

Gene Classification Using Codon Usage and Support Vector Machines

Jianmin Ma, Minh N. Nguyen, and Jagath C. Rajapakse

Abstract—A novel approach for gene classification, adopting codon usage bias as feature inputs to support vector machines (SVMs) is proposed. The DNA sequence is first converted to a 59-dimensional feature vector, where each element corresponds to the relative synonymous usage (RSCU) frequency of a codon. Since the input to the classifier is independent of sequence length, the approach is especially useful when sequences to be classified are of differing lengths and homology-based methods tend to fail. The method is demonstrated with 1,841 Human Leukocyte Antigen (HLA) sequences, which are classified into two major classes, HLA-I and HLA-II. Each major class is further classified into subgroups. Using codon usage frequencies, binary SVM achieved an accuracy rate of 99.3 percent for HLA major class classification and multiclass SVM achieved accuracy rates of 99.73 percent and 98.38 percent for the subclass classification of HLA-I and HLA-II molecules, respectively. Comparisons with *K*-Means clustering and other classifiers and homology-based features are given. Results indicate that the classification based on codon usage bias is consistent with biological functions of HLA molecules.

Index Terms—Codon usage bias, gene classification, Human Leukocyte Antigen (HLA), Major Histocompatibility Complex (MHC), relative synonymous codon usage (RSCU), Support Vector Machines (SVMs).

1 INTRODUCTION

GENETIC information encoded in nucleic acids is transferred to proteins through triplet genetic codes, referred to as *codons*. Codons that code for the same amino acid are referred to as *synonymous codons*, which are used at different relative frequencies during translation [1], [2], a phenomenon referred to as *codon usage bias*. Codon usage bias has been found to be highly variable among different species [3] and is primarily related to gene function [4], [5]. However, codon usage may be different in different species even when the function of the gene is similar [5]. Furthermore, codon bias pattern is closely related to protein tertiary structure [6]. Some researchers have shown that there is a high correlation between codon usage bias and tRNA abundance [7], [8] and gene expression level [9], [10]. The codon usage pattern of chloroplast genes in plants shows a significant divergence compared to most of the other genes of modern land plants [8], [11], perhaps reflecting their eubacterial origin.

The majority of codon usage analysis has been conducted on “deep-branching species” such as viruses [12], bacteria [3], [7], [10], yeast [13], [14], *Caenorhabditis elegans* [15], and *Arabidopsis thaliana* [16], [17]. In contrast, few researches have focused on more recent vertebrate species such as mammals [18], [19], including rodents [20], despite the obvious practical significance of studies on codon usage patterns of these species, especially primate species [21], including human beings. The aim of the present study is to

use codon usage bias for the classification of DNA sequences into different functional groups. We represent the pattern of the codon usage of a given DNA sequence with a feature vector of 59 elements. That is, the input to the classifier provides a compact representation of the input sequence, which is independent of the length of the sequence.

Classification of a large set of genes into a few biologically meaningful groups facilitates the prediction of functions of genes. Early methods of gene classification include homology-based approaches through multiple sequence alignment of protein or nucleic acid sequences [22], [23]. Because of time and space complexities of multiple sequence alignment, homology-based approaches are impractical for classification of lengthy sequences. Furthermore, if the lengths or evolutionary conservation of the sequences differ significantly, correct alignment is difficult to achieve, resulting in a low accuracy of gene classification. More importantly, the information from synonymous mutations is neglected in homology-based approaches, despite the importance of synonymous mutations in gene and protein evolution. Structural features of proteins have been used to classify genes [24], which also neglects synonymous mutations. The use of codon usage bias for gene classification was rarely mentioned earlier except Kanaya who used the species-specific characteristics of codon usage to classify genes from 18 different species, mainly from prokaryotes and unicellular eukaryotes [25].

The use of experimental approaches using microarrays [26] and gene expression data [27] for gene classification are costly and cumbersome. Computational techniques such as artificial neural networks [28] and independent component analysis (ICA) [29], [30] have recently been used for gene classification directly from DNA sequences. Such methods however suffer from the curse

- The authors are with the Bioinformatics Research Center, Nanyang Technological University, Singapore, 637553.
E-mail: {jmma, nmnguyen, asjagath}@ntu.edu.sg.

Manuscript received 14 Mar. 2006; revised 19 Mar. 2007; accepted 5 June 2007; published online 2 Aug. 2007.

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0055-0306.
Digital Object Identifier no. 10.1109/TCBB.2007.70240.

of dimensionality and not using evolutionary information of the input sequences. By using codon usage bias, the present approach reduces the input dimension to 59, irrespective of the length of the sequence, providing an efficient approach for gene classification.

Support Vector Machines (SVMs), having strong foundations in statistical learning theory [31], have been successfully applied in numerous areas of computational biology [32], [33], [34], [35]. As shown by Vapnik [36], [37], SVM implements an optimal marginal classifier minimizing the structural risk and offers several associated computational advantages such as the lack of local minima in the optimization. Furthermore, scalability and the generalization capability of SVM [31] make it more suitable for the classification of genes. Lin et al. applied SVM to study the conserved codon composition of ribosomal protein coding genes in *E. coli*, *M. tuberculosis*, and *S. cerevisiae* [38]. Bhasin and Raghava used SVM in the prediction of HLA-DRB1*0401 binding protein and Cytotoxic T lymphocyte (CTL) epitopes [39], [40]. Donnes and Elofsson used SVM to predict MHC class I binding peptides [41]. Zhao et al. applied SVM in the prediction of T-cell epitopes [42]. In the present work, we show that SVM using codon usage bias as input features demonstrates the best classification performance on Human Leukocyte Antigen (HLA) molecules. The preliminary results of the present work were presented in a conference paper [43].

Recently, there has been a rapid increase of the number of nucleic acid and protein sequences in the international immunogenetics databases [44], [45], [46], [47], which have enabled computational biologists to study human and primate immune systems. We demonstrate our classification method on HLA genes obtained from HLA ImmunoGenetics (IMGT/HLA) database of European Bioinformatics Institute (EBI) (<http://www.ebi.ac.uk/>). The Major Histocompatibility Complex (MHC) is determined by a suite of genes located on a specific chromosome (e.g., HLA is located on chromosome 6, while mouse MHC is located on chromosome 11), which produce glycoprotein products to initiate the immune response of the body [48]. MHC molecules are a vital component of immune response and take part in the selection process of thymus cells, genetic control of immunological reaction, and interactions between immunocytes. The primary function of MHC molecules is to bind and present antigens on cell surfaces for recognition by antigen-specific T-cell receptors (TCR) of lymphocytes. Immune reactions involve interactions between MHC molecules and T lymphocytes [49]; T-cell response has subsequently been restricted not only by the antigen but also by the MHC molecule [50]. Furthermore, MHC molecules are involved in the production of antibodies, which process is also MHC restricted by gene products from the class II region [51], [52]. MHC gene products are involved in the pathogenesis of many diseases, including autoimmune disorders. The exact mechanisms behind MHC associated risk of autoimmune diseases remain to be fully understood.

