Contents lists available at ScienceDirect

Journal of Structural Biology

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

Prediction of protein crystallization outcome using a hybrid method

Frank H. Zucker [1], Christine Stewart [1], Jaclyn dela Rosa, Jessica Kim, Li Zhang, Liren Xiao, Jenni Ross, Alberto J. Napuli, Natascha Mueller, Lisa J. Castaneda, Stephen R. Nakazawa Hewitt, Tracy L. Arakaki, Eric T. Larson, Easwara Subramanian, Christophe L.M.J. Verlinde, Erkang Fan, Frederick S. Buckner, Wesley C. Van Voorhis, Ethan A. Merritt, Wim G.J. Hol *

Medical Structural Genomics of Pathogenic Protozoa (MSGPP), School of Medicine, University of Washington, Seattle, WA 98195-7742, United States

ARTICLE INFO

Article history:
Received 11 December 2009
Received in revised form 18 March 2010
Accepted 23 March 2010
Available online xxxx

Keywords:
Crystal growth
Protein characterization
Thermal shift assay
Dynamic light scattering
Limited proteolysis
Regression partition tree

ABSTRACT

The great power of protein crystallography to reveal biological structure is often limited by the tremendous effort required to produce suitable crystals. A hybrid crystal growth predictive model is presented that combines both experimental and sequence-derived data from target proteins, including novel variables derived from physico-chemical characterization such as R_{30}, the ratio between a protein's DSF intensity at 30 °C and at T_m. This hybrid model is shown to be more powerful than sequence-based prediction alone – and more likely to be useful for prioritizing and directing the efforts of structural genomics and individual structural biology laboratories.

1. Introduction

Detailed knowledge of protein and nucleic acid structures is of central importance for understanding life at its molecular and atomic level, and benefits human health by guiding design of therapeutics, vaccines and diagnostics. For decades protein crystallography has been the primary technique for obtaining structural information of biomacromolecules but, despite huge technical advances, obtaining crystals of good diffraction quality often remains a major bottleneck. Data from 17 structural genomics projects in TargetDB indicate that only 13% of soluble proteins yield crystals suitable for structure determination (Chayen and Saridakis, 2008). Protein crystallization is a complex, relatively poorly understood process driven by many thermodynamic, kinetic, and stochastic factors (Rupp and Wang, 2004). However, certain properties of a protein sample that are expected to impact crystallizability, e.g. homogeneity, solubility, stability and flexibility (Ericsson et al., 2006), can be characterized by biophysical methods available to most laboratories. Several of these methods, including dynamic light scattering (DLS) (D'Arcy, 1994), limited proteolysis (LP) (Gao et al., 2005), differential scanning fluorimetry (DSF) (Ericsson et al., 2006; Price et al., 2009) and size-exclusion chromatography (SEC) (Price et al., 2009; Graslund et al., 2008) assays have been suggested singly as predictors of success in crystal growth. However, there is still considerable scope for improvement in prediction of crystallization outcome (Rupp, 2003).

The wealth of data capturing the success or failure of crystallization attempts by large structural genomics efforts has provided a basis for analyses that attempt to correlate crystallization success with variables derived from amino acid sequence. Sequence-based variables such as size, hydrophobicity, and isoelectric point have long been used to predict solubility (Bertone et al., 2001), which appears to be inversely related to crystallizability (Price et al., 2009). In addition, newer algorithms examine additional variables such as homology to proteins in TargetDB (Slabinski et al., 2007; Jaroszewski et al., 2008), amino acid composition (Overton et al., 2008), co-location of amino acids (Chen et al., 2007; Kurgan et al., 2009), side chain entropy and buried glycines (Price et al., 2009). Significant limitations of such methods include reduced accuracy for proteins larger than 200 residues (Chen et al., 2007; Kurgan et al., 2009), reliance on availability of previously-studied homologs (Slabinski et al., 2007), or *a priori* assumptions about structure (Price et al., 2009). For example, the predictive value of

Abbreviations: DLS, dynamic light scattering; DS, diffraction score; DSF, differential scanning fluorimetry; HyXG-1, hybrid crystal growth prediction model-1; I_{30}, intensity at 30 °C in DSF; I_{Tm}, intensity at the inflection point of a DSF curve; LP, limited proteolysis; R_{30}, ratio of I_{30} to I_{Tm}; R_{MT}, ratio of the intensity of minor transition(s) to the total intensity transition in a DSF curve; SCE, side chain entropy; SEC, size-exclusion chromatography.

* Corresponding author. Fax: +1 206 685 7002.
E-mail address: wghol@u.washington.edu (W.G.J. Hol).

[1] These authors contributed equally to this work.

1047-8477/$ - see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.jsb.2010.03.016

homology appears to drop rapidly below 90% sequence identity (Jaroszewski et al., 2008). This is not surprising, given that changes to only a few residues may introduce or remove favorable protein:protein interaction surfaces that stabilize the formation of a crystal lattice. Indeed, deliberate introduction of small changes in sequence constitutes an established strategy for addressing difficulty in crystallization (Cooper et al., 2007; Klock et al., 2007). Variation in sequence, position and cleavage of affinity tags is also widely used to improve crystallization, an effect confirmed in this study (Supplementary Table 1a, e.g. for targets Cpar071490AAB and Tbru022584AAA).

A possible further concern is that a disproportionate number of structural genomics target sequences are derived from prokaryotic and archeal genomes, which may reduce the predictive power of TargetDB when applied to predicting the crystallizability of eukaryotic target proteins. Indeed, a recent sequence-based predictor of crystallization for expressed proteins did not have the same predictive power for overall success of human proteins (Price et al., 2009), an observation confirmed by our studies reported below.

Quantitative comparison of existing crystal growth prediction methods is difficult for several reasons including the fact that the criteria for judging a prediction as 'correct' varies (Price et al., 2009; Slabinski et al., 2007; Overton et al., 2008; Chen et al., 2007; Kurgan et al., 2009). In several cases only overall success from expression to crystal growth is scored [(Slabinski et al., 2007), $P_{\mathrm{XS-C-Hs}}$ in Price et al. (2009)], rather than distinguishing between success in protein expression and success in crystallization of purified protein. In the current paper we focus on the latter step.

The hypothesis underlying the current paper is that a more powerful approach to predicting crystallizability of a given protein sample is to combine sequence-derived information with multiple experiments that measure a range of biophysical properties of the actual sample to be crystallized. The reasoning is that multiple factors regarding the proteins sample under consideration determine jointly the success of a crystal growth experiment. Since during crystal growth protein–protein contacts need to be established, the nature of the surface of a protein is obviously of special importance. Hence in addition to the homogeneity and stability of individual folded proteins, it makes sense to consider (i) the average physico-chemical properties of the atoms making up the surface of the protein, such as charged versus uncharged, hydrophilic versus hydrophobic, etc.; (ii) the degree of deviations from that average, e.g. the flexibility of side chains, loops, motifs and domains; and (iii) the degree of uniformity in the association of the protein molecules in solution, i.e. whether or not the protein forms well-defined single chain entities or well-defined multi-chain particles.

Estimates of the nature and flexibility of exposed side chains can be derived from sequence information provided that a good prediction of which residues are at the surface can be obtained (Price et al., 2009). Flexible loops are the subject of several sequence-based prediction methods (Price et al., 2009; Slabinski et al., 2007), while limited proteolysis also gives information about the dynamics of surface loops (Hubbard, 1998). The mobility of motifs and domains of a protein with respect to each other is likely reflected in the accessibility of hydrophobic pockets measured by fluorescent probes which increase in quantum yield when the probe is shielded from the solvent, i.e. when the probe interacts with hydrophobic patches of the protein in DSF assays (Ericsson et al., 2006). Homogeneity of a protein sample with regards to aggregation state and impurities can be assessed by combining information from DLS measurements (D'Arcy, 1994; Niesen et al., 2008), SDS–PAGE and SEC (Kawate and Gouaux, 2006). These complementary classes of information should be considered together, as suggested by a survey of SPINE quality assessment data (Geerlof et al., 2006). Some of the parameters derived from sequence and from biophysical data might be overlapping. For example, it was reported that side chain entropy (SCE) could replace individual experimental measures of stability for predicting crystallization of expressed prokaryotic proteins in a recent predictor (Price et al., 2009). Therefore statistical methods are to be used to discover the best combination of parameters for optimal prediction of crystallization results.

We describe here the use of statistical analysis methods to develop a predictor of crystallization and diffraction quality that is based on several types of biophysical experiments combined with protein sequence analysis. New variables are derived for several of the biophysical measurements of protein solutions. The value of these variables is explored in combination with variables derived from sequence to find an optimal combination of variables for predicting the outcome of crystallization experiments. Although we expect that performance of the prediction model will continue to improve as larger training sets and additional categories of physical data are brought to bear, our current best hybrid crystal growth prediction model, HyXG-1, already demonstrates the power of this approach. In contrast to previous work (Price et al., 2009), the resultant hybrid crystal growth prediction method obtained, HyXG-1, is substantially better than methods based on sequence alone in predicting outcome for our validation set.

