ความสัมพันธ์ระหว่างลักษณะของ LINE-1 กับระดับการแสดงออกของยีนในมะเร็ง โดยใช้เทคนิคการทำเหมืองข้อมูล

นางสาวนฤมล ประทานวณิช

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมชีวเวช (สหสาขาวิชา)
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

ASSOCIATION BETWEEN LINE-1 CHARACTERISTICS AND GENE EXPRESSION IN CANCERS USING DATA MINING TECHNIQUES

Miss Naruemon Pratanwanich

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Biomedical Engineering

(Interdisciplinary Program)

Faculty of Engineering

Chulalongkorn University

Academic Year 2011

Thesis Title	ASSOCIATION BETWEEN LINE-1 CHARACTERISTICS AND GENE EXPRESSION IN CANCERS USING DATA MINING TECHNIQUES
By	Miss Naruemon Pratanwanich
Field of Study	Biomedical Engineering
Thesis Advisor	Assistant Professor Chatchawit Aporntewan, Ph.D.
Thesis Co-advisor	Professor Apiwat Mutirangura, M.D., Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

……. Dean of the Faculty of Engineering
(Associate Professor Boonsom Lerdhirunwong, Dr.Ing.)

THESIS COMMITTEE

…….. Chairman
(Associate Professor Mana Sriyudthsak, Ph.D.)

……. Thesis Advisor
(Assistant Professor Chatchawit Aporntewan, Ph.D.)

…….. Thesis Cgggsao-advisor
(Professor Apiwat Mutirangura, M.D., Ph.D.)

…….. Examiner
(Assistant Professor Krung Sinapiromsaran, Ph.D.)

…….. Examiner
(Assistant Professor Sukree Sinthupinyo, Ph.D.)

…….. External Examiner
(Associate Professor Nachol Chaiyaratana, Ph.D.)

นฤมล ประทานวณิช : ความสัมพันธ์ระหว่างลักษณะของ LINE-1 กับระดับการแสดงออกของยีนในมะเร็ง โดยใช้เทคนิคการทำเหมืองข้อมูล. (ASSOCIATION BETWEEN LINE-1 CHARACTERISTICS AND GENE EXPRESSION IN CANCERS USING DATA MINING TECHNIQUES) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร. ชัชวิทย์ อาภรณ์เทวัญ, อ. ที่ปรึกษาวิทยานิพนธ์ร่วม : ศ. ดร. นพ. อภิวัฒน์ มุทริรางกูร, 117 หน้า.

ในเซลล์มะเร็งหลายชนิดพบว่าระดับดีเอ็นเอเมทิลเลชันที่กระจายทั่วไปบน L1 ลดลง เรียกว่าโกลบอลเมทิลเลชัน และมีการค้นพบว่ายีนของเซลล์มะเร็งบางยีนมีการแสดงออกที่น้อยลงอย่างมีนัยสำคัญ ดังนั้นเพื่อวิเคราะห์ลักษณะของ L1 ที่มีผลต่อการแสดงออกของยีน งานวิจัยนี้จึงวิเคราะห์ลักษณะของ L1 ทีละลักษณะโดยการทดสอบไคว์สแควร์และวิเคราะห์การถดถอยโลจิสติก พร้อมทั้งใช้เทคนิคการทำเหมืองข้อมูลโดยใช้ต้นไม้ตัดสินใจและกฎเชื่อมโยง เพื่อวิเคราะห์หลายตัวแปรร่วมกันของ L1 ที่มีผลต่อการแสดงออกของยีนในเซลล์มะเร็ง ผลการวิเคราะห์ข้อมูลลักษณะของ L1 ตัวแปรเดียวแสดงให้เห็นว่าบางลักษณะของ L1 โดยเฉพาะจำนวน L1 มีผลต่อการแสดงออกของยีนในทิศทางที่น้อยลงอย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ α = 0.05 นอกจากนี้ผลจากต้นไม้ตัดสินใจที่มีขนาดใหญ่ทำให้ยากที่จะแปลความหมาย แต่อย่างไรก็ตามกฎเชื่อมโยงสามารถแยกวิเคราะห์ตามชนิดของมะเร็ง โดยกฎที่ได้จากชุดข้อมูลมะเร็งกระเพาะปัสสาวะและมะเร็งตับสนับสนุนสมมติฐานที่ว่า การทรานสคริปชันของ L1 อาจจะควบคุมระดับการแสดงออกของยีนที่ลดลงได้ ซึ่งกฎทั้งสองชุดข้อมูลเสนอปัจจัยในการทรานสคริปชันของ L1 แตกต่างกัน คือ จำนวน L1 มากกว่าสองตัว และลำดับของ SRY Site1 ที่ไม่เปลี่ยนแปลง ตามลำดับ ส่วนกฎที่ได้จากมะเร็งต่อมลูกหมากเสนอปัจจัยที่มีผลต่อการย้ายตำแหน่งของ L1 ซึ่งรวมถึงกระบวนการทรานสคริปชันของ L1 ความเสถียรและกระบวนการของอาร์เอ็นเอ ทรานสเลชัน การตัดต่อดีเอ็นเอ และรีเวอร์สทรานสคริปชันและกระบวนการแทรก ปัจจัยเหล่านั้นคือ ลำดับบน ORF1 และ/หรือ ORF2 ที่ไม่เปลี่ยนแปลง สำหรับมะเร็งศีรษะและลำคอ ปัจจัยที่สำคัญที่ทำให้ระดับการแสดงออกของยีนลดลงคือลำดับของ TF-nkx-2.5 ที่ไม่เปลี่ยนแปลง นอกจากนี้กฎที่ได้จากชุดข้อมูลจำลองมะเร็งปอดโดยใช้สารเคมี 5-AZA แสดงให้เห็นว่าทิศทางการทรานสคริปชันของ L1 ทั้งสองทิศทางสามารถควบคุมระดับการแสดงออกของยีนได้

สาขาวิชา วิศวกรรมชีวเวช ลายมือชื่อนิสิต .....

ปีการศึกษา 2554 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5287180221 : MAJOR BIOMEDICAL ENGINEERING
KEYWORDS : LINE-1 / GENE EXPRESSION / CANCER / HYPOMETHELATION / DATA MINING /

NARUEMON PRATANWANICH : ASSOCIATION BETWEEN LINE-1 CHARACTERISTICS AND GENE EXPRESSION IN CANCERS USING DATA MINING TECHNIQUES. ADVISOR : ASST. CHATCHAWIT APORNTEWAN, Ph.D., CO-ADVISOR : PROF. APIWAT APORNTEWAN, M.D., Ph.D. 117 pp.

Global hypomethylation has been found on L1 in cancer cells. Moreover, having L1 is significantly associated with down regulation of hosting genes for some cancers. Nonetheless, not all genes that possess L1 are down regulated. To identify L1 characteristics that mediate gene expression in cancers, we performed chi-square test and logistic regression for each variable along with decision tree and classification association rules mining for multivariate data analysis. The results from statistical methods indicated the significant L1 characteristics, especially the number of L1, individually associated with gene expression using at significance level α = 0.05. For data mining, the size of the decision tree was too large to be useful. However, rules mining could generate interesting rules. Each cancer dataset has special characteristic rules. Firstly, the derived rules from bladder and liver cancer dataset support the hypothesis that L1 transcription may control down regulation. Both groups of rules suggest the mechanism to promote L1 transcription but different L1 characteristics, the number of L1 > 2 and conserved SRY Site1, respectively. Secondly, the rules derived from prostate cancer represent L1 retrotranspositional activities (conserved ORF1 and/or ORF2) which include L1 transcription, RNA stability and processing, translation, DNA restriction, reverse transcription and insertion. Finally, conserved TF-nkx-2.5 may control down regulation of head and neck cancer. Moreover, the derived rules from the dataset emulating lung cancer by 5-AZA shows that sense and antisense L1 can probably control the expression of genes by either directions of L1 transcription.

Field of Study : Biomedical Engineering Student's Signature .....
Academic Year : 2011 Advisor's Signature .....
Co-advisor's Signature .....

Acknowledgement

I have been supported in so many ways throughout my research and everything leading up to this point. Foremost, I wish to express my faithful to my advisor, Associate Professor Dr. Chatchawit Aporntewan and my co-advisor, Professor Dr. Apiwat Mutirangura for their assistance and encouragement in conducting this research.

I also gratefully acknowledge the members of my thesis committee, Associate Professor Dr. Mana Sriyudthsak, Assistant Professor Dr. Krung Sinapiromsaran, Assistant Professor Dr. Sukree Sinthupinyo and Associate Professor Dr. Nachol Chaiyaratana for their discussion and guidance.

In addition, many thanks go as well to the whole Intelligent System Laboratory (ISL) group in the Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, especially to Professor Dr. Prabhas Chongstitvattana, for fruitful discussions and ideas. Thanks to all of them and the working environment is exemplary.

Furthermore, I am grateful to the ChulalongKorn University Graduate Scholarship to Commemorate the 72nd Anniversary of His Majesty King Bhumibol Adulyadej for my financial support.

Last but not least, special thanks to my family who support and encourage me throughout this study and my every day life. I am very thankful to my friends for their friendship and their help during my graduate study.

Contents

Chapter Page

Chapter Page

List of Tables

List of Figures

CHAPTER I

Introduction

1.1 Background and motivation

According to World Health Organization reports [1], cancer is considered the second leading cause of death worldwide after cardiovascular disease. As the matter of fact, the total number of global cases is rapidly increasing, especially in developing countries. WHO estimates that the total number of the global deaths caused by cancer will have increased 45% in the next two decades, from 7.6 million to over 11 million deaths. Fortunately, it is known that cancer cells have epigenetic characteristics different from normal cells. Therefore, better understanding of the differences between cancer cells and normal cells could assist biologists to discover effective methods to reduce this terrible death rate.

DNA methylation is one of the best known epigenetic differences between cancer cells and normal cells. DNA methylation is a fundamental molecular characteristic for regulation of transcriptional process by attaching methyl groups (CH_3) to DNA molecules. If the DNA methylation on gene promoter regions is altered, it may probably effect on gene expression or the quantity of messenger RNA (mRNA). Consequently, the effected cells are likely to function aberrantly.

In 1983, the important epigenetic abnormality in cancer cells was first discovered [2]. Such anomaly is that cancer cells are usually unmethylated at CpG islands, DNA regions that contain plenty of CG sequences and usually appear in gene promoter regions. It is well known that when these methylated islands become unmethylated, they often cause the activation of genes nearby.

