Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Meta prediction of protein crystallization propensity

Marcin J. Mizianty, Lukasz Kurgan *

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

ARTICLE INFO

Article history:
Received 4 September 2009
Available online 13 September 2009

Keywords:
Structural genomics
X-ray crystallography
Crystallization propensity prediction
Protein structure
Protein crystallization

ABSTRACT

Production of high-quality crystals is one of the main bottlenecks in the X-ray crystallography driven protein structure determination. Availability of structure determination data repositories, such as TargetDB and PepcDB, and flexibility in target selection in structural genomics motivate development of methods that predict crystallization propensity from a given protein sequence. We introduce a novel linear model tree-based meta-predictor, MetaPPCP, which takes advantage of the complementarity of state-of-the-art protein crystallization propensity predictors to provide predictions with about 80% accuracy. Our method combines predictions of XtalPred and CRYSTALP2 with information concerning isoelectric point, hydropathy and number of solved structures for similar sequences. Empirical comparison shows that MetaPPCP outperforms current predictors including OB-Score, XtalPred, ParCrys, and CRYSTALP2. MetaPPCP obtains over 92% accuracy for over a half of its predictions that have probability (propensity to be predicted as crystallizable or crystallization resistant) of above 0.8. The proposed method could provide useful input for target selection procedures of current structural genomics efforts.

Introduction

Proteins are organic compounds composed of amino acids arranged in a linear chain polymer. They adopt an immense variety of shapes which allows them to implement a wide variety of functions such as transportation, signaling, catalysis, formation of the cell cytoskeleton, immune responses, etc. Knowledge of the tertiary protein structure is vitally important for understanding and manipulating their biochemical and cellular functions, which is used in drug design [1], to gain insights into various diseases [2] and to decipher protein–ligand interactions [3]. We currently know over 8 millions non-redundant protein chains while the corresponding structure is known for "only" about 55 thousand proteins deposited into the Protein Data Bank (PDB) database [4]. This wide sequence-structure gap calls for increased efforts in acquiring protein structures, such as structural genomics (SG) [5]. The SG initiatives perform protein family-directed structure analyses in which a group of proteins is targeted and structure(s) of representative members are determined and used to represent the entire group [6].

The most popular method for determination of the protein structure, which accounts for approximately 86% of the solved and deposited structures, is X-ray crystallography [7]. One of the main challenges for the SG initiative it that only about 2–10% of pursued targets yield high-resolution structures [8]. Analysis of data published in the TargetDB database [9], a world-wide repository for information on the experimental progress and status of targets selected for structure determination, shows that only about 8.6% of input chains are successfully crystallized and 4.6% gives diffraction quality crystals [7]. Estimates show that failed attempts account for more than 60% of the structure determination costs [10]. Several strategies have been proposed to improve the success rate [11,12], but the production of high-quality crystals is still one of the major bottlenecks in the protein structure determination [13–15].

The fact that SG allows for certain flexibility in selection of the chains for the crystallization-based structure determination motivates development of methods that predict/assess crystallization propensity for a given protein sequence. The existing crystallization propensity predictors include SECRET [16], OB-Score [17], CRYSTALP [18], XtalPred [10,19], ParCrys [20], and CRYSTALP2 [21]. Some of them were already successfully used to improve structure production at SG centers [17,19]. Two early methods, namely SECRET and CRYSTALP, accept only sequences between 46 and 200 amino acids in length. The remaining predictors are characterized by similar prediction accuracy of about 70% [21] and differ in their design and the information used as their input [7]. Recent results, which are empirically confirmed in this study, show that although these predictors provide similar predictive performance, their predictions are complementary with each other. For instance, CRYSTALP2 is shown to provide correct predictions for about 15% and 13% of protein chains that XtalPred and ParCrys, respectively, predict incorrectly and XtalPred and ParCrys provide

* Corresponding author. Address: Department of Electrical and Computer Engineering, ECERF, 9107 116 Street, University of Alberta, Edmonton, Alta., Canada T6G 2V4. Fax: +1 780 492 1811.
E-mail address: lkurgan@ece.ualberta.ca (L. Kurgan).

