Survey of predictors of propensity for protein production and crystallization with application to predict resolution of crystal structures

Jianzhao Gao[a], Zhonghua Wu[a], Gang Hu[a], Kui Wang[a], Jiangning Song[b,c], Andrzej Joachimiak[*d,e,f], and Lukasz Kurgan[*g]

[a]*School of Mathematical Sciences and LPMC, Nankai University, Tianjin, People's Republic of China*
[b]*Infection and Immunity Program, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia*
[c]*ARC Centre of Excellence in Advanced Molecular Imaging, Monash University, Melbourne, Australia*
[d]*Midwest Center for Structural Genomics, Argonne, USA*
[e]*Structural Biology Center, Biosciences, Argonne National Laboratory, Argonne, USA*
[f]*Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, USA*
[g]*Department of Computer Science, Virginia Commonwealth University, Richmond, USA*

Abstract: Selection of proper targets for the X-ray crystallography will benefit biological research community immensely. Several computational models were proposed to predict propensity of successful protein production and diffraction quality crystallization from protein sequences. We reviewed a comprehensive collection of 22 such predictors that were developed in the last decade. We found that almost all of these models are easily accessible as webservers and/or standalone software and we demonstrated that some of them are widely used by the research community. We empirically evaluated and compared the predictive performance of seven representative methods. The analysis suggests that these methods produce quite accurate propensities for the diffraction-quality crystallization. We also summarized results of the first study of the relation between these predictive propensities and the resolution of the crystallizable proteins. We found that the propensities predicted by several methods are significantly higher for proteins that have high resolution structures compared to those with the low resolution structures. Moreover, we tested a new meta-predictor, MetaXXC, which averages the propensities generated by the three most accurate predictors of the diffraction-quality crystallization. MetaXXC generates putative values of resolution that have modest levels of correlation with the experimental resolutions and it offers the lowest mean absolute error when compared to the seven considered methods. We conclude that protein sequences can be used to fairly accurately predict whether their corresponding protein structures can be solved using X-ray crystallography. Moreover, we also ascertain that sequences can be used to reasonably well predict the resolution of the resulting protein crystals.

Keywords: X-ray crystallography; diffraction quality crystallization; protein production; resolution of protein crystals; meta prediction; protein structure; prediction.

1. INTRODUCTION

The unique protein sequences are sequenced and accumulated at an exponentially increasing pace, as evidenced by the rapid growth and current size of resources such as UniProt (88.03 million as of July 2017) and RefSeq (88.39 million as of July 2017) [1]. Moreover, statistics from the UniProt as of July 2017 reveal that only about 86 thousand non-redundant proteomes (sets of proteins expressed by an organism) have been completely sequenced and about 560 thousand organisms have the corresponding protein sequence data. These numbers suggest that the growth will continue in the foreseeable future, given that recent estimates suggest that the Earth is home to anywhere between 8.7 million [2] and 1 trillion [3] of organisms. More importantly, only a small fraction of these proteins have been functionally characterized to date. The latter number can be estimated based on the amount of proteins in the Swiss-Prot resource, which is at 0.56 million as of July 2017 [4]. Functions of new proteins are learnt by first producing them, i.e., they typically have to cloned, expressed, solubilized, and purified [5], and ultimately for some by acquiring their structures [6].

The emergence of structural genomics (SG) in the early 2000s [7] has resulted in rapid technological advances in protein production and structure determination by taking advantage of robotics, parallelization, new expression vectors, affinity tags, semi-automated purification and crystallization screening methods [5, 8]. Although there were a number of different protocols, a set of specific protocols were shown to be widely applicable to a large set of proteins, and this also suggested alternative strategies for non-conforming proteins [5]. Many of the purified proteins are structurally solved since knowledge of structures is essential for understanding protein functions and related biological processes [9] and facilitates drug discovery efforts [10]. According to the worldwide repository of protein structures, Protein Data Bank (PDB) [11], slightly over 90% of the currently known protein structures (111 thousand out of 123 thousand proteins structures as of July 2017) were

*Address correspondence to these authors at Midwest Center for Structural Genomics, Argonne, USA & Structural Biology Center, Biosciences, Argonne National Laboratory, Argonne, USA & Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, USA, Tel: 630-252-3926; E-mails: andrzejj@anl.gov; Department of Computer Science, Virginia Commonwealth University, Richmond, USA Tel: 804-827-3986; E-mails: lkurgan@vcu.edu

determined by X-ray crystallography. This is also the most commonly used methods by the SG centers and the focus of this article. However, the X-ray crystallography-based protein structure determination efforts are hampered by very low success rates due to the cumulative attrition along the protein production and crystallization pipelines. Studies show that the success rates range between 2 and 10% [12], depending on the source and time of the analysis. Importantly, failed attempts are estimated to account for over 60% of the structure determination costs [12a, 13].

Data originated from these protein production and structure determination experiments were deposited into the TargetTrack [14] database, which superseded the TargetDB [15] and PepcDB [16] databases. TargetTrack includes information about successful and failed protein production and crystallization trials for over 300 thousand protein targets that were provided by dozens of SG centers across the globe. These proteins span all kingdoms of life and cover a diverse range of protein structural and functional classes. The low success rates and availability of the large databases prompted efforts to characterize the determining factors of the amenability of proteins to the production and structure determination. Although the success rates depend on the experimental protocols used, the premise of these studies is that certain intrinsic characteristics of protein sequences generalize across protocols. Indeed, the results collected across many studies of the factors demonstrate that this is the case [17]. These results in turn motivate the development and widespread use of sequence-based target selection tools, developed based on computational methods that use protein sequences to accurately predict the propensity of these proteins to be successfully produced and structurally determined. While propensities for successful gene cloning experiments do not necessarily need to be predicted as they boast nearly 100% success rate, the other steps including protein expression, solubilization, purification, crystallization, and structure determination, require accurate predictive tools. The efforts to build these predictive models so far have been largely focused on the final structure determination step, resulting in the release of over a dozen of predictors [12d, 18]. This article reviews a comprehensive set of (to the best of our knowledge) all 22 currently available predictive models. These include computational models that predict propensity of several production/structure determination steps. To compare, another two most recent reviews on this subject have surveyed 18 [18a] and 10 tools [12d]. Moreover, we also comparatively assessed the predictive performance of several tools, and for the first time, we empirically analyzed the relationship between the predicted propensities for structural determination using X-ray crystallography and the diffraction resolution limits of the successfully solved protein structures. Such analysis could guide the selection of these tools, beyond the typical target choice, in terms of selecting proteins that may lead to higher resolutions of the crystal structures directly from their amino acid sequences.

2. OVERVIEW OF PREDICTORS OF PROPENSITY FOR PROTEIN PRODUCTION AND STRUCTURE DETERMINATION

Sequence-based target selection tools are typically developed based on empirical fitting of predictive models using sequences that were annotated experimentally as either amenable or recalcitrant to a given production/structure determination step. The annotations of these proteins are usually collected from the public databases, such as TargetTrack, TargetDB and PepcDB. The predictive models are used to predict propensity of new protein sequences to undergo the successful completion of a given production/structure determination step. The predictors take a protein sequence as the input and generate numerical score(s) that denotes the likelihood (propensity) that this sequence will complete specific step(s) of the protein production and structure determination pipeline. The typically considered steps include gene cloning, protein production, purification, crystallization and crystallization to obtain diffraction-quality crystals; the latter is synonymous with the successful determination of the structure.