Recently, not only is gene-specific hypomethylation on promoter regions of gene identified in cancer cells, global hypomethylation is also found in cancers cells relative to their normal counterparts [2]. The loss of genome-wide DNA methylation is

found mostly on repetitive regions including LINE-1 or long interspersed nuclear element-1, in various types of cancer cells. Unlike the hypomethylation at CpG islands, the mechanism resulted from the global hypomethylation has been in doubt over these recent decades. Nevertheless, C. Aporntewan *et al.* [3] discovered that some of the genes with LINE-1 are significantly downregulated (see Table 1.1). The chi-square test shows that LINE-1 is associated with down regulation of the hosting genes at significance level of 0.01 resulted from chi-square test.

Table 1.1 The association between the existence of LINE-1 in genes and gene expression in bladder cancer on the 2×2 contingency table [3]

	Down (p-value < 0.01)	Not down
LINE-1	382	537
No LINE-1	3377	8762
	p-value = 9.83E-19	
	Odds = 1.85	

From Table 1.1, a gene with LINE-1 is denoted by “LINE-1” and a gene without LINE-1 is denoted by "No LINE-1". The down regulation of a gene determined by unpaired t-test (p-value < 0.01) is denoted by “Down”. The entries in the 2×2 contingency table show the corresponding number of genes. The statistical values, both p-value resulted from chi-square test and odds ratio, is shown under the table.

Genes with LINE-1 have the high risk to be downregulated almost twice as much as genes without LINE-1 (odds = 1.85). However, it should be noticed that not all of the genes with LINE-1 are downregulated (see Table 1.1). This observation implies that the existence of LINE-1 in genes is not enough to utterly describe down regulation of gene expression. Therefore, to help biologists understand gene expression regulated by LINE-1 in cancer, the study of LINE-1 characteristics is needed.

1.2 Objectives

In this thesis, we determine LINE-1 characteristics which are significantly associated with the down-expression of the genes with LINE-1 in cancers. First, we examine a LINE-1 characteristic by using chi-square test. Second, to simultaneously analyze LINE-1 characteristics requires a large amount of computational power and time consuming; therefore, we apply data mining techniques to mine interesting sets of LINE-1 characteristics from huge cancer data sets within given time. Here we choose classification association rules to represent the sets of LINE-1 characteristics which are significantly related to gene expression in cancers. In addition, we apply a decision tree algorithm to classify gene expression in cancers by LINE-1 characteristics. Finally, we compare the resulting rules and the derived tree.

1.3 Scope

In the beginning, we start by the experiment with parameters for data mining techniques on the gene expression data of bladder cancer. After that, we are applied the similar settings on the other gene expression data sets shown in Table 1.2.

Table 1.2 The list of all gene expression data sets in cancer explored in the thesis

Dataset ID	Description
GSE6631	head and neck squamous cell carcinoma vs normal oral epithelium.
GSE9750	cervical cancer cells vs cervical cancer epithelium.
GSE5816	lung adrenocarcinoma vs human bronchial epithelium.
GSE14811	liver cancer vs normal liver.
GSE1299	Breast Cancer cells vs Normal Breast Epithelium.
GSE5764	ductal and lobular breast cancer vs normal breast.
GSE3167	bladder carcinoma situ vs normal bladder epithelium.
GSE13911	microsatellite instable gastric cancer vs normal stomach epithelium.
GSE6919	metastasis prostate cancer.
GSE9764	5-azadeoxycytidine treated vs untreated human mesenchymal stem cells.
GSE5816	hBEC high dose vs human bronchial epithelium.
GSE4246	HEK293T Ago2sh.
GSE14537	AGO2IP vs Control.
GSE14054	Importin8si-AGO2IP vs control-AGO2IP.

1.4 Contribution

Since the possible combination of LINE-1 characteristics associated with gene expression in cancers will possibly produce a large number of candidates for testing hypotheses, this research helps biologists reduce the number of hypotheses using data mining techniques. With the small number of hypotheses, it is possible for biologists to test the hypotheses in their future experiments.

1.5 Thesis Organization

Next chapter of this thesis covers the biological concepts and computational techniques, both data mining and statistical concepts, used throughout this study. Besides, we explore the literatures related to this work. In chapter III, we demonstrate our methodology to mine the interesting classification rules with statistical testing. Later on, Chapter IV shows the results of our experiments. Finally, we discuss and compare all results in chapter V.

CHAPTER II

Literature Reviews

In this chapter, we describe the fundamental biological and computational background. Next, we review the literatures related to the association rules mining on gene expression data and the studies of LINE-1 in cancers.

2.1 Biological Background

2.1.1 Gene expression

Deoxyribonucleic acid (DNA) comprises four types of nucleotides with its bases: adenosine monophosphate (AMP) with adenine (A), guanosine monophosphate (GMP) with guanine (G), cytidine monophosphate (CMP), cytosine (C), and thymidine monophosphate (TMP), thymine (T). DNA locates in the nucleus and has responsibility to carry genetic information. The different information is made of the different order of these nucleotides. In human genome, there are about 3,300 millions base pairs (bp) distributed on 23 chromosomes. DNA consists of genes and non genes. A gene or a coding region, has three parts: promoter (5') region, gene body (open reading frame) which contains exons (translated regions) and intron (untranslated regions) regions alternately, and terminator (3') region. Any two genes are seperated by a non-coding region called intergenic DNA (see Figure 2.1).

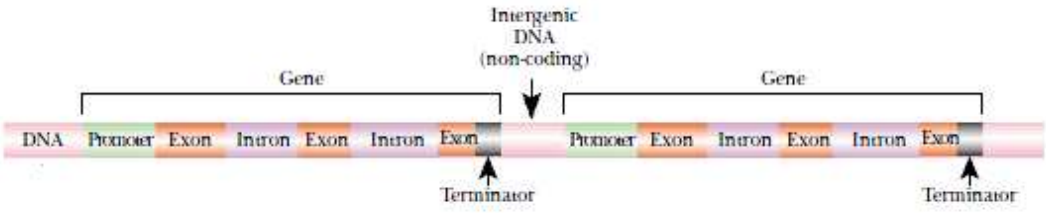
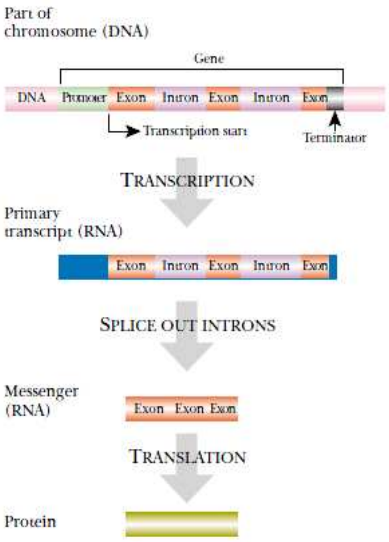


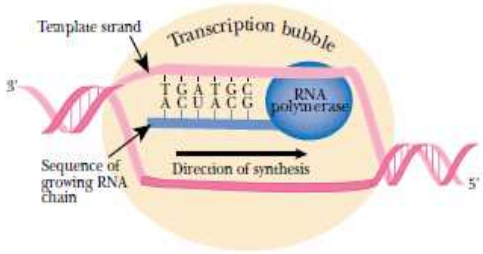
Figure 2.1 The structure of gene [4]

Genes are expressed by generating ribonucleic acid (RNA) through transcriptional process as following steps (see Figure 2.2). The promoter is first

recognized by the enzyme RNA polymerase. Consequently, double helix DNA is locally separated to allow transcription. Next, the RNA polymerase produces primary transcript RNA (tRNA) by using one of DNA single strands as a template in the direction of 5' (upstream) to 3' (downstream). The RNA polymerase generates pre-mRNA until it meets the terminator. The signal to terminate is often consecutive sequences of adenine (A) bases called Poly-A tail. Finally, the intron regions of pre-mRNA are filtered by splicing process. Messenger RNA (mRNA) is the product of this process. Such mRNAs are responsible to transfer genetic information from nucleus to cytoplasm. These mRNAs are finally translated into proteins as the final gene products for running cells normally. The quantity of mRNA, or gene expression level, are able to be measured by microarray technology.



a) The stage of converting gene to protein



b) The transcriptional process

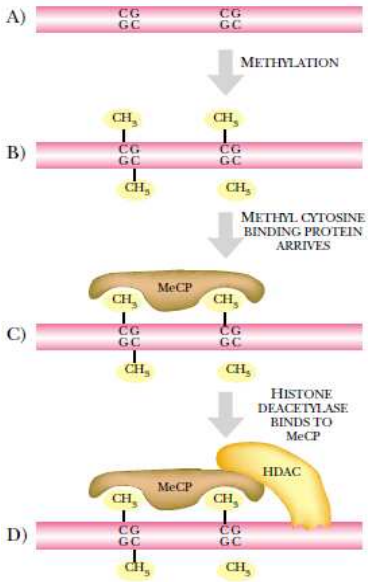
Figure 2.2 The process of decoding from gene to protein [4]

2.1.2 DNA Methylation

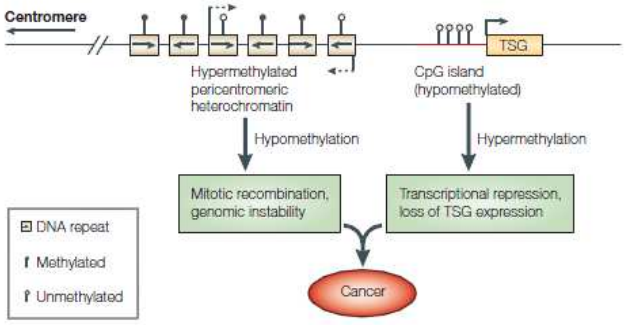
DNA methylation is an epigenetic process of attaching nucleotides with methyl (CH_3) groups in genome. In human, the methylation is considered as a

mechanism to regulate gene expression. Principally, if CpG islands or promoter regions are methylated, such methyl groups control gene expression by silencing that gene (see Figure 2.3a). Since the methylated CpG islands are recognized by methylcytosine binding proteins and some other proteins, this mechanism impedes RNA polymerase to transcribe that methylated gene. As a result, the gene with methylation is finally silenced. Therefore, the methylation on promoter regions has the responsible for discriminating different cells in different organs by switching genes on and off.

