Predicting the Phosphorylation Sites Using Hidden Markov Models and Machine Learning Methods

Pasak Senawongse,[†] Andrew R. Dalby,[†] and Zheng Rong Yang*,[‡]

Department of Biological Science, Exeter University, U. K., and Department of Computer Science,
Exeter University, U. K.

Received February 9, 2005

Accurately predicting phosphorylation sites in proteins is an important issue in postgenomics, for which how to efficiently extract the most predictive features from amino acid sequences for modeling is still challenging. Although both the distributed encoding method and the bio-basis function method work well, they still have some limits in use. The distributed encoding method is unable to code the biological content in sequences efficiently, whereas the bio-basis function method is a nonparametric method, which is often computationally expensive. As hidden Markov models (HMMs) can be used to generate one model for one cluster of aligned protein sequences, the aim in this study is to use HMMs to extract features from amino acid sequences, where sequence clusters are determined using available biological knowledge. In this novel method, HMMs are first constructed using functional sequences only. Both functional and nonfunctional training sequences are then inputted into the trained HMMs to generate functional and nonfunctional feature vectors. From this, a machine learning algorithm is used to construct a classifier based on these feature vectors. It is found in this work that (1) this method provides much better prediction accuracy than the use of HMMs only for prediction, and (2) the support vector machines (SVMs) algorithm outperforms decision trees and neural network algorithms when they are constructed on the features extracted using the trained HMMs.

INTRODUCTION

Protein phosphorylation, performed by the protein kinases or phosphotransferases (Enzyme Commission classification 2.7), is a post-translational modification that is important to the good control of cellular processes. Those cellular processes are cell signaling,[1,2] metabolism,[3] cellular proliferation,[1] and apoptosis.[4−6] Furthermore, the association has also been found with many diseases, including cancer[1,7] and Alzheimer's disease.[6] Hence, it is important to develop quick and competent computational tools to recognize potential phosphorylation sites, which then result in an improved efficiency of characterizing novel protein sequences.[8] There is, however, a problematic issue when identifying the site of phosphorylation because the substrate specificity and phosphorylation mechanism of some protein kinases might be both broad and complex.[9,10] It has been indicated that the prediction of phosphorylation sites should not be carried out using solely a consensus sequence even when its structure was examined.[9] Nevertheless, their work showed that conserved regions, where almost complete serine, threonine, or tyrosine specificity exists, were identified between positions −4 and +4 around the phosphorylation site.

Similar to many other protein function predictions, information in the genome sequence databases has increased substantially with the completion of many eukaryotic genome sequences in this postgenomics era. However, the rate that the annotation of protein functions through experiments occurs is extremely slow. To speed up the annotation, particularly for protein functional site predictions, many computational algorithms have been developed with some success.

The *h* function method known as a frequency-based method was one of the earliest methods for functional site prediction.[11] With the *h* function, the frequency of the amino acids at each residue of a sequence is counted from the training functional sequences only. A novel sequence is classified in terms of the integration of the frequencies from all the residues. Such a method is very simple and straightforward, but the accuracy is often biased toward the functional sequences with high sensitivity and low specificity.

Subsequent studies employed neural networks. As neural networks are not able to recognize non-numerical attributes directly, a feature extraction process must be conducted before using neural networks for constructing a predictor. The common method for extracting features is to use the distributed encoding method.[12] With this method, each amino acid is encoded using a 20-bit-long binary vector, in which only one bit is assigned a unity, leaving the remaining 19 bits zeros. With this encoding method, various pattern recognition algorithms have been successfully applied to functional site prediction subjects, for instance, HIV protease cleavage site prediction,[13−15] signal peptide cleavage site prediction,[16−18] and phosphorylation site prediction.[8] The

* Corresponding author e-mail: z.r.yang@ex.ac.uk; tel.: 44-1392-26405.
[†] Department of Biological Science.
[‡] Department of Computer Science.

