

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
ABSTRACT (ENGLISH)	v
ABSTRACT (THAI)	vii
LIST OF TABLES	xii
LIST OF FIGURES	xiv
CHAPTER 1 INTRODUCTION	
1.1 Principle and rationale of research topic	1
1.2 Objectives and outline of this thesis	3
CHAPTER 2 LITERATURE REVIEW	5
2.1 General information of Northern Thailand	5
2.2 Ethnic groups in Northern Thailand	6
2.3 Single nucleotide polymorphisms	13
2.3.1 SNP discovery	15
2.3.2 Pattern of human SNP variation	18
2.4 Variable selection	20
2.4.1 Introduction to variable and feature selection	20
2.4.2 Variable ranking	21
2.4.3 Information theoretic ranking criteria	21
2.5 Data classification and decision tree	23
2.5.1 Basic in data classification	23

2.5.2	Principle of decision tree	26
2.5.3	Evaluation measurement using confusion matrix	27
2.6	Categorical data and correspondence analysis	29
2.6.1	Basic concept in categorical data analysis	29
2.6.2	Correspondence analysis	30
2.7	Population genetic distance	33
2.7.1	Principle of genetic distance	34
2.7.2	Estimating genetic distance	36
2.8	Related work	37
2.8.1	Study related to the genetic study of northern Thai population.	37
2.8.2	The study of population genetic using SNPs	39
CHAPTER 3 RESEARCH METHODOLOGY		42
3.1	Studied population	42
3.2	SNP data	44
3.3	Computational methods	46
3.3.1	Mutual information with SNP	46
3.3.2	Decision tree	49
3.3.2.1	Problems description	49
3.3.2.2	SNP data set in decision tree	49
3.3.2.3	Population classification using R program	50
3.3.2.4	Decision tree definition	51
3.3.2.5	Decision tree construction	51
3.3.2.6	Divide and conquer algorithms	53

3.3.2.7 Candidate test	54
3.3.2.8 Selecting test	54
3.3.3.9 Performance of evaluation measure	55
3.3.3 Correspondence analysis	57
3.3.3.1 Contingency table in this research	58
3.3.3.2 Correspondence analysis using R program	58
3.3.3.3 Visualization of correspondence analysis result	63
3.3.4 Population genetic distance and relationship visualization	64
3.3.4.1 Genetic distance using PEAS program	64
3.3.4.2 Phylogenetic tree and multidimensional scaling	64
CHAPTER 4 RESULTS AND DISCUSSION	69
4.1 Specific SNP selection	69
4.1.1 Mutual information technique	69
4.1.2 Population classification using decision tree	70
4.1.3 SNP with correspondence analysis	74
4.2 Population genetic distances using SNP data	76
CHAPTER 5 CONCLUSION	82
REFERENCES	83
APPENDICES	93
APPENDIX A Top 60 SNP ranking list and information	94
APPENDIX B Decision tree illustration	115
APPENDIX C Correspondence analysis in R language	121
APPENDIX D Genetic distance calculation and visualization	123
CURRICULUM VITAE	130

LIST OF TABLES

Table	Page
2.1 Example of summary data from Affymetrix raw file	19
2.2 Confusion matrix example	28
2.3 Example of a two-way table	30
2.4 Table of science doctorate in USA	32
3.1 Description of samples in 13 ethnic groups	43
3.2 Affymetrix export file example	44
3.3 Example of the SNP information from Affymetrix SNParray	45
3.4 Genotype frequency in any locus.	47
3.5 Available SNP dataset for decision tree analysis	50
3.6 Testing data example for decision tree analysis	56
3.7 Calculation of accuracy	56
3.8 Example of contingency of SNP genotype	58
3.9 Matrix of rows profile	60
3.10 Matrix of columns profile	61
3.11 Distance matrix example	67
4.1 Mutual information value for the 13 ethnic groups	70
4.2 Number of SNP loci at mutual information value ranges	71
4.3 Classification accuracy for different SNP number from ranking list	73
4.4 Confusion matrix for genotype data of 60 SNP loci	74

4.5	Cavalli-Sforza and Nei's standard genetic distance	78
A.1	Top 60 SNP ranking list	94
B.1	Example data using in Rweka package	115
B.2	Confusion matrix from R program	116
C.1	Example of SNP genotype frequency using in correspondence analysis	121
C.2	Result of correspondence analysis from R program	122
D.1	Input file example using in PEAS software	123
D.2	Input file example using in MEGA5	127

LIST OF FIGURES

Figure		Page
2.1	Northern Thailand map	6
2.2	Single nucleotide polymorphisms in an individual	13
2.3	Single nucleotide polymorphisms in population	14
2.4	Sample preparation and array processing	16
2.5	Overview of SNP array technology	18
2.6	Distribution of human variation within and between populations.	20
2.7	Illustration of classification task	24
2.8	Simple decision tree	28
2.9	Correspondence analysis of doctorate data	33
3.1	Diagram of research design	46
3.2	Example of information gain calculation	52
4.1	Mutual information value distribution of all SNP	72
4.2	Two dimensional visualizing correspondence analysis	75
4.3	Unrooted neighbor-joining tree based on Nei's standard distance	79
4.4	Unrooted neighbor-joining tree based on Cavalli-Sforza and Edward distance	79
4.5	Two-dimensional from classical multidimensional scaling based on Nei's standard distance	80

4.6	Two-dimensional from classical multidimensional scaling based on Cavalli-Sforza and Edward distance	81
B.1	Decision tree summary from Rweka package.	116
B.2	Decision tree visualization from Graphviz	117
D.1	PEAS component program	124
D.2	Procedure for PEAS program running	125
D.3	Phylogenetic tree construction in MEGA5	126
D.4	Relationship visualization and two-dimension graph	129