Di-codon Usage for Gene Classification

Minh N. Nguyen[1], Jianmin Ma[1], Gary B. Fogel[2], and Jagath C. Rajapakse[3,4,5]

[1] BioInfomatics Institute, Singapore
[2] Natural Selection Inc. San Diego, USA
[3] BioInformatics Research Centre, Nanyang Technological University, Singapore
[4] Singapore-MIT Alliance, Singapore
[5] Department of Biological Engineering,
Massachusettes Institutes of Technology, USA

Abstract. Classification of genes into biologically related groups facilitates inference of their functions. Codon usage bias has been described previously as a potential feature for gene classification. In this paper, we demonstrate that di-codon usage can further improve classification of genes. By using both codon and di-codon features, we achieve near perfect accuracies for the classification of HLA molecules into major classes and sub-classes. The method is illustrated on 1,841 HLA sequences which are classified into two major classes, HLA-I and HLA-II. Major classes are further classified into sub-groups. A binary SVM using di-codon usage patterns achieved 99.95% accuracy in the classification of HLA genes into major HLA classes; and multi-class SVM achieved accuracy rates of 99.82% and 99.03% for sub-class classification of HLA-I and HLA-II genes, respectively. Furthermore, by combining codon and di-codon usages, the prediction accuracies reached 100%, 99.82%, and 99.84% for HLA major class classification, and for sub-class classification of HLA-I and HLA-II genes, respectively.

1 Introduction

Genetic information encoded in nucleic acids is transferred to proteins via codons. The study of codon usage is important as it is an integral component of translation of nucleic acids to their functional forms or proteins and its relevance to mutation studies. When a synonymous mutation occurs, the codon usage varies, but the resulting protein product remains unchanged. Therefore, codon usage is a good indicator for studies of mutation and molecular evolution. The pattern of codon usage has been found to be highly variable [1] and is implicated in the function of genes in different species. The use of codon usage bias for gene classification was rarely explored in the past except Kanaya et al. [2] who used the species-specific characteristics of codon usage to classify genes from 18 different species, mainly prokaryotes and unicellular eukaryotes. We recently showed that codon usage is a potential feature for gene classification [3]. Furthermore, using human leukocyte antigen (HLA) molecules, classification based on codon usage bias was shown to be inconsistent with molecular structure and biological function of the genes.

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 211–221, 2009.

Experimental approaches for gene classification often use microarray data, yet such methods are costly and tedious. Researchers have begun to use computational approaches such as machine learning techniques to extract features and thereby classify gene expressions from microarray experiments to identify genes belonging to biologically meaningful groups [4]. Because of the large dimension and the limited sample sizes, these methods have limited utility on larger datasets. Sequence-based gene classification provides an alternate to expression-based methods of gene classification. Other sequence-based methods of gene classification includes homology-based approaches through multiple sequence alignment [5]. Because of time and space complexities in multiple sequence alignment, such approaches are relatively difficult to use on a large number of sequences. Moreover, if the lengths or evolutionary distances of sequences differ, correct alignments are difficult to achieve, resulting in lower gene classification accuracy. More importantly, the information from synonymous mutations is often neglected in homology-based approaches despite their importance in evolution. The classification of genes based on structural features also neglects synonymous mutations [6].

In this paper, we demonstrate the use of di-codon usage as a promising feature for gene classification. Di-codon usage patterns contain additional information for gene classification to those given by codon usage as di-codon usage patterns encapsulate more global (di-codon frequency) information of a DNA sequence. Given that ribosomes actually reside over two codon positions when they slide along mRNA, di-codon usage has a biological rationale to translation of genes. Noguchi et al. developed a prokaryotic gene-finding program, MetaGene, which utilizes di-codon frequencies estimated by the GC content of a given sequence with other various measures [7]. By using di-codon frequencies, their method achieved a higher prediction accuracy than by using codon frequencies alone [7]. A hidden Markov model with self-identification learning for finding protein coding regions from un-annotated genome sequences has been studied and shown that the di-codon model outperforms other competitive features such as amino-acid pairs, codon usage, and G+C content in terms of sensitivity as well as specificity [8]. The gene finding program, DicodonUse, is based on frequencies of di-codons and used for identification of open reading frames that have a high probability of being genes [9]. Uno et al. demonstrated that the main reading frame of Chi sequences (5'-GCTGGTGG-3') increased as a result of the di-codon CTG-GTG increasing under a genomewide pressure for adapting to the codon usage and base composition of the *E. coli* K-12 strain [10].