In this paper, we demonstrate our approach on classification of HLA genes into major classes and their subgroups. HLA molecules are generally classified into three classes, HLA-I, HLA-II, and HLA-III, according to their specific

functions in the immune system [51], [52]. The major classes are further divided into subclasses: HLA-I molecules are classified into HLA-A, HLA-B, HLA-Cw, HLA-E, HLA-F, and HLA-G types, and HLA-II molecules are classified into HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB. The expression of HLA-I genes is constitutive and ubiquitous in most cell types. This is inconsistent with the protective function of cytotoxic T lymphocytes (Tc), which continuously survey cell surfaces and destroy cells harboring metabolically active microorganisms. HLA-II molecules are expressed only within cells that present antigens such as antigen-presenting macrophages, dendritic cells, and B cells. This is in accordance with the functions of helper T lymphocytes (Th) activated locally wherever they encounter antigen presenting cells that have internalized and processed antigens produced by pathogens.

With the present method, we report near perfect test and cross-validation accuracies in the classification of HLA genes into HLA-I and HLA-II classes and their corresponding subclasses. The results are compared with those obtained with unsupervised method of *K*-Means clustering [53], [54] and with classification methods: Linear Discriminant Analysis (LDA) [55] and *k*-nearest neighbors (*k*-NN) [56]. All three methods gave high accuracies for major HLA class classification, indicating the effectiveness of codon usage bias in gene classification. It should be noted here that even unsupervised *K*-Means is capable of accurately clustering HLA molecules into HLA-I and HLA-II groups using codon usage bias frequencies. SVM reported the highest accuracies in both major class classification as well as subclass classification. Further, the present method using codon usage bias outperformed methods using homology-based features.

2 MATERIALS AND METHODS

2.1 Materials

HLA genes were extracted from IMGT/HLA Sequence Database [44], [45], [46], [47] of EBI (Release 2.7, 10/08/2004, <http://www.ebi.ac.uk/imgt/hla/>), which is a part of the international ImMunoGeneTics project (IMGT), providing specialist databases of the sequences of HLA molecules, including official sequences for the Nomenclature Committee for Factors of HLA System of the World Health Organization. HLA gene sequences extracted were checked individually for errors such as incorrectly assigned translation initiation sites, inconsistencies with the reference sequences in EMBL or GenBank nucleotide databases, etc. Moreover, the errors were then manually curated.

Because there are 61 different codons coding for amino acids, only coding sequences consisting of more than 50 amino acids were included in the analysis. Thus, 1,841 HLA genes were finally selected for analysis; the information of the HLA genes in each class are given in Table 1. Note the high standard deviations of the lengths of sequences in most of the subclasses of HLA-I and HLA-II molecules.

TABLE 1
Information on HLA Sequences Used in This Analysis

Main class	Sub-class	Number Of sequences	Mean length (bp)	Standard Deviation (bp)	
HLA Class I	HLA-A	335	741.1	246.9	
	HLA-B	610	725.2	241.3	
	HLA-C	179	743.6	260.0	
	HLA-E	5	795.6	375.5	
	HLA-F	2	1062	38.2	
	HLA-G	15	795.8	155.9	
	Total		1146	734.6	245.7
HLA Class II	HLA-DMA	4	476.3	245.6	
	HLA-DMB	6	447.5	217.3	
	HLA-DOA	8	697.5	43.5	
	HLA-DOB	8	819.0	0.0	
	HLA-DPA1	20	395.7	221.1	
	HLA-DPB1	110	280.1	105.7	
	HLA-DQA1	27	618.0	206.8	
	HLA-DQB1	58	428.4	226.3	
	HLA-DRA	3	737.0	48.5	
	HLA-DRB	DRB1	381	320.4	158.5
		DRB3	41	351.1	196.5
		DRB4	11	511.6	239.1
		DRB5	18	360.8	198.8
		Total		695	356.5
Total		1841	591.9	291.8	

2.2 Methods of Gene Classification

Our approach for gene classification consists of two steps: 1) computation of relative synonymous codon usage (RSCU) frequency pattern as the feature vector representing each sequence and 2) classification using SVM (see Fig. 1). Binary SVM was used to classify a given HLA molecule into major classes, HLA-I and HLA-II, and multiclass SVM for the subclass classification of HLA-I and HLA-II molecules.

2.2.1 Computation of Relative Synonymous Codon Usage Frequency (RSCU)

To evaluate synonymous codon usage without confounding the influence of amino acid compositions of different sequence samples, the RSCU was adopted [13], [17], [57], [58]. For a given coding sequence, RSCU value r_k of synonymous codon k is calculated as

$$r_k = n_k \times \frac{obs_k}{tot_k}, \quad (1)$$

where obs_k and tot_k are the observed number of codon k and the total observed number of codons coding for the amino acid coded by codon k , respectively, and n_k denotes the number of synonymous codons of codon k . Though there are 64 codons in the genetic code, two codons, UGG and AUG, are unique codons for the amino acids Tryptophan (Trp, W) and Methionine (Met, M), respectively, and not considered as their RSCU values equal to unity; three stop codons (UGA, UAA, and UAG) are also not considered. Therefore, RSCU values of only 59 codons are used as input features for classification.

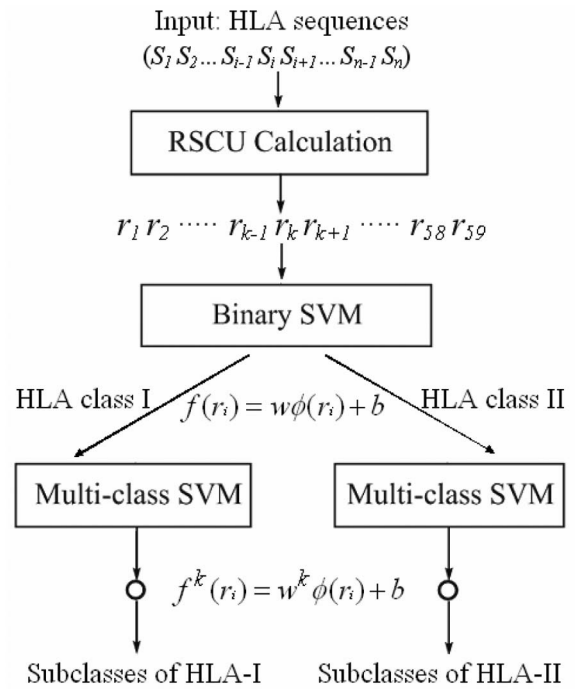


Fig. 1. Illustration of gene classification using codon usage as input features and SVM.

2.2.2 Binary SVM

A binary SVM is adopted to classify sequences into two major classes, HLA-I and HLA-II. Let $s = (s_1, s_2, \dots, s_n)$ denote a DNA sequence of length n , where $s_i \in \{A, U, C, G\}$ and $r = (r_1, r_2, \dots, r_{59})$ denote the input feature vector, where $r_k \in \mathbf{R}$ is the RSCU value of codon k . The classification of sequence s into HLA-I or HLA-II class finds an optimal mapping from \mathbf{R}^{59} space of RSCU values into $\{-1, +1\}$, where -1 corresponds to HLA-I and $+1$ to HLA-II classes, respectively.

