

# Recognition and Classification of Histones Using Support Vector Machine

MANOJ BHASIN, ELLIS L. REINHERZ, and PEDRO A. RECHE

## ABSTRACT

**Histones are DNA-binding proteins found in the chromatin of all eukaryotic cells. They are highly conserved and can be grouped into five major classes: H1/H5, H2A, H2B, H3, and H4. Two copies of H2A, H2B, H3, and H4 bind to about 160 base pairs of DNA forming the core of the nucleosome (the repeating structure of chromatin) and H1/H5 bind to its DNA linker sequence. Overall, histones have a high arginine/lysine content that is optimal for interaction with DNA. This sequence bias can make the classification of histones difficult using standard sequence similarity approaches. Therefore, in this paper, we applied support vector machine (SVM) to recognize and classify histones on the basis of their amino acid and dipeptide composition. On evaluation through a five-fold cross-validation, the SVM-based method was able to distinguish histones from nonhistones (nuclear proteins) with an accuracy around 98%. Similarly, we obtained an overall >95% accuracy in discriminating the five classes of histones through the application of 1-versus-rest (1-v-r) SVM. Finally, we have applied this SVM-based method to the detection of histones from whole proteomes and found a comparable sensitivity to that accomplished by hidden Markov motifs (HMM) profiles.**

**Key words:** histones, classification, support vector machine.

## INTRODUCTION

**I**N EUKARYOTES, THE DNA OF THE NUCLEUS is complexed with an equal mass of proteins forming the chromatin (van Holde, 1989). The major and chief protein components of chromatin are histones. Histones determine the structure of chromatin and play a central role in gene regulation. There are five kinds of histones in the chromatin: H1/H5, H2A, H2B, H3, and H4. Two copies of each of the H2A, H2B, H3, and H4 histones ensemble to form the core of the nucleosome, the building unit of chromatin. The DNA is wrapped around this eight-histone complex in two turns, with each turn consisting of about 80 base pairs. Nucleosomes are linked to each other by 10–90 bp of DNA which is occupied by a single copy of a linker histone—most often H1, although it can be replaced by H5 in some cell types

---

Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber Cancer Institute and Department of Medicine, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115.

(Georgel and Hansen, 2001). Linker histones are pivotal for compacting the chromatin. Stacked nucleosomes make the 10 nm chromatin fiber, which in the presence of H1 can arrange into a helical structure to form a 30 nm fiber.

Overall, histones are well conserved proteins, have a relatively low molecular weight, and have a high content of arginine/lysine residues. The nucleosomal core histones (H2A/B, H3, and H4) are structurally related (Luger *et al.*, 1997), sharing a common fold that is optimal for both histone–histone and histone–DNA interactions. A variety of DNA-binding and multimeric proteins share this same fold, whose origin can be traced back to archaeobacteria (Arents and Moudrianakis, 1995). On the other hand, linker histones (H1/H5) are longer than nucleosomal core histones and are not related to them (Kasinsky *et al.*, 2001). The fold of the linker histones consists of a winged helix followed by a nonordered carboxyl-terminal region (Ramakrishnan *et al.*, 1993). As for the nucleosome core histones, linker histones are also present in protists, but they have evolved independently of the chromosomal core histones found in archaeobacteria (Kasinsky *et al.*, 2001).

With the advent of large amounts of genomic data, highly accurate identification and classification of proteins is of prime importance. Unfortunately, annotation of histone sequences may be complicated by their highly biased amino acid sequence composition towards basic residues. Thus, sequence similarity tools such as BLAST frequently ignore histones as low-complexity regions. In response to these limitations, we developed an alternative approach for the recognition and classification of histones based on their amino acid and dipeptide composition using support vector machine (SVM).

Amino acid composition-based SVM modules were able to discriminate between histone and nonhistone proteins with an accuracy of 97.9%. The accuracy of recognition of histones improved to 98.5% using SVM models trained on dipeptide composition. For the classification of histones, SVM modules trained on dipeptide composition also outperformed those trained on amino acid composition, achieving during cross-validation an overall Matthew’s correlation coefficient (MCC) of 0.98 and an accuracy of 98.4% (MCC and accuracy of amino acid composition-based SVM modules was 95.2% and 0.93%, respectively). Furthermore, we found that detection sensitivity/specificity of histones from whole proteomes using the relevant SVM models was similar to that of hidden Markov model (HMM) profiles. The method has been implemented online for public use at <http://bio.dfci.harvard.edu/dachis/>.