In this paper, we use binary and multi-class support vector machines (SVM) for the classification of genes based on codon and di-codon usage features. Their good generalization capabilities in classification [11,12,13] make them ideal for gene classification. We have used SVMs successfully for classifying protein features [14,15,16], gene expressions [17], mass spectra [18], and genes based on codon usage [3]. Others have also demonstrated their use in other bioinformatics problems: Lin et al. [19] to study conserved codon composition of ribosomal protein coding genes in *E. coli*, *M. tuberculosis*, and *S. cerevisiae*; Bhasin and

Raghava [20,21] for the prediction of HLA-DRB1*0401 binding protein and Cytotoxic T lymphocyte (CTL) epitopes; Donnes and Elofsson for the prediction of MHC class I binding peptides [22]; and Zhao et al. for the prediction of T-cell epitopes [23].

By using di-codon usage pattern as input feature for SVM, we demonstrate our method for gene classification on a dataset of 1,841 HLA gene sequences collected from the IMGT/HLA Sequence Database. The proposed approach achieved substantial improvement in classification accuracies of HLA molecules into HLA-I and HLA-II classes, and their subclasses. We compare our results when using codon usage alone as input feature, and with homology-based methods.

2 Materials and Methods

2.1 Data

Recently, there has been an increase of the number of nucleic acid and protein sequences in the international immunogenetics databases [24,25,26], which has enabled computational biologists to study human and primate immune systems. In order to demonstrate our method, we use a set of 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) and produces glycoprotein products to initiate the immune response of the body [27]. HLA or human 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 HLA 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 HLA molecules and T lymphocytes [28]; T-cell response has subsequently been restricted not only by the antigen but also by HLA molecule [29]. Furthermore, HLA molecules are involved in the production of antibodies, which process is also HLA restricted by gene products from the class II molecules [30,31]. HLA gene products are involved in the pathogenesis of many diseases including autoimmune disorders. The exact mechanisms behind HLA associated risk of autoimmune diseases remain to be fully understood.

We first demonstrate our approach through the classification of HLA genes into major classes HLA-I and HLA-II. The major classes are then divided into sub-classes: HLA-I molecules are classified into HLA-A, HLA-B, HLA-C, 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, HLA-DRB1, HLA-DRB3, HLA-DRB4, and HLA-DRB5. Expression of HLA-I genes is constitutive and ubiquitous in most cell types. This is consistent 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.

HLA genes were extracted from the IMGT/HLA Sequence Database [24,25,26] of EBI (Release 2.7, 10/08/2004, http://www.ebi.ac.uk/imgt/hla/) which is part of the international ImMunoGeneTics project (IMGT) providing specialist databases of the sequences of HLA molecules, including official sequences for Nomenclature Committee for Factors of HLA System of the World Health Organization. Extracted HLA gene sequences were checked individually for errors such as incorrect assignment of translation initiation sites, inconsistencies with the reference sequences in EMBL or GenBank nucleotide databases, etc. and the errors were then curated manually.

Because there are 61 different codons coding for amino acids, in order to have a sufficient sampling of codons for computation, coding sequences of less than 50 amino acids were excluded from this analysis [3], resulting in 1,841 HLA genes. The details of this dataset are available in [3]. Di-codon usage patterns were calculated for each sequence and used as input features for SVM in classifying input HLA sequences into main- and sub-classes. The input to SVM was a 4096-dimensional vector derived from di-codon usage values. Binary SVM was adopted for classification of main classes and multi-class SVM was adopted for sub-class identification of HLA-I and HLA-II molecules.