10.1021/ci050047+ CCC: $30.25
Published on Web 06/18/2005

most important issue with this method is the difficulty of coding biological content in sequences.[19] For instance, the Hamming distance between two binary vectors encoded from any two different amino acids is always 2, whereas the similarity quantified as mutation or substitution probability varies.[20−22].

The bio-basis function was proposed recently as an alternative encoding.[19,23] The concept of the bio-basis function is the merging of two important principles, that is, the principle of homology alignment used in biology sciences[20−22] and the principle of kernel methods used in pattern recognition algorithms.[24] The former provides a method to quantify the similarity between two sequences, whereas the latter provides a framework for learning with collected data. The core of the bio-basis function is using a set of prototype sequences to formulate a numerical feature vector for a sequence, each element of which is the similarity (homology alignment score) with one of the prototype sequences. This method overcomes the difficulty of coding biology content in sequences when using the distributed encoding method. The bio-basis function neural networks developed on the basis of this principle have been successfully applied to a number of applications including HIV protease cleavage site prediction,[19,23] trypsin cleavage site prediction,[19] the prediction of the O-linkage site in glycoproteins,[25] the prediction of phosphorylation sites,[8] the prediction of caspase cleavage sites,[26] the prediction of disordered proteins,[27] and signal peptide cleavage site prediction.[28] However, the selection of the best prototype sequences is not an easy task. This has led to considering the use of mutual information for the selection of the best prototype sequences,[29] the genetic algorithm for the selection of motifs,[8] or the orthogonal algorithm.[19] However, the increase on the prediction accuracy is not always as expected,[8] and computational cost is comparatively large.

In this paper, we have employed hidden Markov models (HMMs) to extract features for the prediction of the phosphorylation sites in proteins using various machine learning algorithms. The simulation shows that the prediction accuracy has been significantly improved.

METHODS

Hidden Markov Models. HMMs are kinds of mathematical models for modeling observation sequences that contain hidden process.[30] Because of its solid theoretic background, HMM has been widely used in many applications. There are two major components, that is, a set of states and transition probabilities between them (including the probability of being the starting state) and a signal with its probability to emit from a certain state. In applications, there are three common issues when using HMMs for modeling a family of protein sequences.[31] First, the likelihood that a sequence belongs to a model is considered. Second, an alignment between one sequence and others with a family of sequences is produced. Third, the trained model based on the known sequences is generated. A detailed description of the topology of standard linear HMMs[32] has been given.[33]

When mapping a sequence through states in a HMM model, a number of paths will eventually generate an associated probability, which is the product of all probabilities found in a path. We refer to a sequence of L amino acids as $\mathbf{x} = (x_1, x_2, ..., x_L)$ and the path of states as $\mathbf{q} = q_1, q_2, ..., q_L$. The probability that the state q_i emits the symbol x_i is denoted by $P(x_i|q_i)$, which is given by a state probability profile. Each transition connecting a state has a probability denoted by $T(q_i|q_{i-1})$. Thus, the probability of an alignment of sequences with a family of sequences by a HMM model can be represented as

$$P(\mathbf{x}|\mathbf{q}) = \prod T(q_i|q_{i-1})\, P(x_i|q_i) \qquad (1)$$

By summing up the probabilities of all paths corresponding to the sequence given a model, we can determine how well the model fits the sequences[30]

$$P(\mathbf{x}) = \sum \prod T(q_i|q_{i-1})\, P(x_i|q_i) \qquad (2)$$

Because our interest is to extract features from functional and nonfunction sequences for using a pattern recognition algorithm to construct a predictor, HMMs are used to extract features by scoring the similarity between the trained functional HMMs and those of all sequences. To derive such HMMs, we have used the HMMER package.[34]

Phosphorylation Clusters. It has been experimentally determined that phosphorylation will not happen without the presence of three amino acids, serine, threonine, or tyrosine.[8] On the basis of this biological knowledge, the sequences can be grouped into three clusters with the presence of serine, threonine, or tyrosine. We refer to these clusters as phosphorylation clusters.

