

# Di-Codon Usage for Classification of Genes

Minh N. Nguyen<sup>a</sup>, Jianmin Ma<sup>a</sup>, Gary B. Fogel<sup>b</sup>,  
Jagath C. Rajapakse<sup>c,d,e,\*</sup>

<sup>a</sup>*BioInfomatics Institute, Singapore*

<sup>b</sup>*Natural Selection Inc., San Diego, USA*

<sup>c</sup>*BioInformatics Research Centre, Nanyang Technological University, Singapore*

<sup>d</sup>*Singapore-MIT Alliance, Singapore*

<sup>e</sup>*Department of Biological Engineering, Massachusetts Institutes of Technology,  
USA*

---

## Abstract

Genes are often classified into biologically-related groups so that inferences on their functions can be made. This paper demonstrates that the di-codon usage is a useful feature for gene classification and gives better classification accuracy than the codon usage. Our experiments with different classifiers show that support vector machines performs better than other classifiers in classifying genes by using di-codon usage as features. The method is illustrated on 1,841 HLA sequences which are classified into two major classes, HLA-I and HLA-II, and further classified into the subclasses of major classes. By using both codon and di-codon features, we show near perfect accuracies in the classification of HLA molecules into major classes and their subclasses.

*Key words:* Di-codon usage; gene classification; Human Leukocyte Antigen (HLA); Major Histocompatibility Complex (MHC); Support Vector Machines (SVM)

---

## 1 INTRODUCTION

Genetic information encoded in nucleic acids is transferred to proteins through 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. In addition, it could also be very useful for mutation studies. When a synonymous

---

\* Corresponding author: asjagath@ntu.edu.sg

mutation occurs, codon usage varies while the resulting protein remains unchanged. Therefore, codon usage is a good indicator for the studies of mutation and molecular evolution. The pattern of codon usage has been found to be highly variable among different species and is mainly attributed to gene function (Sharp et al., 1988). The species-specific characteristics of codon usage was used to classify genes from 18 different species, mainly prokaryotes and unicellular eukaryotes Kanaya et al. (1999). We have recently shown that the codon usage can be used as a potential feature for classification of HLA molecules (Ma et al., 2009). Furthermore, the classification of HLA molecules based on codon usage bias was consistent with the structure and function of the molecules.

Experimental approaches for gene classification often use microarray data, yet gathering of such data is often limited because of the cost and tediousness involved in the experiments. Researchers have begun to use computational techniques such as neural networks, support vector machines (SVM), independent component analysis, etc., which use features of gene expressions to classify gene expressions and identify important genes into biologically meaningful groups (Zhang and Rajapakse, 2009; Eisen et al., 1998; Tamayo et al., 1999; Rajapakse et al., 2005; Hori et al., 2001). Because of the large dimensions of microarray data and the limited sample sizes, these methods have limited utility on larger datasets though high-throughput gene selection methods are beginning to appear.

Other methods of assigning proper functions to genes include homology-based approaches through multiple sequence alignment of proteins or nucleic acid sequences (Wallace et al., 2005). Because of the complexities of multiple sequence alignment in time, it is relatively difficult to classify a large number of genes or proteins with this approach. Furthermore, if the sequences in the set vary in length or in evolutionary conservation, a correct alignment is hard to achieve. This can also affect downstream uses of the alignment data such as for gene classification. More importantly, the information from synonymous mutations is often neglected in homology-based approaches despite the importance of synonymous mutations in evolution. Structural features of proteins have been used to classify genes (Shatsky et al., 2006), which also neglect the importance of synonymous mutations.