## MATERIAL AND METHODS

### Datasets

Two main primary protein sequence sets were used in this study: a histone dataset and a nonhistone dataset. Only nonredundant and complete amino acid sequences were considered. The histone dataset was obtained from the histone database at NHGRI (Sullivan *et al.*, 2002) and consisted of 652 histones comprehending the five distinct histone classes (Table 1). On the other hand, the nonhistone set consisted of the amino acid sequences of 1,014 nuclear proteins other than histones. This set had been previously used for developing prediction methods of subcellular localization such as NNPSL (Reinhardt and Hubbard, 1998), SubLoc (Hua and Sun, 2001), and ESLPred (Bhasin and Raghava, 2004a).

TABLE 1. NUMBER OF FAMILY MEMBERS BELONGING TO THE DIFFERENT HISTONE CLASSES

<i>Family</i>	<i>Number</i>
H1	190
H2A	167
H2B	178
H3	37
H4	73
H5	7
Total	652

### *Amino acid and dipeptide composition*

The amino acid composition,  $f_i$ , of proteins sequences was computed using Equation (1):

$$f_i = n_i/N \quad (1)$$

where  $i$  can be any of the 20 natural amino acids,  $N$  is the total number of amino acids (sequence length), and  $n_i$  is the number of  $i$  amino acids. Thus, for any given protein, the amino acid composition calculations yield a fix length vector of 20 values.

The dipeptide composition of proteins sequence was obtained using Equation (2):

$$f_{ij} = n_{ij}/Nd. \quad (2)$$

where  $n_{ij}$  is the number of dipeptide  $ij$  in the protein, and  $Nd$  is the total number of all dipeptides in the sequence. There are  $20 \times 20$  possible dipeptide combinations ( $ij$ ), and therefore for each protein sequence the dipeptide composition calculations results in a fix vector of 400 values.

### *Support vector machine: Training and prediction*

Support vector machine was applied using the freely downloadable software package SVM<sup>light</sup> (Joachims, 1999). Computational simulations were conducted using polynomial and RBF kernels with different parameters. SVM-based models were trained on the basis of the amino acid and dipeptide composition of protein sequences. The decision function implemented by SVM during training can be defined by Equation (3):

$$f(x) = \text{sign} \left( \sum_{i=1}^N y_i \alpha_i k(x_i, x) + b \right) \quad (3)$$

where  $k$  is kernel function that defines the feature space;  $b$  is the bias value, and  $\alpha_i$  is the number obtained by solving a quadratic programming (QP) problem that gives the maximum margin hyperplane.

### *Cross-validation and performance measures*

The performance of SVM models was evaluated through a five-fold cross-validation (Hua and Sun, 2001). In this five-fold cross-validation, the positive and negative sequence datasets were randomly partitioned into five equal-sized sets. Subsequently, SVM was trained on four sets and tested on the remaining set. This procedure was carried out five times, each time using a distinct set for testing and the remaining four sets for training. The final performance was obtained by averaging the performance of the five tests. The performance of the SVM models was estimated by determining the accuracy (ACC) and Matthew's correlation coefficient (MCC) of the predictions. ACC and MCC were computed using Equations (4) and (5), respectively:

$$\text{ACC} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (4)$$

$$\text{MCC} = [(\text{TP} * \text{TN}) - (\text{FN} * \text{FP})]/\text{sqrt}[(\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TP} + \text{FP}) * (\text{TN} + \text{FN})] \quad (5)$$

where TP are true positives (e.g., histones sequences predicted as histones); FN are false negatives (e.g., histones predicted as nonhistones); TN are true negatives (e.g., nonhistones predicted as nonhistones) and FP are false positives (e.g., nonhistones predicted as histones). In order to asses the confidence of the classification of histones into their various groups using the SVM modules, we also computed the reliability index (RI) of the classification. RI was assigned according to the difference between the highest and second highest SVM output scores. The RI is defined by Equation (6).