Extract Features. Two types of HMM scores are used for the investigation. First, the total scores (see eq 2) obtained from the trained HMMs are used as the features. For instance, there will be K features for each novel sequence if the training sequences are clustered into K clusters and one HMM is constructed for each cluster. Second, each emission score obtained from the match states in the trained HMMs is used as a feature. For instance, there will be KL features for each novel sequence if each sequence with L amino acids and sequences are partitioned into K clusters. One HMM is constructed for each cluster. On the basis of these two types of HMM scores, three feature extraction strategies are used in our study.

(1) A feature vector of three elements is generated for each sequence. Each element is a total score (see eq 2) obtained when a sequence is aligned with one of the trained HMMs for three phosphorylation clusters.

(2) A feature vector of 12 elements is generated for each sequence. Each element is a total score obtained when a sequence is aligned with four trained HMMs for each phosphorylation cluster. Up to four different HMMs for each cluster can be built based on four kinds of alignments that

Table 1. Four Different Styles of Alignments Allowed When Generating HMMs with the *hmmbuild* Command

command	wrt sequence	wrt model	multidomain	HMMER ver.1 command
hmmbuild	local	global	yes	*hmmls*
hmmbuild -f	local	local	yes	*hmmfs*
hmmbuild -g	local	global	no	*hmms*
hmmbuild -s	local	local	no	*hmmsw*

hmmbuild provides (Table 1). For example, in the first feature extraction strategy, the default alignment style is used to perform alignments that are global with respect to the HMMs and local with respect to the sequence and allows multiple domains to hit per sequence. Such HMMs will only find complete domains.[34]

(3) A feature vector of 27 [three clusters (K) and nine residues (L)] elements is generated for each sequence. Each element is an emission score extracted when each amino acid of a sequence is aligned with each match profile in one of the trained HMMs for three clusters. Each trained HMM contains nine match profiles, and one HMM is built for one cluster on the basis of the default alignment parameter.

Classification Methods. We have two classification methods: (1) use HMMs only for classification (referred to as the *base* model) and (2) use machine learning algorithms (referred to as the *enhanced* model).

For the base model, functional and nonfunctional features obtained from aligning functional and nonfunctional sequences to trained HMMs are assumed to follow two multivariate Gaussians. In training a classifier, the functional training sequences are first used to construct HMMs. From this, the functional and nonfunctional sequences are aligned with the constructed HMMs to generate training feature vectors. Finally, a parametric method is used to estimate the density functions on the basis of all the training feature vectors. In testing, a novel sequence is aligned with the constructed HMMs to generate a novel feature vector, which is fed into the trained classifier. The Bayes rule is used for decision making.

For the enhanced model, a machine learning algorithm is used to train a classifier using the extracted features. The classifier is then used for prediction. In training a classifier, the functional training sequences are used to construct HMMs. From this, the functional and nonfunctional training sequences are then aligned with the constructed HMMs to generate training feature vectors. These training feature vectors are then used to train a classifier. In testing, a novel sequence is aligned with the trained HMMs to obtain a novel feature vector. The novel feature vector is fed to the trained classifier for prediction. The machine learning algorithms used in this study include support vector machines (SVMs),[35] back-propagation neural networks (BPNNs),[36] and decision trees.[37] The SVMs package used is the SVM*light* (http://svmlight.joachims.org).[38] The decision tree algorithm used in this study is C4.5, a free software package (http://www.mkp.com/c45).

Model Assessment. The prediction accuracy (ACC), Matthews correlation coefficient (MC),[39] true positive fraction (TPf), and true negative fraction (TNf) are used for assessment. The prediction accuracy is given by

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

where TP, FP, TN, and FN are the number of true positives, false positives, true negatives, and false negatives, respectively. MC is given as follows:

$$\mathrm{MC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})}}$$

The true positive and true negative fractions are defined as All these assessment measurements are averaged, and a

$$\mathrm{TPf} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \text{ and } \mathrm{TNf} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

standard deviation is calculated for each of them.

