OXFORD

Briefings in Bioinformatics, 19(5), 2018, 838–852

doi: 10.1093/bib/bbx018
Advance Access Publication Date: 27 February 2017
Paper

Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity

Huilin Wang, Liubin Feng, Geoffrey I. Webb, Lukasz Kurgan, Jiangning Song and Donghai Lin

Corresponding authors: Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, USA. Tel.: +1 804-827-3986; Email: lkurgan@vcu.edu; Jiangning Song, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. Tel.: +61-3-9902-9304; Fax: +61-3-9902-9500; Email: Jiangning.Song@monash.edu; Donghai Lin, The Key Laboratory for Chemical Biology of Fujian Province, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China. Tel.: +86-592-2186078; Fax: +86-592-2186078; Email: dhlin@xmu.edu.cn

Abstract

X-ray crystallography is the main tool for structural determination of proteins. Yet, the underlying crystallization process is costly, has a high attrition rate and involves a series of trial-and-error attempts to obtain diffraction-quality crystals. The Structural Genomics Consortium aims to systematically solve representative structures of major protein-fold classes using primarily high-throughput X-ray crystallography. The attrition rate of these efforts can be improved by selection of proteins that are potentially easier to be crystallized. In this context, bioinformatics approaches have been developed to predict crystallization propensities based on protein sequences. These approaches are used to facilitate prioritization of the most promising target proteins, search for alternative structural orthologues of the target proteins and suggest designs of constructs capable of potentially enhancing the likelihood of successful crystallization. We reviewed and compared nine predictors of protein crystallization propensity. Moreover, we demonstrated that integrating selected outputs from multiple predictors as candidate input features to build the predictive model results in a significantly higher predictive performance when compared to using these predictors individually. Furthermore, we also introduced a new and accurate predictor of protein crystallization propensity, Crysf, which uses functional features extracted from UniProt as inputs. This comprehensive review will assist structural biologists in selecting the most appropriate predictor, and is also beneficial for bioinformaticians to develop a new generation of predictive algorithms.

Huilin Wang received his MEng in Bioengineering from Jiangnan University, Wuxi, China. He is currently pursuing his PhD in the Department of Chemical Biology, College of Chemistry and Chemical Engineering, Xiamen University, China. His research interests are structural bioinformatics, data mining and machine learning.

Liubin Feng received his BSc and MSc in Physics from Sichuan University, Chengdu, China. He is an engineer with the NMR Center, Xiamen University, China.

Geoffrey I. Webb is a Professor at the Monash Centre for Data Science, Faculty of Information Technology, Monash University, Australia. His research interests are machine learning, data mining, bioinformatics and protein design.

Lukasz Kurgan is the Qimonda Endowed Professor and Vice Chair of the Department of Computer Science at the Virginia Commonwealth University, USA. His research interests include structure and function prediction and modeling of proteins and small RNAs, characterization of sequence-structure/disorder-function relationships in proteins, and prediction and functional characterization of intrinsic disorder. More details on the Web site of his laboratory at http://biomine.cs.vcu.edu/.

Jiangning Song is a Senior Research Fellow at the Biomedicine Discovery Institute and the Department of Biochemistry and Molecular Biology, Monash University, Australia. He is also a Principal Investigator at the Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences. He conducted his postdoctoral research at The University of Queensland, Australia and Kyoto University, Japan. His research interests are bioinformatics, systems biology, machine learning, systems pharmacology and enzyme engineering.

Donghai Lin received his PhD in Physical Chemistry in 1993 from Xiamen University, China. He is a professor and a group leader in the High-Field Nuclear Magnetic Resonance Research Centre and the Department of Chemical Biology, College of Chemistry and Chemical Engineering, Xiamen University, China. His laboratory uses multi-disciplinary approaches to understand protein structure, dynamics and function, involving protein crystallography, nuclear magnetic resonance spectroscopy, molecular simulation and bioinformatics.

Submitted: 19 October 2016; **Received (in revised form):** 19 January 2017

Key words: structural genomics; protein crystallization propensity; target selection; bioinformatics; sequence analysis; machine learning

Introduction

X-ray crystallography is the main approach that is used to solve three-dimensional protein structures, accounting for about 90% of the proteins in the Protein Data Bank (PDB) [1]. Unfortunately, this method is characterized by a high attrition rate related to a difficult task of obtaining diffraction-quality protein crystals. Protein crystallization experiments often involve 'trial-and-error' attempts, and are composed of several laborious steps including sequence cloning, protein expression, solubility analysis, purification and ultimately production of diffraction-quality protein crystals [2–4]. Depending on the source, the success rate of the overall crystallization process was reported to range between 2% and 10% [5–9]. Consequently, a large portion of the overall cost (up to 70%) is spent on the failed attempts [10]. Moreover, availability of a large and growing number of sequenced proteins (the UniProt database [11] includes over 66 million protein sequences) puts pressure to increase the high-throughput of structural determination pipelines. To bridge this widening sequence-structure gap, the Structural Genomics Consortium (SGC) was founded following the successful footsteps of sequencing efforts [12–14]. The primary task of SGC is to solve at least one representative structure for each biologically important protein-fold family by using primarily high-throughput X-ray crystallography [15–17]. To expedite this process, it is necessary to select feasible target proteins for structural determination, which cover novel fold families [18]. Both the selected target proteins and their experimental progresses and statuses are available in the target-registration database TargetTrack (http://sbkb.org/tt/) [19, 20]. As of 8 October 2016, only 4.7% of cloned proteins (10 763/230 782 proteins) have been structurally solved according to the data in TargetTrack. This suggests that the currently existing prediction algorithms for target selection could be further improved and optimized.

Target selection remains a challenging task [21, 22], owing to complex sequence-structure relationships and our limited understanding of biophysical properties of proteins as well as complicated factors controlling protein crystallization [23, 24]. Bioinformatics tools capable of predicting the propensity of a given protein to yield diffraction-quality crystals, have been developed to facilitate target selection and help streamline laborious trial-and-error experimental steps. These bioinformatics tools can be used to select alternative structural orthologues with high likelihoods of successful crystallization, thereby potentially reducing the attrition rates. They were also used to perform large-scale assessments of protein crystallization propensities (CPs) of entire proteomes [21] and collections of proteomes [25]. A variety of bioinformatics tools have been developed in the past decade. They mostly focus on predicting protein solubility [26–28], and crystallization [2–4, 10, 29–34], based on protein crystallization-dependent factors, including transmembrane regions [35, 36], secondary structure elements [37, 38], signal peptides [39] and disordered segments [40]). These bioinformatics tools can provide useful information to guide protein crystallization [23, 41].

In this comprehensive review, we discuss recent progresses in developing protein sequence-based bioinformatics tools for the prediction of protein CP. These prediction tools were developed by exploiting significant differences between collections of crystallizable and non-crystallizable proteins using statistical and machine-learning algorithms. They can be categorized into three major classes. The first class is based on physicochemical properties predicted from protein sequences, which are used to differentiate crystallizable and non-crystallizable proteins. The commonly used properties include sequence length, isoelectric point (p*I*), 20 standard amino acid composition, hydrophobicity, frequencies of hydrophobic, hydrophilic, positively charged, negatively charged and neutral amino acid residues and dipeptide composition as well as tripeptide composition. These bioinformatics tools predict protein CP solely based on amino acid sequence, and are characterized by relatively low runtime, in the order of < 1 s per protein. They include SECRET [30], CRYSTALP [31], OB-Score [29], ParCrys [32], CRYSTALP2 [33], MCSG-Z score [42], SCMCRYS [43] and Crysalis [3].

The second class uses putative structural features predicted from protein sequences including secondary structures, intrinsically disordered regions and solvent accessibilities of residues, in addition to physicochemical properties. These bioinformatics tools use a large set of structural features, such as hydrophobicity and side-chain entropy of putative protein surface, frequency of buried amino acid residues, number of disordered regions and composition of secondary structures. These features are in line with the recent results obtained on a large set of proteins [3]. These results suggested that the sites of protein crystallization-dependent non-optimality might involve the biases associated with side-chain entropy, disordered regions, solvent-exposed regions and C- and N-termini localization [3]. The extension to use putative structural features aims to improve the predictive performance as a trade-off for a longer runtime. The increase in runtime is owing to the reason that computation of the putative structural features requires generation of multiple sequence alignment with the PSI-BLAST program [44]. Typically, these bioinformatics tools take 0.5–3 min to process a query protein sequence on a modern desktop computer. The corresponding predictors include XtalPred [45], P_{XS} [24], SVMCRYS [34], PPCPred [4], XANNPred [46], RFCRYS [47], CRYSpred [48], XtalPred-RF [10] and PredPPCrys [2].

A new bioinformatics tool named Crysf constitutes the third class, which will be introduced in this article. Crysf uniquely applies functional features of protein sequences extracted from UniProt as the input for predicting protein CP. Thus, Crysf can be used for selecting feasible target proteins from UniProt for structural determination.

As nearly 20 bioinformatics tools have been developed for predicting protein CP, it might be difficult to choose the most suitable tool for a given protein. Thus, we herein review a comprehensive group of protein sequence-based CP predictors. We focus on several practical aspects of these predictors including their usability and utilities. In particular, we discuss the used predictive models, benchmark training and test data sets and the inputs. We also describe how to select these predictors and their availability. Furthermore, we will evaluate and compare their predictive performances using two up-to-date benchmark data sets, which include the proteins with recently annotated statuses derived from TargetTrack [19]. We believe that this comprehensive review will assist structural biologists in selecting the most suitable tools for their projects.

Materials and methods

Collection of annotations of crystallization trials

TargetTrack (http://sbkb.org/tt/) is a structural genomics target-registration database serving as a successor of the PepcDB database [19]. It provides detailed annotations regarding experimental progresses and statuses of target proteins deposited by >40 structural genomics centers worldwide. We downloaded the most recent experimental records (9 September 2015) from TargetTrack, comprising 335 993 proteins and 944 479 experimental trials. Each protein is associated with potentially multiple trials representing a set of experimental procedures, which were used to determinate the three-dimensional (3D) structure of this protein. Inspired by previous works [2, 4], we constructed a new data set (named 'TTdata') based on the data extracted from TargetTrack using the following four criteria:

1. We only extracted X-ray crystallography-based experimental trials annotated with the most advanced experimental statuses. These statuses include 'selected', 'cloned', 'expressed', 'soluble', 'purified', 'crystallized', 'diffraction', 'crystal structure' or 'in PDB'. We grouped the proteins with the status of 'crystal structure' or 'in PDB' as crystallizable proteins (defined as the 'CRYS' class), and grouped those with other statuses as non-crystallizable proteins (defined as the 'NCRYS' class).
2. We only selected the experimental trials annotated with two states: 'work stopped'; 'in PDB' or 'crystal structure'.
3. We did not extract the experimental trials both before 1 January 2009 and after 31 December 2014. This could ensure that we only extracted recent data and excluded trials that are potentially still ongoing at present.
4. We eliminated non-crystallizable proteins sharing >100% sequence identity with crystallizable proteins. The sequence identity was quantified by the CD-Hit program [49].