2.2 Di-codon Usage

Let the coding sequence of the gene in terms of codons be denoted by $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ where $s_i \in \Omega$, n is the length of the sequence in codons and $\Omega = \{c_1, c_2, \ldots c_{64}\}$ is the alphabet of codons. The di-codon usage pattern is given by the fractions of di-codon types within the coding sequence and captures the global information about the gene sequence. The di-codon usage $r_{c_j c_k}$ is measured by the fraction of di-codons $(c_j, c_k) \in \Omega^2$ of the sequence $\mathbf{s}$:

$$r_{c_j c_k} = \frac{1}{n-1} \sum_{i=1}^{n-1} \delta(s_i = c_j)\delta(s_{i+1} = c_k) \tag{1}$$

where $\delta(\cdot) = 1$ if the argument inside is satisfied, otherwise is 0. Di-codon patterns have a fixed length of 4096 (64×64) irrespective of the length of the sequence. Let $\mathbf{r} = (r_1, r_2, \ldots, r_k, \ldots, r_{4096})$, where $r_k \in [0, 1]$, denote the feature vector consisting of di-codon usages derived from the input sequence $\mathbf{s}$.

2.3 Binary SVM

A binary SVM classifier was adopted to classify HLA gene sequences into two main classes: HLA-I and HLA-II. The problem of classifying HLA sequence, $\mathbf{s}$,

into major classes is seen as to find the optimal mapping from the space of di-codon usage patterns to HLA-I and HLA-II classes, respectively.

Let $\{(\mathbf{r}_j, q_j) : j = 1, 2, \ldots, N\}$ denote the set of all training exemplars where q_j denotes the desired classification, HLA-I or HLA-II, for the input di-codon usage pattern, $\mathbf{r}_j$, so that the output q_j is -1 if the correct class is HLA-I or $+1$ if the class is HLA-II; N denotes the number of training sequences. SVM implicitly projects the input to a higher dimensional space with a kernel function K and then linearly combines them with a weight vector $\mathbf{w}$ to obtain the output. The binary SVM was trained to classify input vectors of di-codon usage patterns to correct major class of HLA 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:

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

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 which decides the trade-off between the training error and the margin of separation [11,12]. The minimization of the above optimization problem was done by solving a quadratic programming problem. And the class corresponding to the input pattern of di-codon usage values is determined by the resulting discriminant function obtained from the optimization [3].

2.4 Multi-class SVM

Multi-class SVM was adopted to classify HLA sequences to sub-classes of HLA-I and HLA-II molecules. A scheme proposed by Crammer and Singer [32] for multi-class SVM was used, which has the capacity to solve the optimization problem in one step while minimizing the generalization error in the prediction [16].

For HLA-I classification, SVM was used to construct 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

$$(\mathbf{w}^{t_j})^T\phi(\mathbf{r}_j) - (\mathbf{w}^c)^T\phi(\mathbf{r}_j) \geq d_j^c - \xi_j \tag{3}$$

where $t_j \in \Omega_1 = \{\text{HLA-A, HLA-B, HLA-C}\}$ denotes the desired subclass for input $\mathbf{r}_j$, N_1 denotes the number of training sequences of HLA-I molecules, slack variables ξ_j represent the magnitude of error in 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 Eq. (3) was done by solving the quadratic programming problem. Based on the resulting discriminant function, the subclass of HLA-I corresponding to the input pattern of di-codon usage values is determined [3]. For HLA-II, five discriminant functions $f^c,\ c \in \Omega_2$, and $\Omega_2 = \{$HLA-DPB1,HLA-DQA1,HLA-DQB1,HLA-DRB1,HLA-DRB3$\}$ are constructed, each obtained by solving one single optimization problem as formulated in Eq. (3). The subclass of HLA-II, corresponding to the input pattern of di-codon usage was determined by the resulting discriminant function obtained from the optimization [3].

3 Results

Binary SVM was implemented using LIBSVM [33] known to have faster convergence properties than other tools available for solving the quadratic programming problem [34]. For sub-class classification of HLA-I and HLA-II molecules, multi-class SVM was implemented using BSVM libraries [34]. Ten-fold cross-validation was used to evaluate the accuracy in HLA major class classification as well as HLA-I and HLA-II subclass classifications. In order to avoid selection of extremely biased partitions in cross-validation, the dataset was divided randomly into ten balanced partitions of equal size. In addition, we also used specificity and sensitivity to assess the performance of the prediction scheme [3].

For binary and multi-class SVM, the Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\sigma\|\mathbf{x}-\mathbf{y}\|^2}$ gave superior performance over linear and polynomial kernels for classification of HLA molecules. This was also observed in the case of gene classification using codon bias as features [3]. The sensitivity parameter γ and the Gaussian kernel parameter σ were determined by using the grid-search method [34]. Grid-search provides useful parameter estimates for multi-class SVM in a relatively short time.