2. Methods

2.1. Protein expression and purification

Proteins were prepared by the SGPP consortium (Fan et al., 2008) (www.sgpp.org) and the MSGPP program project (www.msgpp.org) using N-terminal His$_6$ tags, NiNTA and size-exclusion chromatography as described previously (Mehlin et al., 2006; Arakaki et al., 2006). SGPP targets (as indicated in Supplementary Table 2) were cloned using the BG1861 vector giving an uncleavable tag. MSGPP targets were also cloned using AVA0421 with a cleavable tag. Thus three tag variants of each target were possible: the 8-residue uncleavable tag, the 21-residue uncleaved tag, or the 4-residue cleaved tag. The SGPP procedure for high-throughput soluble expression screening (Mehlin et al., 2006) was modified for MSGPP targets (as indicated in Supplementary Table 2) by the replacement of sonication with freezing at −80 °C and thawing in lysis buffer containing 0.04 g lysozyme, 0.5 g CHAPS, 0.2 g $MgCl_2(H_20)_6$ and 6 μL benzonate per 100 ml SGPP buffer (see below) with 30 mM imidazole. Proteins were stored in SGPP buffer (25 mM HEPES pH 7.25, 500 mM NaCl, 5% Glycerol) except where noted in Supplementary Table 4 and flash frozen (Deng et al., 2004) before further characterization and crystallization.

2.2. Experimental protein characterization

Protein samples were thawed and characterized in the following ways.

2.2.1. SDS–PAGE analysis

Samples were flash thawed in 30 °C water bath, DTT was added to 5 mM and samples were spun at 25,000 g at 4 °C for 30 min prior to sample dilution. SDS dye with 5% β-mercaptoethanol was added and samples were boiled at 90 °C for 4 min and then run on 8–16% Tris–HCl Ready gel (Bio-Rad).

2.2.2. Differential scanning fluorimetry curves

DSF curves were collected using an Opticon 2 real-time PCR detector (Bio-Rad) to measure the fluorescence of SYPRO Orange (Sigma) in the presence of protein at 0.5 mg/ml in SGPP buffer with 5 mM DTT in 96-well plates as the temperature increased from 20 or 30 to 90 °C in increments of 0.2 °C. Proteins were centrifuged for

Please cite this article in press as: Zucker, F.H., et al. Prediction of protein crystallization outcome using a hybrid method. J. Struct. Biol. (2010), doi:10.1016/j.jsb.2010.03.016

30 min at 25,000*g*, 4 °C before sample preparation. SYPRO Orange dye was diluted from initial concentration of "5000×" to "2.5×" in the final sample.

2.2.3. Limited proteolysis

Purified protein at 1 mg/ml in SGPP buffer + 5 mM $CaCl_2$ was exposed to 20 μg/ml trypsin, chymotrypsin, subtilisin A, or endoproteinase Glu-C for 0, 1 and 24 h. After each time period, the reaction was stopped with 0.17 M acetic acid and SDS dye was added. All samples were boiled and run on SDS–PAGE, gels were then stained with Coomassie Blue stain.

2.2.4. Dynamic light scattering

Measurements were made using DynaPro light scattering instrument (Protein Solutions Inc.). All samples were centrifuged 30 min at 4 °C and 25,000*g* immediately before the experiment in order to remove possible dust particles and diluted to 5–10 mg/ml in SGPP buffer + 5 mM DTT. Measurements were performed at 5 and 30 °C readings were taken for each sample.

2.3. Crystallization

Crystallization screening was performed at the Hauptman–Woodward Institute as previously described (Arakaki et al., 2006; Luft et al., 2003) and using the JCSG + Suite of screens (QIAGEN). After rapid thawing samples were centrifuged for 30 min at 25,000*g* at 4 °C to remove possible precipitate, and kept on ice afterwards until used in crystallization experiments. Crystallization leads from initial screens were optimized for pH, precipitant and additive concentrations as well as protein concentration and temperature. MSGPP crystallization trials were set up using a Phoenix crystallization robot (Art Robbins Instruments) using various commercially available screens. Each screen was set up at varying ratios of protein to reservoir volumes. Conditions for the best-diffracting crystals are shown in Supplementary Table 4.

2.4. Determination of diffraction quality

Suitable crystal cryoprotection solutions were determined as needed. Typically, a synthetic mother liquor was prepared that contained an increased amount of precipitants, salts, and/or additives relative to the crystallization solution, and was then diluted with varying concentrations of glycerol, ethylene glycol, low molecular weight polyethylene glycols (MW < 400 Da), or concentrated salt solutions. Crystals were subjected to the cryoprotection solution for varying amounts of time and in some cases had to be transferred gradually from low to high concentration of the cryoprotectant. On occasion, oils such as paratone-N, mineral oil, parfin oil, or mixtures were used for cryoprotection. Following cryoprotection (if needed), crystals were mounted in suitably-sized CryoLoops (Hampton Research) and flash frozen in liquid nitrogen and tested for diffraction at 100 K on our home X-ray source (Rigaku MM007HF, Saturn detector) or on various synchrotron beamlines (SSRL, ALS, and APS).

2.5. Quantification of experimental and sequence variables

2.5.1. Yield

Expression of soluble protein in high-throughput screens was evaluated from the staining of protein from the equivalent of ~8% of a 600 μL culture. $\mathbf{Yld_S}$ was scored on a scale from 1, no detectable soluble protein, to 5, extremely high soluble protein expression (Supplementary Fig. 2. A score of 5 indicates approximately 5 μg of protein from 48 μL of cultured cells or more, i.e. at least 100 mg/L. $\mathbf{Yld_M}$ is the total mass of protein sent from protein production to crystal screening and growth after large scale expression. Large scale expression was carried out using several different aeration methods and volumes were not consistently recorded, so this measure of yield is not normalized for volume of cell culture.

2.5.2. Size-exclusion chromatography

SEC curves obtained during protein purification were exported from PrimeView Evaluation (Amersham Pharmacia Biotech) and analysed using Microsoft Excel and gnuplot (http://gnuplot.sourceforge.net) as described by Kawate and Gouaux (2006). After fitting a linear background and a single Gaussian to the peak with the highest absorbance peak (Fig. 1a), we calculated the total residual R_{abs} in Excel as $R_{abs} = \Sigma|Y_{obs} - Y_{calc}|/\Sigma Y_{obs}$. We then iteratively fit additional Gaussians to the largest residual peaks (Fig. 1b and c) until a plateau in R_{abs} was reached (Fig. 1d). The Gaussian which gave maximal improvement in R_{abs} was taken as the last Gaussian in the optimal model. $\mathbf{SEC_{R1}}$ is R_{abs} with one Gaussian fit (Fig. 1a). $\mathbf{SEC_{PP}}$ is the percent purity of the pooled fractions using the optimal model (Fig. 1b).

2.5.3. SDS–PAGE analysis

Coomassie Blue-stained gels were scored visually on a scale of 1 (lowest purity) to 5 (highest purity); none of the samples scored below 3.

2.5.4. Differential scanning fluorimetry curves

In theory a protein undergoing a two-state unfolding transition (folded to unfolded with no stable intermediate states) should produce a sigmoid fluorescence intensity curve (Ericsson et al., 2006; Niesen et al., 2007):

$$I = I_{\min} + (I_{\max} - I_{\min})/(1 + e^{(T_m - T)/T_w})$$

Ideally, the change in intensity with temperature, d*I*/d*T*, should be maximal at T_m, the temperature at which half the protein is unfolded, also referred to as the melting point (Niesen et al., 2007). T_w is a measure of the width of the transition, proportional to the full width at half the maximal d*I*/d*T* (FWHM). To derive T_w, we calculated FWHM from the data (see Supplementary Methods) and divided this value by the constant $2^* \ln[(2+\sqrt{2})/(2-\sqrt{2})] \approx 3.525$.

In practice the intensity curve for most of the samples in our study followed a sigmoid curve near T_m but deviated in one or more ways at other temperatures. We therefore used the simple estimate of T_m as the temperature at $(\mathrm{d}I/\mathrm{d}T)_{max}$ to avoid dependence on deviations, and quantified the deviations separately. Deviations included high initial intensity, which we quantified as $\boldsymbol{R_{30}}$ (Fig. 2b and d); multiple transitions with increasing intensity, quantified as R_{MT} (Fig. 2c and Supplementary Fig. 1c, right side); and a decrease in intensity at high temperature, seen in all samples. In the cases of samples with multiple transitions, the transition with the highest d*I*/d*T* always had the highest total change in intensity. We therefore assumed that the major intensity transition represented the major unfolding step, or at least the step in which the plurality of hydrophobic pockets were exposed to dye. We took the midpoint in that major unfolding step as T_m rather than attempting to fit a single sigmoid curve to data showing a multi-step transition, or attempting to determine the midpoint of a multi-step transition.