The constructed TTdata includes 81 279 non-crystallizable proteins and 103 247 crystallizable proteins.

Collection of functional annotations

We retrieved functional annotations of the proteins from UniProt (http://www.UniProt.org/), which included 549 008 proteins from the Swiss-Prot database and 50 011 027 proteins from the TrEMBL database (on 14 July 2015). Swiss-Prot is a collection of entries that are reviewed and manually annotated using a literature search and curator-evaluated computational analysis. TrEMBL is not reviewed in which proteins are annotated computationally. We mapped the proteins in TTdata to both Swiss-Prot and TrEMBL via one-by-one matching of sequences sharing 100% sequence identity. Totally, 5849 crystallizable proteins (positive samples) and 4907 non-crystallizable (negative samples) proteins were mapped to the Swiss-Prot database, constituting the Swiss-Prot data set. Additionally, 8491 crystallizable (positive samples) and 21 426 non-crystallizable (negative samples) proteins were mapped to the TrEMBL database, comprising the TrEMBL data set.

Training and benchmark test data sets

Based on the Swiss-Prot and the TrEMBL data sets, we constructed the training and test data sets using the following three steps:

1. We eliminated sequence redundancy (proteins with >25% sequence identity) within crystallizable proteins contained in either Swiss-Prot or TrEMBL, also eliminated that within non-crystallizable proteins contained in each data set. The sequence identity was qualified by using a combination of CD-Hit [49] and BLAST [44]. Eliminating sequence redundancy within each data set was based on the observation that the proteins with similar sequences could possess distinct CPs [2]. Totally, the Swiss-Prot data set contains 2798 crystallizable and 3096 non-crystallizable proteins (denoted as the 'SP' data set), while the TrEMBL data set contains 4994 crystallizable and 9794 non-crystallizable proteins (denoted as the 'TR' data set).
2. Either the SP data set or the TR data set was randomly divided into six equally sized subsets. The first five subsets were merged together to form the training data set (denoted as 'SP_train' or 'TR_train'), while the remaining sixth subset worked as the independent test data set (denoted as 'SP_test' or 'TR_test').
3. We further eliminated the proteins sharing >25% sequence identity with those used in other predictors. The resulting four data sets were named as 'SP_train_nr', 'SP_test_nr', 'TR_train_nr' and 'TR_test_nr', respectively. These data sets can be downloaded from http://nmrcen.xmu.edu.cn/crysf/.

To examine whether the functional features of similar proteins can be used to predict CP, we mapped TTdata-derived sequences to Swiss-Prot and TrEMBL data sets via one-by-one matching of sequences sharing >90% sequence identity. The resultant data sets were named 'SP0.9' and 'TR0.9', respectively. Hence, each protein in SP0.9 or TR0.9 is associated with one or more orthologous proteins in the Swiss-Prot data set or the TrEMBL data set.

Encoding of protein functional features

We extracted the functional annotations from UniProt, and converted them into numeric functional features as the inputs of the predictive models. We executed the same encoding algorithm to extract functional features from Swiss-Prot and TrEMBL databases, both adopting the same data format. Functional annotations of proteins in UniProt are specified by several fields, e.g. 'RN', 'FT', 'CC', 'PE' and 'SQ'. We used a number of different types of functional annotations including FT (e.g. SIGNAL PROPEP, NP_BIND and DISULFID), CC (e.g. FUNCTION, SUBUNIT, INTERACTION, SUBCELLULAR LOCATION and PTM) and PE (i.e. protein existence), as well as annotations of subcellular locations. A detailed list of the UniProt-derived functional features is shown in Supplementary Table S1.

For each protein, a given functional annotation was encoded into a numeric feature according to the following protocol: a value of '1' was assigned if a protein had a given annotation, and '0' if the annotation was not present. If a given annotation was repeated for the same protein, then the corresponding value of the feature was set to the number of repetitions, e.g. number of certain types of binding sites or inclusion of multiple domains (refer to Supplementary Table S1). A total of 91 functional features were encoded for each protein based on the UniProt-derived annotations. These features were normalized as follows:

$$feat_i^N = \frac{1}{1 + feat_i}$$

where $feat_i$ is the real value of the i-th considered feature, and $feat_i^N$ is the normalization value.

Performance evaluation

Motivated by previous works [2–4], we used the area under the curve (AUC) scores as the primary measures to evaluate the predictive performances of these predictors. The AUC value quantifies the area under the receiver operating characteristic curve (ROC) by plotting the true-positive rate against the false-positive rate. Additionally, we used several other measures to quantify the performance when assessing binary predictions (i.e. crystallizable versus non-crystallizable):

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

$$\text{Accuracy (ACC)} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Sensitivity (SENS)} = \frac{TP}{TP+FN}$$

$$\text{Specificity (SPEC)} = \frac{TN}{TN+FP}$$

$$\text{Precision (PRE)} = \frac{TP}{TP+FP}$$

where *TP*, *FP*, *TN* and *FN* are the numbers of true positives, false positives, true negatives and false negatives, respectively. *TP* and *TN* denote the numbers of correctly predicted crystallizable and non-crystallizable proteins, respectively, while *FP* and *FN* represent the numbers of incorrectly predicted crystallizable and non-crystallizable proteins, respectively.

Forward-feature selection

Because some of the 91 features could be irrelevant to the prediction of CP (i.e. have low predictive power), and some could be redundant (cross-correlated with each other), we performed systematic selection of a subset with relevant and non-redundant features [50]. We performed sequential forward-feature selection (FFS), which was also previously used in related studies [51–53]. In FFS, the initial set of the features was divided into two groups: FFS candidate-feature set and FFS selected-feature set. At the beginning, the FFS selected-feature set was empty, and the initial FFS candidate-feature set was the complete feature set. In each round of FFS, each feature from the FFS candidate-feature set was added into the FFS selected-feature set. The latter was used to build a model (i.e. the support vector regression [SVR]-based model, as explained in section 2.7.1), and evaluate its predictive performance with the AUC value. We only added significant features into the FFS selected-feature set capable of improving the AUC value. Such a manner of selecting a set of features enabled maximizing the AUC value. We performed the entire selection process exclusively on the training data sets, and then assessed the predictive performance using 5-fold cross-validation test on these data sets.

Design of two new predictors

SVR models

We used the LIBSVM package 2.82 [54] to train and construct SVR models. There are three types of available kernels implemented in LIBSVM, namely sigmoid, radial basis function (RBF) and polynomial. We used the RBF kernel to perform feature selection and build the SVR models on the training data sets, thereafter evaluated the corresponding predictive performances using 5-fold cross-validation test. We made use of all three available kernels to build the final models with the selected-feature subsets. The respective kernel parameters (C and γ) were optimized by a grid search algorithm. The SVR models with the highest AUC values among the three kernel types were used to predict the CP.

Design of the CrysComb meta-predictor

Several predictors have been developed for predicting protein CPs (Table 1). We assessed whether they could be integrated into a meta-predictor to obtain predictive performances higher than individual predictors. We used 31 selected outputs from eight predictors as candidate input features to build the integrated CrysComb model on SP_train_nr and TR_train_nr data sets. The eight predictors include OB-Score, CRYSTALP2, SVMCRYS, PPCPred, SCMCRYS, XtalPred-RF, PredPPCrys and Crysalis.

Three recently developed predictors, PPCPred [4], PredPPCrys [2] and Crysalis [3], allow predicting propensities for completing selected steps involved in the protein crystallization process. In detail, they predict propensities for sequence cloning failure, protein material production failure, purification failure and crystallization failure, as well as successful crystal structure determination (CRYs). The design of the CrysComb predictor also includes the selection of an optimal subset from these inputs using the abovementioned FFS method. This selection was performed on the training data sets by using 5-fold cross-validation test.

Design of the Crysf predictor

We extracted and encoded UniProt-derived protein functional features to establish the Crysf predictor and its variants including:

1. Crysf, which uses functional features excluding the PE score to build SVR models using SP_train_nr and TR_train_nr data sets extracted from Swiss-Prot (SP) and TrEMBL (TR) data sets, respectively. More specifically, the PE score (1–5) indicates the type of experimental evidence supporting the existence of a given protein, with '1' representing strong existence (refer to the footnote in Supplementary Table S1).
2. Crysf_PE, which uses all functional features including the PE score to build SVR models on SP_train_nr and TR_train_nr data sets.
3. Crysf_Comb, which uses functional features (excluding the PE score) integrated with selected outputs from Crysalis (10 features) [3] to build SVR models on SP_train_nr and TR_train_nr data sets.
4. Crysf_S, which builds SVR models on the SP0.9 and TR0.9 data sets using the functional features (excluding the PE score) by averaging the feature values of target-protein homologs with >90% sequence identity.

We performed feature selection for each version of the Crysf predictor on the corresponding training data sets using 5-fold cross-validation test. Then, we used the selected features to build the SVR models on the training data sets. Finally, we applied these models to evaluate their predictive performances on the test data sets, and compared these versions of Crysf with the currently existing predictors of protein CP.

