Sequence analysis

ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction

Ian M. Overton\textsuperscript{1}, Gianandrea Padovani\textsuperscript{2}, Mark A. Girolami\textsuperscript{2} and Geoffrey J. Barton\textsuperscript{1,*}

\textsuperscript{1}School of Life Sciences Research, University of Dundee, Dow Street, Dundee, DD1 5EH and \textsuperscript{2}Department of Computing Science, University of Glasgow, Glasgow, GL12 8QQ, UK

Received on June 1, 2007; revised on January 21, 2008; accepted on February 6, 2008

ABSTRACT

The ability to rank proteins by their likely success in crystallization is useful in current Structural Biology efforts and in particular in high-throughput Structural Genomics initiatives. We present ParCrys, a Parzen Window approach to estimate a protein’s propensity to produce diffraction-quality crystals. The Protein Data Bank (PDB) provided training data whilst the databases TargetDB and PepcDB were used to define feature selection data as well as test data independent of feature selection and training. ParCrys outperforms the OB-Score, SECRET and CRYSTALP on the data examined, with accuracy and Matthews correlation coefficient values of 79.1\% and 0.582, respectively (74.0\% and 0.227, respectively, on data with a ‘real-world’ ratio of positive:negative examples). ParCrys predictions and associated data are available from www.compbio.dundee.ac.uk/parcrys.

Contact: geoff@compbio.dundee.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Knowledge of protein three-dimensional structure is useful for addressing and formulating key scientific questions, as well as stimulating technological innovation (Hol, 2000; Todd et al., 2005). For example, structure-based drug design has informed the process of developing treatments for human diseases such as HIV (Poppe et al., 1997), influenza (von Itzstein et al., 1993), asthma (Schuttelkopf et al., 2006) and cancer (Davies et al., 2002). Structural insights have revealed molecular mechanisms of cellular function (Shapiro and Harris, 2000; Zarembinski et al., 1998), pathogenesis (Diprose et al., 2001; Singh et al., 2006) and genetic disease (Yard et al., 2007). In addition, the structures of proteins from thermophilic organisms suggest a promising basis for the development of industrial biocatalysts (Hol, 2000).

Structural genomics is a global enterprise that aspires to produce a large scale mapping of protein structure space (Burley et al., 1999; Chandonia and Brenner, 2006; Hol, 2000, http://www.isgo.org/home/index.php; Service, 2002, Stevens et al., 2001). However, it is common for only 5\% of selected protein targets to result in a high-resolution protein model (http://www.ebi.ac.uk/msd-srv/msdtarget; Service, 2002, 2005; Terwillinger, 2000). Various strategies have been proposed to increase the return from structural genomics efforts, including obtaining one representative structure per protein family and working with multiple orthologues (Brenner, 2000; Chandonia and Brenner, 2005; Hiu and Edwards, 2003; Liu et al., 2004; Savchenko et al., 2003). These approaches imply methods to rank proteins within functionally defined groups according to their likely success in the structural genomics pipeline.

As crystallization is a significant bottleneck in structural biology, the relative ease of obtaining diffraction-quality crystals is a key consideration (Biertumpfel et al., 2005; Chanyen, 2004; Puesy et al., 2005). For example, to date, the Structural Proteomics in Europe (SPINE) consortium reports 58\% success for the progression from ‘Expressed’ to ‘Soluble’ and 53\% success for the progression from ‘Purified’ to ‘Crystallized’ (http://www.ebi.ac.uk/msd-srv/msdtarget). Investigations into the relationship between protein sequence properties and progression through the structural genomics pipeline have suggested features relevant to crystallization propensity prediction (Canaves et al., 2004; Goh et al., 2004). Indeed, there are already methods that aim to predict crystallization propensity. These include SECRET (Smialowski et al., 2006), the OB-Score (Overton and Barton, 2006) and more recently, CRYSTALP (Chen et al., 2007). A limitation of both SECRET and CRYSTALP is that they only accept as input sequences of length between 46 and 200 amino acids. While the OB-score has no such limitation, it only exploits two predictive features to estimate crystallization propensity: isoelectric point and hydrophobicity. In this article we describe a Parzen Window density estimator that incorporates additional features not considered by the OB-Score, whilst permitting prediction over the full range of sequence length. The Parzen Window technique affords a non-parametric statistical approach to estimate a density function from a given sample dataset (Duda and Hart, 1973; Parzen, 1962). We took sequences from the Protein Data Bank (PDB) (Berman et al., 2007) culled at 25\% identity, with R factor of $\leq$0.3 and resolution $\leq$3.0 Å downloaded from PISCES (Wang and Dunbrack, 2003) as the model for the Parzen Window density estimate. The propensity of a test protein to progress to the stage of diffraction-quality crystals was estimated according to the density function derived from the PDB sequences;
no negative examples are required for this. In independent tests, we found that this new method (ParCrys) outperforms the OB-Score, SECRET and CRYSTALP.