We quantified minor transitions (Fig. 2c and Supplementary Fig. 1c, right) as R_{MT}, the fraction of intensity change observed outside the major transitions. We fit the above equation to observed intensities at T_m and $T_m - 2T_{w_*}$ to find I_{min}, estimated the major transition intensity ΔI_{main} as $2^*(I_{Tm} - I_{min})$, and calculated R_{MT} as the ratio of the remaining intensity change to the intensity of the major transition (see Supplementary Methods for details). In cases such as Fig. 2d, the major positive transition was dwarfed by the

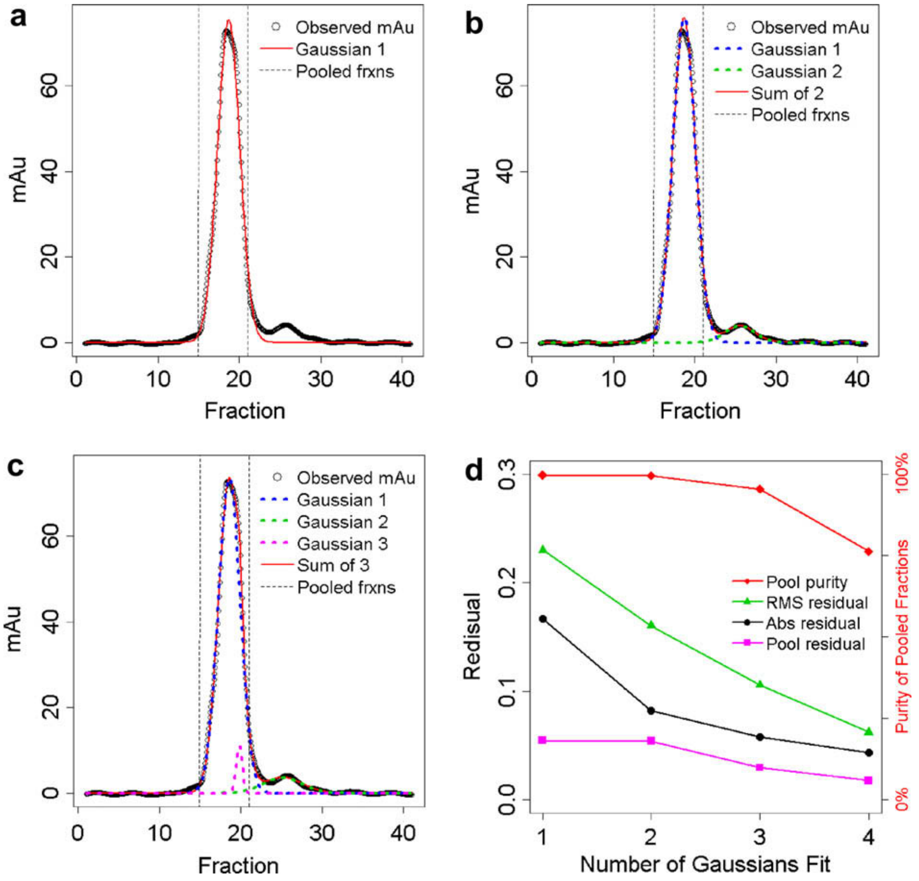

Fig. 1. Analysis of size-exclusion chromatography profiles. Gaussian peaks fit to the SEC curve for *Entamoeba histolytica* aspartate-tRNA ligase batch 24,058. In (a), (b) and (c) open black circles are observed absorbance at 280 nm in milli-absorbance units (mAu); vertical dashes bound the fractions pooled for further characterization and crystallization; red line is calculated mAu using a linear background plus 1, 2 or 3 Gaussian curves fit to the observed mAu using gnuplot. In (b) and (c) dotted lines in blue, green and violet show individual Gaussians. (A 4th Gaussian, not shown, can be fit as another small curve under the main peak.) (d) Residuals and calculated pool purity for fitting 1–4 Gaussians to observed mAu. Left axis: solid black circles, total R_{abs}, the absolute value of the difference between observed and calculated mAu divided by the total observed mAu; magenta squares, R_{abs} for the pooled fractions; green triangles, root mean square of the residuals as a fraction of the mean. Right axis: red diamonds, purity of the pooled fractions i.e. the maximum area under a single Gaussian in the pooled fractions divided by the total pool area. $\mathbf{SEC_{R1}}$ is R_{abs} for one Gaussian: i.e. the area between the red and black curves in (a) over the area under the black curve. For this sample SEC_{R1} = 0.16. $\mathbf{SEC_{PP}}$ is the purity of the pooled fractions calculated in the optimal model. For this sample SEC_{PP} = 0.99 from (b). (Figures prepared in the *R* statistical environment.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

overall negative slope of the curve; here, R_{MT} approached its maximum of 1 while R_{30} was between 1 and its maximum of 2.

Low-temperature fluorescence was quantified using the intensity at 30 °C since this temperature was consistently included in the temperature range of DSF experiments performed in our laboratory. We calculated $\boldsymbol{R_{30}}$ as I_{30}/I_{Tm}, the ratio of the intensity at 30 °C to the intensity at T_m (Fig. 2b), with intensity measured in arbitrary units from the minimum value for each curve. For an ideal sigmoid curve, I_{Tm} would be equal to $I_{max}/2$. For real curves, the intensity decrease at high *T* made it difficult to directly observe I_{max}; I_{Tm} was less sensitive to this common deviation from the ideal. For curves with multiple positive transitions (Fig. 2c, Supplementary Fig. 2c right), using I_{Tm} as the denominator to determine R_{30} gave similar results in most cases to using the overall positive intensity change (ΔI_{total}). Using I_{Tm} resulted in a substantially lower R_{30} compared to using the estimated intensity change of the main transition (ΔI_{main} as described above). In all cases, the ratio using I_{Tm} had the strongest correlation with crystallization outcome.

For curves with overall downward trends (Fig. 2d), any of these denominators ($I_{Tm,}$ I_{total} or I_{main}) would lead to extremely high ratios. Since the intensity was minimal and still dropping at the highest temperature used, the values and thus the ratio of I_{30} and I_{Tm} depended on the highest temperature used. Setting the baseline to the minimum intensity before T_m would have avoided this effect. However, the ratio was still so high in all such cases that this effect did not significantly alter the resulting model or predictions made using R_{30}. Further, this effect was quantified as a high R_{MT} value. In pathological cases where the intensity at 30 °C was far greater than the intensity at T_m, we assigned an arbitrary maximum value of 2 for R_{30}.

In most cases we had at least two measurements of the sample in standard buffer. The average of all valid values was used. Curves with no positive slope above 0.001 raw intensity units per degree were not included in averaging. This threshold is 0.0002 units per 0.2° increment, twice the Opticon Monitor's precision in reporting intensity of 4 decimal places. One sample had no curves with any positive slope; this sample was given arbitrary values of 0 for T_m and T_w, 2 for R_{30} and 1 for R_{MT}.

2.5.5. Limited proteolysis

Each protease was scored visually on a scale of 1–5 (most stable) according to the criteria in Supplementary Table 3, and the scores for the 4 proteases were averaged to calculate $\mathbf{LP_{av}}$.

2.5.6. Dynamic light scattering

Hydrodynamic radius (R_H), polydispersity, intensity and fraction of mass in each peak were recorded. For each sample a