Due to its capability to regulate gene expression, the alteration of the DNA methylation level can effect the quantity of mRNA or gene expression level [4]. In cancer (see Figure 2.3b), it is well known that tumor suppressor gene (TSG), which plays an important role in suppressing the growth of tumor cells, is hypermethylated on its CpG islands or promoter regions. Consequently, the TSG gene cannot be accessible for transcription. When this TSG gene is silenced, the tumor cells are no longer suppressed and are likely to become cancer cells. This gene-specific hypermethylation is well described in many studies, appeared in [2].



a) Silencing gene by DNA methylation



b) DNA methylation in cancer

Figure 2.3 DNA methylation [2]

However, another significant methylation event in cancers is global hypomethylation, particularly on repetitive regions including LINE-1. This molecular phenomenon can result in genomic instability. Recently, the genome-wide hypomethylation in cancer is associated with gene expression level [3]. Nevertheless, the mechanism to regulate gene expression has not been elucidated yet.

2.1.3 Biology behind LINE-1

In addition to normal genes that contribute to our living, human genome consists of many DNA fragments called repetitive sequences [4]. It is believed that these fragments have been evolved from viruses in the past. Unlike the host genetic elements, these fragments can move around from site to site within host deoxyribonucleic acid (DNA) molecules; thus, they are called transposable elements or transposons.

Mainly, 45% of human genome is transposable elements which are characterized into two main classes by the structures and the methods of transposition [5]. One of them is retrotransposons whose mobility relies on autonomous replicative transpositioin while in their ribonucleic acid (RNA) phase. In other words, the retrotransposons move around from place to place by converting their RNA transcripts back into DNA and inserting into the host genome. In this process, they use reverse transcriptase for their movement. However, the only active autonomous retrotransposons in the human genome is Long Interspersed Nuclear Element-1 (LINE-1, L1), a family of the retroelements which is a half of the entire transposons. Currently, L1s are widely spread in both intragenic and intergenic regions in human genome, approximately 20% of the whole genome, due to their evolution and replication in mammals for over millions years [6].

The complete long interspersed nuclear element-1 (LINE-1, L1), approximately 6,000 bp in length, is composed of the 5'-untranslated region (5'-UTR) as an internal promoter for beginning transcription, separately two open reading frames

(ORF1 and ORF2) as coding regions, and 3'-untranslated region (3'-UTR) as a tail containing the signal for ending transcription (see Figure 2.4). As for their movement, L1s use RNA-binding protein encoded from ORF1 and another two proteins, endonuclease and reverse transcriptase, encoded from ORF2 for transpositional activity. In fact, their movement occurs mostly in the embryo phase [4]. Despite the inactive movement after embryo phase, some of L1s (even truncated) are still transcribed to messenger RNA (mRNA) [7].

Figure 2.4 The structure of LINE-1

To sum up, since gene expression can be regulated by DNA methylation, the loss of genome-wide methylation on repetitive regions including L1s in various types of cancer cells may impact on gene expression level. This is the purpose to pursue this thesis.

2.2 Computational Background

2.2.1 Statistical methods

2.2.1.1 Bonferroni correction

The concept of the multiple testing correction is considerable when more than one hypotheses are tested in the similar data set, no matter whether they are independent or not. According to Bonferroni correction, the individual hypothesis should be tested at a significance level of α/n, where α is the critical p-value and n is the number of the hypotheses under study.

2.2.1.2 Logistic regression [9]

Logistic regression analysis uses a set of the independent variables to estimate the probability of the occurrence of the dependent variable Y, where Y is a binary variable. The advantage of this analysis is that it requires neither the normal distribution of the dependent variable nor the linear relationship between the explanatory and response variables.

To estimate the probability that the event (Y) would occur, the logistic regression model uses the Equation (2.7) to describe the relationship between the independent variables and dependent variable in terms of the likelihood, where p is the probability that $Y=1$, $1-p$ is the probability that Y is the other value, and the ratio of both probabilities is known as the Odds value. In other words, the logistic regression model is defined by the log of the Odds of the dependent variable known as logit value.

$$ln(\frac{p}{1-p}) = ln(e^z) = z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (2.7)$$

where

$$p = Prob(Y = 1) = \frac{1}{1 + e^{-z}}$$

$$1 - p = Prob(Y = 0) = 1 - Prob(Y = 1)$$

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

$$Odds = \frac{p}{1-p}$$

Y = The dependent variable X_i = The independent variable *i*

Statistically, the model must pass three criteria as follows.

1. Test the goodness of fit of the proposed model by applying the maximum likelihood concept under the null hypothesis that all coefficients of the explanatory variables are zeroes. To verify this, it starts with calculating -2 times of log likelihood known as $-2LL$ or deviance of both the null model, in which there is no independent variables, and the proposed model. The minimum value for the deviance is 0, which reflects a perfect fit model. Since the distribution of the deviance is like chi-square distribution, the significance of this proposed model is based on the chi-square test where chi-square value is the difference of both deviance and the degree of freedom is the number of the independent variables in the model. The more the reduction of deviance, the better the proposed model.

2. Confirm that each coefficient of the independent variables in the model is statistically significant. The null hypothesis is that the coefficient β_i of the explanatory variables is zero. The method to prove this hypothesis is similar to the approach described above; but, the other independent variables except β_i are considered as constant values and the degree of freedom is always 1.

3. In addition to the significance tests above, the capability of the proposed model to estimate the likelihood of the response variable from the set of the explanatory variables is indicated by $\mathrm{R}^2_{\mathrm{Logit}}$ in the Equation (2.8), where the value of $\mathrm{R}^2_{\mathrm{Logit}}$ is between 0 and 1. If $\mathrm{R}^2_{\mathrm{Logit}}$ value is 1, it indicates that the proposed model is perfect. However, the model must be verified by the significance tests as well.

$$R^2_{Logit} = \frac{-2LL_{null} - (-2LL_{model})}{-2LL_{null}} \tag{2.8}$$

where

$-2LL_{null}$ = The deviance of the null model

$-2LL_{model}$ = The deviance of the proposed model

2.2.2 Data mining techniques

Due to the advent of the high-throughput technology like microarray, data mining is considered as a powerful and essential approach for knowledge discovery. This technique learns the set of the data and constructs the output models to represent the knowledge. These models can be used for many purposes based on the kinds of knowledge mined such as clustering, association, and classification purpose.

Generally, some databases are abound with the interesting information that can be used for describing the characteristics of the data under study in terms of category. Classification data mining is a tool to study the data in this purpose. This technique analyzes the collection of data and create a model for determining what features can be used as classifiers.

The data prepared as an input of the classification model is a single two-dimensional table (see Figure 2.5), where each row is a transaction or a tuple, each column is an item, an attribute, or a feature, and the last column is the class of each row. It is noticed that the class column must be non-numerical or categorical. The output model of the classification can be represented in form of rules.

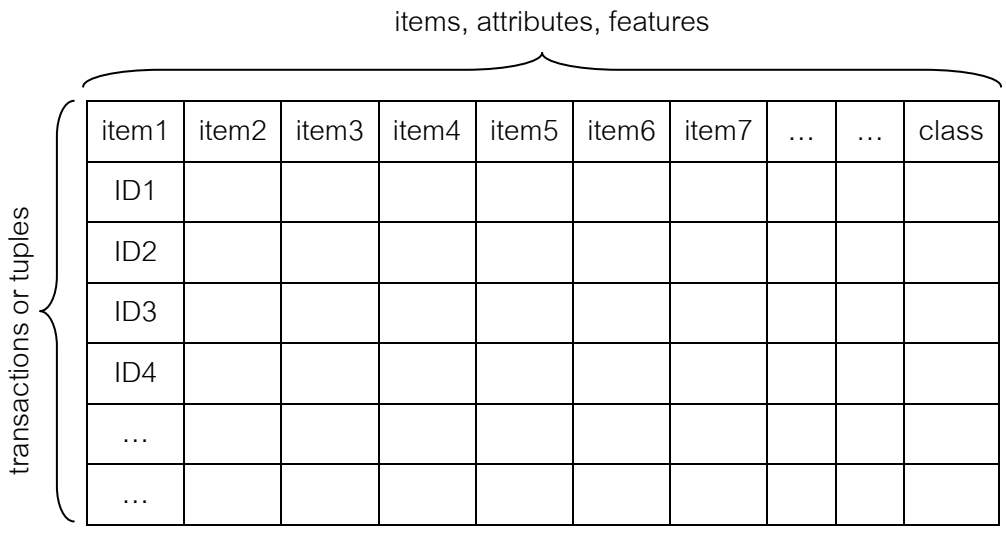


Figure 2.5 The input of the classification model

2.2.2.1 Decision tree [10]

Decision tree mining is a popular classification algorithm due to its simplicity and fastness with little domain knowledge and no any parameter setting. This method uses a tree to represent the classification concept. The individual path belongs to a classification rule, where the inner nodes are the attributes with their values in the branches below and the leaf node is defined as a class. For example, Figure 2.6 demonstrates the classification rules generated from a decision tree.