The classification accuracy of binning 1,841 HLA sequences into either HLA-I or HLA-II classes using binary SVMs was evaluated using ten-fold cross-validation. The optimal estimates of sensitivity parameter $\gamma = 2$ and kernel parameter $\sigma = 0.125$ of the Gaussian kernel achieved an accuracy of 99.95% for 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 sub-classes such as HLA-E, HLA-F, and HLA-G were too small (less than 25 sequences) to be included in the analysis, so the total number of sequences for the experiment was 1,124. For a similar reason, we only considered subclasses of HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1, and HLA-DRB3 for HLA-II subclass classification, so the total number of sequences included in the experiment was 617. For HLA-I sub-class classification of the dataset of 1124 sequences, the parameters $\gamma = 1$ and $\sigma = 0.25$ resulted in the best predictive accuracy of 99.82%, and for HLA-II sub-class classification on the dataset of 617 sequences, the parameters $\gamma = 1$ and $\sigma = 0.25$ gave an accuracy of 99.03%.

The performance of binary SVM for major class classification and multi-class SVM for sub-class classification of HLA-I and HLA-II molecules are presented in Table 1. The standard deviation of cross-validation accuracies of HLA

Table 1. Accuracy (Acc), sensitivity (Sn), and specificity (Sp) of the classification of HLA molecules by using codon and di-codon usage as features for SVM classifier

HLA Classification	**Features***:usage*								
	codon			*di-codon*			*codon + di-codon*		
	Acc	Sn	Sp	Acc	Sn	Sp	Acc	Sn	Sp
Major Class	99.30	98.99	99.48	99.95	99.86	100.0	100.0	100.0	100.0
HLA-I Sub-class	99.73	99.47	99.87	99.82	99.75	99.90	99.82	99.75	99.90
HLA-II Sub-class	98.38	93.82	99.59	99.03	96.35	100.0	99.84	99.40	100

Table 2. Comparison of performances of the present approach using codon and di-codon usage on the dataset of 1841 HLA genes

HLA Classification		Features/Method	Testing Accuracy		Cross-validation Accuracy	
			mean	SD	mean	SD
Major class		Codon	98.72	0.01	99.30	0.01
		Di-codon	99.13	0.01	99.95	0.01
		Codon + Di-codon	**99.78**	**0.01**	**100**	**0.00**
		Homology based method	96.14	0.04	96.65	0.04
Sub-class classification	HLA-I	Codon	98.60	0.03	99.73	0.03
		Di-codon	99.47	0.02	99.82	0.01
		Codon + Di-codon	**99.64**	**0.01**	**99.82**	**0.01**
		Homology based method	97.51	0.23	97.83	0.23
	HLA-II	Codon	97.67	0.03	98.38	0.02
		Di-codon	98.70	0.02	99.03	0.02
		Codon + Di-codon	**99.35**	**0.02**	**99.84**	**0.01**
		Homology based method	96.27	0.24	96.74	0.24

major class classification, HLA-I subclass classification, and HLA-II subclass classification were 0.01, 0.01, and 0.02, respectively, indicating a little effect of data partitioning (referred in Table 2).

We also investigated the combination of codon and di-codon features for the classification of HLA molecules into major classes and HLA-I/HLA-II molecules into their subclasses. A total of 4155 features including relative synonymous

codon usage of 59 codons [3] and 4096 di-codon usage values were used as input for the classification. Table 1 shows the ten-fold cross-validation accuracies, sensitivities, and specificities of binary SVM for major class classification and multi-class SVM for sub-class classification of HLA-I and HLA-II molecules, achieved through best parameter values. By combining codon and di-codon features for HLA sequence classification, the binary SVM achieved the highest accuracy of 100% with sensitivity parameter $\gamma = 2$ and kernel parameter $\sigma = 0.125$ of the Gaussian kernel; multi-class SVM achieved the accuracies of 99.82% and 99.84% for HLA-I and HLA-II sub-class classification, respectively, with parameters $\gamma = 1$ and $\sigma = 0.25$, interestingly, for both classes.

In order to evaluate testing accuracies of the present method, the dataset was randomly divided into two balanced halves of major- and sub-classes of HLA sequences. One partition was selected for training and the other was reserved for testing. SVM was trained with the training dataset and the kernels and parameters were selected based on the best accuracies on the training dataset. The test accuracies were calculated on the testing dataset with the parameters obtained during training. This procedure was repeated 25 times and the mean and standard deviation of accuracy were calculated and given in (Table 2). As seen, the testing and cross-validation accuracies are close, indicating good generalization ability of the method.

