Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Improving the chances of successful protein structure determination with a random forest classifier

Samad Jahandideh,[a,b] Lukasz Jaroszewski[a,b,c] and Adam Godzik[a,b,c]*

[a]Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92307, USA, [b]Joint Center for Structural Genomics, http://www.jcsg.org/, USA, and [c]Center for Research in Biological Systems (CRBS), University of California, San Diego, La Jolla, California USA

Correspondence e-mail: adam@burnham.org

Received 5 June 2013
Accepted 25 November 2013

Obtaining diffraction quality crystals remains one of the major bottlenecks in structural biology. The ability to predict the chances of crystallization from the amino-acid sequence of the protein can, at least partly, address this problem by allowing a crystallographer to select homologs that are more likely to succeed and/or to modify the sequence of the target to avoid features that are detrimental to successful crystallization. In 2007, the now widely used *XtalPred* algorithm [Slabinski *et al.* (2007), *Protein Sci.* **16**, 2472–2482] was developed. *XtalPred* classifies proteins into five 'crystallization classes' based on a simple statistical analysis of the physicochemical features of a protein. Here, towards the same goal, advanced machine-learning methods are applied and, in addition, the predictive potential of additional protein features such as predicted surface ruggedness, hydrophobicity, side-chain entropy of surface residues and amino-acid composition of the predicted protein surface are tested. The new *XtalPred-RF* (random forest) achieves significant improvement of the prediction of crystallization success over the original *XtalPred*. To illustrate this, *XtalPred-RF* was tested by revisiting target selection from 271 Pfam families targeted by the Joint Center for Structural Genomics (JCSG) in PSI-2, and it was estimated that the number of targets entered into the protein-production and crystallization pipeline could have been reduced by 30% without lowering the number of families for which the first structures were solved. The prediction improvement depends on the subset of targets used as a testing set and reaches 100% (*i.e.* twofold) for the top class of predicted targets.

1. Introduction

The high failure rate in experimental protein structure determination by crystallographic methods is still one of the greatest challenges in structural biology. In fully automated gene-to-structure pipelines, success rates between target selection and structure deposition hover around 5% depending on the family, type and source organism (bacteria or eukaryote) of a protein. Laboratories that focus on individual high-profile proteins achieve much higher success rates but typically at a significantly higher cost, both in terms of time and materials, spent on multiple crystallization attempts on a range of construct variants. The cost and time lost on unsuccessful structure-determination attempts impedes the overall progress in structural biology and contributes significantly to the high average cost of determining protein structures. Estimates from JCSG suggest that up to 70% of the average cost of solving any protein structure arises from the costs of failed attempts. The ability to select homologs and/or the design of constructs or mutants that lead to better diffracting crystals would increase the success and lower the cost of protein

Table 1
The preparation of the training and test sets from the PSI TargetTrack database.

Filtering step	Details	No. of targets after this step
The initial positive set (crystallographic structures)	At least at *Crystal Structure* stage as of August 22, 2012. Lengths between 50 and 800 amino acids.	4924
The initial negative set (targets which failed to crystallize)	*Purified* stage as of January 1 2011. Excluded: *crystallized targets (of any quality)*, *NMR targets*, targets *stopped because duplicate target was found*, targets which might have been stopped because of a related structure in the PDB, targets prepared for biological assays. Lengths between 50 and 800 amino acids.	21898
Balancing training and testing data by reducing negative set	Clustering at 66% sequence identity using *CD_HIT* (Li & Godzik, 2006) and random selection of 1/3 of targets	5691
Removing 'trivial' prediction targets	Excluded: targets with predicted signal peptides and transmembrane helices (such targets have practically no chance of crystallizing in standard setups as full-length constructs)	Positive set, 4710; negative set, 4795
Eliminating sequence similarity between training and testing set	The *PSI-BLAST* program was used to put groups of similar sequences into either the training set or the testing set.	Training positive set, 2265; training negative set, 2355; testing positive set, 2445; testing negative set, 2440
Adjusting the percentage of predicted positives by undersampling	Training sets with positive subset reduced by random selection to predict (approximately) 5, 10, 20, 30, 40, 50, 60, 70, 80 and 90% of positives.	

structure determination, allowing structural biology laboratories to tackle a broader range of biologically important targets.

Because of the importance of the problem, efforts to understand and improve the protein crystallization process started decades ago and continue today. Initially, crystallization was viewed as a purely stochastic phenomenon and prediction of crystallization was essentially considered to be impossible. Initial data-mining efforts (Carugo & Argos, 1997), and a growing body of anecdotal evidence collected in individual laboratories and shared between crystallographers, led to the recognition that protein surface properties, unique to each protein, critically affect crystallization. This in turn led to procedures designed to change such properties, for instance by the reduction of surface entropy by protein engineering (Garrard *et al.*, 2001; Goldschmidt *et al.*, 2007; Derewenda, 2011). However, most of these efforts were developed by analysis of biased positive-only training sets, *i.e.* crystallographic structures of proteins stored in the Protein Data Bank (Berman *et al.*, 2000). This situation changed when the Protein Structure Initiative (PSI) started producing and screening large sets of proteins and reporting on both successes and failures. These data were collected in the PSI TargetDB database (Chen *et al.*, 2004), which enabled large-scale data-mining for protein crystallization (Christendat *et al.*, 2000; Canaves *et al.*, 2004; Goh *et al.*, 2004; Smialowski *et al.*, 2006).

In 2007, our laboratory developed the *XtalPred* algorithm (Slabinski *et al.*, 2007) for prediction of crystallization success from the statistical analysis of seven physicochemical features. Since then, *XtalPred* and other similar algorithms such as *ParCrys* (Overton *et al.*, 2008), *CRYSTALP*2 (Kurgan *et al.*, 2009), *MetaPPCP* (Mizianty & Kurgan, 2009), *PXS* (Price *et al.*, 2009), *SVMCRYS* (Kandaswamy *et al.*, 2010), the MCSG *Z*-score (Babnigg & Joachimiak, 2010) and *PPCpred* (Mizianty & Kurgan, 2011) have allowed users to assess the probability of successful structure determination prior to performing any experimental work and to modify or adjust their target-selection strategies. This advance has resulted in a significant enhancement of the efficiency and productivity of Structural Genomics efforts (Jaroszewski *et al.*, 2008; Savitsky *et al.*, 2010; Gabanyi *et al.*, 2011; Xiao *et al.*, 2010) and importantly also helped many individual structural biology groups (Lee *et al.*, 2010; Gómez García *et al.*, 2011, 2012; Oyenarte *et al.*, 2011). A myriad of users have submitted thousands of potential targets to the *XtalPred* server, with requests coming both from Structural Genomics centers and from small structural biology laboratories.