2 METHODS

2.1 Datasets

Selection of appropriate training and test datasets is critical in the development of a predictor. In this study we developed 13 primary datasets, which were filtered by stringent sequence similarity thresholds in order to minimize bias. The Parzen Window density estimate was always based on the PDB data. Feature selection was conducted using TargetDB datasets. Additional datasets for method evaluation were created to be independent of the PDB and feature selection data. The datasets are summarized in Table 1. Details of the dataset creation is as are follows.

2.1.1 Training and feature selection datasets The Parzen Window density estimate was derived from a dataset of 3958 PDB (Berman et al., 2007) structures with a resolution of at least 3.0 Å, maximum R-factor of 0.3 and filtered at 25% sequence identity (PDB3958), obtained from the PISCES server (Wang and Dunbrack, 2003) on August 4, 2006.

For feature selection, positive and negative datasets DIF728 and WS6025, respectively, were taken from TargetDB as described previously (Overton and Barton, 2006). Briefly, DIF728 comprises 728 sequences that gave diffraction-quality crystals, and WS6025 comprises 6025 sequences where work had been stopped before crystals were obtained. The negative feature selection dataset WS728 of the same size as DIF728 was created by randomly selecting 728 sequences from WS6025. The WS728 and DIF728 pair is referred to as FEAT. The WS6025 and DIF728 pair is referred to as FEAT-W, which was used only to define a ParCrys threshold for predicting over ‘real-world’ distributions.

2.1.2 Independent test datasets Datasets were constructed to provide a test independent of the training and feature selection data. The test datasets were made to be both non-redundant and not to show similarity to each other or to the FEAT and PDB3958 sets, according to stringent thresholds. This was achieved by filtering with PSIBLAST (Altschul et al., 1997), using HMMER to cluster against Pfam (Eddy, 1998; Finn et al., 2006) and finally Z-score clustering with AMPS (Barton and Sternberg, 1987) as explained below.

In order to provide a search database for PSIBLAST, the training and feature selection datasets (PDB3958, DIF728 and WS6025) were combined with UniRef50 (Apweiler et al., 2004), low-complexity filtered using SEG (Wan and Wootton, 2000) and helixfilt (D. Jones personal communication) to produce the database DEVIEL_U50. The positive test dataset, representing proteins that readily crystallize, was based on TargetDB (Chen et al., 2004) sequences downloaded on April 17, 2007 and with a date stamp later than April 2006; the date filtering aims to result in 614 sequences (T_NEG614). In order to safeguard against overlap between the positive and negative test datasets, it was necessary to filter T_NEG614 against T_POS72. This step aims to eliminate sequences where work was stopped due to target deselection (Chandonia et al., 2006), although it may also concomitantly remove examples of ‘work stopped’ data that are relatively similar to the ‘diffraction-quality crystals’ data. However, only 4 ‘work stopped’ sequences were removed by this step (see below). To provide a database for this purpose, T_POS72 and UniRef50 were combined, and processed with SEG and helixfilt as for DEVIEL_U50 to generate TPOS_U50. T_NEG614 was searched against TPOS_U50 using PSIBLAST (five iterations, default settings) and matches determined according to ‘similar structure’ thresholds (Rost, 1999). The 4 sequences from T_NEG614 that matched a T_POS72 sequence were eliminated to produce T_NEG610. In order to derive a negative test set of the same size as the positive test set, 72 sequences were randomly selected from T_NEG610 (T_NEG72). The TNEG72 and TPOS72 pair is termed TEST. For testing ‘real-world’ distributions, the T_POS72 and T_NEG610 pair is termed TEST-W.

