CHAPTER 3

Prediction of Human Leukocyte Antigen Gene Using k-Nearest Neighbor Classifier Based on Spectrum Kernel

3.1 Introduction

In this work, we focused on the problem of the prediction of HLA genes. In this work, the prediction of HLA genes can be further subdivided into two related sub-problems, i.e., the prediction of the major and subclasses of the HLA genes. 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 [4-7]. In the case of subclass HLA-I, the genes are classified into HLA-A, HLA-B, HLA-Cw, HLA-E, HLA-F, and HLA-G. The sub-classes of the 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.

Ma *et al* first proposed the use of support vector machine (SVMs) based codon usage, where each element corresponds to the relative synonymous codon usage (RSCU) frequency of a codon [57]. This particular one is known as a feature based classification method. The SVM classifier based on codon usage was compared with k-means clustering, linear discriminant analysis (LDA) [58], and *k*-NN classifier [23,59]. In 2009, di-codon usage [60] was used to compare codon usage for improving HLA gene prediction. This work showed that using di-codon usage as a feature input outperforms using codon usage alone.

Table 3.1 Existing methods for predicting HLA genes

Author	Classifier	Kernel	Dataset		Remove[a]	Features	#feature
			training	testing			
Ma et al, 2009 [57]	SVM	RBF	1,840	-	No	condon usage	59
Nguyen et al, 2009 [60]	SVM	RBF	1,840	-	No	di-condon usage	4096
Our work [1]	***k*-NN**	**String kernel**	**4,398**	**-**	**No**	**sequence similarity score**	**>4000**

[a]remove sequence identity

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 [1]. Since, the k-NN method is the simple method for classification problems. And, a spectrum kernel is one of the most widely used kernels for sequence classification. In our experiments, we designed a series of combination parameters, which allowed us to determine relative contributions from a spectrum kernel and a k-NN method.

3.2 Dataset

Recently, there has been an increase in the number of nucleic acid and protein sequences in the international immunogenetics databases [61-63], 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 (MHC). This complex of over four megabases is located

within the 6p21.3 region of the short arm of human chromo-some 6 and contains an excess of 220 genes [64]. 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 [65].

Table 3.2 Numbers of HLA genes and their subclasses (%)

Major Class	Subclass	Number of sequences	Percentage (%)
HLA Class I	HLA-A	965	30.1
	HLA-B	1,540	48.0
	HLA-C	626	19.5
	HLA-E	9	0.3
	HLA-F	21	0.7
	HLA-G	45	1.4
	Total	3,206	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	1,197	100.0
Total		4,403	

The IMGT/HLA database 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 detail about the data set can be found in Robinson J *et al* [62, 64]. Table 3.2 shows the numbers and percentages of HLA genes and their subclasses used in our experiment.

3.3 HLA gene prediction using *k*-NN classifier based on spectrum kernel

In this work, we developed an approach to predict HLA genes by forming a *k*-NN classifier based on a spectrum kernel. The system flowchart of the *k*-NN classifier based on a spectrum kernel is shown in Figure. 3.1. The HLA gene is first transformed as suitable feature vector by using the spectrum kernel. 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. Currently, although, there are many kinds of distance functions, the simple Euclidean distance is the most widely used option. However, this measure can be inappropriate for high dimensional problems, due to only a few of the features that can effectively capture the characteristic information. Furthermore, the Euclidean distance is sensitive to distortions in the time dimension [16]. These drawbacks suggest that the Euclidean distance may not always be suitable for some tasks of sequence classification.