$$\text{RI} = \begin{cases} \text{INT}(\Delta * 5/3) + 1 & \text{if } 0 \leq \Delta < 4 \\ 5 & \text{if } \Delta \geq 4 \end{cases} \quad (6)$$

*BLAST-based approach for the detection and classification of histones: Strategy and evaluation*

The detection and classification of histones using BLAST (Altschul *et al.*, 1997) was assessed through a five-fold cross validation test as follows. The positive and negative sequence datasets were partitioned in five sets as was done in the case of SVM. Subsequently, a target database was constructed from the sequences in the four sets and queried using the remaining distinct set. The process was repeated five times, so that each set can act as the query test and can be part of the target database. BLAST searches were performed at  $E = 0.001$  and default settings. The performance of the BLAST-based classification was measured by determining the ACC and MCC of the classification results. For this calculations, only the first hit was analyzed per query. Also, queries yielding a nonsignificant hit or a wrong hit (histones hitting nonhistones, or nonhistones hitting histones) were considered either as false negatives (histones hitting nonhistones) or false positives (nonhistones hitting histones). Note that when using BLAST, detection and classification of histones can be analyzed in a single step.

*Proteomewide detection of histones*

The proteomes corresponding to the translated genomes of *Anopheles gambiae*, *Caenorhabditis elegans*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Pan troglodytes*, *Rattus norvegicus*, and *Xenopus tropicalis* were obtained from ENSEMBL (Hubbard *et al.*, 2005). Both known and novel proteins from these organisms were searched for nucleosomal and linker histones using the relevant dipeptide composition-based SVM modules and HMM profiles. Detection of histones using HMM profiles was carried out using the HMMSEARCH utility from the HMMER 2.1 package (Wistrand and Sonnhammer, 2005) in combination with the HMM profiles PF00125 and PF00538—for nucleosomal core histones and linker histones, respectively—obtained from the Pfam database (Bateman *et al.*, 2004). A threshold  $E$  value of 0.01 was used for HMM profile searches.

## RESULTS AND DISCUSSION

Histones are nuclear DNA-binding proteins involved in chromatin structure and dynamics (Ehrenhofer-Murray, 2004). Consistent with their DNA-binding function, amino acid (aa) sequence composition of histones is clearly biased towards basic residues (arg/lys). This bias can complicate the detection and classification of histones using standard similarity-based tools such as BLAST and FASTA. Therefore, we have applied SVM for the detection and classification of histones. SVM is an elegant machine learning tool which has increasingly been used to capture complex biological patterns. Thus, SVM has been used for the classification of microarray data (Furey *et al.*, 2000), the prediction of subcellular localization of proteins (Hua and Sun, 2001), and the classification of protein families (Bhasin and Raghava, 2004b). Application of SVM to the recognition and classification of histones require translating the global information of their sequences into a fixed length format required by SVM. In this study, we obtained the fixed length patterns from sequences of variable length by computing their aa and dipeptide composition (see Material and Methods).

*Recognition and classification of histones using SVM: General approach*

SVM-based methods are binary classifiers, and therefore we approached the multiclass classification of histones as follows. First, we developed a binary SVM-based module for the discrimination of histones from other nuclear proteins. Subsequently, we constructed six SVM-based binary modules for the classification of histones into their respective class (H1, H2A, H2B, H3, H4, and H5). Specifically, for each histone class,  $H_i$ : H1, H2A, H2B, H3, H4, and H5, the relevant SVM module was trained with all the histones in the  $H_i$  class as positive examples and all other histones as negative examples. The SVM modules trained in this manner are known as one-versus-rest (1-v-r) SVM modules. During classification, the sequences are scored with these six distinct 1-v-r SVM modules and are assigned to the class of the SVM providing the highest output score.

TABLE 2. PERFORMANCE OF AA AND DIPEPTIDE COMPOSITION-BASED SVM MODULES IN DISCRIMINATING HISTONE FROM NONHISTONE PROTEINS

Threshold	<i>aa composition</i>		<i>Dipeptide composition</i>	
	Accuracy	MCC <sup>a</sup>	Accuracy	MCC
-1.2	88	0.78	87.04	0.77
-1	92.16	0.85	93.78	0.88
-0.8	95.16	0.9	96.94	0.94
-0.6	96.52	0.93	98.38	0.96
-0.4	97.12	0.94	98.86	0.98
-0.2	97.66	0.95	98.74	0.97
0	97.96	0.96	98.5	0.97
0.2	97.72	0.95	98.08	0.96
0.4	97.84	0.95	97.78	0.96
0.6	96.76	0.93	97.12	0.94
0.8	95.32	0.9	96.52	0.93
1	91.52	0.83	94.28	0.88
1.2	86.18	0.72	89.92	0.8

<sup>a</sup>MCC is Matthew's correlation coefficient.