In this paper, we demonstrate the use of di-codon usage as a promising feature for gene classification. Preliminary work of this study has been presented in Nguyen et al. (2009). Di-codon usage patterns contain additional or more information for gene classification than the codon usage because di-codon usage patterns encapsulate local information as well as global information (di-codon frequency) of a DNA sequence. Given that the ribosomes actually reside over two codon positions when they slide along mRNA, di-codon usage has a direct biological rationale. Noguchi et al. (2006) developed a prokaryotic gene-finding

program, MetaGene, which utilizes di-codon frequencies estimated by the GC content of a given sequence along with other various measures. By using di-codon frequencies, their method achieved a higher prediction accuracy than using codon frequencies alone (Noguchi et al., 2006). A hidden Markov model with self-identification learning for protein coding region identification was studied by Kim et al. (1999) and demonstrated that models using di-codon features outperformed other models using "standard" features such as amino acid pair, codon usage, and GC content in terms of both specificity and sensitivity. The DicodonUse program based on frequencies of di-codons is aimed at a fast and simple assessment of genes present in prokaryotic nucleotide sequences (Paces and Paces, 2002). It is used for identification of open reading frames that have a high probability of being genes.

We demonstrate our method on HLA molecules and extract dicodon usage as features for classification. Binary and multi-class SVM were used for the classification of HLA genes into HLA-I and HLA-II classes, and their subclasses, respectively. SVMs have successfully been used in many bioinformatics applications: for example, predicting protein features (Nguyen and Rajapakse, 2005, 2006, 2007) and classifying gene and protein expressions (Rajapakse et al., 2005; Duan and Rajapakse, 2005). Furthermore, SVM have been used for classification of genes because of their scalability and generalization: for example, Lin et al. used them to study the conserved codon composition of ribosomal protein coding genes in *E. coli*, *M. tuberculosis*, and *S. cerevisiae* (Lin et al., 2002). Bhasin and Raghava used SVM in the prediction of HLA-DRB1\*0401 binding protein and cytotoxic T lymphocyte (Tc) epitopes (Bhasin and Raghava, 2004a,b). Tonnes and Elofsson used SVM to predict MHC class I binding peptides (Tonnes and Elofsson, 2002), and Zhao et al. also applied SVM in the prediction of T-cell epitopes (Zhao et al., 2003). SVM were found to give the best accuracy in gene classification using codon usage bias as input features (Ma et al., 2009). The success of SVM is mainly because of its generalization abilities and its capability of handling high-dimensional data.

The proposed approach achieved substantial improvement in classification accuracy on a dataset of 1,841 HLA gene sequences collected from the IMGT/HLA Sequence Database. We compare our results with homology-based gene classification methods as well as SVM using only codon usage as an input feature. We also experimented with different classifiers such as linear discriminant analysis (LDA) and K-nearest neighbors (KNN) ( $k$ -NN). The SVM provides the optimal margins of separation of the classifiers and uses only the boundary points or support vectors for classification. The LDA assume Gaussianity of classes while KNN assumes no distribution regarding the classes. This enables us explore which classifier better fits the distribution of codon usage.

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. By combining codon and di-codon features, the prediction accuracies of 100%, 99.82%, and 99.84% were achieved for HLA major class classification, and for sub-class classification of HLA-I and HLA-II genes, respectively.

## 2 MATERIALS AND METHODS

### 2.1 Data

Recently, there has been a rapid increase of the number of nucleic acid and protein sequences in the international immunogenetic databases (Robinson et al., 2001, 2003; Galperin, 2004). These sequences have enabled computational biologists to study human and primate immune systems. 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 produces glycoprotein products to initiate the immune response of the body (Bodmer et al., 1995). HLA and human MHC molecules are a vital component of immune response and take part in the selection process of thymus cells, genetic control of immunological reactions, and immunocyte interactions. The primary function of HLA molecules is to bind and present antigens on the cell surface for recognition by antigen-specific T-cell receptors (TCR) of lymphocytes. Immune reactions involve interactions between HLA molecules and T lymphocytes (Rosenthal and Shevach, 1973); T-cell response has subsequently been restricted not only by the antigen but also by HLA binding (Zinkernagel and Doherty, 1974). Furthermore, HLA molecules are involved in the production of antibodies, a process which also involves class II gene products restricted by gene products from the class II molecules (Katz et al., 1973; Han et al., 2000). 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. To this end, classification of HLA molecules into functionally related groups could provide important clues.

