BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 23 2008, pages 2691–2697
doi:10.1093/bioinformatics/btn538

Sequence analysis

Protease substrate site predictors derived from machine learning on multilevel substrate phage display data

Ching-Tai Chen[1,2,†], Ei-Wen Yang[1,†], Hung-Ju Hsu[3,4], Yi-Kun Sun[3], Wen-Lian Hsu[1,*] and An-Suei Yang[3,*]

[1]Institute of Information Science, Academia Sinica, Taipei 115, [2]Institute of bioinformatics, National Chiao Tung University, Hsin Chu 300, [3]Genomics Research Center, Academia Sinica, Taipei 115 and [4]Graduate Institute of Life Sciences, National Defense Medical University, Taipei 114, Taiwan

Received on August 27, 2008; revised on October 9, 2008; accepted on October 10, 2008

Advance Access publication October 29, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Regulatory proteases modulate proteomic dynamics with a spectrum of specificities against substrate proteins. Predictions of the substrate sites in a proteome for the proteases would facilitate understanding the biological functions of the proteases. High-throughput experiments could generate suitable datasets for machine learning to grasp complex relationships between the substrate sequences and the enzymatic specificities. But the capability in predicting protease substrate sites by integrating the machine learning algorithms with the experimental methodology has yet to be demonstrated.

Results: Factor Xa, a key regulatory protease in the blood coagulation system, was used as model system, for which effective substrate site predictors were developed and benchmarked. The predictors were derived from bootstrap aggregation (machine learning) algorithms trained with data obtained from multilevel substrate phage display experiments. The experimental sampling and computational learning on substrate specificities can be generalized to proteases for which the active forms are available for the *in vitro* experiments.

Availability: http://asqa.iis.sinica.edu.tw/fXaWeb/

Contact: hsu@iis.sinica.edu.tw; yangas@gate.sinica.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Proteases—a class of enzyme coded in about 2% of the genes in the human genome—participate in large varieties of biological activities both in maintaining health and in causing diseases (see for example, Hedstrom, 2002; Jin and El-Deiry, 2005; Marnett and Craik, 2005; Packard and Komoriya, 2008; Pissarnitski, 2007). The members in this class of enzyme recognize specific substrate protein sequences and catalyze the hydrolysis of designated peptide bonds to activate or degrade the substrate proteins. The effects of the hydrolysis reactions are frequently amplified, resulting in rapid and substantial change of the biological systems through modulating the balance of proteomic dynamics. Despite their biological importance, the substrate specificities for the majority of the proteases remain incompletely understood, hampering critical understanding of the biological functions of the proteases and the capabilities in designing inhibitors for the proteases as therapeutics.

The origin of the substrate specificities of the proteases can be rationalized through the interaction of the peptidyl substrates with the active sites (Tyndall *et al*., 2005), but complete characterization of a protease's specificity requires extensive *in vitro* experiments. Although high-throughput methods have been developed to quantify protease specificity against a large number of substrates (Gosalia *et al*., 2005; Salisbury *et al*., 2002), enumerating all the possibilities in the substrate sequence space remains experimentally intractable. Combinatorial peptide libraries have been constructed for protease specificity analyses (Harris *et al*., 2000; Marnett and Craik, 2005), but the experimental premise has to rely upon the assumptions that the protease substrate recognition subsites are independent and that the substrate protease binding modes are unchanged regardless of the variation of the substrate sequences. Increasing body of evidence suggests that both assumptions are problematic (Brandstetter *et al*., 1996; Coombs *et al*., 1999; Ding *et al*., 2006; Hsu *et al*., 2008; Laskowski and Qasim, 2000). Consequently, it met with limited success in generalizing the experimental data to computational models capable of predicting potential substrate sites in proteomes.

Substrate phage display (Deperthes, 2002; Hsu *et al*., 2008; Matthews and Wells, 1993; Ohkubo *et al*., 2001; Smith *et al*., 1995) as a platform to characterize substrate protease specificity is not subject to the assumptions pertinent to the subsite independency and the substrate binding modes. Moreover, the phage display approach is advantageous over other high-throughput methods in that this method can effectively exclude from further investigation the amino acid sequences that have little activity as substrates for the protease. Although the scope of the specificity measurement is still subject to experimental limitation even with high-throughput platforms, machine learning processes, when adequately designed, can most effectively make use of the dataset suitable for elucidating optimal rules governing the substrate specificity.

*To whom corrrespondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

The essence of machine learning algorithms requires that the dataset for training contains both positive and negative examples implicitly covering the majority of the rules in substrate specificity recognition. But due to restriction of experimental resources, the dataset tends to be limited to a few hundred experimentally characterized learning cases. Thus, specifically designed substrate phage display procedures were required for optimal machine learning. Two levels of substrate phage display experiments (Hsu *et al*., 2008) were carried out to collect not only the most effective substrate sequences (informative positive examples)—the substrate sequences with seemingly cleavable pattern but could not be cleaved by the enzyme (informative negative examples) were also identified and used in the training procedures. Moreover, for each of the positive examples, the specificities were quantified with quantitative enzyme-linked immunosorbent assay (ELISA) (Hsu *et al*., 2008; Sharkov *et al*., 2001). The dataset containing quantitative positive and negative cases provided a novel opportunity enabling machine learning algorithms to predict regression substrate specificity matching the quantitative ELISA specificity scale.