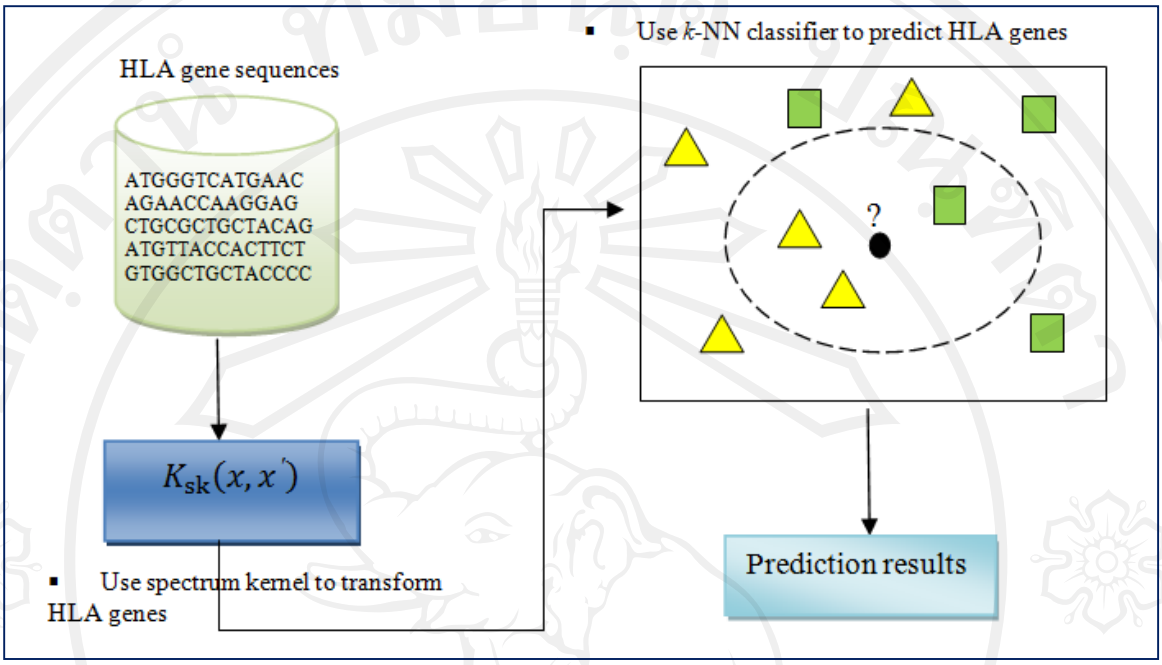

Figure 3.1 - The system flowchart of the *k*-NN classifier based on spectrum kernel to classify HLA gene.

Thus, we proposed a spectrum kernel which is well-known for many tasks of sequence classification [18-19, 38-40]. In this work, we used the combined to approach integrating the *k*-NN classifier based on the spectrum kernel to predict HLA gene, since the spectrum kernel is a conceptually simple and efficient way to perform a sequential classification [18-19].

3.4 Results and discussion

3.4.1 Description of the experiments

The ten-fold cross-validation procedure was performed on 4,275 HLA genes using our combination method to predict HLA genes into major classes and HLA-I/HLA-

II genes into their subclasses. In Table 3.3, the major rows show each sk-spectrum considered individually for different lengths of spectrum, and the major columns show how each *k*-NN classifier is considered individually for different *k* nearest neighbors. We set the sk length of the spectrum kernel (sk= 3-6), and also set kNN (or k-NN) of the k nearest neighbor (k=3 and 5) to compare the predictive performances. Given a sequence of observations $x = (x_1, x_2, \dots, x_n)$, the major classes $y = (y_1, y_2, \dots, y_n)$ are obtained by using the combination method, where $y_i \in$ {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 on 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 (ACC), precision (Prec), and sensitivity (Sens).

The spectrum kernel was implemented in the R language using the String Kernel Methods' package, named stringkernels [66], which is a simple, customizable, and open-source implementation of the string kernel methods. The program stringkernels was designed as free software distributed under a GNU-style copyright based string kernels for use with kernlab [67]. The *k*-NN classifier was implemented using a flexible *k*-NN's package, named knnflex [68]. To compare with other computational methods, we used the same criterion to analyze the selected data. We also selected a well-known method (i.e. *k*-NN classifier). For other well-known methods, the kernel support vector machine (KSVMs) spectrum kernel, and NNs [69] 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 [57].

3.4.2 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 unit in molecular evolution. Ma *et al* [57] proposed that this feature can be represented to have utility in the molecular characterization of a species. For our experiments, the combined 3-NN and 3-spectrum was first considered as a simple combination to examine in the prediction of HLA genes. This combination gave 98.7% accuracy for the major class classification, and reached 98.8% accuracy for the HLA-II sub-class classification. We then increased the sk-spectrum lengths to 4 and 5 to compare the simple combination. These two sk-spectrum lengths gave a decrease in the average predictive performances of both the major-class and sub-class classification (from 99.0 to 97.0). These results showed that the two sk-spectrum lengths might not be enough to capture or represent all information in HLA genes. Thus, using 3-NN with 6-spectrum was then proposed to improve predictive accuracy. In our experiment, the performance increased reaching 99.7 % (from 97.0 to 99.7) for the major class classification when using such selected parameters. Moreover, these optimum parameters also gave 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 also a fundamental unit of

gene transcription and molecular evolution. Moreover, this feature could be reported as a good indicator in gene expression and molecular evolution studies and also provides a rich feature set for gene classification [70-71].