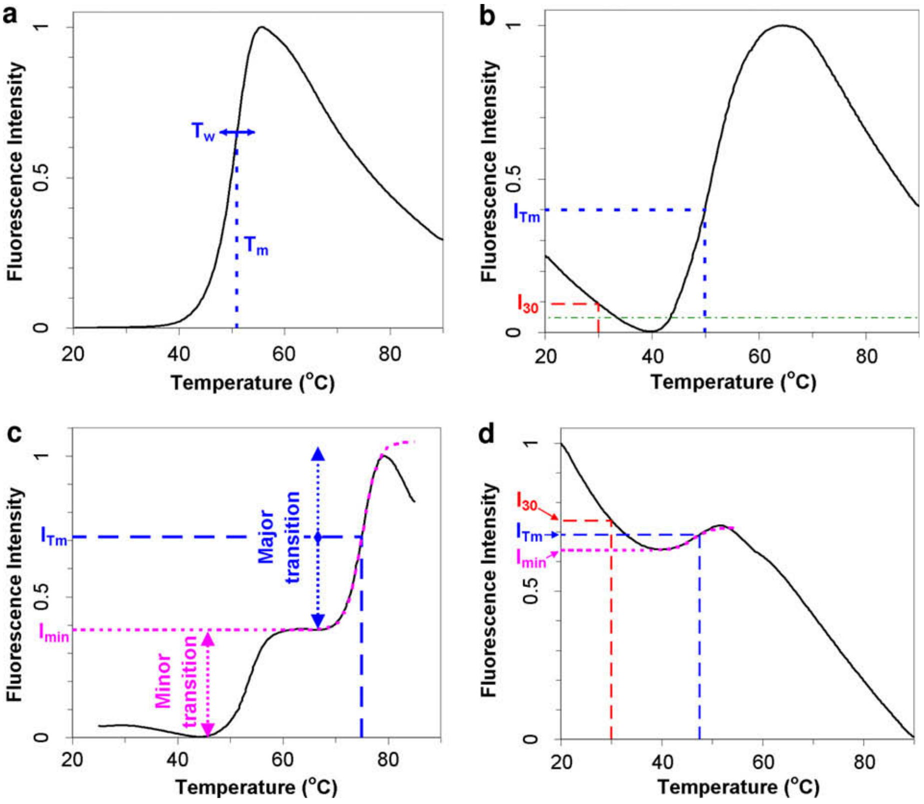

Fig. 2. Analysis of differential scanning fluorimetry curves. Four protein samples illustrate different curve shapes. Black solid lines: fluorescence intensity of SYPRO Orange dye vs. temperature, smoothed over 15 points (3 °C) and normalized to the minimum and maximum observed intensities. Blue dashed vertical lines: T_m, the temperature with the steepest positive slope, $(dI/dT)_{max}$. Blue horizontal dashes: $\boldsymbol{I_{Tm}}$, the intensity at T_m. (a) *Leishmania guyanensis* 6-phosphogluconolactonase with ideal shape: low intensity at low temperature and a single transition. Blue horizontal arrow: temperature range over which the slope is at least ½ of $(dI/dT)_{max}$ i.e. full width at half maximum (FWHM) of the derivative, proportional to the melting transition width $\boldsymbol{T_w}$. (b) *E. histolytica* aspartate-tRNA ligase batch 21,516 with high intensity at low temperature and a single transition. Red horizontal dashes: $\boldsymbol{I_{30}}$, intensity at 30 °C. $\boldsymbol{R_{30}}$ is the ratio of I_{30} to I_{Tm}. Green dot-dash line: I_{30} threshold based on the R_{30} criterion in the decision tree, Fig. 3b, i.e. $I_{30}/I_{Tm} = 0.105$. (c) *Toxoplasma gondii* porphobilinogen synthase amino acids 320–658, with two distinct transitions. Magenta dotted line: sigmoid curve fit to observed intensity at T_m and at $2{\cdot}T_w$ below T_m. At low temperatures this curve approaches $\boldsymbol{I_{min}}$, the estimated starting intensity of the major transition. Since in many cases intensity decays above T_m, and in others a minor transition is seen above T_m, the amplitude of the major transition is estimated as twice the intensity change between I_{min} and I_{Tm}. When there is a minor transition below T_m as in this case, I_{min} is also used as an estimate of the amplitude of that minor transition. R_{MT}, the transition fraction, is calculated as the amplitude of the minor transition(s) over the total amplitude of all transitions. (d) *L. major* methionyl-tRNA synthetase, amino acids 206–747, with high R_{30} and high R_{MT}. Both I_{30}, red dashes, and I_{min} from the curve fit to the transition, magenta dots, are near I_{Tm}, blue dashes. (Figures prepared in Excel.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dominant peak was chosen as the consistent peak with the highest fraction of mass. $\mathbf{DLS_P}$ was assigned as the polydispersity of that peak. $\mathbf{DLS_I}$ was calculated as the intensity of that peak over the total intensity of that peak and all peaks with larger R_H. Smaller peaks were assumed to be salts and other small molecules. $\mathbf{DLS_{MW}}$ was derived from R_H for that peak according to the formula from the Dynamics Version 5 software: $DLS_{MW} = (1.68 \times R_H)^{2.3398}$. $\mathbf{DLS_{MR}}$ is the ratio of DLS_{MW} to the molecular weight of the monomer calculated from the sequence of the expressed protein. An additional categorical score $\mathbf{DLS_{SC}}$ was assigned: 4 (<30% polydispersity in a single major peak), 3 (⩾30% polydispersity in a single major peak, or 2 (more than one peak, regardless of polydispersity); none of the proteins in this study were in category 1 (unmeasurable).

2.5.7. Sequence variables

We explored a limited set of parameters derived directly from the protein sequence: **MW**, calculated molecular weight of the monomer; $\mathbf{HYD_{av}}$, average hydropathy using Kyte and Doolittle values (1982); $\mathbf{Dis_{max}}$, number of amino acids in the longest contiguous stretch of disorder predicted by DisEMBL (Linding et al., 2003) (http://dis.embl.de/); $\mathbf{Dis_{-t}}$, longest stretch of predicted disorder excluding the N-terminal His tag; and **XP**, the score of 1–5, optimal to difficult, from XtalPred, a predictor based on 9 sequence parameters (http://ffas.burnham.org/XtalPred-cgi/xtal.pl) (Slabinski et al., 2007). Other summary metrics such as P_{XS} and $P_{C\text{–}XS\text{–}Hs}$ (Price et al., 2009) were also tested but did not contribute to the predictive power of the models.

2.6. Statistical analysis

2.6.1. Development of predictive model

Predictive models were constructed and tested in the *R* statistical environment (http://www.R-project.org) version 2.8.0. For recursive regression partition trees, parameters were tuned using leave-one-out cross-validation on the training set to optimize predictive power for biophysically valid trees. For SVM, variables were selected using 10-fold cross-validation on the training set by cycles of incremental variable addition and automated combinatorial surveys; parameters were retuned after each round of variable selection.

2.6.2. Analysis of predictive model

Predictive power for regression models was measured by $\mathbf{DS_{Pred}}$ **error,** the root mean squared error = $\sqrt{[\Sigma(O\text{–}P)^2/N]}$ where *O* and *P* are observed and predicted diffraction scores, respectively; by Pearson's correlation coefficient, and by area under the ROC curve of true positive rate versus false positive rate. Since *P* and *O* had

Please cite this article in press as: Zucker, F.H., et al. Prediction of protein crystallization outcome using a hybrid method. J. Struct. Biol. (2010), doi:10.1016/j.jsb.2010.03.016

bimodal rather than normal distributions, probability of observed correlations were estimated using synthetic data. For binary classifications Matthews correlation coefficient, accuracy, sensitivity and selectivity were also measured. Standard deviations for measures of predictive power were calculated using cross-validation results and synthetic data. See Supplementary Methods for further details on model development and analysis.

3. Results

3.1. Quantification of experimental and sequence variables

We considered 107 eukaryotic protein samples (Supplementary Tables 1 and 2, Supplementary Fig. 1) originating from the Structural Genomics of Pathogenic Protozoa (SGPP; www.sgpp.org) and Medical Structural Genomics of Pathogenic Protozoa (MSGPP; www.msgpp.org) pipelines, described in Supplementary Methods. This sample set includes both widely divergent genes and minor sequence variations, and represents the full range of diffraction outcomes, from failure to crystallize to diffraction better than 2 Å resolution. The full set was divided into a training set of 77 samples and a test set of 30 samples, such that the two sets contained similar distributions of crystallization outcome. The training set contained 41 sequences with less than 90% sequence identity to each other. Training set samples with similar sequences but distinct experimental characteristics and outcomes included multiple batches of the same sequence, tag variants, truncations, and homologs from related organisms. All 30 sequences in the test set had less than 85% identity to other proteins in either set.

We derived and quantified 21 experimental and sequence variables based on biophysical characterizations using SDS–PAGE, SEC, DSF, DLS and LP (Table 1). Novel quantitative measures were developed for SEC profiles, DSF curves and LP gels as described in Figs. 1 and 2 and Supplementary Table 3. Crystallization outcome, ranging from 0 to 6, was quantified as diffraction score (DS): no mountable protein crystals after extensive crystal screening (DS = 0), no diffraction (DS = 1), diffraction worse than 10 Å (DS = 2), 10 Å or better (DS = 3), 4 Å or better (DS = 4), 2.8 Å or better (DS = 5), or 2.0 Å or better (DS = 6).

3.2. Development of best predictive model

Many statistical methods can in principle be used to develop predictive models based on experimental and sequence variables (Fig. 3a). We evaluated linear regression, naïve Bayesian, several varieties of support vector machines (SVM), clustering, and recursive regression partition trees as described in Supplementary Methods. Regression partitioning and SVM gave the best results in cross-validation tests using only training data (Supplementary Results). However, regression partitioning gave the best results in predicting test set diffraction scores of the protein samples and will therefore be discussed here further.