In addition, we use the receiver operating characteristic (ROC) curves[40] for the comparison between SVM and BPNN models because we do not have quantitative output from a decision tree algorithm. For each ROC curve, the area under the ROC curve (AUR) is calculated as a quantitative indicator for measuring the robustness of a trained model.

Data. The data are the same as that used in Berry et al.[8] The data are composed of 1840 sequences, 50% of which contain serine, threonine, and tyrosine phosphorylation sites and are referred to as functional or positive sequences and 50% of which are nonfunctional or negative sequences. Each sequence is composed of nine amino acids.

However, when inspected, there were 18 positive sequences with consistently incorrect predictions (data are not shown). These sequences have no serine, threonine, and tyrosine phosphorylation at the phosphorylation site. As it is known that phosphorylation will not happen without the presence of one of these three amino acids,[8] we then removed those 18 positive sequences from the dataset. Another two positive sequences were also randomly selected and removed in order to allow us to partition the data into five equal subsets for 5-fold cross validation. After removing the 20 positive sequences, 20 negative sequences were also randomly removed. Finally, the total data size for this study was 1800 sequences.

Because the basic principle of this study is to use features extracted from HMMs as inputs to pattern recognition algorithms, this approach requires four separate steps:

- Determining training and test sequence: Training sequences and test sequences are selected randomly. The training dataset contains 80% (5-fold cross validation) of the total data, whereas 20% is contained by the test dataset.
- HMMs: Functional sequences in each training dataset are used to construct HMMs. Each HMM consists of nine match states with the emission score for the different amino acids.
- Extracting features: This depends on different feature extraction strategies; a total score or an emission score is extracted after aligning each sequence with each trained HMM.
- SVMs, BPNN, and C4.5: Alignment scores of both functional and nonfunctional sequences in a training dataset

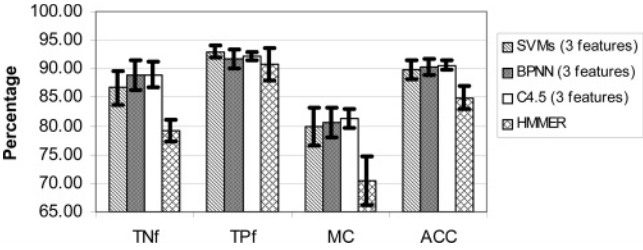

Figure 1. Bar chart comparing the mean true negative fraction (TNf), mean true positive fraction (TPf), Matthews coefficient (MC), and mean total accuracy (ACC) for the three-feature dataset for SVMs, BPNN, C4.5, and HMMER. The error bars stand for the standard deviation.

Table 2. Distribution of Types of Phosphorylated Sites among Five Datasets

	phosphorylated residues					
fold	S training	S testing	T training	T testing	Y training	Y testing
1	470	113	102	31	148	36
2	462	121	112	21	146	38
3	466	117	100	33	154	30
4	463	120	108	25	149	35
5	471	112	110	23	139	45

are used to construct predictors of SVMs, BPNN, and C4.5. Alignment scores of both functional and nonfunctional sequences in a test dataset are used to assess each predictor's performance.

RESULTS AND DISCUSSION

A five-fold cross validation was applied to this dataset. This generated five training datasets and five test datasets. The distribution of the data is shown in Table 2.

On the basis of each dataset, three HMMs are constructed using the functional sequences for three phosphorylation clusters. Because each sequence is nine amino acids long, the generated model contains nine match states with the emission score for each amino acid.

With the first feature extraction strategy, a total score generated when aligning a sequence with one of three constructed HMMs using the *hmmsearch* command is regarded as one of three features. Three features are used by the base and enhanced models. Figure 1 shows the comparison of the testing performances between the two models. It can be seen that the enhanced model consistently outperformed the base model. Among three pattern recognition algorithms, the decision tree model performed the best. In contrast, directly using HMMER for prediction yielded both the lowest performance and the least consistency. The first four rows in Table 3 give the details.

Note that the results of using SVMs are based on the use of the linear kernel (our simulation shows that the total testing accuracy is maximized when $C = 1$) for all the models with different numbers of features. Other nonlinear kernel functions did not work, with the total accuracy around 55%.

