RESEARCH ARTICLE
doi: 10.2306/scienceasia1513-1874.2013.39.042

ScienceAsia 39 (2013): **42–49**

Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel

Watshara Shoombuatong[a,b], Panuwat Mekha[c], Kitsana Waiyamai[d], Supapon Cheevadhanarak[e], Jeerayut Chaijaruwanich[a,b,*]

[a] Department of Computer Science, Chiang Mai University, Thailand
[b] Bioinformatics Research Laboratory, Chiang Mai University, Thailand
[c] Department of Computer Science, Maejo University, Thailand
[d] Department of Computer Engineering, Kasetsart University, Thailand
[e] Division of Biotechnology, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Thailand

*Corresponding author, e-mail: jeerayut.c@cmu.ac.th

Received 1 Aug 2012
Accepted 13 Nov 2012

ABSTRACT: Human Leukocyte Antigen (HLA) plays an important role in the control of self-recognition including defence against microorganisms. The efficient performance of classifying HLA genes facilitates the understanding of the HLA and immune systems. Currently, the classification of HLA genes has been developed by using various computational methods based on codon and di-codon usages. Here, we directly classify the HLA genes by using the k-nearest neighbour (k-NN) classifier. To develop an efficient k-NN classifier, we propose the use of a spectrum kernel to investigate HLA genes. Our approach achieves an accuracy as high as 99.4% of the HLA major classes prediction measured by ten-fold cross-validation. Moreover, we give a maximum accuracy of 99.4% in the HLA-I subclasses. These results show that our proposed method is relatively simple and can give higher accuracies than other sophisticated and conventional methods.

KEYWORDS: gene classification, machine learning, computational method

INTRODUCTION

The human leukocyte antigen system or human lymphocyte antigen (HLA) is the molecular name of a group of molecules in the human major histocompatibility complex (MHC) region on human chromosome 6, which encode the cell-surface antigen-presenting proteins[1]. The HLA, a class of proteins found on the surface membranes of cells, serve the purpose of presenting possible antigens to T and B cells.

The MHC contains a group of molecules that play a crucial role in immune recognition and for the tolerance of tissue grafting. In mice and humans, the MHC molecules have also been found to influence body odours, body odour preferences, and mate choice[2,3]. These sequences are also some of the most polymorphic regions of the genome and are known to play a central role in controlling immunological self and non-self recognition[4]. There are different types of HLA, e.g., HLA-I, and HLA-II. These two gene types are important in the matching of tissues and organs for donation and organ transplantation under outdated immunesuppression protocols. In addition, the major HLA antigens are essential elements for immune function. The two different classes have different functions. The principle function of HLA-I, is to present virally induced peptides on the surface of the cell by linking to the T-Cell receptor of a cytotoxic (CD8) T Cell. This allows the identification of viruses. As HLA-IIs initiate a molecular immune response, they are only present on "immunologically active" cells (B lymphocytes, macrophages, etc.) and not on all tissues[5].

A fundamental problem in computational biology is biological sequence classification. In this paper, we focus on the problem of the classification of the HLA genes which can be further subdivided into two related sub-problems, i.e., the HLA genes are classified into their major classes and subclasses. In the major class, the HLA genes are generally subdivided into three classes, i.e., HLA-I, HLA-II, and HLA-III, according to their specific functions in the immune system[6,7]. The subclasses of the HLA-I genes are classified into HLA-A, HLA-B, HLA-Cw, HLA-E, HLA-F, and HLA-G. The sub-classes of HLA-II genes are classified into HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB.

A number of computational approaches have been developed for the sequence classification problem, including methods based on pairwise similarity of sequences (such as BLAST[8]), generative approaches (such as profile HMMs[9,10]) and discriminative approaches (such as kernel SVMs[11], the Fisher kernel[12], profile kernels[13], the mismatch kernel[14], and pairwise-SVMs[15,16]). Ma et al used the support vector machines (SVMs) based codon usage, where each element corresponds to the relative synonymous codon usage frequency of a codon[17]. This particular one is known as feature based classification method. Their SVM method has been compared with K-Means clustering, Linear Discriminant Analysis[18], and k-NN classifier[19]. In 2009, the di-codon usage l[20] was used to compare codon usage for the HLA genes classification. This work showed that using di-codon usage as feature inputs outperforms using codon usage alone.