We demonstrate our approach of gene classification through the classification of HLA molecules into major classes and their sub-groups. HLA molecules are generally classified into three classes: HLA-I, HLA-II, and HLA-III, according to their specific functions in the immune system (Katz et al., 1973; Han et al., 2000). The major classes are further 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 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, protecting 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, inside the 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 (Robinson et al., 2001, 2003; Galperin, 2004) of EBI (<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. The errors were then curated manually (Ma et al., 2009).

Because there are 61 different codons coding for amino acids, in order to have a sufficient number of codons for computation of codon usage, coding sequences of less than 50 amino acids were excluded from our analysis (Ma et al., 2009), resulting in 1,841 HLA genes. More details regarding this dataset are available in (Ma et al., 2009). Di-codon usage patterns were calculated for each sequence and used as input features for SVM for classification of HLA molecules into their main classes and sub-classes.

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.

## 2.2 Codon Usage

Let the coding sequence of the gene in terms of codons be  $s = (s_1, s_2, \dots, s_n)$  where  $s_i \in \Omega$ ,  $n$  is the length of the sequence in codons, and  $\Omega = \{c_1, c_2, \dots, c_{64}\}$  is the alphabet of the codons. The usage  $r_{c_j}$  of the codon  $c_j$  is measured by

the fraction of codons  $c_j$  of the sequence  $s$ :

$$r_{c_j} = \frac{1}{n} \sum_{i=1}^n \delta(s_i = c_j) \quad (1)$$

where  $\delta$  is the Kronecker delta:  $\delta(\cdot) = 1$  if the argument inside is satisfied, otherwise is 0. The codon usage  $r_c$  of a given gene sequence  $s$  is a vector  $r_c = (r_{c_j} : j = 1, 2, \dots, 64)$  of length 64, regardless of the length of the sequence.

### 2.3 Di-Codon Usage

The di-codon usage pattern is given by the fractions of di-codons in the coding sequence and captures more global information about the gene sequence than the codon usage pattern. The di-codon usage  $r_{c_j c_k}$  of a di-codon  $c_j c_k$  is measured by the fraction of codon pairs  $(c_j, c_k) \in \Omega^2$  in the sequence  $s$ :

$$r_{c_j c_k} = \frac{1}{n-1} \sum_{i=1}^{n-1} \delta(s_i = c_j) \delta(s_{i+1} = c_k) \quad (2)$$

The di-codon usage of a DNA sequence  $s$  is given by a vector  $r_{cc} = (r_{c_j c_k} : j = 1, 2, \dots, 64; k = 1, 2, \dots, 64)$  of 4096 length irrespective of the length of the sequence.

### 2.4 Support Vector Machines (SVM)

The SVM offers good generalization capabilities suitable for many bioinformatics applications. An SVM is characterized by its kernel function  $K$  and connection weights  $w$ . Let  $r$  denote the di-codon pattern of the HLA sequence  $s$ . Then, the discriminant function of a binary SVM is given by:

$$f(r) = h^T(r)w + w_0 \quad (3)$$

where  $K(r, r') = h^T(r)h(r')$  and  $w_0$  is the constant weight term. In order to classify sequences into major HLA classes, a binary classifier was adopted and trained such that  $f(r) = +1$  for HLA-I molecules and  $f(r) = -1$  for HLA-II molecules.

The SVM weights and the parameters of the function  $h$  were derived through supervised learning with training data. Let  $\{(r^j, q^j) : j = 1, 2, \dots, N\}$  denote the training dataset where  $q^j$  is the desired classification for an input

di-codon pattern  $r^j$ . The SVM first transforms the input features to a higher-dimensional space by using the transform function  $h$  to maximize the separability and then linearly combines using the weight vector  $w$  to obtain the output. The optimal classification by the SVM is achieved by tuning the sensitivity parameter  $\gamma$  and the kernel parameter  $\sigma$ . For more details on SVM classifiers, the readers are kindly referred to Hastie et al. (2009).