Table 3. Prediction Results When Applying Different HMM Feature Extraction Strategies Using the Enhanced and Base Methods[a]

	TNf (%)	TPf (%)	MC (%)	ACC (%)
SVMs (3 features)	87(2.99)	93(1.08)	80(3.36)	90(1.74)
BPNN (3 features)	89(2.58)	92(1.62)	81(2.67)	90(1.35)
C4.5 (3 features)	89(2.24)	92(0.68)	81(1.72)	91(0.89)
HMMER	79(1.91)	91(2.79)	70(4.28)	85(2.07)
SVMs (12 features)	87(2.38)	93(0.56)	81(2.75)	90(1.42)
BPNN (12 features)	90(2.21)	92(2.52)	81(3.66)	91(1.82)
C4.5 (12 features)	89(2.44)	91(2.50)	80(1.27)	90(0.65)
SVMs (27 features)	83(2.57)	100(0)	85(2.25)	92(1.29)
BPNN (27 features)	85(3.42)	95(4.59)	80(3.99)	90(1.97)
C4.5 (27 features)	88(2.62)	90(3.28)	78(4.75)	89(2.37)

[a] The numbers outside the parentheses stand for the measurements of the indicators, whereas the ones inside the parentheses stand for the standard deviation. We have multiplied MC by 100 because it can be easily visualized.

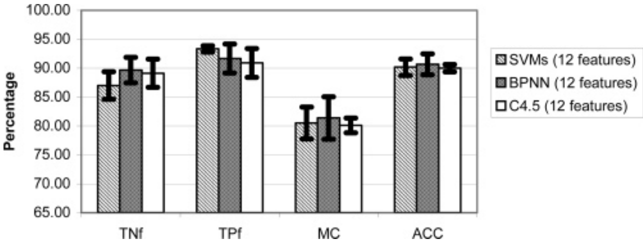

Figure 2. Bar chart comparing TNf, TPf, MC, and ACC for the 12-feature dataset for SVMs, BPNN, and C4.5. The error bars stand for the standard deviation.

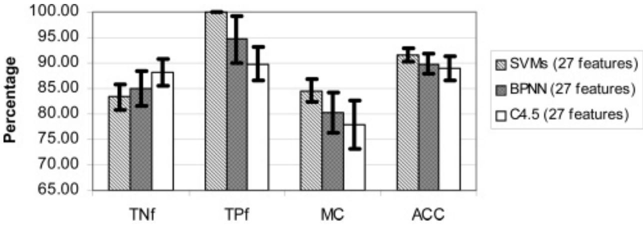

Figure 3. Bar chart comparing TNf, TPf, MC, and ACC for the 27-feature dataset for SVMs, BPNN, and C4.5. The errror bars stand for the standard deviation.

With the second feature extraction strategy, it is expected that sensitivity can be improved by increasing the number of features. Therefore, we applied different alignment styles when training HMMs using the command *hmmbuild*. We then had four HMMs for each of the three protein clusters. This led to 12 features for every sequence. Figure 2 shows the comparison of the testing performances. It can be seen that, although BPNN slightly outperformed other algorithms, C4.5 showed the most consistent prediction accuracy. The prediction accuracy of BPNN is 90.67 ± 1.82%, and the standard deviation of C4.5 is 0.65% (see the fifth, sixth, and seventh rows in Table 3).

With the third feature extraction strategy, 27 features are extracted from emission scores after aligning each sequence with nine match profiles of the trained HMMs for the three protein clusters. Figure 3 presents the testing results. It can be seen that the performance of the SVMs is the best. In particular, SVMs achieved 100% for TPf. The standard deviation of the prediction accuracy has been reduced to 1.29%. However, it should be noted that the variation between TPf and TNf in the SVM models was the largest. On the other hand, C4.5 has the least variation between TNf and TPf. The last three rows in Table 3 give the details.