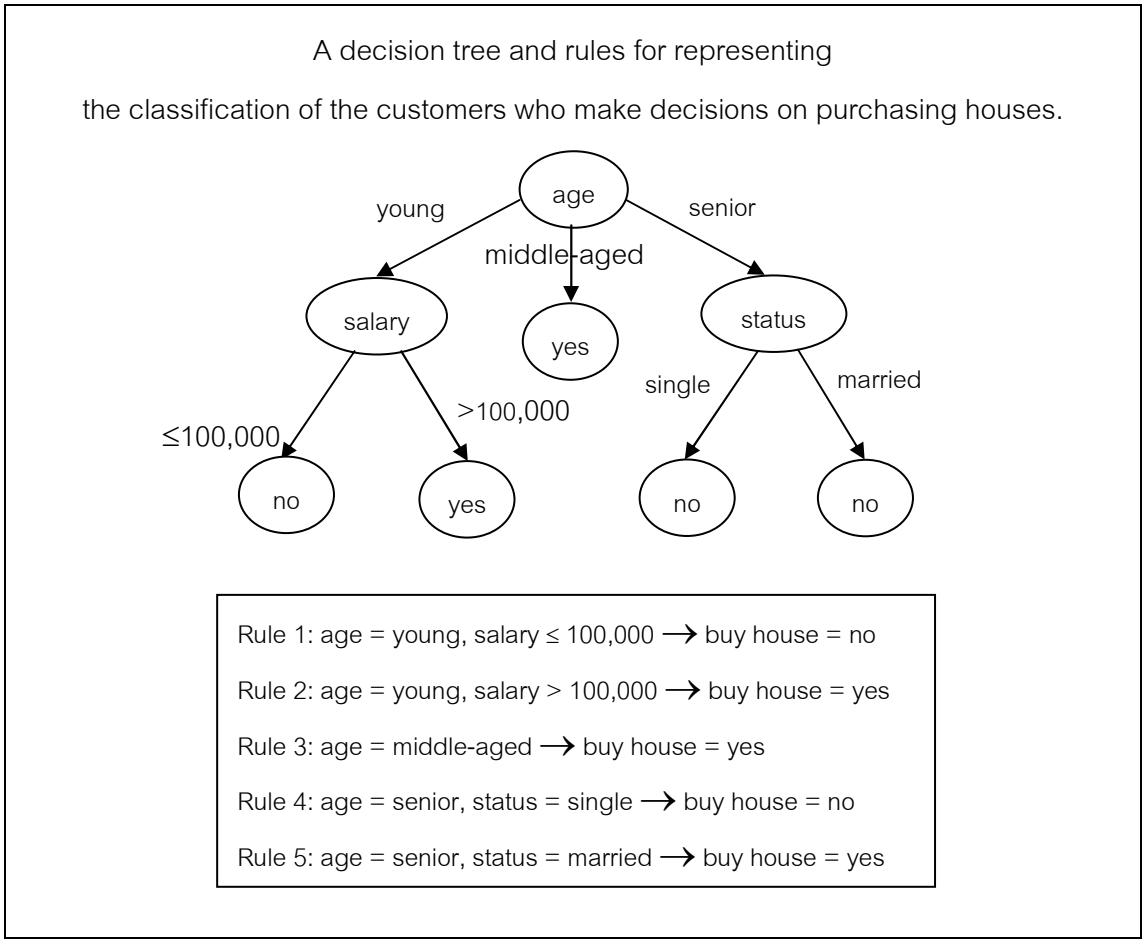


Figure 2.6 The example of decision tree with its generated rules

2.2.2.1.1 C4.5 algorithm

Of decision tree induction, C4.5 is a fundamental algorithm allowing both numeric and non-numeric data types as its input. This algorithm constructs trees by a

top-down recursive divide-and-conquer manner based on a greedy approach. The C4.5 algorithm is described in Figure 2.7.

C4.5 algorithm

Input: 1. A data partition D
2. The attribute list of D
3. The attribute selection method

Output: A decision tree

```
        C4.5 (D, attribute_list, attribute_selection_method)
(1)     {   create a node N
(2)         if all tuples belong to the same class C
(3)         {return node N as a leaf node labeled with the class C          }
(4)         if attribute_list is empty
(5)         {  reture node N as a leaf node labeled with the majority class in D      }
(6)         apply attribute_selection_method(D, attribute_list)
(7)         label node N with the splitting_criteriaon
(8)         If splitting_attribute is discrete-valued and multiway splits allowed
(9)         {  Attribute_list = attribute_list – splitting_attribute }
(10)        for each outcome j of splitting_criterion
(11)        {  let Dj be the set of data tuples in D satisfying outcome j
(12)           if Dj is empty
(13)              {  attach a leaf labeled with the majority class in D to node N  }
(14)           else {  attach the node returned by generate_decision_tree(Dj,attribute_list) to node N  }  }
            return node N    }
```

Figure 2.7 C4.5 algorithm [10]

There are three parameters as the input of this algorithm, the data partition D, attribute_list as the list of attributes, and the attribute selection method which performs the heuristic function to choose the best attribute for branching. The C4.5 algorithm is described as follows (see Figure 2.7).

1. Create a node N (Line 1) and if the terminating criteria is not met (Line 2-5), apply the attribute selection method, which performs one of the attribute selection measure, either gain ratio or information gain. After its computation, this heuristic method returns

the splitting criterion which consists of the splitting attribute and perhaps a splitting point if the selected attribute is continuous-valued or numeric (Line 6).

2. Label node N with the splitting attribute (Line 7).

3. If the selected attribute is categorical or nominal, then remove the selected splitting attribute from the attribute list since it is not needed for future partitioning (Line 8-9).

4. Split each branch from node N with each result of the splitting criterion (Line 10-11).

5. Let D_j be the result of partitioning D with outcome j. Call the C4.5 function recursively until meeting the terminating conditions below (Line 14).

- If all tuples in the partition D_j are in the same class (Line 2) or no attributes in the attribute list (Line 4), then assign the node N as the leaf node with the majority class in D (Line 3,5).
- If the partition D_j has no tuples, then create the leaf node labeled with the majority class in D and attach it to node N (Line 13).

2.2.2.1.2 Attribute selection measures

To split branches in each level of the decision tree, the best attribute is chosen as a node to classify each tuple by computing one of the following attribute selection measures [10].

2.2.2.1.2.1 Information gain

The information gain applies the information theory to seek the suitable splitting attribute. By definition, the best splitting attribute is the attribute whose values partition the data into groups such that the sum of the purity in each group is maximized under this partitioning. The purity of the data group is evaluated by the similarity of the class labeled in each tuple. For example, if the tuples in a group have the same class label, the purity of this group is 100%. Indeed, the purity of the data in each group is measured by the entropy value where the purity is inversely proportional to the entropy. In other words, the purer the data, the less the entropy. According to the information theory, the entropy can be alluded as the expected information needed to classify a tuple in terms of the number of bits. Therefore, the expected information can be calculated in the Equation (2.9).

$$Info(D) = -\sum_{i=1}^{m}(p_i log_2 p_i) \quad (2.9)$$

Let node N have the tuples of partition D with m classes, p_i is the ratio of class and the total number of tuples in D ($\frac{C_i}{D}$). When the data of Node N is divided into v groups by attribute A, the new expected information after partition is calculated in the Equation (2.10), where the term $\frac{|D_j|}{D}$ plays as the weight of the j^{th} partition.

$$Info_A(D) = \sum_{j=1}^{v}\frac{|D_j|}{|D|} \times Info\,(D_j) \quad (2.10)$$

To select the best splitting attribute, the expected information before and after partitioning by every attribute is computed. Next, the difference of every couple of the expected information known as the information gain is calculated and compared (see Equation (2.11)).

$$Information\ Gain(A) = Info(D) - Info_A(D) \quad (2.11)$$

$Info(D)$ and $Info_A(D)$ is the expected information before and after partitioning by attribute A, respectively. Finally, the attribute with the highest information gain is selected as the splitting attribute for branching.

2.2.2.1.2.1 Gain ratio

The gain ratio is an extension of the information gain by applying a normalization approach. Since the previous measure is sometimes biased by selecting the attribute with a lot of values but useless for classification purpose such as ID of the tuples, the gain ratio uses the split information value to normalize the information gain. The split information value is computed according to the Equation (2.12).

$$Split\ Info_A(D) = -\sum_{j=1}^{v}\frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|}) \quad (2.12)$$

The split information slightly differ from the information where it considers the total number of tuples in the j^{th} partition instead of the number of tuples that respect to the class C_i. Next, the gain ratio is computed in the Equation (2.13).

$$Gain\ Ratio(A) = \frac{Gain(A)}{Split\ Info(A)} \quad (2.13)$$

Finally, like the information gain, the attribute with the highest gain ratio is chosen as the splitting attribute.

2.2.2.2 Classification association rules

Association rules mining is a searching technique for the frequent patterns in a given data. For classification purpose, this method uses a rule in form of X $\rightarrow$ Y, where X and Y are the set of items and a class label respectively, to represent the association between the items in the antecedent and the class in the consequence of the rule.

2.2.2.2.1 The important definitions

Let X and Y be any subsets of the items, then

- **Itemset or pattern** is the nonempty subset of items in a given dataset.
- **Support(X)** is the frequency of X appeared in a given dataset.
- **Support(X $\rightarrow$ Y)** is the ratio of the frequency of X U Y appeared in the data and the total number of samples in a given dataset.
- **Confidence(X $\rightarrow$ Y)** is the ratio of the support(X $\rightarrow$ Y) and the support(X) known as the probability of Y given X.
- **Minimum support** is the threshold for finding frequent patterns.
- **Minimum confidence** is the threshold for generating association rules.
- **Frequent itemset** is the itemset that appear in the given data that has support bigger than or equal to the minimum support.

2.2.2.2.2 The processes in rule mining

In general, association rules mining can be divided into two steps. Firstly, this method finds the itemsets whose supports satisfy the minimum support. Such itemsets are known as frequent itemsets. Secondly, all rules are generated from all frequent itemsets. The rules which do not satisfy the minimum support and minimum confidence are pruned.

2.2.2.2.3 FP-growth

FP-growth proposed in the previous work [11] is one of the efficient techniques to mine frequent patterns without candidate generation. This concept can reduce the memory usage even when the minimum support threshold is low, since it uses tree as its data structure to keep the frequent patterns instead of generating all candidates.

The structure of the FP-tree begins with the null root node, each item collected as a child node, and a frequent-item-header table where each entry consists of the name of frequent item and the head of node-link pointing to the first node which carries the similar name. Each node comprises three values-, its name, its frequency, and its link for connecting the next node. To discover the frequent patterns, there are two algorithms used in this technique. One is to construct FP-tree (see Figure 2.8), the other is to generate the frequent patterns (see Figure 2.9).

FP-tree construction algorithm

Input: 1. A database D in form of a two-dimensional table
2. Minimum support threshold

Output: A FP-tree of the data

```
     FP_tree(Database, Minimum_support)
(1)  {  FP_tree = new tree(null)
(2)     F_List = new list(null)
(3)     Foreach Transaction in Database
(4)     {  Item[i].frequency ++ }
(5)     Foreach itemi
(6)     {  If(item[i].frequency > Minimum_support)
(7)        {  F_List.add(item[i]) }
(8)        Sort(F_List, Support_descending)
(9)        Create_frequent_item_headertable(F_List)
(10)       Foreach Transaction in Database
(11)       {  Insert([head|tail], FP_tree)        }}
```

```
(12)  Insert(Node,Tree)
(13)  {   If(Tree.contain(Node))
(14)        {  Tree.increment(Node,1)      }
(15)     Else
(16)        {
(17)           Initial_support = 1
(18)           Tree.add(Node,Initial_support)
(19)        }
      }
```

Figure 2.8 FP-tree construction algorithm [10]

The algorithm to construct the FP-tree is performed as follows (see Figure 2.8).

