Using Random Forest Based on Codon Usage for Predicting Human Leukocyte Antigen Gene

Panuwat Mekha
Program of Computer Science,
Faculty of Science, Maejo University
Chiang Mai, Thailand
panuwat_m@mju.ac.th

Nutnicha Teeyasuksaet
The Fifth Regional Livestock Office,
Department of Livestock Development
Chiang Mai, Thailand
nutnichavet@gmail.com

Abstract

Predicting of Human Leukocyte Antigen (HLA) gene can provide procedure into the human immune system. The classification of HLA genes has been developed by using various computational methods random forest based on codon usage. And ten-fold cross-validation to evaluate the models. Here, we propose methods of amino acid composition (AAC), dipeptide compositions (DPC) and p-collocated to investigate for major class/sub class HLA genes and to achieve high accuracy 96.24%, 98.25% and 99.25%, respectively, compared with the existing method. Finally, we shown nucleotide triplets code for a specific amino acid affect to predicting HLA gene.

Keywords: *human leukocyte antigen, computational methods, gene prediction*

I. Introduction

Human leukocyte antigen or human lymphocyte antigen (HLA) are highly polymorphic cell membrane associated glycoproteins encode in a group of genes, which resides on the short arm of chromosome 6 [1]. And The HLA is the molecule of human known as the Major Histocompatibility Complex (MHC). That is important for controlling immune function of self and non self recognition including defense against microorganisms [2]. These molecules have major function in responsive immunity by presenting antigenic peptides to T-cells [3]. Based on the structure of the antigens produced and their function, there are divided into two classes of HLA e.g., HLA-I and HLA-II [4].

Two major classes of HLA molecules have different process and functions. The principle function of HLA-I is to present viral protein, autogenetic and tumor antigen on the surface and are recognized by (CD8) T-Cell [5]. HLA-II is the helper-function for immune reaction is called "immunologically active" cells that found on antigen presenting cells (APC) such as B-cells, macrophages, dendritic cells and activated T-cells). [6].

978-1-4673-7825-3/15/$31.00 ©2015 IEEE

We focus on the problem in computation biological sequence classification of the HLA genes which can divide into two related sub-problems. The first part is classified of HLA genes into major classes. In the major class, HLA genes are divided into two classes, i.e., HLA-I, HLA-II, according to their specific functions and essential element in immune system [7] [8]. And the second part is classified subclasses of the HLA-I genes 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, HLADPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB.

Random forest (RF) is a robust classifier term of an ensemble of unpruned classification or decision trees developed by Breaiman (2001), was demonstrated to high performance better than machine learning approach others [9]. The advantages of RF classifier are less overfittng problems [10][11] and estimating feature importance. This method is appropriate for small dataset. In this paper, we employed RF to predict the Human Leukocyte Antigen Gene, based on codon usage and feature selections. In comparison with Ma et al presented a method using gene classification based on codon usage method bias as feature inputs by using each of elements correspond to the relative synonymous codon usage frequency (RSCU) of a sequence [12]. In 2009, the di-codon usage was used classify the HLA genes and compare with codon usage. This work shown that using di-codon usage as feature inputs will give outperformer than just using codon usage [13].

In this work, we developed an approach to HLA genes Classification using Random forest method based on codon usage. And we used amino acid composition (AAC), dipeptide compositions (DPC) and p-collocated to investigate for major class/sub class HLA genes. Finally, our approach developed to HLA protein classification using random forest method that predicts whether a sequence can be HLA class I or class II. Our approach was compared with other classification methods, such as k-NN, KSVM. Our experimental results show high accuracy score for major HLA classification method, which simplifier and faster than other methods.

II. MATERIALS AND METHODS

A. Data set

HLA genes data were downloaded from IMGT/HLA Sequence Database of EBI (Release3.6 14/04/2014, available at http://www.ebi.ac.uk/imgt/hla) [14]. This website provides HLA sequence for DNA sequence. The number of unfiltered dataset in Release3.6 contained 8,576 HLA-I and 2,649 HLA-II sequences is more redundant data. Thus, in this work, we filtered sequence identity of such unfiltered datasets was reduced to 30% with CD-HIT [15]. So the remaining dataset consists of 147 HLA-I and 115 HLA-II sequences, summarized in Table I.

TABLE I. Numbers of HLA genes and their subclasses.