The first predictors of propensity for X-ray crystallography-based protein structure determination, SECRET [19] and OB-Score [20], were developed in 2006. The development of these two methods happened shortly after the release of the TargetDB and PepcDB resources in 2001 and 2004, respectively. Eight other methods that predict propensity for diffraction-quality crystallization were published between 2007 and 2010. In chronological order they are CRYSTALP [21], XtalPred [13, 22], ParCrys [23]; PXS [17d], CRYSTALP2 [24], MetaPPCP [25], SVMCrys [26], and MCSG-Z score [27]. The year 2011 marks the release of the first tool, PPCPred [28], that predicts both the propensity for the protein production and the propensity for the structure determination that was also addressed by previous methods. More specifically, PPCpred predicts the propensities for multiple steps including material production, purification, crystallization and diffraction-quality crystallization. Since 2011, three other tools that similarly cover multiple steps were developed: PPCinter [29] and PredPPCrys [30] in 2014 and Crysalis [31] in 2016. The latter two methods also generate the propensity for protein cloning in addition to the propensity for the four steps that are covered by PPCpred and PPCinter. Moreover, eight more tools that address the prediction of the propensity for the diffraction-quality crystallization were made available over the last six years: XANNpred [32], RFCRYS [33], CRYSPred [34], SCMCRYS [35], fDETECT [36], XtalPred-RF [37], TargetCrys [38], and Crysf [18a]. In total, 22 methods have been developed over the last decade. They are summarized in **Table 1**.

Table 1. Summary of protein sequence-based predictors of the propensity for protein cloning, material production, purification, crystallization and structure determination using X-ray crystallography (diffraction-quality crystallization). Bold font identifies methods that were used to perform empirical analysis.

Method	Year published	Availability[1]	Batch prediction (max number of proteins allowed)	Output[2]	SG center[3]	Number of citations[4]		URL[5]
						total	annual	
OB-score	2006	SA	No	DCR	SSPF	52	4.7	http://www.compbio.dundee.ac.uk/obscore/
SECRET	2006	WS	Yes (25; 46<chain length<200)	DCR		82	7.5	http://mips.helmholtz-muenchen.de/secret/secret.seam#
CRYSTALP	2007	NA	No	DCR		69	6.9	NA
XtalPred	**2007**	**WS**	**Yes (10)**	**DCR**	**JCSG**	**158**	**15.8**	**http://ffas.burnham.org/XtalPred-cgi/xtal.pl**
ParCrys	2008	WS	Yes (no limit)	DCR		57	6.3	http://www.compbio.dundee.ac.uk/parcrys/cgi-bin/input.pl
P_{XS}	2009	WS	No	DCR	NESG	115	12.7	http://nmr.cabm.rutgers.edu:8080/PXS/
CRYSTALP2	2009	WS	Yes (100)	DCR		50	6.3	http://biomine.cs.vcu.edu/servers/CRYSTALP2/
MetaPPCP	2009	NA	No	DCR		27	3.4	NA
SVMCrys	2010	SA	No	DCR		30	4.3	http://www3.ntu.edu.sg/home/EPNSugan/index_files/svmcrys.htm
MCSG Z-score	2010	WS	No	DCR	MCSG CSGID	26	3.7	http://bioinformatics.anl.gov/cgi-bin/tools/pdpredictor/
PPCpred	**2011**	**WS**	**Yes (5)**	**MP PF CR DCR**		**48**	**8.0**	**http://biomine.cs.vcu.edu/servers/PPCpred/**
XANNpred	2011	WS	Yes (5)	DCR	SSPF	17	2.8	http://www.compbio.dundee.ac.uk/xtal/cgi-bin/xannpred_in.pl
RFCRYS	2012	NA	No	DCR		19	3.8	NA
CRYSPred	2012	NA	No	DCR		17	3.4	NA
SCMCRYS	2013	SA	No	DCR		21	5.3	http://iclab.life.nctu.edu.tw/SCMCRYS/
fDETECT	**2014**	**WS**	**Yes (1000)**	**DCR**		**6**	**2.0**	**http://biomine-ws.ece.ualberta.ca/fDETECT/**
PredPPCrys	**2014**	**WS**	**No**	**CL MP PF CR DCR**		**8**	**2.7**	**http://www.structbioinfor.org/PredPPCrys/**
PPCinter	2014	NA	No	MP PF CR DCR		5	1.7	NA
XtalPred-RF	**2014**	**WS**	**Yes (10)**	**DCR**		**20**	**6.7**	**http://ffas.burnham.org/XtalPred-cgi/xtal.pl**
Crysalis	**2016**	**WS**	**Yes (10000)**	**CL MP PF CR DCR**		**6**	**6.0**	**http://nmrcen.xmu.edu.cn/crysalis/**
TargetCrys	**2016**	**WS**	**Yes (no limit)**	**DCR**		**3**	**3.0**	**http://csbio.njust.edu.cn:8080/TargetCrys/**
Crysf	2017	WS	Yes (10000)	DCR		0	0.0	http://nmrcen.xmu.edu.cn/crysf/

[1] Availability: webserver (WS); standalone application (SA); and implementation/webserver is not available (NA).

[2] Output: propensity for cloning (CL); propensity for material production (MP); propensity for purification (PF); propensity for crystallization (CR); and propensity for diffraction-quality crystallization (DCR).

[3] Structural genomics (SG) centers that use a given tool: Midwest Center for Structural Genomics (MCSG); Center for Structural Genomics of Infectious Diseases (CSGID); Joint Center for Structural Genomics (JCSG); Scottish Structural Proteomics Facility (SSPF); and NorthEast Structural Genomics consortium (NESG).

[4] The number of citations was collected using Google Scholar in May 2017.

[5] URL: The URL address of the developed method; Not available (NA) denotes that the implementation/webserver is not available.

The early methods, including SECRET, OB-Score and CRYSTALP, were built using relatively small datasets of proteins accompanied by low quality annotations of propensities for structure determination. Newer methods use well-annotated and larger datasets sourced from TargetTrack, TargetDB and PepcDB, and their predictive quality has improved over time [18a, 28]. These improvements were driven by the inclusion of a larger number of indicators/determining factors of amenability of proteins for the production and structure determination and implementation of more sophisticated predictive models. The predictive models utilized by recent predictors include neural network (XANNpred), random forest (XtalPred-RF and RFCRYS), support vector machine (PPCpred, CRYSPred, PredPPCrys, PPCinter, and TargetCrys) and support vector regression (Crysalis and Crysf).