Let $\{(\mathbf{r}_j, q_j) : j = 1, 2, \dots, N\}$ denote the set of training exemplars, where q_j denotes the desired class, HLA-I or HLA-II, for the input feature vector \mathbf{r}_j of RSCU values of sequence s_j ; N denotes the number of training sequences. SVM first transforms the input to a higher dimensional space with a kernel function \mathcal{K} and then linearly combines them with a weight vector \mathbf{w} to obtain the output. The binary SVM is trained to classify the input vectors of RSCU values to the correct major class of HLA.

For HLA major class classification, SVM constructs a discriminant function by solving the following optimization problem:

Minimize

$$\frac{1}{2} \mathbf{w}^T \mathbf{w} + \gamma \sum_{j=1}^N \xi_j,$$

subject to the constraints

$$q_j(\mathbf{w}^T \phi(\mathbf{r}_j) + b) \geq 1 - \xi_j \text{ and } \xi_j \geq 0, \quad (2)$$

where slack variables ξ_j represent the magnitude of error in the classification, ϕ represents the mapping function to a higher dimension, b is the bias used to classify samples,

and $\gamma (> 0)$ is the sensitivity parameter that decides the trade-off between the training error and the margin of separation [36], [37].

The minimization of the above optimization problem is equivalent to maximizing the following quadratic function:

$$\max_{\alpha} \sum_{j=1}^N \alpha_j - \frac{1}{2} \sum_{j=1}^N \sum_{i=1}^N \alpha_j \alpha_i q_j q_i \mathcal{K}(\mathbf{r}_j, \mathbf{r}_i), \quad (3)$$

subject to $0 \leq \alpha_j \leq \gamma$ and $\sum_{j=1}^N \alpha_j q_j = 0$.

Function $\mathcal{K}(\mathbf{r}_j, \mathbf{r}_i) = \phi(\mathbf{r}_j)^T \phi(\mathbf{r}_i)$ is the kernel function and the weight vector $\mathbf{w} = \sum_{j=1}^N q_j \alpha_j \phi(\mathbf{r}_j)$.

Once the parameters α_j are obtained from the optimization, the resulting discriminant function f is given by

$$f(\mathbf{r}_i) = \sum_{j=1}^N q_j \alpha_j \mathcal{K}(\mathbf{r}_j, \mathbf{r}_i) + b = \mathbf{w}^T \phi(\mathbf{r}_i) + b, \quad (4)$$

where bias b is chosen so that $q_j f(\mathbf{r}_j) = 1$ for all j with $0 < \alpha_j < \gamma$. The class corresponding to input pattern \mathbf{r}_i of RSCU values is HLA-I if $f(\mathbf{r}_i) < 0$ or HLA-II if $f(\mathbf{r}_i) \geq 0$.

2.2.3 Multiclass SVM

Multiclass SVM was adopted to classify DNA sequences to subclasses of HLA-I and HLA-II molecules. A scheme proposed by Crammer and Singer [59] for multiclass SVM was used, which has the capacity to solve the optimization problem in one step while minimizing the generalization error in the prediction [60].

For HLA-I classification, SVM constructs three discriminant functions all of which are obtained by solving one single optimization problem:

Minimize

$$\frac{1}{2} \sum_{c \in \Omega_1} (\mathbf{w}^c)^T \mathbf{w}^c + \gamma \sum_{j=1}^{N_1} \xi_j,$$

subject to the constraints

$$(\mathbf{w}^{t_j})^T \phi(\mathbf{r}_j) - (\mathbf{w}^c)^T \phi(\mathbf{r}_j) \geq d_j^c - \xi_j, \quad (5)$$

where $t_j \in \Omega_1 = \{\text{HLA-A, HLA-B, HLA-C}\}$ denotes the desired subclass for input r_j , N_1 denotes the number of training sequences of HLA-I molecules, slack variables ξ_j represent the magnitude of error in the classification, $c \in \Omega_1$ denotes the predicted subclasses of HLA-I sequence, and

$$d_j^c = \begin{cases} 0 & \text{if } t_j = c \\ 1 & \text{if } t_j \neq c. \end{cases}$$

The minimization of the above optimization problem in (5) is done by solving the following quadratic programming problem:

$$\max_{\alpha_j^c} - \frac{1}{2} \sum_{j=1}^{N_1} \sum_{i=1}^{N_1} \mathcal{K}(\mathbf{r}_j, \mathbf{r}_i) \sum_{c \in \Omega_1} \alpha_j^c \alpha_i^c - \sum_{j=1}^{N_1} \sum_{c \in \Omega_1} \alpha_j^c d_j^c$$

such that $\sum_{c \in \Omega_1} \alpha_j^c = 0$ and

$$\alpha_j^c \leq \begin{cases} 0 & \text{if } t_j \neq c, \\ \gamma & \text{if } t_j = c, \end{cases} \quad (6)$$

where $\mathcal{K}(\mathbf{r}_i, \mathbf{r}_j) = \phi(\mathbf{r}_i)^T \phi(\mathbf{r}_j)$ denotes the kernel function and the weight vector $\mathbf{w}^c = \sum_{j=1}^N \alpha_j^c \phi(\mathbf{r}_j)$.

Once the parameters α_j^c are obtained from the optimization, the resulting discriminant function f_c of a test input vector \mathbf{r}_i is given by

$$f_c(\mathbf{r}_i) = \sum_{j=1}^{N_1} \alpha_j^c \mathcal{K}(\mathbf{r}_i, \mathbf{r}_j) = (\mathbf{w}^c)^T \phi(\mathbf{r}_i). \quad (7)$$

The subclass of HLA-I, corresponding to the input feature vector \mathbf{r}_i is determined by $\arg \max_{c \in \Omega_1} f_c(\mathbf{r}_i)$.

For HLA-II classification, five discriminant functions f^c ,

$$c \in \Omega_2 = \{\text{HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1, HLA-DRB3}\},$$

were obtained by solving one single optimization problem, which is formulated, as in (5). The subclass of HLA-II, corresponding to the input pattern of RSCU values \mathbf{r}_i is determined by $\arg \max_{c \in \Omega_2} f^c(\mathbf{r}_i)$.

3 RESULTS

RSCU values were computed by using our own toolkit for Codon Analysis. Binary SVM was implemented using LIBSVM [61] known to have faster convergence properties than other tools available for solving the quadratic programming problem [62]. For the subclass classification of HLA-I and HLA-II molecules, multiclass SVM was implemented using BSVM libraries [61]. LDA and k -NN methods were implemented by using the LNKnet package [63]. K -Means clustering and Principal Component Analysis (PCA) were performed by using SPSS [64].

We use a 10-fold cross-validation procedure to evaluate the accuracy in the HLA major class classification and the HLA-I and HLA-II subclass classifications on data sets containing more than 25 sequences, respectively. In order to avoid the selection of extremely biased partitions in cross validation, the data set was randomly divided into 10 balanced partitions of same size. The cross-validation accuracies were evaluated with different types of kernels and different values of sensitivity parameter γ and kernel parameters. The kernel type and parameters were set based on the highest accuracy.

3.1 Classification of HLA Molecules Using SVM and Codon Bias

In this section, we report the cross-validation accuracies of our approach in the classification of HLA molecules into major classes and HLA-I/HLA-II molecules into their subclasses, illustrating the selection of kernel and the estimation of kernel and sensitivity parameters.

