

ห้องสมุดงานวิจัย สำนักงานคณะกรรมการวิจัยแห่งชาติ



E46292



DEVELOPMENT OF A WEB-BASED TOOL FOR VISUALIZING AND
PROFILING HUMAN ENDOGENOUS RETROVIRUSES (HERVS)
NEIGHBORING GENES

MR. THITIPONG KAWICHAI

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
(BIOINFORMATICS AND SYSTEMS BIOLOGY)
SCHOOL OF BIORESOURCES AND TECHNOLOGY AND
SCHOOL OF INFORMATION TECHNOLOGY
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY THONBURI

2011



Development of a Web-based Tool for Visualizing and Profiling
Human Endogenous Retroviruses (HERVs) Neighboring Genes

Mr. Thitipong Kawichai B.Sc. (Mathematics)

A Thesis Submitted in Partial Fulfillment of the Requirements for
the Degree of Master of Science (Bioinformatics and Systems Biology)
School of Bioresources and Technology and School of Information Technology
King Mongkut's University of Technology Thonburi

2011

Thesis Committee




.....
(Asst. Prof. Jonathan Hoyin Chan, Ph.D.)

Chairman of Thesis Committee


.....
(Lecturer, Santitham Prom-on, Ph.D.)

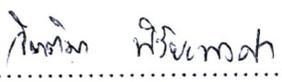
Member and Thesis Advisor


.....
(Asst. Prof. Asawin Meechai, Ph.D.)

Member and Thesis Co-advisor


.....
(Asst. Prof. Kanokwan Poomputsa, Ph.D.)

Member


.....
(Researcher, Jittima Piriyaongsa, Ph.D.)

Member

PREFACE

This thesis is a part of the requirements to accomplish my master degree in Bioinformatics and Systems Biology, King Mongkut's University of Technology Thonburi. The topic of the study is "Development of a web-based tool for visualizing and profiling human endogenous retroviruses (HERVs) neighboring genes" or "การพัฒนาเครื่องมือบนเว็บสำหรับการแสดงผลและวิเคราะห์เอนโดจีนัสรีโทรไวรัสที่อยู่รอบยีน" in Thai. This work has been done with an aim to develop a novel bioinformatics tool, so-called HERV Profiler, for facilitating a study of human endogenous retroviruses or HERVs neighboring human genes. Furthermore, the case studies illustrating the utilization of HERV Profiler were also included in this work.

The thesis report is composed of five chapters. The first chapter is about background and rationale, objectives, scope of works, and expected outputs. The second chapter is related to backgrounds and literature reviews. For the third chapter, it is about materials and research methods. The fourth chapter is the results and discussions. Lastly, the conclusions and suggestions would be written in the fifth chapter.

Thesis Title	Development of a web-based tool for visualizing and profiling human endogenous retroviruses (HERVs) neighboring genes
Thesis Credits	12
Candidate	Mr. Thitipong Kawichai
Advisor	Dr. Santitham Prom-on
Co-advisor	Asst. Prof. Dr. Asawin Meechai
Program	Master of Science
Field of Study	Bioinformatics and Systems Biology
Faculty	School of Bioresources and Technology and School of Information Technology
B.E.	2554

Abstract

E46292

Human endogenous retroviruses (HERVs) are the remnants of ancient retroviral infections dispersed throughout the human genome. In particular cases, HERV regulatory sequences still remain active, and this allows them to influence the transcriptions of neighboring genes. Although HERV resources have been available for more than a decade, there is still not a database or tool that really supports the studies on HERVs nearby genes. To address this concern, this thesis aims to develop a web-based tool, so-called HERV Profiler, for facilitating investigations on neighboring HERVs of human genes. Several HERV characteristics were compiled in the HERV Profiler database, such as location relative to genes, family, superfamily, type of truncation patterns, orientation, and intactness ratio. HERV Profiler provides two main features, including visualizing and profiling the neighboring HERVs. In the visualizing, there are two different views to display the neighboring HERVs, including displaying in tabular and graphical format. In the HERV profiling, users can construct HERV profiles from the input gene list, find over-represented HERV types among the gene list, and rank input genes based on their over-represented HERVs. In the case studies, HERV Profiler can discover potential HERV types and genes under the condition of Systemic Lupus Erythematosus (SLE). In conclusion, HERV Profiler is a powerful tool that can efficiently facilitate HERV studies. Currently, HERV Profiler is freely accessible at http://sigma.cpe.kmutt.ac.th/herv_profiler.