Major class	**subclass**	**Identity <30 %**
HLA-I	HLA-A	54
	HLA-B	60
	HLA-C	33
	Total	147
HLA-II	HLA-DQA	10
	HLA-DQB	37
	HLA-DRB	68
	Total	115
All Total		262

B. Random forests (RF)

RF is trademark term for ensemble of decision trees widely used machine learning method [16]. RF is used by bagging and number of tree for random feature selection in tree induction or splitting. In bagging, sampling the original dataset on each tree get the different training based on a bootstrapping sample. To obtain a low-bias tree, RF random selecting a various parameter of features to split at each tree which useful to estimate prediction errors and correlation for feature importance. To evaluate the prediction high performance, RF shows the cross-validation prediction of model for reducing number of sequence [17].

The proposed method is an efficient and generalized method for creating many kinds of methods for predicting protein functions from all sequences or dataset. The appropriate prediction problems are the amino acid, dipeptide composition and p-collocated. Moreover, number of selected feature selection set to 59 codons from all value 64 codons. This is nucleotide triplet's code for a specific amino acid. Each amino acid has a three-letter code to present the important role in function of significantly effective features [18].

The system flowchart of the method with propensity analysis is shown in Fig. 1. The description of HLA classification consists of the following parts:

1) Get dataset is DNA sequence from the IMGT/HLA Sequence Database of EBI. Then all sequence translates DNA to protein sequences.

2) Feature vector representation by amino-acid composition (AAC), dipeptide composition (DPC) and p-collocated.

a) Amino acid Compositions (AAC)

Amino acid compositions are the simple sequence representation that used as features for prediction of various structural aspects. Given 20 *AAs* (*A,C,..., W, Y*), which ordered alphabets and denoted as AA_1, AA_2,..., AA_{19}, AA_{20} and n_i is the number of occurrences of AA_i in the sequence [19] . The amino acid composition (AAC) is represented by:

$$ACC = (\frac{n_1}{k}, \frac{n_2}{k}, \ldots\ldots, \frac{n_{19}}{k}, \frac{n_{20}}{k}) \qquad (1)$$

where *k* is the length of the protein sequence.

b) Dipeptide compositions (DPC)

In addition, dipeptide compositions are defined as a feature vector of the number of occurrence these pairs in amino acids. AA_i, AA_j, (i.e., *AA, AC,..., YW, YY*), which are denoted as AA_1AA_1, AA_1AA_2,..., $AA_{20}AA_{19}$, $AA_{20}AA_{20}$ where AA_iAA_j as $n_{i,j}$. there are 400 possible *AA* pairs. The dipeptide composition (DPC) of AA_i and AA_j is defined as:

$$DPC = (\frac{n_{1,1}}{k}, \frac{n_{1,2}}{k}, \ldots\ldots, \frac{n_{19,20}}{k}, \frac{n_{20,20}}{k}) \qquad (2)$$

where $k = n_{1,1} + n_{1,2}+, \ldots, +n_{20,20}$

c) P-collocated

The p-collocated representation considers collocated pairs of *AAs* that are separated by *p* other *AAs*. Collocated pairs for *p* = 0, 1,..., 4 are considered which are known as the dipeptides with gaps. For each value of *p*, there are 400 corresponding number of feature values. The *p*-collocated amino acid pairs predict protein and improve predictive performance for HLA dataset [20].

3) Designing random forest classifier using a value vector sampled independently number of random feature selection in tree splitting and Predicting that results that each protein sequence can be HLA class I or class II.

4) Selecting feature vector by Gini index, which use feature between AAC and DPC. And analysis of importance feature in the method.

C. Codon Usage

There are a large number of kernel methods. For example, some string kernels count the exact matching of *n* characters between the strings, others allow gaps (mismatch kernel) etc. Our work emphasized on using a spectrum kernel for classifying HLA genes.

The first kernel function defined for strings can be found [21] and Lodhi et al. who proposed a novel kernel and its application to text classification tasks by using a string subsequence kernel [22]. The string kernel is most closely related to characteristic approach. The advantage of string kernel is the frequency of n grams which the number of appearances of substrings of length n in a sequence. Leslie at el. introduced a second spectrum kernel (sequence-similar kernel) for their application to protein classification. The idea behind this spectrum kernel approach relates to the similarity of two strings based on the number of common subsequences. Saunders et al. reported on the computational advantages of the codon usage due to the method being simple and fast. If using suitable data structure, prediction can be done in linear time [23].