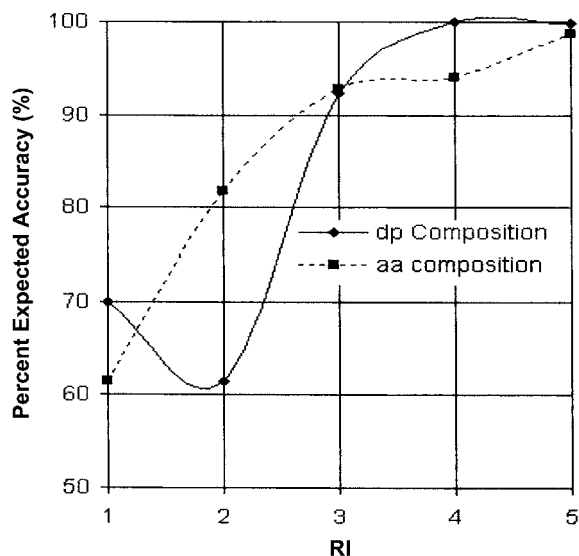
### *Recognition and classification of histones using SVM: Performance*

The SVM modules were trained on the basis of the aa and dipeptide composition of positive and negative protein samples, and their performance was evaluated through a five-fold cross-validation (see Material and Methods for details). The aa and dipeptide composition-based SVM modules achieved an ACC of 97.9% and 98.5%, respectively, for the recognition of histones at a default threshold (0) where sensitivity (percentage of correctly predicted histones) and specificity (percentage of correctly predicted nonhistones) are nearly equal (Table 2). The best results were obtained using a polynomial kernel of second degree with default regulatory parameters (C). For the classification of histones, the aa composition-based SVM modules achieved an overall ACC and MCC of 95.2% and 0.93%, respectively. The overall ACC and MCC of dipeptide composition-based SVM modules were 98.4% and 0.98%, respectively (Table 3). The SVM modules trained on dipeptide composition were able to classify the H5 histones (ACC and MCC were 85.6% and 0.65%, respectively), whereas the aa composition-based SVM modules failed to do so. This classification failure of the aa composition-based SVM modules may be due to the small number of protein members included in the H5 family. The summary of results of aa and dipeptide composition-based modules for classification of histones is shown in Table 3. In summary, dipeptide composition-based SVM modules performed consistently better than aa composition-based SVM modules at both recognition and classification of histones (Tables 2 and 3). This result indicate that dipeptide composition is a great feature to represent protein sequences in a fix format, encapsulating both the global amino acid frequency and the local amino acid order. Nevertheless, aa composition-based modules were also able to recognize and classify the histones with good accuracy, reflecting the high degree of conservation of histones within a given family.

TABLE 3. PERFORMANCE OF AA COMPOSITION AND DIPEPTIDE (DP) COMPOSITION-BASED SVM MODULES IN THE CLASSIFICATION OF HISTONES

Approach	<i>H1</i>		<i>H2A</i>		<i>H2B</i>		<i>H3</i>		<i>H4</i>		<i>H5</i>	
	ACC <sup>a</sup>	MCC <sup>a</sup>	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
aa composition	94.2	0.93	97.6	0.94	96.0	0.93	97.3	0.99	98.6	0.97	0.00	0.00
dp composition	96.8	0.97	99.4	0.99	98.8	0.99	100	1.00	100	1.00	85.7	0.65

<sup>a</sup>ACC: Accuracy of predictions; MCC: Matthew's correlation coefficient.



**FIG. 1.** Plot between the reliability index (RI) and expected accuracy. Dotted line corresponds to the aa composition-based SVM module, whereas the solid line is for the dipeptide composition-based SVM module.