Several of these tools are used directly by SG centers: MCSG-Z score at the Midwest Center for SG (MCSG) and Center for SG of Infectious Diseases (CSGID), XtalPred at the Joint Center for SG (JCSG), XANNpred and OB-Score by the Scottish Structural Proteomics Facility (SSPF) and PXS by the NorthEast SG consortium (NESG). Many of these tools are made available as webservers or standalone applications to provide service to a wider structural biology community. Some of them enjoy relatively strong uptake by the community. **Table 1** shows that SECRET, CRYSTALP, XtalPred, ParCrys, Pxs, CRYSTALP2, PPCpred, SCMCRYS, XtalPred-RF and Crysalis are cited on average at least five times per year since they were published. The top three most cited methods are XtalPred (16 citations per year), PXS (13 per year) and PPCpred (8 per year). We note that these citation counts should be interpreted with caution since some of the articles discuss other aspects beyond the methodology (e.g., articles that introduce the PXS and fDETECT methods) and some were published very recently and do not yet have sufficient citation record.

The webservers, which are provided for 14 out of 22 tools, appeal to less computer savvy end users that require predictions in *an ad* hoc manner. To use a webserver, a user needs to arrive at a specific URL (see **Table 1**) using any of the major web browsers, enter their protein sequence(s) and request the prediction. The prediction is performed on the server side, using a fully automated and free service that delivers the results into a browser window and/or via email. Moreover, several of these webservers allow for batch predictions of multiple proteins. They include SECRET, XtalPred, ParCrys, CRYSTALP2, PPCpred, XANNpred, fDETECT, XtalPred-RF, Crysalis, TargetCrys, and Crysf. **Table 1** provides the upper limit of the size of the protein sets that can be predicted with these methods. In contrast to the webservers, the standalone applications are geared towards users who need to use these tools more frequently. In this case, a given predictor must be installed and run on the user's computer. Three tools (OB-score, SVMcrys and SCMCRYS) offer this option. Moreover, five of the 22 tools were not made publically available by the authors.

We have recently empirically demonstrated that the use of these tools leads to a substantial improvement in the quality of the selection of proteins for structure determination when compared with an *ad hoc* target selection [36]. For example, currently about 14% of modeling families of human proteins have structures; the modeling family is a group of proteins for which accurate structures can be obtained through computational modeling if at least one of the proteins in that group has known structure. We estimated that this coverage can be substantially improved to 49% assuming that all human proteins that can be crystallized, based on their putative propensities, will be solved. Moreover, we showed that solving the structures of the same number of human proteins selected at random would lead to an estimated coverage of 5%, which is significantly lower than the current coverage.

3. EMPIRICAL EVALUATION OF SELECTED PREDICTORS OF PROTEIN CRYSTALLIZATION

We evaluated and compared the predictive quality of seven representative predictors of the propensity for diffraction-quality crystallization on a new benchmark dataset. We also performed first-of-its-kind large-scale evaluation of the relation between these propensities and the resolution of proteins that are successfully crystalized.

3.1. Setup of the empirical evaluation

The benchmark dataset includes an equal number of proteins that were successfully solved using X-ray crystallography and proteins for which attempts to obtain diffraction-quality crystals failed, which we collected in July 2016. We balanced the size of these two protein sets to ensure that the corresponding empirical assessment, including statistical analysis, is straightforward and reliable. We limited to the dataset size to about 500 proteins to ease the burden of collecting the prediction outputs of the seven predictors that we considered. The set of 267 proteins that could not be crystallized was obtained from the TargetTrack database (http://sbkb.org/). They include an equal number of randomly chosen proteins that fail the three main steps of the crystallization pipeline: 89 proteins that fail protein production step, 89 that fail purification step and 89 that fail to form crystals. The annotation of these step was made using the trial stop statuses available in TargetTrack as follows: sequencing failed, cloning failed and expression failed stop statuses were used to annotate proteins that failed material production; purification failed stop status to denote the failure to purify; and crystallization failed and poor diffraction stop statuses to annotate proteins that fail to form crystals. We matched this set of 267 proteins with the other 266 proteins that were crystalized and for which the resolution of these crystals is uniform, i.e., we have sufficient and similar numbers of proteins over the entire range of the resolution. To derive this set of crystallizable proteins, we first downloaded the protein structures that were deposited between July 2006 and July 2016 in PDB [39]. We filtered the corresponding 204,440 protein chains based on five following steps. First, we removed peptides, i.e., protein chains with ≤ 30 amino acids (194,674 chains remain). Second, we deleted chains which contain over 1% of the non-standard amino acid X in the sequence (193,902 chains remain). Third, we eliminated sequences that cannot be matched with proteins in the UniProt (41,120 chains are left); this is necessary to ensure that we do not use multiples of the same protein in the dataset and that we can select the structure with the highest resolution to represent a given protein. Correspondingly, in the fourth step we clustered the remaining sequences using BLASTCLUST [40] with the coverage 100% (-L 1) and identity score (-S 100) for one of the sequences (-b F); for each resulting cluster, the sequence with maximal length and the highest resolution was selected to represent the cluster (7,254 chains are left). Fifth, we randomly chosen 19 proteins for each of the following 14 resolution bins defined in ref. [36]: below 1.19Å, [1.19Å, 1.22Å), [1.22Å, 1.26Å), [1.26Å, 1.30Å), [1.30Å, 1.36Å), [1.36Å, 1.42 Å), [1.42Å, 1.49Å), [1.49Å, 1.59Å), [1.5 Å, 1.71Å), [1.71Å, 1.88Å), [1.88Å, 2.15Å), [2.15Å, 2.71Å), [2.71Å, 3.50Å], and over 3.50Å. This ensures that the resolution of the selected proteins is distributed over a wide resolution range, which is important for testing the relation between putative propensities for the structural determination and the diffraction resolution. The final dataset includes 533 proteins and is provided in the Supplementary **Table S1**.

We selected a set of representative methods for the prediction of propensity for structure determination to be included in this empirical analysis. We focused on modern and popular methods that are implemented as webservers and we used these webservers to collect their predictions. We also required that the considered methods output real-valued propensities, in contrast to methods that generate only binary outcomes (crystallizable vs. non-crystallizable protein). The real-valued score is necessary to analyze the relation between the putative propensities and the resolution of protein structures. We included all methods that were published after 2013 that offer webservers (**Table 1**), such as fDETECT, PredPPCrys, XtalPred-RF, Crysalis and TargetCrys. We could not include Crysf since this method uses extra inputs, beyond the protein sequence that is used by all other methods, in the form of a functional TrEMBL profile. Consequently, Crysf can be applied to a subset of proteins that have sufficiently complete set of annotations in TrEMBL, which excludes some of the proteins in our benchmark dataset. In addition, we supplemented the five new predictors with two older and commonly cited methods which are available as webservers that offer batch predictions: XtalPred (published in 2007; 16 citations per year) and PPCpred (published in 2011; 8 citations per year). The availability of the batch prediction option has eased the efforts of securing predictions for our benchmark dataset. To sum up, we included seven methods in our empirical study.