Given a space X of all finite length sequences of characters from an alphabet A , $|A| = l.$ The spectrum kernel is a convolution kernel specialized for the string comparison problem. For a number $k \geq 1,$ the k-codon usage is the set of sequence with k-length (k-mers) of contiguous or matching neighborhood. Such as, 3-mers = 3 codons or 6-mers = 6-codons etc. The feature map is indexed by all possible subsequences of k length from the alphabet $A.$ The feature map from X to $\mathbb{R}^{l^k}$ is defined by

$$\Phi_{\mathrm{k}}(x) = (\phi_a(x))_{a \in A^k} \qquad (3)$$

where $\phi_a(x)$ = number of times a occurs in x

Thus, the pattern of a sequence x under the feature map is a weighted representation of its k-spectrum. We can assign to the a-th coordinate a binary value a; 0 if it does not occur in x and 1 if it occurs in x. The kernel is defined by

$$K_k(x,y) = \langle \Phi_{\mathrm{k}}(\mathrm{x}), \Phi_{\mathrm{k}}(\mathrm{y}) \rangle.$$

III. RESULTS AND DISCUSSION

For the result, we show the cross-validation performance of the random forest methods to classify HLA molecules into their major classes and subclass.

In Table II, the first major row reports that each random forest methods are considered individually for AAC, DPC and P-collocated. We set the algorithm of the random forest to compare the performances.

The result for this P-collocated was 99.25 % accuracy. The predictive performances of DPC and AAC were 98.51 % and 96.24 % accuracy, respectively. Using a P-collocated resulted in the best performance. The results increase reaching 100.00 % sensitivity and 99.8 % specificity of HLA-I. For classifying HLA molecules and can build valuable features that classified HLA molecules into major classes and subclasses. Therefore, P-collocated could be the best indicator for gene expression and molecular evolution studies, as shown in Table II provides a performance feature for gene classification.

TABLE II. Prediction performance of the proposed Random forests

Methods	**Major class**			**Sub-class**	
	Acc	***Sen***	***Spec***	***HLA-I***	***HLA-II***
Random forests (AAC)	96.24	100.00	91.94	80.28	88.71
Random forests (DPC)	98.51	100.00	96.83	90.14	88.71
Random forests (p-collocated)	99.25	100.00	98.41	90.14	90.14

Several computational biology approaches have been previously developed to predict the HLA genes using SVM and codon usage and di-codon usage method appeared to have the best performance [24]. SVM method has been performed both a codon usage and di-codon usage to consider features for the HLA molecules classification. Nguyen *et al* demonstrated that using di-codon usage as feature inputs outperforms using codon usage because the di-codon frequencies are a better indicator of gene and molecular information and developed high accuracy value when used Random Forest Based on Codon Usage [25]**.**

The accuracy for the HLA gene prediction and other classifier are shown in the Table III. The proposed HLA gene perdition, which are random forest method used by number parameter equal 100 parameter (n=100), provides the best accuracy equals to SVM uses kernel function with polynomial n=2 and polynomial n=4 that equals 99.25%. The other methods provide lower accuracies. Thus, results of RF method give the best edge over k-NN and k-NN based on spectrum kernel in HLA major class accuracy and highest accuracy for HLA sub-class.

TABLE III. Prediction performance of the proposed method and the other classifiers

Methods	Kernel function/ parameter	Major class	Sub-class	
			HLA-I	*HLA-II*
k-NN	k=5	91.73	78.87	85.48
SVM	linear	98.51	84.51	88.71
	poly n=2	97.74	77.46	88.71
	poly n=3	99.25	84.51	88.71
	poly n=4	99.25	88.73	88.71
	RBF	96.24	63.38	85.48
k-NN/sk	k=3 (Codon) /sk=6 (di-codon)	95.49	83.10	88.71
RF+Codon	n=100	99.25	90.14	90.14

A. *Feature important of HLA using Random Forests based on codon usage*

Feature selection is able to choose important attribute from many relevant and irrelevant features. Moreover, it can reveal the relationship and correlation between features selection and predictions by an explicit model. Feature importance measure for random forest based on codon usage which is contribution to the prediction performance. It can be provide to increase mean of accuracy for classification that is calculated in the course of training [26]. In our RF method for predicting HLA based on codon usage. The best codon of the 20 top ranked informative features is shown in Fig. 1

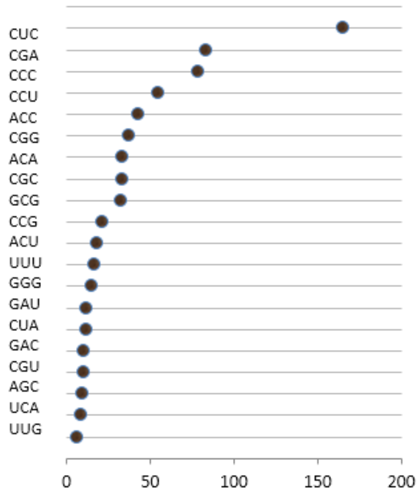

Fig. 1, Feature importance of codon usage

B. *Feature important of Composition Feature of HLA*

In this work, we analyse the feature importance for each class of sequence. It can provide better understanding of HLA property sequences. The efficiency feature importance estimator of the random forests method is applied to indicate informative features in the dataset for each feature type. Moreover, the mean decrease of Gini index (MDGI) is used for measuring prediction accuracy that can be available for ranking feature importance based on measure approaches [27]. Thus, we used the largest value MDGI with AAC for assigning to rank feature importance. So, the prediction result show superiority of the 20 top ranked informative features in amino acids are distinguishing for HLA.