0006-291X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2009.09.036

correct predictions for 15% and 14% of chains for which CRYSTALP2 makes mistakes [21]. Also, large scale tests demonstrate that about 90% of the chains can be correctly predicted by at least one of the four methods, which suggests that an ensemble of these methods could provide improvements when compared with the individual predictors [7]. A simple vote-based combination of these predictors provides only relatively minor improvements, e.g. accuracy of 73.6% vs. 70.6% was obtained with a vote-based ensemble and best performing individual method, respectively [7], which motivates development of more advanced meta-predictors. Advanced ensembles, which utilize classification models on the outputs generated by atomic (base) predictors, were already found useful in related studies including prediction of protein folds [22], secondary structure [23], and gene function [24], to name just a few. To this end, we introduce a novel linear model tree-based meta-predictor for protein crystallization propensity, named MetaPPCP, which takes advantage of the complementarity of the state-of-the-art protein crystallization propensity predictors to improve the quality of the prediction. The proposed method is characterized by a novel design in which protein chains are partitioned into subsets using a decision tree, where for each subset a different logistic regression model is used to predict the crystallization propensity.

Materials and methods

Datasets. We use a dataset composed of 2000 protein chains which was originally introduced in [21] and which was developed using procedure proposed in [20]. The crystallizable proteins were extracted from sequences deposited in TargetDB and they include the last 1000 depositions as of December 2008. The non-crystallizable sequences, which correspond to the actual construct sequences used, were extracted from the last 1000 trial sequences as of December, 2008, deposited into PepcDB [25]. Duplicate sequences were removed and the remaining sequences were processed to remove the N-terminal hexaHis tag and LEHHHHHH tag at the C-terminus, which are introduced to ease purification. This dataset was randomly divided into two disjoint subsets, TRAINING dataset composed of 1500 chains that is used to train and parameterize the proposed method (using 5-fold cross validation procedure) and TEST500 dataset that is used to compare MetaPPCP with existing method. We also use TEST144 dataset which is the largest test set from [20] and which consists of 72 crystallizable and 72 non-crystallizable chains. The datasets are available from http://biomine.ece.ualberta.ca/MetaPPCP/MetaPPCP.html.

Quality measures. The predictions were compared with the original annotations from the TargetDB to assess the prediction quality. Four outcomes are possible: TP (true positive)/FN (false negative) which corresponds to crystallizable chains that were correctly/incorrectly predicted as crystallizable/non-crystallizable, respectively, and FP (false positive)/TN (true negative) which indicates that non-crystallizable chains were incorrectly/correctly predicted as crystallizable/non-crystallizable, respectively. The predictions were assessed based on the following quality indices:

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

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

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

The accuracy measures the fraction of correct predictions among all predictions. The Matthews Correlation Coefficient (MCC) is confined to $\langle -1,1 \rangle$ interval where 0 corresponds to prediction equivalent to a random classification. Higher MCC value corresponds to better performance of the prediction method. TPR (true positive rate) and TNR (true negative rate) quantify the fraction of correctly predicted crystallizable (positive) and non-crystallizable (negative) proteins, respectively. We also report receiver-operator characteristics (ROC) curves that present a graphical plot of the TP rate = TP/(TP + FN) against FP rate = FP/(FP + TN). This is performed by thresholding the probabilities (confidence values) that are generated together with the predicted classes (crystallizable vs. non-crystallizable). These plots are also used to compute the area under the ROC curve (AROC). The higher the AROC value is the better the predictive power of the corresponding method.

