JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 5, 2009
© Mary Ann Liebert, Inc.
Pp. 691–702
DOI: 10.1089/cmb.2008.0167

Predict Collagen Hydroxyproline Sites Using Support Vector Machines

ZHENG RONG YANG

ABSTRACT

Collagen hydroxyproline is an important posttranslational modification activity because of its close relationship with various diseases and signaling activities. However, there is no study to date for constructing models for predicting collagen hydroxyproline sites. Support vector machines with two kernel functions (the identity kernel function and the bio-kernel function) have been used for constructing models for predicting collagen hydroxyproline sites in this study. The models are constructed based on 37 sequences collected from NCBI. Peptide data are generated using a sliding window with various sizes to scan the sequences. Fivefold cross-validation is used for model evaluation. The best model has specificity of 70% and sensitivity of 90%.

Key words: algorithms, combinatorial proteomics, learning, machine learning, proteins.

1. INTRODUCTION

Hydroxyl groups ($-$OH) is an important chemical in nature for introducing a hydroxyl into a compound. This chemical process is called hydroxylation and is completed by enzymes called hydroxylases. Hydroxylation is a posttranslational modification activety that is oxygen-dependent. Hydroxylation can take place in proteins hydroxylating prolines and lysines. A hydroxylated proline happens to CI^3 atom and results in a hydroxyproline. The enzyme that makes hydroxyproline is called prolyl hydroxylase.

One of the hydroxyprolines is in collagens, which are long and fibrous proteins. Collagens have tough bundles which are the major components of the extracellular matrix proteins based on which most tissues are formed. Collagens are the basis for bones and teeth. The degradation of them therefore causes skin wrinkles and aging. They also play a very important role in blood vessels, the cornea, and the lens of the eye.

Hyrdoxyproline is the key to the stability of collagens (Improta et al., 2008; Krane, 2008; Palfi and Perczel, 2008). The stability of a collagen is made through hydroxylation at the β residue of a triad structure (α-β-G). In terms of biochemistry, it has been recognized that, during a hydroxylation process, a hydrogen bond is formed with water molecules, and the hydration restricts the peptide molecules surrounded by water molecules (Bella et al., 1995; Miles and Bailey, 2001; Kawahara et al., 2005).

Abnormal activity of lung collagen hydroxyproline has been found being linked with lung cancer (Sunila and Kuttan, 2006; Guruvayoorappan and Kuttan, 2008). Collagen turnover has been linked with stomach cancer (Guszczyn and Sobolewski, 2004). The high turnover of collagens was also found associated with pancreatic cancer (Palka et al., 2002). Because collagens have tensile strength, they are important for cell

School of Biosciences, University of Exeter, Exeter, United Kingdom.

adhesion and malignant tumor invasion. This has been found in colon cancer where Type I and II collagens show different patterns in healthy and malignant tissues (Bode et al., 2000). In human tumor cell line (HT-1080), the reasons for the high extent of intracellular posttranslational modifications in type IV collagens were investigated, and it was found that 4-hydroxyproline activity is higher in this cell line compared with normal human skin fibroblasts (Pihlajaniemi et al., 1981).

In studying how collagen receptor glycoprotein VI plays a role in phosphatidylinositol 3-kinase signaling, it was found that collagen-related peptide with hydroxyproline selectively induces phosphatidylinositol 3,4,5-trisphosphate and phosphatidylinositol 3,4-bisphosphate in platelets (Pasquet et al., 1999). It has also been shown that synthetic collagen-like peptides in the structure of Gly-Pro-HyP can induce various signaling pathways that regulate human platelet function (Achison et al., 1996).

Despite the importance of collagen hydroxyproline, there is no report of constructing machine learning models for predicting hydroxyproline sites in novel collagen proteins. One report predicts aromatic hydroxylation sites of cytochrome P450 substrates using structural information (Kharchevnikova et al., 2005), which may not be easily generalized to a wide range of hydroxyproline site prediction when structural information is not available.

The support vector machine (SVM) (Vapnik, 1995) has been shown to be a powerful tool in machine learning and has been successfully applied to various bioinformatics projects; therefore, it is employed in this study. In using SVM, a proper kernel function needs to be designed for a specific application. A kernel function is normally associated with a distance or similarity metric by which an input vector (or an input peptide in peptide classification) is compared with a support vector (or a support peptide in peptide classification). It outputs a similarity measure between an input vector and a support vector. The prediction made by a SVM model is a weighted similarity between an input vector and all the support vectors. Note that, in most applications, the training input vectors are the candidate support vectors. Through learning, a subset of training input vectors (candidate support vectors) are automatically selected for testing. An identity kernel and a bio-kernel are proposed for using SVM in this study.

The data used in this study are collected from NCBI. There are 37 sequences with 6920 prolines, of which 537 are experimentally annotated collagen hydroxyproline sites and 207 are inferred by similarity.

All the experimentally annotated collagen hydroxyproline sites are treated as positive data and treat all the prolines which are not yet annotated and which are not inferred as collagen hydroxyprolines by similarity as negative data. Therefore, there are two classes of data for model construction. A fivefold cross-validation approach is used for model evaluation. Because there is a large imbalance between the positive and the negative data, the randomized negative data are divided into size similar to the positive data, and each of these divisions is combined with the positive data to generate a training data set. For each such training data set, a model is built and tested. The final prediction capability of the models is assessed by the averaged prediction accuracy of all models.