Table 1. Comprehensive comparison of bioinformatics tools for sequence-based prediction of protein CP

Name	Year published	Data sets (number of crystallizable/ non-crystallizable proteins)	Availability	Batch prediction	Predictive model	Input features used	Runtime per sequence	Format of input	Output	URL
OB-Score	2006 [29]	6182/6025	Source code (Linux)	Yes	Z-score matrix	pI, GRAVY	<1 s	FASTA sequence	Propensity for crystallization	http://www.compbio.dundee.ac.uk/obscore/
CRYSTALP2	2009 [33]	728/728	Webserver	Yes (<100 proteins)	NRBF	pI, GRAVY, AAC, PAAC	<1 s	FASTA sequence	Propensity for crystallization	http://biomine.ece.ualberta.ca/CRYSTALP2/
SVMCRYS	2010 [34]	728/728	Source code (Linux)	Yes	SVM	pI, GRAVY, AAC, TPC, PCP	<1 s	FASTA sequence	Propensity for crystallization	http://www3.ntu.edu.sg/home/EPNSugan/index_files/svmcrys.htm
PPCPred	2011 [4]	2408/4760	Webserver	Yes (<5 proteins)	SVM	pI, GRAVY, PCP, AAC, pSS, pDISO, pRSA	~2 min	FASTA sequence	Propensity for production, purification, crystallization and diffraction-quality crystallization	http://biomine.ece.ualberta.ca/PPCpred/
SCMCRYS	2013 [43]	1197/2378	Source code (Linux)	Yes	SCM	PAAC	<1 s	Special format	Propensity for crystallization	http://iclab.life.nctu.edu.tw/SCMCRYS/
XtalPred-RF	2014 [10]	2265/2355	Webserver	Yes (<10 proteins)	RF	pRSA, surface gravy, entropy and ruggedness	~1.5 min	FASTA sequence	Propensity for crystallization	http://ffas.burnham.org/XtalPred-cgi/xtal.pl
fDETECT	2014 [25]	2408/4760	–	Yes	SVM	AAC, PCP, instability index, complexity segments, etc	≪1 s	FASTA sequence	Propensity for crystallization	N/A
PredPPCrys	2014 [2]	5383/23 348	Webserver	No	SVM	pI, GRAVY, AAC, PAAC, TPC, PCP, pSS, pDISO, pRSA, PROFEAT	~3 min	FASTA sequence	Propensity for cloning, production, purification, crystallization and diffraction-quality crystallization	http://www.structbioinfor.org/PredPPCrys/
Crysalis	2016 [3]	5383/23 348	Webserver	Yes (up to 10 000 proteins)	SVR	AAC, PCP, KSAAP, GKSAAP	<1 s	FASTA sequence	Propensity for cloning, production, purification, crystallization and diffraction-quality crystallization	http://nmrcen.xmu.edu.cn/crysalis/
Crysf	This study	2798/3096	Webserver	Yes (up to 10 000 proteins)	SVR	Crysalis results, UniProt functional features	~5 s per query	UniProt ID	Propensity for crystallization	http://nmrcen.xmu.edu.cn/crysf/

NRBF, normalized radial basis function; SVM, support vector machine; SCM, scoring card method; RF, random forest; SVR, support vector regression; pI, isoelectric point; GRAVY, grand average of hydropathy; AAC, amino acid composition; PAAC, p-collocated amino acid pair composition; TPC, tripeptide composition; PCP, physicochemical properties based on AAindex1 indices; pSS, predicted secondary structures; pDISO, predicted intrinsic disorder; pRSA, predicted relative solvent accessibility; PROFEAT, features generated by PROFEAT web server; KSAAP, *k*-spaced amino acid pairs; GKSAAP, grading *k*-spaced amino acid pairs.

Performance comparison for the predictors of protein CP

We evaluated and compared the predictive performances of a comprehensive set of the currently existing predictors for protein CP. Eight predictors previously developed are showed in Table 1. These predictors are available to the end users via either a webserver or a standalone program, including OB-Score [29], CRYSTALP2 [33], SVMCRYS [34], PPCPred [4], SCMCRYS [43], XtalPred-RF [10], PredPPCrys [2] and Crysalis [3]. We compared the predictive performances of our newly developed functional annotation-based predictor Crysf and the CrysComb meta-predictor with other predictors. Additionally, we also reviewed the fDETECT tool [25] which provided a large-scale analysis of all known complete proteomes, even though it is not accessible for most of the end users.

Results and discussion

Review of the currently existing predictors for protein CP

We sorted chronologically and compared nine currently existing predictors and our new Crysf predictor in Table 1. We reviewed and discussed the architectures of these predictors including the used input features, used predictive models and the data sets used to train and test these models.

The nine predictors were developed between 2006 and 2016, several of which were developed in the past 3 years. The data sets used to develop these tools were derived from a number of relevant public databases, including TargetDB [20], PepcDB [19] and TargetTrack [20]. Table 1 indicates that the most popular predictive model is either support vector machine (SVM) or its variant, SVR. The two most popular predictive models were used by six of ten considered predictors. One of the important differences between these predictors is their prediction scopes. While the majority of these tools predict the protein CP, or more precisely the propensity for the diffraction-quality crystallization, three tools (PPCPred, PredPPCrys and Crysalis) also predict the propensity for successfully completing some of the crucial steps during the crystallization process. We herein discuss the prediction scopes and architectural details of these tools in a chronological order.

OB-Score [29] uses the Z-score scale to estimate protein CP, which measures the similarity of the grand average of hydrophobicity calculated by the Kyte–Doolittle scale (termed as GRAVY) [55], and that of isoelectric point (p*I*) computed from the sequence using Bioperl [56], to the GRAVY-p*I* distributions of previously annotated crystallizable and non-crystallizable proteins. Structural genomics data have demonstrated that both p*I* and GRAVY could be used as markers of protein CP [29, 57], which has motivated their uses in OB-score. The p*I* value can work as a marker for protein solubility, purification and crystallization, while GRAVY is an indicator of transmembrane proteins that are more difficult to be crystallized.

CRYSTALP2 [33] is an updated version of CRYSTALP [31], which eliminates the limitation of the earlier version applicable only for short protein chains (i.e. $<$200 AAs). It uses a kernel-based predictive model to predict the propensity of a query protein sequence to produce diffraction-quality crystals. CRYSTALP2 considers a large number of input features including the composition and collocation of AAs, p*I* and hydrophobicity. It uses a small subset of these features selected empirically to maximize predictive performance. This predictor was established based on a data set extracted from the TargetDB [20] and PepcDB [19] databases, which contains 728 crystallizable and 728 non-crystallizable proteins.

SVMCRYS was developed in 2010 [34], which was trained on the same data set as CRYSTALP2. This predictor uses sequence-derived information as the input of the SVM model, including physicochemical properties extracted from the AAindex database [58] and secondary structures predicted with PSIPRED [59].

PPCPred was developed in 2011 [4]. It is the first tool that not only predicts the CP, but also predicts the propensities for completing several selected steps involved in the crystallization process. These steps include protein production, purification, crystallization and finally the typically considered production of diffraction-quality crystals. This tool assesses the likelihood of a query protein to be crystallized, and also identifies potential bottlenecks in the crystallization process. PPCPred uses a large training data set to train the SVM models, which contains 2408 crystallizable and 4760 non-crystallizable proteins extracted from PepcDB [19]. Besides using the sequence-derived features, this predictor also uses structural features related to predicted secondary structures (pSS), intrinsically disordered regions and relative solvent accessibility. This is the first approach that exploited the new type of the inputs by integrating sequence-derived features with structural features.

SCMCRYS uses a scoring card-based predictive model optimized with a genetic algorithm, and input features derived directly from the sequence in the form of *p*-collocated amino acid pairs [43]. SCMCRYS trained and tested the SCM models on the same data sets as PPCPred.

XtalPred-RF [10] extends the original XtalPred predictor [45], which uses a random forest-based predictive model selected from those constructed by several machine-learning methods. The random forest model was trained and tested on the data sets extracted from PSI's TargetTrack database. Its test data set covers 2265 crystallizable and 2335 non-crystallizable proteins. The input features used in this predictor include the predicted protein surface ruggedness, hydrophobicity, side-chain entropy of surface residues and amino acid composition of the predicted protein surface [10]. The new feature, protein surface ruggedness, is defined as a ratio of absolute solvent accessibility to the total accessible area of a protein. This feature denotes the extent to which the surface of a protein is regarded as more 'rugged' or less 'rugged' than an average expected for a protein of a given size.

fDETECT was developed as a part of a large-scale study that assessed X-ray crystallography-derived structural coverage of 8.7 million non-redundant proteins from $\sim$2000 complete proteomes across all kingdoms of life [25]. fDETECT extends the architecture of PPCPred by removing computationally intensive features based on putative structural information and adding a large set of runtime-efficient features. fDETECT was designed to be computationally fast (0.8 ms per sequence) and can be used to process large sets of proteins.

Inspired by PPCPred, PredPPCrys was developed to predict the propensities for sequence cloning, protein material production, purification, crystallization and ultimately, production of diffraction-quality crystals [2]. It uses a comprehensive set of sequence-derived features, some of which were calculated with the PROFEAT webserver [60]. PredPPCrys considers 2924 features and the two-layer SVM architecture. It has been demonstrated that the two-layer design has an improved predictive performance over a single-layer design [2].

Crysalis is a CP predictor, which can be also used to computationally assess crystallizability of protein mutants [3]. This tool has been successfully used to enhance the expression level and purity of the ATP binding domain of the EspI protein from *Mycobacterium tuberculosis* [61]. More precisely, Crysalis can be used to (1) select crystallizable target proteins, (2) identify potential single-point mutations that might enhance protein CP and (3)

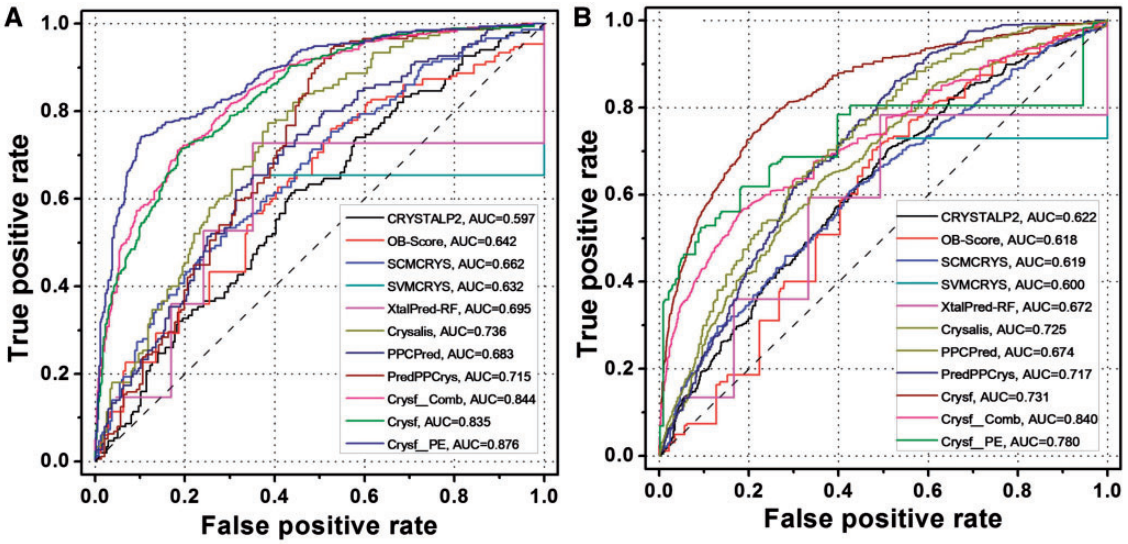

Figure 1. ROC curves of nine predictors of protein CP based on (**A**) the SP_test_nr data set; (**B**) the TR_test_nr data set. The AUC values of these predictors are shown. Note that the Crysf predictor has three versions: Crysf, Crysf_Comb and Crysf_PE.

annotate target proteins based on their predicted structural properties. The latter include putative transmembrane segments, functional domains and conserved residues, putative secondary structure, putative disordered regions and putative solvent accessibility. Similar to PredPPCrys, this tool is based on a two-layer SVM model using sequence-derived features as its inputs.