Design of meta-predictor. The considered inputs (features) for the proposed meta-predictor encompass outputs generated by OB-Score, ParCrys, XtalPred, and CRYSTALP2 predictors. They include predicted class (crystallizable vs. non-crystallizable) and prediction score (estimated prediction probability) for the four methods. We also consider information generated by the XtalPred server, which includes length and isoelectric point (p*I*) of the input sequence, its Gravy and instability index values, average number of insertions in the alignment compared to homologs (structures with similar sequence) in non-redundant (NR) database, number of homologs in NR and PDB databases, and predicted percentage of coils, coiled coils, longest disorder region, transmembrane helices, and signal peptides [10,19]. We evaluated a wide range of prediction models implemented in WEKA platform [26] that include linear logistic regression (LOG) and nonlinear Support Vector Machine (SMV), probabilistic Naïve Bayes (NB), C4.5 decision tree (C4.5), and logistic model tree (LMT). Each of the classification models was parameterized using the full set of features and 5-fold cross validation on the TRAINING dataset. The parameters for SVM include kernel types (RBF kernel with different widths, polynomial and normalized polynomial with different degrees) and complexity parameter *C*, ridge value for LOG, tree pruning factor and minimal number of instances at leaf nodes for C4.5, and minimal number of instances in leaves for LMT. The parameterized classifiers were used to perform two best-first search based feature selections. The features were either added one at the time starting with empty set (forward search) or they were removed one at the time starting with the set of all features (backward search). The features were added/removed based on the average, over the 5-folds, MCC value of a given parameterized classifier that uses the selected features. We repeated the cross validations for up to five times using randomized division into 5-folds for as long as the coefficient of variation (the ratio of the standard deviation to the mean) was below 0.02 to assure a robust estimate of the MCC value. Next, each classifier was parameterized again using the selected feature set and 5-fold cross validation on the TRAINING dataset. As a result, we have 15 designs where five types of prediction models are executed on three different feature sets.

The above designs were compared against a baseline predictor based on a majority-vote in which the output is the most frequent prediction of its base methods. Since we consider four base methods (OB-Score, ParCrys, XtalPred, and CRYSTALP2), we selected the best performing, according to MCC values, configuration among four combinations of three methods and four designs in which all four methods are used and where the tie-break (2 vs. 2 split decision) is resolved by applying the prediction of one of the methods. We also estimated an upper limit of the prediction quality for a meta-predictor by assuming that a given prediction is correct if at least one of the four methods provides a correct prediction for the corresponding protein sequence.

Table 1
Summary of results, ordered by MCC values, on the TRAINING, TEST500, and TEST144 datasets. Results on TRAINING set include the best configurations of ensembles based on SVM, LMT (MetaPPCP), NB, LOG, and C4.5 classifiers, baseline majority vote-based ensemble, and existing predictors including ParCrys, CRYSTALP2, XtalPred, and OB-Score. The results on the TEST500 and TEST144 datasets compare MetaPPCP, SVM-based meta-predictor, ParCrys, CRYSTALP2, XtalPred, and OB-Score.

Dataset	Prediction method				No. of features	Accuracy	MCC	TRP	TNR
	Type	Classifier	Feature selection	Parameters					
TRAINING	Meta	SVM[a]	Forward best-first	RBF kernel, *width* = 2.5, *C* = 2	12	79.33	0.59	0.88	0.71
		LMT[a]	Forward best-first	*# inst.* = 15	5	78.40	0.58	0.88	0.69
		NB[a]	Backward best-first	Not applicable	12	77.67	0.57	0.90	0.65
		LOG[a]	Forward best-first	*Ridge* = 1.0E-8	8	77.47	0.55	0.84	0.71
		C4.5[a]	Backward best-first	*Pruning* = 0.05, *# inst.* = 20	6	76.80	0.55	0.89	0.65
		Majority vote-based ensemble[b]			4	73.12	0.48	0.86	0.60
		At least one correct[c]			4	90.53	0.82	0.99	0.82
	Base	ParCrys[d]			8	69.73	0.41	0.83	0.56
		CRYSTALP2[e]			88	69.60	0.40	0.77	0.62
		XtalPred[f]			9	69.27	0.39	0.75	0.63
		OB-Score[d]			2	68.80	0.39	0.85	0.53
TEST500	Meta	MetaPPCP			5	81.00	0.63	0.89	0.73
		SVM-based ensemble			12	79.80	0.60	0.87	0.73
	Base	OB-Score[d]			2	73.00	0.49	0.89	0.58
		ParCrys[d]			8	73.40	0.48	0.84	0.63
		XtalPred[f]			9	72.40	0.45	0.77	0.68
		CRYSTALP2[e]			88	68.40	0.37	0.73	0.64
TEST144	Meta	MetaPPCP			5	80.56	0.61	0.82	0.79
		SVM-based ensemble			12	75.69	0.51	0.78	0.74
	Base	OB-Score[d]			2	67.36	0.38	0.88	0.47
		ParCrys[d]			8	68.75	0.38	0.79	0.58
		XtalPred[f]			9	79.17	0.58	0.79	0.79
		CRYSTALP2[e]			88	75.69	0.52	0.79	0.72