2. METHODS

All the peptides generated will have the same length (i.e., the same number of residues). Denoted by Θ is a set of 20 amino acids and by R the peptide length. Each peptide is then an R-fold chain of amino acids, $\mathbf{s}_i \in \Theta^R$. A collection of all the $\in$ peptides is denoted by $\Omega = \{\mathbf{s}_i\}_{i=1}^N$.

To use SVM, a proper kernel function should be considered first. In dealing with sequence homology alignment, there are two simple metrics to score the similarity or distance between two sequences: (1) the Needleman-Wunsch score (Needleman and Wunsch, 1970), and (2) the Dayhoff score as well as its variants (Dayhoff, 1978; Altschul, 1990). The Dayhoff score is also called a "mutation matrix," which is a 20×20 matrix for protein sequences, where each entry measures the possibility that one amino acid is mutated to the other. It therefore measures the similarity between two amino acids. The Needleman-Wunsch score is binary, while the Dayhoff score is based on probability estimation. Based on these two scoring methods, two different kernel functions can be derived for using SVM.

The kernel function based on the Needleman-Wunsch score is called an "identity kernel." The core part of the identity kernel function is the residue identity between two pair-wise residues from two peptides, defined as

$$\delta(s_{ir}, s_{jr}) = \begin{cases} 0 & \text{if } s_{ir} \neq s_{jr} \\ 1 & \text{if } s_{ir} = s_{jr} \end{cases}$$

TABLE 1. IDENTITY MATRIX

	A	*C*	*D*	…	*Y*
A	1	0	0	…	0
C	0	1	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮
Y	0	0	0	…	1

TABLE 2. DAYHOFF MUTATION MATRIX

	A	*C*	*D*	…	*Y*
A	40	24	32	…	20
C	24	80	12	…	0
⋮	⋮	⋮	⋮	⋮	⋮
Y	20	32	16	…	72

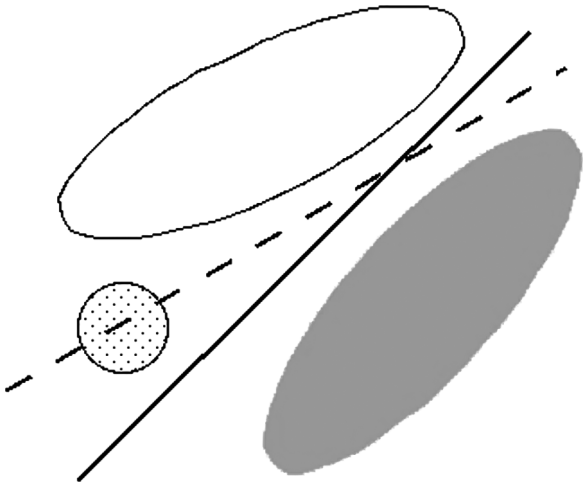

FIG. 1. An illustration of the impact of decision boundary by removing sequences with inferred hydroxyprolines. The filled ellipse indicates the data swarm of hydroxyproline peptides. The open ellipse means the data swarm of non-annotated prolines from sequences which are reserved from sequence identity check. The small circle with dots represents the data swarm of non-annotated prolines from sequences which are removed from sequence identity check. The solid line is produced by using the non-annotated proline peptides in the open ellipse and the non-annotated proline peptides in the circle as the negative data and the hydroxyproline peptides in the filled ellipse as the positive data. The dashed line is produced by using the non-annotated proline peptides in the open ellipse as the negative data and the hydroxyproline peptides in the filled ellipse as the positive data.

TABLE 3. FINAL DATA SETS FOR MODEL CONSTRUCTION

Window	*Neg*	*Pos*	*Pos∼Pos*	*Neg∼Pos*	*Neg∼Neg*	*Mod*
8	4657	416	121	535	985	11
10	5147	433	104	390	640	11
12	5373	456	81	282	522	11
14	5523	462	75	258	396	11
16	5619	468	69	238	320	12
18	5750	478	59	183	244	12
20	5791	484	53	162	224	11

Neg, number of the non-annotated peptides; Pos, number of the annotated peptides; Pos∼Pos, number of identical annotated peptides; Neg∼Pos, number of identical peptides from the opposite categories; Neg∼Neg, number of identical non-annotated peptides; Mod, number of models by the random division.

Here, s_{ir} and s_{jr} are the r^{th} residues of two peptides, $\mathbf{s}_i$ and $\mathbf{s}_j$, respectively. The identity kernel between two peptides is then defined as a polynomial function of the residue identities shown below

$$\phi^{\tau}(\mathbf{s}_i, \mathbf{s}_j) = [\alpha\rho(\mathbf{s}_i, \mathbf{s}_j) + \beta]^d$$

where α, β, and d are the parameters of the polynomial function and

$$\rho(\mathbf{s}_i, \mathbf{s}_j) = \sum_{r=1}^{R} \delta(s_{ir}, s_{jr})$$

For instance, the similarity between PRGLGPPG and LPGPGAPG is ρ(PRGLGPPG, LPGPGAPG) = 4 and ϕ^{τ} (PRGLGPPG, LPGPGAPG) = $(\alpha 4 + \beta)^d$. It must be noted that, for binary data, three kernel functions are available in SVM. They are dot product, polynomial, and sigmoid functions. In this study, it has been found that both dot product and sigmoid function did not work, leading to very low prediction accuracy.