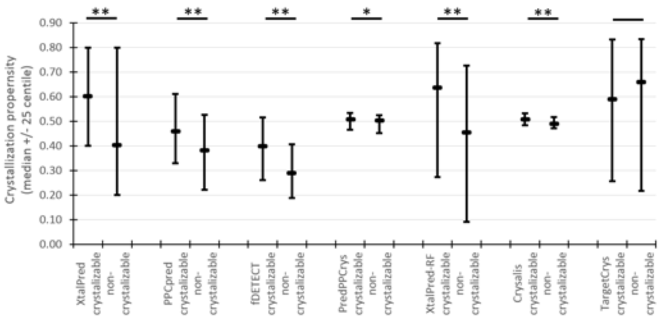

Fig. 1. Values of putative propensity for diffraction-quality crystallization generated by the seven considered predictors for crystallizable and non-crystallizable proteins. The distributions of the values of propensities are represented using median and 25th and 75th centiles (error bars). The top of the plot summarizes the significance of the differences between the propensities for the crystallizable and non-crystallizable proteins generated by the same predictor. Significance was quantified with the Wilcoxon rank sum test: ** when *p*-value<0.001, * when *p*-value<0.05. Values generated by XtalPred (XtalPred-RF), which are integers between 1 and 5 (1 and 11), were divided by 5 (11) to fit into the [0, 1] range.

3.2. Assessment of the prediction of propensity for diffraction-quality crystallization

The seven predictors were used to predict propensity for crystallization for the proteins from the benchmark dataset and these predictions were compared with the experimental annotations. **Fig. 1** summarizes the differences in the values of propensities generated by each of the seven predictors between the proteins that were crystallized and those that could not be crystallized. We also assessed the statistical significance of the differences between these two sets of propensities by performing the Wilcoxon rank sum test. We used this non-parametric test since the values of propensities are not normal, which we verified using the Anderson-Darling test at the 0.05 significance. **Fig. 1** shows that the values of putative propensities for obtaining diffraction-quality crystals are significantly different (*p*-value < 0.05) between crystallizable vs. non-crystallizable proteins for six out of the seven predictors on our benchmark dataset. This is expected given that these tools were designed specifically to differentiate between these two classes of proteins.

Table 2. Predictive performance of protein sequence-based predictors of the propensity for diffraction-quality crystallization. The first row gives AUC values when the propensities are used to predict crystalizable vs. non-crystallizable proteins. The second and third rows provide AUC values when the propensities are used to predict crystallizable proteins with above average (<2.2) resolution and crystallizable proteins with high (<1.6) resolution from other crystallizable and non-crystallizable proteins. The last two rows provide the values of Pearson correlation coefficient (PCC) between the putative propensity and resolution of crystallizable proteins, and the mean absolute error (MAE) measured between the putative propensity and resolution. When computing MAE we converted the values of putative propensities to fit them into the distribution of resolutions for proteins deposited in PDB. Bold font identifies methods that are used to implement the MetaXXC predictor.

Type of analysis	Metric	TargetCrys	PredPPCrys	**XtalPred**	**XtalPred-RF**	PPCpred	**Crysalis**	fDETECT	**MetaXXC**
Prediction of crystalizable vs. non-crystallizable proteins	AUC	0.50	0.56	0.59	0.60	0.61	0.62	0.64	0.61
Prediction of crystallizable proteins with above average resolution (<2.2 Å) vs. other proteins	AUC	0.53	0.57	0.62	0.63	0.62	0.63	0.62	0.65
Prediction of crystallizable proteins with high resolution (<1.6 Å) vs. other proteins	AUC	0.50	0.58	0.60	0.61	0.65	0.65	0.62	0.63
Relation between putative propensity and resolution of crystallizable proteins	PCC	-0.18	-0.06	-0.23	-0.25	-0.14	-0.09	0.02	-0.29
	MAE [Å]	0.79	0.84	1.66	1.46	0.78	0.82	0.89	0.77

Similar to recent articles [18a, 28-32, 34, 36], we measured the predictive performance of these predictors with the area under the receiver operating characteristic (ROC) curve (AUC). The AUC values range between 0.5 (for a random predictor) and 1 (for a perfect predictor) and are used to assess the real-valued propensities against the experimental annotations of crystallizable vs. non-crystallizable proteins. The first row in **Table 2** summarizes these results. Since these methods use protein sequence as the only input they achieve relatively modest values of AUC. The most accurate methods that obtain AUC ≥ 0.6 are XtalPred-RF, PPCpred, Crysalis and fDETECT. This is in agreement with **Fig. 1** where the distributions of the propensities generated by these four methods are the most different between the crystallizable and non-crystallizable proteins. As expected XtalPred-RF is better than XtalPred, given that both were developed by the same group and the former superseded the latter. We note that these predictions are considered as reasonably accurate and can be used in a practical context, given the fact that the overall success rates of the X-ray crystallography are below 10% [12].

3.3. Application of the putative propensity for the diffraction-quality crystallization to predict resolution of protein crystals

The resolution quantifies the quality of the data collected from the crystal. If proteins in the crystal are perfectly aligned, then they will all scatter X-rays the same way and the diffraction pattern will show fine details of the protein structure. Higher quality structures have higher resolution but lower values of resolution. The resolution of the structures derived with X-ray crystallography typically ranges between about 0.5 Å and several Å. The average resolution of protein structures equals to 2.2 Å with the standard deviation of 1.2 Å (source: www.rcsb.org/). Roughly speaking, high quality structures (resolution < 1.2 Å) provide accurate structural information for all atoms and amino acids, and are generally error-free. Good quality structures (resolution between 1.2 Å and 2.2 Å) may include inaccurate rotamers for some amino acids but the overall structure is typically correct. Acceptable-quality structures (resolution between 2.2 Å and 3.4 Å) have some amino acids with incorrect rotamer information and usually structure of parts of the protein surface is inaccurate. Low-quality structures (resolution above 3.4 Å) frequently have incorrect rotamer information and their surface is largely inaccurate. At over 4 Å resolution the secondary structure of the protein cannot be determined and as a result the 3-D structure is inaccurate. Apart from the observable impact on the quality of the structure, at least good quality is desirable when modeling with the protein structure, for instance for rational drug design [41] and to computationally model protein-protein interactions [42].

Interestingly, a couple of recent works hypothesize that the putative propensities for diffraction-quality crystallization can be also used to predict the resolution of the resulting solved protein structures [18a, 36]. In 2014, the authors of ref. [36] observed that proteins that have higher resolution structures also on average have higher values of putative propensity generated with fDETECT. A similar conclusion was reached in 2016 for the predictions that combine results from Crysf and Crysalis [18a]. We followed up on these findings by evaluating and combining multiple predictors of propensities for structural determination.

Fig. 2 summarizes the differences in the values of propensities generated by each of the seven predictors between the crystallizable proteins that have high resolution (38 proteins with the highest resolution < 1.22 Å) and low resolution (38 proteins with the lowest resolution > 2.71 Å) in our benchmark dataset. We also assessed the significance of the differences between these two sets of propensities using the Wilcoxon rank sum test (the distributions of the propensity values are not normal). We note that three methods, namely XtalPred, XtalPred-RF and Crysalis, produce propensities that are significantly higher (*p*-value<0.05) for proteins that have high resolution structures when compared to the proteins that have low resolution structures. We combined the outputs of these three methods using a simple average to devise a meta predictor, MetaXXC. As expected, the differences in the propensity values between high and low resolution proteins that are produced by MetaXXC are significant with *p*-value = 0.003 (**Fig. 2**). This p-value is lower than the *p*-values for the base (input) predictors; *p*-value = 0.007 for Crysalis, 0.009 for XtalPred, and 0.015 for XtalPred-RF. Expectedly, this is in agreement with **Fig. 2** where the distributions of the propensities generated by MetaXXC are the most different between the high resolution and low resolution structures.