1. Create a tree with the null root node and a null frequent items list (Line 1 and 2).
2. Scan all transaction in the database and compute the frequency of each item (Line 3 and 4).
3. Collect the frequent items, which appear in the database that passed the minimum support test, into F_List (Line 6 and 7).
4. Sort the items in F_List by descending respect to the support (Line 8).
5. Create the frequent-item-header from F_List (Line 9).
6. For each transaction in the database, insert each frequent items in the transaction into the tree according to the order in the F_List one by one (Line 10 and 11). To insert any item i, the condition of the existence of the item i in the prefix path of the tree. If tree contains the node of item i, then increment the node i's frequency by 1; otherwise, create a new node where its initial frequency is set to 1 (Line 12-19) and the parent link connects to its parent and the node link points to the node carrying the similar name.

FP-growth algorithm	
Input: FP-tree	Output: The complete set of frequent patterns
	FP-growth(tree, α)
(1)	{if Tree contains a single prefix path
(2)	{let P be the single prefix-path part of Tree
(3)	let Q be the multipath part with the top branching node replaced by a null root
(4)	foreach combination (denoted as β) of the nodes in the path P
(5)	{generate pattern $\beta \cup \alpha$ with support = minimum support of nodes in β
(6)	let frequent_pattern_set(P) be the set of patterns so generated}
(7)	else { let Q be tree
(8)	Foreach item ai in Q
(9)	{generate pattern β = ai $\cup \alpha$ with support = ai.support
(10)	construct β's conditional pattern-base and then β's conditional FP-tree $Tree_\beta$
(11)	if $tree_\beta \neq \varnothing$
(12)	{call FP-growth($tree_\beta$, β)
(13)	Let frequent_pattern_set(Q) be the set of patterns so generated
(14)	return(frequent_pattern_set(P) $\cup$ frequent_pattern_set(Q) $\cup$ (frequent_pattern_set(P) $\times$
(15)	frequent_pattern_set(Q)))

Figure 2.9 FP-growth algorithm [11]

The algorithm to generate the frequent patterns from FP-tree is divided into three portions (see Figure 2.9): the single prefix-path P (Line 3), the multiple path Q (Line 4), and their combinations (Line 15). Besides, this algorithm is considered in three conditions:

1. If the FP-tree contains a single prefix-path, enumerate every combination patterns with the support being the minimum support of the nodes in the subpath (Line 1-5).
2. If the FP-tree is the multiple path that does not contain the single path denoted as Q, construct the conditional pattern-based and for each item in the FP-tree (Line 8-11).
3. The terminating condition checks if the FP-tree is empty (Line 11). Unless the stopping criteria is met, call the function FP-growth recursively.

Lastly, this algorithm returns the complete set of frequent pattern sets in Line 15.

2.3 Literature reviews

2.3.1 Association rules mining on gene expression

DNA microarray, a high-throughput biotechnology, can currently measure the expression levels of thousands of genes or even the entire genome under different experimental conditions at a single test. In general, the microarray generates gene expression data in form of M × N matrix, where M and N are the columns of genes under study and the rows of experimental conditions respectively. Due to the advent of the sophisticated technology, researchers has been changed the method to analyze gene expression data from gene-specific to genome-wide.

Association rules mining is considered as a powerful methods for uncovering the relationship of gene expression data from microarray technology as a whole genome. For decades, many researchers have studied such relationship in different ways. Nonetheless, the analysis of gene expression data by using association rules mining can be divided into three primarily goals [12][13][14]:

1.) To examine the association between the set of genes.

2.) To discover co-expressed genes associated with certain biological conditions.

3.) To identify what genes in cancer are expressed differentially from normal cells.

Principally, association rules mining on gene expression data is composed of five phases (see Figure 2.10) [13]:

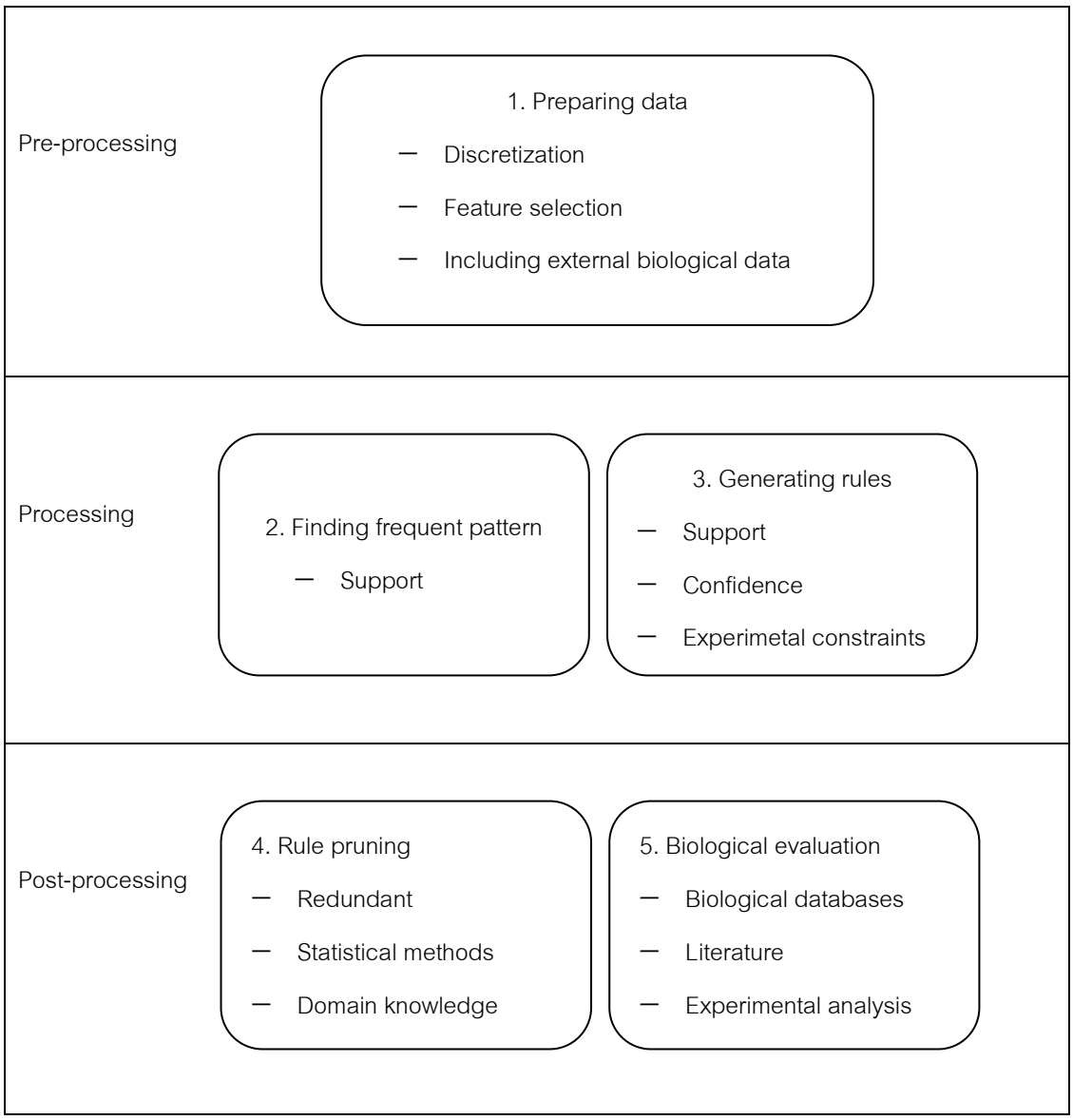


Figure 2.10 The five phases of mining association rules on gene expression data [13]

2.3.1.1 Preprocessing phase

To mine association rules on gene expression data, discretization is an inevitable task in preprocessing phase. There are two strategies of discretization - supervised and unsupervised methods. U. M. Fayyad and K. B. Irani applied the supervised strategy based on information theory to discretize gene expression data [15]. They divided numeric values into a number of disjoint ranges by using recursive binary partition algorithm so as to minimize the entropy of each bin. As for the unsupervised strategy, the basic approach used in the previous work [16] is equal width partitioning. This method divides the measured features into k bins equally, where k is a parameter defined by the researcher. Another method is to use the thresholds from some prior knowledge in discretization process [17][18]. In most genetic studies, statistical approaches like t-test are applied to detect the gene expression mean difference in cancer cells and normal counterparts [3].

In addition, biological information other than gene expression data can be included in this process [14][18]. Similarly, if there are some correlations between items or there are too many attributes to mine, feature selection should be operated before mining phase. Finally, the input data must be prepared in form of a single two-dimensional table before processing in the next phase.

2.3.1.2 Frequent patterns (itemsets) mining phase

This phase is considered as a bottle neck of association rules mining. Many researches prove that finding frequent patterns is NP-hard problem because memory space and time consumption are exponentially grew by the number of items in the data set. In the earlier, APRIORI proposed by J. F. Hair *et al.* is widely used [9]. This column-enumeration-based algorithm is suitable when the frequent pattern is short; however, this method is impractical in case of mining long frequent patterns. Therefore, a new algorithm FP-Growth is an alternative [11]. The advantage of this algorithm is to use FP-Tree data structure for keeping frequent patterns instead of generating all candidates like APRIORI. This structure uses less memory space than APRIORI. Another column-enumeration algorithms rely on closed-based algorithms. The closed itemset is an

itemset which has no super sets with similar support. Thus, mining such itemsets can cope with redundant rules which will be generated in the next step.

Microarray data has the number of columns (genes under study) much more than the number of rows (experiments or samples); therefore, the previous column-enumeration-based methods might not be appropriate. Unlike the column-based methods, The row-enumeration-based algorithms discover itemsets in row space instead of column space [19][20]. Like column-enumeration-based techniques, there are many algorithms which solve a huge number of generated rules problem like closed itemsets mining techniques.

However, both column and row enumeration approaches are support based. Therefore, no matter what techniques are applied, we may still uncover very rare itemsets with high confidence. In this case, confidence based method is a solution. This technique uses confidence threshold in finding frequent patterns phase instead of support threshold. This technique was proposed by T. Mcintosh and S. Chawla [21].