This identity kernel function is similar to the widely used orthogonal encoding method (Qian and Sejnowski, 1998), which explicitly codes peptides to binary input vectors for using machine learning approaches for modeling. The identity kernel function codes peptides implicitly using an identity matrix (Table 1).

In fact, this identity matrix is an extreme case of many mutation matrices. The Needleman-Wunsch algorithm, which was originally developed for molecular sequence homology alignment, has been replaced by many advanced algorithms such as the Smith-Waterman algorithm (Smith and Waterman, 1981), as well as some database sequence homology alignment tools such as FASTA (Wilbur and Lipman, 1993) and BLAST (Altschul, 1990). All of these new algorithms or tools are using mutation matrices (the Dayhoff score and its variants) rather than the identity matrix for scoring sequence similarity. The simplified Dayhoff matrix (Dayhoff, 1978) is shown in Table 2. It can be seen that the relationship between any pair of amino acids is not *hard*. Instead, it becomes *softer*. The residue identity using a mutation matrix is then defined as

$$\delta(s_{ir}, s_{jr}) = M(s_{ir}, s_{jr})$$

Here, $M(s_{ir}, s_{jr})$ is a value from a mutation matrix. The bio-kernel is an exponential function defined as below

$$\phi^{\upsilon}(\mathbf{s}_i, \mathbf{s}_j) = \exp\left(\frac{\rho(\mathbf{s}_i, \mathbf{s}_j) - \rho(\mathbf{s}_j, \mathbf{s}_j)}{\rho(\mathbf{s}_j, \mathbf{s}_j)}\right)$$

where

$$\rho(\mathbf{s}_i, \mathbf{s}_j) = \sum_{r=1}^{R} M(s_{ir}, s_{jr})$$

The other difference between the two kernel functions is that the identity kernel function may not be bounded within an interval between zero and one, where one means that two peptides are identical and zero means that two peptides are completely different. However, the bio-kernel function is bounded within this interval. For a completely identical pair of peptides with R residues, the identity kernel function will output a value as $\phi^{\tau}(identical) = (\alpha R + \beta)^d$. If $\alpha = 1$, $\beta = 0$, and $d = 3$, ϕ^{τ} $(identical) = R^3$. If $R = 30$, ϕ^{τ} $(identical) = 27000$. If two peptides are completely different, ϕ^{τ} $(distinct) = 0$. However, in the bio-kernel function, ϕ^{υ} $(identical) \rightarrow 1$ and ϕ^{υ} $(distinct) \rightarrow 0$.

The bio-kernel function has been used in the bio-basis function neural networks (Thomson et al., 2003; Yang and Thomson, 2005) and has been successfully used for many peptide classification problems.

It is not intended in this study to use sequence identity check to remove sequences, because it is inappropriate for peptide classification. Sequence identity check was originally a technique used in sequence homology alignment. In order to increase specificity, one of two sequences which have identity over a certain threshold is removed. In peptide classification, peptides are the sole data and are normally much shorter than whole sequences. What it is needed is to avoid identical peptides in model construction. Two sequences may have one or more identical conserved segments, but they still have some non-conserved areas

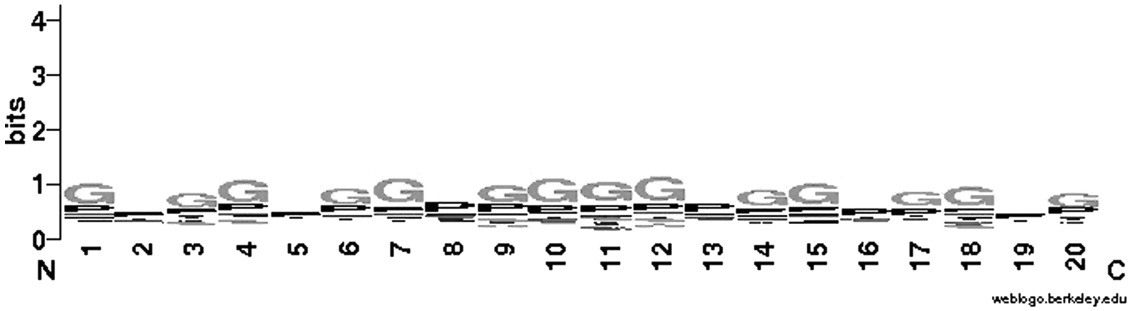

FIG. 2. The logo of 20-mer non-annotated peptides.

FIG. 3. The logo of 20-mer annotated peptides. X means unknown amino acid, which occurs when peptides are generated at two terminals of sequences.