It has recently been shown that the random forest method is highly suitable for crystallizability prediction (Jahandideh & Mahdavi, 2012), suggesting a path to improve the prediction accuracy of *XtalPred*. A new large data set extracted from the new PSI TargetTrack database (http://sbkb.org/tt/) has allowed us to also test additional physicochemical features for their correlation with structure-determination success. In this study, we describe the new *XtalPred-RF* algorithm and test it on several benchmarks, as well as illustrate its benefits by estimating how applying it to our original target selection procedures would have increased the productivity of our center.

It is important to note that while *XtalPred* and other algorithms for prediction of crystallization success were developed by groups involved in and using data from high-throughput crystallization projects, their application is not limited to such groups. Individual laboratories dealing with groups of homologous proteins can benefit from optimizing their selection strategy, and save valuable time and cost.

Note: in the following sections, we use the term 'crystallizability' as a shortcut for the 'probability of yielding well diffracting crystals that allow structure determination'.

2. Materials and methods

We followed three strategies to improve the *XtalPred* algorithm. First, we tested several machine-learning approaches on

the original data used to develop the *XtalPred* method. Second, we updated the training and testing sets, as the previous sets only contained data collected before 2005, by adding all data collected up to 2011 or 2012 (see Table 1). Finally, we introduced and evaluated several additional physicochemical features of the (predicted) protein surface. In subsequent sections, we describe the details of each strategy. The following definitions of the accuracy, specificity, sensitivity and Matthews correlation coefficient (MCC) were used in this publication,

$$\begin{aligned}\text{Accuracy} &= (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}),\\ \text{Specificity} &= \text{TN}/(\text{FP}+\text{TN}),\\ \text{Sensitivity} &= \text{TP}/(\text{TP}+\text{FN}),\\ \text{MCC} &= \frac{\text{TP}*\text{TN}-\text{FP}*\text{FN}}{[(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{NT}+\text{FN})]^{1/2}},\end{aligned} \tag{1}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

2.1. Machine-learning methods

XtalPred (Slabinski *et al.*, 2007) used a simplistic approximation of independent probabilities, *i.e.* the expert pool method. Here, we test several machine-learning methods: (i) the support vector machine (SVM), (ii) the artificial neural network (ANN) and (iii) the random forest (RF) methods, which, at least in principle, are better suited to take into account complex multi-dimensional interactions between different physicochemical variables used in crystallization prediction.

Support vector machine (SVM) is a statistical-learning-theory-based classification algorithm (see Vapnik, 1995, 1998). In this study, we applied the tune function using the e1071 package of the *R* environment (v.2.11-1) to develop a binary SVM-based method. The tune function uses a grid search to find the optimum structure of SVM.

Artificial neural network (ANN) is a powerful nonlinear predictor inspired by biological neural networks (Bishop, 1995). In order to determine the optimum structure of the network, we constructed a large number of networks, varying the number of hidden neurons, the number of iterations and the learning rate. Several training algorithms such as gradient descent, resilient back-propagation, quasi-Newton and conjugate gradient were tested. The optimization resulted in an architecture characterized by one output neuron representing the crystallizability of a protein (0 for noncrystallizable protein and 1 for a protein with solved structure: see the note at the end of §1), one hidden layer containing six neurons, and an input layer containing eight neurons corresponding to the sequence parameters used in the original *XtalPred* training set. The best result was obtained using the conjugate-gradient training algorithm. The network was trained perfectly after 2000 iterations. The optimal learning rate was found to be 0.2. The program used to construct the neural networks was written in *MATLAB* v.7.14 (R2012a).

The RF algorithm (Breiman, 2001) is an advanced machine-learning method that has been successfully applied to various biological problems (Díaz-Uriarte & Alvarez de Andrés, 2006; Svetnik *et al.*, 2003; Jiang *et al.*, 2007; Kandaswamy *et al.*, 2011). RF utilizes hundreds or thousands of independent decision trees to perform classification. Each of the member trees is built on a bootstrap sample from the training data using a random subset of available variables. RF is particularly suitable for mining high-dimensional and noisy data (Fang *et al.*, 2008, 2009). In this study, we used the RF algorithm implemented by the *randomForest* (v.4.6-2) *R* package (Liaw & Wiener, 2002). The number of trees and stepFactor were set to 1000 and 2, respectively. For the other parameters of the RF method, we used default values as provided by the *R* package.

2.2. The training and testing data sets

The original *XtalPred* was developed in 2007 (Slabinski *et al.*, 2007), and today much larger data sets are available through the PSI TargetTrack database (http://sbkb.org/tt/). However, to make them useful for the purposes of the training of machine-learning methods, they need to be appropriately processed.

2.2.1. Balancing the data sets. In classification problems, the proper selection of the training data considerably affects the classification accuracy. In most cases, balanced data sets with equal counts of all classes are optimal for training. However, the data in real applications often have an imbalanced class distribution, *i.e.* most of the data are in one class (the majority class). Unfortunately, if the data used for training have a specific ratio between classes, the classifier's predictions tend to have a similar ratio between classes. Moreover, if the crystallization success is 5%, a classifier could achieve 95% accuracy by predicting failure for all of the targets. Therefore, it is important to adopt methods suitable for classification in imbalanced data problems (Yen & Lee, 2009). The two most common approaches to deal with the class-imbalance problem are over-sampling and under-sampling techniques. The over-sampling approach increases the number of minority class samples to reduce the degree of imbalanced distribution. In contrast, in the under-sampling approach one reduces the number of samples in the majority class (Yen & Lee, 2009). The under-sampling technique has recently been used in structural bioinformatics projects (Zhang *et al.*, 2012; Yu *et al.*, 2013). Generally, the performances of over-sampling approaches are worse than those of under-sampling approaches, so here we applied the under-sampling approach to address the data imbalance. In the case of crystallizability prediction, the majority class corresponds to a negative training set (crystallization failures). In order to obtain a balanced training set, we reduced the negative set by clustering and random selection (see §2.2.3 below and Table 1).

