Average value of hydrophobicity index	Hydrophobicity
0.299	-0.322	CHAM830108	Average value of AA index describing a parameter of charge transfer donor capability	Charge
0.299	0.287	FAUJ880112_bur	Average value of AA index describing negative charge for buried AAs	Charge
0.292	0.205	FAUJ880112	Average value of AA index describing negative charge	Charge

Case Studies

We present three case studies to demonstrate how the features selected in this study can contribute to an improved prediction of the crystallization propensity. We choose three crystallizable proteins from the TEST_NEW dataset (PDB ids: 3DBO, 3I59 and 3IHU) that were correctly predicted by CRYSpred but predicted as non-crystallizable by the other methods, except for the 3I59 and 3IHU proteins that were correctly predicted by the SVMCRYS and XtalPred, respectively. Table 4 summarizes the input feature values and the predicted crystallization propensities by the CRYSpred, SVMCRYS, CRYSTALP2, ParCrys, OB-Score, and XtalPred methods for these targets.

Table **4** Values of input features and predicted crystallization propensities for three selected protein targets (PDB ids: 3DBO, 3I59 and 3IHU). Features values were normalized using min-max method and the maximal/minimal values of a given feature on the FEAT dataset. Values of the average hydropathy score and pI were taken from TargetDB and the value of the instability index was taken from the XtalPred predictions. The crystallization propensities are used to predict whether a given chain is crystallizable using a cut-off value that for the CRYSpred and CRYSTALP 2 is 0.5, for ParCrys is 3564600, for OB-Score is 0.809, and for XtalPred is 3. The XtalPred generates five propensities including optimal, suboptimal, and average which are assumed as the crystallizable, and the difficult and very difficult which are considered as the non-crystallizable The SVMCRYS does not produce the propensity scores.

		3DBO	3I59	3IHU
15 features used by CRYSpred	NAKH900113	0.263	0.211	0.219
	KUMS000103	0.574	0.679	0.472
	KUMS000104	0.576	0.695	0.482
	GRAR740101	0.151	0.250	0.176
	DIS_MAX_norm	0.030	0.042	0.136
	QIAN880103_exp	0.412	0.667	0.470
	PARJ860101	0.277	0.482	0.181
	WERD780101	0.582	0.422	0.688
	DIS_REAL	0.017	0.019	0.077
	BIOV880101	0.688	0.516	0.788
	BAEK050101	0.532	0.407	0.637
	COWR900101	0.688	0.588	0.736
	CHAM830108	0.245	0.192	0.405
	FAUJ880112_bur	0.241	0.188	0.040
	FAUJ880112	0.226	0.558	0.166
Features relevant to crystallization success that are used by other methods	Avg. Hydropathy Score	-0.431	-0.271	-0.209
	pI Value	11.9	10.2	9.1
	Instability Index	77.03	43.45	38.41
Crystallization propensity predictions for a given predictor in the format: binary prediction (propensity value) where C stands for crystallizable and N for non-crystallizable	CRYSpred	C (0.580)	C (0.747)	C (0.705)
	SVMCRYS	N (n/a)	C (n/a)	N (n/a)
	CRYSTALP2	N (0.329)	N (0.380)	N (0.407)
	ParCrys	N (192000)	N (2080000)	N (3360000)
	OB-Score	N (-2.79)	N (-4.11)	N (-2.22)
	XtalPred	N (very difficult)	N (difficult)	C (average)