Shown in Figure 4 is a summary of the comparison between different feature extraction strategies using SVMs. It can be seen that the extracting features using the emission score (27 features) gave a substantial improved performance

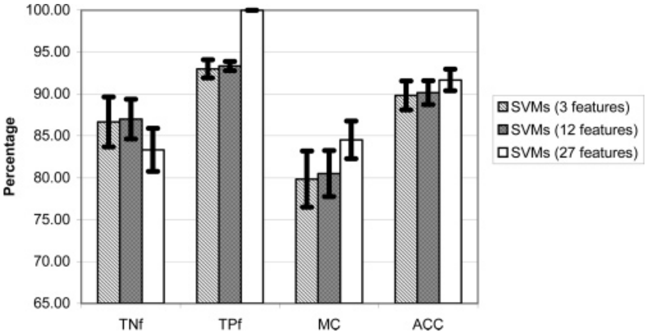

Figure 4. Bar chart comparing TNf, TPf, MC, and ACC for the 3-, 12-, and 27-feature datasets with SVMs. The error bars stand for the standard deviation.

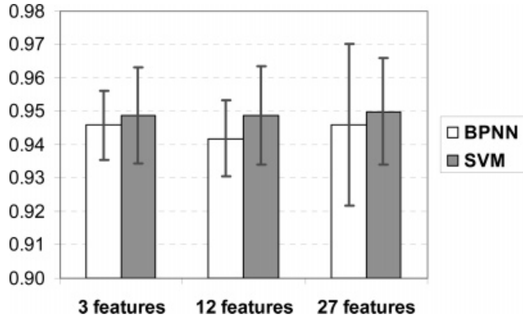

Figure 5. AUR for two groups of models.

Table 4. Comparison among Different Pattern Recognition Algorithms with HMM Features, the Distributed Encoding Method, and the Bio-Basis Functions (rBBFNN)

	TNf (%)	TPf (%)	MC (%)	ACC (%)
SVMs (HMM 27 features)	83(2.57)	100(0)	85(2.25)	92(1.29)
BPNN (HMM 12 features)	90(2.21)	92(2.52)	81(3.66)	91(1.82)
C4.5 (HMM 3 features)	89(2.24)	92(0.68)	81(1.72)	91(0.89)
BPNN (Berry et al.)[8]	88(1.94)	92(2.80)	No data	90(1.64)
C4.5 (Berry et al.)[8]	83(3.83)	98(1.06)	No data	90(2.03)
rBBFNN (Berry et al.)[8]	89(1.54)	86(2.61)	No data	88(1.50)

compared with that using the total score (3 features and 12 features). Not only were the highest ACC and MC achieved, but also, the consistency of the prediction has been improved.

The results of the prediction presented in Table 4 show a comparison between pattern recognition algorithms with features extracted using HMMs, the BPNN with the distributed encoding method (BPNNd), and bio-basis function neural networks (BBFNN). It can be seen that the use of HMMs for extracting features provided better prediction accuracy.

It should be noted that some protein sequences could contain nonexperimentally determined phosphorylation sites; because of this, some of the sequences in the study might be functional, but they are considered nonfunctional at the moment. The exclusion of these sequences may bias the constructed HMMs. The final development is to compare current nonspecific protein classification (serine, threonine, and tyrosine) with different protein kinase classification schemes.[41] This will result in more elements in the feature vector that would increase the discrimination ability between functional and nonfunctional sequences.

Shown in Figure 5 is the comparison using ROC curves. It can be seen that SVM performed better than the BPNN model, and the 27-features model performed the best. Shown in Figure 6 are the ROC curves for the BPNN and SVM models using 27 features.

SUMMARY

In this paper, we have proposed a novel method that uses the features extracted using HMMs for applying a machine learning algorithm to generate a classifier for the prediction of phosphorylation sites in proteins. It has been shown that the enhanced models outperformed the base models. This is as expected because only the functional sequences are used in the base models, which has the problem of a large number of false positives. Although the use of the machine learning algorithms in the enhanced model has encouraged the use of the nonfunctional sequences for discrimination rather than the construction of HMMs, the improvement on prediction accuracy indicates that the use of both functional and nonfunctional sequences is critically important for a good classifier that is able to discriminate between functional and nonfunctional sequences well.