2.2.2. Using RF for multiple crystallizability class prediction. The original *XtalPred* method based on the expert pool

approach provided a numerical score, which made it possible to rank targets by their predicted crystallizability (see the note at the end of §1) score and subsequently to group targets into multiple target 'classes'. This would not be directly available from the *XtalPred-RF* method since the RF classifier provides only two tiered (positive and negative) predictions. However, for practical purposes, the ability to select the top 5, 10 or 20% of the most promising targets is very important. In order to make it possible to group targets into multiple crystallizability classes, we retrained the RF classifier, adjusting the percentage of predicted positives to a desired level using the under-sampling technique described in the previous section. Thus, the *XtalPred-RF* method consists of a series of independent RF classifiers trained on differently balanced training sets, which are then used to identify classes of targets with different probabilities of successful structure determination. This approach allowed direct comparison with target classes calculated using the existing *XtalPred* method (see Fig. 4*a*) and provides more useful information to the user.

2.2.3. Procedure for preparation of the training and testing sets. The detailed preparation of the training and testing sets from the TargetTrack (http://sbkb.org/tt/) database involved several steps (see Table 1). The preparation of the positive set is straightforward (targets at the Crystallographic Structure stage from the PSI TargetTrack database), while the preparation of the negative set requires the use of additional filters to eliminate targets that did not fail in crystallographic trials but were abandoned for other reasons. The size of the negative set was also reduced by the under-sampling method. For details, see Table 1. Lists of the training and testing sets of targets used here are available from the *XtalPred* server at http://ffas.burnham.org/XtalPred/data.tar.

2.2.4. Eliminating the risk of overfitting. The application of advanced machine-learning methods increases the risk of overfitting, or in other words obtaining artificially good results that would not be reproduced in real-life applications. The most likely causes of overfitting include: (i) a small training set, (ii) noise in the data and (iii) the inclusion of irrelevant features in the data. To avoid these problems, we (i) used a larger data set in comparison with the original *XtalPred* data set and (ii) avoided irrelevant features by testing the effect of adding novel features on the performance of the method. In the case of sequence data, overfitting would correspond to direct 'memorization' of individual sequences and 'predicting' crystallizability based on close sequence similarity (preferably the machine-learning method 'learns' and then recognizes protein features rather than individual sequences). Thus, we split the initial data set into training and testing subsets using different sequence-similarity cutoffs and then retrained and retested the same prediction method (RF) with the same set of protein features. We did not observe any substantial changes in performance as measured by the MCC when we used a less stringent cutoff for separating the training and testing sets (MCC changes were below 3% and did not show any systematic trend). In order to completely eliminate the risk of overfitting in all of our tests, we used the most stringent separation cutoff based on the *PSI-BLAST* (Altschul *et al.*, 1997) algorithm. This means that the sequences in our training set do not have any similarity detectable with *PSI-BLAST* to any sequence in our testing set.

2.3. Calculation of features of the predicted protein surface

We predicted the relative surface accessibility (RSA) for each residue of each target sequence using the *NetSurfP* algorithm (Petersen *et al.*, 2009) and tested two methods of calculating features of the predicted surface: (i) simple averaging over residues with exposed (E) status as predicted by *NetSurfP* and (ii) averaging over all residues but using values of the predicted relative surface area (RSA) as weights (see equation 2). Method (ii) led to significantly better crystallizability predictions and was consequently used in the final version of *XtalPred-RF*.

$$\bar{f} = \sum_{i=1}^{N} f_i \mathrm{RSA}_i / \sum_{i=1}^{N} \mathrm{RSA}_i, \qquad (2)$$

where $\bar{f}$ is the value of the protein feature averaged over the protein surface, f_i is the feature's value for residue i and RSA_i is the relative surface accessibility for residue i. N is the total number of residues in the protein.

2.3.1. Introducing surface ruggedness. One can anticipate that the number of protrusions and cavities on the protein surface may have an impact on crystallizability. We usually do not know the shape of the protein surface prior to structure determination but we can, to some extent, predict whether it is more or less 'rugged' than the average expected for a protein of a given size. We introduced surface 'ruggedness' defined by a simple formula: as a ratio between surface area calculated as a sum of absolute solvent accessibilities of individual residues (as predicted by *NetSurfP*) and the total accessible area expected for a protein of a given molecular mass (statistics calculated by Miller *et al.*, 1987; see equation 3). For proteins for which structure can be predicted with some accuracy by fold prediction or comparative modeling, the predicted

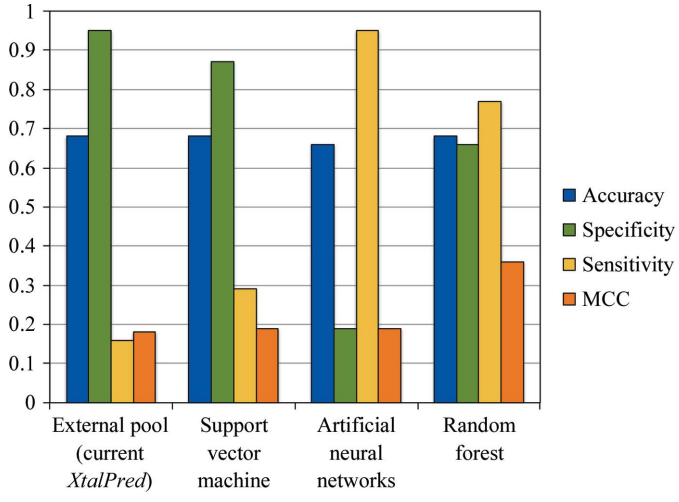

Figure 1
Performance comparison of machine-learning methods *versus* the expert pool method used in *XtalPred*. The definitions of sensitivity, specificity, accuracy and MCC are given at the beginning of §1 (equation 1).