Additional datasets were required to allow comparison with the SECRET and CRYSTALP methods, because these methods

### Table 1. Primary datasets

<table>
<thead>
<tr>
<th>Combined dataset</th>
<th>Purpose</th>
<th>Dataset</th>
<th>Length statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Min</td>
</tr>
<tr>
<td>PDB3958</td>
<td>Training</td>
<td>PDB3958</td>
<td>40</td>
</tr>
<tr>
<td>TEST</td>
<td>Test</td>
<td>T_POS72</td>
<td>72</td>
</tr>
<tr>
<td>TEST-RL</td>
<td>Test (RL)</td>
<td>T_NEG72</td>
<td>122</td>
</tr>
<tr>
<td>FEAT</td>
<td>Feature Selection</td>
<td>DIF728</td>
<td>72</td>
</tr>
<tr>
<td>FEAT-RL</td>
<td>Comparator set (RL)</td>
<td>WS728</td>
<td>72</td>
</tr>
<tr>
<td>TEST-SOL</td>
<td>Additional test (RL)</td>
<td>T_POS12</td>
<td>63</td>
</tr>
<tr>
<td>FEAT-W</td>
<td>Threshold ('real-world')</td>
<td>DIF728</td>
<td>72</td>
</tr>
<tr>
<td>TEST-W</td>
<td>Test ('real-world')</td>
<td>WS6025</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T_POS72</td>
<td>42</td>
</tr>
</tbody>
</table>

A summary of the training, feature selection and test datasets is shown. Each combined dataset is made up of a positive and negative pair; for example, FEAT comprises DIF728 (+ve) and WS728 (–ve). ‘RL’ indicates restricted length datasets, with sequence length minima and maxima of 46 and 200, respectively. FEAT-RL was not used in feature selection, but was derived for comparison with FEAT. FEAT-W was only used to select the threshold for ‘high-scoring’ predictions.

Sternberg, 1987) with a Z-score threshold of 5. This left 72 sequences to form the positive test dataset (T_POS72).

The negative test dataset was based on PDB trial sequences, parsed from the XML (http://pepdb.pdb.org). These sequences ostensibly represent the actual construct sequence used. Sequences were taken if annotated as Status ‘work stopped’ and with Status History including ‘Cloned’ but not including an indicator of crystallization (e.g. ‘Crystals’). Sequences were excluded if annotated as ‘test target’, if the ‘stopDetails’ included ‘duplicate target found’, and DNA sequences were filtered out leaving 12070 sequences. These were processed by a protocol similar to that applied in generating T_POS72, resulting in 614 sequences (T_NEG614). In order to safeguard against overlap between the positive and negative test datasets, it was necessary to filter T_NEG614 against T_POS72. This step aims to eliminate sequences where work was stopped due to target deselection (Chandonia et al., 2006), although it may also concomitantly remove examples of ‘work stopped’ data that are relatively similar to the ‘diffraction-quality crystals’ data. However, only 4 ‘work stopped’ sequences were removed by this step (see below). To provide a database for this purpose, T_POS72 and UniRef50 were combined, and processed with SEG and helixfilt as for DEVIEL_U50 to generate TPOS_U50. T_NEG614 was searched against TPOS_U50 using PSIBLAST (five iterations, default settings) and matches determined according to ‘similar structure’ thresholds (Rost, 1999). The 4 sequences from T_NEG614 that matched a T_POS72 sequence were eliminated to produce T_NEG610. In order to derive a negative test set of the same size as the positive test set, 72 sequences were randomly selected from T_NEG610 (T_NEG72). The TNEG72 and TPOS72 pair is termed TEST. For testing ‘real-world’ distributions, the T_POS72 and T_NEG610 pair is termed TEST-W.