Keywords: HERVs / neighboring genes / web-based tool / HERV profiling

หัวข้อวิทยานิพนธ์	การพัฒนาเครื่องมือบนเว็บสำหรับการแสดงผลและวิเคราะห์เอนโดจีนรีโทรไวรัสที่อยู่รอบยีน
หน่วยกิต	12
ผู้เขียน	นายจิตติพงษ์ กาวิชัย
อาจารย์ที่ปรึกษา	ดร.สันติธรรม พรหมอ่อน
อาจารย์ที่ปรึกษาร่วม	ผศ.ดร.อัศวิน มีชัย
หลักสูตร	วิทยาศาสตร์มหาบัณฑิต
สาขาวิชา	ชีวสารสนเทศและชีววิทยาระบบ
คณะ	ทรัพยากรชีวภาพและเทคโนโลยี และเทคโนโลยีสารสนเทศ
พ.ศ.	2554

บทคัดย่อ

E46292

เอนโดจีนรีโทรไวรัสในมนุษย์หรือเอชอีอาร์วี (Human Endogenous Retroviruses or HERVs) เป็นลำดับดีเอ็นเอ (DNA sequences) ที่หลงเหลือจากการติดเชื้อของไวรัสในอดีต ซึ่งสามารถพบได้ทั่วไปในจีโนมของมนุษย์ ในบางกรณีลำดับดีเอ็นเอของตัวควบคุมการแสดงออกของเอชอีอาร์วียังคงทำงานได้อยู่อย่างมีประสิทธิภาพ ส่งผลให้เอชอีอาร์วีสามารถส่งผลกระทบต่อการแสดงออกของยีนที่อยู่บริเวณข้างเคียงได้ ถึงแม้ว่าข้อมูลที่เกี่ยวข้องกับเอชอีอาร์วีจะมีอยู่มาเป็นระยะเวลานานกว่าหนึ่งทศวรรษแล้วก็ตาม แต่ก็ยังไม่มีฐานข้อมูลหรือเครื่องมือใดที่ช่วยในการศึกษาเกี่ยวกับเอชอีอาร์วีที่อยู่ใกล้ยีนได้อย่างแท้จริง ดังนั้นวิทยานิพนธ์นี้จึงมีวัตถุประสงค์เพื่อพัฒนาเครื่องมือบนเว็บ หรือเอชอีอาร์วีโปรไฟล์เลอร์ (HERV Profiler) เพื่อช่วยอำนวยความสะดวกในการศึกษาเอชอีอาร์วีที่อยู่ข้างเคียงยีนในมนุษย์ เอชอีอาร์วีโปรไฟล์เลอร์ได้เก็บรวบรวมลักษณะของเอชอีอาร์วีไว้มากมาย ได้แก่ บริเวณของยีนที่เอชอีอาร์วีแทรกตัวอยู่ แฟมิลี (family) ซุปเปอร์แฟมิลี (superfamily) รูปแบบการขาดหายของส่วนประกอบของเอชอีอาร์วี (type of truncation patterns) และสัดส่วนความสมบูรณ์ของเอชอีอาร์วี (intactness ratio) การทำงานของเอชอีอาร์วีโปรไฟล์เลอร์สามารถแบ่งออกได้เป็นสองส่วนหลักๆ คือ การแสดงผลและการทำโปรไฟล์ลิงของเอชอีอาร์วีที่อยู่ใกล้ยีน ผู้ใช้สามารถแสดงผลของเอชอีอาร์วีได้สองรูปแบบ คือ ในรูปแบบของตารางและรูปภาพ การทำโปรไฟล์ลิงจะช่วยให้ผู้ใช้สามารถสร้างโปรไฟล์ หาชนิดของเอชอีอาร์วีที่ปรากฏอย่างเด่นชัดในชุดของยีน และเรียงลำดับยีนตามความสำคัญของเอชอีอาร์วีที่อยู่ข้างเคียงยีนนั้นได้ จากกรณีศึกษาของโรคเอสแอลอี (SLE or Systemic Lupus Erythematosus) พบว่าเครื่องมือนี้สามารถใช้วิเคราะห์หาชนิดของเอชอีอาร์วีและยีนที่น่าจะเกี่ยวข้องกันได้ ดังนั้นเครื่องมือนี้ นับว่าเป็นเครื่องมือที่มีประสิทธิภาพอย่างยิ่งในการช่วยอำนวยความสะดวกแก่ผู้ใช้ที่ต้องการศึกษาเกี่ยวกับเอชอีอาร์วีที่อยู่ข้างเคียงยีนในมนุษย์ เอชอีอาร์วีโปรไฟล์เลอร์เปิดให้บริการผ่านทางเว็บไซต์ http://sigma.cpe.kmutt.ac.th/herv_profiler