Besides these predictors described above, several other predictors can be also applied to predict protein CP. However, these predictors are not currently accompanied by either a standalone program or a webserver, which makes them less accessible to the end users. They include SECRECT (2006) [30], CRYSTALP (2007) [31], MCSG-Zscore (2010) [42], CRYSpred (2012) [48] and RFCRYS (2012) [47]. Additionally, the SERp tool (available at http://services.mbi.ucla.edu/SER/) was developed to assess protein crystallizability based on the concept of surface-entropy reduction [62, 63].

Evaluation of the performances of selected predictors for protein CP

We evaluated predictive performances of the nine tools based on two benchmark data sets. Figure 1 illustrates ROC curves of the eight currently existing predictors and the new Crysf predictor on the SP_test_nr and TR_test_nr data sets. The predicted CP indicates a likelihood of obtaining diffraction-quality crystals for a query protein. Expectedly, predictors with larger AUC scores would provide higher performances for predicting protein CP.

We also evaluated the performances of these predictors for a binary classification (crystallizable proteins versus non-crystallizable proteins) based on several popular measures: Matthew's correlation coefficient (MCC), accuracy, specificity, sensitivity and precision. As shown in Tables 2 and 3, the AUC values of the eight currently existing tools range between 0.597 and 0.736 on SP_test_nr, and between 0.600 and 0.725 on TR_test_nr. The top-ranking predictors were Crysalis, PredPPCrys, XtalPred-RF and PPCPred. Similarly, the arguably most representative single measure of the binary prediction, MCC, ranges between 0.172 and 0.381 on SP_test_nr, and between 0.134 and 0.381 on TR_test_nr. The top three predictors were Crysalis, PredPPCrys and PPCPred, which provided relatively accurate prediction results with MCC $\geq$ 0.3. Overall, these data indicated that the majority of these predictors offered relatively good predictive performances.

Table 2. Performance comparison of the available predictors of protein CP evaluated on the SP_test_nr data set

Method	AUC	MCC	ACC (%)	SPEC (%)	SENS (%)	PRE (%)
OB-Score	0.642	0.214	60.6	59.3	62.7	49.5
CRYSTALP2	0.597	0.172	59.3	56.4	61.3	47.2
SVMCRYS	0.632	0.257	62.7	61.0	65.3	51.6
PPCPred	0.683	0.301	66.1	68.6	62.0	55.7
SCMCRYS	0.662	0.220	61.4	61.9	60.7	50.3
XtalPred-RF	0.695	0.216	64.8	83.1	36.0	57.4
PredPPCrys	0.715	0.381	69.7	55.3	**81.3**	69.4
Crysalis	0.736	0.371	67.7	62.7	75.3	56.2
Crysf	0.835	0.512	75.4	79.4	71.8	79.9
Crysf_Comb	0.844	0.516	75.6	79.4	72.3	80.1
Crysf_PE	**0.876**	**0.650**	**82.0**	**89.2**	75.7	**88.9**

Values in bold highlight the best predictive performance for the corresponding measure. MCC: Matthew's correlation coefficient; ACC: accuracy; PRE: precision; SEN: sensitivity. SPE: specificity.

Table 3. Performance comparison of the available predictors of protein CP evaluated on the TR_test_nr data set

Method	AUC	MCC	ACC (%)	SPEC (%)	SENS (%)	PRE (%)
OB-Score	0.618	0.223	59.5	52.4	70.2	49.1
CRYSTALP2	0.622	0.179	58.4	55.5	62.9	48.0
SVMCRYS	0.600	0.201	57.3	47.2	**72.9**	47.4
PPCPred	0.674	0.267	63.5	63.1	64.2	53.1
SCMCRYS	0.619	0.180	58.8	57.5	60.9	48.3
XtalPred-RF	0.672	0.134	62.4	94.3	13.4	60.6
PredPPCrys	0.717	0.312	66.0	65.7	66.3	55.1
Crysalis	0.725	0.320	66.2	65.4	67.3	55.9
Crysf	0.731	0.318	67.9	72.7	59.5	55.6
Crysf_Comb	**0.840**	**0.490**	75.5	76.2	74.1	64.2
Crysf_PE	0.780	0.478	**76.4**	**95.4**	43.5	**84.3**

Values in bold highlight the best predictive performance for the corresponding measure. MCC: Matthew's correlation coefficient; ACC: accuracy; PRE: precision; SEN: sensitivity. SPE: specificity.

Given that these predictors are organized chronologically in Tables 2 and 3, we observed a strong trend that the newer tools provide improved predictive performances compared with the older tools. The corresponding Pearson correlation coefficients between the year of publication and the AUC values obtained on TR_test_nr and SP_test_nr are 0.75 and 0.81, respectively. Note that, these tools were evaluated and compared on both TR_train_nr and SP_train_nr, and the corresponding results are shown in Supplementary Tables S2 and S3. The primary difference between these data sets is the quality of functional annotations, which is relevant to the assessment of the Crysf predictor as it uses these annotations as the inputs of the SVR model.

Ease of use and functionality of these predictors

The functionality and ease of use of the nine predictors were discussed and compared, focusing on their utilities as bioinformatics tools that can be readily used by non-bioinformaticians. More specifically, we considered the following aspects: (1) availability and usefulness of the webserver; (2) availability and usefulness of the standalone software; (3) ability of the tool to support batch predictions consisting of multiple query sequences; (4) runtime; and (5) the format of the prediction output. The tools and their major relevant characteristics are summarized in Table 1.

Most of the nine predictors are available as webservers. They require only an Internet connection and a web browser. The computations are performed on the server side. The end user simply needs to go to the corresponding URL (Table 1), copy and paste the FASTA-formatted sequence(s) and click the 'start' button. After the job is processed, the prediction output will be returned either to the web page or to the user-provided email address. This means that these tools are accessible even for less computer savvy users. A few tools, including OB-Score, SVMCRYS and SCMCRYS are provided in a form of source code. These tools allow users to execute batch computations on Linux operating systems. However, non-bioinformaticians might find it difficult to install and run these predictors on their local computers. For example, SVMCRYS requires users to install the SVM software on the top of the source code. It also only accepts the input protein sequence in a specific format.

Similarly to the tools offered as standalone software, most of the webservers provide the ability to perform predictions on multiple query protein sequences at the same time. This makes it convenient to compare the predicted CPs for a set of protein sequences. Both CRYSTALP2 and PPCPred can support batch prediction for a maximum of 100 and 5 query protein sequences per run in FASTA format, respectively. The XtalPred-RF's webserver allows submission of up to 10 protein sequences, while Crysalis and Crysf enable large-scale proteome-wide prediction analysis by allowing users to run a batch of up to 10 000 protein sequences. The PredPPCrys's webserver does not support batch prediction.

The runtime required for processing a query sequence is another major factor, particularly when these predictors are used for selecting targets from a large number of protein candidates. As described above, features used as the inputs by these tools can be grouped into three major categories: sequence-based features, structural features and functional features (Table 1). The structural features are composed of the pSS, disordered regions and relative solvent accessibility. Calculation of these structural features requires a longer runtime (in minutes) as compared with that of the other categories of features (in seconds). The prediction on a typical query sequence takes 1–3 min for PPCPred, XtalPred-RF and PredPPCrys, and takes $<$1 s for the other predictors. Both Crysf and fDETECT are among the fastest predictors with a runtime of milliseconds for completing a prediction on a single query sequence.

Overall, three important factors including the availability of webservers, ability to process batch predictions of multiple protein sequences and the low runtime, make these predictors friendly for non-bioinformaticians. Consequently, these tools enjoy a relatively heavy workload. For example, as a webserver released in 2011, PPCPred has been used by $>$4600 unique users from 73 countries (source: Google Analytics as of Sept 2016).

We also compared the prediction outputs of these predictors. OB-Score [29] requires a user to run the prediction in the command-line, and returns its prediction output in a tab-delimited format. The output includes the protein ID, OB-Score, GRAVY, pI, mol_wt (molecular weight) and sequence length. The value of OB-Score ranges from −5.3 to 9.1. CRYSTALP2 [31] classifies a query protein as either crystallizable or not-crystallizable, and assigns a confidence score ranging from 0 to 1, e.g. 'Protein T0404 is non-crystallizable with 0.403 confidence'.

SVMCRYS is available as a standalone software [43], which requires the users to provide the training data set for constructing the predictive model. This predictor produces a binary output, i.e. 'resistant-to-crystallization' or 'amenable-to-crystallization' without a numeric propensity score.

Once a user runs a prediction, SCMCRYS produces three output files: 'testfile.results', 'testfile_ori.csv' and 'testfile_result.csv'. The 'testfile.results' file stores the predicted propensity score varying between around 300 to 600. The 'testfile_ori.csv' and 'testfile_result.csv' files store the original score from the predictive model and summarize predictive performances on user-input training and test data sets, respectively.

PPCPred [4] maps a query protein into one of several categories: 'fail to produce protein material', 'fail to purify', 'fail to crystallize' and 'yield diffraction-quality crystals', along with the predicted CP ranging from 0 to 1. The output results also include predicted propensities for several selected steps involved in the crystallization process, including production of protein material fails, purification fails, crystallization fails and yielding diffraction-quality crystals. These scores are used to anticipate which steps of the crystallization process are more likely to fail for the proteins that cannot be crystallized. The prediction results from PPCPred can be downloaded as a CSV file.

XtalPred-RF [10] identifies a query protein as one of several categories that describe the likelihood of protein crystallization. Its output encompasses protein sequence length, gravy index, instability index, pI, coiled coils, percentage of coil structure, transmembrane helices, signal peptide and insertion score. These predicted physicochemical and structural characteristics are of benefit to follow-up experimental design and target selection. In addition, XtalPred-RF performs the BLAST [44] search for homologs of the query protein in both NCBI NR database (a non-redundant database of NCBI clustered at 60% sequence-identity level) [64] and PDB [1].

PredPPCrys [2] offers the propensity scores for five major steps in the crystallization process. Similarly to PPCPred, PredPPCrys identifies a query protein as either crystallizable protein or non-crystallizable protein, and then predicts the propensities for cloning, production, purification, crystallization and structure determination.