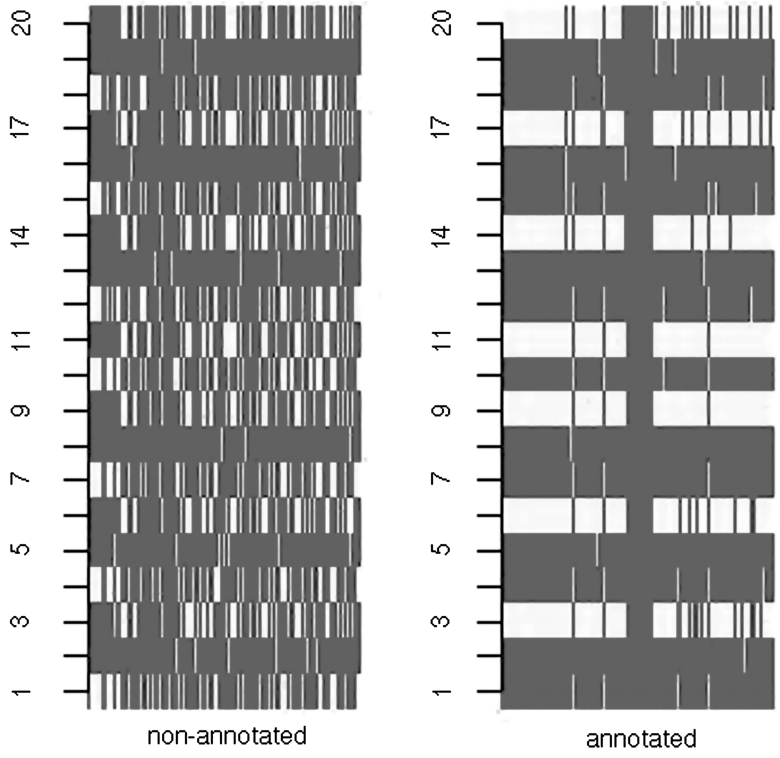

FIG. 4. Binary glycine image patterns in the non-annotated peptides and the annotated hydroxyproline peptides.

that may contain dramatic differences from each other. This means that, in a removed sequence, there might be some valuable short peptides with prolines for good discrimination. In fact, what is wanted is to look at a short peptide composed of flanking residues of a proline to predict if it is a hydroxyproline. Losing some valuable non-hydroxyprolines may bias the decision boundary of a model and therefore lead to more false positives. If the data do distribute, as shown in Figure 1, it can be seen that by removing non-annotated proline peptides from the sequences, which are removed in sequence identity check (the peptides in the circle in Fig. 1), the decision boundary will be biased (the dashed line in Fig. 1). By using this biased decision boundary, more false positives will be generated because the decision boundary at the bottom-left corner has been heavily pulled towards the area of non-annotated proline peptides. Any novel non-hydroxyprolines at this corner will be misclassified as hydroxyprolines. In Figure 1, some possible increase of misclassified hydroxyprolines is found. Because the decision boundary also has a small bias towards the area of hydroxyprolines (the top-right corner in Fig. 1), any novel hydroxyprolines in this area will also be misclassified.

3. RESULTS

3.1. Data

Sequences with hydroxyproline sites were collected from NCBI. The search keyword was "hydroxyproline." The search resulted in 295 sequences, with 10,789 prolines. Among these prolines, 980 (9%) are experimentally verified (annotated) hydroxyproline sites, and 684 (6%) are inferred by similarity. The rest, 9135 (85%), are non-annotated prolines. Among 295 sequences, only 37 (13%) were collagens. Therefore, the rest of the 258 sequences were discarded without consideration in this study. Within these 37 sequences, there are 6920 prolines. The density of prolines in collagens was 187 compared with the 15 that appeared in the rest of the 258 non-collagen sequences. Among 980 experimentally determined hydroxyproline residues, 537 (55%) were found in 37 collagens. There were 207 inferred hydroxyprolines in these 37 collagens. In 37 collagens, the ratio of experimentally verified hydroxyprolines over the inferred ones is 2.6, while it is 0.5 in the rest of the 258 non-collagen sequences. All the inferred hydroxyprolines were removed from the study, but not the whole sequences. The Supplementary Tables show all the proteins and their sites used in this study (see supplementary material online at *www.liebertonline.com*).

Among 37 sequences, 21 have no inferred hydroxyproline sites, while 16 have no experimentally verified hydroxyproline sites. Table S1 (see supplementary material online at *www.liebertonline.com*) gives the details of these sequences.

3.2. Experimental design

An odd-sized sliding window is used to scan the whole protein sequences to generate peptides with a proline in the middle residue. A peptide is denoted by

$$P_{R/2} - P_{R/2-1} - \cdots - P_1 - P_0 - P_{1'} - \cdots - P_{R/2-1'} - P_{R/2}$$

where P_0 is a hydroxyproline. Because hydroxyproline always uses a proline, the middle residue is therefore removed for computational efficiency. The peptide used for modeling is then defined as follows (N for the N-terminal and C for the C-terminal):

$$N_{R/2} - N_{R/2-1} - \cdots - N_1 - C_1 - \cdots - C_{R/2-1} - C_{R/2}$$

The sliding window size varies from 8 to 20, with a gap of 2 for this revised peptide structure.

It is obvious that there will be some identical peptides when applying the sliding window techniques to scan the sequences to generate peptides. If two identical peptides are in the same category (both are annotated collagen hydroxyproline peptides or both are non-annotated proline peptides), either one is removed. However, it is nearly impossible to avoid having two peptides from opposite categories sharing completely identical amino acids. This will happen for two reasons.