In this work, factor Xa (fXa) was used as a model system. FXa has been an attractive target for antithrombotic therapeutics (Guertin and Choi, 2007; Hertzberg, 1994). In addition, recombinant fusion proteins are frequently expressed with a fXa cutting site designed for removing fusion domains (Jenny *et al*., 2003). Hence, prediction of the fXa specificity has important scientific and technological applications. Using artificial neural network (ANN) and support vector machines (SVM) algorithms to model experimental data from multilevel substrate phage display experiments and high-throughput enzymatic kinetics measurements, we constructed fXa specificity predictors applicable for scanning windows of sequences for potential fXa substrate sites. Ineffective learning due to imbalanced learning data was overcome with bootstrap aggregation (also known as 'bagging') (Breiman, 1996) and data resampling machine learning techniques. High prediction scores for fXa cleavage sites in known substrate proteins demonstrated that the machine learning algorithms were capable of predicting *in vivo* substrate sites with data from *in vitro* experiments. The methodology can be generalized to the proteases cleaving peptidyl substrates in the *in vitro* experiments.

2 METHODS

2.1 Experimental data from multilevel substrate phage display and quantitative substrate specificity ELISA

The experimental details can be found in a recent publication (Hsu *et al*., 2008) and in Supplementary Methods. In the data collected for this study, the largest observed k_{obs} was $70\,200\,\mathrm{min}^{-1}$. The positive substrates were defined by the k_{obs} threshold of $5000\,\mathrm{min}^{-1}$. Below this threshold, the linear regression in determining k_{obs} from the quantitative ELISA experiments yielded fluctuating coefficient of determination (R^2) (Supplementary Table 1), indicating that these k_{obs}'s were too small to be confidently determined. The k_{obs} was determined by following the enzymatic kinetics over a time course of 60 min—a labor-intensive procedure for large number of substrates. A simplified procedure of measuring the reaction completeness after 60 min of fXa cleavage was first applied to each of the sampled substrate phages so as to prescreen the phage substrates for accurate but elaborative measurement of k_{obs}. In the total 312 sequences sampled, 187 k_{obs}'s were determined. The k_{obs}'s for the 125 substrate sequences that did not pass the prescreening procedure were assigned as $0\,\mathrm{min}^{-1}$. Out of the observed k_{obs}'s, 132 sequences were determined as positive substrate sequences, defined by the k_{obs} threshold of $5000\,\mathrm{min}^{-1}$. The rest of the sequences were negative substrate sequences. This is the DS-312 dataset (Supplementary Table 1). The cost for DNA sequencing and quantitative ELISA was the only limiting factor for data collection.

2.2 Training and benchmarking machine learning algorithms

2.2.1 Ten-fold cross validation test with the DS-312 dataset The positive and negative cases in the DS-312 dataset were randomly divided into 10 equal portions. One portion was used as test set while the rest of the dataset was used as training set. After the training procedure converged to preset criteria for the machine learning algorithm, the prediction capabilities of the trained predictors were benchmarked with the test set. The process took turns to benchmark prediction accuracy on the 10 non-overlapping test sets with the predictors trained with the corresponding training set. The results of the 10-fold cross validation test were the average of the accuracy benchmarks (see below for the prediction accuracy measurements as predictability benchmarks).

2.2.2 Artificial neural network Standard feed-forward back-propagation neural network (Rumelhart *et al*., 1986) was used to learn the weight of the network by employing gradient descent to minimize the sum of squared error between the network output values and the target values. Six-residue sequence fragments were used as inputs for the ANN. Each of the amino acid types was encoded with 11 physiochemical properties (Liu *et al*., 2006) from AAindex (Kawashima and Kanehisa, 2000): The input layer consisted of 66 nodes (11 properties for each residue in a 6mer peptide). The only hidden layer contained eight nodes. The output layer had a single node for normalized k_{obs} value. Two types of normalized k_{obs} value were used to encode the output feature—first, two-class (classification) prediction: if k_{obs} was greater or equal to $5000\,\mathrm{min}^{-1}$, the output was 1, otherwise the output was encoded as 0; second, real-value (regression) prediction: if k_{obs} was smaller than or equal to 0, the output is 0, otherwise k_{obs} was normalized by a sigmoid-like function, *sf*.

$$sf(k_{\mathrm{obs}}) = \left[1 + \exp - \left(\frac{k_{\mathrm{obs}}}{5000\,\mathrm{min}^{-1}}\right)\right]^{-1} \quad (1)$$

The normalization function exaggerated the separation of the $k_{obs} > 5000\,\mathrm{min}^{-1}$ from the k_{obs} that were equal to 0, while still distinguishing larger k_{obs} from smaller k_{obs}. Learning rates of the hidden layer and the output layer were 0.055 and 0.035, respectively. The training iteration was stopped as the mean absolute error between the ANN output values and the target values converged below the threshold of 0.05. The parameter set, normalization function and the architecture of ANN were determined empirically for optimal performance.