In this work, we developed an approach to classify HLA genes using a combined method which integrates the k-nearest neighbour (k-NN) classifier with a spectrum kernel. We designed a series of combination parameters, which allowed us to determine relative contributions from a spectrum kernel and a k-NN method. Finally, our approach was then compared with other conventional methods. Our experimental results show that our simpler combining approach is comparable to other sophisticated and conventional methods for major class classification.

Dataset

Recently, there has been an increase in the number of nucleic acid and protein sequences in the international immunogenetics databases[21–23], which has enabled computational biologists to study human and primate immune systems in greater depth. The IMGT/HLA database was established to provide a locus-specific database (LSDB) for the allelic sequences of the genes in the HLA system, also known as the human major histocompatibility complex. This complex of over four megabases is located within the 6p21.3 region of the short arm of human chromosome 6 and contains an excess of 220 genes[24].

HLA genes were extracted from the IMGT/HLA Sequence Database of EBI (Release 2.28 15/01/2010, available at http://www.ebi.ac.uk/imgt/hla/). The name of these HLA genes and alleles, and their quality control is the responsibility of the WHO Nomenclature Committee for Factors of the HLA System[25]. The IMGT/HLA data-base contains entries for all HLA alleles, and alleles of some related genes, officially named by the Nomenclature Committee. These entries are derived from expertly annotated copies of the original EMBL-Bank/GenBank/DDBJ entries. More details about the data set can be found in Robinson et al[22,24]. Table 1 shows the numbers and percentages of HLA genes and their subclasses used in our experiment.

Table 1 Numbers of HLA genes and their subclasses.

Major Class	Subclass	Number of sequences	Percentage (%)
HLA Class I	HLA-A	965	30.1
	HLA-B	1540	48.0
	HLA-Cw	626	19.5
	HLA-E	9	0.3
	HLA-F	21	0.7
	HLA-G	45	1.4
	Total	3206	100.0
HLA Class II	HLA-DMA	4	0.3
	HLA-DMB	7	0.6
	HLA-DOA	12	1.0
	HLA-DOB	9	0.8
	HLA-DPA1	27	2.3
	HLA-DPB1	138	11.5
	HLA-DQA1	35	2.9
	HLA-DQB1	107	8.9
	HLA-DRA	3	0.3
	HLA-DRB	855	71.4
	Total	1197	100.0
Total		4403	

SEQUENCE CLASSIFICATION METHOD BASED ON k-NN CLASSIFIER AND SPECTRUM KERNEL

Overview of kernel method

Using machine learning, there are techniques called kernel methods which are used to construct a maximum-margin separating hyperplane between two separated classes. This particular kernel method is known as a support vector machines (SVMs). The SVM is one of the best-known and most frequently used kernel methods[26]. Vapnik first introduced the kernel method with the principle of structure risk minimization in statistical learning theory[27,28]. In general, a data set is formally represented as

$$D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \{1, -1\}\}; \quad i = \{1, 2, \ldots, n\},$$

where x_i is the ith input vector and y_i is the class of x_i. Each x_i is an n-dimensional vector. Principally, the idea of the kernel method is to construct a maximum-margin hyperplane separating the classes of x.

In the learning process of kernel methods such as SVMs, a hard-margin separation is usually performed, even though labelling errors are unavoidable in many practical problems. To deal with this problem, a soft-margin separation is introduced to mitigate these errors by finding a maximum margin separator which allows misclassifying a training data set[29,30]. In general, when training data sets are nonlinearly separable, the basic idea is to retain the simplicity of linear methods by using mapping functions to map the original data set into a higher dimensional space, called feature space, where linear methods can classify them. The mapping function $\Phi(x)$ is performed by defining the inner product between each pair of data points in the data set of the feature space through the kernel function. Thus if $\Phi(x)$ denotes the mapping function, the kernel function can be expressed as a similarity measurement between the training data set, which is defined as:

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle = \Phi(x)^{\mathrm{T}} \Phi(x').$$

In the classification problem with training data set D, we can predict the class of unknown data x_{n+1} by using a linear decision function determined by the kernel function of inner product between feature vectors. The decision function can be expressed as follows:

$$f(x_{n+1}) = \sum_{i=1}^{N} y_i \alpha_i K(x_{n+1}, x_i)$$

where y_i is the class of x_i, $K(\cdot, \cdot)$ is the kernel function, and α_i is a weighted parameter. More details about the kernel methods can be found in Refs. 26–29.

