Prediction of protein crystallization outcome using a hybrid method


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

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 stoichiometric 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...
homology appears to drop rapidly below 90% sequence identity (Jaroszewska 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. 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). Variations 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 CparO71490AAB and TbruO22584AAA).

A possible further concern is that a disproportionate number of structural genomics target sequences are derived from prokaryotic and archael 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), 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 multiplex 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 sets of data. Hubbard (1998) describes the dynamics of surface loops, while limited proteolysis also provides information about the stability and 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 SGRP consortium (Fan et al., 2008) (www.sgrp.org) and the MSGGP program project (www.msggp.org) using N-terminal His6 tags, NINTA and size-exclusion chromatography as described previously (Mehlin et al., 2006; Arakaki et al., 2006). SGRP targets (as indicated in Supplementary Table 2) were cloned using the BG1861 vector giving an uncleavable tag. MSGGP targets were also cloned using AAVA0421 with a cleavable tag. Thus three tag variants of each target were possible: the 8-residue uncleavable tag, the 21-residue uncleavable tag, or the 4-residue cleaved tag.

The SGRP procedure for high-throughput soluble expression screening (Mehlin et al., 2006) was modified for MSGGP 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 MgCl2(H2O)6 and 6 μl benzene per 100 ml SGRP buffer (see below) with 30 mMimidazole. Proteins were stored in SGRP 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 Optricon 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 SGRP buffer with 5 mM DTT in 96-well plates as the temperature increased from 20 to 30 or 30 to 90 °C in increments of 0.2 °C. Proteins were centrifuged for...
257 30 min at 25,000g, 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

258 Purified protein at 1 mg/ml in SGPP buffer + 5 mM CaCl₂ was exposed to 20 μg/ml trypsin, chymotrypsin, subtilisin A, or endo-

259 proteinase 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.

260 All samples were boiled and run on SDS–PAGE, gels were then stained with Coomassie Blue stain.

2.2.4. Dynamic light scattering

261 Measurements were made using DynaPro light scattering instrument (Protein Solutions Inc.). All samples were centrifuged 30 min at 4°C and 25,000g 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

262 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).

263 After rapid thawing samples were centrifuged for 30 min at 25,000g 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

264 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 cryo-

265 protectant. 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

266 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. Yld₅₀ 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. Yld₅₀ 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

267 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 \( R_{\text{abs}} \) in Excel as \( R_{\text{abs}} = \Sigma |Y_{\text{obs}} - Y_{\text{calc}}|/\Sigma Y_{\text{obs}} \). We then iteratively fit additional Gaussians to the largest residual peaks (Fig. 1b and c) until a plateau in \( R_{\text{abs}} \) was reached (Fig. 1d). The Gaussian which gave maximal improvement in \( R_{\text{abs}} \) was taken as the last Gaussian in the optimal model. \( SEC_{\text{G1}} \) is \( R_{\text{abs}} \) with one Gaussian fit (Fig. 1a).

2.5.3. SDS–PAGE analysis

268 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

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

\[
I = I_{\text{min}} + (I_{\text{max}} - I_{\text{min}})/(1 + e^{(T - T_{w})/T_{w}})
\]

Ideally, the change in intensity with temperature, \( dI/dT \), should be maximal at \( T_{\text{m}} \), 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 dI/dT (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 \( T_{\text{m}} \) but deviated in one or more ways at other temperatures. We therefore used the simple estimate of \( T_{\text{m}} \) as the temperature at which \( (dI/dT)_{\text{max}} \) to avoid dependence on deviations, and quantified the deviations separately. Deviations included high initial intensity, which we quantified as \( R_{\text{G1}} \) (Fig. 2b and d); multiple transitions with increasing intensity, quantified as \( R_{\text{G2}} \) (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 dI/dT 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 \( T_{\text{m}} \) 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 \( R_{\text{MT}} \), the fraction of intensity change observed outside the major transitions. We fit the above equation to observed intensities at \( T_{\text{m}} \) and \( T_{\text{m}} - 2 \Delta T_{\text{m}} \) to find \( I_{\text{min}} \), estimated the major transition intensity \( \Delta I_{\text{max}} \) as \( 2/(I_{\text{m}} - I_{\text{min}}) \), and calculated \( R_{\text{MT}} \) 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

overall negative slope of the curve; here, \( R_{30} \) 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 \( R_{30} \) 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 (\( \Delta I_{total} \)). Using \( I_{Tm} \) resulted in a substantially lower \( R_{30} \) compared to using the estimated intensity change of the main transition (\( \Delta I_{main} \) 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 \( LP_{av} \).