We used multi-class SVMs proposed by Crammer and Singer (2002) to classify HLA sequences to sub-classes of HLA-I and HLA-II molecules. A multi-class classifier with  $L$  output classes was achieved using  $L$  discriminant functions  $\{f_l : l = 1, 2, \dots, L\}$  which were derived in a single optimization step (Ma et al., 2009). The output class  $\hat{l}$  for an input di-codon pattern  $r$  is assigned to the class giving the maximum discriminant function value:

$$\hat{l} = \arg \max_l f_l(r). \quad (4)$$

A three-class SVM was used to find the sub-class of HLA-I molecules: HLA-A, HLA-B, or HLA-C, given its di-codon usage pattern. For HLA-II sub-class classification, a five-class SVM was used to determine the sub-class label: HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1, or HLA-DRB3.

## 2.5 Implementation

For the classification of major HLA classes, binary SVM was implemented using LIBSVM (Chang and Lin, 2001). For sub-class classification of HLA-I and HLA-II molecules, multi-class SVM was implemented using BSVM libraries (Hsu and Lin, 2002). In order to compare with other classifiers, LDA and  $k$ -NN methods were implemented by using LNKnet package (LNKnet, 2004).

## 3 EXPERIMENTS AND RESULTS

Ten-fold cross-validation was used to evaluate the accuracies of HLA major class classification as well as HLA-I and HLA-II subclass classification. In order to avoid the selection of extremely biased partitions, the dataset was divided randomly into ten balanced partitions of equal sizes for cross-validation. In addition, we also report specificity and sensitivity as well as the standard deviation of performance measures to assess the performance of prediction schemes.

### 3.1 Parameter Estimation

The kernel type and the sensitivity parameters were decided heuristically by evaluating cross-validation performances with different types of kernels, and sensitivity and kernel parameters. For binary and multi-class SVM, the Gaussian kernel  $K(r, r') = e^{-\sigma \|r-r'\|^2}$  provided superior performance over linear and polynomial kernels for classification of HLA molecules. This was also observed earlier for gene classification using codon bias as features (Ma et al., 2009). The sensitivity parameter  $\gamma$  and the Gaussian kernel parameter  $\sigma$  were determined by using the grid-search method (Hsu and Lin, 2002). Grid-search provides reasonable estimates of parameters for multi-class SVM in a relatively short period of time.

### 3.2 Results

[Table 1 is to be included here.]

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 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 major class classification, HLA-I subclass classification, and HLA-II subclass classification were 0.01, 0.01, and 0.02, respectively, indicating only very minor effects from data partitioning. (referred in Table 2).

[Table 2 is to be included here.]

We also investigated the combination of codon and di-codon features for the classification of HLA molecules into major classes, and HLA-I and HLA-II molecules into their subclasses. A total of 4155 features including relative synonymous codon usage of 59 codons (Ma et al., 2009) 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 100% accuracy 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 divided randomly 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 can be observed, the testing and cross-validation accuracies are similar, indicating reasonable generalization.

### 3.3 Comparison with Other Classifiers

The SVM was compared to two other classifiers: linear discriminant analysis (LDA) and K-nearest neighbors ( $k$ -NN). Test and cross-validation accuracies are given in Table 2, comparing performances of the three classifiers. For major class classification of HLA molecules, the results in Table 2 demonstrate that the SVM using codon and di-codon features yielded an overall accuracy of 100% which is higher than results of  $k$ -NN (96.63%), LDA (93.32%), and classifiers using codon usage bias, SVM (99.30%),  $k$ -NN (93.95%), LDA (89.29%) and di-codon usage, SVM (99.95%),  $k$ -NN (95.17%), LDA (91.96%).