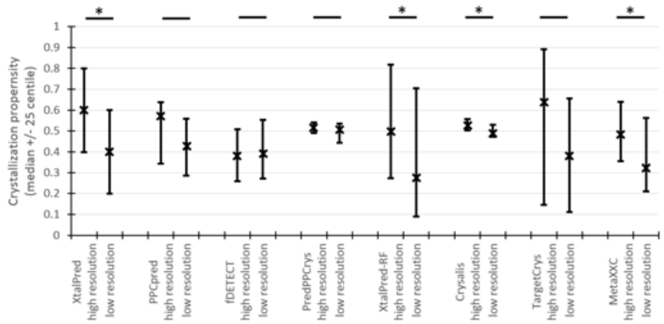

Fig. 2. Values of putative propensity for diffraction-quality crystallization generated by the seven considered predictors for crystallizable proteins with high resolution (< 1.22 Å) and low resolution (> 2.71 Å). The distributions of the values of propensities are represented using the median and 25th and 75th centiles (error bars). The top of the plot indicates the significance of the differences between propensities for the high and low resolution proteins generated by the same predictor. The significance was quantified with the Wilcoxon rank sum test: ** when *p*-value<0.001, * when *p*-value<0.05. The integers values between 1 and 5 (1 and 11) generated by XtalPred (XtalPred-RF) were transformed to 1, 0.75, 0.5, 0.25, 0 (1, 0.9, 0.8, …, 0.1, 0) to fit into the [0, 1] range.

Next, we investigated whether the putative propensity for the diffraction-quality crystallization generated by the seven predictors and MetaXXC can be used to predict proteins with high resolution (<1.6 Å) and above average (<2.2 Å) resolution. The 1.6 Å cutoff is computed as the average resolution of structures in PDB, which equals 2.2 Å, minus 0.5*standard_deviation of the resolution of proteins in PDB, which is 1.2 Å. The second and third rows in **Table 2** summarize these results. These AUC values are for the prediction of the crystallizable proteins with above average or with high resolution from other crystallizable and non-crystallizable proteins. We observed that several methods, such as Crysalis, XtalPred-RF, PPCpred, XtalPred, and fDETECT, can relatively accurately (AUC > 0.6) predict proteins that have structures with the above average resolutions. The same five methods have AUC > 0.60 for the prediction of proteins with the high resolution structures. Moreover, MetaXXC secures the highest AUC = 0.65 for the former prediction and the third highest AUC = 0.63 for the latter prediction. These results suggest that targeting proteins which score high putative propensities for obtaining diffraction-quality crystals generated by these methods is likely to result in high-resolution structures.

To further investigate this point, we studied a relation between the propensities and the resolution of the structures of the 266 crystallizable proteins in our benchmark dataset. We computed the Pearson correlation coefficient (PCC) between the propensities and resolution (the fourth row in **Table 2**). The PCC values of all methods, except for fDETECT that secures near zero PCC, are negative. This reveals that higher propensities are characteristic of proteins with lower values of the resolution (i.e., proteins with higher resolution structures). The highest PCC = -0.29 is secured by MetaXXC, suggesting a modest degree of correlation between propensities and resolution. Moreover, propensities produced by three other methods, XtalPred-RF, XtalPred and TargetCrys, have correlation near or below -0.2. Next, we computed the mean absolute error (MAE) between the predicted and experimentally determined resolutions. The predicted resolution was computed from the putative propensities output by each of the eight methods, including MetaXXC, by normalizing it to the range of values of the experimental resolutions. The calculation process is as follows. First, propensities generated by each method are scaled to the unit range using the min-max normalization. Second, the score is inverted by subtracting the normalized value from 1; this results in a score that has a positive correlation with the resolution. Lastly, the resulting scores for a given predictor are fit into the distribution of resolutions of structures in PDB (http://www.rcsb.org/pdb/statistics/histogram.do?mdcat =refine&mditem=ls_d_res_high&minLabel=0&maxLabel=7 .01&numOfbars=700&name=Resolution). We fit the distribution using discrete intervals of size 0.01 Å to cover the range of resolution of the proteins in our benchmark dataset, which is between 0.48 and 7.01 Å. The last row in **Table 2** summarizes these results. The MAE values range from 0.77 Å (for MetaXXC predictor) to 1.66 Å (for XtalPred method). Overall, we observe that the meta predictor secures the lowest value of MAE and the highest PCC. The higher values of MAE for XtalPred and XtalPred-RF can be explained by the fact that these methods output only a small set of 5 and 11 possible values, respectively. Besides being impractical given that the resolution is a real number, this results in larger errors compared to methods that generate the real-valued outputs. The errors in the 0.8 Å range seem to be reasonably low given that the range of values of resolution in our benchmark dataset is about 6.5 Å and that the standard deviation of resolutions of protein structures in PDB 1.2 Å.

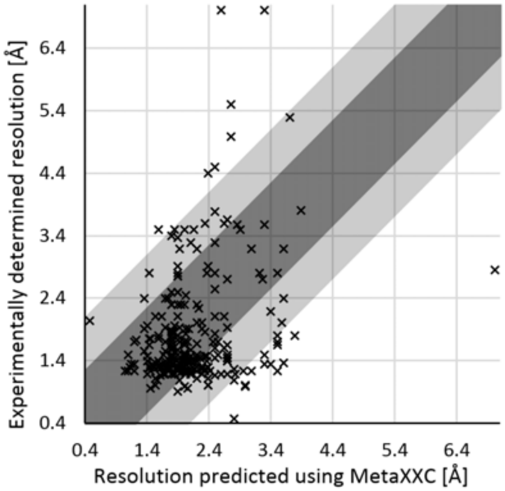

Fig. 3. Relation between the putative resolution values generated with the MetaXXC method and the experimental structure resolution values for the crystallizable proteins. Each point denotes a protein from the benchmark dataset. The putative resolution was computed by normalizing the putative propensities for obtaining the diffraction-quality crystals output by MetaXXC to the range of values of experimental resolutions in PDB. The shaded diagonal areas represent the space where the prediction error is smaller than half of the standard deviation of resolution of proteins structures in PDB (0.6 Å, dark shade) and smaller than the standard deviation of resolution of proteins structures in PDB (1.2 Å, light shade).

Fig. 3 offers a closer look at the relation between the experimentally determined resolution and the resolution predicted by the best performing meta-predictor MetaXXC for our benchmark proteins. The shaded diagonal areas represent the space where the prediction error is smaller than half of the standard deviation of resolution of proteins structures in PDB (0.6 Å, dark shade) and smaller than the standard deviation (1.2 Å, light shade). The figure reveals that about 50% of predictions are within half of the standard deviation from the experimentally determined resolution and about 80% are within the standard deviation.