The main issue in modeling protein sequences is the testing time complexity. When we use any nonparametric machine learning algorithm, a serious concern is the calculation with all the collected protein sequences. Suppose there are 10 000 sequences, we then need 10 000 calculations for obtaining 10 000 similarities. From this, an output can be made using a trained model. The principle proposed in this study is to use hidden Markov models to extract features. When a feature model has been produced, all the collected training sequences are no longer necessary for use in the testing stage. We can then regard this proposed method as a parametric model, and it is widely unknown that a parametric model is computationally cheaper.

In this work, we aimed to compare different feature extraction strategies using HMMs. Our future work will

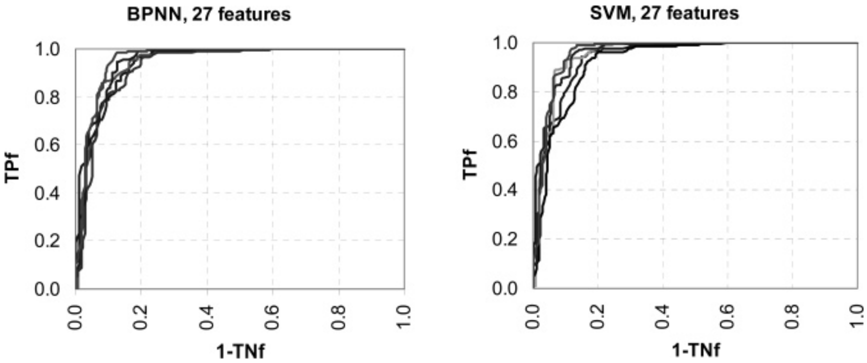

Figure 6. ROC curves for the BPNN and SVM models using 27 features.

investigate the introduction of feature selection methods to reduce the number of features and, hence, reduce the testing time further.

ACKNOWLEDGMENT

P.S. thanks Emily Berry for her help in providing the data and for advice.

REFERENCES AND NOTES

(1) Matthews, H. R. Protein kinases and phosphatases that act on histidine, lysine or arginine residues in eukaryotic prote+ins: a possible regulator of the mitogen activated protein kinase cascade. *Pharmacol Ther.* **1995**, *67*, 323−350.

(2) Li, L.; Shakhnovich, E. I.; Mirny, L. A. Amino acids determining enzyme substrate specificity in prokaryotic and eukaryotic protein kinases. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 4463−4468.

(3) Salway, J. *Metabolism at a Glance*, second ed.; Blackwell Science Ltd: Cambridge, MA, 1999.

(4) Rang, H. P.; Dale, M. M.; Ritter, J. M. *Pharmacology*, fourth ed.; Churchill Livingstone: Edinburgh, Scotland, 1999.

(5) Lewin, B. *Genes VII*; Oxford University Press Inc.: New York, 2000.

(6) Kobayashi, K.; Nakano, H.; Hayashi, M.; Shimazaki, M.; Fukutani, Y.; Sasaki, K.; Koshino, Y. Association of phosphorylation site of tau protein with neuronal apoptosis in Alzheimer's disease. *J. Neurol. Sci.* **2003**, *208*, 17−24.

(7) King, R. J. B. *Cancer Biology*, second ed.; Pearson Education Ltd: Edinburgh Gate, U. K., 2000.

(8) Berry, E. A.; Dalby, A. R.; Yang Z. R. Reduced bio-basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms. *Comput. Biol. Chem.* **2004**, *28*, 75−85.

(9) Kreegipuu, A.; Blom, N.; Brunak, S.; Järv, J. Statistical analysis of protein kinase specificity determinants. *FEBS Lett.* **1998**, *430*, 45−50.

(10) Pinna, L.; Ruzzene, M. How do protein kinases recognize their substrates? *Biochim. Biophys. Acta*. **1996**, *1314*, 191−225.