The first reason is the size of the sliding window. In the absence of proper knowledge of substrate specificity for proper binding between an enzyme and a substrate, it is hard to judge the proper size of a

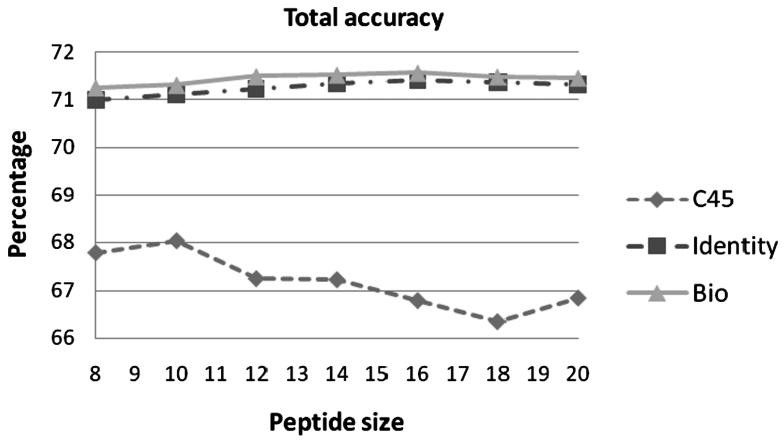

FIG. 5. The comparison between three models.

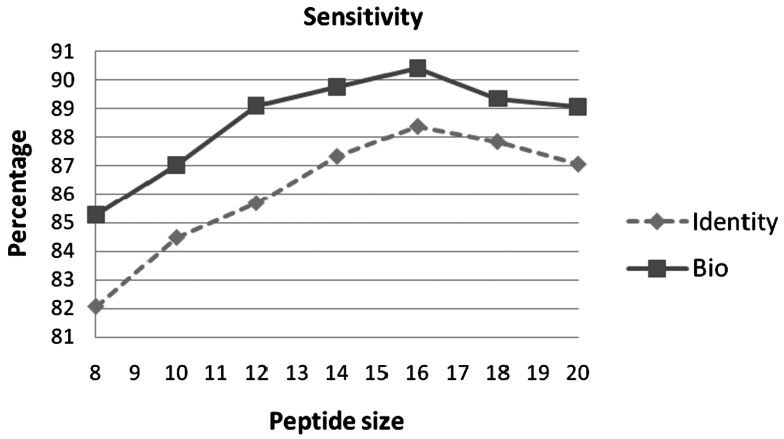

FIG. 6. Sensitivity comparison between two SVM models.

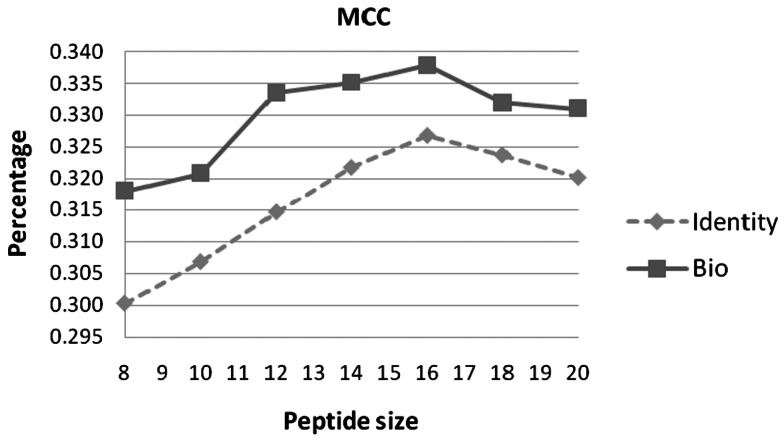

FIG. 7. MCC comparison for the two SVM models.

sliding window. It is necessary to vary the size of the sliding window to optimize the model prediction capability. When a small sliding window size is used, two peptides from the opposite categories will very likely share completely identical amino acids.

The second reason is that it is well known that some non-annotated prolines may be hydroxyprolines if more experiments can be done. Because of this, some identical peptides from different categories *may* imply that the non-annotated proline in the non-annotated peptide would be a hydroxyproline if two peptides are completely identical.

Removing both is not reasonable, as the experimentally annotated hydroxyprolines have already been in a small percentage. The non-annotated peptide is therefore removed, and the annotated peptide remains if two are identical. Table 3 details the organized data sets for model construction and evaluation (see supplementary material online at *www.liebertonline.com*).

This study aimed to test all the randomly selected non-annotated peptide data sets (i.e., each non-annotated peptide must take part in model construction and evaluation). The annotated and non-annotated peptides are first randomly divided into five folds separately. One fold of annotated peptides and one fold of non-annotated peptides are then pulled together to form one testing data set. Next, the remaining four folds of non-annotated peptides are pulled together. These non-annotated peptides are then randomly divided into 11 subsets based on the rough ratio between the non-annotated and the annotated peptides (Table 3). Each of these 11 subsets of non-annotated peptides is combined with the remaining four folds of annotated peptides to generate one training data set. There will be 11 such training data sets and 11 models in each run fivefold cross-validation. Each of these 11 models will be evaluated on the testing data set independently. This process is repeated for five times for fivefold cross-validation. It is understood that the Jackknife test is the most robust one for model evaluation, as indicated in Chou and Shan (2007) and Shan and Chou (2007). Cross-validation rather than the Jackknife test was chosen bared solely on computation burden. For fivefold cross-validation, it took 1 week to complete all simulation in a PC with 4-GC RAM and 3-GHz speed. Using the Jackknife test for this data set would have been approximately 1100 times more demanding in terms of CPU time. This work could not have been completed affordably.