The considered targets have high pI and low average hydrophaty values. These values produce a low OB-Score which suggests difficulties with the crystallization [28]. The pI and hydrophaty values are also most likely responsible for the incorrect predictions from the ParCrys and CRYSTALP2 methods, and have, along with the high values of the instability index, strong impact on the predictions generated by the XtalPred. The correct XtalPred prediction for 3IHU has low confidence (the average propensity value). The factor that resulted in reducing the propensity to average, when compared with the very difficult and difficult propensities generated for the 3DBO and 3I59 chains, is the fact that the 3IHU protein has a relatively large number of homologs, equal 8, to the targets with solved structures. Considering the features used by CRYSpred, the three targets are characterized by high values of indices describing secondary structures (including KUMS000103, KUMS000104, and QIAN880103_exp features) and solvent accessibility (including WERD780101 and BIOV880101) and low values of the disorder content related features (including DIS_MAX_norm and DIS_REAL features). We observe that the KUMS000103, KUMS000104, QIAN880103, WERD780101 and BIOV880101 features have positive correlations with the outcomes (the annotation of the crystallizable and non-crystallizable chains), see Table 3, which means that their high positive values indicate higher likelihood of successful crystallization. On the other hand, the DIS_MAX_norm and DIS_REAL features have negative correlations with the outcomes, see Table 3, which is why their lower values are indicative of the successful crystallization. Moreover, in spite of the low average hydropathy score, which is unfavorable for the successful crystallization, the values for the hydrophobicity-based indices which are used in the CRYSpred, i.e., the relatively low values for the PARJ860101 feature that has negative correlation with the outcomes and the high values of the COWR900101 feature that is positively correlated with the outcomes, suggest that these targets are crystallizable. Finally, the relatively low values for the features describing AA content (GRAR740101) and charge (CHAM830108), which have negative correlations with the outcomes, see Table 3, suggest that the input chains are feasible crystallization targets. The values of the remaining features overall do not agree with the observations on training dataset; however the SVM classifier that we apply compensates for these discrepancies and produces correct predictions for all three targets.

Conclusions

We propose a novel in-silico method for the crystallization propensity prediction, CRYSpred. Our method outperforms the existing predictors on three test datasets. CRYSpred improves accuracy over the second best ParCrys method by about 3% for the largest test set that includes new crystallization targets. Our method uses the SVM classifier and a set of fifteen novel features which quantify information about AA composition, hydrophobicity, charge, propensity of AAs to form helical conformation, and the predicted solvent accessibility and disorder. The features used by the CRYSpred are well-grounded in the prior studies that investigated factors related to the propensity for protein crystallization. We observe that the predictive qualities obtained for the newer test data are lower, which motivates continuing development of new predictors which would accommodate for the new advances in the crystallization protocols. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [54], we shall make an effort in our future work to provide a web-server for the method presented in this paper.

Acknowledgements

This work was supported in part by the Discovery grant from NSERC Canada to LAK and the Izaak Walton Killam Memorial scholarship to MJM.

References

[1] Chou, K. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **2004**, *11*(21):5-34.

[2] Chou, K.; Wei, D.; Du, Q.; Sirois, S.; Zhong, W. Progress in computational approach to drug development against SARS. *Curr. Med. Chem.*, **2006**, *13*(32), 63-70.

[3] Norin, M.; Sundström, M. Protein models in drug discovery. *Curr Opin Drug Discov Devel*, **2001**, *4*, 284-290.

[4] Moult, J.; Fidelis, K.; Kryshtafovych, A.; Rost, B.; Hubbard, T.; Tramontano, A. Critical assessment of methods of protein structure prediction—Round VII. *Proteins*, **2007**, *69*(S8), 3–9.

[5] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.*, **2000**, *28*, 235-242.

[6] Chandonia, J.M.; Brenner, S.E. The impact of structural genomics: expectations and outcomes. *Science*, **2006**, *311*, 347-351.

[7] Terwilliger, T.C.; Waldo, G.; Peat, T.S.; Newman, J.M.; Chu, K.; Berendzen, J. Class-directed structure determination: Foundation for a protein structure initiative. *Protein Sci.*, **1998**, *7*(9), 1851-1856.

[8] Brenner, S.E. Target selection for structural genomics. *Nat. Struct. Biol.*, **2000**, *7*, 967–969.

[9] Dessailly, B.H.; Nair, R.; Jaroszewski, L.; Fajardo, J.E.; Kouranov, A.; Lee, D.; Fiser, A.; Godzik, A.; Rost, B.; Orengo, C. PSI-2: structural genomics to cover protein domain family space. *Structure*, **2009**, *17*(6), 869-881.

[10] Hui, R.; Edwards, A. High-throughput protein crystallization. *J. Struct. Biol.*, **2003**, *142*, 154-161.

[11] Service, R. Structural genomics, round 2. *Science*, **2005**, *307*, 1554–1558.