Additional datasets were required to allow comparison with the SECRET and CRYSTALP methods, because these methods
only give predictions for sequences with length between 46 and 200 (Chen et al., 2007; Smialowski et al., 2006). T_POS72 and TNEG610, respectively, have 43 (T_POS43) and 197 (TNEG197) sequences with length between 46 and 200. In order to derive a negative test set of the same size as T_POS43, 43 sequences were randomly selected from TNEG197 (TNEG43). The T_POS43 and TNEG43 pair is termed TEST-RL.

Also, WS6025 and DIF728 were used as a basis to provide length-restricted datasets independent of the TEST-RL dataset. DIF728 and WS6025, respectively, have 246 (DIF246) and 3103 (WS3103) sequences with length between 46 and 200. In order to derive a negative test set of the same size as DIF246, 246 sequences were randomly selected from WS3103 (WS246). The DIF246 and WS246 pair is termed FEAT-RL.

Additional test datasets were developed from the PepDB trial sequences. These datasets allow testing on actual construct sequences for both the positive and negative test data, as well as allowing examination of predictive power for the crystallization of soluble proteins. For the positive set, sequences were taken if Status History included ‘Diffraction-quality crystals’ but not including ‘In PDB’ and with length ≥30. These 569 sequences were processed by a protocol similar to that applied in generating T_POS72, resulting in 30 sequences, of which 16 had length between 46 and 200 (T_POS16). The negative test set was based on TNEG614 identifiers, with ‘trialSequences’ taken if work had been stopped after ‘status’ ‘soluble’, but before ‘status’ ‘crystallized’; thus 28 sequences were identified, of which 12 had length between 46 and 200. To derive a positive test set of the same size as the negative test set 12 sequences were randomly selected from TPOS16 (TPOS12). The TPOS12 and TNEG12 pair is termed TEST-SOL. For TEST-SOL no filter was applied to remove potential overlap between the positive and negative data.

2.2 Feature calculation

For each protein sequence, the number of low complexity regions identified by SEG (Wan and Wootton, 2000), hydrophobicity, isoelectric point (pI) and the standard amino acid frequencies were calculated. Hydrophobicity was calculated as the sum of Goldmann-Engleman-Steiz (GES) hydrophobicity values (Engelman et al., 1986) for all residues, divided by the sequence length. Isoelectric point was calculated using the Bioperl (Stajich et al., 2004) pI calculator module with EMBoss-defined pKa values (Rice et al., 2000).

2.3 Parzen window probability density function estimate

The Parzen Window density estimator technique (Duda and Hart, 1973; Parzen, 1962) aims to define an unknown probability density p(x) from a set of observations, in this case the observations are provided by the PDB3958 dataset. The Parzen Window density estimator was implemented in C++ using the Gnu Scientific Library (Galassi et al., 2006). Let us suppose that the values of the considered observations are defined from p(x) in a D-dimensional space. The probability P that a vector x belongs to a specific region R can be expressed as:

\[ P = \int_R p(x) \, dx \]  

(1)

Supposing N observations are drawn from p(x), the points K falling inside the region R can be estimated by:

\[ K \approx NP \]  

(2)

Assuming the region R is sufficiently small so that the probability density p(x) can be considered constant within R, then P can be approximated to:

\[ P \approx p(x) \cdot V \]  

(3)

where V is the volume of the region R. Combining Equations 2 and 3 we can write:

\[ p(x) = \frac{K}{NV} \]  

(4)

Equation (4) depends on two contradictory assumptions. The region R is assumed to be small enough in order to consider the probability density p(x) constant, yet Equation (2) assumes R is large enough to contain a sample of K of points falling inside it. Nevertheless, the Parzen Window density probability estimator exploits this result to model p(x). Thus, K can be defined given a fixed value of V. As N tends to infinity, the estimate of p(x) tends towards the true value (Duda and Hart, 1973).