Ten-fold cross validation was performed on classifying 1,841 HLA sequences into HLA-I and HLA-II classes with a binary SVM. We first set the sensitivity parameter γ to a particular value from the set $\{0.5, 0.75, 1.0, \dots, 2.0\}$, and then, the cross-validation performances of classification with a binary SVM using a Gaussian kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\sigma \|\mathbf{x} - \mathbf{y}\|^2}$ with the parameter σ in the set $\{0.01, 0.02, 0.03, \dots, 0.1\}$, a linear kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$, and polynomial kernels $\mathcal{K}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^d$ with $d = 2, 3, 4$ were obtained. Table 2 shows the best accuracies obtained. As seen, with parameters $\sigma = 0.1$ and $\gamma = 1.5$, the Gaussian kernel achieved

TABLE 2

Best 10-Fold Cross-Validation Accuracies for the Classification of HLA Molecules into HLA-I and HLA-II Classes Using Binary SVM with Different Kernel Functions

Kernel Function and Parameters				
Gaussian	Linear	Polynomial	Polynomial	Polynomial
$\gamma = 1.5; \sigma = 0.1$	$\gamma = 1.0$	$\gamma = 1.5; d=2$	$\gamma = 1.5; d=3$	$\gamma = 1.5; d=4$
99.30	98.89	98.26	97.89	95.84

the highest accuracy of 99.30 percent for the classification of HLA molecules.

For HLA-I subclass classification, we first considered the subclasses of HLA-A, HLA-B, and HLA-C, as the numbers of sequences in other subclasses such as HLA-E, HLA-F, and HLA-G are less than 25, to be included in the analysis, so the total sequence number for the experiment is 1,124. For similar reason, we only consider the subclasses of HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1, and HLA-DRB3 for HLA-II subclass classification; so, the total sequence number for the experiment is 617.

As in a two-class classification, we empirically determined the kernel type and the best parameters in a manner similar to the major class classification. The performance of the subclass classification was high with sensitivity γ in the range [0.5, 1.5] and Gaussian kernel parameter σ in [0.01, 0.1]. Table 3 shows the best accuracies achieved with multiclass SVM using different kernels: Gaussian, linear, and polynomial. As seen, the Gaussian kernel achieved the best cross-validation accuracies. On the data set of 1,124 HLA-I sequences, the parameters $\gamma = 1.0$ and $\sigma = 0.1$ resulted in the best predictive accuracy of 99.73 percent, and for HLA-II subclass classification on the data set of 617 sequences, the parameters $\gamma = 1.5$ and $\sigma = 0.1$ gave the highest accuracy of 98.38 percent.

Table 4 gives the sensitivity and specificity values at the best accuracies of the SVM classification of HLA molecules. The classifications of HLA molecules into major classes and subclasses show near perfect performances with the present approach. The standard deviation of cross-validation accuracies of HLA major class classification, HLA-I subclass classification, and HLA-II subclass classification were 0.010, 0.028, and 0.024, respectively, indicating little effect by data partitioning.

We also performed multiclass SVM on the original data set for the subclass classification of HLA-I and HLA-II molecules without excluding subclasses having less than 25 samples. Two-fold cross validation was adopted as the

TABLE 4

Performance of the Classification of HLA Molecules by Using Codon Usage Bias and SVM

Classification	Accuracy	Sensitivity	Specificity
HLA major class classification	99.30	98.99	99.48
Sub-class classification of HLA-I	99.73	99.47	99.87
Sub-class classification of HLA-II	98.38	93.82	99.59

numbers of sequences in the subclasses of HLA-E, HLA-F, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, or HLA-DRA were small. The multiclass SVM achieved the accuracies of 99.39 percent for the classification of six subclasses of HLA-I and 98.71 percent for 13 subclasses of HLA-II, indicating a little effect on removing classes with a small number of classes.

In order to evaluate the testing accuracies of the present method, the data set was randomly divided into two balanced halves of major and subclasses of HLA molecules. One partition was selected for training, and the other was reserved for testing. The SVM was trained with the training data set. The kernels and parameters were selected empirically based on the best accuracies on the training data set. The testing accuracies were calculated on the testing data set with the parameters obtained during training. This procedure was repeated 25 times, and the mean and standard deviation of accuracy were reported in Table 5. The accuracies on training data were reported as well. As seen, the testing and training accuracies were close.

3.2 Comparison with Other Classifiers

We compared the present approach with the unsupervised method of K -Means clustering [53], [54], and two other classifiers, LDA [55] and k -nearest neighbors (k -NN) [56], using RSCU features as inputs. K -Means clustering was performed with different k values and starting seeds for HLA major and subclass classifications. The value of k was empirically determined by the best accuracies. For k -NN classifiers, the appropriate value of k was empirically found in the range [1, 10] for the best accuracy. The train, test, and cross-validation accuracies were computed as for SVM and given in Table 5 for all methods.

TABLE 3

Best 10-Fold Cross-Validation Accuracies for the Classification of HLA-I and HLA-II Molecules into Their Subclasses Using Multiclass SVM with Different Kernel Functions

Multi-class SVM	Kernel Function and Parameters				
	Gaussian	Linear	Polynomial	Polynomial	Polynomial
HLA-I sub-class classification	$\gamma = 1.0; \sigma = 0.1$	$\gamma = 1.0$	$\gamma = 1.0; d = 2$	$\gamma = 1.0; d = 3$	$\gamma = 1.0; d = 4$
	99.73	98.56	98.47	98.19	96.25
HLA-II sub-class classification	$\gamma = 1.5; \sigma = 0.1$	$\gamma = 1.0$	$\gamma = 1.5; d = 2$	$\gamma = 1.5; d = 3$	$\gamma = 1.5; d = 4$
	98.38	98.00	97.85	97.44	95.85

TABLE 5
Comparison of Accuracies of Different Classifiers Using Codon Usage Bias as Inputs

Classification	Classifier	Training Accuracy		Testing Accuracy		Cross-validation Accuracy	
		mean	SD	mean	SD	mean	SD
HLA major class classification	SVM	98.91	0.010	98.72	0.011	99.30	0.010
	<i>k</i> -means	94.83	0.035	93.12	0.039	94.02	0.037
	LDA	90.27	0.025	88.02	0.027	89.29	0.023
	<i>k</i> -NN	94.81	0.013	92.30	0.022	93.95	0.015
Sub-class classification of HLA-I	SVM	99.64	0.027	98.60	0.028	99.73	0.028
	<i>k</i> -means	53.32	0.028	52.98	0.024	53.05	0.028
	LDA	89.51	0.012	87.72	0.021	88.31	0.023
	<i>k</i> -NN	93.92	0.018	93.68	0.015	93.82	0.018
Sub-class classification of HLA-II	SVM	98.88	0.024	97.67	0.032	98.38	0.024
	<i>k</i> -means	54.32	0.029	52.23	0.025	54.53	0.029
	LDA	90.11	0.015	88.37	0.023	89.03	0.018
	<i>k</i> -NN	94.83	0.011	93.02	0.014	94.00	0.012

As seen, SVM outperformed all other methods in major class and subclass classifications of HLA molecules. *K*-Means clustering gives high accuracy on the major class classification, and both LDA and *k*-NN report higher accuracies for HLA major class, as well as HLA-I/HLA-II subclass classifications. This indicates that RSCU is an effective feature for classification of HLA molecules.