2.2.3 Support vector machines SVM is a two-class classification approach with a maximized-margin hyperplane, where margin is the distance from the separating hyperplane to the closest data point (Burges, 1998). Each natural amino acid type was encoded with 20 bits of binary string (1 for the bit corresponding to the amino acid type; 0 for others)—120 bits encoded a six-residue input sequence i as a vector X_i. The multidimensional hyperplane was created to separate a set of complex feature vectors X_i into binary labeled classes Y_i. Here, the Y_i output was encoded as either positive or negative fXa substrate. In non-linear separable cases, a maximum-margin hyperplane can be obtained after uniquely transforming the input variables, via a non-linear mapping, to a high-dimensional kernel space. The radial basis function (RBF) kernel implemented in the LIBSVM software package (Chang and Lin, 2001):

$$\Phi(\mathrm{X}_i, \mathrm{X}_j) = \exp(-\gamma ||\mathrm{X}_i - \mathrm{X}_j||^2) \quad (2)$$

was used to calculate the maximum-margin hyperplane in the kernel space. The cost (c) and gamma (γ) parameters of the SVM were optimized with

grid searching for the optimal accuracy using only the training dataset. The trained SVM predictors were then applied to the test set for accuracy benchmarking. The confidence level probabilities estimated by the associated LIBSVM utility (Wu *et al*., 2004) were used as the normalized k_{obs} real-value (regression) predictions.

2.2.4 Rule-based (Naïve) predictors Three different rule-based (naïve) predictors (NP) were constructed to perform prediction by pattern matching: XXXRXX (NP^1), XXGRXX (NP^2) and XRXRXX (NP^3), where X represents any natural amino acid. These patterns were derived from the positive substrate sequences in DS-312. Only the sequences matched the patterns were predicted to be positive from the corresponding naïve predictor. NP^1 covered almost all the positive substrate sequences in DS-312 dataset and was expected to predict all the true positives along with many false positives, while NP^2 and NP^3 were more specific in predicting true positives along with many false negatives. The naïve predictors were useful in classification prediction and the results were comparable with ANN and SVM classification predictions. But the naïve predictors were not applicable in realistic sequence scanning for protease substrate sites in natural protein sequences; such task requires real-value (regression) predictors derived from ANN or SVM (see above).

2.2.5 Prediction capacity benchmarking For two-class classification tasks, performances were benchmarked by accuracy (Acc), recall (Rec), precision (Pre), *F*-score (Fsc) (Manning *et al*., 2007) and Matthew's correlation coefficient (MCC) (Matthews, 1975).

$$\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN} \times 100; \tag{3a}$$

$$\text{Rec} = \frac{TP}{TP+FN} \times 100; \tag{3b}$$

$$\text{Pre} = \frac{TP}{TP+FP} \times 100; \tag{3c}$$

$$\text{Fsc} = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre}+\text{Rec}} \tag{3d}$$

$$\text{MCC} = \frac{TP \times TN + FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3c}$$

where TP is the number of true positives; TN the number of true negatives; FP the number of false positives and FN the number of false negatives. Recall can be viewed as a measurement of completeness, whereas precision is a measurement of exactness or fidelity. There is an inverse relationship between recall and precision; therefore, *F*-score, the weighted harmonic mean of the two, combines them with equal weight. MCC, as a measure of the quality of two-class classifications, is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. Its value ranges between 0 and 1; random correlation gives 0 MCC while perfect correlation yields 1 MCC. AUC is the area under ROC (receiver operating characteristic) curve, where the *y*-axis is Rec/100 and the *x*-axis is (1-Pre)/100. AUC of 1 represents a perfect predictor with both maximal specificity and sensitivity; random prediction with no discrimination power yield AUC of 0.5.

For real-value regression prediction, the accuracy results of the predictors were benchmarked by mean absolute error (MAE), root mean squared error (RMSE) and Pearson's correlation coefficient (PCC)

$$\text{MAE} = \frac{1}{N}\sum_i |X_i - Y_i|; \tag{4a}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_i (X_i - Y_i)^2}; \tag{4b}$$

$$\text{PCC} = \frac{\sum_i X_i Y_i - \frac{1}{N}\sum_i X_i \sum_i Y_i}{\sqrt{\left(\sum_i X_i^2 - \frac{1}{N}\left(\sum_i X_i\right)^2\right) \times \left(\sum_i Y_i^2 - \frac{1}{N}\left(\sum_i Y_i\right)^2\right)}} \tag{4c}$$

X_i and Y_i stand for the real and the predicted value, respectively. PCC ranges between -1 and 1; perfect correlation yields 1 PCC while negative PCC represents inverse correlation between the real and predicted values.

2.3 Predicting protease substrate sites

2.3.1 Predictors for scanning fXa substrate sites in protein sequences The accuracy benchmarks of the predictors trained on the DS-312 dataset revealed the feasibility of the multilevel phage display strategy and the extent of the prediction power of the computational models. However, since the DS-312 dataset contained substrate sequences mostly biased toward the positive sequence pattern with Arg at the P1 position, the predictors trained so far were not suitable for scanning protein sequences where on average ∼19 out of 20 six-residue sequences did not contain Arg at the P1 position. This problem was remedied by including simulated negative cases to both training and test datasets to mimic the realistic protein sequence compositions encountered in scanning for potential fXa substrate sites, such that the sequence information for most negative cases that had been excluded from the surveys of the multilevel phage display experiments could be included.