The Crysalis webserver offers two modes for the services, including 'Prediction mode' and 'Design mode' [3]. The end users can submit jobs under the 'Prediction mode' to predict protein CPs. The 'Design mode' is used for computational design and analysis of non-crystallizable proteins predicted with low CPs.

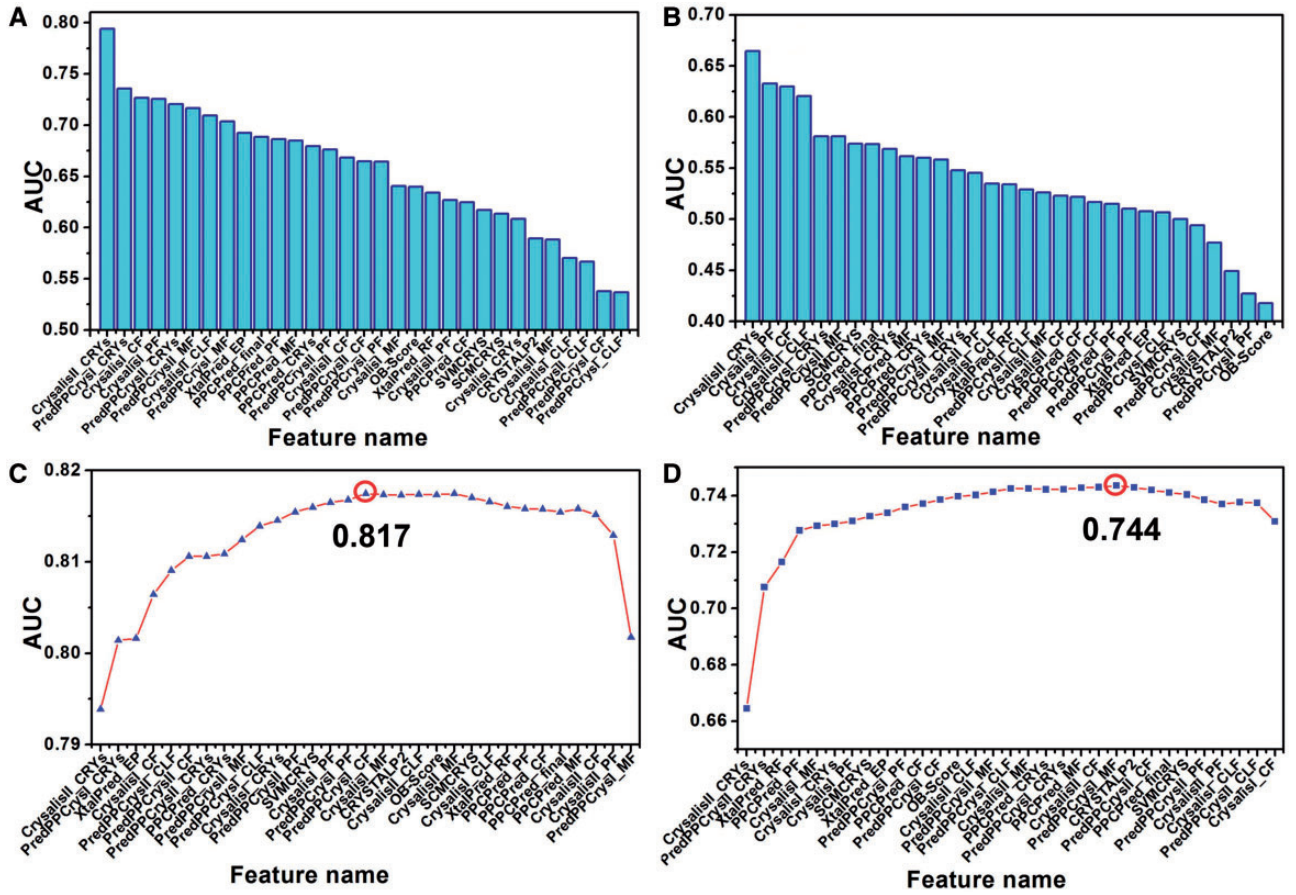

Figure 2. Relative importance and contributions of 31 individual features (i.e. the outputs from eight predictors) selected for building the integrated CrysComb model on SP_train_nr and TR_train_nr data sets. (**A** and **B**) AUC values of the SVR models built with each of the individual features on (**A**) SP_train_nr and (**B**) TR_train_nr. (**C** and **D**) Cumulative AUC values of the SVR models selected during the feature selection process based on (**C**) SP_train_nr and (**D**) TR_train_nr.

Crysalis produces three types of outputs: 'prediction results' (equivalent to 'Prediction mode'), 'computational design' and 'protein structural and functional annotation'. The computational design results suggest some residues to undergo single mutation analysis, and illustrate them on a hotspot map, and also provide a list of top 20 ranked single mutations potentially capable of enhancing the likelihood of successful crystallization. This webserver also provides predicted structural characteristics, including secondary structures, intrinsically disordered regions, solvent accessibility and functional domains as well as conserved sites. These results can be downloaded from the Crysalis webserver for the further analysis.

Crysf exploits the UniProt-derived functional annotations to predict the protein CP. While this tool could offer more accurate prediction results, as shown in Tables 2 and 3, it can only be used to assess CPs for the proteins available in UniProt. Thus, Crysf would be more suitable to guide structural biologists to select feasible proteins from UniProt, but less appropriate for guiding the rational design of truncated constructs. Users can submit either UniProt IDs or FASTA-formatted sequences to the Crysf webserver for the prediction. This webserver has pre-calculated CPs for the proteins available in UniProt, which is updated periodically to reflect new contents of the UniProt resource. Note that the predictive performance of the Crysf predictor is dependent on the quality and reliability of the functional annotations derived from the database. As expected, Crysf could provide prediction results more accurate for the Swiss-Prot entries but relatively less accurate for the TrEMBL entries.

The improved performance of the CrysComb meta-predictor

We first calculated the AUC values of the SVR models built with each of 31 outputs from the eight predictors, and found that all these outputs were contributory for the prediction of protein CP on both the SP_train_nr and TR_train_nr data sets (Figure 2A and B). We also found that the CrysallisII_CRYs model achieved the highest AUC value in the 31 SVR models (Figure 2A and B) on either SP_train_nr or TR_train_nr (Figure 2A and B).

Thus, we used the 31 outputs as candidate input features to optimize the CrysComb models. On SP_train_nr, we selected 16 features as the inputs to train the CrysComb model, obtaining a AUC value (0.817). The 16 features were derived from four predictors including Crysalis, PredPPCrys, SVMCRYS and PPCPred (Figure 2C). Note that the AUC value of the CrysComb model was higher than those of the 16 SVR models built with each of the 16 features as an individual input (Figure 2C), including that of the best-performing individual model CrysallisII_CRYs (AUC: 0.794; Figure 2A; Supplementary Table S4).

Similarly, we selected 22 features as the inputs to train the CrysComb models on TR_train_nr, obtaining a high AUC value (0.744). The 22 features were derived from six predictors including Crysalis, PredPPCrys, XtalPred_RF, PPCPred, SCMCRYS and OB-Score predictors (Figure 2D). The CrysComb model achieved a significantly increased AUC value compared with the 22 SVR models built with individual features (Figure 2D). Note that the best-performing individual model CrysallisII_CRYs only obtained an AUC value of 0.665 (Figure 2B; Supplementary Table S4).

The Crysf predictor

Four versions of the Crysf predictor

The Crysf predictor has four versions: Crysf, uses functional annotations excluding PE scores; Crysf_PE, uses all functional annotations; Crysf_Comb, uses functional annotations excluding PE scores, which are integrated with the outputs from Crysalis; Crysf_S, uses the functional features of target proteins derived from their homologs. For each version of the Crysf predictor, the feature subsets were constructed by performing feature selection experiments on SP_train_nr and TR_train_nr using 5-fold cross-validation test. The curves of AUC against the number of FFS-selected features (FFS order) are showed in Supplementary Figure S1. The highest AUC values associated with the feature subsets used in Crysf_Comb (Supplementary Figure S1B and D) were larger than those used in Crysf (Supplementary Figure S1A and C), i.e. 0.852 versus 0.838 based on SP_train_nr, 0.817 versus 0.735 based on TR_train_nr.

Performance comparison among Crysf and other predictors

Crysf, Crysf_PE and Crysf_Comb were assessed and compared with other current predictors (Tables 2 and 3). Crysf outperformed other currently existing predictors on SP_test_nr, while it showed the similar performance as Crysalis and PredPPCrys on TR_test_nr. This was owing to this reason that SP_test_nr was constructed based on higher quality functional annotations from Swiss-Prot, while TR_test_nr based on lower quality functional annotations from TrEMBL.

By exploiting the PE score as an extra functional feature, Crysf_PE could obtain improved predictive performances on both SP_test_nr and TR_test_nr. However, as a portion of PE = 1 entries is associated with available crystal structures for given proteins, Crysf_PE is potentially related to the bias of using *a priori* information regarding crystallizability as input features. Thus, Crysf_PE might be not suitable to the prediction of CPs for those proteins.

By integrating the functional features (without the PE score) with the outputs from Crysalis, Crysf_Comb achieved better predictive performances on two test data sets SP_test_nr and TR_test_nr (Tables 2 and 3). Notably, the AUC values of Crysf_Comb were 0.844 on SP_test_nr and 0.840 on TR_test_nr, while those of Crysalis were 0.736 on SP_test_nr and 0.725 on TR_test_nr, respectively. The MCC values were increased from 0.371 (Crysalis) to 0.516 (Crysf_Comb) on SP_test_nr, and enhanced from 0.320 (Crysalis) to 0.490 (Crysf_Comb) on TR_test_nr. These results demonstrated that Crysf_Comb had improved predictive performances compared with Crysalis. Furthermore, Crysf_Comb also displayed AUC values higher than Crysf, i.e. 0.844 versus 0.835 on SP_test_nr, and 0.840 versus 0.731 on TR_test_nr. Similarly, Crysf_Comb showed enhanced MCC values compared with Crysf on both SP_test_nr (0.516 versus 0.512) and TR_test_nr (0.490 versus 0.318). Similar results were also obtained on two training data sets SP_train_nr and TR_train_nr (Supplementary Tables S2 and S3).

Predictive performances of Crysf_S were assessed on both the SP0.9 and TR0.9 data sets, which used the functional features to describe target-protein homologs with > 90% sequence identities. Compared with Crysalis, Crysf_S achieved higher AUC values, in detail, 0.755 versus 0.673 on SP0.9, 0.742 versus 0.672 on TR0.9 (Supplementary Table S6).