3.3. Analysis of hybrid experimental characterization and sequence model

The best partition tree (Fig. 3b, hereafter also called the HyXG-1 tree) obtained from consideration of all 21 variables (Table 1) applies four experimental and two sequence criteria. Experimental variables used in the model are: (i) the ratio of intensity at 30 °C to intensity at the melting point in differential scanning fluorimetry curves (R_{30}); (ii) soluble protein expression level in high-throughput screening ($\mathrm{Yld_S}$); (iii) residual after fitting one Gaussian to a SEC curve ($\mathrm{SEC_{R1}}$); and (iv) ratio of molecular weight from hydrodynamic radius to calculated weight of the monomer ($\mathrm{DLS_{MR}}$), while, in addition, sequence variables incorporated into the model are: (v) calculated monomer molecular weight (MW) in Daltons; and (vi) number of amino acids in the longest disordered region predicted by DisEMBL (Linding et al., 2003) ($\mathrm{Dis_{max}}$). The model predicts good diffraction for samples with low MW (i.e. monomer under 36.3 kDa) and low R_{30} (i.e. I_{30}/I_{Tm} less than 0.105), but poor outcomes for samples with low MW and high

Table 1
Experimental and sequence variables tested.

Source	Variable	Description (see Supplementary Methods for full definitions)	Range[a]	Mean (SD)[b]	Correlation[c]
Protein production	**$\mathbf{Yld_S}$**	Score for soluble expression screening gels	1–5	3.4 (1.0)	0.16
	$\mathrm{Yld_M}$	Total mass of protein produced (mg)	>0	52 (39)	0.18
SDS–PAGE	SDS	Average of 4 visual scores; reducing conditions	1–5	4.4 (0.6)	−0.01
Limited proteolysis	**$\mathbf{LP_{av}}$**	Average of scores for 4 proteases	1–5	3.3 (0.9)	0.39
Size-exclusion chromatography	$\mathrm{SEC_{hu}}$	Visual scoring of chromatogram image	1–5	3.4 (1.0)	0.08
	$\mathbf{SEC_{R1}}$	Residual (R_{abs}) with 1 Gaussian fit, as fraction of total area	0–1	0.4 (0.3)	−0.11
	$\mathbf{SEC_{PP}}$	Percent purity of pooled fractions at plateau of R_{abs}	0–1	0.8 (0.2)	−0.17
Dynamic light scattering	$\mathrm{DLS_P}$	Percent polydispersity	0–100	23 (14)	−0.09
	$\mathrm{DLS_I}$	Percent intensity in major peak	0–100	92 (11)	0.05
	$\mathrm{DLS_{SC}}$	Composite score: 4, $\mathrm{DLS_P} \leqslant 30$ and $\mathrm{DLS_I} = 100$; 3, $\mathrm{DLS_P} > 30$ and $\mathrm{DLS_I} = 100$; 2, $\mathrm{DLS_I} < 100$	2–4	2.6 (0.8)	0.19
	$\mathbf{DLS_{MW}}$	MW calculated from hydrodynamic radius (kDa)	>0	190 (332)	−0.01
	$\mathbf{DLS_{MR}}$	MW from hydrodynamic radius/predicted monomer MW	>0	4 (7)	0.04
Differential scanning fluorimetry	T_{m}	Melting temperature (°C) or 0 if no valid melting point	20–90	53 (10)	0.08
	TW	Melting width (°C)	⩾0	7 (3)	0.07
	$\boldsymbol{R_{30}}$	Ratio of intensity at 30 °C to intensity at T_{m}	0–2	0.4 (0.5)	−0.37
	R_{MT}	Fraction of intensity change in other transitions	−1 to 1	0.28 (0.24)	−0.31
Sequence analysis	**MW**	Predicted molecular weight of monomer including tag (Da)	>0	49 K (16 K)	−0.34
	$\mathbf{Hyd_{av}}$	Average hydropathy (GRAVY)	±4.5	−0.32 (0.14)	0.05
	$\mathbf{Dis_{max}}$	Longest stretch of disordered residues	⩾0	19 (9)	−0.19
	$\mathrm{Dis_{-t}}$	Longest stretch of disorder excluding N-terminal tag	⩾0	8 (8)	−0.07
	XP	Score from XtalPred web server	1–5	3.4 (1.3)	−0.23

Large, bold variables are those used in partition trees in Table 2.
[a] Range of possible values.
[b] Mean (and standard deviation) of values for training set of 77 samples.
[c] Correlation of training set values to diffraction score.

Please cite this article in press as: Zucker, F.H., et al. Prediction of protein crystallization outcome using a hybrid method. J. Struct. Biol. (2010), doi:10.1016/j.jsb.2010.03.016

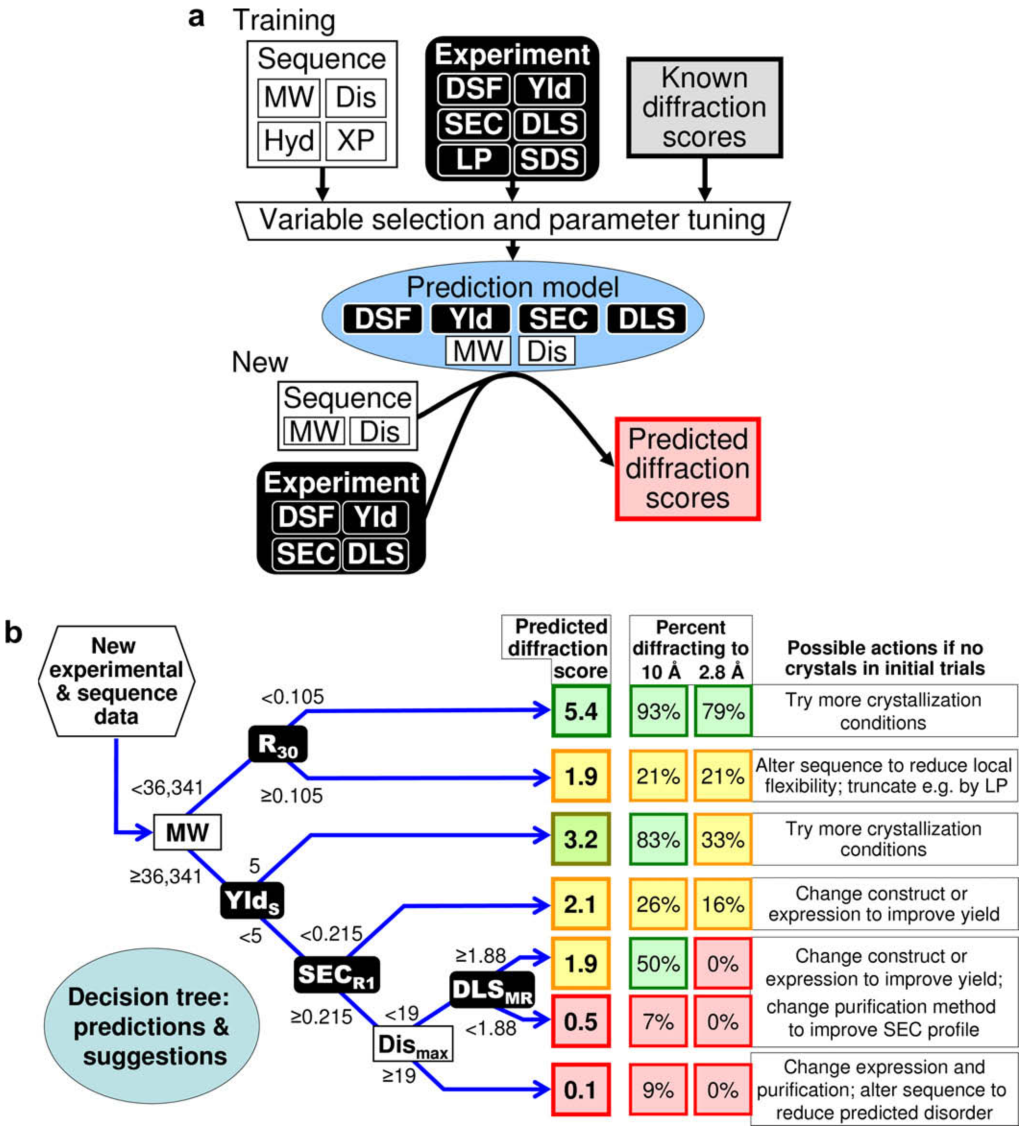
a
Training
Sequence
MW
Dis
Hyd
XP
Experiment
DSF
Yld
SEC
DLS
LP
SDS
Known diffraction scores
Variable selection and parameter tuning
Prediction model
DSF
Yld
SEC
DLS
MW
Dis
New
Sequence
MW
Dis
Experiment
DSF
Yld
SEC
DLS
Predicted diffraction scores
b
New experimental & sequence data
<36,341
MW
≥36,341
R30
<0.105
≥0.105
Ylds
5
<5
SECR1
<0.215
≥0.215
Dismax
<19
≥19
DLSMR
≥1.88
<1.88
Decision tree: predictions & suggestions
Predicted diffraction score
Percent diffracting to
10 Å
2.8 Å
Possible actions if no crystals in initial trials
5.4
93%
79%
Try more crystallization conditions
1.9
21%
21%
Alter sequence to reduce local flexibility; truncate e.g. by LP
3.2
83%
33%
Try more crystallization conditions
2.1
26%
16%
Change construct or expression to improve yield
1.9
50%
0%
0.5
7%
0%
Change construct or expression to improve yield; change purification method to improve SEC profile
0.1
9%
0%
Change expression and purification; alter sequence to reduce predicted disorder