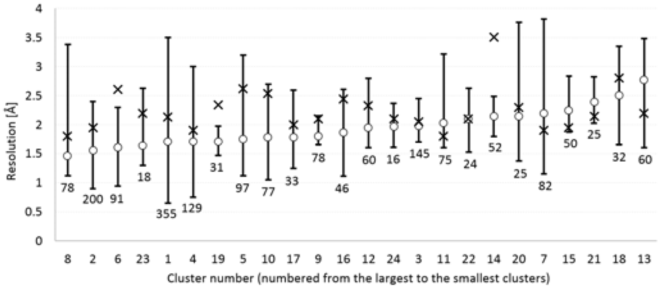

Fig. 4. Relation between the putative resolution values generated with the MetaXXC method and the range of experimentally determined resolutions for protein clusters collected from PDB. Each cluster includes the same protein chain that has multiple solved structures in PDB. The distributions of the values of experimentally determined resolutions are represented using the median (hollow circle marker) and maximum and minimum values (error bars). The number of protein structures in a given cluster is given below the error bars. Clusters are sorted in the ascending order according to the experimentally determined value of resolution.

Interestingly, PDB includes multiple structures of certain protein chains. Often these proteins are complexed with different ligands (e.g., carbonic anhydrase II in structures 2NNG, 2NNO, 3CYU, etc.; transthyretin in structures 3TCT, 4QXV, 4IIZ, etc.) or part of multimer mutants where mutations are present in the other proteins involved in the complex (e.g., proteins in the 20S proteasome in structures 5FG7, 4QV9, 4QV3, etc.). Consequently, although these structures are for the same proteins, they may have different resolutions. We collected clusters of structures of the same protein chains to investigate how well the putative propensity for the diffraction-quality crystallization generated for the same chain represent the distributions of the resolutions of these protein structure clusters. First, we collected and clustered all identical chains in PDB structures. Next, we selected clusters that include at least ten chains. Finally, for each cluster we computed an average number of chains in the PDB structures that include the corresponding chain (size of the protein assembly) and we removed clusters where this average number of chains > 4. The latter aims to exclude cases where the resolution of the protein chains in the cluster is heavily dependent on the many other chains in the PDB structure. Consequently, we collected 24 clusters of structures of identical protein chains, whose size ranges from 16 structures (in the case of chaperone heat shock protein 90) to 355 structures (in the case of lysozyme C). The clusters are provided in the Supplementary **Table S2**. **Fig. 4** shows the range of experimentally determined resolutions of the structures in each cluster, including the median value represented by the circles, together with the resolution predicted with the best performing MetaXXC method (denoted by cross shaped markers). Each cluster is accompanied by one prediction since all corresponding protein sequences are identical. The predicted resolution is within the range of experimentally determined resolution ranges of a given cluster for 87% of clusters (21 out of 24). The mean absolute difference between the putative resolution and the median resolution of the structures in clusters equals 0.43 Å and difference is below 1 Å for all but two clusters. For 16 out of 25 clusters, the difference is below 0.50 Å. Overall, we conclude that putative resolution produced by the meta method agrees with the range of experimentally determined resolutions, suggesting that the information extracted from the protein sequence can indeed be used to quantitatively characterize the resolution of protein crystals.

4. CONCLUSIONS

We reviewed a comprehensive set of 22 computational methods that predict the propensity of successful completion of several protein production and structure determination steps solely from the protein sequences. We found that most of these models are easily accessible to the end users, as either webservers and/or standalone software, and that some methods are well cited, suggesting that they are being reasonably well utilized by the scientific community. An empirical analysis of the predictive performance of a selected set of seven representative predictors reveals that they output quite accurate values of propensity for the diffraction-quality crystallization. Using a new benchmark dataset, we found that six methods generated values that are significantly larger for the crystallizable proteins when compared to the proteins that could not be crystallized. Moreover, we summarized results of a first-of-its kind study of a relation between the putative propensities for the diffraction-quality crystallization and the resolution of the crystallizable proteins. We found that the Crysalis, XtalPred and XtalPred-RF methods produce propensities that are significantly higher for the proteins that have high resolution structures available in PDB compared to the low resolution structures. Based on this observation, we devised a meta predictor, MetaXXC, by averaging the propensities generated by these three methods. The propensities generated by each of the eight methods, including MetaXXC, can be converted into putative values of resolution by normalizing them to the range of experimentally determined values of resolution. Our empirical analysis demonstrates that the putative propensities computed from the output of MetaXXC have modest correlation with the experimentally determined resolutions and the mean absolute difference between the experimental and putative propensities equals 0.77 Å. We observed that the meta predictor offers the lowest value of MAE and the highest PCC when compared with the other seven considered methods. We also found that the resolutions predicted with MetaXXC agree well with the distributions of resolutions of multiple structures for identical protein chains that can be found in PDB. The putative resolution values are on average within 0.43 Å from the median resolution of the clusters of structures of identical protein chains. Altogether, we conclude that the information extracted from the protein sequences can be used to quite accurately predict whether the structure of a given protein sequence can be solved using X-ray crystallography and to predict resolution of the resulting protein crystals.

LIST OF ABBREVIATIONS

SG: Structural gGnomics; PDB: Protein Data Bank; MCSG: Midwest Center for Structural Genomics; CSGID: Center for Structural Genomics of Infectious Diseases; JCSG: Joint Center for Structural Genomics; SSPF: Scottish Structural Proteomics Facility; NESG: NorthEast Structural Genomics Consortium; URL: Uniform Resource Locator; ROC: Receiver Operating characteristic Curve; AUC: Area Under receiver operating characteristic Curve. PCC: Pearson Correlation Coefficient; MAE: Mean Absolute Error;WS: WebServer; SA: Standalone Application; NA: Not Available; CL: propensity for CLoning; MP: propensity for Material Production; PF: propensity for PuriFication; CR: propensity for CRystallization; DCR: propensity for Diffraction-quality CRystallization.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) [grant 20130031120001], the National Natural Science Foundation of China (NSFC) [grants 11101226 and 11701296], the International Development Research Center of Canada [grant 104519-010], the Qimonda Endowed Chair funding, and the National Institutes of Health grant [GM115586 to A.J]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary.docx:

Table S1: the benchmark dataset curated in this study.

Table S2: Clusters of structures of identical protein chains collected from PDB.

REFERENCES

[1] (a) O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; Astashyn, A.; Badretdin, A.; Bao, Y.; Blinkova, O.; Brover, V.; Chetvernin, V.; Choi, J.; Cox, E.; Ermolaeva, O.; Farrell, C. M.; Goldfarb, T.; Gupta, T.; Haft, D.; Hatcher, E.; Hlavina, W.; Joardar, V. S.; Kodali, V. K.; Li, W.; Maglott, D.; Masterson, P.; McGarvey, K. M.; Murphy, M. R.; O'Neill, K.; Pujar, S.; Rangwala, S. H.; Rausch, D.; Riddick, L. D.; Schoch, C.; Shkeda, A.; Storz, S. S.; Sun, H.; Thibaud-Nissen, F.; Tolstoy, I.; Tully, R. E.; Vatsan, A. R.; Wallin, C.; Webb, D.; Wu, W.; Landrum, M. J.; Kimchi, A.; Tatusova, T.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D. *Nucleic Acids Res* **2016,** *44* (D1), D733-45; (b) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. *Nucleic Acids Res* **2004,** *32* (Database issue), D115-9; (c) The UniProt, C. *Nucleic Acids Res* **2017,** *45* (D1), D158-D169.