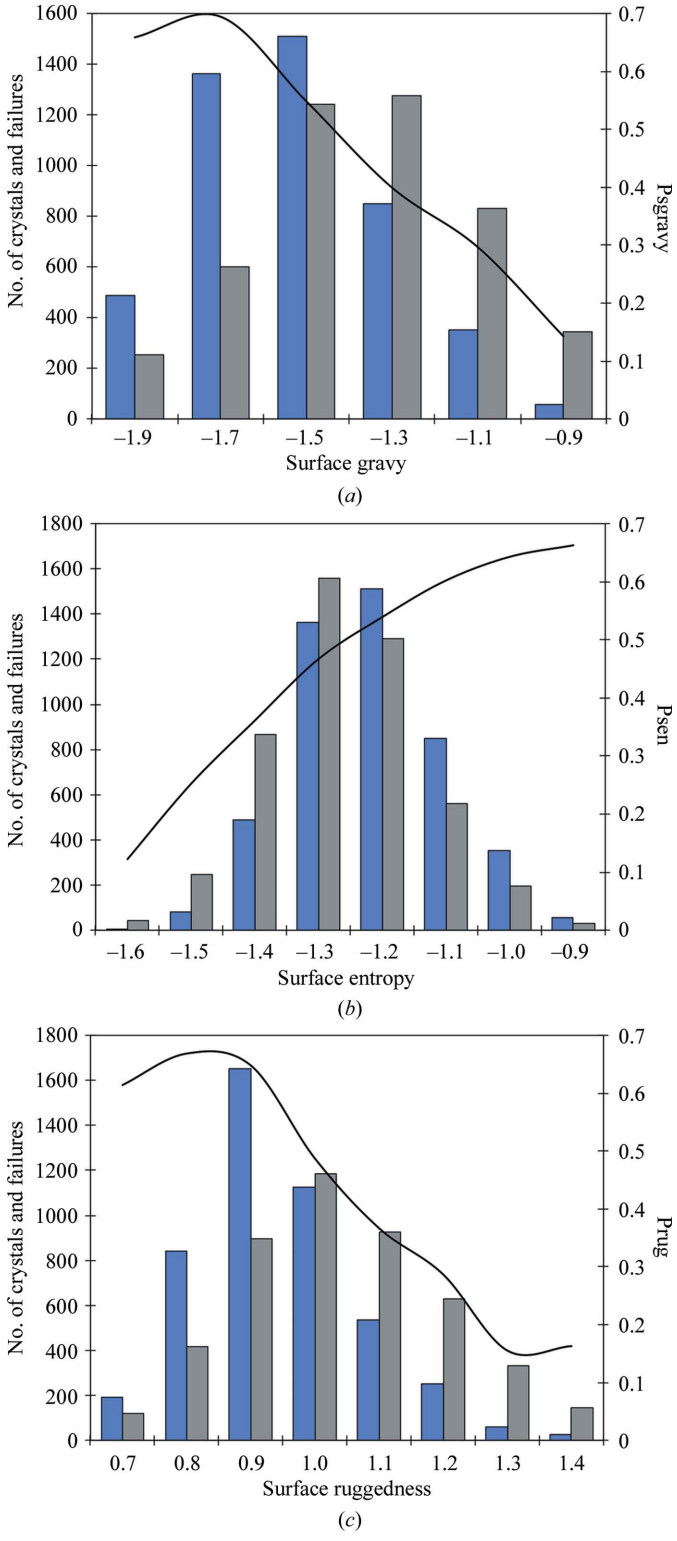

Figure 2
Correlation of the features of the predicted protein surface with the likelihood of crystallization. (*a*) Surface hydrophobicity. (*b*) Surface side-chain entropy. (*c*) Surface ruggedness. Targets from the data set were grouped into bins according to the values of these features. The bar graphs (associated with the left axis of the graph) show the number of successfully determined X-ray structures (blue) and the number of crystallization failures (gray) in each bin, respectively. The estimated likelihood of crystallization in each bin is depicted by a black line (associated with the right axis of the graph).

solvent accessibilities can in principle be improved, but as a first approximation, we decided to use the *NetSurfP* predicted values.

$$\mathrm{SR} = A_\mathrm{N}/A_\mathrm{MJLC}, \tag{3}$$

where SR is the surface ruggedness, $A_\mathrm{N} = \sum_{i=1}^{N} \mathrm{SA}_i$ is the total accessible surface calculated by *NetSurfP*, $A_\mathrm{MJLC} = 6.3M^{0.73}$ is the accessible surface predicted based on molecular mass (Miller *et al.*, 1987), SA_i is the predicted absolute solvent accessibility of residue i and M is the molecular mass of the protein.

3. Results

3.1. Application of advanced machine-learning methods

The current version of the *XtalPred* algorithm (Slabinski *et al.*, 2007) uses the expert pool method (Genest *et al.*, 1984) to combine probability distributions calculated for individual protein features and thus relies on the assumption that these probabilities are independent. This is a rather crude approximation, since most of the proteins features are likely to be correlated (for instance, more hydrophilic proteins may contain more structural disorder, larger proteins tend to be, on average, more hydrophobic *etc.*) and it is possible that some effects may be conditional (the impact of pI is more significant for small proteins *etc.*). In such situations, machine-learning approaches typically provide better prediction accuracy. Indeed, in our tests, all of the machine-learning techniques used (*i.e.* SVM, ANN and RF), when trained on the same set of protein features and the same training set as the original *XtalPred* algorithm (Slabinski *et al.*, 2007), at least slightly surpassed the original *XtalPred* results, with the RF method yielding the best results. The MCC (Matthews, 1975) improved twofold compared with the original *XtalPred* results (0.36 *versus* 0.18). Fig. 1 compares the performance of machine-learning methods *versus* the expert pool method (as used in the original *XtalPred*). The MCC (Matthews, 1975; see equation 1) is universally used in machine learning as a measure of the quality of binary (two-class) classifications. The MCC varies between −1 and 1, where −1 indicates that all predictions are wrong, 0 means that the predictions are comparable to random selection and an MCC of 1 means prefect prediction. The MCC does not strongly depend on the overall percentage of positive or negative predictions given by the method and, because of this, it can be used to compare predictions using different methods. [Note: the MCC is not related to the Matthews coefficient (Matthews, 1968) used in crystallography, other than being proposed by the same author.]