Fig. 3. Development of diffraction predictor using experimental results and sequence. (a) Predictive model design. (Top) train the model on experimental and sequence data and known crystallization outcomes quantified as diffraction scores (DS). (Bottom) use the model to predict DS for new samples from new experimental and sequence data. (b) Hybrid crystal growth predictor (HyXG-1) decision tree prediction trained on 77 samples: start with experimental and sequence data for a new protein sample (top left); travel to the right across the tree branching according to criteria shown; arrive at the predicted DS for each category (center). Predicted DS is the mean DS for all training samples in that category; from top to bottom, there were 9, 7, 10, 14, 7, 12 and 18 training samples in each category. To the right are the percent of all test and training samples in each category diffracting to at least 10 Å or at least 2.8 Å, and suggestions for actions if no crystals are seen in initial trials. Possible changes include: change construct tag, tag placement or promoter; change expression host, scale-up volume, aeration method, or time and temperature regime; change purification columns (e.g. add ion exchange), tag cleavage, lysis and column buffers, or final concentration step.

R_{30}. Moderate outcomes are predicted for samples with high MW and very high Yld_S scores (over 100 mg/L soluble expression in HT screening). Poor outcomes are predicted for other high MW samples, with slightly better outcomes for samples with low SEC_{R1} (less than 21.5% of A_{280} outside a single Gaussian curve) or with low Dis_{max} (fewer than 19 amino acids in the longest stretch of predicted disorder) and high DLS_{MR} ($MW_{RH}/MW_{monomer}$ greater than 1.88).

The predictive power of this HyXG-1 tree was evaluated by applying the model to the test set of 30 samples (Fig. 4 and Table 2 row A). With success defined as 2.8 Å or better diffraction (DS ⩾ 5), 25 samples (83%) were correctly predicted. With success defined as better than 10 Å diffraction (DS > 3, dotted line in Fig. 4a), 26 samples were correctly predicted, 6 as successful, 20 as unsuccessful. The resulting Matthews correlation coefficient is 0.67; selectivity is high, 20/21 = 95%; sensitivity is moderate, 6/9 = 67%; and the overall accuracy of the prediction model is high, 26/30 = 87%. For comparison, the highest Matthews correlation coefficient on our test set using previously reported sequence-only predictors (Price et al., 2009; Slabinski et al., 2007) was 0.48, with an accuracy of 60%.

3.4. Relative importance of experimental and sequence variables

In order to test the relative importance of two classes of variables, those from experimental results and those from sequence analysis, new decision trees based on only one of the two classes were constructed. First, we considered only those variables of one class that contributed to the best hybrid tree. Next, we constructed trees from all variables of one class from the full set of 21 variables. In each case we used the same parameters and training set as for the best hybrid tree. There is a substantial increase in predictive power of the best

Please cite this article in press as: Zucker, F.H., et al. Prediction of protein crystallization outcome using a hybrid method. J. Struct. Biol. (2010), doi:10.1016/j.jsb.2010.03.016

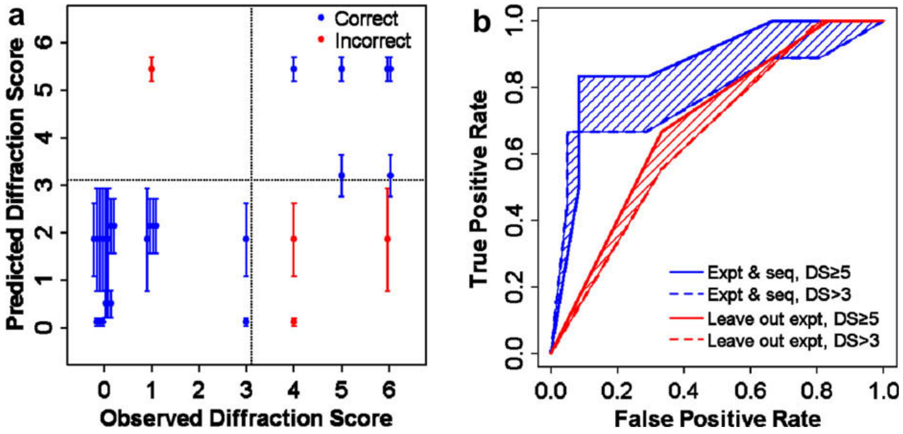

Fig. 4. Diffraction score predictions using experimental results and sequence. (a) DS observed vs. DS predicted by the HyXG-1 model shown in (3b) for the test set of 30 new samples. DS is: 0, no mountable protein crystals after extensive crystal screening; 1, no diffraction; 2, diffraction worse than 10 Å; 3, 10–4.01 Å diffraction; 4, 4.80–2.81 Å diffraction; 5, 2.80–2.01 Å diffraction; 6, 2.00 Å or better diffraction. Bars: ±1 standard deviation based on the deviation of training DS. Dotted lines and coloring based on success threshold of better than 10 Å (DS > 3). (b) Receiver operating characteristic (ROC) curves: area under curve is a measure of predictive power. Blue lines, predictions from combined experimental and sequence data (Table 2, row A); red, predictions leaving out experimental data (row C). Dashes, ROC curve for success threshold of better than 10 Å (DS > 3); solid, success threshold of 2.8 Å or better (DS ⩾ 5). Shading added to visually clarify the association of lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Effects of experimental and sequence variables on prediction power.

Model	Variables used in prediction model									DS_{Pred} error[f]	Correlation[g]	ROC area[h]	
	Experimental variables					Sequence variables						DS > 3	DS ⩾ 5
A. Best with expt. & seq.[a]	R_{30}	Yld_S	SEC_{R1}	DLS_{MR}		MW	Dis_{max}			1.96 (0.13)	0.56 (0.06)	0.77 (0.04)	0.87 (0.05)
B. Leave out seq. from A[b]	R_{30}	(Yld_S)	SEC_{R1}	DLS_{MR}						2.73 (0.08)	−0.07 (0.06)	0.61 (0.05)	0.49 (0.06)
C. Leave out expt. from A[c]						MW	Dis_{max}			2.46 (0.10)	0.18 (0.07)	0.65 (0.05)	0.69 (0.06)
D. Best with expt. only[d]	R_{30}	Yld_S	SEC_{PP}[d]	DLS_{MW}[d]	LP_{av}[d]					1.90 (0.06)	0.57 (0.04)	0.70 (0.08)	0.71 (0.08)
E. Best with seq. only[e]						MW	Dis_{max}	Hyd_{av}[e]	XP[e]	2.58 (0.12)	0.17 (0.08)	0.64 (0.05)	0.63 (0.06)

For descriptions of variables see Table 1.

[a] Best partition model combining experimental and sequence variables from 77-sample training set.

[b] The 4 experimental variables from model A were supplied to the partition algorithm. The algorithm discarded Yld_S as a criterion.

[c] The 2 sequence variables from A were supplied to the algorithm; the algorithm used both as criteria.

[d] All experimental variables were supplied. The algorithm used 2 of the same variables as in A, replaced SEC_{R1} and DLS_{MR} with related variables SEC_{PP} and DLS_{MW}, and added LP_{av}.

[e] All sequence variables were supplied; hydropathy (Hyd_{av}) and XtalPred score (XP) were added to the sequence variables used in A.

[f] Three measures of predictive power for the 30-sample test set (parentheses: standard deviation estimated from synthetic data). Square root of the mean square difference between predicted and observed diffraction scores (DS).

[g] Three measures of predictive power for the 30-sample test set (parentheses: standard deviation estimated from synthetic data). Pearson's correlation coefficient for predicted and observed DS.

[h] Three measures of predictive power for the 30-sample test set (parentheses: standard deviation estimated from synthetic data). Area under ROC curves as in Fig. 4b, with success defined as "better than 10 Å diffraction" (DS > 3) or as "2.8 Å or better diffraction" (DS ⩾ 5).

hybrid tree compared to trees without experimental variables (Fig. 4b and Table 2, row A compared to C or E). For example, the correlation rose from 0.18 ($p > 0.16$) to 0.56 ($p < 0.0014$) with the addition of experimental variables. The improvement in predictive power is more than twice the estimated standard deviation for prediction error, for correlation and also for the area under the receiver operating characteristic (ROC) curve with a diffraction score cutoff of DS ⩾ 5 (Fig. 4b). Interestingly, the error and correlation for the best experiment-only tree (Table 2, row D) were significantly better than the best sequence-only tree (Table 2, row E).