คำสำคัญ: เอชอีอาร์วี / ยีนที่อยู่ข้างเคียง / เครื่องมือบนเว็บ / เอชอีอาร์วีโปรไฟล์ลิง

ACKNOWLEDGEMENTS

At first, I would like to express my sincere gratitude and deep appreciation to my advisor, Dr. Santitham Prom-on, for his kindness, invaluable guidance, and encouragement. He gave assistances and supported me everything along the time that the thesis has been done. Also, I would like to thank my co-advisor, Asst. Prof. Dr. Asawin Meechai, who gave me a lot of advices and useful suggestions. Frequently, he also assisted me to correct and improve my works. Furthermore, I would like to acknowledge my committees, Asst. Prof. Dr. Jonathan Hoyin Chan, Asst. Prof. Dr. Kanokwan Poomputsa, and Dr. Jittima Piriyaongsa for their valuable suggestions and comments, which were very useful in improving the works.

In addition, I would like to thank Mr. Pumipat Tongyoo for his suggestions, comments, and assistances. Those suggestions led me to solve several problems during the course of the thesis. Furthermore, I would like to thank my mother for her encouragement and taking care of me. Although there is no any technical suggestion from her, only the cheering words could inspire me to be able to overcome all of the problems occurred at that time. An important one in my family that I would like to specially acknowledge here is my father. Even though he is not here anymore, I believe that he is always beside me and give me his assistances until I have reached my goal now.

I am grateful to Bioinformatics and Systems Biology, King Mongkut's University of Technology Thonburi and National Center for Genetic Engineering and Biotechnology (BIOTEC) for supporting on scholarships and resources that facilitate me until the thesis have been completed. In addition, I would like to thank Development and Promotion of Science and Technology Talents Project (DPST) for financial support as well as a great opportunity to study and research.

Lastly, I would like to thank all of my friends here who have encouraged and supported me along the way until I finally reached my destination now.

CONTENTS

	PAGE
PREFACE	i
ENGLISH ABSTRACT	ii
THAI ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
CONTENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	x
LIST OF TECHNICAL VOCABULARIES AND ABBREVIATIONS	xiv
CHAPTER	
1. INTRODUCTION	1
1.1 Background and rationale	1
1.2 Objectives	2
1.3 Scope of work	2
1.4 Expected outputs	2
2. BACKGROUNDS AND LITURATURE REVIEWS	3
2.1 Human endogenous retroviruses (HERVs)	3
2.1.1 What is HERV	3
2.1.2 Genomic structure of HERVs	4
2.1.3 Classification of HERVs	5
2.1.4 Biological functions of HERVs	6
2.2 Evidences linking HERVs to SLE	9
2.3 Databases and tools related to HERVs	11
2.3.1 Repbase Update (RU)	11
2.3.2 HERVd	12
2.3.3 RetroSearch	12
2.3.4 TranspoGene	13
2.3.5 Tools for detection of HERVs	14
2.4 Fisher's exact test	14

3. MATERIALS AND METHODOLOGY	16
3.1 Hardware and software	16
3.1.1 REannotate	16
3.1.2 Python programming language	16
3.1.3 MySQL database	16
3.1.4 HTML	16
3.1.5 PHP script language	16
3.1.6 Apache web server	17
3.1.7 R programming language	17
3.2 Data resources	17
3.2.1 Human repeat annotation data	17
3.2.2 Human gene annotation data	17
3.2.3 Cross-reference gene ID data	17
3.2.4 SLE microarray data (GSE20864)	18
3.3 System design and requirements	18
3.3.1 System flow design	18
3.3.2 Functional requirements	19
3.3.3 Overview of system development processes	23
3.4 Database construction	24
3.4.1 Data collection	24
3.4.2 Data preparation	25
3.4.3 Database design and implementation	32
3.5 Tool implementation and tool testing	33
3.6 Case studies related to SLE	36
3.6.1 Case study 1: Significance of the number of HERVs under the SLE condition	36
3.6.2 Case study 2: Significance of HERV characteristics under the SLE condition	37
4. RESULTS AND DISCUSSIONS	39
4.1 Consideration of the maximum distance parameter	39
4.2 Web interface	42
4.2.1 Visualizing neighboring HERVs of genes in a gene list	45
4.2.2 Profiling neighboring HERVs of genes in a gene list	48