3.2. Including features of the predicted protein surface in crystallization predictions

The features of the protein surface are expected to have a greater impact on protein crystallizability than features of the protein core. For instance, we expect the percentage of serine residues on the protein surface to have a greater impact on

Table 2
The effect of adding new features on the prediction of crystallizability (see the note at the end of §1).

Definitions of accuracy, specificity, sensitivity and MCC are given at the beginning of §2 (equation 1).

	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC (%)
Initial set of features				
Length, gravy, pI, instability index, predicted disorder, insertion score	68.0	72.0	66.0	36.0
Added input features for random forest method	$\Delta_{Acc.}$	$\Delta_{Spe.}$	$\Delta_{Sens.}$	Δ_{MCC}
Hydrophobicity of the predicted surface ('surface gravy') (averaged using solvent accessibility; see §2)	0.6	4.5	−2.0	2.8
Surface entropy (averaged using solvent accessibility; see §2)	1.0	0.4	1.7	3.6
Surface ruggedness (see §2)	1.9	0.3	3.7	6.7
Amino-acid composition of the predicted surface (averaged using solvent accessibility; see §2)	2.9	4.1	1.6	10.3
Overall amino-acid composition	1.4	1.7	1.1	4.9
Current best set of features				
Length, pI, instability index, predicted disorder, insertion score, surface hydrophobicity, surface entropy, surface ruggedness, surface amino-acid composition, overall amino-acid composition	74.0	78.0	69.0	47.0

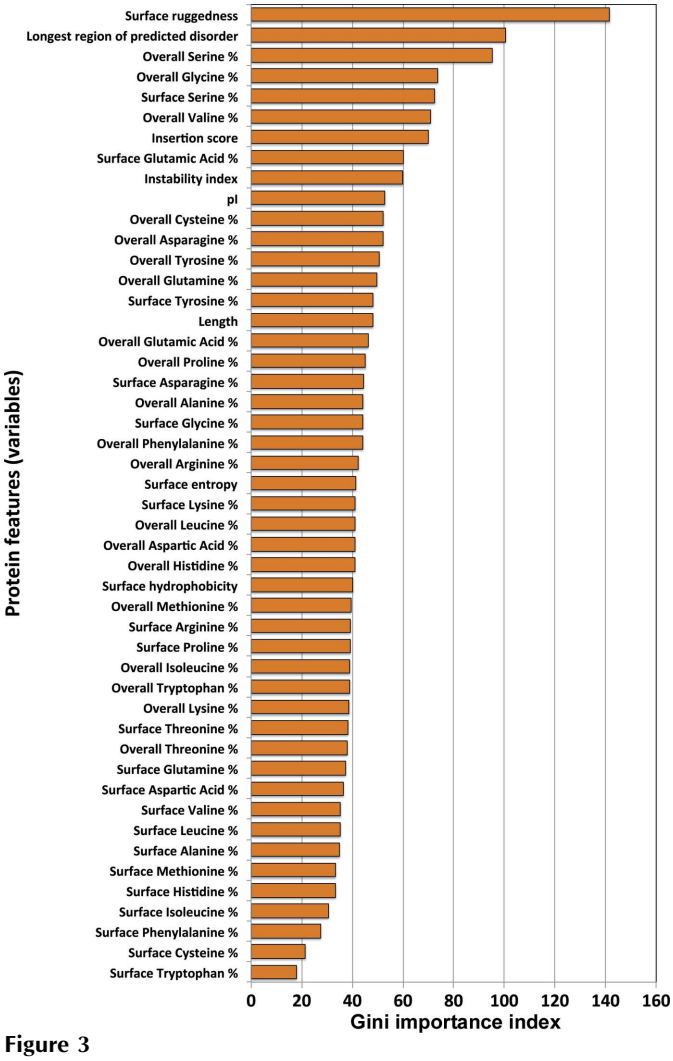

Figure 3
The importance of different variables (protein features) used as predictors in the random forest method provided by the importance function (Gini Index) from the random forest algorithm.

protein crystallization than the percentage of serine residues in the protein interior. Unfortunately, the protein surface is usually unknown prior to structure determination. However, as we show here, features of the predicted surface that could be calculated directly from the sequence of the protein can provide a good approximation, at least in a statistical sense, to those of the actual surface. The histograms shown in Fig. 2 illustrate the correlation of three features of the protein surface with protein crystallizability and Table 2 shows the cumulative contributions of these features to the prediction using the RF method.

The initial set of features included the variables used in the original *XtalPred* algorithm, *i.e.* sequence length, isoelectric point, gravy index, the longest disordered region, instability index, percentage of coil structure, coiled coils and insertion score (see Table 1 of Slabinski *et al.*, 2007). When this initial set of features was used to train the RF classifier, it resulted in an accuracy, sensitivity, selectivity and MCC of 68, 72, 66 and 36% (see equation 1), respectively, when tested on the testing set. Subsequently, we used this result as the reference for evaluating the improvement in prediction from adding novel features as predictive variables. We decided to focus on the features of the predicted protein surface (here predicted using the *NetSurfP* method; Petersen *et al.*, 2009). We tested two ways of calculating (averaging) features of the protein surface as described in §2. The surface features calculated using the weighting method led to a better improvement in the prediction (as measured by the MCC and other parameters) compared with averaging only over predicted exposed residues. The improvement from adding surface entropy, hydrophobicity and ruggedness calculated as weighted averages (see §2) was indicated by increases in the MCC and other measures of prediction performance. The single feature which provided the largest contribution to the improvement in prediction was surface ruggedness (as shown in Table 2, amino-acid composition had a greater impact on the prediction improvement but is described by 20 parameters, while surface ruggedness is described by a single parameter).

3.3. Evaluating the importance of individual features

The RF method provides several measures of the importance of individual predictors (variables). Here, we used the Gini importance index to evaluate the importance of 48 variables included in the optimized version of *XtalPred-RF* (see Fig. 3). The most important individual variables include surface ruggedness, the length of the longest predicted

Table 3
Application of different crystallizability prediction methods to the problem of solving the first structures from 271 Pfam families assigned to the JCSG in 2005.