[a] Results based on 5-fold cross validation on the TRAINING dataset.
[b] The best performing majority-vote ensemble with 4 base predictors and CRYSTALP2 as the tie-breaker.
[c] Estimate of the upper limit on prediction quality of a meta-predictor in which a prediction is assumed correct if any of the four base predictors provides correct result.
[d] Results computed using the ParCrys/OB-Score server at http://www.compbio.dundee.ac.uk/xtal/.
[e] Results computed using the CRYSTALP2 model at http://biomine.ece.ualberta.ca/CRYSTALP2/CRYSTALP2.html.
[f] Results computed using the XtalPred server at http://ffas.burnham.org/XtalPred/.

Table 1 compares the best, with respect to the MCC values, results from the 5-fold cross validation on the TRAINING dataset obtained with each of the five prediction model types (among the results on the three features sets), with the results obtained on the whole TRAINING dataset (without cross validation) of the baseline ensemble and the four base predictors. As expected, each of the five classifier-based ensembles provides predictions that outperform results from the four state-of-the-art existing predictors as well as the results with the majority vote-based method. The improvements in accuracy range between 3.5 and 10 percentage points. The best performing meta-predictors achieve around 79% accuracy, while the upper limit is estimated to be 90.5%. The two top scoring ensembles are based on SVM and LMT classifiers where the former uses 12 features and the latter only 5. Subsequent analysis on the test datasets, see Table 1, shows that the SVM-based solution does not generalize into other dataset as well as the LMT-based method. Also, the LMT model is white-box (can be analyzed and understood by users) and is less complex as it uses fewer features. The LMT [27], which is selected to implement the proposed MetaPPCP method, is a binary decision tree, which is built using C4.5 algorithm [28], with linear regression models, which are derived with the LogitBoost algorithm [29], at the leaves. Predictions are obtained by descending the tree branches to a leaf and using the associated linear model to compute class membership probabilities.

$$P(\text{crystallizable}) = e^{LRcryst}/(e^{LRcryst} + e^{LRnoncryst})$$

$$P(\text{non-crystallizable}) = e^{LRnoncryst}/(e^{LRnoncryst} + e^{LRcryst})$$

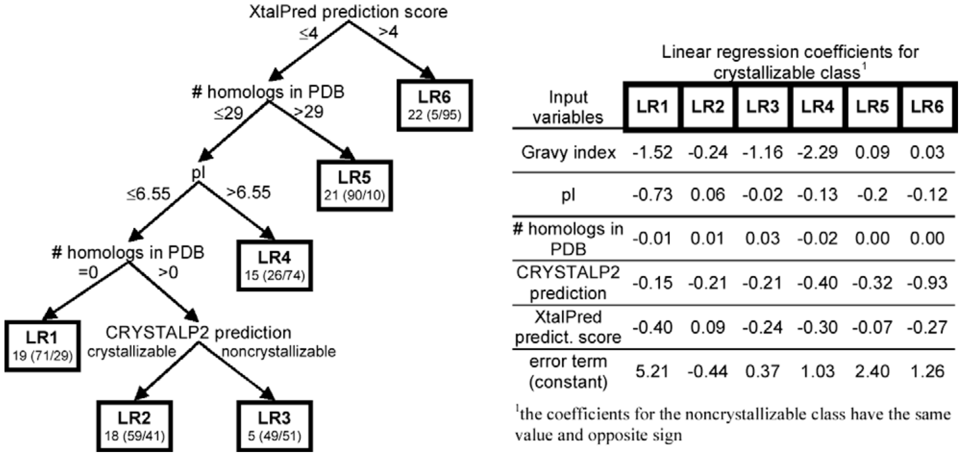
where *LRcryst* and *LRnoncryst* are the values produced by the linear regression models for the crystallizable and non-crystallizable classes, respectively. Since in our binary prediction problem *LRcryst* = −*LRnoncryst*, see Fig. 1, the probability of a given chain to be predicted as crystallizable/non-crystallizable increases as the value its corresponding linear model (*LRcryst*/*LRnoncryst*) increases.