A simulated negative sequence was a randomly generated six-residue sequence segment based on the occurrence probabilities for each of the amino acid types in natural protein sequences; the associated k_{obs} of the simulated negative sequence was set to $0\,\text{min}^{-1}$. Two machine learning aspects were consequential due to the inclusion of the simulated negative cases: First, there were false negatives in the simulated negative substrate sequences where the actual k_{obs} could be larger than the threshold of $5000\,\text{min}^{-1}$. The fraction of the false negative cases had an upper limit of ∼1/20 in the simulated negative cases (with the Arg at the P1 position), and thus the noise due to the false negative cases was expected to be largely rectified by the positive cases in the training set during the machine learning iterations. A series of control machine learning experiments with increasing portion of XXXRXX cases in the training set were conducted to investigate the relationship between the noise level and the machine learning benchmarks and to identify the threshold fraction of the false negative cases resulting in deteriorating predictor benchmarks. Second, the introduction of overwhelmingly large number of simulated negative substrate sequences would cause imbalance of the training data, resulting in computational models ignoring the minority positive cases. This problem was circumvented with bootstrap aggregation and data resampling techniques as described below.

2.3.2 Bootstrap aggregation (BAGGING) This is a machine learning technique for generating multiple versions of a predictor and using them to produce an ensemble of prediction results (Breiman, 1996). Each individual classifier in the predictor ensemble was trained with a different sampling (bag) of the training set, and the final prediction was calculated by averaging with equal weight the output values from the predictors (Manning *et al*., 2007). Data resampling was carried out first by randomly dividing DS-312 dataset into 10 bins for 10-fold cross-validation tests. In contrast to the previous training/testing procedure, the training set comprised a set of bags (50 bags for ANN and 10 bags for SVM). Each bag contained all the 119 positive sequences and randomly sampled 30 negative ones from the nine bins designated for training. Simulated negative cases twice as many as the positive ones were also included in each of the training sets (bags). As such, 100 times of simulated negative cases over the positive substrates were used to train the set of ANN predictors, while still avoided learning problems resulting from imbalance dataset. The proportion of different kinds of training cases resampled in this study was found to be the most effective for predictor optimization. The test set comprised 100 simulated amino acid sequences of 150 residues each; each of the sequences contained five positive

and two negative substrate sequences randomly obtained from the test dataset (the collection of 10-folds of the 100 simulated test sequences was dubbed DS-Exp52), or three positive and six negative substrate sequences from the test dataset (this collection of 10-folds of the 100 simulated sequences was dubbed DS-Exp36). The goal for the machine learning predictors was to discriminate the true positive substrate sequences from true negative cases with as few false positives and false negatives as possible.

3 RESULTS AND DISCUSSION

3.1 Multilevel substrate phage display

In principle, the training dataset for machine learning algorithms needs to contain as many data points as possible for both positive and negative cases. In reality, the cost for unlimited DNA sequencing and quantitative ELISA measurement is prohibitive, and thus the strategy for positive and negative data sampling is critically consequential for the prediction capabilities of the trained computational models. We set our limit to ∼300 experimental data points for both positive and negative cases combined for the protease. This experimental limit was considered acceptable in most experimental setup. Figure 1 depicts the procedure of the substrate phage display. The first-level substrate phage display selection–amplification cycles used pC-4X library (Supplementary Methods) with consecutive four random residues for enzymatic cleavage—there was no preconception in designing the display library. A modest group of 32 sequences were sampled and the k_{obs}'s were determined at the end of the selection–amplification cycles. At this level of substrate phage display experiment, the goal was to effectively identify positive substrate sequence patterns, and thus only the selection procedure for positive substrate sequences was carried out. Consequently, 29 out of 32 sampled sequences had $k_{obs} > 0$, and 18 out of these 29 (62%) substrate sequences belonged to only two sequence patterns: XXGR and XRXR, where the peptide bond after the C-terminal Arg side-chain was the scissile bond (Hsu *et al*., 2008). Herein, we adapted a more stringent criterion of $k_{obs} > 5000\,\text{min}^{-1}$ for positive substrates; 16 sequences were defined as positive substrates and 14 (89%) sequences belonged to the two most prominent patterns for positive substrate sequences.