For the sub-class classification of HLA-I molecules, the accuracy of multi-class SVM using di-codon usage patterns reached 99.82% which is again the highest of the three classifiers using codon usage bias and the two classifiers using di-codon feature, LDA and  $k$ -NN. For sub-class classification of HLA-II molecules, the present method achieved the highest accuracy of 99.84% which is 10.81% and 7.78% higher than LDA method, 5.84% and 4.86% higher than  $k$ -NN, and 1.46% and 0.81% higher than SVM using codon usage bias and di-codon feature, respectively. These results show that di-codon usage pattern is an important feature for gene classification.

In addition, LDA assumes data (codon and di-codon features) as Gaussian and  $k$ -NN attempts to minimize training error, and, therefore, they offer wiggly decision boundaries, resulting in high variance to test data. Here our aim is to improve test accuracies, and our experiments with different classifiers shows that SVM generalizes well and performs better than other classifiers in classifying genes.

### 3.4 Comparison with Homology Based Methods

Moreover, the prediction accuracies of the present method outperformed classification on homology (Ma et al., 2009) for both major class and sub-class classification of HLA molecules (see Table 2). In order to compare the 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. ClustalX was used to generate a multiple sequence alignment and a distance matrix was constructed using all pairwise similarities (Thompson et al., 1997). The distance matrix has been shown previously as an effective feature for clustering or classification of aligned sequences (Grishin et al., 2002). Using this distance matrix as a set of 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 improves classification accuracy and is an effective feature for classification of HLA genes.

### 3.5 Error Analysis

We examined the set of misclassified sequences to better understand the errors generated by this approach. Sequence HLA-A\*2445N was classified incorrectly into HLA-II when using binary SVM with di-codon usage for major class classification. The length of HLA-A\*2445N protein sequence, 72aa, is very short compared to the length of its nucleotide sequence of 820nt. For sequences HLA-A\*2445N of total length 820nt, the length and location of the CDS was given as a partial coding sequence. Only this partial CDS was used in the study, which may have led to the incorrect classification. For the sub-class classification of HLA-I molecules, two sequences HLA-A\*2445N and HLA-Cw\*0507N were classified incorrectly into HLA-B by the present approach. Similar to HLA-A\*2445N, the sequence HLA-Cw\*0507N had only a partial CDS provided.

## 4 DISCUSSION AND CONCLUSION

Our study showed that codon and di-codon usage are useful features for gene classification. 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 therefore, as seen in the 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 was further explored for finer classification of the molecules. For the classification of major HLA classes and HLA subclasses, the present approach using di-codon usage patterns achieved better overall accuracies than those obtained by the classifiers using codon usage bias. Furthermore, by combining codon and di-codon features, near perfect accuracies were achieved with binary and multi-class SVM. The method is independent of the length of sequences and thus useful when homology-based methods tend to fail on datasets having genes of varying length (Ma et al., 2009).

It is evident for the mechanistic basic that base composition varies at all levels of the phylogenetic hierarchy and throughout the genome. This variation possibly makes reconstructing phylogenetic trees and inferring evolutionary processes difficult (Mooers and Holmes, 2000). Therefore, explicit phylogenetic studies of base composition are rare. Christianson proposed a simple analysis of codon usage in the parallel divergence of phytochromes in three model plants. This method can find identical bias for all family members within each taxon and increasingly divergent patterns of bias between increasingly divergent taxa (Christianson, 2005). Wang and Hickey (2007) studied codon usage patterns among rice genes and indicated that the differences in codon usage reflect a relatively rapid evolutionary increase in the nucleotide content of some rice genes. Recently, Yang and Nielsen (2008) modeled selection on codon usage for phylogenetic analysis by introducing codon-fitness and mutation-bias parameters. These models were applied to predict optimal codon frequencies for the gene as well as compare mitochondrial and nuclear genes from several mammalian species. Zhao et al. (2008) used relative synonymous codon usage (RSCU) and hierarchical clustering method to generate a cluster tree for 44 genes of 11 Human Bocavirus (HBoV) isolates.