These results demonstrated that: (1) UniProt-derived functional features could be used to accurately predict protein CP; (2) higher predictive performance could be achieved by using higher quality functional features extracted from the Swiss-Prot database; (3) an integrated predictor using both sequence-based features and functional features could obtain improved predictive performance compared with the predictors using solely either sequence-based features or functional features; (4) Crysf and its variant Crysf_Comb provide higher predictive performances than other predictors. Note that different versions of Crysf are limited to the proteins with available functional annotations in either Swiss-Prot or TrEMBL.

Quantitative analysis of functional features and obstacles related to protein crystallization

The UniProt-derived functional features used in Crysf are categorized into six major types: protein polymerization, UniProt feature (FT), posttranslational modification (PTM), database resources (DR), protein subcellular location and others including reference number and PE. Supplementary Table S7 shows a list of functional features selected for Crysf from SP_train_nr and TR_train_nr data sets.

DR is the largest type of the selected functional features (18 selected features as shown in Supplementary Table S8). DR is used to describe the presence of a given protein in protein-annotation DR, such as EMBL, Gene3D, KEGG and ProteinModelPortal. FT is the second largest type of functional features (11 selected features). Protein polymerization is considered as a major obstacle in the crystallization process, and might result in protein aggregation or lead to various forms of microheterogeneity [65, 66]. This explains why Crysf includes the functional features describing the polymerization of a given protein. Furthermore, protein subcellular localization is also an important factor for crystallization. Nine and six protein subcellular localization-related features are included in Crysf based on SP_train_nr and TR_train_nr, respectively. In addition, protein crystallization trials have indicated that PTMs could influence the success rate of crystallization [23, 66–69]. Thus, Crysf also includes PTM-related features, which quantify whether a given query protein undergoes PTM such as glycosylation and acylation.

We calculated statistical significances of differences in all functional feature values between crystallizable and non-crystallizable proteins on SP_train_nr and TR_train_nr (Supplementary Table S9). These results indicate whether and which functional features were useful to discriminate between crystallizable and non-crystallizable proteins. We also rank predictive performances (AUC values) and statistical significances (*P*-values) of individual functional features (Figure 3). The selected functional features are labeled with red asterisks. Furthermore, we analyzed five selected functional features with the lowest *P*-values (Supplementary Figure S2), including FT_TRANSMEM, DR_SMR, various polymerization state, homo/hetero-polymerization state and dimer.

The first discriminative feature (Supplementary Table S9), FT_TRANSMEM, describes the number of membrane-spanning regions. As expected, Supplementary Figure S2A demonstrates that the more membrane-spanning regions a full-length protein has, the more difficult it is to crystallize this protein. The second discriminative feature, DR_SMR, denotes the number of annotation records in the SWISS-MODEL repository, which quantifies the number of 3D structural models generated by automated homology modeling for a given protein [70]. Supplementary Figure S2B illustrates that proteins with more homology-based models tend to be easier to be crystallized compared with those without or with fewer homology-based models. This observation is in good agreement with the previous results [26].

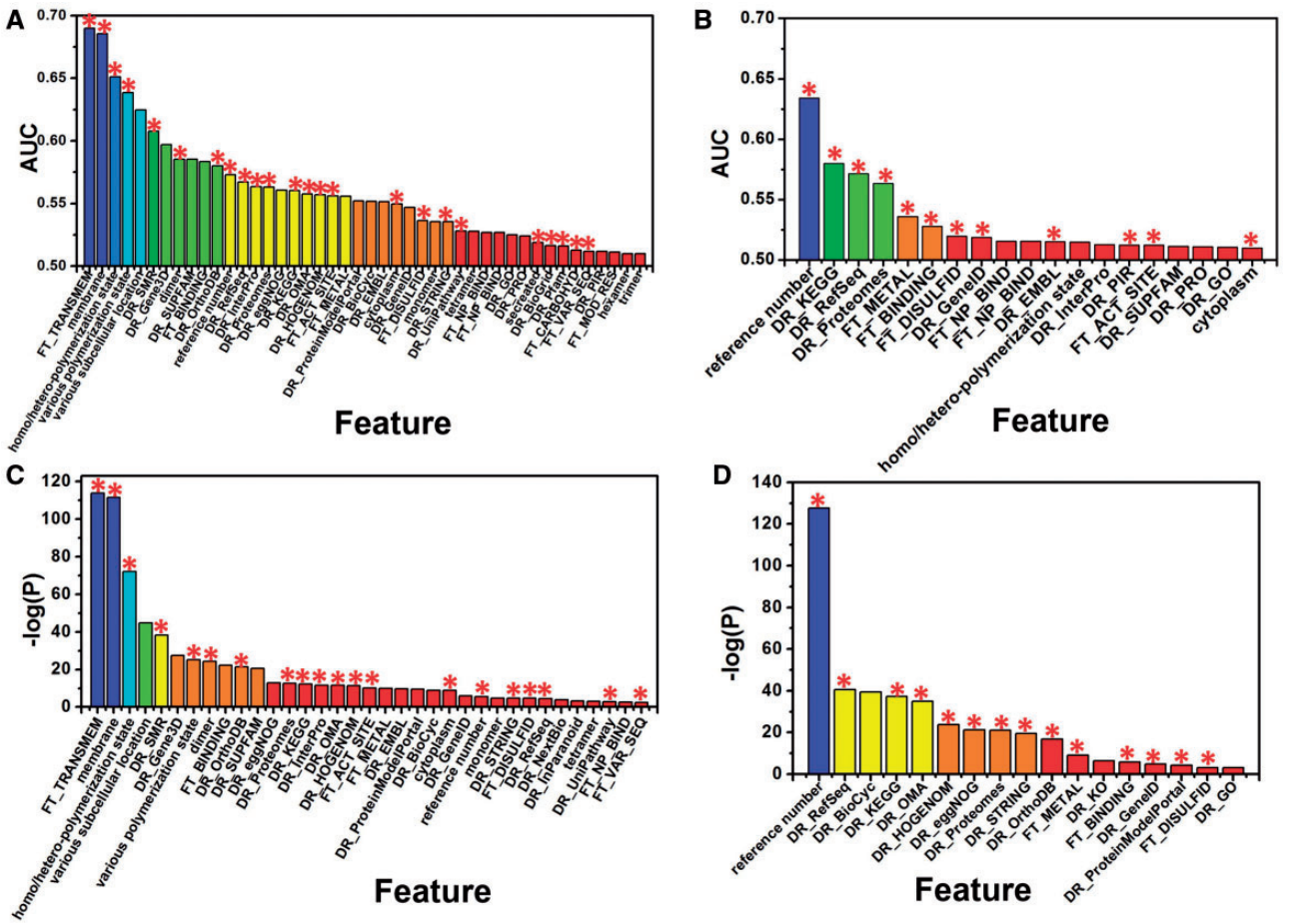

Figure 3. Statistical significances and predictive performances of individual features computed from UniProt-derived functional annotations based on SP_train_nr and TR_train_nr. (**A** and **B**) Features are ranked according to their AUC values obtained using 5-fold cross-validation test. The features with AUC > 0.51 are shown. (C and D) Features are ranked according to their *P*-values calculated by the Wilcoxon rank-sum test. The features with *P*-value < 0.01 are shown. Features selected for building the Crysf model are labeled with red asterisks. A detailed description of these functional features can be found in Supplementary Table S1.

The last three discriminative features depict protein polymerization, including various polymerization state (single state, multiple states, no-statement), homo/hetero-polymerization state (homo-polymer, hetero-polymer, no-statement) and dimer (yes, no). Our analysis shows that proteins with various polymerization states and hetero-polymeric proteins are more difficult to be crystallized (Supplementary Figure S2C and D). In addition, we also demonstrate that homo-polymeric proteins and dimeric proteins tend to be more crystallizable (Supplementary Figure S2D and E).

Large-scale prediction of CP for proteins from UniProt and PDB

We applied Crysf_Comb, the most accurate variant of the Crysf predictor, to conduct a large-scale prediction of CP for >50 million proteins derived from UniProt, including Swiss-Prot and TrEMBL. The prediction results can be downloaded from http://nmrcen.xmu.edu.cn/crysf/data/reviewed.results (549 008 proteins derived from Swiss-Prot, 44.9 MB) and http://nmrcen.xmu.edu.cn/crysf/data/tremb.results (50 011 027 proteins derived from TrEMBL, 4.0 GB). We calculated the distribution of protein CP scores ranging from −1 to 1 (Figure 4) with a threshold at zero. The CP scores for Swiss-Prot-derived proteins are skewed to higher values (skewness = 0.458) compared with those for TrEMB-derived proteins. This might result from the evidence that TrEMBL contains more proteins with incomplete functional annotations than Swiss-Prot [11]. However, Swiss-Prot-derived proteins show an almost similar fraction of crystallizable proteins as TrEMB-derived proteins (27% versus 29%).

In a recent study, the fDETECT tool predicted that about 25% of modeling families of proteins were crystallizable, which were derived from ∼2000 fully sequenced genomes in UniProt (8 652 940 non-redundant proteins) [25]. The protein families mean the clusters of proteins sharing high sequence similarity, for which structural models can be obtained through homology modeling. This proportion of crystallizable proteins predicted by fDETECT is similar to that predicted by Crysf_Comb.

It should be noted that a portion of proteins with available crystal structures in PDB were predicted as 'non-crystallizable' proteins. We analyzed the CP scores of 12 162 non-redundant proteins shared by Swiss-Prot and PDB databases. We further classified these proteins in PDB into five groups according to their structural resolutions: A, <1.5 Å; B, 1.5–2.0 Å; C, 2.0–2.5 Å; D, 2.5–3.0 Å; E, >3.0 Å. As shown in Supplementary Figure S3, these classified proteins in PDB have high likelihoods of being successfully crystallized and structurally solved compared with those unclassified proteins in Swiss-Prot. The median CP scores of the A, B, C, D and E groups are 0.284, 0.182, 0.121, 0.082 and −0.087, respectively. In contrast, the median CP score of all proteins in Swiss-Prot is −0.204, which is distinctly lower than those of the five groups. The average CP scores are 0.265, 0.206,

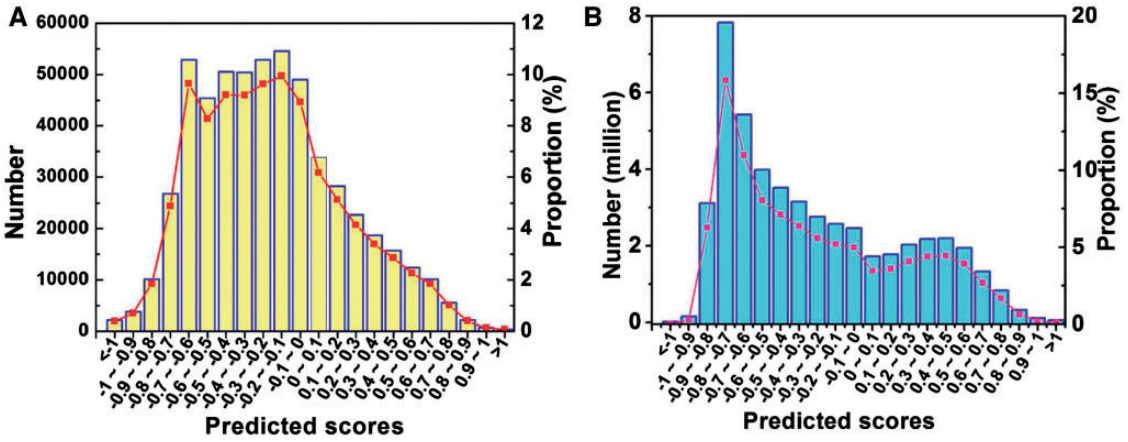

Figure 4. Distributions of CP scores calculated by Crysf_Comb for proteins in (**A**) Swiss-Prot and (**B**) TrEMBL. The red squares and simulated lines indicate the fraction of the proteins with given putative CPs to all proteins in Swiss-Prot and TrEMBL.