(*a*) Random, *XtalPred* and *XtalPred-RF*.

	Solved structures/'solved' families		
Top scoring targets used (%)	Random	*XtalPred*	*XtalPred-RF*
5.3	3/3	9/7	18/12
10.2	5/5	13/8	28/22
19.7	12/12	19/11	34/26
30.0	18/18	30/21	43/34
39.7	23/22	35/26	48/37
50.2	30/27	45/35	50/39
60.3	37/31	49/38	54/43
70.2	45/34	54/42	60/49
79.7	52/41	58/46	65/52
89.9	60/49	63/50	65/52
100.0	65/52	65/52	65/52

(*b*) Multistep *XtalPred-RF*.

	Solved structures/'solved' families
Top scoring targets used (%)	Multistep *XtalPred-RF*
5.3	18/12
9.4	26/22
17.5	30/26
27.4	38/34
35.3	40/37
44.2	43/40
52.9	46/43
61.8	53/50
70.6	55/52

disordered fragment and the overall percentage of serine residues in the protein sequence. The percentage of glycine and valine residues as well as the insertion score (the percentage of gaps calculated in the multiple sequence alignment of sequences homologous to the target; see Slabinski *et al.*, 2007) were also recognized as important predictors of the crystallizability of a protein (see §4).

3.4. Revisiting the PSI-2 Pfam family draft

As described in §2, we defined target classes by *XtalPred-RF* using the under-sampling technique. This allowed a direct comparison of target classifications by *XtalPred* and *XtalPred-RF* on the testing set (prepared as described in §2.2.3 and in Table 1). As expected, classification by *XtalPred-RF* is clearly better, as indicated by significantly higher success rates in the top target classes and lower success rates in the bottom classes (see Fig. 4*a*). However, the testing set contained a high percentage of positive cases, and the impact of using *XtalPred* can be more realistically evaluated by applying it to the real target-selection problem.

To fully illustrate the benefits of using *XtalPred-RF* for target selection in a 'real-life' application, we performed an *in silico* experiment in which different target-selection methods were applied to targets from 271 Pfam families assigned to JCSG by the PSI 'Pfam Target draft' which took place in 2005. In this draft, the PSI Centers selected over 1300 largest Pfam families with no structural coverage and distributed them between four PSI production centers (Dessailly *et al.*, 2009). Between 2005 and 2010, JCSG selected 3471 targets, solving 65 structures, including the first representatives of 52 Pfam families (for some families, JCSG solved more than one representative). The overall success rate for this group of targets was 1.9%, but this also includes a substantial number of targets for which the work was stopped because a structure of a homologous protein had already been solved. Table 3

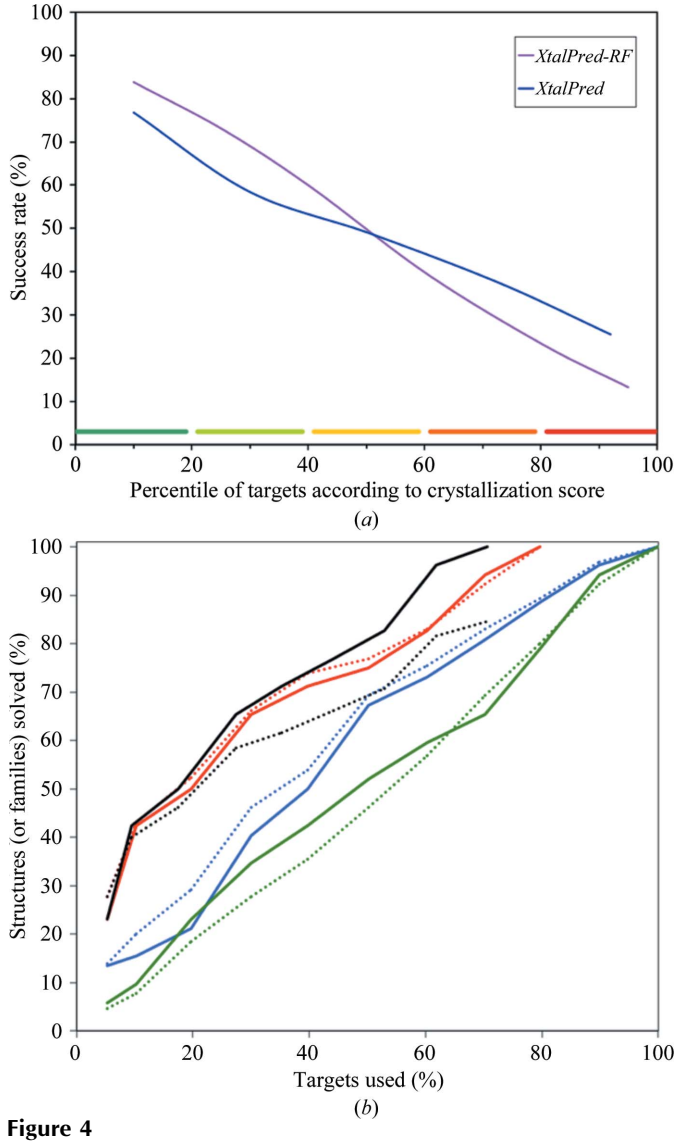

Figure 4
(*a*) Comparison of the success rates obtained on the testing set for target classes defined by *XtalPred* (blue line) and *XtalPred-RF* (magenta line). Percentile ranges for five target classes of the original *XtalPred* are depicted as colored bars above the *x* axis. Target classes defined by *XtalPred-RF* were grouped into pairs to allow direct comparison with five classes from *XtalPred*. (*b*) Revisiting the PSI-2 target draft, where the JCSG solved 65 structures including the first structural representatives of 52 Pfam families using a target pool of 3471 proteins. The graph shows the percentage of families (solid lines) and structures (dotted lines) as a function of the percentage of selected protein targets. Green lines, targets selected by random; blues lines, targets selected using *XtalPred*; red lines, targets selected using *XtalPred-RF* as introduced in this manuscript; black lines, targets selected in multiple steps using *XtalPred-RF* (selection of the top 5% of targets, elimination of solved families from the target pool, selection of the next best 5% of targets *etc.*; see §4).

shows the number of targets that had to be processed to achieve a specific number of structures in three different scenarios: if targets were selected randomly, by the *XtalPred* method and by the *XtalPred-RF* method described in this publication. The results indicate that both target-selection methods have a clear advantage over random target selection, as indicated by a lower number of targets needed to achieve the same number of solved structures and first representatives of protein families for the same number of selected targets (see Fig. 4*b*). Targets selected with *XtalPred-RF* reached a success rate approximately two times higher than targets selected with *XtalPred* and up to six times higher than randomly selected targets (top row in Table 3*a*). In order to eliminate the risk of overfitting, all 3471 JCSG targets used in the 2005 target draft were excluded from the training set used in this experiment.