[2] Mora, C.; Tittensor, D. P.; Adl, S.; Simpson, A. G. B.; Worm, B. *Plos Biol* **2011,** *9* (8).

[3] Locey, K. J.; Lennon, J. T. *Proc Natl Acad Sci U S A* **2016,** *113* (21), 5970-5.

[4] Boutet, E.; Lieberherr, D.; Tognolli, M.; Schneider, M.; Bansal, P.; Bridge, A. J.; Poux, S.; Bougueleret, L.; Xenarios, I. *Methods Mol Biol* **2016,** *1374*, 23-54.

[5] Graslund, S.; Nordlund, P.; Weigelt, J.; Bray, J.; Hallberg, B. M.; Gileadi, O.; Knapp, S.; Oppermann, U.; Arrowsmith, C.; Hui, R.; Ming, J.; Dhe-Paganon, S.; Park, H. W.; Savchenko, A.; Yee, A.; Edwards, A.; Vincentelli, R.; Cambillau, C.; Kim, R.; Kim, S. H.; Rao, Z.; Shi, Y.; Terwilliger, T. C.; Kim, C. Y.; Hung, L. W.; Waldo, G. S.; Peleg, Y.; Albeck, S.; Unger, T.; Dym, O.; Prilusky, J.; Sussman, J. L.; Stevens, R. C.; Lesley, S. A.; Wilson, I. A.; Joachimiak, A.; Collart, F.; Dementieva, I.; Donnelly, M. I.; Eschenfeldt, W. H.; Kim, Y.; Stols, L.; Wu, R.; Zhou, M.; Burley, S. K.; Emtage, J. S.; Sauder, J. M.; Thompson, D.; Bain, K.; Luz, J.; Gheyi, T.; Zhang, F.; Atwell, S.; Almo, S. C.; Bonanno, J. B.; Fiser, A.; Swaminathan, S.; Studier, F. W.; Chance, M. R.; Sali, A.; Acton, T. B.; Xiao, R.; Zhao, L.; Ma, L. C.; Hunt, J. F.; Tong, L.; Cunningham, K.; Inouye, M.; Anderson, S.; Janjua, H.; Shastry, R.; Ho, C. K.; Wang, D. Y.; Wang, H.; Jiang, M.; Montelione, G. T.; Stuart, D. I.; Owens, R. J.; Daenke, S.; Schutz, A.; Heinemann, U.; Yokoyama, S.; Bussow, K.; Gunsalus, K. C.; Consortium, S. G.; Macromol, A. F.; Ctr, B. S. G.; Consortium, C. S. G.; Function, I. C. S.; Ctr, I. S. P.; Genomics, J. C. S.; Genomics, M. C. S.; Ctr, N. Y. S. G. R.; Consortium, N. S. G.; Facility, O. P. P.; Facility, P. S. P.; Med, M. D. C. M.; Proteomics, R. S. G.; Complexes, S. *Nat Methods* **2008,** *5* (2), 135-146.

[6] (a) Zhang, C.; Kim, S. H. *Curr Opin Chem Biol* **2003,** *7* (1), 28-32; (b) DePietro, P. J.; Julfayev, E. S.; McLaughlin, W. A. *BMC Struct Biol* **2013,** *13*, 24; (c) Grabowski, M.; Niedzialkowska, E.; Zimmerman, M. D.; Minor, W. *J Struct Funct Genomics* **2016,** *17* (1), 1-16.

[7] Chandonia, J. M.; Brenner, S. E. *Science* **2006,** *311* (5759), 347-51.

[8] Kim, Y.; Babnigg, G.; Jedrzejczak, R.; Eschenfeldt, W. H.; Li, H.; Maltseva, N.; Hatzos-Skintges, C.; Gu, M.; Makowska-Grzyska, M.; Wu, R.; An, H.; Chhor, G.; Joachimiak, A. *Methods* **2011,** *55* (1), 12-28.

[9] (a) Harrison, S. C. *Nat Struct Mol Biol* **2004,** *11* (1), 12-5; (b) Chang, R. L.; Andrews, K.; Kim, D.; Li, Z.; Godzik, A.; Palsson, B. O. *Science* **2013,** *340* (6137), 1220-3.

[10] (a) Grabowski, M.; Chruszcz, M.; Zimmerman, M. D.; Kirillova, O.; Minor, W. *Infect Disord Drug Targets* **2009,** *9* (5), 459-74; (b) Drinkwater, N.; McGowan, S. *Biochem J* **2014,** *461* (3), 349-69; (c) Weigelt, J. *Exp Cell Res* **2010,** *316* (8), 1332-8; (d) Anderson, W. F. *Infect Disord Drug Targets* **2009,** *9* (5), 507-17.

[11] (a) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res* **2000,** *28* (1), 235-42; (b) Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L. *Structure* **2012,** *20* (3), 391-6.

[12] (a) Jahandideh, S.; Jaroszewski, L.; Godzik, A. *Acta Crystallogr D Biol Crystallogr* **2014,** *70* (Pt 3), 627-35; (b) Zimmerman, M. D.; Grabowski, M.; Domagalski, M. J.; Maclean, E. M.; Chruszcz, M.; Minor, W. *Methods Mol Biol* **2014,** *1140*, 1-25; (c) Terwilliger, T. C.; Stuart, D.; Yokoyama, S. *Annu Rev Biophys* **2009,** *38*, 371-83; (d) Smialowski, P.; Wong, P. *Methods Mol Biol* **2016,** *1415*, 341-70.

[13] Slabinski, L.; Jaroszewski, L.; Rodrigues, A. P.; Rychlewski, L.; Wilson, I. A.; Lesley, S. A.; Godzik, A. *Protein Sci* **2007,** *16* (11), 2472-82.

[14] (a) Berman, H. M.; Westbrook, J. D.; Gabanyi, M. J.; Tao, W.; Shah, R.; Kouranov, A.; Schwede, T.; Arnold, K.; Kiefer, F.; Bordoli, L.; Kopp, J.; Podvinec, M.; Adams, P. D.; Carter, L. G.; Minor, W.; Nair, R.; La Baer, J. *Nucleic Acids Res* **2009,** *37* (Database issue), D365-8; (b) Gabanyi, M. J.; Adams, P. D.; Arnold, K.; Bordoli, L.; Carter, L. G.; Flippen-Andersen, J.; Gifford, L.; Haas, J.; Kouranov, A.; McLaughlin, W. A.; Micallef, D. I.; Minor, W.; Shah, R.; Schwede, T.; Tao, Y. P.; Westbrook, J. D.; Zimmerman, M.; Berman, H. M. *J Struct Funct Genomics* **2011,** *12* (2), 45-54.

[15] Chen, L.; Oughtred, R.; Berman, H. M.; Westbrook, J. *Bioinformatics* **2004,** *20* (16), 2860-2.

[16] Kouranov, A.; Xie, L.; de la Cruz, J.; Chen, L.; Westbrook, J.; Bourne, P. E.; Berman, H. M. *Nucleic Acids Res* **2006,** *34* (Database issue), D302-5.