Results

Prediction model

The LMT model of the proposed MetaPPCP method is shown in Fig. 1. The model is based on five features that include the CRYSTALP2 prediction, and prediction score, number of homologs in PDB, Gravy hydropathy index [30], and p*I* values provided by the XtalPred server. Using HR1946 target sequence from TargetDB as an example, the corresponding Gravy index = −0.35, p*I* = 7.97, number homologs in PDB = 15, CRYSTALP2 prediction = 0 (non-crystallizable) and XtalPred score = 5. Since the XtalPred score equals 5, we use the LR6 model to compute *LR6cryst* = −1.0569 and *LR6noncryst* = 1.0569. These values are used to calculate *P*(crystallizable) = 0.11 < *P*(non-crystallizable) = 0.89, and thus this target is predicted as non-crystallizable with probability of 0.89.

The most likely reason for inclusion of the XtalPred and CRYSTALP2 predictions is that they are characterized by stronger complementarity when compared with the other two base methods. More specifically, 84.9% of proteins in the TRAINING dataset are correctly predicted by XtalPred or CRYSTALP2, while 77.4% are successfully predicted by OB-Score or ParCrys. This agrees with the results shown in [7]. The isoelectric point and hydrophobicity were used, similarly as in MetaPPCP, to implement both OB-Score and ParCrys methods, which also explains why output from these two methods was not utilized.



Input variables	Linear regression coefficients for crystallizable class[1]					
	LR1	LR2	LR3	LR4	LR5	LR6
Gravy index	-1.52	-0.24	-1.16	-2.29	0.09	0.03
pI	-0.73	0.06	-0.02	-0.13	-0.2	-0.12
# homologs in PDB	-0.01	0.01	0.03	-0.02	0.00	0.00
CRYSTALP2 prediction	-0.15	-0.21	-0.21	-0.40	-0.32	-0.93
XtalPred predict. score	-0.40	0.09	-0.24	-0.30	-0.07	-0.27
error term (constant)	5.21	-0.44	0.37	1.03	2.40	1.26

[1]the coefficients for the noncrystallizable class have the same value and opposite sign

Fig. 1. The proposed prediction model; the decision tree is shown on the left and the linear regression (LR) models from the leaf nodes are shown on the right. The left most "LR1 19 (71/29)" leaf node denotes that its corresponding linear regression is LR1, and that this node concerns 19% of the input proteins among which 71% are crystallizable and 29% are non-crystallizable.

Comparison with existing predictors

The MetaPPCP method was compared with OB-Score, ParCrys, XtalPred, and CRYSTALP2 on the two test datasets, TEST500 and TEST144, see Table 1. The results demonstrate that MetaPPCP outperforms the existing methods by 7.6 and 1.4 percentage points on the two datasets, respectively. The smaller difference on the TEST144 dataset can be explained by its relatively small size. The proposed model consistently provides predictions with about 80% accuracy, while the second best on the TEST144 dataset XtalPred obtains 72% accuracy and is ranked behind OB-Score and ParCrys on the TEST500 dataset. All considered predictors are characterized by TPR higher than TNR, which indicates that they perform better when predicting propensity of crystallizable chain when compared with the predictions for the crystallization resistant chains. Depending on the test dataset used, MetaPPCP correctly predict 82% or more crystallizable chains and 73% or more non-crystallizable proteins. To compare, XtalPred correctly predicts 79/68% or more of the crystallizable/crystallization resistant chains, respectively.

Comparison using ROC curves, see Fig. 2, shows that MetaPPCP outperforms all competing methods for the entire range of FP and TP rates on the TEST500 dataset. The AROC of MetaPPCP, which equals 0.88, is larger by 0.11 when compared with the second best predictor. The results on the smaller TEST144 dataset show that MetaPPCP provides favorable TP rates for small FP rates and is outperformed only by XtalPred for larger FP rates. At the same time, the proposed method still obtains the largest AROC value on this dataset.