For the classification of histones using the SVM modules, we have also computed the reliability index (RI) (Fig. 1). In case of the aa composition-based modules, 82.2% of sequences are predicted with  $RI \geq 4$ . Furthermore, these sequences are classified with an ACC of 96.4%. Likewise, for the dipeptide composition-based modules,  $\geq 94\%$  of the total histone sequences have a  $RI \geq 4$ , and are classified with an ACC of 99.9%.

#### *Recognition and classification of histones using BLAST*

A similarity search based module for the detection and classification of histones was constructed using BLAST. The construction and evaluation of the performance of the BLAST module is described elsewhere (Material and Methods). Under the experimental conditions, detection of histones using BLAST reached an ACC of only 89% (BLAST failed to detect 176 proteins from a total set of 1,666 proteins), lower than that of the SVM-based modules designed in this study. With regard to the classification of histones, BLAST misclassified four histones and no hits were observed for six proteins. The overall accuracy of BLAST during classification was 98.4%, close to that of the SVM-based modules, indicating that, at the classification level, the performance of the BLAST-based module is equivalent to that of the SVM module. However, it is important to note that the comparison of the SVM- and BLAST-based modules is not straightforward. BLAST-based annotation of query proteins works by the detection of hits with sequence similarity in a target database and then transferring the annotation from the protein hits to the query. Thus, the low performance of the BLAST module during the recognition experiment may in part be linked to the small size of the target database used during the five-fold cross-validation. Nevertheless, in this analysis, BLAST also failed to return hits due to low complexity filtering (especially with H1 histones). Furthermore, errors in annotation using BLAST may also occur when the target hit is not annotated properly. In contrast, the SVM-based modules developed in this study take full advantage of the annotations provided from a highly curated database. These curated annotations empower the SVM-based modules to recognize and classify the histones with high accuracy.

#### *Proteome-wide analysis of histones using SVM-based modules*

The SVM module trained on dipeptide composition was used for the detection of histone in the translated genomes of *Anopheles gambiae*, *Caenorhabditis elegans*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Pan troglodytes*, *Rattus norvegicus*, and *Xenopus tropicalis* (Table 4). The prediction results of

TABLE 4. THE PROTEOMEWIDE RECOGNITION OF HISTONES USING SVM AND HMM PROFILES<sup>a</sup>

Genome	SVM		HMM	
	Nucleosome	Linker	Nucleosome	Linker
<i>Anopheles gambiae</i>	0	5	0	5
<i>Caenorhabditis elegans</i>	1	9	0	7
<i>Gallus gallus</i>	1	6	1	6
<i>Homo sapiens</i>	1	2	1	1
<i>Mus musculus</i>	1	6	1	6
<i>Pan troglodytes</i>	5	6	5	6
<i>Rattus norvegicus</i>	3	3	3	3
<i>Xenopus tropicalis</i>	2	3	2	3

<sup>a</sup>The detection was performed at high specificity (cutoff = 1.0). The nucleosome includes H2A, H2B, H3, and H4 and the linker includes H1 and H5 histones.