Sensitivity, specificity, total accuracy, Matthews' correlation coefficient (Matthews, 1975), and receiver operating characteristics (Metz, 1978) are used for the evaluation.

SVM^{light} is used (Joachims, 1998), and two optional parameters (C and J) of SVM are varied. The C optional parameter is a trade-off between the training error and the testing error. The J optional parameter is for handling the imbalance of two categories of peptides in a data set. Through large trail-and-error simulations, it was determined to use 100 for both optional parameters of SVM^{light}. For C4.5, there is nearly no optional parameter like those in SVM^{light} that can be tuned.

3.3. Simulation results

Figure 2 shows the logo using WebLogos (Schneider and Stephens, 1990) for all the non-annotated peptides. Although glycine and proline have high frequencies compared with the other 18 amino acids, the difference between these two frequencies and the frequencies of other 18 amino acids is not significant. This means that the background information is generally random, as expected.

Figure 3 shows the logo for all the annotated hydroxyproline peptides. Glycine has a very high frequency, compared with the frequencies of the other 19 amino acids. Glycine is the dominating amino acid, particularly for N_8, N_5, N_2, C_1, C_4, C_7, and C_{10}. It is indicated in Krane (2008) that a collagen needs glycine at every third residue for the stability.

All the non-annotated and all the annotated peptides are then scanned. If a residue is glycine, a one is recorded; otherwise, a zero is recorded. This has resulted in the two glycine images shown in Figure 4. It can be seen that the annotated peptides show a quite symmetrical pattern regarding glycine. However, it is hard to see such a pattern in the non-annotated peptide glycine image.

Figure 5 shows the comparison of the total accuracy among three the models: one C4.5 and two SVM models with different kernel functions. It can be seen that the SVM models much outperformed the C4.5 model. The C4.5 model produced biased prediction accyracy, where the mean specificity was about 65%, whereas the sensitivity was about 92%. Both SVM models had specificity of about 70% and sensitivity over 85%.

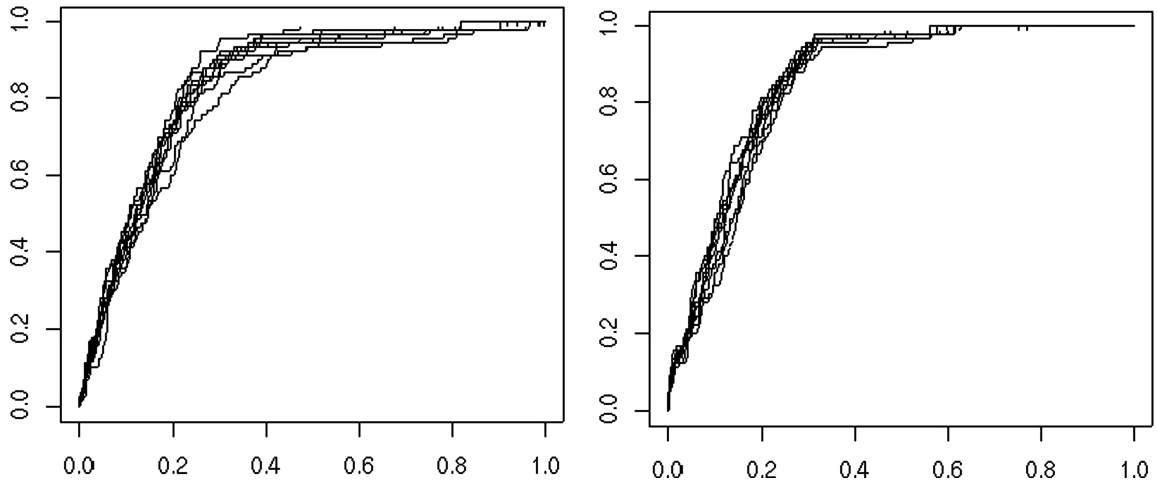

FIG. 8. The ROC curves of the two SVM models. **(Left)** ROC curves for the identity kernel SVM model. **(Right)** ROC curves for the bio-kernel SVM model.

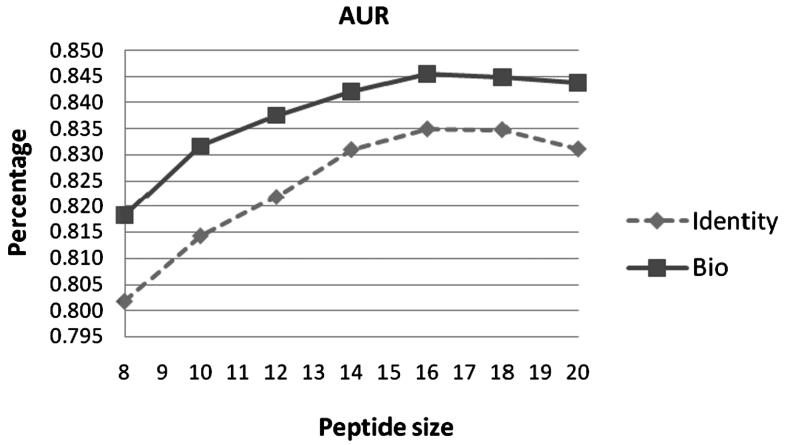

FIG. 9. AUR comparison for both the identity kernel model and the bio-kernel model for all the window sizes.