3.3 Significance Analysis of Codon Usage

We used *t*-test to compare the codon bias usage of each of 59 codons in HLA-I and HLA-II molecules. We used Bonferroni correction to account for multiple statistical comparisons of codons in testing the significance [65], [66]. The results of the *t*-test with adjusted *P* values (uncorrected values divided by 59) are shown in Table 6. As seen, the usage of most of the 59 codons are significantly different between classes at the level of $p < 0.01$ except for a few cases: GCU (coding for Alanine, $p = 2.36$), GCA (coding for Alanine, $p = 11.59$), GGA (coding for Glycine, $p = 1.26$), GUC (coding for Valine, $p = 0.46$), CUG (coding for Leucine, $p = 11.0970$), CCA (coding for proline, $p = 6.27$), UCU (coding for Serine, $p = 22.41$), GAU (coding for Glutamic acid, $p = 0.07$), GAC (coding for Glutamic acid, $p = 0.11$), AAA (coding for lysine, $p = 17.75$), and AAG (coding for lysine, $p = 8.27$). In another words, except for the usage of these codons, the codon usage bias of all other codons are significantly different in HLA-I and HLA-II molecules and effective in the classification.

3.4 HLA Classification Using Homology-Based Features

In order to compare discriminating power of codon usage bias, homology-based distance matrices were used for the classification of HLA sequences, HLA-I sequences, and HLA-II sequences. The multiple sequence alignment on sequences was performed by using ClustalX (<http://bips.u-strasbg.fr/fr/Documentation/ClustalX/>, version: 1.83) [67], and the distance matrix was constructed by pairwise similarities of aligned sequences. The distance matrix has been previously shown as an effective feature for clustering or classification of aligned sequences [23]. Using the

distance matrix as input features, SVM was used to classify the sequences; and 10-fold cross-validation accuracies are reported in Table 7. As seen, RSCU features showed improvement in classification accuracy. Moreover, the alignment of 1,841 HLA sequences with ClustalX took about 24 hours on a workstation with a 3.0-GHz CPU and 2 Gbytes memory, while it took only several minutes to prepare the codon usage values. This indicates that codon usage bias is not only a superior feature but also provides an efficient approach to gene classification.

3.5 Principle Component Analysis of HLA Molecules

Principle component analysis (PCA) [68] was performed to obtain a low-dimensional representation of the space of codon usage for the visualization of the distribution of HLA sequences. The analyses involved first measuring the codon usage, computing the eigenvectors, and then deriving the orthogonal projections of codon space by using the eigenvectors. Genes with similar codon usage bias are neighbors on high-variance principal component projections.

The first and second principal components yielded 24.3 percent and 20.1 percent of the total variance, respectively. As shown in Fig. 2, red squares represent HLA-I molecules and green squares HLA-II molecules. There is only one misclassification representing HLA-A*2445N sequence. The subclasses of 1,124 HLA-I sequences and 617 HLA-II sequences were also visualized using the first two principal components of codon usage bias. The two components accounted for 45.0 percent and 45.3 percent of variance, respectively, for HLA-I and HLA-II sequences. The distribution of codon usage bias in the subclasses of HLA-I and HLA-II molecules is illustrated using the first two principal components in Figs. 3 and 4, respectively.

4 DISCUSSION

We demonstrated codon usage bias as an effective feature for gene classification and that SVM gives the best classification accuracies. The method is independent of the lengths of sequences and useful when homology-based

TABLE 6
Significance of Codon Usage Bias of 59 Codons in Discriminating HLA-I and HLA-II Molecules

Codon	t value	p-value (Corrected)	Codon	t value	p-value (Corrected)
GCU_A	2.0565	2.3611	UCC_S	48.1206	0.0000
GCC_A	6.9453	0.0000	UCA_S	11.8163	0.0000
GCA_A	-1.2925	11.5964	UCG_S	16.4022	0.0000
GCG_A	-6.9692	0.0000	AGU_S	14.7692	0.0000
GGG_G	-26.4082	0.0000	AGC_S	-36.7864	0.0000
GGA_G	2.3033	1.2641	ACU_T	15.5710	0.0000
GGC_G	51.4533	0.0000	ACC_T	36.3207	0.0000
GGU_G	-14.7486	0.0000	ACG_T	-23.6012	0.0000
GUU_V	-27.5032	0.0000	ACA_T	-12.4351	0.0000
GUC_V	2.6597	0.4654	AAU_N	-30.7248	0.0000
GUA_V	-4.3235	0.0010	AAC_N	28.4501	0.0000
GUG_V	12.8698	0.0000	CAA_Q	-18.2763	0.0000
AUU_I	-5.9631	0.0000	CAG_Q	17.9226	0.0000
AUC_I	15.1793	0.0000	UAU_Y	6.1789	0.0000
AUA_I	36.6197	0.0000	UAC_Y	-6.1789	0.0000
UUA_L	-6.0518	0.0000	CAU_H	-6.6976	0.0000
UUG_L	-28.4911	0.0000	CAC_H	7.1577	0.0000
CUU_L	-11.5128	0.0000	GAU_D	-3.2741	0.0651
CUC_L	44.9089	0.0000	GAC_D	3.1081	0.1147
CUG_L	-1.3173	11.0970	GAA_E	-21.9313	0.0000
CUA_L	4.4146	0.0007	GAG_E	21.9313	0.0000
UUC_F	12.5897	0.0000	AAA_K	1.0351	17.7491
UUU_F	-12.5897	0.0000	AAG_K	-1.4761	8.2713
CCU_P	-53.6192	0.0000	CGU_R	-13.3971	0.0000
CCC_P	52.6717	0.0000	CGC_R	64.1302	0.0000
CCA_P	-1.6156	6.2869	CGA_R	-4.7490	0.0001
CCG_P	38.5147	0.0000	CGG_R	-37.2104	0.0000
UGU_C	-32.2672	0.0000	AGA_R	-14.7242	0.0000
UGC_C	31.1255	0.0000	AGG_R	17.8578	0.0000
UCU_S	-0.8788	22.4059			

methods tend to fail. The efficacy of the method was demonstrated on a set of HLA genes collected from the IMGT/HLA database. Once the HLA genes were classified according to major classes, codon usage bias was further explored for a finer classification of the molecules. In major class classification of HLA molecules and subclass classifications of HLA-I and HLA-II molecules, codon usage bias was more effective than the homology-based features; SVM outperformed *K*-Means clustering and LDA and *k*-NN classifiers.

TABLE 7
Accuracies of Homology- and Codon-Usage-Based Features for the Classification of HLA Genes with SVM

Feature	HLA major class	HLA-I sub-class	HLA-II sub-class
Codon Usage Bias	99.30	99.73	98.38
Homology	96.65	97.83	96.74

Interestingly, unsupervised clustering method such as *K*-Means gave a high accuracy in classifying HLA sequences into HLA-I and HLA-II classes; all HLA-I sequences were correctly clustered into class-I, and only a

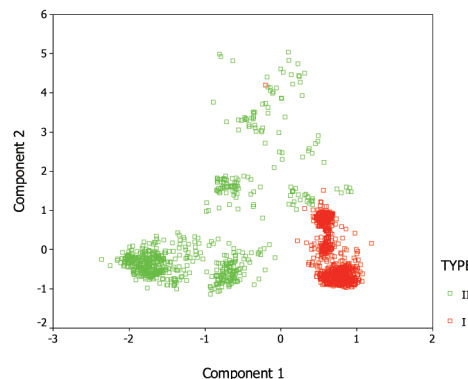


Fig. 2. Illustration the of distribution of HLA-I and HLA-II molecules in the set of 1,841 HLA molecules by using the two largest principal components of codon usage bias: red squares in the plot represent HLA-I sequences and green squares represent HLA-II sequences.