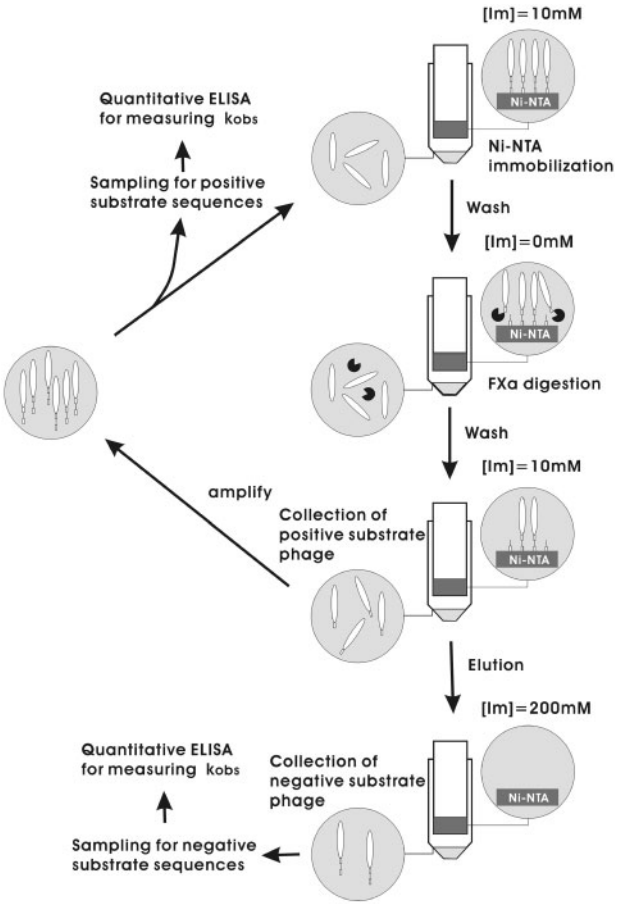

Fig. 1. Flow chart for the multilevel substrate phage display selection–amplification cycle. Details of the methodology can be found in Supplementary Methods (Hsu *et al*., 2008).

Based on the positive substrate patterns above, two second-level substrate phage display libraries were constructed: pC-XXGRXX and pC-XRXRXX (Supplementary Methods). Two extra amino acids C-terminal to the scissile bond added additional information on the P1′ and P2′ sites. The selection procedure for positive substrates yielded 116 sequences that had k_{obs} greater than the threshold of $5000\,\text{min}^{-1}$: 40% of these sequences were obtained from the pC-XRXRXX library and the rest were from the pC-XXGRXX library. Equally important, the selection procedure for negative substrate sequences yielded 125 negative substrate sequences from the two libraries: 46% of the negative substrate sequences were derived from the pC-XXGRXX library and the rest from the pC-XRXRXX library. Overall, 180 negative substrate sequences and 132 positive substrate sequence, as well as the corresponding k_{obs}'s, were sampled with the multilevel substrate phage display experiments, and the dataset was named DS-312.

3.2 Cross-validation tests on DS-312

These cross-validation tests were carried out to evaluate not only the information completeness of the dataset but also the limitation of the predictors' accuracy imposed by the information content of the training dataset. The 10-fold cross-validation benchmarks, which measured the average predictability of random 10% of the cases in a dataset with the predictors trained with the rest 90% of the dataset, on the dataset DS-312 are summarized in Tables 1 and 2 for classification and regression predictions, respectively: As expected, naïve predictors have higher recall as most of the substrate sequences in DS-312 match these patterns based on the phage display experimental design. On the other hand, ANN and SVM predictors yielded predictions with reasonable MCC. Both ANN and SVM models captured substantial substrate sequence rules—judging by the AUC of approximately approaching 0.7 (not covered by the

Table 1. Classification benchmarks from 10-fold cross-validation test on various predictors with the DS-312 dataset

Method	Acc	Rec	Pre	Fsc	MCC	AUC
ANN	68.27	49.24	67.01	56.77	0.346	0.729
SVM	67.20	48.48	65.30	55.60	0.313	0.682
NP[1]	43.27	100.00	42.72	59.87	0.084	–
NP[2]	58.65	71.97	50.80	59.56	0.210	–
NP[3]	50.32	59.09	43.58	50.16	0.030	–

naïve predictors) and MCC larger than 0.3. In addition, the PCC outcomes for the ANN and SVM are all approaching 0.5, suggesting that regression prediction models could provide reasonably accurate real-value predictions, which could not be derived from the naïve predictors.

In summary, although the accuracy of the prediction is subject to the coverage of the dataset, which had been completed only to certain extent, the 10-fold cross-validation tests suggested that the set of data points collected in DS-312 contained sufficient information to train reasonably accurate predictors to carry out realistic sequence scanning for fXa substrate sites.

3.3 Cross-validation tests on DS-Exp52 and DS-Exp36

Since the DS-312 dataset contained substrate sequences with Arg at the P1 position, the predictors trained with DS-312 were not suitable to scanning natural protein sequences where on average ∼19 out of 20 six-residue sequence segments do not contain Arg at the P1 position. This problem was remedied by including simulated negative cases to the training dataset to mimic the realistic protein sequence compositions encountered in a proteome. A simulated negative sequence was a randomly generated six-residue sequence based on the occurrence probabilities for each of the amino acid types in natural protein sequences. Two machine learning aspects were of concern due to the inclusion of the simulated negative cases in the training set: First, there were false negatives in the simulated negative substrate sequences where the actual k_{obs} could be larger than the threshold of 5000 min^{-1}. Second, the introduction of overwhelmingly large number of simulated negative substrate sequences would cause imbalance of the training data, resulting in computational models ignoring the minority positive cases. The former was proven to be insignificant; the latter was circumvented with bootstrap aggregation and data resampling (see below).