2.3.1.3 Association rules generation phase

From all frequent itemsets, association rules are generated with support and confidence. Normally, a frequent itemset can generate 2^{k-2} rules, where k is the size of each itemset. Certain constraints are sometimes used in this phase, depending on the purposes of the study. For example, the size of the left-hand or right-hand side of the rule is fixed [17][22]. In addition, if the goal is to apply association rules as classification rules; therefore, they generated rules whose the right-hand side of the rules were one of the class labels [19][20][23].

2.3.1.4 Rules filtering

In principal, the minimum support and confidence of the rule are primary parameters rules filtering. Besides, some studies [22] used Certainty factor (CF) as another filter to avoid some shortcomings of the support and confidence framework, since the certainty factor is calculated from both support and confidence. Furthermore, statistical methods such as chi-square test are also used in this process to determine which rules are statistically significant. Nevertheless, the number of generated rules is

still too tremendous to be interpreted in terms of biological information; therefore, redundant rules have to be eliminated. By definition, for classification purpose, a redundant rule is a rule whose left-hand side is a superset of another rule which has the similar class in the right-hand side. Some techniques such as top-K covering rule groups technique proposed by G. Cong require the less number of rules; therefore, not only are redundant rules pruned, low-rank rules are also filtered [20]. On the other hand, when the association rules are used as classification rules, the methods to filter rules are different. In this sense, an uninteresting rule is a rule whose antecedent is a superset of that in other rules but whose confidence is lower.

2.3.1.5 Biological evaluation

A rule is significant in biological point of view if and only if that rule is a significant set of genes which share some biological pathways. Therefore, the association rules must be evaluated in terms of biological perspective. To begin with, the consequent rules with the results from previous clustering methods on the similar gene expression data sets were compared [17][24]. Some rules confirmed the prior biological knowledge, others were considered as hypotheses for future investigation. It is noticed that the biological evaluation phase focused on rule by rule analysis not in the point of the entire rules investigation.

In practical, since mining association rules is limited by some threshold such as minimum support threshold, minimum confidence threshold, and thresholds used in discretization procedure, varying these thresholds is a technique to find the proper threshold in a particular data set [24]. It is noted that the lower the support threshold is set, the more the number of rules is generated. Thus, if the support threshold is set too low, the algorithms are unable to extract the rules because of memory bloat.

Although it is widely accepted that association rule mining is one of the efficient data mining techniques to disclose the intrinsic information in gene expression data, many thresholds in the mining association rules algorithms must be defined by

users, especially support and confidence threshold. The users should vary these thresholds until the result meet their satisfaction. Since most of the medical data sets naturally rare cases, the minimum support threshold may sometimes be set less than 10% of data [18]. However, the suitable thresholds depend on the size and the complexity of the data set where the latter factor we cannot measure.

2.3.2 The application of association rules on gene expression data

Association rules can usually unveil the information in the data set even with little prior knowledge. In the beginning, the research of C. Creighton and S. Hanash [17] and the research of C. Becquet et al. [24] showed that the gene-expression association rules mining is a promising complementary technique of previous prevailing clustering algorithms. They cited that a gene may normally related to others genes more than one biological pathways. Hence, the clustering algorithm cannot cover this event. On the contrary, this phenomenon can be represented by association rules. In other words, a gene can exist in many rules but belong to only one cluster. Moreover, C. Creighton and S. Hanash [17] generated randomized data set to validate that the mined rules were not occurred by chance, since they found only a few rules relative to the real data set.

The large number of generated rules is challenging issue; therefore, many following studies have developed the techniques in finding frequent patterns stage to alleviate the explosion of the rules. Those techniques are described in section 2.3.1.

Mostly, the purpose of mining gene expression association rules is to study how the expression of a gene is associated with those of other genes. Thus, the rules mined in many research works are always in the form of $X \rightarrow Y$, where X and Y are the set of genes with their expression levels (over-expressed or under-expressed). Some of the researchers applied certain constraints in rules pruning phase (see section 2.3.1). The examples of association rules in gene expression data are following:

- {ORT1} $\rightarrow$ {ADH5, ARG4, CTF13, …} [17] means that when gene ORT1 is over-expressed, the set of genes in the right-hand side of the rule are over-regulated as well.
- {ARO3} $\rightarrow$ {ARG1,ARG4,CTF13,HIS5,LYS1, …} [17] means that when gene ARO3 is over-expressed, the set of genes in the right-hand side of the rule are over-regulated as well.
- $\{\overline{\text{ESC8}}\} \rightarrow \{\overline{\text{IMD1}}, \overline{\text{IMD2}}\}$ [21] means that when gene ESC8 is under-expressed, the set of genes in the right-hand side of the rule are under-regulated as well.

2.3.3 The application of association rules on gene expression data integrated with external biological information

In the previous association rules mining on gene expression data analysis, biological information has been used only in the posterior procedure to evaluate the association of the expression levels among genes. Nevertheless, not only has gene expression data been collected nowadays, other molecular information such as biological process, cellular component, and molecular function has been also published by more than 300 journals [13]. Like gene expression databases, these biological data sources are accessible through online databases like Gene Ontology (GO) [26] database and many databases provided by National Center for Biotechnology Information (NCBI) [27]. Moreover, gene expression is theoretically related to other biological factors. Therefore, to fully understand molecular function, these external information should be integrated with gene expression data in mining association rules phase to analyze simultaneously. Besides, some experiments studied gene expression in various time points called temporal gene expression analysis or under biological conditions such as Heat shock, Sporulation, and Diauxic shift. These include many studies [18], [25] and [28]. However, such external information was selected from the literatures in the past.

There are two forms of rules. One is in form of biological features → gene expression, the other is gene expression → biological features. The examples of association rules in gene expression data are following:

- {Citrate cycle (TCA cycle)} → {[+] T6, [+] T7} [18] means that genes involved in Citrate -cycle are overexpressed at time point 6 and 7.
- {PIR3_up, PIR1_up} → {(21 minutes) HTB2_up} [25] means that when PIR3 and PIR1 are over-expressed, HTB2 is subsequently over-regulated in 21 minutes.
- {pr:RAP1, pr:FHL1} → {heat3↓} [28] means that genes controlled by regulators of ribosome pathways RAP1 and FHL1 are under-expressed in the heat shock process at time point 3
- {heat3↓, heat4↓, heat5↓} → {go:0006412 (translation)} [28] means that down-regulated genes under the heat shock process at time point 3, 4, and 5 are involved in translational process.

2.3.4 The application of association rules on gene expression data in cancer

It is widely know that the expression levels of some genes in cancer cells differ from those in normal adjacent cells. Thus, these genes are used as marker genes or signature genes to classify cancer cells from normal cells. When over-expressed or under-expressed genes are able to be located, this outcome can help biologists diagnose various types of cancer.

The research works [19][20] developed the performance of association rules mining algorithm and generated the rules in form of X → C, where X was a set of genes with their expression level and C is a class label representing whether the cell was cancerous or not. Later on, they grouped rules for classification purpose.

Furthermore, K. R. Seeja *et al.* [12] studied gene expression data in pancreas cancer cells by mining X → Y rules, where both X and Y are a set of genes

with their expression levels. These rules showed co-expressed signature genes in pancreas cancer cells. The usefulness of mining such rules is to improve better therapeutics.

Recently, F. J. Lopez used heterogeneous sources in their analysis in breast cancer [22]. They integrated prognostic factors such as ki67 (proliferation rate), metastasis, and tumor stage with gene expression data in association rules mining stage. The goal of this work is to discover the relationship of these two types of data. The generated rules from this experiment are in form of X $\rightarrow$ Y, where X and Y are an attribute in either gene expression data or prognostic factors. Specifically, this study was interested in rules whose both sides contained only one item.

The examples of association rules on gene expression data in cancers are following:

- {IBRDC2 [+]} $\rightarrow$ {AASS [+], FLJ20160 [-], … } [12] means that when IBRDC2 is over-expressed, the group of genes in the right hand side of the rule is also regulated as shown, where [+], [-] is over-expressed and under-expressed respectively.
- {GREB1 = under} $\rightarrow$ {ki67 = +} [22] means that while gene GREB1 is under-regulated, the proliferation rate (ki67) is likely to increase.

2.3.4 The studies of LINE-1 in cancer

Both gene-specific hypermethylation and global hypomethylation are common and crucial epigenetic events in cancers. Unlike gene-specific hypermethylation, genome-wide hypomethylation has been unclear for years. Still, It is known that most of the loss of DNA methylation is found on repetitive sequences including LINE-1 [2]. Since LINE-1s are widely interspersed in human genome (about 17% of the whole genome), LINE-1s are used to study genome-wide methylation in cancers [29]. K. Chalitchagorn *et al.* [30] analyzed the methylation levels on LINE-1 in many types of cancer: colon, bladder, head and neck, liver, lung, renal, prostate, breast,

esophagus, thyroid, and stomach cancer. They found that DNA methylation in cancer cells is significantly lower than that in normal cell counterparts (p-value < 0.01).

Recently, C. Aporntewan *et al.* [3] analyzed gene expression data in a variety of malignant tumors of stomach, breast, liver, lung, cervix, head and neck squamous cells, prostrate, and bladder. The results showed that not only are genes with LINE-1 hypomethylated, but such genes are also likely to down-regulated and the degree of down-regulation based on the level of hypomethylation on LINE-1. However, the mechanism how LINE-1s control gene expression in cancer remains in question. Therefore, the questionable mechanism of LINE-1 to regulate gene expression is the motivation of this thesis.

CHAPTER III

Methodology

This chapter describes the materials used in the experiments and demonstrates the methods applied in the analysis.

3.1 Datasets

Four data sources used in this study are following:

3.1.1 Gene information database

This data stores the general information of all genes in human genome. The important attributes used in this study are shown in the Table 3.1.

Table 3.1 The important attributes from gene information database

Attribute names	Description
GeneID	The unique identifier for a gene
Symbol	The default symbol for the gene
Chromosome	The chromosome on which this gene is placed

3.1.2 Gene reference sequence database

This data stores the comprehensive information of all genes in human genome. The important attributes used in this study are shown in the Table 3.2.