In our study, a larger collection of 1,841 Human leukocyte antigen (HLA) gene sequences was used for gene classification. Further, investigating usage patterns of codons and di-codons with a clustering method can generate cluster trees that are useful to understand the processes governing the evolution of genes in the immune system. Although our demonstration was limited to HLA

molecules, the approach is dataset independent and can be applied to any molecular family for classification. As SVM generalizes well in the experiments, it could also help the prediction of the function of novel genes.

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 correlations between di-codon usage patterns and biological phenotypes, and finding their relations and interactions in biology could unfold valuable biological information and functions of molecules 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. Technically, incorporating SVM with patterns of tri-codons or quad-codons (higher-dimensional features), could be considered to improve the classification accuracy further. However it appears that the combination of codon and di-codon usage is sufficient for near perfect accuracy on the HLA data we have examined here.

## References

- Bhasin, M. and Raghava, G. P., 2004. SVM based method for predicting HLA-DRB1\*0401 binding peptides in an antigen sequence, *Bioinformatics*, vol 20, pp 421-423.
- Bhasin, M. and Raghava, G. P., 2004. Prediction of CTL epitopes using QM, SVM and ANN techniques, *Vaccine*, vol 22, pp 3195-3204.
- Bodmer, J. G., Marsh, S. G. E., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Charron, D., Dupont, B., Erlich, H. A., Mach, B., Mayr, W. R., Parham, P., Sasazuki, T., Schreuder, G. M. T., Strominger, J. L., Svejgaard, A., and Terasaki, P. I., 1995. Nomenclature for factors of the HLA system, 1995, *Tissue Antigens*, vol 46, pp 1-18.
- Chang, C. C. and Lin, C. J. LIBSVM : a library for support vector machines, <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Christianson, M. L., 2005. Codon usage patterns distort phylogenies from or of DNA sequences, *American Journal of Botany.*, vol 92, pp 1221-1233.
- Crammer K. and Singer, Y., 2002. On the Learnability and Design of Output Codes for Multiclass Problems, *Machine Learning*, vol 47, pp 201-233.
- Cristianini, N. and Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press.
- Donnes, P. and Elofsson, A., 2002. Prediction of MHC class I binding peptides, using SVMHC, *BMC Bioinformatics*, vol 3(1), pp 25-32.
- Duan, K. B. and Rajapakse, J. C., 2005. Multiple SVM-RFE for gene selection

- in cancer classification with expression data, *IEEE Trans Nanobioscience*, vol 4(3), pp 228-234.
- Eisen, M. B., Spellman, P. T., Brown, P. Q., and Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U.S.A.*, vol 95, pp 14863-14868.
- Galperin, M., 2004. The Molecular Biology Database Collection: 2004 update, *Nucleic Acids Res.*, vol 32, pp D2-D22.
- Grishin, V.N., Grishin, N.V., 2002. Euclidian space and grouping of biological objects, *Bioinformatics*, vol. 18, pp. 1523-1534.
- Han, H. X., Kong, F. H., and Xi, Y. Z., 2000. Progress of studies on the function of MHC in immuno-recognition, *J. Immunol. (Chinese)*, vol 16(4), pp 15-17.
- Hastie, T., Tibshirani, R., and Friedman, J., 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd Edition.
- Hori, G., Inoue, M., Nishimura, S., and Nakahara, H., 2001. Blind gene classification based on ICA of microarray data. In Lee, Jung, Makeig, and Sejnowski (eds), *ICA2001: 3rd International Conference on Independent Component Analysis and Signal Separation*, vol. 3, pp. 332-336, San Diego, California, U. S. A.
- Hsu, C. W. and Lin, C. J., 2002. A comparison on methods for multi-class support vector machines, *IEEE Transactions on Neural Networks*, vol 13, pp 415-425.
- Katz, D. H., Hamoaka, T., and 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.*, vol 137, pp 1405-1418.
- Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T., 1999. 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.
- Kim, C., Konagaya, A., and Asai K., 1999. A generic criterion for gene recognitions in genomic sequences, *Genome Inform Ser Workshop Genome Inform*, vol 10, pp 13-22.
- Lin, K., Kuang, Y., Joseph, J. S., and Kolatkar, P. R., 2002. 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.*, vol 30, pp 2599-2607.
- LNKnet software package: <http://www.ll.mit.edu/IST/lknnet/>.
- Ma, J. M., Nguyen, M. N., and Rajapakse, J. C., 2009. Gene Classification using codon usage and support vector machines, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6 (1), pp 134-143.
- Mooers, A. O. and Holmes, E. C., 2000. The evolution of base composition and phylogenetic inference, *Trends Ecol Evol.*, vol 15(9), pp 365-369.