Discussion

The proposed prediction model demonstrates that isoelectric point, Gravy index, and availability of solved homolog structures can be successfully used to augment and combine predictions from the XtalPred and CRYSTALP2 methods. The utility of this information is supported by prior works which demonstrate strong relation between hydropathy/isoelectric point and the crystallization propensity [15,17,19,31] and which use these indices in related studies,

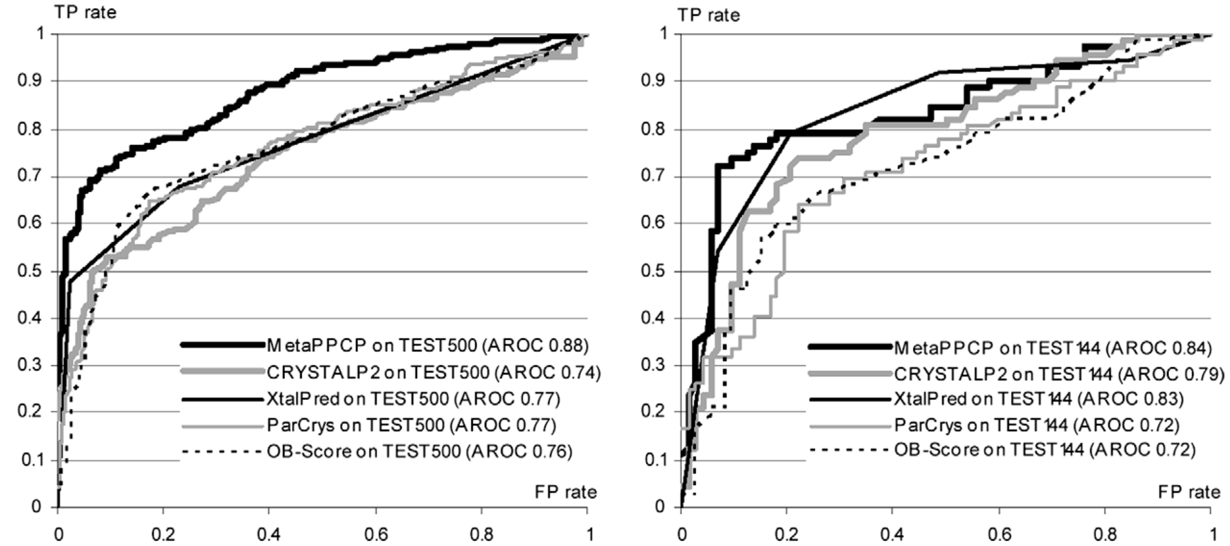


Fig. 2. The ROC curves for the predictions of MetaPPCP, ParCrys, CRYSTALP2, XtalPred, and OB-Score on TEST500 (left panel) and TEST144 (right panel) datasets. The AROC values are given in figure legends.

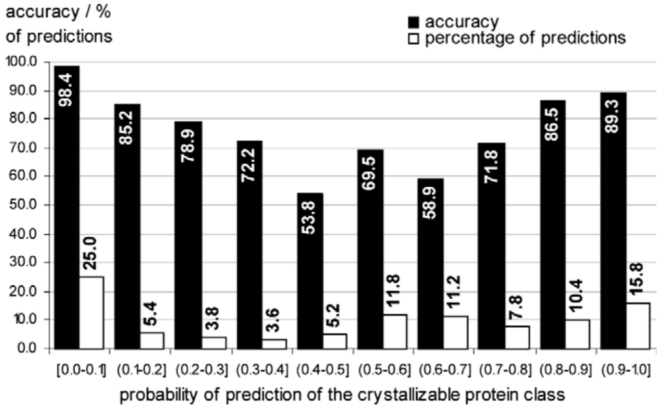


Fig. 3. The distribution of the prediction accuracy (black bars) and percentage of predictions (white bars) in the function of the probability of the prediction of crystallizable proteins for the TEST500 dataset.

such as for suggesting optimal pH ranges for crystallization screening [32,33] and prediction of the protein production success (which covers processes between DNA cloning and protein purification) [19].