Spectrum kernel for sequence classification

One of the most widely used kernels for sequence classification is a spectrum kernel or string kernel, which transforms a sequence into a feature vector[30]. The kernel function of strings was first proposed by Watkins[31]. In 2002, Lodhi et al[32] introduced a powerful string subsequence kernel for text classification. Leslie et al[33] showed that the spectrum kernel can effectively be applied to protein classification. Saunders et al[34] also reported on the computational advantages of the spectrum kernel for its fast and simple calculation. If a suitable data structure is used, the prediction can be done in linear time.

The idea behind the spectrum kernel approach is based on the similarity of two strings containing common subsequences. The spectrum kernel is a convolution kernel specialized for the string comparison problem. For a number $sk \geqslant 1$, the sk-spectrum of a sequence x consists of all the possible subsequences of length sk that it contains. Given the alphabet A, a sequence x is transformed into a feature space by a transformation function or feature mapping function.

$$\Phi_{sk}(x) = (\phi_a(x))_{a \in A^{sk}}$$

where $\phi_a(x)$ is the number of times a occurs in x. The kernel function is the inner product of the features vectors:

$$K_{sk}(x, x') = \langle \Phi_{sk}(x), \Phi_{sk}(x') \rangle.$$

The k-NN classifier based on spectrum kernel for sequence classification

The sequence classification methods can be divided in to three main categories: feature based classification, sequence distance base classification, and model based classification. In this work, we develop an approach to classify HLA genes combining feature based classification (i.e., spectrum kernel) and sequence distance base classification (k-nearest neighbour). The simple method of feature based classification used here is the k-NN method. The k-NN method is conceptually based on a distance function to measure the similarity between a pair of objects. The classifier is an instance-based learning algorithm that has been shown to be very effective for a variety of problem domains[19]. Given a labelled sequence data set D, a positive integer k, and a new sequence x to be classified, the k-NN classifier finds the k nearest neighbours of x in D, $\mathrm{knn}(x)$, and returns the dominating class label in $\mathrm{knn}(x)$ as the label of x[35,36].

Basically, the definition of the distance function demonstrates that an appropriate distance function is obviously crucial for the effective performance of a k-NN classifier. The Euclidean metric is the most common distance measure. For classifying HLA genes, there are many computational methods which have been presented for the classification of the HLA genes, such as the SVM method based codon[17] and dicodon usage[20]. Obviously, the feature transformation method was considered to be the crucial process[17,20].

Currently, there are many kinds of distance functions. However, this measure can be inappropriate for high dimensional problems, due to only a few of the features that effectively capture the characteristic information. Furthermore, the well-known distance is sensitive to distortions of the time dimension. In our work, therefore, we propose a combined approach which integrates the k-NN classifier based on spectrum kernel for HLA genes classification. The spectrum kernel is a conceptually simple and efficient way

to perform a sequential classification[37]. In addition, the k-NN method is currently a very well-known and popular classifier.

RESULTS AND DISCUSSION

Description of the experiments

The ten-fold cross-validation procedure was performed on 4275 HLA genes using our combination method to classify HLA genes into major classes and HLA-I/HLA-II genes into their subclasses. In Table 2, the major rows show each k-spectrum considered individually for different k lengths, and the major columns show how each k-NN classifier is considered individually for different k nearest neighbours. We set the k length of the spectrum kernel (sk = 3–6), and set k-NN of the k nearest neighbour (3-NN and 5-NN) to compare the predictive performances.