4. Discussion

The HyXG-1 decision tree suggested by recursive regression partition (Fig. 3b) is consistent with correlations of individual protein characteristics to crystallization found in previous work (Ericsson et al., 2006; Price et al., 2009; Slabinski et al., 2007; Kawate and Gouaux, 2006) and in this study (Table 1). For instance, low initial intensity followed by a sharp increase on melting in DSF has been reported as favorable for crystallization (Ericsson et al., 2006). High fluorescence intensity at 30 °C indicates existence of hydrophobic pockets, possibly due to flexibility of loops, secondary structure elements or motifs, in which the fluorophore can bind. Upon increasing the temperature, unfolding of the environment of these pockets may lead to increased exposure of the fluorophore to the surrounding solvent and concomitant decreased fluorescence intensity. When the temperature is sufficiently high to initiate unfolding of one or more major domains, an increase in fluorescence intensity is observed when new binding sites for the fluorophore become available. Determining the precise mechanism leading to high R_{30} is beyond the scope of this paper, but it appears

Please cite this article in press as: Zucker, F.H., et al. Prediction of protein crystallization outcome using a hybrid method. J. Struct. Biol. (2010), doi:10.1016/j.jsb.2010.03.016

from our analysis that R_{30} quantifies a property of proteins which is more significant than the T_m, which might be due to the fact that R_{30} reports on features of the target protein at a temperature generally closer to the conditions of crystallization than T_m.

Though the DSF properties of some proteins are sensitive to buffer conditions (Vedadi et al., 2006), results in our lab (unpublished) and others (Lavinder et al., 2009; Yeh et al., 2006; Jarvest et al., 2003) suggest that for many proteins DSF results are consistent across a variety of buffers and protein concentrations. This may partially explain why characterization experiments done in one buffer have considerable power in predicting crystallization, even though crystallization conditions essentially always differ from any buffer used to test solution properties of the protein (Supplementary Table 4).

While it is not clear precisely what roles overall protein stability and local flexibility play in crystallization (Price et al., 2009), low predicted disorder has been shown to be important for crystallographic success (Price et al., 2009; Slabinski et al., 2007). High predicted stability, moderate fraction of predicted loops and no long stretches of predicted disorder were favorable for crystallization in one set of mostly prokaryotic proteins (Slabinski et al., 2007). In another set of proteins, no predictive power was seen for either experimentally measured overall stability or limited proteolysis which may monitor loop flexibility, but low predicted disorder was important for success in crystallizing soluble prokaryotic proteins and also in expressing and crystallizing soluble eukaryotic proteins (Price et al., 2009). These finding are in agreement with our results showing that proteins with smaller predicted disordered regions (low Dis_{max}) tend to crystallize better.

Most proteins require relatively pure solutions to crystallize. Gaussian SEC profiles indicate homogeneous protein solutions, or at least homogeneity of protein size. In some cases, protein crystallization requires SEC profiles close to Gaussian (Kawate and Gouaux, 2006). Our measure of SEC_{R1} quantifies the purity of the protein sample in terms of hydrodynamic radius, which reflects the homogeneity of monomer or oligomer size and shape. A value of SEC_{R1} less than 0.215 is incorporated in the partition tree obtained (Fig. 3b).

Our DLS_{MR} threshold near 2 in the partition tree is consistent with the finding that dimers and oligomers are favored for crystallization over monomers (Price et al., 2009). Other DLS-derived variables do not contribute to predictive power, possibly because the properties they measure were already accounted for by other variables used in the model. Our samples did not show the strong negative correlation between multidispersity and well-diffracting crystals seen in other work (Niesen et al., 2008). The Yld_S criterion of the decision tree is consistent with the high success rate observed in our structural genomics work for proteins that express very well, probably due to the relative ease of selecting highly purified fractions from purification columns (unpublished results). Thus for the decision tree from regression partitioning on combined experimental and sequence variables, the criteria are plausible given the known and expected correlates of those biophysical properties.

The reason why combined consideration of several variables enhances prediction of crystallization outcome is likely due to the fact that multiple factors play a role in determining the success in crystal growth. The molecular weight criterion in the predicting partition tree might reflect that larger proteins tend to contain multiple domains some of which may have a tendency to be flexible with respect to each other. R_{30} from DSF experiments likely indicate a degree of flexibility of loops, motifs and domains. The symmetry of sizing chromatographic peaks is related to the homogeneity of the molecular species in the sample and its state of oligomerization. Long stretches of amino acids that are predicted to be disordered decrease the likelihood of forming regular crystal contacts. From the results obtained it appears that the well-crystallizing protein tends to be – in general – one with homogenous particle size, stable folding at 30 °C, and few flexible domains, motifs or loops.

The analysis presented here was necessarily limited to protein samples for which full biophysical characterization data was available. Despite this relatively small set as compared to the number of targets available for sequence-only analysis, it is clear that joint consideration of multiple experimental variables in addition to sequence significantly improves prediction of crystallization and diffraction (Table 2), yielding higher accuracy than previously reported for methods based on sequence alone (Price et al., 2009; Slabinski et al., 2007; Overton et al., 2008). The improved predictive power gained by joint consideration of multiple experimental variables stands in contrast to relatively poor correlation with success reported for single experimental measures (Price et al., 2009). It is quite possible that incorporating other experimental methods such as mass spectroscopy (Jeon et al., 2005), NMR data (Page et al., 2005) and static light scattering (Wilson, 2003), may further increase the predictive power of hybrid models.

The HyXG-1 hybrid predictor may be most useful in cases where proteins fail to crystallize on initial setup and the prediction is strongly positive or negative. The prediction can then help investigators prioritize their efforts towards an increased likelihood of success in producing diffracting crystals (Fig. 3b, right side). For instance, if the protein sample prepared has a high R_{30} and a molecular weight less than 36 kDa, strategies to lower the R_{30} are likely to be most effective. This might be achieved in several ways such as removing flexible termini by limited proteolysis; or by designing, cloning and expressing new truncations of the protein; or by switching to other species which contain fewer stretches of predicted disorder; or by replacing flexible segments by shorter linkers or by domains of known structure with little disorder.

We are developing a web site which will provide researchers with tools for assigning standardized quantitative descriptions to their experimental results, and for using these results to predict crystallization outcome and prioritize further efforts. Researchers will be invited to upload sets of protein characterizations and crystallization outcomes to help improve the predictive model by increasing the number of samples in the training set and adding new experimental methods to be considered.

5. Conclusion

We have developed a set of novel variables derived from biophysical data. Several of these such as R_{30} and DLS_{MR} appear to be useful in predicting crystallization outcome. A predictive hybrid model, combining multiple biophysical characterization and sequence-derived data, such as the HyXG-1 decision tree derived by regression partition (Fig. 3b), is more powerful than sequence-based prediction alone – and therefore likely to be useful in guiding crystallization efforts.

6. Author contributions

Database development and statistical analysis: Frank H. Zucker, Christine Stewart, Wim G.J. Hol.

Bioinformatics and target selection: Frank H. Zucker, Christophe Verlinde, Easwara Subramanian, Fred Buckner.

Protein production: Alberto J. Napuli, Natascha Mueller, Lisa J. Castaneda, Stephen Nakazawa Hewitt, Wesley C. Van Voorhis.

Protein characterization: Jaclyn dela Rosa, Jessica Kim, Li Zhang, Liren Xiao, Jenni Ross, Alberto J. Napuli, Natascha Mueller, Lisa J. Castaneda, Stephen Nakazawa Hewitt.

Please cite this article in press as: Zucker, F.H., et al. Prediction of protein crystallization outcome using a hybrid method. J. Struct. Biol. (2010), doi:10.1016/j.jsbi.2010.03.016

Protein crystallization: Jaclyn dela Rosa, Jessica Kim, Li Zhang, Liren Xiao, Jenni Ross, Wim G. J. Hol.

X-ray data collection: Tracy Arakaki, Eric Larson, Ethan Merritt.

Project coordination: Erkang Fan, Wim G.J. Hol.

Manuscript writing: Frank H. Zucker, Christine Stewart, Ethan Merritt, Wim G. J. Hol.

Acknowledgments

We thank Drs. Chris Mehlin and Juergen Bosch for their contributions to this work in protein production and structure determination, respectively. We are indebted to Thomas E. Kammeyer for assistance in deciphering the Opticon Monitor 3 raw data format for DSF data. Funded by NIH Grants Nos. GM64655, GM088518 and P01AI067921.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2010.03.016.