[12] Kurgan, L.A.; Mizianty, M.J. Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis. *Natural Science*, **2009**, *1*, 93-106.

[13] Slabinski, L.; Jaroszewski, L.; Rodrigues, A.P.C.; Rychlewski, L.; Wilson, I.A.; Lesley, S.A.; Godzik, A. The challenge of protein structure determination - lessons from structural genomics. *Protein Sci.*, **2007**, *16*(11), 2472-2482.

[14] McPherson, A. Protein crystallization in the structural genomics era. *J. Struct. Funct. Genomics*, **2004**, *5*(1-2), 3-12.

[15] Chayen, N.E. Turning protein crystallisation from an art into a science. *Curr. Opin. Struct. Biol.*, **2004**, *14*(5), 577-583.

[16] Biertumpfel, C.; Basquin, J.; Suck, D. Practical implementations for improving the throughput in a manual crystallization setup. *J Appl Crystallogr*, **2005**, *38*, 568-570.

[17] Pusey, M.L.; Liu, Z.J.; Tempel, W.; Praissman, J.; Lin, D.; Wang, B.C.; Gavira, J.A.; Ng, J.D. Life in the fast lane for protein crystallization and X-ray crystallography. *Prog. Biophys. Mol. Biol.*, **2005**, *88*, 359-386.

[18] Rodrigues, A.; Hubbard, R.E. Making decisions for structural genomics. *Brief. Bioinformatics*, **2003**, *4*, 150-167.

[19] Chen, L.; Oughtred, R.; Berman, H.M.; Westbrook, J. TargetDB: A target registration database for structural genomics projects. *Bioinformatics*, **2004**, *20*, 2860–2862.

[20] Kouranov, A.; Xie, L.; de la Cruz, J.; Chen, L.; Westbrook, J.; Bourne, P.E.; Berman, H.M. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **2006**, *4*, D302-305.

[21] Brenner, S.E.; Barken, D.; Levitt, M. The PRESAGE database for structural genomics. *Nucleic Acids Res.,* **1999**, *27*(1), 251-253.

[22] Rupp, B.; Wang, J.W. Predictive models for protein crystallization. *Methods*, **2004**, *34*, 391-408.

[23] Canaves, J.M.; Page, R.; Wilson, I.A.; Stevens, R.C. Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: Maximum clustering strategy for structural genomics. *J. Mol. Biol.*, **2004**, *344*, 977–991.

[24] Goh, C.S.; Lan, N.; Douglas, S.M.; Wu, B.; Echols, N.; Smith, A.; Milburn, D.; Montelione, GT.; Zhao, H.; Gerstein, M. Mining the structural genomics pipeline: Identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.*, **2004**, *336*, 115–130.

[25] Oldfield, C.J.; Ulrich, E.L.; Cheng, Y.; Dunker, A.K.; Markley, J.L. Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins*, **2005**, *59*, 444–453.

[26] Chandonia, J.M.; Kim, S.H.; Brenner, S.E. Target selection and deselection at the Berkeley Structural Genomics Center. *Proteins*, **2006**, *62*, 356–370.

[27] Smialowski, P.; Schmidt, T.; Cox, J.; Kirschner, A.; Frishman, D. Will my protein crystallize? A sequence-based predictor. *Proteins*, **2006**, *62*, 343-355.

[28] Overton, I.M.; Barton, G.J. A normalised scale for structural genomics target ranking: the OB-Score. *FEBS Lett.*, **2006**, *580*, 4005-4009.

[29] Chen, K.; Kurgan, L.A.; Rahbari, M. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem. Biophys. Res. Commun.*, **2007**, *355*, 764-769.

[30] Overton, I.M.; Padovani, G.; Girolami, M.A.; Barton, G.J. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics*, **2008**, *24*, 901-907.

[31] Slabinski, L.; Jaroszewski, L.; Rychlewski, L.; Wilson, I.A.; Lesley, S.A.; Godzik, A. XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics*, **2007**, *23*(24), 3403-3405.

[32] Kurgan, L.A.; Razib, A.A.; Aghakhani, S.; Dick, S.; Mizianty, M.J.; Jahandideh, S. CRYSTALP2: Sequence-based Protein Crystallization Propensity Prediction. *BMC Struct. Biol.*, **2009**, *9*, 50.