Supposing the region R to be a D-dimensional hypercube centred on the point x, with length of the hypercube sides given by h, V is defined:

\[ V = h^D \]  

(5)

We can count K, the number of points falling inside R according to the ‘Window Function’, k(u):

\[ k(u) = \begin{cases} 1 & |u_i| \leq \frac{1}{2} \quad ; \quad i = 1, \ldots, D \\ 0 & \text{otherwise} \end{cases} \]  

(6)

The term |u_i| refers to the Euclidian distance in dimension i for the considered data points. Given a data point x_n, the quantity k(\(\begin{pmatrix} x - x_n \end{pmatrix}\)) will be one if x_n falls within the hypercube centred on x with side h, and zero otherwise. The value of K can therefore be defined by:

\[ K = \sum_{n=1}^{N} k\left(\frac{x - x_n}{h}\right) \]  

(7)

Substituting Equation (5) and Equation (7) into Equation (4) defines a new expression for the estimated probability density at x:

\[ p(x) = \frac{1}{N h^D} \sum_{n=1}^{N} k\left(\frac{x - x_n}{h}\right) \]  

(8)

From Equation (6), k(u) is symmetrical, therefore Equation (8) can be reinterpreted as the sum over N cubes centred on the N data points x_n. The use of a step function creates the presence of artificial discontinuities in the probability density function estimate, so to obtain a smoother density model we chose a Gaussian kernel, which gives rise to the following density model:

\[ p(x) = \frac{1}{N h^D \sqrt{2\pi}} \sum_{n=1}^{N} \exp \left[ -\frac{1}{2} \left( \frac{|x - x_n|^2}{h^2} \right) \right] \]  

(9)

where h now represents the SD of the Gaussian components. Thus, the final density function estimate is obtained by adding up the contributions of several Gaussian models placed over the whole data set and then normalizing. Therefore, the value of h is crucial to correctly describing the PDB3958 data. The model becomes too sensitive to noise when h is too small, whilst information loss occurs when h is too large.

The optimal value of h was first estimated using PDB3958, with 2000 search increments between values h_{max} and h_{min} of 0.8 and 0.0001, respectively. The Area under the Receiver-Operator Characteristic (AROC) curve was determined by 10-fold crossvalidation, identifying a value of h equal to 0.0157 with the best PDB3958 AROC. Five additional values of h around this initial h-value were examined to maximize the AROC for the Parzen Window density estimate derived from PDB3958 and predicting over all of the FEAT dataset, finding a final value of h equal to 0.040. To facilitate separation of test data into the classes ‘amenable to crystallization’ and ‘recalcitrant to crystallization’ a probability density cutoff value was determined by optimizing accuracy [Equation (10)] over the FEAT predictions. This cutoff was applied to give ‘non-optimized’ accuracy and Matthews
correlation coefficient [Equation (11)] for predictions over TEST, TEST-RL and TEST-SOL.

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}
\]

\[
\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

\(TP = \text{true positives, } TN = \text{true negatives, } FP = \text{false positives, } FN = \text{false negatives.}\)

3 RESULTS AND DISCUSSION

3.1 Selection of features

In order to rank the amino acids by their effectiveness in prediction, the FEAT dataset was analysed using three features; isoelectric point (pI), hydrophobicity and each of the 20 amino-acid frequencies in turn (Table S1, supplementary data). The AROC was calculated for each feature combination. Random predictions will give AROC values of 0.5 whilst a perfect predictor would give an AROC value of 1. From the ranking given in Table S1, amino acid frequencies were successively included as features (Table 2). ParCrys outperforms the OB-Score when using only pI and hydrophobicity as features. This gain is likely due to ParCrys providing a smoother and non-parametric representation of the data, compared with the OB-Score. The improvement upon addition of ‘S’ and ‘C’ frequencies provides a similar performance gain, with additional features providing small improvements up to the best performing combination of features ‘pI, hydrophobicity, S, C, G, F, Y, M’. A similar process was repeated with FEAT but for pI, hydrophobicity and the number of low-complexity regions predicted by SEG (Wan and Wootton, 2000). However, the AROC values obtained were worse than those obtained without including the SEG information (see Supplementary Material). SEG has previously been used in structural genomics target selection to eliminate targets that are ‘...likely to be intractable for HT study’ (Chandonia et al., 2006), so this pre-filtering of the structural genomics targets may be the underlying reason that the SEG predictions are not found to be informative with the FEAT dataset.