The decision tree from Fig. 1 divides the protein chains into 6 groupings that are formed by descending along the branches to the leaf nodes. Three of these sets, which correspond to LR1, LR2, and LR5 models, cover chains that are predominantly crystallizable, another two (LR4 and LR6) include proteins that are predominantly resistant to crystallization and the remaining LR3 includes approximately equal number of both protein classes. Using the subset corresponding to LR6 as an example, we observe (see regression coefficients in Fig. 1) that larger values of the Gravy index and smaller p*I* are associated with proteins that are more likely to crystallize. On the contrary, for the data in the LR2 leaf node, the probability of successful crystallization increases as the Gravy index decreases and the p*I* increases. The complex nature of the relations between crystallization propensity and p*I* and Gravy indices, which are nonlinear and have multiple optima, was observed for data generated at the Joint Center for Structural Genomics [19]. Our model attempts to partition the sequence space into subspaces that allows for an accurate linear multivariate approximation of these relations.

The proposed meta-predictor has an explicit human-readable model, which requires information coming from two based predictors, XtalPred and CRYSTALP2. In spite of its simple design, MetaPPCP provides predictions with over 80% accuracy and is shown to outperform existing methods on both considered test datasets. Our model also outputs probability of the prediction, *P*(crystallizable) = 1 – *P*(non-crystallizable), which indicates confidence in the prediction outcome that corresponds to the class (crystallizable vs. non-crystallizable) associated with higher probability. Fig. 3 shows that predictions with high probabilities (for either class) are characterized by better performance than the predictions where probabilities for the two classes are similar. For instance, predictions with probabilities >0.9 for crystallizable/non-crystallizable class are 89.3/98.4% accurate, respectively. The MetaPPCP obtains the accuracy of 92.2% for the approximately 57% of predictions that have probability >0.8 for one of the classes.

We note that all investigated crystallization propensity predictors consider only intra-molecular factors encoded in the protein chain. They may not provide accurate predictions when intermolecular factors such as protein–protein and/or protein–precipitant interactions, buffer composition, etc. must be considered. These methods are limited to predictions for non-redundant chains. We recommend the use of the surface entropy reduction server [34] when assessing crystallization of close homologs.

Conclusion

We introduce a novel white-box sequence-based protein crystallization propensity predictor. The proposed method provides predictions with about 80% accuracy and is shown to outperform current methods. Our model confirms that isoelectric point and hydropathy are important for the crystallization prediction. The probabilities produced by the proposed method can be used to indicate higher quality predictions. For instance, predictions with probabilities of above 0.8 are characterized by over 92% accuracy. We believe that our model could provide useful input for target selection procedures utilized by structural genomics centers and structural biologists.

References

[1] R.V. Guido, G. Oliva, A.D. Andricopulo, Virtual screening and its integration with modern drug design technologies, Current Medicinal Chemistry 15 (1) (2008) 37–46.

[2] X. Fernàndez-Busquets, N.S. de Groot, D. Fernandez, S. Ventura, Recent structural and computational insights into conformational diseases, Current Medicinal Chemistry 15 (2008) 1336–1349.

[3] K. Chen, L. Kurgan, Investigation of atomic level patterns in protein–small ligand interactions, PLoS ONE 4 (2) (2009) e4473.

[4] 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.

[5] S.E. Brenner, A tour of structural genomics, Nature Reviews Genetics 2 (10) (2001) 801–809.

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

[7] L. Kurgan, M.J. Mizianty, Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis, Natural Science 1 (2) (2009) 93–106.

[8] R. Service, Structural genomics, round 2, Science 307 (2005) 1554–1558.

[9] L. Chen, R. Oughtred, H.M. Berman, J. Westbrook, TargetDB: a target registration database for structural genomics projects, Bioinformatics 20 (16) (2004) 2860–2862.

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

[11] A. Savchenko, A. Yee, A. Khachatryan, T. Skarina, E. Evdokimova, M. Pavlova, A. Semesi, J. Northey, S. Beasley, N. Lan, R. Das, M. Gerstein, C.H. Arrowmith, A.M. Edwards, Strategies for structural proteomics of prokaryotes: quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches, Proteins 50 (2003) 392–399.