3.1 Comparison with Homology-Based Methods

In order to compare discriminating power of di-codon usage pattern, 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 [35] and the distance matrix was constructed by pairwise similarities of aligned sequences. The distance matrix has been shown previously as an effective feature for clustering or classification of aligned sequences [36]. Using the distance matrix as input features, SVM was used to classify the sequences; and ten-fold cross-validation accuracies are reported in Table 2. These results show that di-codon usage pattern gives improvement in classification accuracy and is an effective feature for classification of HLA genes.

4 Discussion and Conclusion

Codon and di-codon usage are useful features in synonymous mutation studies in molecular evolution because when a synonymous mutation occurs, though the phenotype (the coded protein) does not change, the codon usage pattern as well as features such as the gene expression level are affected. Di-codon usage patterns provide additional information on codon usage as ribosomes actually reside over two codon positions during translation. Therefore, di-codon usage is a good indicator in gene expression and molecular evolution studies and, as seen in our experiments, provides a good feature for gene classification.

The efficacy of our method was demonstrated on a set of HLA genes collected from IMGT/HLA database. Once HLA genes were classified according to major classes, di-codon usage were further explored for more precise classification of the molecules. In major class classification of HLA molecules and subclass classifications of HLA-I and HLA-II molecules, the present approach using di-codon usage patterns achieved better overall accuracies than obtained by the classifiers using codon usage bias. The results in classification of HLA genes, using codon and di-codon usage as features for SVM were near perfect. The method is independent of the lengths of sequences and useful when homology-based methods tend to fail on datasets having genes of varying length. Also, in case of SVM, testing and cross-validation accuracies were close, indicating that the parameter estimation and kernel selection procedures were not sensitive to data.

Since the classifications of HLA molecules into their subclasses were accurately achieved with di-codon usage patterns, the functions of HLA molecules should be closely related to di-codon usage. Although our demonstration was limited to HLA molecules, the approach could be generalized and applicable for the classification of other groups of molecules as well. As the method generalized well in the experiments, it could also help in the prediction of the function of novel genes. The authors are unaware of any public datasets for benchmarking gene classification algorithms such as the approach presented here.

Di-codon usage 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 di-codon usage patterns and biological phenotypes and finding the relationships and interactions can result in unfolding valuable biological information from nucleic acid sequences. For novel genes, di-codon usage patterns could be used for their classification and helpful in inferring their function. Therefore, analyses of di-codon usage patterns with computational techniques that capture inherent rules of translation could be useful for both basic and applied research in life sciences. Investigating usage patterns of which codons and di-codons most affect the classification of genes is worthy of further exploration. Recently, error-correcting output codes (ECOC) provide a general-purpose method for improving the performance of inductive learning programs on multi-class problems. Therefore, a comparison of the multi-class SVM with ECOC methods for multi-class gene classifications could be helpful and is reserved for future work.

References

1. Sharp, P.M., Cowe, E., Higgins, D.G.: Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapien*s: a review of the considerable within-species diversity. Nucleic Acids Res. 16, 8207–8211 (1988)
2. Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T.: 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 238, 143–155 (1999)