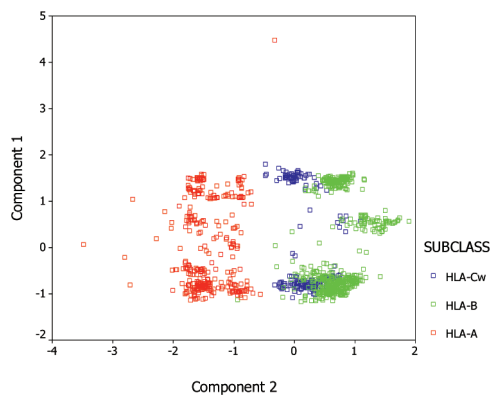


Fig. 3. Illustration of the distribution of 1,124 HLA-I sequences in the codon usage bias space using the two largest principal components.

small portion of HLA-II sequences were misclassified into class-I. This is consistent with the fact that the HLA-II molecules have more complicated structure and functions than the HLA-I molecules. However, *K*-Means was not successful in grouping the subclasses of HLA-I and HLA-II molecules, which maybe due to the small number of sequences in the subclasses.

Codon usage pattern-based approach for the classification of HLA sequences is relatively fast and efficient as the input sequences are transformed to a smaller dimensional RSCU space. Homology-based approaches needs high computing and storage requirements, especially when the sequences are of varying lengths and of large numbers. Codon usage is related to sequence homology and carries important information of evolution. For example, when a mutation occurs, homology analysis only shows the similar evolutionary rates of divergence between sequences, whereas codon usage analysis reveals changes in the relative frequency of encoded amino acids and the usage patterns of related codons, etc.

Since the classifications of HLA molecules into their subclasses were accurately achieved with codon usage bias, it can be concluded that the functions of HLA molecules are closely related to synonymous codon usage bias. Although our demonstration was limited to HLA molecules, the approach should be generalized and applicable for the classification of other groups of molecules as well. Our method could also help in the prediction of the function of a novel gene.

Codon usage analysis is a useful parameter in synonymous mutation studies in molecular evolution as when synonymous mutations occur, though the phenotype (the coded protein) does not change, codon usage patterns and features such as the gene expression level are affected. Therefore, the codon usage can be used as a good indicator in studies of gene expression and molecular evolution. The evolution of genes in the immune system is closely related to various types of mutations, of which synonymous mutation is of great importance. Furthermore, the codon usage of MHC genes is helpful in revealing the phylogeny and the regulation of the immune system and has the potential to assist research in mammalian immunogenetics.

Codon usage bias is a complicated phenomenon affected by many factors such as species, gene function, protein structure, gene expression level, tRNA abundance, etc. Building a correlation between codon bias pattern and

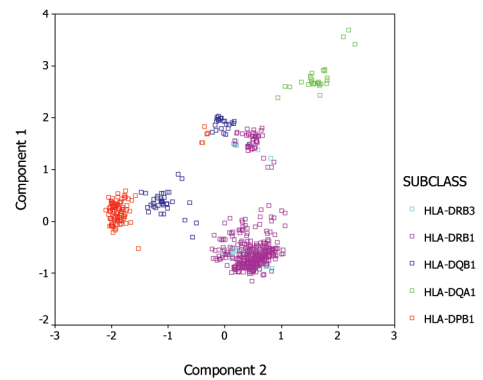


Fig. 4. Illustration of the distribution of 617 HLA-II sequences in the codon usage bias space using the two largest principal components.

biological phenotypes and finding the relationships and interactions can result in an unfolding valuable biological information from nucleic acid sequences. For novel genes, synonymous codon usage patterns could be used for their classification and helpful in inferring their function. Therefore, the analyses of codon usage patterns with computational techniques that capture inherent rules of translation could be useful for both basic and applied research in life sciences.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gary B. Fogel and Mr. Gavyn W.L. Pang for their valuable advice on this project. This work was partially supported by a grant to J.C. Rajapakse by the Biomedical Research Council (Grant 04/1/22/19/376) of the Agency of Science and Technology Research, administered through the National Grid Office, Singapore.

REFERENCES

- [1] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pave, "Codon Catalog Usage and the Genome Hypothesis," *Nucleic Acids Research*, vol. 8, pp. r49-r62, 1980.
- [2] T.C. Ghosh, S.K. Gupta, and S. Majumdar, "Studies on Codon Usage in *Entamoeba histolytica*," *Int'l J. Parasitology*, vol. 30, pp. 715-722, 2000.
- [3] P.M. Sharp, E. Cowe, and D.G. Higgins, "Codon Usage Patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, and *Homo sapiens*: A Review of the Considerable Within-Species Diversity," *Nucleic Acids Research*, vol. 16, pp. 8207-8211, 1988.
- [4] J.M. Ma, T. Zhou, W.J. Gu, X. Sun, and Z.H. Lu, "Cluster Analysis of the Codon Use Frequency of MHC Genes from Different Species," *Biosystems*, vol. 65, pp. 199-207, 2002.
- [5] J.M. Ma, N.M. Nguyen, G.B. Fogel, and J.C. Rajapakse, "Determination of the Relative Importance of Gene Function or Taxonomic Grouping to Codon Usage Bias Using Cluster Analysis and SVMs," *Proc. IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology*, Sept. 2006.
- [6] W.J. Gu, T. Zhou, J.M. Ma, X. Sun, and Z.H. Lu, "The Relationship between Synonymous Codon Usage and Protein Structure in *Escherichia coli* and *Homo sapiens*," *Biosystems*, vol. 73, pp. 89-97, 2004.
- [7] T. Ikemura, "Correlation between the Abundance of *Escherichia coli* Transfer RNAs and the Occurrence of the Respective Codons in Its Protein Genes: A Proposal for a Synonymous Codon Choice That Is Optimal for the *E. coli* Translational System," *J. Molecular Biology*, vol. 151, pp. 389-409, 1981.