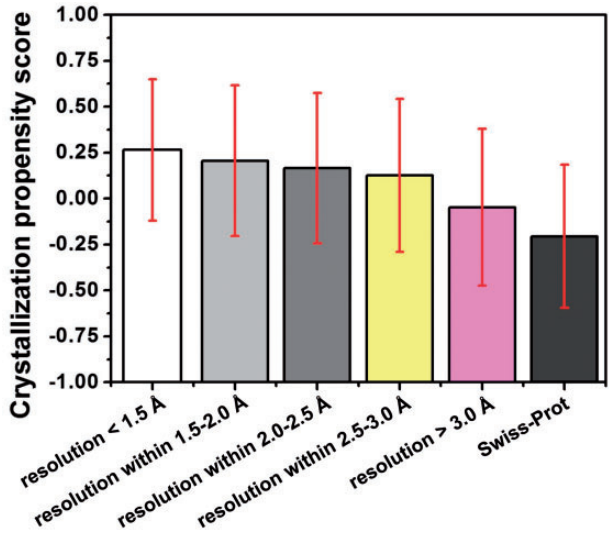

Figure 5. Distributions of CP scores calculated by Crysf_Comb for the proteins shared by Swiss-Prot and PDB databases. The proteins in PDB were classified into five groups according to their structural resolutions.

0.167, 0.126 and −0.048 for the A, B, C, D and E groups, respectively (Figure 5). However, the average CP score is −0.205 for all proteins in Swiss-Prot. Our large-scale analysis suggests a potential relation between the resolution of a PDB entry and its putative CP score. Using the CP score of −0.205 as a threshold, we calculated the proportions of the five groups of proteins predicted as crystallizable proteins: A, 88.3%; B, 83.4%; C, 81.2%; D, 78.2%; E, 63.5%. However, only 46.8% of the proteins in Swiss-Prot were predicted as crystallizable proteins.

On the other hand, other predictors also generated false negatives for the proteins in PDB. Even if these proteins have been structurally solved, these predictors still predicted them as 'non-crystallizable' proteins. There are several potential reasons for the false negatives. First, while a given protein is wrongly predicted as 'non-crystallizable' protein, it might have still a relatively high CP score. Note that the binary prediction results are dependent on a threshold of the CP score (e.g. a zero threshold is used in Crysf_Comb). Proteins with CP scores > the threshold will be predicted as crystallizable proteins. Expectedly, lowering the threshold will increase the number of putative crystallizable proteins, and correspondingly decrease the false negatives. Second, the predictors such as Crysf were developed mostly based on high-throughput protein crystallization data from Structural Genomics Centers. However, a portion of crystal structures deposited in PDB was determined traditionally by structural biologists using specialized equipment, unique protocols and laborious trial-and-error efforts. Potentially, these structures could not be determined by using high-throughput crystallization-to-structure pipelines. Thus, the prediction of CP primarily based on the high-throughput protein crystallization data would potentially result in false negatives. In addition, the quality and scope of the training data set would limit predictive performances of Crysf and other predictors.

The Crysf webserver

We provide an online webserver (http://nmrcen.xmu.edu.cn/crysf/) for applying the Crysf predictor to predict CPs based on user-submitted protein sequences of interest. The query protein sequences are specified in either the UniProt-ID or the UniProt-FASTA format. The Crysf webserver provides examples of properly formatted inputs. To facilitate target selection for a large set of candidate proteins, this webserver also supports batch prediction on multiple protein sequences (up to 10 000) in a single request. It allows two types of functional annotations as the inputs for the predictive model: reviewed annotations derived from Swiss-Prot; unreviewed annotations derived from TrEMBL. Different runtime is required for the prediction on one query protein sequence using different types of functional annotations, i.e. 5 s and 2 min for annotations derived from Swiss-Prot and TrEMBL, respectively. Furthermore, this webserver also allows submission of mixed target queries using functional annotations derived from both Swiss-Prot and TrEMBL. The runtime is about 2 min per query. In this case, the Crysf webserver will select a suitable predictive model to conduct the prediction based on the entries in UniProt associated with the query protein sequences.

Limitations and future directions

To date, nearly 20 sequence-based bioinformatics tools have been developed for predicting protein CP. Eight currently exiting predictors are accessible to the users via either the standalone software or the webservers. While some of these predictors offer relatively accurate prediction results with ease of use, they also have certain drawbacks. These predictors basically take

protein sequences as the inputs, without consideration of some other factors which would potentially influence protein CP. We herein list these factors as follows: the used experimental protocols; lack of resources that could lead to abandoning certain proteins; human elements such as potentially lacking consistency and oversight; dynamic nature of annotations related to CP; quality of data from public repositories including TargetTrack and PDB, which are used to build and validate predictive models. These factors are difficult to be appropriately considered in predictive models, which might bring down the upper limit of the predictive performance. Some of these factors are even virtually impossible to be considered. For instance, more resources are available for certain proteins of biological significance or practical importance, thus increasing their chances to be structurally solved, while other more routine proteins might be scraped with lesser amount of efforts.

In particular, it is necessary to consider the differences among experimental protocols used by different structural biology laboratories. These differences are associated with the development of laboratory-specific protein production protocols and crystallization screens, including use of specific vectors, expression systems and inclusion of metal ions, detergents and molecular chaperones [16]. Ability to consider these factors will ultimately depend on the availability of well-annotated data. While these data are available in TargetTrack, they are currently relatively sparse (incomplete). Nevertheless, it could be expected that we would obtain more high-quality data with sufficient quantities for building predictive models. We thus anticipate the release of a next generation of CP predictors, which would incorporate these intrinsic factors with the predictive models to provide more accurate predictions and to answer 'what-if' queries. These queries would simulate different scenarios for the same protein sequence with the assumption of using different experimental protocols. It could be expected that the next generation of CP predictors will predict which of these setups are the most suitable to maximize CPs.

Another relevant factor is the quality of the data set. First, some of inaccurate and unreliable data from the source databases were used to build and validate predictive models [71, 72]. Moreover, as the technology progresses over time, some proteins previously considered as non-crystallizable proteins, would be identified as crystallizable proteins. Functional features of these proteins would be updated in predictive models. A recent research has demonstrated that, to some extent the predictive performances of protein CP predictors degrade over time [4]. Thus, we strongly suggest that the developers of these bioinformatics tools periodically re-optimize the predictive models using up-to-date data sets.

On the other hand, we could develop new and improved predictors through exploiting new types of the inputs for predictive models. Note that the currently exiting tools rely primarily on both the features derived directly from the protein sequence and certain structural features predicted from the sequence. Here we present a new predictor Crysf, which explores functional features extracted from the UniProt resource as a new type of inputs. We have demonstrated that the Crysf_Comb predictor, which integrates functional features used in Crysf with the prediction outputs from Crysalis, could provide higher predictive performance than Crysf and Crysalis individually. To further improve the prediction performance, we would exploit new types of information related to the functional, structural or taxonomic features deduced from query protein sequences. Given that protein CP varies to a large extent in the context of taxonomic classification [25], we thus suggest that the taxonomic features could be integrated into the inputs extensively used in the currently exiting predictors, which would significantly enhance predictive performances. Furthermore, the predictors of protein CP could be extended to perform design of constructs that are more viable for structural determination. Several attempts to computationally score mutations have been made to potentially enhance the protein CP [3, 63].

Conclusions

We contribute a comprehensive review and assessment of bioinformatics tools for predicting protein CP. This comparative review provides useful guidance for both bioinformaticians and non-technical end users of these tools. We discuss new research directions that would spur development of more accurate predictors. These include importance of inclusion of details concerning experimental protocols into the predictive models and use of high-quality and recent data to train and validate predictive models.

We also furnish useful hints to assist structural biologists in selecting appropriate predictors suitable for their research needs. We discuss the availability of nine predictors, and detail their inputs, outputs and architectures. Our results indicate that different predictors have different strengthens and weakness, in particular by trading off predictive performance for runtime. We find that Crysf, Crysalis, CRYSTALP2, fDETECT, OB-Score and SVMCRYS are the most suitable tools for rapid (runtime-efficient) target selection for large candidate proteins, with fDETECT being the fastest tool. On the other hand, PPCPred, XtalPred-RF and PredPPCrys are slower owing to the use of putative structural features. Notably, four predictors including Crysalis, PPCPred, PredPPCrys and XtalPred-RF provide more accurate prediction results. Moreover, several recently published predictors of CP, in particular PPCPred, PredPPCrys and Crysalis, can also predict the propensities to complete a set of selected steps involved in the crystallization process. These steps include cloning, material production and purification. In addition, we demonstrate that the CrysComb meta-predictor, which integrates the selected outputs from eight predictors as candidate input features to build the predictive model, provides significantly higher predictive performance than the individual predictors. We also contribute a new predictor of CP, Crysf, which uses a novel type of inputs that rely on functional annotations of proteins extracted from UniProt. We show that Crysf offers better predictive performance with high runtime-efficiency. However, this predictor can be used only for the proteins with functional annotations available in UniProt, in contrast to other predictors that use just a readily available protein sequence as input. A freely available webserver that implements the Crysf predictor is available at http://nmrcen.xmu.edu.cn/crysf/, and the webpage provides details about this webserver.