SVM-based modules were compared with those obtained with profile hidden Markov models (HMM) using the HMMER package (<http://hmm.wustl.edu>). This comparative analysis revealed that the 92.8% of nucleosome histones and 92.5% of linker histones were recognized by both HMM profiles and SVM-based modules (Table 4). The SVM-based module was able to recognize three additional linker histones, plus one additional nucleosome histone (H2A). Two of the three SVM-predicted linker histones are from *Caenorhabditis elegans* (F36A2.10 and Y73B6BL.9b) and the other (ENSP00000272019) is from *Homo sapiens*. The additional nucleosome histone detected by the SVM-based module (ZC477.1) is from *C. elegans* (sequences found in supplemental file supp.doc). Annotation of the additional histones detected by SVM was checked by blasting the sequences against the nonredundant database (NR) of GenBank. Following this analysis, we determined that the human protein ENSP00000272019 detected by SVM as an H1-linker histone represents a false positive as it had no similarity to any annotated histone and contained the three cornifin domain, which are not found in histones. Of the two remaining H1-linker proteins, Y73B6BL.9b was confirmed by BLAST to be a linker histone (closer hit was GI:24636128 which is annotated as H1-isoform protein), whereas F36A2.10 BLAST hits were hypothetical proteins (best hit was GI:3876772). Likewise, BLAST hits for the additional H2A histone (ZC477.1) detected by the SVM were also hypothetical proteins (best hit was GI:7510796). In addition, after reblasting these hypothetical hits against NR, no further annotation was possible either for F36A2.10 or ZC477.1. Given that SVM was able to detect a new H1-like protein not detected by HMM profiles (Y73B6BL.9b), it would appear possible that Y73B6BL.9, ZC477.1 might be new histones not detected by HMM or BLAST. Given that these proteins were detected from virtual proteomes obtained by computer-generated translations of ensembled genomes, we carried out a BLAST search against the EST dataset of GenBank using these proteins. In this manner, we discovered that Y73B6BL.9b, ZC477.1, F36A2.10 had representative ESTs in the database (GI:30748124, GI:47594434, GI:40961907, respectively), thus discarding the possibility that these proteins are computational artifacts. In sum, despite that the SMV modules were trained in a very restricted set of positive and negative examples (only nuclear proteins were considered), they were able to detect the histones from a large and diverse set of unseen data as efficiently as HMM profiles. These results advocate for the general development and application of specialized SVM-based methods for sequence annotation.

#### *Dachis: An online web server for detection and annotation of histones*

Based on the SVM modules developed in this study, we have implemented a server for the detection and classification of histones (Fig. 2A) under the name of Dachis. The web server is hosted by the Dana-Farber Cancer Institute, and it is available at the site <http://bio.dfci.harvard.edu/dachis/>. A representative prediction result is shown in Fig. 2B. The results indicate whether the sequence is a histone. Furthermore, if the protein query is a histone, it indicates the class to which it belongs along with the confidence of the prediction results in term of a RI and expected ACC values.

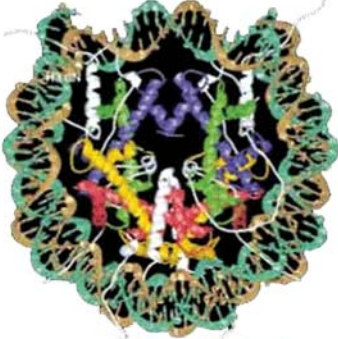
**DACHIS**  
Dipeptide Composition Based Method For Recognition & Classification of Histones

HOME | HELP | SUBMIT | CONTACT

- ▶ HOME
- ▶ SUBMIT
- ▶ HELP
- ▶ ALGORITHM
- ▶ CONTACT

### Summary

***DACHIS***  
It is a novel method for recognition & classification of histones. It is based on amino acid composition & dipeptide composition. The method has been developed using SVM. The method is able to discriminate between histone & non-histones with more than 99% accuracy. The method is further able to classify histones to families with >98% accuracy.



Histones in Nucleosome

(A)

**DACHIS**  
Dipeptide Composition Based Method For Recognition & Classification of Histones

HOME | HELP | SUBMIT | CONTACT

- ▶ HOME
- ▶ SUBMIT
- ▶ HELP
- ▶ ALGORITHM
- ▶ CONTACT

### Prediction Results

**Date of Prediction**      **Wed Jan 12 12:24:32 2005**

**Prediction Approach**      **Amino acid composition**

Sequence	Prediction	
MADATAAPAVAPAKSPKPKAAAKPKKPSAHPKYSEMIGKAIAALKERGGSS RQAILKYIMANFNVVKDAKSVNAHLKALRAGVKNNSLKQSKGTGASGSF RIGEANVVKGGPAKAKKAAPKAAKPKKAKSTPKKGGPAAKKPAAGEKAA KPKAKKPAAKKAAKPKKPAKSPAKKKAAPKAKKTPKPKK	<b>Histone 1 (H1)</b>	Reliability Index= 3 Expected Accuracy=~ 92.8%
MADATAAPAVAPAKSPKPKAAAKPKKPSAHPKYSEMIGKAIAALKERGGSS RQAILKYIMANFNVVKDAKSVNAHLKALRAGVKNNSLKQSKGTGASGSF RIGEAQVVKGGPAKAKKAAPKAAKPKKAKSTPKKGGPAAKKPAAGEKAA KPKAKKPAAKKAAKPKKPAKSPAKKKAAPKAKKTPKPKK	<b>Histone 1 (H1)</b>	Reliability Index= 4 Expected Accuracy=~ 94.0%
ASSDDGGKKLLLLPQHEERERYYYYYYGHGGGGGHHHHHKKKKHHHH	<b>Non-Histone</b>	