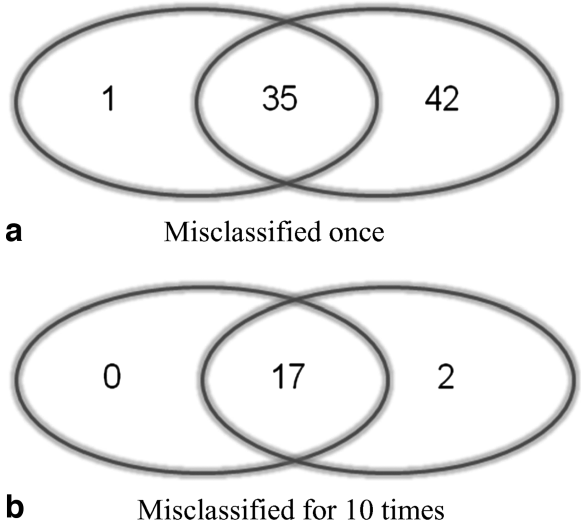

FIG. 10. The misclassified hydroxyprolines from two SVM models for the protein P30754.

Figure 6 shows the sensitivity comparison between two SVM models. It can be seen that the bio-kernel model outperformed the identity kernel model all the way through.

Figure 7 shows the comparison in terms of MCC, and the bio-kernel SVM model still outperformed the identity kernel SVM model.

Both SVM models showed that model performance was the best when the peptide size was 16. Figure 8 shows the ROC curves of the two SVM models of peptide size 16, where the horizontal axis is the false positive rate (i.e., the ratio of misclassified non-annotated peptides) and the vertical axis is the true positive rate (i.e., the ratio of correctly classified annotated peptides). Each curve was formed by connecting all the models with different threshold values for classification. It can be seen that the curves are close to the top-left corner, meaning that the models are robust.

In order to have a quantitative view of model robustness, the area under the ROC curve (AUR) was calculated for each model. It is obvious that the larger the AUR, the better the model robustness is. Figure 9 shows the AUR values for all the models, where the bio-kernel model still outperformed the identity kernel model. It also shows that model robustness was the best when the peptide size was 16.

The misclassified hydroxyproline sites using the identity kernel SVM model are analyzed. Among 21 sequences with full experimentally verified hydroxyproline sites, the hydroxyproline sites in H56978 and P02467 were never misclassified, while the hydroxyproline sites in B38623, Q28084, and P12108 were always misclassified. The hydroxyproline sites in Q25460, P85154, I56978, P85153, Q02388, P0C2W2, P12111, P05997, P08123, and P08125 (10 sequences) were misclassified with various percentages, but were never misclassified more than six times (from six models). Note that there are 11 or 12 models in each run of cross-validation a (Table 3). The misclassification rate of the hydroxyproxyline sites in P30754 remained high. It was 17% for 10 models. The details can be seen in Table S2 (see supplementary material online at *www.liebertonline.com*).

In analyzing the misclassified hydroxyproline sites using the identity kernel SVM model, different patterns are found. Three sequences (H56978, P08123, and P02467) never had their hydroxyproline sites misclassified. Only one sequence (B38623) had its hydroxyproline sites always misclassified. The hydroxyproline sites in 11 sequences (Q25460, P85154, P25508, I56978, P85153, Q02388, P12111, P05997, Q28084, P08125, and P12108) were never misclassified more than six times (models). Details can be seen in Table S3 (see supplementary material online at *www.liebertonline.com*).

Table S4 shows 78 of 105 misclassified hydroxyproline in protein P30754 (see supplementary material online at *www.liebertonline.com*). The other 27 hydroxyproline sites were never misclassified by both. Among these 78 misclassified sites, the bio-kernel SVM model only had 36. If considering a site as a misclassified one unless it has been misclassified 10 times, the identify SVM kernel model had 19 misclassified sites and the bio-kernel SVM model had 17 misclassified sites. Details are shown in Figure 10.

4. DISCUSSION

This article studied the prediction capability of collagen hydroxyprolines using the SVM with two kernel functions. For the evaluation of this prediction capability, 37 protein sequences with annotated collagen hydroxyprolines are collected from NCBI. The peptides generated from these sequences are organized using various sliding window sizes from eight to 20, with a gap of two. The prediction evaluation is based on the fivefold cross-validation approach using randomly matched training data sets to ensure every peptide anticipating the whole modeling process. The final prediction accuracy is averaged on these models. The result shows that the 16-mer peptide data is able to achieve the best prediction accuracy (with up to 70% specificity and 90% sensitivity). Further analysis has found that the bio-kernel SVM model has fewer misclassified hydroxyproline sites compared with the identity kernel SVM model.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Achison, M., Joel, C., Hargreaves, P.G., et al. 1996. Signals elicited from human platelets by synthetic, triple helical, collagen-like peptides. *Blood Coagul. Fibrinolysis* 7, 149–152.