(11) Poorman, R. A.; Tomasselli, A. G.; Heinrikson, R. L.; Kezdy, F. J. A cumulative specificity model for protease from human immunodeficiency virus types 1 and 2, inferred from statisticalanalysis of an extended substrate data base. *J. Biol. Chem*. **1991**, *22*, 14554−14561.

(12) Qian, N.; Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol*. **1988**, *202*, 865−84.

(13) Cai, Y. D.; Chou, K. C. Artificial neural network model for predicting HIV protease cleavage sites in protein. *J. Protein Chem*. **1998**, *17*, 607−15.

(14) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support Vector Machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.* **2002**, *23*, 267−74.

(15) Narayanan, A.; Wu, X.; Yang, Z. R. Mining viral protease data to extract cleavage knowledge. *Bioinformatics* **2002**, *18*, s5−s13.

(16) Nielsen, H.; Engelbrecht, J.; Brunak, S.; von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **1997**, *10*, 1−6.

(17) Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*. **2004**, *340*, 783−95.

(18) Vert, J. P. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. *Pac. Symp. Biocomput. 2002* **2002**, 649−60.

(19) Thomson, R.; Hodgman, C.; Yang, Z. R.; Doyle, A. K. Characterising Proteolytic Cleavage Site Activity Using Bio-Basis Function Neural Network. *Bioinformatics* **2003**, *19*, 1741−1747.

(20) Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. A model of evolutionary change in protein. *Atlas Protein Seq. Struct.* **1978**, *5*, 345−352.

(21) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−10.

(22) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.

(23) Yang, Z. R.; Thomson, R. A novel neural network method in mining molecular sequence data. *IEEE Trans. Neural Networks* **2005**, *16*, 263−274.

(24) Duda, R. O.; Hart, P. E. *Pattern Classification and Scene Analysis*; John Wiley and Sons Inc: New York, 2002.

(25) Yang, Z. R.; Chou, K. C. Predicting the linkage sites in glycoproteins using bio-basis function neural network. *Bioinformatics* **2004**, *20*, 903−908.

(26) Yang, Z. R. Prediction of caspase cleavage sites using Bayesian bio-basis function neural networks. *Bioinformatics*, in press.

(27) Yang, Z. R.; Thomson, R.; Esnouf, R. RONN: use of the bio-basis function neural network technique for the detection of natively disordered regions in proteins. *Bioinformatics*. Accepted for publication.

(28) Yang, Z. R. Orthogonal kernel machine in prediction of functional site in protein. *IEEE Trans. Syst., Man Cybernetics B* **2005**, *35*, 100−106.

(29) Yang, Z. R.; Berry, E. A. Reduced bio-basis function neural networks for protease cleavage site prediction. *J. Bioinf. Comput. Biol.* **2004**, *2*, 1−21.

(30) Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257−286.

(31) Grundy, W. N.; Timothy, L. B.; Charles, P. E.; Michael, E. B. Meta-MEME: Motif-based Hidden Markov Models of Protein Families. *Comput. Appl. Biosci.* **1997**, *13*, 397−406.

(32) Krogh, A.; Brown, M.; Mian, I. S.; Sjölander, K.; Haussler, D. Hidden Markov models in computational biology: applications to protein modelling. *J. Mol. Biol.* **1994**, *235*, 1501−1531.

(33) Olsson, B.; Laurio, K.; Gudjonsson, L. A hybrid method for protein sequence modelling with improved accuracy. *Inf. Sci.* **2001**, *139*, 113−138.

(34) Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755−763.

(35) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.

(36) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing: Exploration in the Cognition*; MIT Press: Cambridge, MA, 1986.

(37) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Francisco, CA, 1993.

(38) Joachims, T. Making large-scale SVM learning practical. In *Advances in Kernel Methods − Support Vector Learning*; Scholkopf, B., Burges, C., Eds.; MIT Press: Cambridge, MA, 1993.

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

(40) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912−1934.

(41) Metz, C. E. Basic principles of ROC analysis. *Semin. Nucl. Med.* **1978**, *8*, 283−298.

CI050047+