Table 3.2 The important attributes from gene reference sequence database

Attribute names	Description
GeneID	The unique identifier for a gene
start_position_on_the_genomic_accession	The start position of the gene
end_position_on_the_genomic_accession	The end position of the gene
Orientation	The orientation of the gene

3.1.3 Gene Expression Omnibus (GEO) datasets database

GEO datasets stores many kinds of high-throughput functional genomic data submitted by the scientific community. From this database, two types of data were used in the study. One kind of data was microarray platforms data. This data defines a list of probes along with what set of genes may be detected by these probes. Each platform data is assigned by a unique identifier with a “GPL” prefix. For example, Table 3.3 shows a part of the platform data GPL570. The first column presents probe identifiers (ID) and the second column presents the set of genes (Gene Symbol) detected by those probes in the first column belong. From the Table 3.3, probe ID “1552277_a_at” and “1552303_a_at” are corresponding to multiple genes, so these probes are called heterogeneous probes. On the contrary, probes which belongs to a single gene such as “1007_s_at”, “1552286_at”, and “1552289_a_at” are called homogeneous probes.

Table 3.3 A part of microarray platform annotation data (GPL570)

ID	Gene Symbol
1007_s_at	DDR1
1552277_a_at	C9orf30 /// TMEFF1
1552286_at	ATP6V1E2
1552289_a_at	CILP2
1552303_a_at	FLJ77644 /// TMEM106A
206432_at	HAS2
206433_s_at	SPOCK3
206434_at	SPOCK3

The other kind of data was gene expression series data. This data contains gene expression levels of samples made up for an experiment, where a gene expression level is corresponding to a reference probe in a reference platform. Each gene expression dataset is assigned by a unique identifier with a “GSE” prefix. For

example, Table 3.4 shows a part of the gene expression data of bladder cancer (GSE3167) generated by the GPL570 platform. The first row presents sample or subject identifiers which begin with "GSM" prefix, the first column presents probe identifiers (ID_REF) in the reference platform GPL570 and the $value_{ij}$ in the table is a gene expression level of the subject in j^{th} column and probe in i^{th} row.

Table 3.4 A part of bladder cancer gene expression data (GSE3167)

ID_REF	GSM134899	GSM134901	GSM134902	GSM134904	GSM134906
1007_s_at	2791.1	4428.5	2440.6	3178	3508.1
1552277_a_at	1495.1	2840.6	1780.6	806	777
1552286_at	346.4	565.8	238	297.8	332.2
1552289_a_at	109.9	84	88.3	107.8	173.3
1552303_a_at	48	56.1	40.9	19.9	52.1
206432_at	17.2	36.9	31.1	299.9	220.3
206433_s_at	16.2	18.4	24.3	4.6	18.7
206434_at	16.5	11.8	12.8	18.3	22

The GSE datasets, including the reference GPL data of each dataset, analyzed in the study with the lists of sample identifiers, where each is assigned by a unique identifier with a "GSM" prefix, used as controls and tests of each GSE dataset.

3.1.4 LINE-1 characteristics database (L1Base)

The L1Base database stores the characteristics of 11,901 putatively active human long interspersed nuclear element-1s (LINE-1s or L1s), which are categorized into three main types: intact in the two ORFs, full length L1s (FLI-L1s), L1s with intact ORF2 but disrupted ORF1 (ORF2-L1s), and full length (>6000bp) non-intact L1s (FLnI-L1s). The number of each category is shown in the Table 3.5. In addition, the description of each LINE-1 characteristics is the in Appendix A.

Table 3.5 LINE-1 classification

LINE-1 classification (based on L1Base online database)	The number of LINE-1s
Human Full-Length, Intact LINE-1 Elements [FLI-L1]	145
Human ORF2 Intact LINE-1 Elements [ORF2-L1]	103
Human Full-Length >4500nt LINE-1 Elements [FLnI-L1]	11,653
Total	**11,901**

The four sources of each database described above are shown in the Table 3.6.

Table 3.6 The references of the databases used in the study

Types of database	References
Gene information	ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/ Mammalia/Homo_sapiens.gene_info.gz
Gene reference sequence	ftp://ftp.ncbi.nih.gov/gene/DATA/gene2refseq.gz
GEO datasets	http://www.ncbi.nlm.nih.gov/geo
L1 characteristics	http://l1base.molgen.mpg.de

3.2 Computer specification and tools

In this study, we performed the experiments on an operating system Windows 7 Professional 64-bit on a HP Z800 workstation with dual Intel Xeon E5520 2.26 GHz and 16 GB of memory. For all codes, we implemented on C# programming language in Microsoft Visual Studio 2010, an integrated development environment (IDE). Besides, we utilized the following tools:

- **Microsoft SQL Database Server 2008 Express:** A free database management system
- **R Statistics (version 2.12.0):** A free statistical software
- **RapidMiner (64-bit, version 4.6):** An open-source data mining software

3.3 The methods in classification association rules mining

There were four main parts in this analysis: preprocessing, bivariate data analysis using statistical methods, multivariate data analysis where decision tree and classification association rules mining were applied. (see Figure 3.1).

<table>
<tr><td colspan="3">Preprocessing</td></tr>
<tr><td colspan="3">Constructing the table containing genes with LINE-1
Discretizing gene expression level of each gene.
Constructing the final two dimensional table</td></tr>
<tr><td colspan="3">Bivariate data analysis</td></tr>
<tr><td>Nominal LINE-1 characteristics
Chi-square test</td><td colspan="2">Numeric LINE-1 characteristics
Bivariate logistic regression</td></tr>
<tr><td colspan="3">Decision tree mining</td></tr>
<tr><td colspan="3">C4.5 algorithm + 10 folds
— Information gain
— Gain ratio</td></tr>
<tr><td colspan="3">Classification association rules mining</td></tr>
<tr><td>Discretization
Binominal transformation
Feature selection</td><td>Pre-processing</td><td></td></tr>
<tr><td>Frequent patterns mining (FP-growth)
Rules generation</td><td>Processing</td><td></td></tr>
<tr><td>Rules filtering
Biological evaluation</td><td>Post-processing</td><td></td></tr>
</table>

Figure 3.1 The overview of methodology in this study

3.3.1 Preprocessing

In this phase, necessary data from various sources was downloaded from the four online databases as text files and imported to our own single database. Moreover, certain data values were transformed. Ultimately, the final output of this process was a single two dimensional table where its rows were a Cartesian product of a set from the multiplication of a set of experimental genes in GSE data set and a set of LINE-1s in genes and its columns comprised gene symbols, LINE-1 characteristics, and gene expression levels (either down-regulated or not down-regulated) as class labels (see Figure 3.2). It should be noted that some rows had the similar gene symbol because that gene contains multiple LINE-1s. Still, these rows with the similar gene symbol had the similar gene expression level.

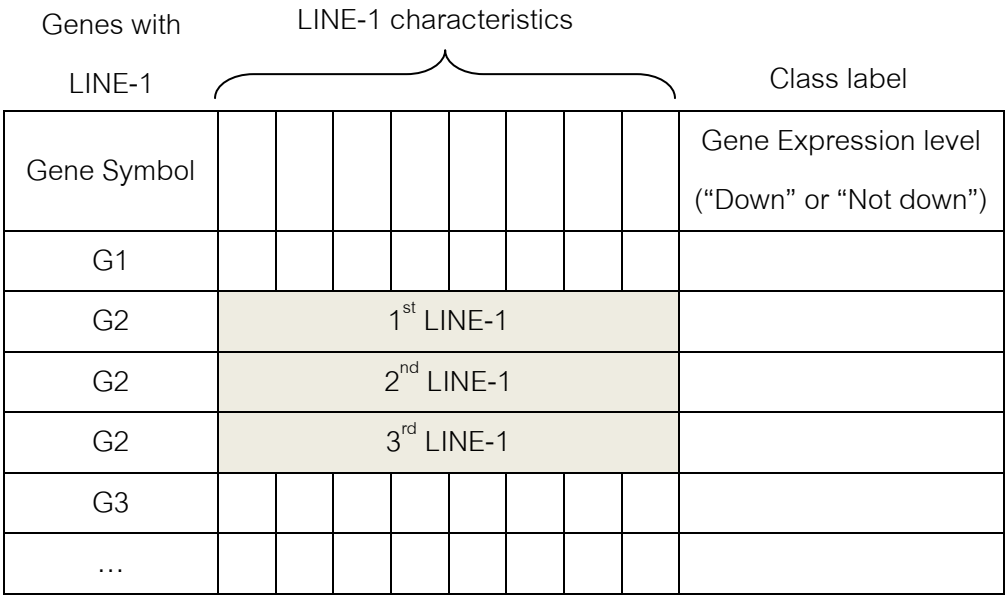


Figure 3.2 The structure of the final table before performing analysis

To obtain the final table, the following three steps were required:

3.3.1.1 Constructing the table containing genes with LINE-1

Due to a variety of the data sources and the intricacy to manage the data in text files, all relevant files (see Section 3.1) were first imported into the database

management system. To discover genes with LINE-1 on the entire genome, the start and end positions of each gene and each LINE-1 from gene reference sequence data and L1Base data respectively were compared because a gene with LINE-1must overlap at least a LINE-1.

After what genes had LINE-1s was known, a table containing these genes was then constructed. In addition to gene symbols with their corresponding LINE-1 identifiers (LINE-1 IDs), the table consisted of the orientations of these genes which were derived from the gene reference sequence data. Table 3.7 shows the structure of such table. It is noticed that a gene may have multiple LINE-1s. Similarly, a LINE-1 may lie over multiple genes.

Table 3.7 A part of the table containing genes with LINE-1 and gene orientations

Gene Symbol	Orientation of the gene	LINE-1 ID
LEPR	+	831
LEPR	+	832
CFH	+	898
COL24A1	-	828
ST6GAL2	-	1484
UGT1A3	+	1636
UGT1A4	+	1636

Next, this table was used as a reference table for every GSE data set to prepare the final two-dimensional table which comprised the symbols of genes having LINE-1, the characteristics of LINE-1s lying on those genes and the expression levels of those genes.

3.3.1.2 Discretizing gene expression level of each gene

Before determining whether a gene is either down-regulated or not, gene expression levels of all probes in each GSE data set was first categorized. To discretize gene expression level of each probe, the statistical method t-test was performed to

compare the difference of means between the control groups and the experimental groups of each GSE data set. The controls and tests of each GSE data sets are relied on the earlier work [3]. For any probe, if the means of both groups were significantly different and the mean of the experimental group was less than that of the control group, the gene expression of such probe was considered down-expressed or down-regulated. Otherwise, its gene expression was treated as not down-regulated.