Given a sequence $x = (x_1, x_2, \ldots, x_n)$ of observations, the major classes $y = (y_1, y_2, \ldots, y_n)$ are obtained by using the combination method, where $y_i \in \{\text{HLA-I and HLA-II}\}$. Moreover, the present method has also been applied to classify subclasses of HLA-I/HLA-II. Here, HLA-I subclasses are focused to classify into HLA-A, HLA-B, HLA-Cw, HLA-F, and HLA-G. The HLA-II subclasses are then classified into HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, and HLA-DRB1. The prediction performances are evaluated with accuracy, precision (Prec), and sensitivity (Sens)[38].

The spectrum kernel method was implemented in the R language using the String Kernel Methods package, named 'stringkernels'[39], which is a simple, customizable, and open-source implementation of string kernel methods. The program 'stringkernels' was designed as free software distributed under a GNU-style copyright based string kernels for use with 'kernlab'[40]. The k-NN classifier was implemented using a more flexible k-NN's package, named 'knnflex'[41]. To compare with the other computational methods, we used the same criterion to analyse the selected data. We selected simple and efficient approaches, i.e., BLAST and profile-HMM, respectively. We also selected the well-known method (i.e., k-NN classifier). For the well-known methods, the kernel support vector machine (KSVMs) spectrum kernel, and NNs[42] were selected to compare against our combination method. For our experiment, linear and polynomial kernel functions were evaluated in the KSVM method. Basically, in the learning process of the KSVMs and NNs, the HLA genes are converted into 59-feature vectors based on the codon usage property[17].

Table 2 Performances of the classification of HLA genes by using the combination method with different sk lengths and k nearest neighbours.

		k-NN	
k-spectrum	Classification	3-NN	5-NN
	All	98.7	86.5
$sk = 3$	Class I	98.9	82.6
	Class II	98.2	97.9
	All	97.7	96.7
$sk = 4$	Class I	97.8	96.7
	Class II	79.3	96.7
	All	97.0	96.7
$sk = 5$	Class I	97.3	96.7
	Class II	96.4	96.7
	All	99.7	98.7
$sk = 6$	Class I	99.4	98.9
	Class II	98.8	98.2

Results and discussion of the combination method

In our experiments, we started with a 3-spectrum, since the 3-spectrum corresponds to a codon which is a fundamental property in molecular evolution. Currently, Ma et al and Nguyen et al[17,20] proposed that this feature can directly represent a utility in molecular characterization of species.

For our experiment, the combined 3-NN and 3-spectrum was first considered as a simple combination. This combination was found to have 98.7% accuracy for the major class classification, and reached 98.8% accuracy for the HLA-II sub-class classification. We then increased sk-spectrum lengths to 4 and 5 to compare the simple combination. These two sk-spectrum lengths give a decrease in the average predictive performances of both the major-class and sub-class classification (from 99.0 to 97.0). These results show that two sk-spectrum lengths may not be enough to represent all information in HLA genes. In our experiment, the high performance increases reaching 99.7% (from 97.0–99.7) for major class classification when using 3-NN with 6-spectrum. Moreover, these optimized parameters also give the predictive performance of subclass classification of HLA-I and HLA-II as high as 99.4% and 98.8%, respectively. In biological knowledge, the 6-spectrum is known as a di-codon which is a fundamental unit of gene transcription and molecular evolution. Moreover, this feature could be a good indicator in gene expression and molecular evolution studies and provide a rich feature set for gene classification[43,44].

We also compared the predictive performance

Table 3 The ten-fold cross validation accuracies of the classification of HLA major classes using different classification methods.

Methods	Acc	HLA-I		HLA-II		Chi squared
		Prec	Sens	Prec	Sens	(p-value)
profile-HMM	98.5	98.3	99.8	99.3	94.8	135.7 (2.2×10^{-16})
BLAST	98.2	97.9	99.7	99.1	93.3	164.2 (2.2×10^{-16})
NNs	87.1	94.9	87.5	68.7	86.0	1.6 (0.21)
KSVM (linear)	98.5	98.9	99.2	97.7	96.8	33.3 (7.6×10^{-9})
KSVM (poly)	98.4	99.1	98.4	96.3	97.2	10.7 (1×10^{-3})
KNN classifier	95.7	97.7	96.9	90.9	92.0	45.69 (1.4×10^{-11})
String kernel	99.4	99.5	99.7	99.1	98.6	14.6 (1×10^{-4})
Combined method	99.4	99.2	100.0	100.0	97.8	164.2 (2.2×10^{-16})

for sk spectrum lengths equal to 7, 8, and 9 (data not shown). The results show that the performances of these increasing lengths are close to using the 6-spectrum. As indicated, a larger k length does not necessary obtain a better performance for HLA gene classification. It can be concluded that the 6-spectrum is long enough to capture all informative features of HLA genes; or the over-estimated k-spectrum may introduce useless information or features into the learning process.