Altschul, S.F., Gish, W., Miller, W., et al. 1990. Basic local alignment search tool. *J. Mol. Biol*. 215, 403–410.

Bella, J., Brodsky, B., and Berman, H.M. 1995. Hydration structure of a collagen peptide. *Structure* 3, 893–906.

Bode, M.K., Karttunen, T.J., Makela, J., et al. 2000. Type I and III collagens in human colon cancer and diverticulosis. *Scand. J. Gastroenterol.* 35, 747–752.

Chou, K.C., and Shan, H.B. 2007. Signal-CF: a subset-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun*. 357, 633–640.

Dayhoff, M.O. 1978. Observed frequencies of amino acid replacements between closely related proteins. *Atlas of Protein Sequence and Structure. Vol. 5. Suppl. 3.* National Biomedical Research Foundation, Washington, DC.

Guruvayoorappan, C., and Kuttan, G. 2008. Anti-metastatic effect of Biophytum sensitivum is exerted through its cytokine and immunomodulatory activity and its regulatory effect on the activation and nuclear translocation of transcription factors in B16F-10 melanoma cells. *J. Exp. Ther. Oncol*. 7, 49–63.

Guszczyn, T., and Sobolewski, K. 2004. Deregulation of collagen metabolism in human stomach cancer. *Pathobiology* 71, 308–313.

Improta, R., Berisio, R., and Vitagliano, L. 2008. Contribution of dipole-dipole interactions to the stability of the collagen triple helix. *Protein Sci*. 17, 955–961.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. *Proc. Eur. Conf. Mach. Learn.* 137–142.

Kawahara, K., Nishi, Y., Nakamura, S., et al. 2005. Effect of hydration on the stability of the collagen-like triple-helical structure of [(4R)-hydroxyprolyl-4(R)-hydroxyprolylglycine]. *Biochemistry* 44, 15812–15822.

Kharchevnikova, N.V., Dmitriev, A.V., Borodina, I.V., et al. 2005. Quantum chemical model for prediction of the site of hydroxylation of aromatic substances mediated by cytochrome P450. *Biomed. Khim*. 51, 341–355.

Krane, S.M. 2008. The importance of proline residues in the structure, stability and susceptibility to proteolytic degradation of collagens. *Amino Acids* 35, 703–710.

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

Metz, C.E 1978. Basic principles of ROC analysis. *Semin. Nuclear Med.* 8, 283–298.

Miles, C.A., and Bailey, A.J. 2001. Thermally labile domains in the collagen molecule. *Micron* 32, 325–332.

Needleman, S., and Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

Palfi, V.K., and Perczel, A. 2008. How stable is a collagen triple helix? An ab initio study on various collagen and beta-sheet forming sequences. *J. Comput. Chem*. 29, 1374–1386.

Palka, J., Surazynski, A., Karna, E., et al. 2002. Prolidase activity disregulation in chronic pancreatitis and pancreatic cancer. *Hepatogastroenterology* 49, 1699–1703.

Pasquet, J.M., Bobe, R., Gross, B., et al. 1999. A collagen-related peptide regulates phospholipase C-gamma2 via phosphatidylinositol 3-kinase in human platelets. *Biochem. J*. 342, 171–177.

Pihlajaniemi, T., Myllyl, A.R., Alitalo, K., et al. 1981. Posttranslational modifications in the biosynthesis of type IV collagen by a human tumor cell line. *Biochemistry* 20, 7409–7415.

Qian, N., and Sejnowski, T. 1998. Predicting the secondary structure of globular proteins using neural network models. *Proc. Int. J. Conf. Neural Networks* 865–884.

Semenza, G.L. 2007. HIF-1 mediates the Warburg effect in clear cell renal carcinoma. *J. Bioenerg. Biomembr*. 39, 231–234.

Schneider, T.D., and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100.

Shen, H.B., and Chou, K.C. 2007. Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem. Biophys. Res. Commun.* 363, 297–303.

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.

Sunila, E.S., and Kuttan, G. 2006. A preliminary study on antimetastatic activity of *Thuja occidentalis* L. in mice model. *Immunopharmacol. Immunotoxicol*. 28, 269–280.

Telang, S., Clem, A.L., Eaton, J.W., et al. 2007. Depletion of ascorbic acid restricts angiogenesis and retards tumor growth in a mouse model. *Neoplasia* 9, 47–56.

Thomson, R., Hodgman, T., Yang, Z.R., et al. 2003. Characterising proteolytic cleavage site activity using bio-basis function neural networks. *Bioinformatics* 19, 1741–1747.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Wilbur, W.J., and Lipman, D.J. 1993. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* 80, 726–730.

Yang, Z.R., and Thomson, R. 2005. Bio-basis function neural network for prediction of protease cleavage sites in proteins. *IEEE Trans. Neural Networks* 16, 263–274.

Address reprint requests to:
Dr. Zheng Rong Yang
School of Biosciences
University of Exeter
Prince of Wales Road
Exeter EX4 4PS, UK

E-mail: z.r.yang@ex.ac.uk