[33] Mizianty, M.J.; Kurgan, L.A. Meta prediction of protein crystallization propensity. *Biochem. Biophys. Res. Commun.*, **2009**, *390*, 10-15.

[34] Price, W.N. 2nd; Chen, Y.; Handelman, S.K.; Neely, H.; Manor, P.; Karlin, R.; Nair, R.; Liu, J.; Baran, M.; Everett, J.; Tong, S.N.; Forouhar, F.; Swaminathan, S.S.; Acton, T.; Xiao, R.; Luft, J.R.; Lauricella, A.; DeTitta, G.T.; Rost, B.; Montelione, G.T.; Hunt, J.F. Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nat. Biotechnol.*, **2009**, *27*, 51-57.

[35] Kandaswamy, K.; Pugalenthi, G.; Suganthan, P.N.; Gangal, R. SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence. *Protein Pept. Lett.*, **2010**, *17*, 423-430.

[36] Babnigg, G.; Joachimiak, A. Predicting protein crystallization propensity from protein sequence. *J. Struct. Funct. Genomics*, **2010**, *11*, 71-80.

[37] Vapnik, V., *The nature of statistical learning theory*. Springer, New York **1995**.

[38] Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **2008**, *36*, D202-D205.

[39] Derewenda, Z.S. Rational protein crystallization by mutational surface engineering. *Structure*, **2004**, *12*, 529-535.

[40] Goldschmidt, L.; Cooper, D.R.; Derewenda, Z.; Eisenberg, D. Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. *Protein Sci.*, **2007**, *16*, 1569-1576.

[41] Ahmad, S.; Gromiha, M.M.; Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **2003**, *50*, 629-635.

[42] Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **1999,** *12*(2), 85-94.

[43] Dor, O.; Zhou, Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins*, **2007**, *66*, 838-845.

[44] Ward, J.J.; McGuffin, L.J.; Bryson, K.; Buxton, B.F.; Jones, D.T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics,* **2004**, *20*, 2138-2139.

[45] John, G.; Langley, P. In: *Estimating continuous distributions in bayesian classifiers*, Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Montreal, Quebec, Canada, August 18-20, 1995; Morgan Kaufmann, pp. 338–345.

[46] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update; *SIGKDD Explor*, **2009**, *11*(1), 10-18.

[47] Xindong, W.; Kumar, V.; Quinlan, J.R.; Ghosh,, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; Zhou, Z.H.; Steinbach, M.; Hand, D.J.; Steinberg, D.; Wu, X. Top 10 algorithms in data mining. *Knowl Inf Syst,* **2008**, *14*(1), 1-37.

[48] Mizianty, M.; Kurgan, L. Improved identification of outer membrane beta barrel proteins using primary sequence, predicted secondary structure and evolutionary information. *Proteins* **2011**, *79*(1), 294-303.

[49] Mizianty, M.; Stach, W.; Chen, K.; Kedarisetti, K.D.; Miri Disfani, F.; Kurgan, L. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics* **2010**, *26*(18), i489-i496.

[50] Zhang, T.; Zhang, H.; Chen, K.; Shen, S.; Ruan, J.; Kurgan, L. Accurate sequence-based prediction of catalytic residues. *Bioinformatics* **2008**, *24*(20), 2329-2338.

[51] Zhang, T.; Zhang, H.; Chen, K.; Ruan, J.; Shen, S.; Kurgan, L. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Prot. Pept. Sci*. **2010**, *11*(7), 609-628.

[52] Fogg, M.J.; Wilkinson, A.J. Higher-throughput approaches to crystallization and crystal structure determination. *Biochem. Soc. Trans.*, **2008**, *36*(4), 771-775.

[53] Grey, J.; Thompson, D. Challenges and opportunities for new protein crystallization strategies in structure-based drug design. *Expert Opin Drug Discov*, **2010**, 5(11), 1039-1045.

[54] Chou, K.C.; and Shen, H.B. Review: recent advances in developing web-servers for predicting protein attributes. *Natural Science*, **2009**, *2*, 63-92