Comparison of the HLA major class classification with other classification methods

In the previous section, our optimized feature is the 6-spectrum and 3-NN that give the highest predictive performance. To compare with the other computational methods, we used the same criterion to analyse selected data.

Table 3 lists the predictive performance and chi squared (p-value) with various computational methods. BLAST and profile-HMM were first considered which yield 98.2% and 98.4% accuracies, respectively. These two simple approaches yielded high accuracies, because of their high identity scores in the HLA major classes. Therefore, the BLAST method can provide accuracy reaching 99.0%. However, this method considers only the most similarity, but the full length alignment is ignored to take advantage for analysing. The k-NN classifier was used to compare our approach, this classifier gave an accuracy of 95.7%. In our experiment, there are two classifiers based codon usage properties, i.e., SVMs, and NNs. The SVM classifiers readily yielded 98.0% accuracy when using both linear and polynomial kernel functions. When classifying by using NNs, we obtain accuracies lower than 90.0%. As indicated, the SVM classifier is suitable for classifying HLA genes, when the HLA genes are represented with the codon usage property. We further compared the spectrum kernel which is a powerful measurement[32] for text classification. The spectrum kernel considerably increases the accuracy (from 98.2 or 98.4–99.4). However, there is one disadvantage of the spectrum kernel in that it is hard to interpret and hard for the user to gain additional knowledge besides the classification results. In the HLA genes classification, our approach achieved maximum performance results in the cases of sensitivity of HLA-I and precision of HLA-II. For the other performance tests, the string kernel method yielded performance results which were better than our approach.

In Table 3 on the last column, we show the chi squared test (p-value). There is one computational method which has a p-value > 0.005, i.e., the NNs. This result shows that this model consistently has high predictive performance for the two classes of HLA genes. Our combined method yields a p-value of 2.2×10^{-22}.

Comparison of the HLA-I/HLA-II subclasses classification with other classification methods

In Table 4, we further compare our combination with the other computational methods for the HLA-I/HLA-II subclass classification problem. Most of the computational methods gave accuracy values as high as 90% on both the HLA-I and HLA-II sequences. However, there are two computational methods which yielded accuracies higher than 95% for both the HLA-I and HLA-II, i.e., string kernel and our combined method. We found that the k-NN classifier can be suitable for classifying the major classes, but this classifier could not efficiently classify the HLA-II sequences, because this classifier is sensitive to imbalances in the samples and to the distance measure. Our approach can increase the accuracy in HLA-II subclass (from 88.9–98.8) when the k-NN

Table 4 Ten-fold cross validation accuracies of the HLA-I/HLA-II subclasses classification using different classification methods.

Classification	profile HMM	BLAST	NN	KSVM (linear)	KSVM (poly)	K-NN	String kernel	Combined method
Sub-class of HLA-I	97.1	96.6	85.6	97.9	98.9	93.8	98.0	99.4
Sub-class of HLA-II	92.2	91.8	87.6	92.6	90.6	88.9	99.6	98.8

classifier was integrated with the spectrum kernel. Since the spectrum kernel can represent HLA genes with suitable vectors, especially 6-spectrum. In the HLA-I subclass, our combined approach obtained the maximum performance as high as 99.4% accuracy. In the HLA-II subclass, the string kernel gave 99.6% accuracy which is better than our approach. However, the average accuracy of our approach is similar to the string kernel in both of the HLA-I/HLA-II subclass classifications.