4.3 Tool testing	54
4.3.1 Testing of sub-module 2.2	54
4.3.2 Testing of sub-module 2.3	55
4.4 Case studies related to SLE	56
4.4.1 Case study 1: Significance of the number of HERVs under the SLE condition	56
4.4.2 Case study 2: Significance of HERV characteristics under the SLE condition	58
5. CONCLUSIONS AND SUGGESTIONS	63
5.1 Conclusions	63
5.2 Suggestions	63
REFERENCES	64
APPENDIX	
A. List of HERV families and names/groups	69
B. Data dictionary of HERV Profiler database	74
C. HERV Profiler User's Guide	77
CURRICULUM VITAE	108

LIST OF TABLES

TABLE		PAGE
2.1	List of example human genes affected by HERV regulatory sequences	9
2.2	A 2×2 contingency table	15
3.1	An example of a contingency table of sub-module 2.2	22
3.2	The number of records in the original downloaded data according to the categories of assembled sequences	24
3.3	The number of selected records from the data resources	26
3.4	Detailed proportions of each HERV superfamily in the HERV annotation data	26
3.5	List of the family names assigned by the mistakes in REannotate's text-processing, and their new edited names	30
3.6	The numbers and percentages of HERV elements according to each type of truncation patterns	32
3.7	Summary of both the numbers of genes and HERV elements resulting from mapping HERVs on the human genes	32
4.1	The number of the defragmented elements, resulting from the HERV defragmentation using REannotate with different values of the distance parameters, and their coverage percentages	41
4.2	Results of gene ranking in all four gene sets	56
4.3	P-values of the z-tests performing on the number of HERVs at each location between a group of significant and insignificant genes	57
4.4	Top ten genes ranked by the number of all neighboring HERVs throughout the genes	60
4.5	Top ten genes ranked by the number of in-gene HERVs	60
4.6	Top ten genes ranked by the number of intron HERVs	60
C.1	List of the categories of each HERV characteristics	97
C.2	Descriptions of terms shown in result summary of constructed profiles	98

C.3	Descriptions of terms shown in the menu after constructing profiles	99
C.4	Descriptions of terms shown in the result summary	102
C.5	Descriptions of terms shown in the result table	103
C.6	Descriptions of limiting parameters and option	103
C.7	Descriptions of terms shown in the result summary	105
C.8	Descriptions of terms shown in the result table	105
C.9	Descriptions of the available parameter and option	105

LIST OF FIGURES

FIGURE		PAGE
2.1	Diagram showing classification of the transposable elements and their fractions of the human genome	3
2.2	Genomic structures of retroviral proviruses and HERVs	5
2.3	Five potential mechanisms of the HERVs for modulating the expression of the neighboring genes	8
2.4	Summarization of purposed mechanisms used by human endogenous retroviruses in the etiopathogenesis of SLE and other autoimmune diseases	11
2.5	An example of the search page in HERVd	12
2.6	The first page of RetroSearch	13
2.7	The first page of TranspoGene	14
3.1	Diagram of system flow design	18
3.2	An example to construct HERV profile	21
3.3	Overview processes of system development	23
3.4	Flow chart showing all of the steps in the data selection	25
3.5	The number and proportions of each HERV fragment type	27
3.6	Flow chart showing the processes in the HERV defragmentation	27
3.7	An example of the results obtained from the HERV defragmentation by using REannotate	28
3.8	Flow chart illustrating the way to determine the type of truncation pattern for an HERV element	29
3.9	Flow chart showing all of the steps to manipulate HERV family names in the REannotate output	31
3.10	Summary of entity relationship diagram in HERV Profiler database	33
3.11	Work flow diagram illustrating the way to generate gene lists for testing the sub-module 2.2	34