- [8] B.R. Morton, "Chloroplast DNA Codon Use: Evidence for Selection at the PSB A Locus Based on tRNA Availability," *J. Molecular Evolution*, vol. 37, pp. 273-280, 1993.
- [9] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, "Codon Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity," *Nucleic Acids Research*, vol. 9, pp. r43-r74, 1981.
- [10] M. Gouy and C. Gautier, "Codon Usage in Bacteria: Correlation with Gene Expressivity," *Nucleic Acids Research*, vol. 10, pp. 7055-7074, 1982.
- [11] B.R. Morton, "Codon Use and the Rate of Divergence of Land Plant Chloroplast Genes," *Molecular Biology and Evolution*, vol. 11, pp. 231-238, 1994.
- [12] W.J. Gu, T. Zhou, J.M. Ma, X. Sun, and Z.H. Lu, "Analysis of Synonymous Codon Usage in SARS Corona Virus and other Viruses in the Nidovirales," *Virus Research*, vol. 101, pp. 155-161, 2004.
- [13] P.M. Sharp, T. Tuohy, and K. Mosurski, "Codon Usage in Yeast: Cluster Analysis Clearly Differentiates Highly and Lowly Expressed Genes," *Nucleic Acids Research*, vol. 14, pp. 5125-5143, 1986.
- [14] M.A. Freire-Picos, M.I. Gonzalez-Sisco, A.M. Rodriguez-Torres, E. Ramil, and M.E. Cerdan, "Codon Usage in *Kluyveromyces lactis* and in Yeast Cytochrome C-Encoding Genes," *Gene*, vol. 139, pp. 43-49, 1994.
- [15] M. Stenico, A.T. Lloyd, and P.M. Sharp, "Codon Usage in *Caenorhabditis elegans*: Delineation of Translational Selection and Mutational Biases," *Nucleic Acids Research*, vol. 22, pp. 2437-2446, 1994.
- [16] H. Chiapello, F. Lisacek, M. Caboche, and A. Henaut, "Codon Usage and Gene Function Are Related in Sequences of *Arabidopsis thaliana*," *Gene*, vol. 209, pp. GC1-GC38, 1998.
- [17] C. Mathe, A. Peresetsky, P. Dehais, M. Van Montagu, and P. Rouze, "Classification of *Arabidopsis thaliana* Gene Sequences: Clustering of Coding Sequences into Two Groups According to Codon Usage Improves Gene Prediction," *J. Molecular Biology*, vol. 285, pp. 1977-1991, 1999.
- [18] A.C. Eyre-Walker, "An Analysis of Codon Usage in Mammals: Selection or Mutation Bias," *J. Molecular Evolution*, vol. 33, pp. 442-449, 1991.
- [19] X. Pan and J. Fu, "Molecular Evolution of MHC DQA Genes. II. Phylogenetic Analysis Based on Nucleotide Substitution and SCU Bias," *Yi Chuan Xue Bao (Chinese)*, vol. 24, pp. 394-402, 1997.
- [20] N.G. Smith and L.D. Hurst, "The Causes of Synonymous Rate Variation in the Rodent Genome. Can Substitution Rates Be Used to Estimate the Sex Bias in Mutation Rate," *Genetics*, vol. 152, pp. 661-673, 1999.
- [21] S.K. McWeeney and A.M. Valdes, "Codon Usage Bias and Base Composition in MHC Genes in Humans and Common Chimpanzees," *Immunogenetics*, vol. 49, pp. 272-279, 1999.
- [22] I.M. Wallace, G. Blackshields, and D.G. Higgins, "Multiple Sequence Alignments," *Current Opinion in Structural Biology*, vol. 15, pp. 261-266, 2005.
- [23] V.N. Grishin and N.V. Grishin, "Euclidian Space and Grouping of Biological Objects," *Bioinformatics*, vol. 18, pp. 1523-1534, 2002.
- [24] M. Shatsky, R. Nussinov, and H.J. Wolfson, "Optimization of Multiple-Sequence Alignment Based on Multiple-Structure Alignment," *Proteins*, vol. 62, pp. 209-217, 2006.
- [25] S. Kanaya, Y. Yamada, Y. Kudo, and T. Ikemura, "Studies of Codon Usage and tRNA Genes of 18 Unicellular Organisms and Quantification of *Bacillus subtilis* tRNAs: Gene Expression Level and Species-Specific Diversity of Codon Usage Based on Multivariate Analysis," *Gene*, vol. 238, pp. 143-155, 1999.
- [26] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Sciences*, vol. 95, pp. 14863-14868, 1998.
- [27] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proc. Nat'l Academy of Sciences*, vol. 96, pp. 2907-2912, 1999.
- [28] L. Lancashire, O. Schmid, H. Shah, and G. Ball, "Classification of Bacterial Species from Proteomic Data Using Combinatorial Approaches Incorporating Artificial Neural Networks, Cluster Analysis and Principal Components Analysis," *Bioinformatics*, vol. 21, pp. 2191-2199, 2005.
- [29] E. Oja and A. Hyvaerinen, "A Fast Fixed-Point Algorithm for Independent Component Analysis," *Neural Computation*, vol. 9, pp. 1483-1492, 1997.
- [30] X.W. Zhang, Y.L. Yap, D. Wei, F. Chen, and A. Danchin, "Molecular Diagnosis of Human Cancer Type by Gene Expression Profiles and Independent Component Analysis," *European J. Human Genetics*, vol. 13, pp. 1303-1311, 2005.
- [31] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge Univ. Press, 2000.
- [32] M.N. Nguyen and J.C. Rajapakse, "Prediction of Protein Relative Solvent Accessibility with a Two-Stage SVM Approach," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, pp. 30-37, 2005.
- [33] M.N. Nguyen and J.C. Rajapakse, "Two-Stage Support Vector Regression Approach for Predicting Accessible Surface Areas of Amino Acids," *Proteins: Structure, Function, and Bioinformatics*, vol. 63, pp. 542-550, 2006.
- [34] K.B. Duan, J.C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data," *IEEE Trans. Nanobiotechnology*, vol. 4, pp. 228-234, 2005.
- [35] J.C. Rajapakse, K.B. Duan, and W.K. Yeo, "Proteomic Cancer Classification with Mass Spectra Data," *Am. J. Pharmacology*, vol. 5, pp. 281-292, 2005.
- [36] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.
- [37] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.
- [38] K. Lin, Y. Kuang, J.S. Joseph, and P.R. Kolatkar, "Conserved Codon Composition of Ribosomal Protein Coding Genes in *Escherichia coli*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*: Lessons from Supervised Machine Learning in Functional Genomics," *Nucleic Acids Research*, vol. 30, pp. 2599-2607, 2002.
- [39] M. Bhasin and G.P. Raghava, "SVM Based Method for Predicting HLA-DRB1*0401 Binding Peptides in an Antigen Sequence," *Bioinformatics*, vol. 20, pp. 421-423, 2004.
- [40] M. Bhasin and G.P. Raghava, "Prediction of CTL Epitopes Using QM, SVM and ANN Techniques," *Vaccine*, vol. 22, pp. 3195-3204, 2004.
- [41] P. Donnes and A. Elofsson, "Prediction of MHC Class I Binding Peptides, Using SVMHC," *BMC Bioinformatics*, vol. 3, pp. 25-32, 2002.
- [42] Y. Zhao, C. Pinilla, D. Valmori, R. Martin, and R. Simon, "Application of Support Vector Machines for T-Cell Epitopes Prediction," *Bioinformatics*, vol. 19, pp. 1978-1984, 2003.
- [43] J.M. Ma, N.M. Nguyen, W.L. Pang, and J.C. Rajapakse, "Gene Classification Using Codon Usage and SVMs," *Proc. IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology*, pp. 435-442, 2005.
- [44] J. Robinson, A. Malik, P. Parham, J.G. Bodmer, and S.G.E. Marsh, "IMGT/HLA Sequence Database—A Sequence Database for the Human Major Histocompatibility Complex," *Tissue Antigens*, vol. 55, pp. 280-287, 2000.
- [45] J. Robinson, M.J. Waller, P. Parham, J.G. Bodmer, and S.G.E. Marsh, "IMGT/HLA Sequence Database—A Sequence Database for the Human Major Histocompatibility Complex," *Nucleic Acids Research*, vol. 29, pp. 210-213, 2001.
- [46] J. Robinson, M.J. Waller, P. Parham, N. de Groot, R. Bontrop, L.J. Kennedy, P. Stoehr, and S.G.E. Marsh, "IMGT/HLA and IMGT/MHC: Sequence Databases for the Study of the Major Histocompatibility Complex," *Nucleic Acids Research*, vol. 31, pp. 311-314, 2003.
- [47] M. Galperin, "The Molecular Biology Database Collection: 2004 Update," *Nucleic Acids Research*, vol. 32, pp. D3-D22, 2004.
- [48] J.G. Bodmer, S.G.E. Marsh, E.D. Albert, W.F. Bodmer, R.E. Bontrop, D. Charron, B. Dupont, H.A. Erlich, B. Mach, W.R. Mayr, P. Parham, T. Sasazuki, G.M.T. Schreuder, J.L. Strominger, A. Svejgaard, and P.I. Terasaki, "Nomenclature for Factors of the HLA System, 1995," *Tissue Antigens*, vol. 46, pp. 1-18, 1995.
- [49] A.S. Rosenthal and E. Shevach, "Function of Macrophages in Antigen Recognition by Guinea Pig T Lymphocytes: I. Requirement for Histocompatible Macrophages and Lymphocytes," *J. Experimental Medicine*, vol. 138, pp. 1194-1212, 1973.
- [50] R.M. Zinkernagel and P.C. Doherty, "Restriction of in Vitro T Cell-Mediated Cytotoxicity in Lymphocytic Choriomeningitis within a Syngeneic or Semiallogeneic System," *Nature*, vol. 248, pp. 701-702, 1974. B.Kindred, D.C. Shreffler, "H-2 Dependence of Cooperation between T and B Cells In Vivo," *J. Immunology*, vol. 109, pp. 940-943, 1972.