Sequence classification on human genes

The human (*Homo sapiens*) genome consists of 23 chromosome pairs and the small mitochondrial DNA. Chromosome 6 is one of these human chromosomes which contain the Major Histocompatibility Complex, with over 100 genes related to the immune response, and plays a vital role in organ transplantation. We used our combined approach to classify the human genes based on the latest release (NCBI Build 37.3) of 32 185 genes[45]. Given a human gene $x = (x_1,\ x_2, \ldots, x_n)$, the classes $y = (y_1, y_2, \ldots, y_n)$ are obtained by using the combination method, where $y_i \in \{\text{HLA, other}\}$. In the experimental results, our approach gives predictive performances more than 90%, which a considerable decrease in precision for the HLA (76.3%) sequences. Since the k-NN classifier is sensitive to both the positive data set and negative data set.

CONCLUSIONS

We have proposed a combination of the k-nearest neighbour method based on the spectrum kernel which is a simple approach for gene classification. The combination method is performed on the problem of classification of HLA genes. For our experimental results, our approach gives a maximum of 99.4% accuracy in the major class classification, and achieves as high as 100.0% in cases of sensitivity of the HLA-I and precision of HLA-II sequences. Moreover, in subclass classification, our approach still provides a higher performance for HLA-I subclass (99.4%) compared with the other computational methods. For our experiment, we found that the k-NN classifier gives the highest accuracy for the major class classification (95.7%), but, this method considerably decreases in subclass classification, i.e., 88.9% accuracy on HLA-II. Since, this subclass has a lower identity score. The string kernel yields the highest results for the HLA-II subclass classification and is also close to our approach. In practice, our approach is simple to understand, but surprisingly is able to accurately classify the HLA genes. Our predictive performances are better than the other computational methods for the HLA-I subclass classification, but the prediction accuracy of the string kernel is better than our approach for the classification of the HLA-II subclass. However, our approach already provides highly predictive performances; it may be improved by applying mismatch kernels which allow inexact matching of substrings.

Acknowledgements: We thank R. Cutler for helpful comments and suggestion on this manuscript. This project was supported by the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Programme (Grant No. PHD/0209/2552), Chiang Mai University's Graduate School, and the National Centre for Genetic Engineering and Biotechnology, Thailand.