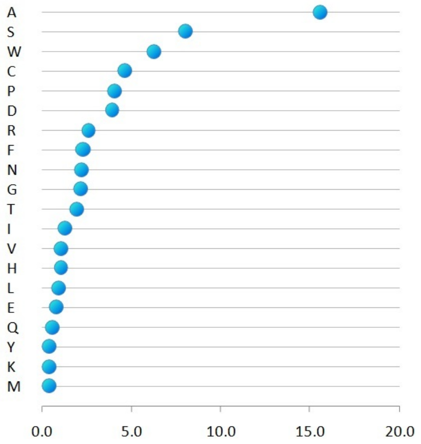

Fig. 2, Feature importance of Amino Acid Composition (AAC)

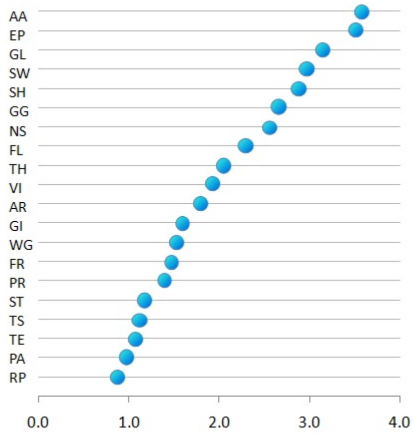

Fig. 3, Feature importance of Dipeptide Composition (DPC)

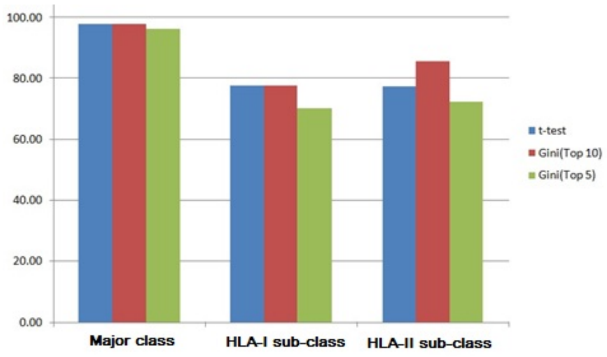

Fig. 4, Performances of comparison with Gini index

The result performances of comparison with Gini index different classification methods are presented in Fig 4. When using HLA molecules with t-test, 10-Top Gini (red) and 5-Top Gini (green) we achieve the best results of nearly 100.00% of major class which is better than any of the other subclass of HLA-I and HLA-II . For this study, we found that using this method by applied with Gini index can get the best results.

IV. CONCLUSION

We have shown that the random forest method can be applied to the HLA molecules classification problem based on DNA sequences directly. Table II summarizes the performance comparison of the classified the HLA molecules into HLA-I and HLA-II with using AAC, DPC and p-collocated. Initially, we used the p-collocated to classify the HLA molecules, because p-collocated is considers collocated pairs of DNA sequences that are separated by p other. In the same way, codon corresponds to a fundamental unit of molecular information. We increase codon sequence to the six length size. This choice proved to be successful, obtaining the optimal performance. Furthermore, we compare random forest with KSVMs and NN. Table III shows that when using random forest based on codon usage with the actual DNA sequences of HLA molecules, it leads to improved classification performance.

ACKNOWLEDGMENT

We thank Prof.R. Cutler, Dr. Watshara Shoombuatong and Assoc.Prof.Dr. Jeerayut Chaijaruwanich for comments and suggestion on this manuscript. Moreover, this project was supported by Faculty of science, Maejo University and Faculty of science, Chiang Mai University and Mahidol University.

REFERENCES

[1] T.A. Biokowski, S.R. Matino and A. Joachimiak, "Predicting HLA Class I Non-Premissive Amino Acid Residues Substitutions", Volume7, Issue8, e41710, August 2010.

[2] H. Hugh , F. JRL, A.C. Wang and G.B. Ferrara, "Basic Immunogenetics", 3rd Ed. Oxford: Oxford University Press, 1984.