Table 2 summarizes the accuracy and Matthews correlation coefficient [Equation (11)] for predictions over TEST, TEST-RL and TEST-SOL.

The results shown were obtained using the FEAT dataset. The combination ‘pI, Hydrophobicity, S, C, G, F, Y, M’ was found to perform the best. AROC denotes the area under the receiver-operator characteristic curve. These data were calculated in R (R Development Core Team, 2004).

to be informative. This is probably because our amino acid features were examined in combination with pI and hydrophobicity; pI probably serves to represent much of the information present in the individual frequencies of charged amino acids. The sparseness of data increases with the number of selected features eventually causing prediction performance to degrade (Guyon and Elisseeff, 2003), therefore limiting the number of features that can usefully be included.

ParCrys outperformed the OB-Score over both the FEAT and FEAT-RL datasets, with significantly higher AROC values (two-tailed \(P\)-value < 0.01).

3.2 Evaluation and comparison with other methods

In order to classify protein sequences as ‘amenable to crystallisation’, or ‘recalcitrant to crystallisation’, it is necessary to determine a threshold for the ParCrys probability density estimate. A threshold value of 3 564 600 was obtained by optimizing accuracy [Equation (10)] over the FEAT dataset Receiver-Operator Characteristic (ROC) space. The ParCrys accuracy estimate was then obtained by applying the threshold derived over FEAT to the independent test sets. A threshold value of 0.809 for the OB-Score method was similarly defined. FEAT is estimated to be structurally independent of the test datasets (methods), and so thresholds calculated from FEAT can be applied over the test datasets with little risk of bias. TEST-RL, the restricted length subset of the TEST dataset was developed for comparison of ParCrys, OB-Score, SECRET and CRYSTALP, this was necessary because only sequences with length between 46 and 200 are able to be analysed by SECRET and CRYSTALP (Chen et al., 2007; Smialowski et al., 2006). The SECRET predictions for TEST-RL and TEST-SOL were obtained from the web server (http://webclu.bio.wzw.tum.de:8080/secret/). The CRYSTALP predictions for TEST-RL were kindly provided by the CRYSTALP authors.

Table 3 summarizes the accuracy and Matthews correlation coefficient [Equation (11)] for predictions over TEST, TEST-RL and TEST-SOL.