4. Discussion

It has already been demonstrated that the application of advanced machine-learning methods and especially the RF classifier improves crystallizability prediction (Jahandideh & Mahdavi, 2012). However, the RFCRYS method described in that publication used only a limited set of protein features and was tested on a relatively small test set. Our results confirm the usefulness of the RF method on a large updated test set and prove that it is not the result of overfitting or direct memorization of close similarities between sequences of proteins in the training and test sets.

We selected RF as the prediction method and subsequently tested the usefulness of several additional variables as crystallizability predictors, focusing on the features of the predicted protein surface. The inclusion of protein surface features, even if only predicted from sequence information alone, significantly improves the prediction as suggested by crystallization probability distributions for individual variables (Fig. 2) and confirmed by overall prediction improvement (Table 2) and by the importance of individual features as evaluated by the Gini importance index (Fig. 3). It is necessary to note here that the importance measures included in the RF method (and most other importance measures) are unavoidably influenced by interactions between variables (Liaw & Viener, 2002). In fact, the variables (predictors) used by *XtalPred-RF* are expected to be correlated. For instance, surface features such as surface ruggedness, surface hydrophobicity and surface entropy are simple functions dependent on tabularized values for residue types and/or the predicted solvent exposures of individual residues (see equations 1 and 2). In particular, surface entropy is likely to be correlated with surface ruggedness since both tend to assign high values to long and branched side chains. Thus, the impact of individual percentages of such residues on crystallizability may be underestimated by the Gini importance index since they are already taken into account in surface entropy. While RF (in contrast to the previously used expert pool) automatically takes such correlations into account and effectively reduces the individual contributions of correlated variables, the Gini importance of an individual variable may be diminished if a correlated variable is already included in the prediction method. Therefore, to gain independent insight, the impact of individual variables was also assessed by one-dimensional histograms of negative and positive targets (Fig. 2).

As clearly indicated by the histogram shown in Fig. 2(*c*), surface ruggedness has a strong negative impact on the crystallizability of a protein; the probability of target crystallization for low ruggedness values is around 0.7, but for high ruggedness values it drops to 0.1. This is consistent with the nearly 7% increase in the MCC resulting from adding this feature to the predictors used by RF (it is noteworthy that this significant improvement occurs despite that fact that it was added after adding surface entropy, which is likely to describe similar features of the protein surface). One-dimensional histograms (Figs. 2*b* and 2*c*) suggest why surface ruggedness may have a higher overall impact on protein crystallizability than surface entropy; while both features have a high impact on crystallizability, surface ruggedness has a strong impact on a larger number of targets (surface ruggedness below or above 1 in Fig. 2*c*), while surface entropy seems to have strong impact mostly at the extremes of its distribution (surface entropies below -1.3 or above -1.2 in Fig. 2*b*).

As expected, protein surface features have an impact on the crystallization of a protein, and including them (even in a very simple form) in the prediction process leads to significant improvements in crystallizability prediction. By adding these features and by applying advanced data-mining methods, we developed an improved crystallizability prediction method, *XtalPred-RF*. It is publicly available from the *XtalPred* server at http://ffas.sanfordburnham.org/XtalPred-cgi/xtal.pl. *XtalPred-RF* is expected to further reduce the cost of structure determination in cases where an optimal target(s) can be selected from a larger pool of proteins.

The potential impact of the prediction improvements implemented in *XtalPred-RF* for structural characterization of protein families was demonstrated by revisiting (*in silico*) the target-selection and structure-determination efforts which followed the 'Pfam Family Draft' performed by the PSI high-throughput centers in 2005. The advantage of *XtalPred-RF* over *XtalPred* and random target selection is most significant when small numbers of individual protein targets are selected from a very large pool of available targets, since it makes it possible to select targets from the optimal classes where the *XtalPred-RF* resolution is higher. In order to improve the efficiency in targeting protein families even further, one can consider a multistep target selection in which a small number of targets are selected first and families with representatives solved in this first step are then excluded from the pool of targets selected in the next step *etc.* According to our tests, the application of such a procedure with ten steps (corresponding to adding subsets of targets listed in Table 3*a*) would lead to the solution of first structural representatives of all eventually solved Pfam families using just 70% of the individual protein targets used in the real target selection performed between 2005 and 2010 (see Table 3, Fig. 4*b*). It would also allow the solution of first representative structures from half of the

targeted Pfam families by using only about 1/6 of the individual proteins available in the genome pool. The most dramatic improvement of crystallizability prediction was observed for the top class of targets, where the targets selected using the orginal *XtalPred* method yielded nine structures while the same number of targets selected using *XtalPred-RF* yielded 18 structures (see Table 3*a*).

The examples presented in this manuscript focus on the application of *XtalPred-RF* to large-scale crystallographic structure determination efforts, but it can provide useful information for any structural biology group interested in the structural characterization of a protein family. A typical application of the *XtalPred-RF* server for such a user would be to prioritize a series of homologous targets according to likely crystallization success or to search for such targets in other bacterial genomes (the *XtalPred* server has such an option).

At the same time, it is important to note that all training and test sets used in this study consist of data from prokaryotic proteins. Therefore, one can expect that *XtalPred-RF* would reach optimal performance for this type of proteins. Feedback from our users and literature references indicate that *XtalPred* predictions are increasingly being used for construct design in eukaryotic proteins, and we are now working on expanding the training data sets to include eukaryotic proteins and on expanding the *XtalPred-RF* algorithm to enable the design of optimal construct boundaries.