Key Points

- This article provides a useful guide to facilitate selection of protein crystallization prediction tools.
- We review the availability, ability for batch predictions, details of the predictive models, runtime and applications of nine bioinformatics tools for predicting protein crystallization propensity.
- We herein introduce Crysf, a new tool using functional annotation-derived inputs. Integration of Crysf with the currently existing tool Crysalis provides better predictive performances.

- Comparison of predictive performances of the crystallization propensity predictors reveals that Crysf, Crysalis, CRYSTALP2 and OB-Score are best suited for fast proteome-wide target selection, whereas PPCPred, XtalPred-RF and PredPPCrys are suitable for refining the target selection and prioritization after the first round of target selection.
- Independent tests on two up-to-date test data sets indicate that the predictive performances can be further improved by integrating the outputs from multiple predictors as the inputs for the predictive model.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

The National Key Research and Development Program of China (2016YFA0500600), the National Natural Science Foundation of China (31670741, 61202167, 61303169, and 81661138005), National Health and Medical Research Council of Australia (NHMRC) (490989), National Institutes of Health (AI111965) and the Hundred Talents Program of the Chinese Academy of Sciences (CAS). J.S. is an NHMRC Peter Doherty Fellow and recipient of the Hundred Talents Program of CAS. L.K. was supported in part by the Qimonda Endowed Chair position at the Virginia Commonwealth University.

References

1. Rose PW, Bi C, Bluhm WF, *et al*. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res* 2013;**41**:D475–82.
2. Wang H, Wang M, Tan H, *et al*. PredPPCrys: accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. *PloS One* 2014;**9**:e105902.
3. Wang H, Feng L, Zhang Z, *et al*. Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci Rep* 2015;**6**:21383.
4. Mizianty MJ, Kurgan L. Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics* 2011;**27**:i24–33.
5. Service R. Structural biology. Structural genomics, round 2. *Science* 2005;**307**:1554–8.
6. Kurgan L, Mizianty MJ. Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis. *Nat Sci* 2009;**1**:93–106.
7. Terwilliger TC, Stuart D, Yokoyama S. Lessons from structural genomics. *Annu Rev Biophys* 2009;**38**:371–83.
8. Ng JT, Dekker C, Reardon P, *et al*. Lessons from ten years of crystallization experiments at the SGC. *Acta Crystallogr D Struct Biol* 2016;**72**:224–35.
9. Zimmerman MD, Grabowski M, Domagalski MJ, *et al*. Data management in the modern structural biology and biomedical research environment. *Methods Mol Biol* 2014;**1140**:1–25.
10. Jahandideh S, Jaroszewski L, Godzik A. Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallogr D Biol Crystallogr* 2014;**70**:627–35.
11. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
12. Kobe B, Guss M, Huber T. *Structural Proteomics: High-Throughput Methods*. New York, NY: Humana Press, 2008.
13. Terwilliger TC. The success of structural genomics. *J Struct Funct Genomics* 2011;**12**:43–4.
14. Terwilliger TC. Structural genomics in North America. *Nat Struct Mol Biol* 2000;**7**:935–9.
15. Burley SK. An overview of structural genomics. *Nat Struct Mol Biol* 2000;**7**:932–4.
16. Joachimiak A. High-throughput crystallography for structural genomics. *Curr Opin Struct Biol* 2009;**19**:573–84.
17. Grabowski M, Niedzialkowska E, Zimmerman MD, *et al*. The impact of structural genomics: the first quindecennial. *J Struct Funct Genomics* 2016;**17**:1–16.
18. Bertone P, Kluger Y, Lan N, *et al*. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* 2001;**29**:2884–98.
19. Kouranov A, Xie L, de la Cruz J, *et al*. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res* 2006;**34**:D302–5.
20. Chen L, Oughtred R, Berman HM, *et al*. TargetDB: a target registration database for structural genomics projects. *Bioinformatics* 2004;**20**:2860–2.
21. Bray JE. Target selection for structural genomics based on combining fold recognition and crystallisation prediction methods: application to the human proteome. *J Struct Funct Genomics* 2012;**13**:37–46.
22. Smialowski P, Frishman D. Protein crystallizability. In: *Data Mining Techniques or the Life Sciences*. Springer, Humana Press, NY, 2010, 385–400.
23. Overton IM, Barton GJ. Computational approaches to selecting and optimising targets for structural biology. *Methods* 2011;**55**:3–11.
24. Price Ii WN, Chen Y, Handelman SK, *et al*. Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nat Biotechnol* 2009;**27**:51–7.
25. Mizianty MJ, Fan X, Yan J, *et al*. Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr D Biol Crystallogr* 2014;**70**:2781–93.
26. Magnan CN, Randall A, Baldi P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* 2009;**25**:2200–7.
27. Agostini F, Vendruscolo M, Tartaglia GG. Sequence-based prediction of protein solubility. *J Mol Biol* 2012;**421**:237–41.
28. Agostini F, Cirillo D, Livi CM, *et al*. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics* 2014;**30**:2975–7.
29. Overton IM, Barton GJ. A normalised scale for structural genomics target ranking: the OB-score. *FEBS Lett* 2006;**580**:4005–9.
30. Smialowski P, Schmidt T, Cox J, *et al*. Will my protein crystallize? A sequence-based predictor. *Proteins* 2006;**62**:343–55.
31. Chen K, Kurgan L, Rahbari M. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem Biophys Res Commun* 2007;**355**:764–9.
32. Overton IM, Padovani G, Girolami MA, *et al*. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics* 2008;**24**(7):901.
33. Kurgan L, Razib AA, Aghakhani S, *et al*. CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Struct Biol* 2009;**9**:50.
34. Kandaswamy KK, Pugalenthi G, Suganthan P, *et al*. SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence. *Protein Pept Lett* 2010;**17**:423–30.

35. Hennerdal A, Elofsson A. Rapid membrane protein topology prediction. *Bioinformatics* 2011;**27**:1322–3.
36. Tsirigos KD, Hennerdal A, Käll L, *et al*. A guideline to proteome-wide α-helical membrane protein topology predictions. *Proteomics* 2012;**12**:2282–94.
37. Faraggi E, Zhang T, Yang Y, *et al*. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem* 2012;**33**:259–67.
38. Buchan DW, Minneci F, Nugent TC, *et al*. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res* 2013;**41**:W349–57.
39. Petersen TN, Brunak S, von Heijne G, *et al*. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 2011;**8**:785–6.
40. Deng X, Eickholt J, Cheng J. A comprehensive overview of computational protein disorder prediction methods. *Mol Biosyst* 2012;**8**:114–21.
41. Ruggiero A, Smaldone G, Squeglia F, *et al*. Enhanced crystallizability by protein engineering approaches: a general overview. *Protein Pept Lett* 2012;**19**:732–42.
42. Babnigg G, Joachimiak A. Predicting protein crystallization propensity from protein sequence. *J Struct Funct Genomics* 2010;**11**:71–80.
43. Charoenkwan P, Shoombuatong W, Lee H-C, *et al*. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PloS One* 2013;**8**:e72368.
44. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
45. Slabinski L, Jaroszewski L, Rychlewski L, *et al*. XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics* 2007;**23**:3403–5.
46. Overton IM, van Niekerk C, Barton GJ. XANNpred: neural nets that predict the propensity of a protein to yield diffraction-quality crystals. *Proteins* 2011;**79**:1027–33.
47. Jahandideh S, Mahdavi A. RFCRYS: Sequence-based protein crystallization propensity prediction by means of random forest. *J Theor Biol* 2012;**306**:115–19.
48. Kurgan L. CRYSpred: accurate sequence-based protein crystallization propensity prediction using sequence-derived structural characteristics. *Protein Pept Lett* 2012;**19**:40–9.
49. Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
50. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**:2507–17.
51. Wang M, Zhao X-M, Tan H, *et al*. Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics* 2014;**30**:71–80.
52. Wang M, Zhao X-M, Takemoto K, *et al*. FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PloS One* 2012;**7**:e43847.
53. Li F, Li C, Wang M, *et al*. GlycoMine: a machine learning-based approach for predicting N-, C-and O-linked glycosylation in the human proteome. *Bioinformatics* 2015;**31**:1411–19.
54. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**:27.
55. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;**157**:105–32.
56. Stajich JE, Block D, Boulez K, *et al*. The Bioperl toolkit: perl modules for the life sciences. *Genome Res* 2002;**12**:1611–18.
57. Canaves JM, Page R, Wilson IA, *et al*. Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: maximum clustering strategy for structural genomics. *J Mol Biol* 2004;**344**:977–91.
58. Kawashima S, Pokarowski P, Pokarowska M, *et al*. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;**36**:D202–5.
59. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000;**16**:404–5.
60. Rao H, Zhu F, Yang G, *et al*. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* 2011;**39**:W385–90.
61. Chen H, Wang H, Sun T, *et al*. Recombinant preparation and functional studies of EspI ATP binding domain from *Mycobacterium tuberculosis*. *Protein Expr Purif* 2016;**123**:51–9.
62. Cooper DR, Boczek T, Grelewska K, *et al*. Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta Crystallogr D Biol Crystallogr* 2007;**63**:636–45.
63. Goldschmidt L, Cooper DR, Derewenda ZS, *et al*. Toward rational protein crystallization: a web server for the design of crystallizable protein variants. *Protein Sci* 2007;**16**:1569–76.
64. Pruitt KD, Tatusova T, Brown GR, *et al*. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 2012;**40**:D130–5.
65. Kundrot C. Which strategy for a protein crystallization project? *Cell Mol Life Sci* 2004;**61**:525–36.
66. McPherson A. *Crystallization of Biological Macromolecules*. New York, NY: Cold Spring Harbor Laboratory Press, 1999.
67. Derewenda ZS. The use of recombinant methods and molecular engineering in protein crystallization. *Methods* 2004;**34**:354–63.
68. Dong A, Xu X, Edwards AM, *et al*. In situ proteolysis for protein crystallization and structure determination. *Nat Methods* 2007;**4**:1019–21.
69. Walter TS, Meier C, Assenberg R, *et al*. Lysine methylation as a routine rescue strategy for protein crystallization. *Structure* 2006;**14**:1617–22.
70. Kiefer F, Arnold K, Künzli M, *et al*. The SWISS-MODEL repository and associated resources. *Nucleic Acids Res* 2009;**37**:D387–92.
71. Rupp B, Wlodawer A, Minor W, *et al*. Correcting the record of structural publications requires joint effort of the community and journal editors. *Febs J* 2016;**283**:4452–57.
72. Minor W, Dauter Z, Helliwell JR, *et al*. Safeguarding structural data repositories against bad apples. *Structure* 2016;**24**:216–20.