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

dominant peak was chosen as the consistent peak with the highest fraction of mass. **DLS** was assigned as the polydispersity of that peak. **DLS** 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. **DLS** was derived from \( R_H \) for that peak according to the formula from the Dynamics Version 5 software: \( DLS = (1.68 \times R_H)^{2.3398} \).

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; **HYD**, average hydropathy using Kyte and Doolittle values (1982); **Dismax**, number of amino acids in the longest contiguous stretch of disorder predicted by DisEMBL (Linding et al., 2003) (http://disembl.de); **Dis**, 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 \( P_S \) and **P** 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 returned after each round of variable selection.

2.6.2. Analysis of predictive model

Predictive power for regression models was measured by **DSS** error, the root mean squared error = \( \sqrt{\frac{\sum(O-P)^2}{N}} \) 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 \( O \) and \( P \) had
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 other proteins in either set. All 30 sequences in the test set had 85% identity to other proteins in either set. The training and test sets had less than 85% identity to other proteins in either set.

We derived and quantified 21 experimental and sequence variables tested 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 outcome (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).

Table 1

<table>
<thead>
<tr>
<th>Source</th>
<th>Variable</th>
<th>Description (see Supplementary Methods for full definitions)</th>
<th>Range</th>
<th>Mean (SD)</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein production</td>
<td>YldS</td>
<td>Score for soluble expression screening gels</td>
<td>1–5</td>
<td>3.4 (1.0)</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>YldM</td>
<td>Total mass of protein produced (mg)</td>
<td>&gt;0</td>
<td>52 (39)</td>
<td>0.18</td>
</tr>
<tr>
<td>SDS–PAGE</td>
<td>SDS</td>
<td>Average of 4 visual scores; reducing conditions</td>
<td>1–5</td>
<td>4.4 (0.6)</td>
<td>-0.01</td>
</tr>
<tr>
<td>Limited proteolysis</td>
<td>LPavar</td>
<td>Average of scores for 4 proteases</td>
<td>1–5</td>
<td>3.3 (0.9)</td>
<td>0.39</td>
</tr>
<tr>
<td>Size-exclusion chromatography</td>
<td>SECpar</td>
<td>Visual scoring of chromatogram image</td>
<td>1–5</td>
<td>3.4 (1.0)</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>SECini</td>
<td>Residual (Rini) with 1 Gaussian fit, as fraction of total area</td>
<td>0–1</td>
<td>0.4 (0.3)</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>SECPref</td>
<td>Percent purity of pooled fractions at plateau of Rini</td>
<td>0–1</td>
<td>0.8 (0.2)</td>
<td>-0.17</td>
</tr>
<tr>
<td>Dynamic light scattering</td>
<td>DLSi</td>
<td>Percent polydispersity</td>
<td>0–100</td>
<td>23 (14)</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>DLS1</td>
<td>Percent intensity in major peak</td>
<td>0–100</td>
<td>92 (11)</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>DLSNC</td>
<td>Composite score; 4, DLSi ≤ 30 and DLS1 = 100; 3, DLSi &gt; 30 and DLS1 = 100; 2, DLSi &lt; 100</td>
<td>2–4</td>
<td>2.6 (0.8)</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>DLS500</td>
<td>MW calculated from hydrodynamic radius (kDa)</td>
<td>&gt;0</td>
<td>190 (332)</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>DLSmax</td>
<td>MW from hydrodynamic radius/predicted monomer MW</td>
<td>&gt;0</td>
<td>4 (7)</td>
<td>0.04</td>
</tr>
<tr>
<td>Differential scanning fluorimetry</td>
<td>Tm</td>
<td>Melting temperature (°C) or 0 if no valid melting point</td>
<td>20–90</td>
<td>53 (10)</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>TW</td>
<td>Melting width (°C)</td>
<td>≥0</td>
<td>7 (3)</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Rint</td>
<td>Ratio of intensity at 30 °C to intensity at Tm</td>
<td>0–2</td>
<td>0.4 (0.5)</td>
<td>-0.37</td>
</tr>
<tr>
<td></td>
<td>Rintex</td>
<td>Fraction of intensity change in other transitions</td>
<td>−1 to 1</td>
<td>0.28 (0.24)</td>
<td>-0.31</td>
</tr>
<tr>
<td>Sequence analysis</td>
<td>MW</td>
<td>Predicted molecular weight of monomer including tag (Da)</td>
<td>&gt;0</td>
<td>49 k (16 k)</td>
<td>-0.34</td>
</tr>
<tr>
<td></td>
<td>Hydav</td>
<td>Average hydrophathy (GRAVY)</td>
<td>≥4.5</td>
<td>-0.32</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Dismax</td>
<td>Longest stretch of disordered residues</td>
<td>≥0</td>
<td>19 (9)</td>
<td>-0.19</td>
</tr>
<tr>
<td></td>
<td>Dismax</td>
<td>Longest stretch of disorder excluding N-terminal tag</td>
<td>≥0</td>
<td>8 (8)</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td>XP</td>
<td>Score from XtalPred web server</td>
<td>1–5</td>
<td>3.4 (1.3)</td>
<td>-0.23</td>
</tr>
</tbody>
</table>