3.12	Work flow diagram illustrating the way to generate gene lists for testing the sub-module 2.3	35
3.13	Work flow diagram illustrating methods to prepare the microarray data in the case study 1	36
3.14	Work flow diagram illustrating methods to prepare the microarray data in the case study 2	37
4.1	The numbers of the elements resulting from the defragmentations using different values of the distance parameter	39
4.2	The rates of change of all elements resulting from the defragmentation using different values of the distance parameter	40
4.3	Homepage of HERV Profiler	42
4.4	Serach/Profiling page	44
4.5	Summary results after input a gene list	45
4.6	Displaying neighboring HERVs in a tabular format	46
4.7	Displaying neighboring HERVs in a graphical format	47
4.8	Detailed information of the neighboring HERVs of a particular gene linked from the graphical display	48
4.9	The pages of profile construction before and after performing the construction	49
4.10	Full table of the HERV profiles	50
4.11	Full table of the HERV type descriptions	50
4.12	Results of finding over-represented HERV types	51
4.13	Results of ranking genes based on their possessed HERVs	52
4.14	Profile charts of a particular genes	53
4.15	Graphical display linked from the page of gene ranking	53
4.16	Tabular display linked from the page of gene ranking	54
4.17	Reported over-represented HERV types using the gene set 1	54
4.18	Reported over-represented HERV types by using the gene set 2	55
4.19	Reported over-represented HERV types by using the gene set 3	55

4.20	Reported over-represented HERV types by using the gene set 4	55
4.21	Box plots of the number of HERVs located at the downstream location comparing between top 100 significant genes and bottom 100 insignificant genes	58
4.22	Summary results of the 230 significant genes under the SLE condition	59
4.23	Over-represented HERV types among 219 significant gene list studied under SLE condition	61
4.24	Top ten genes related to over-represented HERV types	62
C.1	A bird's eye view describing overall features of HERV Profiler	79
C.2	Search/Profiling page	81
C.3	An example of the summary report	83
C.4	Location of the button of displaying in tabular format	84
C.5	Results of the tabular displaying of neighboring HERVs	85
C.6	Details of the gene isoform and HERV information shown in a rectangular area	86
C.7	An example of the defragmentation details	88
C.8	UCSC gene link	90
C.9	Location of the button of displaying in graphical format	91
C.10	Results of the graphical displaying of neighboring HERVs	92
C.11	Explanation of the components in the graphical display of a gene isoform	93
C.12	An example of the tabular result linked from the graphical displays	94
C.13	Location of the button of HERV profiling	96
C.14	Profile constructing page	96
C.15	Results shown after finishing constructing HERV profiles	98
C.16	Full table of the HERV profiles	99
C.17	An example file of the HERV profiles retrieved from the tool	100

C.18	Table of the defined HERV types	101
C.19	An example file of the HERV types retrieved from the tool	101
C.20	Results of finding over-represented HERV types among a gene list	102
C.21	An example file downloaded from the tool of the finding over-represented HERV types	103
C.22	Results of ranking genes following to possessing of the over-represented HERV types	104
C.23	Profile charts of a particular gene	106
C.24	An example file downloaded from the tool of ranking genes following to possessing over-represented HERV types	106
C.25	An example file of profile charts of a particular gene	107
C.26	Annotation summary results of DAVID	107

LIST OF TECHNICAL VOCABULARIES AND ABBREVIATIONS

BaEV	Baboon Endogenous Virus
CCDS	Consensus Coding Sequence
DAVID	Database for Annotation, Visualization and Integrated Discovery
DNA	Deoxyribonucleic Acid
ERV	Endogenous Retroviruses
GEO	Gene Expression Omnibus database
GIRI	Genetic Information Research Institute
GO	Gene Ontology
HERV	Human Endogenous Retrovirus
HERVd	Human Endogenous Retroviruses Database
ID	Identifier
LTR	Long Terminal Repeat
MMTV	Mouse Mammary Tumor Virus
MaLR	Mammalian apparent LTR-retrotransposon
mRNA	messenger Ribonucleic Acid
MuLV	Murine Leukemia Virus
NCBI	National Center for Biotechnology Information
ORF	Open Reading Frame
RefSeq	Reference Sequence Database
RNA	Ribonucleic Acid
RU	Rebase Update
SLE	Systemic Lupus Erythematosus
TE	Transposable Element
tRNA	transfer Ribonucleic Acid

SWISS-PROT	Swiss Institute of Bioinformatics-Protein Database
UCSC	University of California, Santa Cruz
UniProt	Universal Protein Resource