- Noguchi, H., Park, J., and Takagi, T., 2006. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences, *Nucleic Acids Research*, vol 34(19), pp 5623-5630.
- Nguyen, M. N. and Rajapakse, J. C., 2005. Prediction of protein relative solvent accessibility with a two-stage SVM approach, *Proteins: Structure, Function, and Bioinformatics*, vol 59, pp 30-37.
- Nguyen, M. N. and Rajapakse, J. C., 2006. Two-stage support vector regression approach for predicting accessible surface areas of amino acids, *Proteins: Structure, Function, and Bioinformatics*, vol 63, pp 542-550.
- Nguyen, M. N. and Rajapakse, J. C., 2007. Prediction of protein secondary structure with two-stage multi-class SVM approach, *International Journal of Data Mining and Bioinformatics*, vol 1(3), pp 248-269.
- Nguyen, M. N., Ma, J., Fogel, G. B., and Rajapakse, J. C., 2009. Di-codon usage for gene classification, *Pattern Recognition in Bioinformatics, Lecture Notes of Computer Science*, 2009.
- Paces, J. and Paces, V., 2002. DicodonUse: the programme for dicodon bias visualization in prokaryotes, *Folia Biol (Praha)*., vol 48(6), pp 246-249.
- Rajapakse, J. C., Duan, K. B., and Yeo, W. K., 2005. Proteomic cancer classification with mass spectrometry data, *American Journal of Pharmacology*, vol 5(5), pp 281-292.
- Robinson, J., Waller, M. J., Parham, P., Bodmer, J. G., and Marsh, S. G. E., 2001. IMGT/HLA Sequence Database - a sequence database for the human major histocompatibility complex, *Nucleic Acids Res.*, vol 29, pp 210-213.
- Robinson, J., Waller, M. J., Parham, P., Groot, N. de, Bontrop, R., Kennedy, L. J., Stoehr, P., and Marsh, S. G. E., 2003. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex, *Nucleic Acids Res.*, vol 31, pp 311-314.
- Rosenthal, A. S. and Shevach, E., 1973. Function of macrophages in antigen recognition by guinea pig T lymphocytes. I. Requirement for histocompatible macrophages and lymphocytes, *J. Exp. Med.*, vol 138, pp 1194-1212.
- Sharp, P. M., Cowe, E., and Higgins, D. G., 1988. 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 Res.*, vol 16, pp 8207-8211.
- Shatsky, M., Nussinov, R., and Wolfson, H. J., 2006. Optimization of multiple-sequence alignment based on multiple-structure alignment, *Proteins: Structure, Function, and Bioinformatics*, vol 62, pp 209-217.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci. U.S.A.*, vol 96, pp 2907-2912.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res.*, vol.