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.

Moderate outcomes are predicted for samples with high MW and very high YldS 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_1 (less than 21.5% of A_280 outside a single Gaussian curve) or with low Dismax (fewer than 19 amino acids in the longest stretch of predicted disorder) and high DLSMR (MWRH/MWmonomer 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\textsuperscript{P} >= 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, 20/31 = 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...
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

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables used in prediction model</th>
<th>Sequence variables</th>
<th>DSmax error</th>
<th>Correlation</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Best with expt. &amp; seq.</td>
<td>R_S0, YldS, SEC_R, DLS</td>
<td>MW, DSmax</td>
<td>1.96</td>
<td>0.56</td>
<td>0.77, 0.87</td>
</tr>
<tr>
<td>B. Leave out seq. from A</td>
<td>R_S0, YldS, SEC_R, DLS</td>
<td>MW, DSmax</td>
<td>2.73</td>
<td>0.07</td>
<td>0.61, 0.49</td>
</tr>
<tr>
<td>C. Leave out expt. from A</td>
<td>R_S0, YldS, SEC_R, DLS</td>
<td>MW, DSmax</td>
<td>2.46</td>
<td>0.18</td>
<td>0.65, 0.69</td>
</tr>
<tr>
<td>D. Best with expt. only</td>
<td>R_S0, YldS, SEC_R, DLS</td>
<td>MW, DSmax, LPav</td>
<td>1.90</td>
<td>0.57</td>
<td>0.70, 0.71</td>
</tr>
<tr>
<td>E. Best with seq. only</td>
<td>R_S0, YldS, SEC_R, DLS</td>
<td>MW, DSmax, Hydav, XP</td>
<td>2.58</td>
<td>0.17</td>
<td>0.64, 0.63</td>
</tr>
</tbody>
</table>

For descriptions of variables see Table 1.

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 $R_{50}$ is beyond the scope of this paper, but it appears
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., 2005). These finding are in agreement with our results showing that proteins with smaller predicted disordered regions (low $D_{\text{dismax}}$) 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$_{\text{rat}}$ 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$_{\text{rat}}$ less than 0.215 is incorporated in the partition tree obtained (Fig. 3b).

Our DLS$_{\text{exp}}$ 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 $Y_{3}$, 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 plausibly given the known and expected correlates of those biological 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$_{\text{exp}}$ 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.
Protein production: Alberto J. Napuli, Natasha 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, Natasha Mueller, Lisa J.
Castaneda, Stephen Nakazawa Hewitt.

Protein crystallization: Jaclyn de la Rosa, Jessica Kim, Li Zhen, Li-ren 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.


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. Kammer 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