References

Chayen, N.E., Saridakis, E., 2008. Protein crystallization: from purified protein to diffraction-quality crystal. Nat. Methods 5, 147–153.

Rupp, B., Wang, J., 2004. Predictive models for protein crystallization. Methods 34, 390–407.

Ericsson, U.B., Hallberg, B.M., Detitta, G.T., Dekker, N., Nordlund, P., 2006. Thermofluor-based high-throughput stability optimization of proteins for structural studies. Anal. Biochem. 357, 289–298.

D'Arcy, A., 1994. Crystallizing proteins – a rational approach? Acta Crystallogr. D Biol. Crystallogr. 50, 469–471.

Gao, X., Bain, K., Bonanno, J.B., Buchanan, M., Henderson, D., Lorimer, D., Marsh, C., Reynes, J.A., Sauder, J.M., Schwinn, K., Thai, C., Burley, S.K., 2005. High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. J. Struct. Funct. Genomics 6, 129–134.

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

Graslund, S., Sagemark, J., Berglund, H., Dahlgren, L.G., Flores, A., Hammarstrom, M., Johansson, I., Kotenyova, T., Nilsson, M., Nordlund, P., Weigelt, J., 2008. The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. Protein Expr. Purif. 58, 210–221.

Rupp, B., 2003. High-throughput crystallography at an affordable cost: the TB structural genomics consortium crystallization facility. Acc. Chem. Res. 6, 173–181.

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T., Gerstein, M., 2001. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res. 29, 2884–2898.

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

Jaroszewski, L., Slabinski, L., Wooley, J., Deacon, A.M., Lesley, S.A., Wilson, I.A., Godzik, A., 2008. Genome pool strategy for structural coverage of protein families. Structure 16, 1659–1667.

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

Chen, K., Kurgan, L., Rahbari, M., 2007. Prediction of protein crystallization using collocation of amino acid pairs. Biochem. Biophys. Res. Commun. 355, 764–769.

Kurgan, L., Razib, A.A., Aghakhani, S., Dick, S., Mizianty, M., Jahandideh, S., 2009. CRYSTALP2: sequence-based protein crystallization propensity prediction. BMC Struct. Biol. 9, 1–15.

Cooper, D.R., Boczek, T., Grelewska, K., Pinkowska, M., Sikorska, M., Zawadzki, M., Derewenda, Z., 2007. Protein crystallization by surface entropy reduction: optimization of the SER strategy. Acta Crystallogr. D Biol. Crystallogr. 63, 636–645.

Klock, H.E., Koesema, E.J., Knuth, M.W., Lesley, S.A., 2007. Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. Proteins Struct. Funct. Bioinform. 71, 982–994.

Hubbard, S., 1998. The structural aspects of limited proteolysis of native proteins. Biochim. Biophys. Acta – Protein Struct. Mol. Enzymol. 1382, 191–206.

Niesen, F.H., Koch, A., Lenski, U., Harttig, U., Roske, Y., Heinemann, U., Hofmann, K.P., 2008. An approach to quality management in structural biology: biophysical selection of proteins for successful crystallization. J. Struct. Biol. 162, 451–459.

Kawate, T., Gouaux, E., 2006. Fluorescence-detection size-exclusion chromatography for pre-crystallization screening of integral membrane proteins. Structure 14, 673–681.

Geerlof, A., Brown, J., Coutard, B., Egloff, M.P., Enguita, F.J., Fogg, M.J., Gilbert, R.J., Groves, M.R., Haouz, A., Nettleship, J.E., Nordlund, P., Owens, R.J., Ruff, M., Sainsbury, S., Svergun, D.I., Wilmanns, M., 2006. The impact of protein characterization in structural proteomics. Acta Crystallogr. D Biol. Crystallogr. 62, 1125–1136.

Fan, E., Baker, D., Fields, S., Gelb, M.H., Buckner, F.S., Van Voorhis, W.C., Phizicky, E., Dumont, M., Mehlin, C., Grayhack, E., Sullivan, M., Verlinde, C., Detitta, G., Meldrum, D.R., Merritt, E.A., Earnest, T., Soltis, M., Zucker, F., Myler, P.J., Schoenfeld, L., Kim, D., Worthey, L., Lacount, D., Vignali, M., Li, J., Mondal, S., Massey, A., Carroll, B., Gulde, S., Luft, J., Desoto, L., Holl, M., Caruthers, J., Bosch, J., Robien, M., Arakaki, T., Holmes, M., Le Trong, I., Hol, W.G., 2008. Structural genomics of pathogenic protozoa: an overview. In: John, J. (Ed.), Methods in Molecular Biology, Structural Proteomics – High-throughput Methods, vol. 426. Humana Press, Inc., Totawa, NJ, pp. 497–513.

Mehlin, C., Boni, E., Buckner, F.S., Engel, L., Feist, T., Gelb, M.H., Haji, L., Kim, D., Liu, C., Mueller, N., Myler, P.J., Reddy, J.T., Sampson, J.N., Subramanian, E., Van Voorhis, W.C., Worthey, E., Zucker, F., Hol, W.G., 2006. Heterologous expression of proteins from Plasmodium falciparum: results from 1000 genes. Mol. Biochem. Parasitol. 148, 144–160.

Arakaki, T., Le Trong, I., Phizicky, E., Quartley, E., DeTitta, G., Luft, J., Lauricella, A., Anderson, L., Kalyuzhniy, O., Worthey, E., Myler, P.J., Kim, D., Baker, D., Hol, W.G.J., Merritt, E.A.M., 2006. Structure of Lmaj006129AAA, a hypothetical protein from Leishmania major. Acta Cryst. F62, 175–179.

Deng, J., Davies, D.R., Wisedchaisri, G., Wu, M., Hol, W.G.J., Mehlin, C., 2004. An improved protocol for rapid freezing of protein samples for long-term storage. Acta Crystallogr. D Biol. Crystallogr. 60, 203–204.

Luft, J.R., Collins, R.J., Fehrman, N.A., Lauricella, A.M., Veatch, C.K., DeTitta, G.T., 2003. A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J. Struct. Biol. 142, 170–179.

Niesen, F.H., Berglund, H., Vedadi, M., 2007. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat. Protoc. 2, 2212–2221.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B., 2003. Protein disorder prediction: implications for structural proteomics. Structure 11, 453–459.

Vedadi, M., Niesen, F.H., Allali-Hassani, A., Fedorov, O.Y., Finerty Jr., P.J., Wasney, G.A., Yeung, R., Arrowsmith, C., Ball, L.J., Berglund, H., Hui, R., Marsden, B.D., Nordlund, P., Sundstrom, M., Weigelt, J., Edwards, A.M., 2006. Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. Proc. Natl. Acad. Sci. USA 103, 15835–15840.

Lavinder, J.J., Hari, S.B., Sullivan, B.J., Magliery, T.J., 2009. High-throughput thermal scanning: a general, rapid dye-binding thermal shift screen for protein engineering. J. Am. Chem. Soc. 131, 3794–3795.

Yeh, A.P., McMillan, A., Stowell, M.H., 2006. Rapid and simple protein-stability screens: application to membrane proteins. Acta Crystallogr. D Biol. Crystallogr. 62, 451–457.

Jarvest, R.L., Berge, J.M., Brown, M.J., Brown, P., Elder, J.S., Forrest, A.K., Houge-Frydrych, C.S., O'Hanlon, P.J., McNair, D.J., Rittenhouse, S., Sheppard, R.J., 2003. Optimisation of aryl substitution leading to potent methionyl tRNA synthetase inhibitors with excellent gram-positive antibacterial activity. Bioorg. Med. Chem. Lett. 13, 665–668.

Jeon, W.B., Aceti, D.J., Bingman1, Craig A., Vojtik1, Frank C., Olson1, Andrew C., Ellefson1, Jason M., McCombs1, Janet E., Sreenath1, Hassan K., Blommel1, Paul G., Seder1, Kory D., Burns1, Brendan T., Geetha1, Holalkere V., Harms1, Amy C., Sabat1, Grzegorz, Sussman1, Michael R., Fox1, Brian G., Phillips Jr., George N., 2005. High-throughput purification and quality assurance of arabidopsis thaliana proteins for eukaryotic structural genomics. J. Struct. Funct. Genomics 6, 143–147.

Page, R., Peti, W., Wilson, I.A., Stevens, R.C., Wüthrich, K., 2005. NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. PNAS 102, 1901–1905.

Wilson, W., 2003. Light scattering as a diagnostic for protein crystal growth—A practical approach. J. Struct. Biol. 142, 56–65.

Please cite this article in press as: Zucker, F.H., et al. Prediction of protein crystallization outcome using a hybrid method. J. Struct. Biol. (2010), doi:10.1016/j.jsb.2010.03.016