3. Ma, J.M., Nguyen, M.N., Rajapakse, J.C.: Gene Classification using codon usage and support vector machines. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6(1), 134–143 (2009)
4. Zhang, Y., Rajapakse, J.C. (eds.): Machine Learning in Bioinformatics. John Wiley and Sons Inc., Chichester (2009)
5. Wallace, I.M., Blackshields, G., Higgins, D.G.: Multiple sequence alignments. Curr. Opin. Struct. Biol. 15, 261–266 (2005)
6. Shatsky, M., Nussinov, R., Wolfson, H.J.: Optimization of multiple-sequence alignment based on multiple-structure alignment. Proteins: Structure, Function, and Bioinformatics 62, 209–217 (2006)
7. Noguchi, H., Park, J., Takagi, T.: MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Research 34(19), 5623–5630 (2006)
8. Kim, C., Konagaya, A., Asai, K.: A generic criterion for gene recognitions in genomic sequences. Genome Inform. Ser. Workshop Genome Inform. 10, 13–22 (1999)
9. Paces, J., Paces, V.: DicodonUse: the programme for dicodon bias visualization in prokaryotes. Folia Biol. (Praha) 48(6), 246–249 (2002)
10. Uno, R., Nakayama, Y., Tomita, M.: Over-representation of *Chi* sequences caused by di-codon increase in *Escherichia coli K-12*. Gene 380(1), 30–37 (2006)
11. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
12. Vapnik, V.: Statistical Learning Theory. Wiley and Sons, Inc., New York (1998)
13. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
14. Nguyen, M.N., Rajapakse, J.C.: Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins: Structure, Function, and Bioinformatics 59, 30–37 (2005)
15. Nguyen, M.N., Rajapakse, J.C.: Two-stage support vector regression approach for predicting accessible surface areas of amino acids. Proteins: Structure, Function, and Bioinformatics 63, 542–550 (2006)
16. Nguyen, M.N., Rajapakse, J.C.: Prediction of protein secondary structure with two-stage multi-class SVM approach. International Journal of Data Mining and Bioinformatics 1(3), 248–269 (2007)
17. Duan, K.B., Rajapakse, J.C.: Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans. Nanobioscience 4(3), 228–234 (2005)
18. Rajapakse, J.C., Duan, K.B., Yeo, W.K.: Proteomic cancer classification with mass spectrometry data. American Journal of Pharmacology 5(5), 281–292 (2005)
19. Lin, K., Kuang, Y., Joseph, J.S., Kolatkar, P.R.: 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 Res. 30, 2599–2607 (2002)
20. Bhasin, M., Raghava, G.P.: SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics 20, 421–423 (2004)
21. Bhasin, M., Raghava, G.P.: Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine 22, 3195–3204 (2004)
22. Donnes, P., Elofsson, A.: Prediction of MHC class I binding peptides, using SVMHC. BMC Bioinformatics 3(1), 25–32 (2002)
23. Zhao, Y., Pinilla, C., Valmori, D., Martin, R., Simon, R.: Application of support vector machines for T-cell epitopes prediction. Bioinformatics 19, 1978–1984 (2003)

24. Robinson, J., Waller, M.J., Parham, P., Bodmer, J.G., Marsh, S.G.E.: IMGT/HLA Sequence Database - a sequence database for the human major histocompatibility complex. Nucleic Acids Res. 29, 210–213 (2001)
25. Robinson, J., Waller, M.J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L.J., Stoehr, P., Marsh, S.G.E.: IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. 31, 311–314 (2003)
26. Galperin, M.: The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res. 32, D2–D22 (2004)
27. Bodmer, J.G., Marsh, S.G.E., Albert, E.D., Bodmer, W.F., Bontrop, R.E., Charron, D., Dupont, B., Erlish, H.A., Mach, B., Mayr, W.R., Parham, P., Sasazuki, T., Schreuder, G.M.T., Strom-inger, J.L., Svejgaard, A., Terasaki, P.I.: Nomenclature for factors of the HLA system, 1995. Tissue Antigens 46, 1–18 (1995)
28. Rosenthal, A.S., Shevach, E.: Function of macrophages in antigen recognition by guinea pig T lymphocytes. I. Requirement for histocompatibile macrophages and lymphocytes. J. Exp. Med. 138, 1194–1212 (1973)
29. Zinkernagel, R.M., Doherty, P.C.: Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. Nature 248, 701–702 (1974)
30. Katz, D.H., Hamoaka, T., Benacerraf, B.: 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. Exp. Med. 137, 1405–1418 (1973)
31. Han, H.X., Kong, F.H., Xi, Y.Z.: Progress of studies on the function of MHC in immuno-recognition. J. Immunol. (Chinese) 16(4), 15–17 (2000)
32. Crammer, K., Singer, Y.: On the Learnability and Design of Output Codes for Multiclass Problems. Machine Learning 47, 201–233 (2002)
33. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
34. Hsu, C.W., Lin, C.J.: A comparison on methods for multi-class support vector machines. IEEE Transactions on Neural Networks 13, 415–425 (2002)
35. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G.: The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24, 4876–4882 (1997)
36. Grishin, V.N., Grishin, N.V.: Euclidian space and grouping of biological objects. Bioinformatics 18, 1523–1534 (2002)