To quantitatively identify the consequence of the noises in the training dataset introduced by simulated negative cases, we used a series of control machine learning experiments to test the relationship between the artificial noise level and the accuracy of the predictors. The results indicated that only when the artificial fraction with the pattern of XXXRXX reached 10% in the training set, the prediction benchmarks began to deteriorate with MCC decreased by not more than 0.1 (from 0.766 to 0.68 for DS-Exp52 and from 0.564 to 0.53 for DS-Exp36 in the ANN control experiments). This control experiment suggested that our methodology in expanding the dataset DS-312 to DS-Exp datasets would compromise the prediction accuracy to a small extent; i.e. the MCC shown in Table 3 could not exceed the upper bound of 0.9 judging by the fraction of ∼1/20 of the XXXRXX in the simulated negative sequences.

Classification benchmarks for DS-Exp52 and DS-Exp36 are shown in Table 3. It was obvious from the table that ANN and SVM yielded comparable performance and outperformed the three naïve predictors in all prediction capability benchmarks. AUC values of both ANN and SVM were high because the prediction capabilities to discriminate true negatives from false negatives had been substantially strengthened due to the large volume of simulated negative training cases. Hence, in such a highly imbalanced dataset (only ∼5% of positive cases), Fsc or MCC could better discriminate the prediction capabilities than Acc and AUC (all above 90% in Table 3). NP^1 has recall of 100% because pattern XXXRXX is a necessary but not a sufficient condition for positive substrate sequences.

The most notorious difficulty in imbalanced training would be the undesirable consequence where the predictors neglect positive substrate sequences, resulting in mostly false positive predictions and rare true positive predictions—a situation would yield unrealistic Acc and Pre values. However, the Rec and MCC values shown in Table 3 indicated that this difficulty had been circumvented through the application of the bootstrap aggregation and data resampling techniques (see Section 2). In addition, naïve predictors were not subject to the imbalanced training difficulty as described above, and the superior performance of the ANN and SVM over the naïve predictors also supported that the bootstrap aggregation predictors were successfully trained with balanced prediction capabilities in distinguishing both negative and positive substrate sequences in realistic protein sequences.

The results presented in the previous paragraph also indicated that the ANN and SVM regression predictors could be used to construct a

Table 2. Regression benchmarks from 10-fold cross-validation test on ANN and SVM predictors with the DS-312 dataset

Method	MAE	RMSE	PCC
ANN	0.350	0.418	0.462
SVM	0.340	0.411	0.481

Table 3. Classification benchmarks from 10-fold cross-validation for the bootstrap aggregation predictors based on various algorithms with the DS-Exp52 and DS-Exp36 dataset

Method	Acc	Rec	Pre	Fsc	MCC	AUC
DS-Exp52						
ANN	98.31	83.36	71.95	77.24	0.766	0.993
SVM	98.22	78.38	72.23	75.18	0.732	0.994
NP^1	92.24	100.00	30.77	47.06	0.532	–
NP^2	97.97	72.24	69.96	71.08	0.700	–
NP^3	97.50	59.36	65.06	62.08	0.609	–
DS-Exp36						
ANN	96.99	84.13	39.36	53.63	0.564	0.985
SVM	96.78	77.23	36.82	49.86	0.520	0.984
NP^1	89.01	100.00	15.84	27.35	0.375	–
NP^2	96.86	70.07	36.54	48.03	0.492	–
NP^3	96.37	60.43	30.81	40.81	0.415	–

Table 4. Regression benchmarks from 10-fold cross-validation test for the bootstrap aggregation ANN and SVM predictors with the DS-Exp52 and DS-Exp36 datasets

Method	MAE	RMSE	PCC
DS-Exp52			
ANN	0.044	0.102	0.804
SVM	0.039	0.150	0.790
DS-Exp36			
ANN	0.049	0.116	0.640
SVM	0.065	0.212	0.600

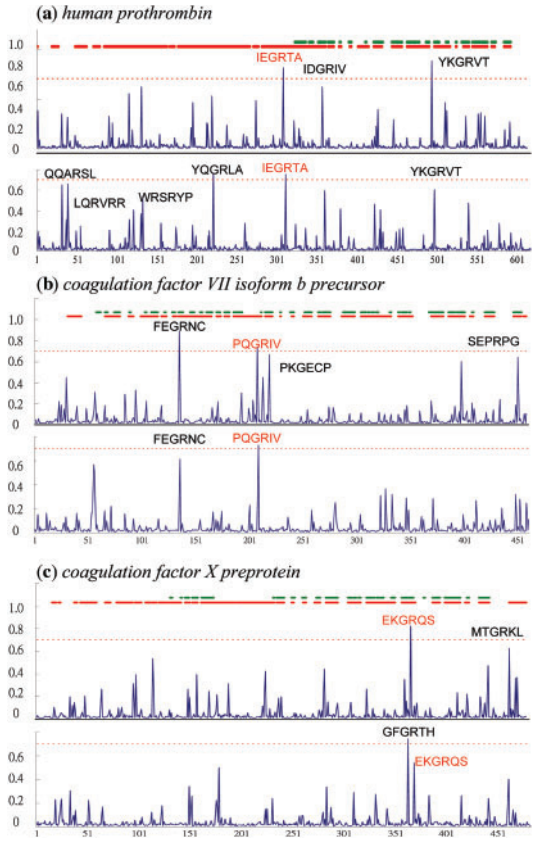

Fig. 2. Normalized k_{obs} real-value predictions for three proteins. Panels (**a**)–(**c**) were the ANN (upper panel) and SVM (lower panel) predictions for human prothrombin, coagulation factor VII isoform b precursor and coagulation factor X preprotein, respectively. Only the coil regions are highlighted: the green bar was assigned with the DSSP program on 3D structures (PDB id: 1A2C, 1DAN, and 1EZQ); the red bar was predicted with HYPROSP II (Lin *et al*., 2005). Known fXa substrate sites (as indicated by substrate sequence colored in red above the peaks) can be identified by ANN and SVM using a cutoff normalized k_{obs} value of 0.7. Substrate sequences with normalized k_{obs} value >0.6 are also labeled.