[12] J.M. Chandonia, S.E. Brenner, Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches, Proteins 58 (2005) 166–179.

[13] N.E. Chayen, Turning protein crystallisation from an art into a science, Current Opinion in Structural Biology 14 (5) (2004) 577–583.

[14] M. Puesy, Z.J. Liu, W. Tempel, J. Praissman, D. Lin, B.C. Wang, J.A. Gavira, J.D. Ng, Life in the fast lane for protein crystallization and X-ray crystallography, Progress in Biophysics and Molecular Biology 88 (2005) 359–386.

[15] W.N. Price, Y. Chen, S.K. Handelman, H. Neely, P. Manor, R. Karlin, R. Nair, J. Liu, M. Baran, J. Everett, S.N. Tong, F. Forouhar, S.S. Swaminathan, T. Acton, R. Xiao, J.R. Luft, A. Lauricella, G.T. DeTitta, B. Rost, G.T. Montelione, J.F. Hunt, Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data, Nature Biotechnology 27 (1) (2009) 51–57.

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

[17] I.M. Overton, G.J. Barton, A normalised scale for structural genomics target ranking: the OB-Score, FEBS Letters 580 (2006) 4005–4009.

[18] K. Chen, L. Kurgan, M. Rahbari, Prediction of protein crystallization using collocation of amino acid pairs, Biochemical and Biophysical Research Communications 355 (2007) 764–769.

[19] L. Slabinski, L. Jaroszewski, A.P.C. Rodrigues, L. Rychlewski, I.A. Wilson, S.A. Lesley, A. Godzik, The challenge of protein structure determination—lessons from structural genomics, Protein Science 16 (11) (2007) 2472–2482.

[20] I.M. Overton, G. Padovani, M.A. Girolami, G.J. Barton, ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction, Bioinformatics 24 (2008) 901–907.

[21] L. Kurgan, A.A. Razib, S. Aghakhani, S. Dick, M.J. Mizianty, S. Jahandideh, CRYSTALP2: sequence-based protein crystallization propensity prediction, BMC Structural Biology 9 (2009) 50.

[22] K. Chen, L. Kurgan, PFRES: protein fold classification by using evolutionary information and predicted secondary structure, Bioinformatics 23 (21) (2007) 2843–2850.

[23] S. Montgomerie, S. Sundararaj, W.J. Gallin, D.S. Wishart, Improving the accuracy of protein secondary structure prediction using structural alignment, BMC Bioinformatics 7 (2006) 301.

[24] Y. Guan, C.L. Myers, D.C. Hess, Z. Barutcuoglu, A.A. Caudy, O.G. Troyanskaya, Predicting gene function in a hierarchical context with an ensemble of classifiers, Genome Biology 9 (Suppl. 1) (2008) S3.

[25] A. Kouranov, L. Xie, J. de la Cruz, L. Chen, J. Westbrook, P.E. Bourne, H.M. Berman, The RCSB PDB information portal for structural genomics, Nucleic Acids Research 4 (2006) D302–D305.

[26] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, San Francisco, 2005.

[27] N. Landwehr, M. Hall, E. Frank, Logistic model trees, Machine Learning 59 (1–2) (2005) 161–205.

[28] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[29] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Annals of Statistic 38 (2) (2000) 337–374.

[30] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology 157 (1982) 105–132.

[31] J.M. Canaves, R. Page, I.A. Wilson, R.C. Stevens, Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: maximum clustering strategy for structural genomics, Journal of Molecular Biology 344 (2004) 977–991.

[32] K.A. Kantardjieff, B. Rupp, Protein isoelectric point as a predictor for increased crystallization screening efficiency, Bioinformatics 20 (2004) 2162–2168.

[33] K.A. Kantardjieff, M. Jamshidian, B. Rupp, Distributions of p*I* vs. pH provide strong prior information for the design of crystallization screening experiments, Bioinformatics 20 (2004) 2171–2174.

[34] L. Goldschmidt, D.R. Cooper, Z. Derewenda, D. Eisenberg, Toward rational protein crystallization: a web server for the design of crystallizable protein variants, Protein Science 16 (2007) 1569–1576.