Table 3.3 Performances of the classification of HLA genes by using the combination method with different k lengths and *k* nearest neighbors.

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

We also compared the predictive performance for sk spectrum lengths equal to 7, 8, and 9 (data not shown). The results showed that the performances of these increasing lengths were close to using the 6-spectrum. As indicated, a larger k length did not necessary obtain a better performance for HLA gene classification. This is a case where the simple

classifier may be better than a complex one. In conclusion, it can be said that the 6-spectrum was powerful enough to capture all informative features of HLA genes; or the over-estimated k-spectrum might introduce useless information into the learning process. Table 3.4 lists the accuracies of the HLA genes subclass classification using our combined method (6-spectrum and only 3-NN). Thus, in this work, we selected suitable parameters (6-spectrum and 3-NN) for future analysis.

In Table 3.3, our approach obtains accuracies as high as 100.0% for the subclass E, F, and DRB1. Since, the characteristics of the three subclasses are high identity within their classes.

Table 3.4 The ten-fold cross validation accuracies for the classification of HLA-I and HLA-II into their subclasses using the combined method.

Major class	Subclass	Accuracy (%)
HLA-I	A	99.6
	B	99.9
	C	99.0
	D	88.9
	E	100.0
	F	100.0
HLA-II	DPA1	88.9
	DPB1	97.1
	DQA1	91.4
	DQB1	97.2
	DRB1	100.0

3.4.3 Comparison of the HLA major class classification with other classification methods

In the previous section, our optimized feature used the 6-spectrum and 3-NN to seek our high predictive performance. To compare with the other computational methods, we used the same criterion to analyze the selected dataset of HLA genes. The *k*-NN classifier was used to compare to our approach; this classifier gave an accuracy of 95.7%. In our experiment, there are two classifiers based condon usage properties, i.e. SVMs and NNs. The SVM classifiers yielded 98.0% accuracy when using both linear and polynomial kernel functions. Using NNs obtained accuracies lower than 90.0%. As indicated, the SVM classifier was suitable for classifying HLA genes, when the HLA gene was represented with the codon usage property. We further compared the spectrum kernel (by using with SVM classifier) which is a known powerful measure [39] for text classification. The spectrum kernel considerably increases the accuracy (from 98.2 or 98.4 to 99.4) as is shown in Table 3.5. 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 our results, our approach obtained maximum predictive results in the cases of sensitivity of HLA-I and precision of HLA-II. For the other performance tests, the spectrum kernel method yielded performance results which were better than our approach.

Table 3.5 The ten-fold cross validation accuracies of the classification of HLA major classes using different classification methods.

Methods	Acc	HLA-I		HLA-II	
		Prec	Sens	Prec	Sens
Neural Network	87.1	94.9	87.5	68.7	86.0
SVM (linear)	98.5	98.9	99.2	97.7	96.8
SVM (poly)	98.4	99.1	98.4	96.3	97.2
k-NN (Euclidean)	95.7	97.7	96.9	90.9	92.0
SVM (spectrum)	**99.4**	**99.5**	99.7	99.1	**98.6**
k-NN (spectrum)	**99.4**	99.2	**100.0**	**100.0**	97.8

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

In Table 3.6, we further compared 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. the 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 also to the distance

measure. Our approach can increase the accuracy in HLA-II subclass (from 88.9 to 98.8) when the *k*-NN classifier was integrated with the spectrum kernel. Since, the spectrum kernel can represent HLA genes with suitable vectors, especially the 6-spectrum.

Table 3.6 Ten-fold cross validation accuracies of the HLA-I/HLA-II subclasses classification using different classification methods.