\begin{table}[h]
\centering
\begin{tabular}{|l|l|}
\hline
Features & AROC \\
\hline
pI, Hydrophobicity & 0.697 \\
pI, Hydrophobicity, S & 0.714 \\
pI, Hydrophobicity, S, C & 0.723 \\
pI, Hydrophobicity, S, C, G & 0.725 \\
pI, Hydrophobicity, S, C, G, F & 0.725 \\
pI, Hydrophobicity, S, C, G, Y & 0.726 \\
pI, Hydrophobicity, S, C, G, F, Y & 0.726 \\
pI, Hydrophobicity, S, C, G, F, Y, M & 0.727 \\
pI, Hydrophobicity, S, C, G, F, Y, M, T & 0.726 \\
pI, Hydrophobicity, S, C, G, F, Y, M, T, H & 0.726 \\
pI, Hydrophobicity, S, C, G, F, Y, M, T, H, D & 0.725 \\
pI, Hydrophobicity, S, C, G, F, Y, M, T, H, D, W & 0.723 \\
\hline
\end{tabular}
\caption{Successive addition of amino acids in combination with isoelectric point (pI) and hydrophobicity}
\end{table}
Table 3. Accuracy and Matthews correlation coefficient for test datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>Accuracy (%)</th>
<th>Matthews correlation coefficient</th>
<th>AROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ParCrys</td>
<td>TEST-RL</td>
<td>79.1</td>
<td>0.582</td>
<td>0.844</td>
</tr>
<tr>
<td>OB-Score</td>
<td>TEST-RL</td>
<td>69.8</td>
<td>0.402</td>
<td>0.711</td>
</tr>
<tr>
<td>ParCrys-W</td>
<td>TEST-RL</td>
<td>67.4</td>
<td>0.384</td>
<td>0.844</td>
</tr>
<tr>
<td>SECRET</td>
<td>TEST-RL</td>
<td>58.1</td>
<td>0.163</td>
<td>0.580</td>
</tr>
<tr>
<td>CRYSTALP</td>
<td>TEST-RL</td>
<td>46.5</td>
<td>-0.070</td>
<td>-</td>
</tr>
<tr>
<td>ParCrys</td>
<td>TEST-SOL</td>
<td>75.0</td>
<td>0.530</td>
<td>0.840</td>
</tr>
<tr>
<td>ParCrys-W</td>
<td>TEST-SOL</td>
<td>66.7</td>
<td>0.385</td>
<td>0.840</td>
</tr>
<tr>
<td>OB-Score</td>
<td>TEST-SOL</td>
<td>58.3</td>
<td>0.302</td>
<td>0.736</td>
</tr>
<tr>
<td>SECRET</td>
<td>TEST-SOL</td>
<td>20.8</td>
<td>-0.642</td>
<td>0.063</td>
</tr>
<tr>
<td>ParCrys</td>
<td>TEST</td>
<td>71.5</td>
<td>0.446</td>
<td>0.752</td>
</tr>
<tr>
<td>ParCrys-W</td>
<td>TEST</td>
<td>68.0</td>
<td>0.370</td>
<td>0.752</td>
</tr>
<tr>
<td>OB-Score</td>
<td>TEST</td>
<td>64.6</td>
<td>0.319</td>
<td>0.681</td>
</tr>
<tr>
<td>ParCrys-W</td>
<td>TEST-W</td>
<td>74.0</td>
<td>0.227</td>
<td>0.738</td>
</tr>
<tr>
<td>ParCrys</td>
<td>TEST-W</td>
<td>55.3</td>
<td>0.217</td>
<td>0.738</td>
</tr>
<tr>
<td>OB-Score</td>
<td>TEST-W</td>
<td>53.2</td>
<td>0.211</td>
<td>0.701</td>
</tr>
</tbody>
</table>

‘AROC’ is the area under the receiver operator characteristic curve. AROC values were not obtained for CRYSTALP, as this method did not return raw scores. The ParCrys and OB-Score accuracy and Matthews correlation coefficient values shown for TEST-RL, TEST-SOL TEST and TEST-W were produced by defining a cutoff value optimal for FEAT (see text). The ParCrys-W method refers to results obtained using the cutoff derived from the ‘real-world’ data (FEAT-W). These data were calculated in R (R Development Core Team, 2004).

Interestingly, ParCrys had a higher AROC value over the restricted length dataset TEST-RL compared with the non-restricted dataset TEST. Indeed, higher accuracy and Matthews correlation coefficient values were also found for TEST-RL compared with TEST. These observations concur with results presented for the SECRET method (Smialowski et al., 2006).

For the data unrestricted by length, the datasets representing sequences where ‘work was stopped before crystals were obtained’ had smaller maximal length values than the datasets representing sequences with diffraction quality crystals (Table 1). Also, the ‘work stopped’ data (WS728) had smaller median and mean values of length, compared with the ‘diffraction-quality crystals’ data (DIF728). Coupled with the observed improvement in ParCrys performance for the restricted length datasets, this suggests that the restriction by length may remove sequences that were successfully crystallized but incorrectly classified by the prediction methods. The ‘diffraction-quality crystals’ sequences with length >600 (DIF-LONG) had median hydrophobicity of −0.83, similar to the WS728 value (−0.78). However, the DIF728 median hydrophobicity was −0.67, similar to the PDB3958 median hydrophobicity (−0.70). Therefore, the DIF-LONG dataset is relatively hydrophobic with a median hydrophobicity even lower than that of the ‘work stopped before crystals’ dataset. Application of the Wilcoxon rank sum test shows the PDB3958 hydrophobicity to be significantly different to that of WS728 and DIF-LONG (respective one-tailed P-values 1.43e−10 and 0.016), whereas DIF728 is not significantly different to PDB3958 (one-tailed P-value 0.989). Therefore, the longest sequences with diffraction-quality crystals have a hydrophobicity distribution that is shifted away from that of PDB3958, compared to the hydrophobicity distribution of the overall ‘diffraction-quality crystals’ dataset (DIFF728). Moreover, this shift is in the same direction as that observed for the ‘work stopped before crystals’ dataset (WS728). These observations may be rationalized by the tendency for larger proteins to have a greater proportion of residues in the hydrophobic core. Accordingly, the improved performance on shorter sequences appears to be partly due to incorrect classification of the longer ‘diffraction-quality crystals’ sequences into the relatively more hydrophobic ‘recalcitrant’ set.