[17] (a) Christendat, D.; Yee, A.; Dharamsi, A.; Kluger, Y.; Savchenko, A.; Cort, J. R.; Booth, V.; Mackereth, C. D.; Saridakis, V.; Ekiel, I.; Kozlov, G.;

Maxwell, K. L.; Wu, N.; McIntosh, L. P.; Gehring, K.; Kennedy, M. A.; Davidson, A. R.; Pai, E. F.; Gerstein, M.; Edwards, A. M.; Arrowsmith, C. H. *Nat Struct Biol* **2000,** *7* (10), 903-9; (b) Goh, C. S.; Lan, N.; Douglas, S. M.; Wu, B.; Echols, N.; Smith, A.; Milburn, D.; Montelione, G. T.; Zhao, H.; Gerstein, M. *J Mol Biol* **2004,** *336* (1), 115-30; (c) Chandonia, J. M.; Kim, S. H.; Brenner, S. E. *Proteins* **2006,** *62* (2), 356-70; (d) Price, W. N.; Chen, Y.; Handelman, S. K.; Neely, H.; Manor, P.; Karlin, R.; Nair, R.; Liu, J. F.; Baran, M.; Everett, J.; Tong, S. C. N.; Forouhar, F.; Swaminathan, S. S.; Acton, T.; Xiao, R.; Luft, J. R.; Lauricella, A.; DeTitta, G. T.; Rost, B.; Montelione, G. T.; Hunt, J. F. *Nat Biotechnol* **2009,** *27* (1), 51-57; (e) Rodrigues, A.; Hubbard, R. E. *Brief Bioinform* **2003,** *4* (2), 150-67; (f) Canaves, J. M.; Page, R.; Wilson, I. A.; Stevens, R. C. *Journal of Molecular Biology* **2004,** *344* (4), 977-991; (g) Kantardjieff, K. A.; Rupp, B. *Bioinformatics* **2004,** *20* (14), 2162-2168; (h) Kantardjieff, K. A.; Jamshidian, M.; Rupp, B. *Bioinformatics* **2004,** *20* (14), 2171-2174; (i) Oldfield, C. J.; Ulrich, E. L.; Cheng, Y.; Dunker, A. K.; Markley, J. L. *Proteins* **2005,** *59* (3), 444-53; (j) Oldfield, C. J.; Xue, B.; Van, Y. Y.; Ulrich, E. L.; Markley, J. L.; Dunker, A. K.; Uversky, V. N. *Biochim Biophys Acta* **2013,** *1834* (2), 487-98.

[18] (a) Wang, H.; Feng, L.; Webb, G. I.; Kurgan, L.; Song, J.; Lin, D. *Brief Bioinform* **2017**, https://doi.org/10.1093/bib/bbx018; (b) Smialowski, P.; Frishman, D. *Methods Mol Biol* **2010,** *609*, 385-400.

[19] Smialowski, P.; Schmidt, T.; Cox, J.; Kirschner, A.; Frishman, D. *Proteins* **2006,** *62* (2), 343-55.

[20] Overton, I. M.; Barton, G. J. *FEBS Lett* **2006,** *580* (16), 4005-9.

[21] Chen, K.; Kurgan, L.; Rahbari, M. *Biochem Bioph Res Co* **2007,** *355* (3), 764-769.

[22] Slabinski, L.; Jaroszewski, L.; Rychlewski, L.; Wilson, I. A.; Lesley, S. A.; Godzik, A. *Bioinformatics* **2007,** *23* (24), 3403-5.

[23] Overton, I. M.; Padovani, G.; Girolami, M. A.; Barton, G. J. *Bioinformatics* **2008,** *24* (7), 901-907.

[24] Kurgan, L.; Razib, A. A.; Aghakhani, S.; Dick, S.; Mizianty, M.; Jahandideh, S. *Bmc Structural Biology* **2009,** *9*.

[25] Mizianty, M. J.; Kurgan, L. *Biochem Biophys Res Commun* **2009,** *390* (1), 10-5.

[26] Kandaswamy, K. K.; Pugalenthi, G.; Suganthan, P. N.; Gangal, R. *Protein Peptide Lett* **2010,** *17* (4), 423-430.

[27] Babnigg, G.; Joachimiak, A. *J Struct Funct Genomics* **2010,** *11* (1), 71-80.

[28] Mizianty, M. J.; Kurgan, L. *Bioinformatics* **2011,** *27* (13), i24-i33.

[29] Gao, J.; Hu, G.; Wu, Z.; Ruan, J.; Shen, S.; Hanlon, M.; Wang, K. *Current Bioinformatics* **2014,** *9* (1), 57-64.

[30] Wang, H.; Wang, M.; Tan, H.; Li, Y.; Zhang, Z.; Song, J. *PLoS One* **2014,** *9* (8), e105902.

[31] Wang, H.; Feng, L.; Zhang, Z.; Webb, G. I.; Lin, D.; Song, J. *Scientific reports* **2016,** *6*.

[32] Overton, I. M.; van Niekerk, C. A. J.; Barton, G. J. *Proteins-Structure Function and Bioinformatics* **2011,** *79* (4), 1027-1033.

[33] Jahandideh, S.; Mahdavi, A. *J Theor Biol* **2012,** *306*, 115-9.

[34] Mizianty, M. J.; Kurgan, L. A. *Protein Peptide Lett* **2012,** *19* (1), 40-49.

[35] Charoenkwan, P.; Shoombuatong, W.; Lee, H. C.; Chaijaruwanich, J.; Huang, H. L.; Ho, S. Y. *Plos One* **2013,** *8* (9).

[36] Mizianty, M. J.; Fan, X.; Yan, J.; Chalmers, E.; Woloschuk, C.; Joachimiak, A.; Kurgan, L. *Acta Crystallogr D Biol Crystallogr* **2014,** *70* (Pt 11), 2781-93.

[37] Jahandideh, S.; Jaroszewski, L.; Godzik, A. *Acta Crystallogr D* **2014,** *70*, 627-635.

[38] Hu, J.; Han, K.; Li, Y.; Yang, J.-Y.; Shen, H.-B.; Yu, D.-J. *Amino acids* **2016,** *48* (11), 2533-2547.

[39] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic acids research* **2000,** *28* (1), 235-242.

[40] Dondoshansky, I.; Wolf, Y. *NCBI, Bethesda, Md* **2002**.

[41] (a) Jazayeri, A.; Dias, J. M.; Marshall, F. H. *J Biol Chem* **2015,** *290* (32), 19489-95; (b) Fernandez-Ballester, G.; Fernandez-Carvajal, A.; Gonzalez-Ros, J. M.; Ferrer-Montiel, A. *Pharmaceutics* **2011,** *3* (4), 932-53; (c) Grey, J. L.; Thompson, D. H. *Expert Opin Drug Discov* **2010,** *5* (11), 1039-45.

[42] (a) Park, H.; Lee, H.; Seok, C. *Curr Opin Struct Biol* **2015,** *35*, 24-31; (b) Movshovitz-Attias, D.; London, N.; Schueler-Furman, O. *Proteins-Structure Function and Bioinformatics* **2010,** *78* (8), 1939-1949.

Received: xxx xx, 2017 Revised: xxx xx, 2017 Accepted: xxx xx, 2017