substrate sequence scanning machine on natural protein sequences with reasonable accuracy. Since naïve predictors cannot perform regression (real-value) predictions by pattern matching; only the benchmarking procedure for ANN and SVM regression predictions was carried out and the results are shown in Table 4. High PCC as well as low MAE and RMSE for both methods indicated that the enhancement of the performance in comparison with the DS-312 experiment was due to the strengthening of the prediction of the true negatives as a result of the inclusion of the simulated negative substrate sequences. The results shown in Table 4 indicated that the ANN and SVM regression predictors were reasonably accurate as a scanning device for potential substrate sequences in natural protein sequences.

3.4 Data-driven fXa substrate sequence scanning machines: case studies on fXa substrate sequences

Three known *in vivo* substrate proteins: prothrombin, coagulation factor VII isoform b precursor and coagulation factor X preprotein, for which the fXa cleavage sites are known, were used as test cases. Sequence scanning results with both ANN and SVM regression predictors are shown in Figure 2. All the previously identified fXa substrate sites: IEGRTA at position 311–316 of prothrombin, PQGRIV at position 209–214 of factor VII precursor and EKGRQS at position 369–375 of factor X preprotein (Brandstetter *et al*., 1996) have been identified with the ANN and SVM regression predictors—the predicted normalized specificity values for these sites were among the highest scorers shown in Figure 2. These predictions were remarkable in that the input information for the predictors were obtained only from the *in vitro* phage display experiments.

Some predicted positive substrate sites could be irrelevant to the function of fXa because these sites are not accessible to the protease due to the steric hindrance of the native protein structures. At least a fraction of these sites can be tentatively isolated by overlapping the substrate sequence scanning results with predicted secondary structure fragments (see Fig. 2 for the coil structure predictions of the test proteins)—only the substrate sequences observed in the tentative coil regions could be more relevant to the biological function of the protease. In addition, since our training dataset contains the majority of the specific substrate sequences but nevertheless is unlikely to include all the specificity information there is for the protease, false predictions could occur. We anticipate that the prediction accuracy can be further improved with larger collection of data points, albeit at the cost of experimental resources. Current predictions are useful as initial screening for the substrates from a large pool of protein sequences.

The prediction utilities for Figure 2 are available from the web server http://asqa.iis.sinica.edu.tw/fXaWeb/ (user's instruction included).

4 CONCLUSIONS

Substrate specificity scanning for proteases of interest are important research tools in biological and technology applications. Computational tools and databases have been established to facilitate the search of connections between the proteases and the corresponding substrates (Backes *et al*., 2005; Boyd *et al*., 2004; Garay-Malpartida *et al*., 2005; Narayanan *et al*., 2002; Rawlings *et al*., 2008; Yang, 2005). Existing applications have been relying on naturally occurring substrate sequences in training computational models for substrate predictions. Difficulties arose as the naturally occurring substrate sequences covered a small fraction of the specificity spectrum of the protease of interest, or worse when the substrate sequence data were seriously skewed or unavailable.

The results in this work supported computational models capable of automatically identifying biologically relevant substrate sequences. Tables 1 and 2 compare the similarities and contrast the differences between the two encoding methods. In two class predictions (Table 1), ANN encoding was superior to SVM encoding, while in the real value predictions (Table 2), SVM outperformed ANN. In particular, the differences are notable in Figure 2. The two encoding methods are complementary and thus the consensual results from the two types of predictions are expected to be more reliable.

The computational models constructed herein are useful proteomic tools, providing connection information between the regulatory proteases and the tentative substrate proteins in the proteome. The methodology was designed to model the substrate specificity data without the need to assume that the sequential

substrate residues are non-cooperative and independent, and thus can be generalized to any protease of interest, provided that the active form (not necessarily purified form) of the protease is available for the *in vitro* experiments.

ACKNOWLEDGEMENTS

We thank Dr KC Tsai for helpful discussions.

Funding: National Health Research Institutes [NHRI-EX95-9525EI to A.-S.Y.]; National Science Council [NSC 96-2311-B-001-030-MY3 to A.-S.Y.]; Genomics Research Center at Academia Sinica Taipei Taiwan (to A.S.Y.); National Science Council [NSC 95-3114-P-002-005-Y to W.-L.H.]; Thematic program of Academia Sinica [AS 95ASIA02 to W.-L.H.].

Conflict of Interest: none declared.

REFERENCES

Backes,C. *et al*. (2005) GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Res.*, **33**, W208–W213.