[3] A.M. Lttile and P. Phrham, "Polymorphism and evolution of HLA class I and II genes and molecules", Rev Immunogenet, 1999, 1, pp. 105-123.

[4] U. Shankarkumar , "The Human Leukocyte Antigent (HLA) system", *Int J Hum Genet.* 2004, 4(2), pp. 91-103.

[5] S. Nail, "The human HLA system", J Indian Rheumatology Assoc. 11, 2003, pp. 79-83.

[6] M. Browning and A. McMichael, "HLA and MHC: Genes, Molecules and Function", Bios Scientific Publishers, Oxford, 1996.

[7] D.H. Katz, T. Hamoaka and B. Benacerraf, "Cell interactions between Histocompatible T and B Lymphocytes. Failure of Physiologic Cooperation Interactions between T and B Lymphocytes from Allogeneic Donor Strains in Humoral Response to Hapten-Protein", Conjugates. J. Experimental Medicine, 1973, pp. 137-141.

[8] H.X. Han, F.H. Kong and Y.Z. Xi, "Progress of Studies on the Function of MHC in Immuno-Recognition", J. Immunology (Chinese), 2000, 16 (4), pp. 15-17.

[9] L. Breiman, "Random forests", Machine Learning, vol. 45, pp. 5-32, Oct 2001.

[10] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling", J. Chem. Inf. Comput. Sci.43, 2003, pp. 1947-1958.

[11] J.L. Milhon and J.W. Tracy, "Updated codon usage in Schistosoma", Exp Parasitol, 1995, 80, pp.353-356.

[12] M.N. Nguyen, J.M. Ma, G.B. Fogel and J.C. Rajapakse, "Di-codon Usage for Gene Classification", The 4th IAPR International Conference on Pattern Recognition in Bioinformatics, 2009, 5780, pp. 11–221.

[13] J. Robinson , A. Malik, P. Parham, J.G. Bodmer and SGE. Marsh, "IMGT/HLA and IMGT/MHC: sequence databases for the human major histocompatibility complex", Nucleic Antigens. Vol. 55, 2001, pp.280-287.

[14] L. Weizhong and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences", Bioinformatics, Vol. 22, 2006, pp. 1658-1659.

[15] M.L. Calle and V. Urrea, "Letter to the editor: Stability of Random Forest importance measures", Brief-bioinformatic, 2011, 12, pp. 86–89.

[16] D. Amaratunga, J. Cabrera, and Y. S. Lee, "Enriched random forests," Bioinformatics, vol. 24, pp. 2010-4, Sep 15 2008.

[17] J.M. Ma, M.N. Nguyen and J.C. Rajapakse, "Gene classification using codon usage and support vector machines", IEEE ACM Trans Comput Biol Bioinformatics, 2009, pp.134–43.

[18] N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao, "Information assessment on predicting protein-protein interactions", BMC Bioinformatics, vol. 5, p. 154, Oct 18 2004.

[19] W.R. Pearson and D.J. Lipman, "Improved tools for biological sequence comparison", Proc. Natl Acad. Sci. USA, 1988, 85, pp. 2444–2448.

[20] L. Liao and W.S. Noble, "Combining Pairwise Sequence Similarity and SupportVector Machines for Detecting Remote Protein Evolutionary and Structural Relationships", J Comp Biol, 2003, 10, 857–868.

[21] C. Watkins, "Kernel from matching operation", Tech. rep. Department of Computer Science, Royal Holloway, University of London,1990.

[22] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins, "Text classification using string kernels", Journal of Machine Learning Research, 2002, Vol. 2, pp. 419–444.

[23] C. Saunders, H. Tschach and J.S. Taylor, "Syllables and other string kernel extensions", ICML, 2002, Vol. 19, pp. 530-537.

[24] M. Mitreva, M.C. Wendl, J. Martin, T. Wylie, Y. Yin, L. Larson, J. Parkinson, R.H. Waterston and J.P. McCarter, "Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species", Genome Biology, 2006, 7: R75.

[25] C. Chen, Y.X. Tian, X.Y. Zou, P.X. Cai and J.Y. Mo, "Using pseudo-amino acid composition and support vector machine to predictprotein structural class", Journal of Theoretical Biology 243 (3), 2006, pp. 444–448.

[26] K.D. Kedarisetti, L. Kurgan and S. Dick, "Classifier ensembles for protein structural class prediction with varying homology", Bio-chemical and Biophysical Research Communications 348, 2006, pp. 981–988.

[27] L. Kurgan and L. Homaeian, "Prediction of structural classes for protein sequences and domains impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, Pattern Recognition", special issue on Bioinformatics 39 (12), 2006, pp. 2323–2343.