REFERENCES

1. Wedekind C, Escher S, Waal MV, de Frei M (2007) The major histocompatibility complex and perfumers descriptions of human body odors. *Evol Psychol* **5**,: 330–43.
2. Nail S (2003) The human HLA system. *J Indian Rheumatol Assoc* **11**, 79–83.
3. Shankarkumar U (2004) The human leukocyte antigent (HLA) system. *Int J Hum Genet* **4**, 91–103.
4. Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. *Crit Rev Immunol* **17**, 179–224.
5. Browning A, Michael MC (1996) HLA and MHC: Genes, molecules and function, *Bios Scientific Publishers*, Oxford.
6. Katz DH, Hamoaka T, Benacerraf B (1973) 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**, 1405–141.
7. Han HX, Kong FH, Xi YZ (2000) Progress of studies on the function of MHC in immuno-recognition. *Chin J Immunol* **16**, 15–7.
8. Altschul SF, Gish W, Miller W, Myers EW, Lipmanl DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403–10.
9. Baldi B, Chauvin Y, Hunkapiller T, McClure MA (1994) Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci Unit States Am* **91**, 1059–63.
10. Krogh A, Brown M, Mian I, Sjölander K, Haussler D (1994) Hidden Markov models in computational biology: applications to protein modeling. *J Mol Biol* **235**, 1501–31.
11. Jaakkola T, Diekhans M, Haussler D (2000) A discriminative framework for detecting remote protein homologies. *J Comput Biol* **7**, 95–114.
12. Jaakkola T, Diekhans M, Haussler D (2000) Using the Fisher kernel method to detect remote protein homologies. *J Comput Biol* **7**, 95–114.
13. Kuang R, Ie E, Wang K, Siddiqi M, Freund Y, Leslie C (2005) Profile-based string kernels for remote homology detection and motif extraction. *J Bioinformatics Comput Biol* **3**, 527–50.
14. Leslie C, Eskin E, Cohen A, Weston J, Noble WS (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20**,: 467–76.
15. Liao L, Noble WS (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J Comput Biol* **10**, 857–68.
16. Chua HN, Sung W-K (2005) A better gap penalty for pairwise SVM. *Proc APBC*. 11–21.
17. Ma JM, Nguyen MN, Rajapakse JC (2009) Gene classification using codon usage and support vector machines. *IEEE ACM Trans Comput Biol Bioinformatics* **6**, 134–43.
18. Han JW, Kamber M (2001) *Data Mining: Concepts and Techniques*. Academic Press.
19. Aha D, Kibler D (1991) Instance-based learning algorithms. *Mach Learn* **6**, 37–66.
20. Nguyen MN, Ma JM, Fogel GB, Rajapakse JC (2009) Di-codon usage for gene classification. *Lect Notes Comput Sci* **5780**, 211–21.
21. Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P, Marsh SGE (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res* **31**, 311–4.
22. Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE (2000) IMGT/HLA and IMGT/MHC: sequence databases for the human major histocompatibility complex. *Tissue Antigens*. 55, 280–287.
23. Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE (2001) IMGT/HLA and IMGT/MHC: sequence databases for the human major histocompatibility complex. *Nucleic Acids Res* **29**, 210–3.
24. Robinson J, Waller MJ, Fail SC, McWilliam H, Lopez R, Parham P, Marsh SEG (2009) The IMGT/HLA database. *Nucleic Acids Res* **37**, 1013–7.
25. Marsh SG, Bodmer JG, Albert ED, et al (2001) Nomenclature for factors of the HLA system. *Tissue Antigens* **57**, 236–83.
26. Vapnik V (1995) *The Nature of Statistical Learning Theory*, Springer.
27. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* **20**, 273–97.
28. Vapnik V (1998) *Statistical Learning Theory*, John Wiley & Sons.
29. Cristianini N, Shawe-Taylor J (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge Univ Press.
30. Gärtner T (2003) A survey of kernels for structured data. *SIGKDD Explor Newslett* **5**, 49–58.
31. Watkins C (1990) Kernel from matching operation, Tech. rep. Department of Computer Science, Royal Holloway, Univ of London.
32. Lodhi H, Saunders C, Taylor JS, Cristianini N, Watkins C (2003) Text classification using string kernels. *J Mach Learn Res* **2**, 419–44.
33. Leslie C, Eskin E, Noble WS (2001) The spectrum kernel: A string kernel for SVM protein classification. *Proc Pac Symp* **7**, 564–75.
34. Saunders C, Tschach H, Shawe-Taylor J (2002) Syllables and other string kernel extensions. In: Sammut C, Hoffmann AG (eds) *Proceedings of the 19th International Conference on Machine Learning (ICML02)*, pp 530–7.
35. Li D, Deogun JS, Wang K (2007) Gene function classification using fuzzy k-nearest neighbor approach. In: *Proceedings of the 2007 IEEE International Conference on Granular Computing*, 644.
36. Gao YJ, LiC, Chen GC, Chen L, Jiang XT, Chen C (2007) Efficient k-Nearest-neighbor search algorithms for historical moving object trajectories. *J Comput Sci Tech* **22**, 232–44.
37. Schölkopf B, Tsuda K, Vert JP (2004) *Kernel Methods in Computational Biology*, The MIT Press.
38. Okori W, Obua J (2011) Machine learning classification technique for famine prediction. In: *Proceedings of the World Congress on Engineering 2011*, Vol II WCE 2011.
39. Kober M (2010) 'stringkernels' (String Kernel Methods for kernlab); software available at http://cran.r-project.org/web/packages/stringkernels.
40. Karatzoglou A, Smolo A, Hornik K (2009) 'kernlab' (Kernel-based Machine Learning Lab); software available at http://cran.r-project.org/web/packages/kernlab.
41. Brooks AD (2009) 'knnflex' (knnflex: A more flexible KNN); software available at http://cran.r-project.org/web/packages/knnflex.
42. Ripley B, Hornik K, Gebhardt A (2009) 'nnet' (Feed-

forward Neural Networks and Multinomial Log-Linear Models); software available at http://cran.r-project.org/web/packages/nnet.
43. Milhon JL, Tracy JW (1995) Updated codon usage in Schistosoma. *Exp Parasitol* **80**, 353–6.
44. Mitreva M, Wendl MC, Martin J, et al(2006) Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species. *Genome Biol* **7**, R75.
45. Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**, (Database issue), D32–6.