We find unequal numbers of sequences in our non-redundant datasets ‘diffraction-quality crystals’ and ‘work stopped before crystals’. Therefore, to evaluate ParCrys in the context of the observed distribution of ‘real-world’ data, we took the full DIF728 and WS6025 sets (FEAT-W) to determine a ParCrys threshold by optimizing Matthews correlation coefficient over the ROC space. Simply optimizing for accuracy is flawed for these data because the negative class represents 89.2% of the data, and so the highest possible threshold value would be selected to always predict into the ‘recalcitrant’ class. Independent test data were provided by the TEST-W set (see methods and Table 1) and gave a ParCrys AROC of 0.738. The accuracy and Matthews correlation coefficient values were 74.0% and 0.227, respectively (Table 3), using the ParCrys threshold determined over FEAT-W. This threshold, derived from ‘real-world’ distributions, defines the class ‘high-scoring crystallization propensity prediction’. Therefore, ParCrys provides a three state classification: ‘recalcitrant’, ‘amenable’
and ‘high-scoring’. We recommend the ‘high-scoring’ threshold is applied when the number of ‘recalcitrant’ sequences is expected to outweigh the number of ‘amenable’ sequences, for example when selecting ‘amenable’ sequence(s) from a structurally uncharacterized group of orthologues. The ratio of positive:negative examples was approximately 1:8 in both the FEAT-W and TEST-W datasets. In general, ParCrys predictions are intended to provide guidance in selecting targets and should be interpreted in conjunction with additional knowledge about the sequence(s) of interest. An interesting further study would be to train a predictor to distinguish proteins that differ in only a few amino acids, yet have quite different crystalization characteristics. Unfortunately, data with which to train such a predictor are currently too sparse, but targeting a predictor at ‘borderline’ examples of this type is likely to become possible in the future.

4 CONCLUSIONS

We present a novel algorithm, ParCrys, for estimating a protein’s propensity to progress to the stage of diffraction-quality crystals via current structural biology techniques. We find the feature combination ‘Hydrophobicity, isoelectric point, S, G, C, F, Y, M’ to be optimal. Independent test data finds ParCrys to outperform the other publicly available methods, SECRET, CRYSTALP and the OB-Score. Also, in agreement with the authors of SECRET, we find that predictions are more accurate for sequences of length between 46 and 200 amino acids. ParCrys predictions and datasets are available from www.compbio.dundee.ac.uk/parcrys.

ACKNOWLEDGEMENTS

We thank Dr L. Kurgan for providing CRYSTALP predictions and Dr T. Walsh and Dr J. Monk for computational advice. UK Biotechnology and Biological Sciences Research Council (BBS/B/14434 to G.J.B.); Scottish Bioinformatics Research Network (HR03021 to G.J.B.); UK Engineering and Physical Sciences Research Council (EP/E052029/1 to M.A.G.).

Conflict of Interest: none declared.

Fig. 1. ROC data for ParCrys, the OB-Score, SECRET and CRYSTALP are shown for the independent test dataset TEST-RL. Also ROC curves for ParCrys and the OB-Score are shown for independent test dataset TEST. ParCrys outperforms the other methods over both TEST-RL and TEST, with AROC values of 0.844 and 0.752, respectively. A full ROC curve was not plotted for CRYSTALP, as this method returned a binary classification of sequences without raw scores. This figure was generated in R (R Development Core Team, 2004).
REFERENCES