(B)

**FIG. 2.** DACHIS web server. (A) The features of the web interface of the method. (B) The prediction results. The results indicate whether the sequence query belongs to the histone superfamily. If the sequence belongs to the histone superfamily, it will subsequently indicate the histone class. The results also display the reliability index and expected accuracy of the prediction results.



## APPENDIX: SUPPLEMENTAL FILE

*Caenorhabditis elegans*

>F36A2.10pep:knownchromosome:CEL130:I:8827628:8828212:-1gene:F36A2.1  
 nscript:F36A2.10:  
 MARKTTVAKVGSKPGSTKKTIIQAVRPTVRTQGAVTRSQAALRGHMGITDSSTSTSSSRI  
 PKEKLVKSRSRSRKSTRSRSRKSTRSRSRSRSRSRSTRGKKRAPKKAVTTKAARSISP  
 VKVKKTEAIKSRGSSRVSAAHK

---

SVM Prediction: Linker Histone  
 BLAST: Hypothetical protein F36A2.10 (GI:3876772)  
 PFAM: No Significant Hit

//

>Y73B6BL.9bpep:knownchromosome:CEL130:IV:6294110:6295206:1gene:Y73B6  
 transcript:Y73B6BL.9b  
 MPFFPSRLFSFGVEHETGYVSRQSGRGRVYKGGYSPFTITFQVSTHRSTMSDVTVAETP  
 AVKTPTKAKSKTTKEPKAKVAAAHPFFINMVTAAATKKPVAKKPVAKKAATGEKKAK  
 KTTVAKKTGDKVKKAKSPKPAKKVAKSPAKKAAPKKAPAKKAAAPKA

---

SVM Prediction: Linker Histone  
 BLAST: Histone h1 like protein (GI:24636128)  
 PFAM: No Significant Hit

//

>ZC477.1pep:knownchromosome:CEL130:IV:7111980:7112844:1gene:ZC477.1t  
 cript:ZC477.1:  
 MTAVGGAPRGASTMTAVGGAPVGGSSSTMTAVGGAPSGASTMTAIGGAPRGASTMTAV  
 GGAPMGGGSTMGAPSGASTMTAVGGAPSGASTMTAIGGAPRGASTMTAVGGAPMGGG  
 STMTAVGGAPIGGSSSTMTAVGGVSTMTAVGGAPGGASTMTAMGGGPSAFGGAPPPPSG  
 SAMGGGGGGGATSAYFVGSGAMGGGGAGAQSVGGGPVGGGGGAKSGGGGGGIPG  
 QSVYMGAGGGGGGGGATSAYFAPR

---

SVM Prediction: Nucleosome Histone (H2A)  
 BLAST: Hypothetical protein ZC477.1 (GI7510796)  
 PFAM: No Significant Hit

//

*Homo sapien*

>ENSP00000272019pep:novelchromosome:NCBI35:1:242463731:242466466:1ge  
 ne:ENSG00000143687transcript:ENST00000272019  
 MTILTPECILNLFHFSLLHLLCQPPKAIIVSGPGKAIIVSGPGKAIIVSGPGKAIIVSGPGKAIIVS  
 GPGKAIIVSGPGKAIIVSGPGKAIIVSGPGKAIIVSGPGKAIIVSGPGKAIIVSGPGKAI  
 VSGPGKAIIVSGPGKAIIVSGPGKAIIVSGPGKAIIVSGPGKAIIVSGPGKAIIVSGK