Boyd,S.E. *et al*. (2004) PoPS: a computational tool for modeling and predicting protease specificity. In *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, pp. 372–381, Stanford, CA.

Brandstetter,H. *et al*. (1996) X-ray structure of active site-inhibited clotting factor Xa. Implications for drug design and substrate recognition. *J. Biol. Chem*., **271**, 29988–29992.

Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, **2**, 121–167.

Chang,C.C. and Lin,C.J. (2001) LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm. (last accessed date February 27, 2007).

Coombs,G.S. *et al*. (1999) Revisiting catalysis by chymotrypsin family serine proteases using peptide substrates and inhibitors with unnatural main chains. *J. Biol. Chem.*, **274**, 24074–24079.

Deperthes,D. (2002) Phage display substrate: a blind method for determining protease specificity. *Biol. Chem*., **383**, 1107–1112.

Ding,X. *et al*. (2006) Direct crystallographic observation of an acyl-enzyme intermediate in the elastase-catalyzed hydrolysis of a peptidyl ester substrate: exploiting the "glass transition" in protein dynamics. *Bioorg. Chem.*, **34**, 410–423.

Garay-Malpartida,H.M. *et al*. (2005) CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics*, **21** (Suppl. 1), i169–i176.

Gosalia,D.N. *et al*. (2005) Profiling serine protease substrate specificity with solution phase fluorogenic peptide microarrays. *Proteomics*, **5**, 1292–1298.

Guertin,K.R. and Choi,Y.M. (2007) The discovery of the Factor Xa inhibitor otamixaban: from lead identification to clinical development. *Curr. Med. Chem.*, **14**, 2471–2481.

Harris,J.L. *et al*. (2000) Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. *Proc. Natl Acad. Sci. USA*, **97**, 7754–7759.

Hedstrom,L. (2002) Serine protease mechanism and specificity. *Chem. Rev*., **102**, 4501–4524.

Hertzberg,M. (1994) Biochemistry of factor X. *Blood Rev.*, **8**, 56–62.

Hsu,H.J. *et al*. (2008) Factor Xa active site substrate specificity with substrate phage display and computational molecular modeling. *J. Biol. Chem.*, **283**, 12343–12353.

Jenny,R.J. *et al*. (2003) A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa. *Protein Expr. Purif.*, **31**, 1–11.

Jin,Z. and El-Deiry,W.S. (2005) Overview of cell death signaling pathways. *Cancer Biol. Ther.*, **4**, 139–163.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.,* **28**, 374.

Laskowski,M. and Qasim,M.A. (2000) What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochim. Biophys. Acta*, **1477**, 324–337.

Lin,H.N. *et al*. (2005) HYPROSP II – a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics*, **21**, 3227–3233.

Liu,W. *et al*. (2006) Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics*, **7**, 182.

Manning,C.D. *et al*. (2007) *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Marnett,A.B. and Craik,C.S. (2005) Papa's got a brand new tag: advances in identification of proteases and their substrates. *Trends Biotechnol*., **23**, 59–64.

Matthews,B.W. (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Matthews,D.J. and Wells,J.A. (1993) Substrate phage: selection of protease substrates by monovalent phage display. *Science*, **260**, 1113–1117.

Narayanan,A. *et al*. (2002) Mining viral protease data to extract cleavage knowledge. *Bioinformatics*, **18** (Suppl. 1), S5–S13.

Ohkubo,S. *et al*. (2001) Substrate phage as a tool to identify novel substrate sequences of proteases. *Comb. Chem. High Throughput Screen*., **4**, 573–583.

Packard,B.Z. and Komoriya,A. (2008) Intracellular protease activation in apoptosis and cell-mediated cytotoxicity characterized by cell-permeable fluorogenic protease substrates. *Cell Res*., **18**, 238–247.

Pissarnitski,D. (2007) Advances in gamma-secretase modulation. *Curr. Opin. Drug Discov. Devel.*, **10**, 392–402.

Rawlings,N.D. *et al*. (2008) MEROPS: the peptidase database. *Nucleic Acids Res.*, **36**, D320–D325.

Rumelhart,D.E. *et al*. (1986) *Learning Internal Representations by Error Propagation*. MIT Press, Cambridge, MA, USA.

Salisbury,C.M. *et al*. (2002) Peptide microarrays for the determination of protease substrate specificity. *J. Am. Chem. Soc.*, **124**, 14868–14870.

Sharkov,N.A. *et al*. (2001) Reaction kinetics of protease with substrate phage. Kinetic model developed using stromelysin. *J. Biol. Chem.*, **276**, 10788–10793.

Smith,M.M. *et al*. (1995) Rapid identification of highly active and selective substrates for stromelysin and matrilysin using bacteriophage peptide display libraries. *J. Biol. Chem.*, **270**, 6440–6449.

Tyndall,J.D. *et al*. (2005) Proteases universally recognize beta strands in their active sites. *Chem. Rev.*, **105**, 973–999.

Wu,T.F. *et al*. (2004) Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.*, **5**, 975–1005.

Yang,Z.R. (2005) Mining SARS-CoV protease cleavage data using non-orthogonal decision trees: a novel method for decisive template selection. *Bioinformatics*, **21**, 2644–2650.