Classification	NN	SVM (linear)	SVM (poly)	k-NN (Euclidean)	SVM (spectrum)	k-NN (spectrum)
Sub-class of HLA-I	85.6	97.9	98.9	93.8	98.0	**99.4**
Sub-class of HLA-II	87.6	92.6	90.6	88.9	**99.6**	98.8

In the HLA-I subclass, our combined approach obtained the maximum performance of 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.

3.4.5 Sequence classification on human genes

The human (Homo sapiens) genome consists of 23 chromosome pairs and the small mitochondrial DNA. Chromosome 6 is the human chromosomes which contains 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 (Build 37.3) of 32,185 genes [72]. Given a human gene $x = (x_1, x_2, \dots, x_n)$, the classes $y = (y_1, y_2, \dots, y_n)$ are obtained by using the combination method, where

$y_i \in$ {HLA, other}. In the experimental results, our approach gives predictive performances over than 90%, which is a considerable decrease in precision for the HLA (76.3%) sequences. Since the *k*-NN classifier is the simple classifier using only a few parameters and has less complex system. It is possible that in this case our approach gave a decrease in precision for the HLA.

3.4.6 Discussion

As seen, Table 3.5 and 3.6 show the predictive performances of existing classifiers to classify HLA genes. Most classifiers obtained higher than 90.0% accuracy. It is not surprising that these results were obtained, because our dataset did not remove high similarity sequence of the HLA genes. Thus, using simplest classifier, i.e. *k*-NN classifier, can give high predictive accuracy. Basically, in the machine learning task, any computational method is proposed; the dataset, which is used to perform, must be removed (high) sequences identity. Therefore, for this thesis, we removed sequence identity (reduced to 30%), and the numbers of new sequences of HLA gene are shown in Table 3.7. We give only the comparison predictive results of *k*-NN classifier, SVM classifier, and our proposed methods for predicting major/subclass, which are shown in Table 3.8-3.9.

Table 3.8 shows that the SVM using a linear kernel (simple kernel function) give higher accuracy for predicting major classes than other classifiers. Our combined method gives lower result than the SVM classifier. And the SVM classifier using polynomial degree=4 obtains higher predictive results for predicting the subclasses of HLA-I and

HLA-II compared to other classifiers. In this work, the classification of sub-class of HLA-I, our approach is comparable to the other classifiers. These results show that the characteristics of the HLA genes in the major classes are more similar as compared to the subclass. Thus, this noticeable point leads to the accuracy in predicting subclass of HLA-II, i.e. all classifiers obtain lower that 90% accuracy. However, the redundant dataset should be removed as < 25% sequence similarity to verify the efficient of this new method.

Table 3.7 Numbers of removed high sequence identity of HLA genes

Major Class	Subclass	Identity <30 %
HLA Class I	HLA-A	28
	HLA-B	33
	HLA-C	10
	Total	71
HLA Class II	HLA-DPB1	3
	HLA-DQA1	8
	HLA-DQB1	15
	HLA-DRB	36
	Total	62
Total		133

Table 3.8 The ten-fold cross validation accuracies of the HLA major classes classification using different classification methods (using sequence identity < 30%).

k-NN (Euclidean)	SVM (linear)	SVM (poly=2)	SVM (poly=3)	SVM (poly=4)	k-NN (spectrum)
91.73	98.5	97.74	92.25	92.25	**95.49**

Bold is the result of our approach to predicting HLA gene

Table 3.9 Ten-fold cross validation accuracies of the HLA-I/HLA-II subclasses classification using different classification methods (using sequence identity < 30%).

Classification	*k*-NN (Euclidean)	SVM (linear)	SVM (poly=2)	SVM (poly=3)	SVM (poly=4)	k-NN (spectrum)
Sub-class of HLA-I	78.87	84.51	77.46	84.51	88.73	**83.10**
Sub-class of HLA-II	85.48	87.71	87.71	87.71	87.71	**87.71**

Bold is the result of our approach to predicting HLA gene

3.5 Conclusion

We have proposed a combination of the *k*-NN classifier based on the spectrum kernel approach to classify HLA genes. For our experimental results, our approach gives a maximum of 99.4% accuracy in the major class classification, and also achieves as high as 100.0% in cases of sensitivity of the HLA-I and precision of HLA-II sequences. Moreover, in the subclass classification problem, our approach still provides a higher performance for the 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 incredibly it 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.