AIVSGPGKAIVSGLGKAIVSSLGKAIVSGPGKAIVSGPGKAIVSGPGKAIVSSLGKAIVCSG  
 KAIVSGPGKAIVSGLGKAIVSGPGKAIVSGPGKAIVSGPGKAIVSGPGKAIVSGPGKAIVSP  
 GKAIVSGPGKAIVSGPGKAIVSGLGKAIVSGPGKAIVSGLGKAIVSSLGKAIVSGPGKAIVS  
 GPGKAIVSGPGKAIVSSLGKAIVCSLGKAIVSGPGKAIVSGLGKAIVSGPGKAIVSGPGKAI  
 VSGPGKAIVSGPGKAIVSGPGKAIVSSLGKAIVCSLGKAIVSGPGKAIVSGPGKAIVSGPG  
 KAIVSGPGKAISSLGKAIVCSLGKAIVSGPGKAIVSGLGKAIVSSLGKAIVSSLGKAIVSSL  
 GKAIVSGPGKAIVSSPGKAIVSGPGKAIVSGPGKAIVSGPGKAIVSGPGKAIVSSPGKAIVS  
 GPGKAIVSSPGKAIVSSPGKAIVSSPGKAIVSSPGKAIVSSPGKAIVSSPGKAIVSSLGKAI  
 VSGPGKAIVSGPGKAIVSGPGKAIVSSPGKAIVSSLGKAIVSGPGKAIVSSPGKAIVSGPGKA  
 IVSSPGKAIVSSPGKAIVSGPGKAIVSSPGKAIVSSPGKAIVSSPGKAIVSSPGKAIVSSLGK  
 AIVSSLGKAIVSSLGKAIVSGPGKAIVSSLGKAIVSSLGKAIVSSLGKAIVSGPGKAIVSSPG  
 KAIVSSPGKAIVSGPGKAIVSSLGKAIVSSPGKAIVSSPGKAIVSSPGKAIVSSPGKAIVSSPG

SVM Prediction: Linker Histone

BLAST: PREDICTED: similar to ba74P14.2 (novel protein) (GI:51467286)

PFAM: Cornifin Cornifin (SPRR) family.

//

## ACKNOWLEDGMENTS

This manuscript was supported by NIH grant AI57330 and the Molecular Immunology Foundation.

## REFERENCES

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389–3402.
- Arents, G., and Moudrianakis, E.N. 1995. The histone fold: A ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proc. Natl. Acad. Sci. USA* 92, 11170–11174.
- Bateman, A., Coin, L., Durbin, R., *et al.* 2004. The Pfam protein families database. *Nucl. Acids Res.* 32, D138–141.
- Bhasin, M., and Raghava, G.P. 2004a. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucl. Acids Res.* 32, W414–419.
- Bhasin, M., and Raghava, G.P. 2004b. GPCRpred: An SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucl. Acids Res.* 32, W383–389.
- Ehrenhofer-Murray, A.E. 2004. Chromatin dynamics at DNA replication, transcription and repair. *Eur. J. Biochem.* 271, 2335–2349.
- Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906–914.
- Georgel, P.T., and Hansen, J.C. 2001. Linker histone function in chromatin: Dual mechanisms of action. *Biochem. Cell. Biol.* 79, 313–316.
- Hua, S., and Sun, Z. 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17, 721–728.
- Hubbard, T., Andrews, D., Caccamo, M., *et al.* 2005. Ensembl 2005. *Nucl. Acids Res.* 33, D447–453.
- Joachims, T. 1999. *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods*, MIT Press, Boston.
- Kasinsky, H.E., Lewis, J.D., Dacks, J.B., and Ausio, J. 2001. Origin of H1 linker histones. *FASEB J.* 15, 34–42.
- Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389, 251–260.
- Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L., and Sweet, R.M. 1993. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* 362, 219–223.
- Reinhardt, A., and Hubbard, T. 1998. Using neural networks for prediction of the subcellular location of proteins. *Nucl. Acids Res.* 26, 2230–2236.
- Sullivan, S., Sink, D.W., Trout, K.L., Makalowska, I., Taylor, P.M., Baxevanis, A.D., and Landsman, D. 2002. The Histone Database. *Nucl. Acids Res.* 30, 341–342.

van Holde, K.E. 1989. *Chromatin*, Springer-Verlag, New York.

Wistrand, M., and Sonnhammer, E.L. 2005. Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC Bioinformatics* 6, 99.

Address correspondence to:

*Pedro A. Reche*  
*Department of Medicine*  
*Harvard Medical School*  
*77 Avenue Louis Pasteur*  
*Boston, MA 02115*

*E-mail:* reche@research.dfci.harvard.edu