- 24, pp. 4876-4882.
- Uno, R., Nakayama, Y., and Tomita, M., 2006. Over-representation of Chi sequences caused by di-codon increase in *Escherichia coli* K-12, *Gene*, vol 380(1), pp 30-37.
- Vapnik, V., 1995. *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.
- Vapnik, V., 1998. *Statistical Learning Theory*, Wiley and Sons, Inc., New York.
- Wallace, I. M., Blackshields, G., and Higgins, D. G., 2005. Multiple sequence alignments, *Curr. Opin. Struct. Biol.*, vol 15, pp 261-266.
- Wang, H. C. and Hickey, D. A., 2007. Rapid divergence of codon usage patterns within the rice genome, *BMC Evol Biol.*, vol 7, suppl 1:S6.
- Yang, Z. and Nielsen, R., 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage, *Mol Biol Evol.*, vol 25(3), pp 568-579.
- Zhang, Y. and Rajapakse, J. C. (Eds.), 2009. *Machine Learning in Bioinformatics*, John Wiley and Sons Inc.
- Zhao, Y., Pinilla, C., Valmori, D., Martin, R., and Simon, R., 2003. Application of support vector machines for T-cell epitopes prediction, *Bioinformatics*, vol 19, pp 1978-1984.
- Zhao, S., Zhang, Q., Liua, X., Wang, X., Zhang, H., Wua, Y., and Jiang, F., 2008. Analysis of synonymous codon usage in 11 Human Bocavirus isolates. *Biosystems*, vol 92(3), pp 207-214.
- Zinkernagel, R. M., and Doherty, P. C., 1974. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system, *Nature*, vol 248, pp 701-702.

Table 1

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

<b>Classification</b>	<b>Features</b>								
	<i>Codon</i>			<i>Di-codon</i>			<i>Codon + Di-codon</i>		
	<i>Acc</i>	<i>Sn</i>	<i>Sp</i>	<i>Acc</i>	<i>Sn</i>	<i>Sp</i>	<i>Acc</i>	<i>Sn</i>	<i>Sp</i>
Major Class	99.30	98.99	99.48	99.95	99.86	100	100	100	100
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	99.84	99.40	100

Preprint

Table 2

Performance comparison of the present approach with other classifiers using codon and di-codon usage on the dataset of 1841 HLA genes.

Classification	Features	Classifier	Testing Accuracy		Cross-validation Accuracy		
			mean	SD	mean	SD	
			Major class classification				
	Codon	SVM	98.72	0.01	99.30	0.01	
		LDA	88.02	0.03	89.29	0.02	
		$k$ -NN	92.30	0.02	93.95	0.02	
	Di-codon	SVM	99.13	0.01	99.95	0.01	
		LDA	90.98	0.14	91.96	0.13	
		$k$ -NN	93.70	0.10	95.17	0.08	
	Codon + Di-codon	SVM	<b>99.78</b>	<b>0.01</b>	<b>100</b>	<b>0.00</b>	
		$k$ -NN	95.76	0.08	96.63	0.06	
	Homology based method			96.14	0.04	96.65	0.04
	Sub-class classification						
	HLA-I	Codon	SVM	98.60	0.03	99.73	0.03
			LDA	87.72	0.02	88.31	0.02
			$k$ -NN	93.68	0.02	93.82	0.02
		Di-codon	SVM	99.47	0.02	99.82	0.01
			LDA	90.30	0.08	91.55	0.07
			$k$ -NN	94.31	0.06	94.66	0.05
		Codon + Di-codon	SVM	<b>99.64</b>	<b>0.01</b>	<b>99.82</b>	<b>0.01</b>
			$k$ -NN	94.57	0.05	95.82	0.04
		Homology based method			97.51	0.23	97.83
	HLA-II	Codon	SVM	97.67	0.03	98.38	0.02
			LDA	88.37	0.02	89.03	0.02
			$k$ -NN	93.02	0.01	94.00	0.01
		Di-codon	SVM	98.70	0.02	99.03	0.02
			LDA	91.57	0.04	92.06	0.05
			$k$ -NN	94.65	0.03	94.98	0.03
		Codon + Di-codon	SVM	<b>99.35</b>	<b>0.02</b>	<b>99.84</b>	<b>0.01</b>
			$k$ -NN	94.81	0.03	95.14	0.03
		Homology based method			96.27	0.24	96.74