

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	iii
ABSTRACT (ENGLISH)	v
ABSTRACT (THAI)	vii
LIST OF TABLES	xiii
LIST OF FIGURES	xv
ABBREVIATIONS	xvi
CHAPTER I INTRODUCTION	
1.1 Bioinformatics	1
1.2 The challenge in protein function prediction	1
1.3 Contributions	2
1.3 Thesis outlines	4
CHAPTER II BACKGROUND	
2.1 Biological background	6
2.1.1 Human Leukocyte Antigen Gene	6
2.1.2 Human Immunodeficiency Virus type 1	8
2.1.3 Protein crystallization	8
2.2 Sequence classification	9
2.2.1 Feature based classification	10
2.2.2 Sequence distance based classification	11
2.2.3 Model based classification	12
2.2.4 Support vector machine	12

2.3 Feature selection methods	14
2.3.1 Filter techniques	16
2.3.2 Wrapper techniques	17
2.3.3 Embedded techniques	17
2.4 Support Vector Machine	18
3.2.1 Notation	18
3.2.2 Functional and geometric margins	19
3.2.3 The optimal margin classifier	19
3.2.4 Kernel function	20
2.5 Scoring Card Method	21
2.5.1 Creation of dataset	22
2.5.2 Constructing an initial scoring matrix	22
2.5.3 Deriving the propensity score of each amino acid	22
2.5.4 Optimizing a scoring matrix	23
2.5.5 Predicting protein crystallization	24
2.6 Logistic Model Tree	25
2.6.1 The model	25
2.6.2 Building logistic model tree	27
2.6.3 Splitting criterion	29
2.6.4 Stopping criterion	30
2.6.5 Pruning the tree	31

**CHAPTER III PREDICTION OF HUMAN LEUKOCYTE ANTIGEN GENE
USING k -NEAREST NEIGHBOR CLASSIFIER BASED ON SPECTRUM
KERNEL**

3.1 Introduction	32
3.2 Dataset	33
3.3 Sequence classification method based on k -NN classifier and spectrum kernel	35
3.4 Results and discussion	36
3.4.1 Description of the experiments	36
3.4.2 Results and discussion of the combination method	38
3.4.3 Comparison of the HLA major class classification with other classification methods	41
3.4.4 Comparison of the HLA-I/HLA-II subclasses classification with other classification methods	42
3.4.5 Sequence classification on human genes	43
3.4.6 Discussion	44
3.5 Conclusion	46

**CHAPTER IV HIV-1 CRF01_AE CORECEPTOR USAGE PREDICTION
USING KERNEL METHODS BASED LOGISTIC MODEL TREES**

4.1 Introduction	48
4.2 Dataset	50
4.3 Computational methods for HIV-1 CRF01_AE coreceptor usage prediction	51

4.3.1 Implementation using kernel methods based LMT	51
4.4 Description of the experiments	53
4.5 Results and discussion	54
4.5.1 Results and discussion of kernel methods	54
4.5.2 Comparison of the kernel methods based LMT to other genotypic predictors and computational classification methods	56
4.5.3 Discussion	59
4.6 Conclusion	60

CHAPTER V PREDICTING PROTEIN CRYSTALLIZATION USING A SIMPLE SCORING CARD METHOD

5.1 Introduction	62
5.2 Dataset	65
5.3 Results and discussion	66
5.3.1 Prediction performance of the scoring card method based on collocated dipeptides	66
5.3.2 Performance of SVM using amino acid and collocated dipeptide compositions	69
5.3.3 Comparison of the scoring card method with existing classifiers	71
5.3.4 Analyzing dipeptide compositions	72
5.3.5 Application of propensity score of amino acids for the crystallization of protein	75

5.4 Conclusion	77
CHAPTER VI CONCLUSION	78
REFERENCES	80
APPENDIX	93
APPENDIX A PUBLICATIONS BY AUTHOR	94
CURRICULUM VITAE	97

LIST OF TABLES

Table	Page
2.1 A taxonomy of feature selection techniques	15
2.2 The characteristics of feature selection techniques	16
3.1 Existing methods for predicting HLA genes	33
3.2 Numbers of HLA genes and their subclasses (%)	34
3.3 Performances of the classification of HLA genes by using the combination method with different k lengths and k nearest neighbors.	39
3.4 The ten-fold cross validation accuracies for the classification of HLA-I and HLA-II into their subclasses using the combined method.	40
3.5 The ten-fold cross validation accuracies of the classification of HLA major classes using different classification methods.	42
3.6 Ten-fold cross validation accuracies of the HLA-I/HLA-II subclasses classification using different classification methods.	43
3.7 Numbers of removed high sequence identity of HLA genes	45
3.8 The ten-fold cross validation accuracies of HLA major classes classification using different classification methods (using sequence identity < 30%).	46
3.9 The ten-fold cross validation accuracies of HLA subclasses classification using different classification methods (using sequence identity < 30%).	46
4.1 Some existing methods for predicting HIV-1 coreceptor usage	49

4.2	Predictive performances of SVM based different features with various kernel functions	55
4.3	Comparison of predictive performances from different features when using SVM based radial basis function kernel	56
4.4	Performances of Kernel method vs. the other genotypic predictors.	58
4.5	Summary of removed high sequence identity of HIV-1 coreceptor usage	59
5.1	Some existing methods for predicting protein crystallization from sequences.	59
5.2	Summary of the dataset for evaluating the performance of protein crystallizability predictors.	64
5.3	Performances of the scoring card method based on initial CSM with various gaps	65
5.4	Performances of the scoring card method based on optimized CSM with zero gaps.	68
5.5	Performances of the scoring card method based on combined CSM with various gaps.	69
5.6	Performances SVM using various compositions.	70
5.7	Comparison of the scoring card method with existing methods.	72
5.8	The propensity scores of amino acids to be crystallizable	74
5.9	The 10-top of highest and lowest propensity scores of dipeptides to be crystallizable	74

LIST OF FIGURES

Figure	Page
2.1 The HLA region of Chromosome 6	7
2.2 The V3 loop region	8
2.3 The system flowchart of filter techniques	16
2.4 The system flowchart of wrapper techniques	17
2.5 The system flowchart of embeddble techniques	18
2.6 The system flowchart of the scoring card method (SCM)	24
2.7 LogitBoost algorithm	28
2.8 Building logistic models	29
3.1 The system flowchart of the k -NN classifier based on spectrum kernel to classify HLA gene.	39
4.1 The system flowchart of the SVM classifier based LMT method to predict HLA-1 CRF_AE coreceptor usage.	52
5.1 The ROC curve of the scoring card method by using various gaps	69
5.2 The propensity scores of dipeptides from averaging results of 10 independent runs	73

ABBREVIATIONS

<i>k</i>-NN	<i>k</i> -Nearest Neighbor
SVMs	Support vector machines
LMT	Logistic model tree
SCM	Scoring Card Method
HLA	Human Leukocyte Antigen
HIV-1	Human Immunodeficiency Virus type 1
MHC	Human major Histocompatibility Complex
CCR5	CC-chemokine receptor 5
CXCR4	CXC-chemokine receptor 4
PDB	Protein Data Bank
PCP	Physicochemical properties
FS	Feature subset
CSM	Crystallizable scoring matrix
IGA	Intelligent genetic algorithm
R value	Pearson's correlation coefficient
ROC	Receiver operating characteristic
LDA	Linear discriminant analysis
KSVMs	Kernel support vector machines
V3	Third variable loop