This work is supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award No. R01 GM095847 (SER/XtalPred) and U54 GM094586 (JCSG). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.

Babnigg, G. & Joachimiak, A. (2010). *J. Struct. Funct. Genomics*, **11**, 71–80.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Breiman, L. (2001). *Mach. Learn.* **45**, 5–32.

Canaves, J. M., Page, R., Wilson, I. A. & Stevens, R. C. (2004). *J. Mol. Biol.* **344**, 977–991.

Carugo, O. & Argos, P. (1997). *Protein Sci.* **6**, 2261–2263.

Chen, L., Oughtred, R., Berman, H. M. & Westbrook, J. (2004). *Bioinformatics*, **20**, 2860–2862.

Christendat, D. *et al.* (2000). *Nature Struct. Biol.* **7**, 903–909.

Derewenda, Z. S. (2011). *Acta Cryst.* D**67**, 243–248.

Dessailly, B. H., Nair, R., Jaroszewski, L., Fajardo, J. E., Kouranov, A., Lee, D., Fiser, A., Godzik, A., Rost, B. & Orengo, C. (2009). *Structure*, **17**, 869–881.

Díaz-Uriarte, R. & Alvarez de Andrés, S. (2006). *BMC Bioinformatics*, **7**, 3.

Fang, J., Dong, Y., Williams, T. D. & Lushington, G. H. (2008). *J. Bioinform. Comput. Biol.* **6**, 223–240.

Fang, J., Koen, Y. M. & Hanzlik, R. P. (2009). *BMC Chem. Biol.* **9**, 5.

Gabanyi, M. J. *et al.* (2011). *J. Struct. Funct. Genomics*, **12**, 45–54.

Garrard, S. M., Longenecker, K. L., Lewis, M. E., Sheffield, P. J. & Derewenda, Z. S. (2001). *Protein Expr. Purif.* **21**, 412–416.

Genest, C., Weerahandi, S. & Zidek, J. V. (1984). *Theory Decis.* **17**, 61–70.

Goh, C.-S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G. T., Zhao, H. & Gerstein, M. (2004). *J. Mol. Biol.* **336**, 115–130.

Goldschmidt, L., Cooper, D. R., Derewenda, Z. S. & Eisenberg, D. (2007). *Protein Sci.* **16**, 1569–1576.

Gómez García, I., Oyenarte, I. & Martínez-Cruz, L. A. (2011). *Acta Cryst.* F**67**, 349–353.

Gómez García, I., Stuiver, M., Ereño, J., Oyenarte, I., Corral-Rodríguez, M. A., Müller, D. & Martínez-Cruz, L. A. (2012). *Acta Cryst.* F**68**, 1198–1203.

Jahandideh, S. & Mahdavi, A. (2012). *J. Theor. Biol.* **306**, 115–119.

Jaroszewski, L., Slabinski, L., Wooley, J., Deacon, A. M., Lesley, S. A., Wilson, I. A. & Godzik, A. (2008). *Structure*, **16**, 1659–1667.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. & Lu, Z. (2007). *Nucleic Acids Res.* **35**, 339–344.

Kandaswamy, K. K., Chou, K.-C., Martinetz, T., Möller, S., Suganthan, P. N., Sridharan, S. & Pugalenthi, G. (2011). *J. Theor. Biol.* **270**, 56–62.

Kandaswamy, K., Pugalenthi, G., Suganthan, P. N. & Gangal, R. (2010). *Protein Pept. Lett.* **17**, 423–430.

Kurgan, L., Razib, A. A., Aghakhani, S., Dick, S., Mizianty, M. & Jahandideh, S. (2009). *BMC Struct. Biol.* **9**, 50.

Lee, C.-K., Cheong, C. & Jeon, Y. H. (2010). *FEBS Lett.* **584**, 675–680.

Li, W. & Godzik, A. (2006). *Bioinformatics*, **22**, 1658–1659.

Liaw, A. & Wiener, M. (2002). *R News*, **2**(3), 18–22.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). *J. Mol. Biol.* **196**, 641–656.

Mizianty, M. J. & Kurgan, L. (2009). *Biochem. Biophys. Res. Commun.* **390**, 10–15.

Mizianty, M. J. & Kurgan, L. (2011). *Bioinformatics*, **27**, i24–i33.

Overton, I. M., Padovani, G., Girolami, M. A. & Barton, G. J. (2008). *Bioinformatics*, **24**, 901–907.

Oyenarte, I., Lucas, M., Gómez García, I. & Martínez-Cruz, L. A. (2011). *Acta Cryst.* F**67**, 318–324.

Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M. & Lundegaard, C. (2009). *BMC Struct. Biol.* **9**, 51.

Price, W. N. *et al.* (2009). *Nature Biotechnol.* **27**, 51–57.

Savitsky, P., Bray, J., Cooper, C. D., Marsden, B. D., Mahajan, P., Burgess-Brown, N. A. & Gileadi, O. (2010). *J. Struct. Biol.* **172**, 3–13.

Slabinski, L., Jaroszewski, L., Rodrigues, A. P., Rychlewski, L., Wilson, I. A., Lesley, S. A. & Godzik, A. (2007). *Protein Sci.* **16**, 2472–2482.

Smialowski, P., Schmidt, T., Cox, J., Kirschner, A. & Frishman, D. (2006). *Proteins*, **62**, 343–355.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. & Feuston, B. P. (2003). *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley–Interscience.

Xiao, R. *et al.* (2010). *J. Struct. Biol.* **172**, 21–33.

Yen, S.-J. & Lee, Y.-S. (2009). *Exp. Syst. Applic.* **36**, 5718–5727.

Yu, D.-J., Hu, J., Tang, Z.-M., Shen, H.-B., Yang, J. & Yang, J.-Y. (2013). *Neurocomputing*, **104**, 180–190.

Zhang, Y., Zhang, D., Mi, G., Ma, D., Li, G., Guo, Y., Li, M. & Zhu, M. (2012). *Comput. Biol. Chem.* **36**, 36–41.