- [51] D.H. Katz, T. Hamaoka, and B. Benacerraf, "Cell interactions between Histocompatible T and B Lymphocytes. Failure of Physiologic Cooperation Interactions between T and B Lymphocytes from Allogeneic Donor Strains in Humoral Response to Hapten-Protein Conjugates," *J. Experimental Medicine*, vol. 137, pp. 1405-1418, 1973.
- [52] H.X. Han, F.H. Kong, and Y.Z. Xi, "Progress of Studies on the Function of MHC in Immuno-Recognition," *J. Immunology (Chinese)*, vol. 16, no. 4, pp. 15-17, 2000.
- [53] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis*. Academic Press, 1979.
- [54] J.W. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Academic Press, 2001.
- [55] P. Winkel and E. Juhl, "Assumptions in Linear Discriminant Analysis," *Lancet*, vol. 2, pp. 435-436, 1971.
- [56] D. Aha and D. Kibler, "Instance-Based Learning Algorithms," *Machine Learning*, vol. 6, pp. 37-66, 1991.
- [57] P.M. Sharp and W.H. Li, "The Codon Adaptation Index—A Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications," *Nucleic Acids Research*, vol. 15, pp. 1281-1295, 1987.
- [58] J.M. Comeron and M. Aguade, "An Evaluation of Measures of Synonymous Codon Usage Bias," *J. Molecular Evolution*, vol. 47, pp. 268-274, 1998.
- [59] K. Crammer and Y. Singer, "On the Learnability and Design of Output Codes for Multiclass Problems," *Machine Learning*, vol. 47, pp. 201-233, 2002.
- [60] M.N. Nguyen and J.C. Rajapakse, "Two-Stage Multi-Class SVMs for Protein Secondary Structure Prediction," *Proc. Pacific Symp. Biocomputing*, 2005.
- [61] C.W. Hsu and C.J. Lin, "A Comparison on Methods for Multi-Class Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 13, pp. 415-425, 2002.
- [62] J. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," *Advances in Kernel Methods—Support Vector Learning*, B. Scholkopf, C.J.C. Burges, and A.J. Smola, eds., pp. 185-208, MIT Press, 1999.
- [63] *LNKnet Software Package*, <http://www.ll.mit.edu/IST/lnknet/>, 2008.
- [64] J.M. Su, R.H. Fu, J.B. Zhou, and L.H. Zhang, *Practical Guide for the Statistical Software of SPSS for Windows*, pp. 465-477. Publishing House of Electronics Industry, 2000.
- [65] C.E. Bonferroni, "Il Calcolo Delle Assicurazioni su Gruppi di Teste," *Studi in Onore del Professore Salvatore Ortu Carboni*, pp. 13-60, 1935.
- [66] W.R. Rice, "Analyzing Tables of Statistical Tests," *Evolution*, vol. 43, pp. 223-225, 1989.
- [67] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins, "The ClustalX Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools," *Nucleic Acids Research*, vol. 24, pp. 4876-4882, 1997.
- [68] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, vol. 2, pp. 559-572, 1901.



Jianmin Ma received the bachelor's degree in medicine from Binzhou Medical College, the master's degree in biochemistry from Suzhou Medical College, and the PhD degree in biomedical engineering from Southeast University, China. His current research interests include the analysis of codon usage bias characteristics by using multivariate statistical methods and computational intelligent methods and their corresponding applications in gene classification and phylogenetics. He is a research fellow in the Bioinformatics Research Centre (BIRC), School of Computer Engineering (SCE), Nanyang Technological University (NTU), Singapore.



Minh N. Nguyen received the BSc degree (first class) in computer science from the Vietnam National University, Hanoi, and the PhD degree in computational biology from Nanyang Technological University (NTU), Singapore. His current research interests include the prediction of protein structure and protein-protein interactions, and gene classification. He is the author of more than 20 research publications in refereed journals, book chapters, and conference proceedings on computational biology and machine learning. He is a research fellow in the Bioinformatics Research Centre (BIRC), Nanyang Technological University (NTU), Singapore.



Jagath C. Rajapakse received the BSc (Eng) degree (with first class honors) in electronic and telecommunication engineering from the University of Moratuwa, Sri Lanka, and the MSc and PhD degrees in electrical and computer engineering from the University of Buffalo. He was a visiting professor at the Biological Engineering Division, Massachusetts Institute of Technology (MIT), a visiting scientist at the Max-Planck-Institute of Cognitive and Brain Sciences, Leipzig, Germany, and a visiting fellow at the National Institute of Mental Health, Bethesda, Maryland. He is a professor of computer engineering and the director of the Bioinformatics Research Centre (BIRC), Nanyang Technological University (NTU), Singapore. His current research interests include gene networks, protein interactions, neural systems, and pathways. He is the author of more than 200 research publications in refereed journals, books, and conference proceedings on brain imaging, computational biology, and machine learning. He serves an associate editor for the *IEEE Transactions on Computational Biology and Bioinformatics*. He is the chair of IAPR Technical Committee on Bioinformatics. He is a senior member of the IEEE.

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.