After the discretization of gene expression had been achieved in each probe, what genes were down-regulated or not was addressed. Since a probe can measure gene expression belonging to multiple genes, the counting technique called simple count was performed. By definition, a gene is down-regulated if and only if it has at least a unique probe, a probe belongs to a single gene, is significantly down-regulated or at least two homology probes are significantly down-regulated; otherwise, the gene is considered not down-regulated. Here, the expression levels of each gene in every GSE dataset were already discretized into only two values: down-regulated and not down-regulated.

3.3.1.3 Constructing the final two dimensional table

In this stage, the final two dimensional table of each GSE dataset was created by integrating the following data. Firstly, the table containing genes having LINE-1s and gene orientations from section 3.3.1.1. Secondly, the GSE dataset containing gene expression levels (GE level), which had been discretized from 3.3.1.2. Finally, the LINE-1 characteristics data from L1Base. Every column in the three data sources was merged by matching corresponding gene symbols for the first two data sources and corresponding LINE-1 identifiers (LINE-1 ID) for the last two data sources. During constructing the final table, a couple of LINE-1 characteristics, the number of LINE-1 and the orientation of each gene, were included. Figure 3.3 is an integration diagram of these three data sources, GSE3167, genes having LINE-1s and their orientations table, and L1Base, to make the final table of GSE3167 (see Table 3.8). It was noted that gene symbols “ST6GAL2” , “COL24A1”, and “ZZZ3” were not included in the final table, since these genes were not in GSE3167 or did not contain any LINE-1. Similarly, LINE-1

identifiers “508”, “509”, and “512” were not included in the final. Lastly, the final table of every GSE dataset had 57 columns including gene expression level. The total number of rows, nevertheless, was varied by GSE datasets.

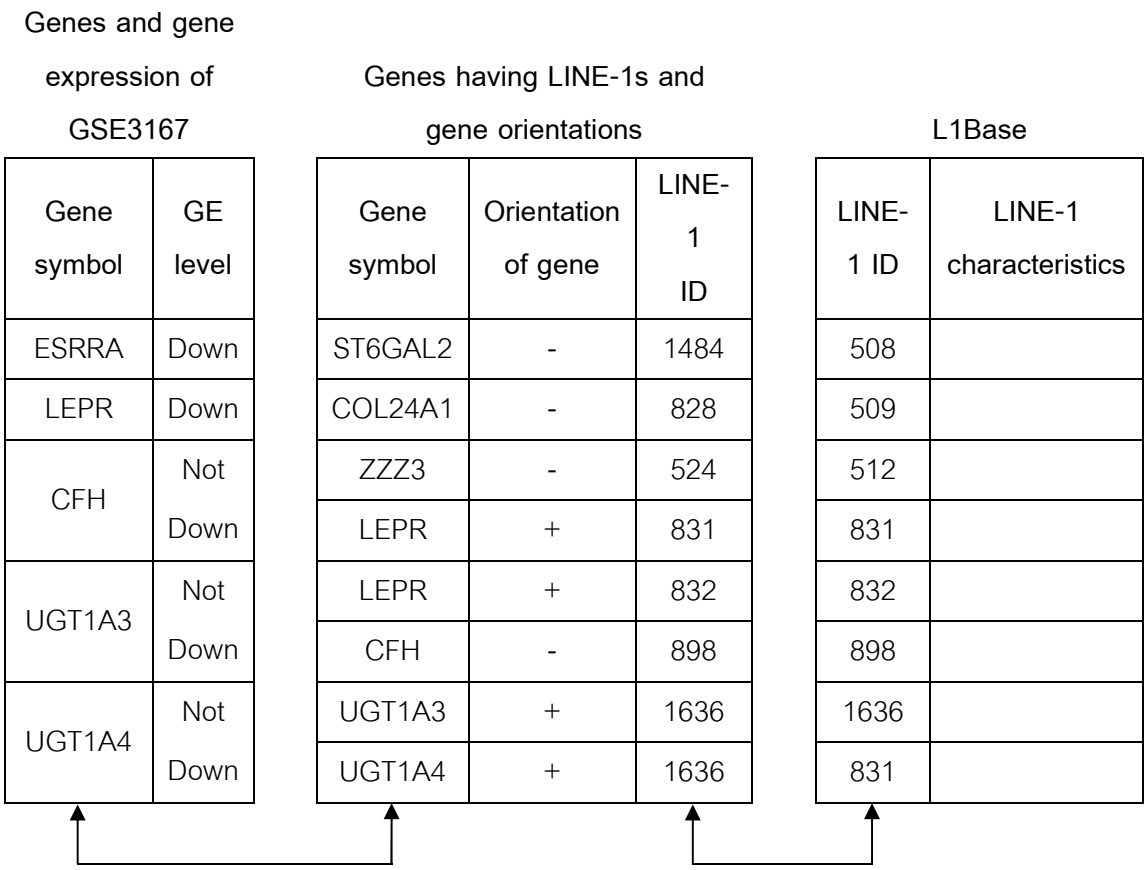


Figure 3.3 The diagram to combine three data sources to make the final table

Table 3.8 A part of the final two dimensional table of bladder cancer dataset (GSE3167)

Gene symbol	LINE-1 ID	LINE-1 characteristics	Orientation of gene	The number of LINE-1	GE level
LEPR	831	…	+	2	Down
LEPR	832	…	+	2	Not Down
CFH	898	…	-	1	Not Down
UGT1A3	1636	…	+	1	Not Down
UGT1A4	1636	…	+	1	Not Down

3.3.2 Bivariate data analysis

After the final table had been completed, the analysis of the association between gene expression data and LINE-1 characteristics is performed. In this stage, only a LINE-1 characteristic was analyzed at a time. First, the distribution of every LINE-1 characteristic was observed. Next, chi-square test was applied for every nominal LINE-1 characteristic. Table 3.9 shows the contingency table created to perform chi-square test. Each entry in the contingency table was the number of corresponding records in the final table built from Section 3.3.1.

Table 3.9 The structure of a contingency table of the association between each nominal LINE-1 characteristic and gene expression (either down regulation or not)

	A LINE-1 characteristic			
	Value 1	Value 2	…	Value n
Down				
Not down				

For every numeric LINE-1 characteristic, bivariate logistic regression was employed. Finally, p-values derived from the statistical tests were considered whether the association between gene expression and any LINE-1 characteristic was statistically significant or not.

3.3.3 Multivariate data analysis by decision tree mining

For multivariate data analysis, in the beginning, we analyzed LINE-1 characteristics by applying a basic decision tree mining technique, C4.5 algorithm. We performed a couple of experiments with different attribute selection measures. We used information gain for one experiment and gain ratio for the other experiment. In both experiments, ten-fold cross validation was applied. Then, we compared the results from both attribute selection measures.

3.3.4 Multivariate data analysis by classification association rules mining

This phase was the core of the thesis. We performed classification association rules mining to search the association between LINE-1 characteristics and down regulation of genes containing LINE-1.

To begin with, further data preparations such as discretization, transformation, and feature selection were required in this stage despite data preprocessing in Section 3.3.1. Next, frequent patterns were mined by using FP-Growth algorithm and classification association rules were generated. After that, redundant rules were filtered out. Finally, LINE-1 characteristics were scored and biological evaluation was discussed. The following sections below describe each step in details.

3.3.4.1 Discretization

We applied two methods to discretize the numeric L1 characteristics. One was equal frequency bins where we used the median of the attribute as a threshold to divides its tuples into two groups. In this sense, we presumed that the down-regulation may be relevant the values either more or less than the median with equal probability.

3.3.4.2 Binominal transformation

Due to the data structure of FP-tree, every attributes needs to have only two values or binominal. Therefore, each LINE-1 characteristic with more than two values was transformed by using both characteristic's name and its value to form a new characteristic. For example, Type, one of LINE-1 characteristics, has three values: FLnI_L1, FLI_L1, and ORF2_L1. This characteristic would be transformed to three characteristics, Type= FLnI_L1, Type=FLI_L1, and ORF2_L1. Each new characteristic has only two values, true and false.

3.3.4.3 Feature selection

Sometimes, interesting patterns or rules are rarely appear in database. In this sense, these patterns or rules have low support but high confidence. Therefore, In association rules mining, whether these rare rules are discovered or not depends on

minimum support threshold. Supposed that the minimum support threshold is set to zero, every possible pattern including rare patterns will completely be found. However, the minimum support threshold cannot be set to zero in a huge data because of the constraint of time and memory space. The more the attributes there are in association rule mining, the more the time and space are consumed. Thus, to minimize support threshold, some features or attributes need to be cut off. However, it should be trade-off between how low support threshold should be set and how many attributes there should be remained in frequent patterns mining.

In this study, logistic regression method was performed to select some LINE-1 characteristics to be used in the next frequent patterns mining stage. This statistical method satisfied the data in this study, since it could predict the probability of a binominal dependent variable like whether gene expression would be down or not. Therefore, the statistical model formed in logistic regression were composed of all LINE-1 characteristics as independent variables and gene expression data as a dependent variable. However, logistic regression only works on metric independent variables. Therefore, all of the nominal LINE-1 characteristics must have been dummy coded. In dummy coding task, we created new metric n+1 characteristics whose values were either 0 or 1, from any characteristic with n values. For example, Type, one of LINE-1 characteristics, has three values: FLnl_L1, FLI_L1, and ORF2_L1. We built new four characteristics as shown in the Table 3.10.

Table 3.10 An example of dummy coding of “Type”, one of nominal LINE-1 characteristics

Original characteristic	New characteristics after dummy coding	
Type	Type1	Type2
FLnl_L1	1	0
FLI_L1	0	1
ORF2_L1	0	0

Here, we used R Statistics to perform logistic regression. This tool showed the outcome model with the coefficient of each LINE-1 characteristic and its p-value. Then we ranked these characteristics by their p-values and used p-value of 0.5 as a threshold. So, if the p-values any LINE-1 characteristics coefficients were lower than 0.5, these characteristics were deleted. Finally, only the remained LINE-1 characteristics were analyzed in the next stage.

3.3.4.4 Frequent patterns mining

In this stage, RapidMiner version 4.6 was employed to mine frequent patterns by performing FP-growth algorithm. The minimum support threshold was set as low as possible. Although FP-growth algorithm used a compact tree data structure to keep all frequent patterns, the minimum support threshold could not be set low enough for creating rules with high confidence because of the limitation of memory space. Therefore, another way to minimize the overall minimum support threshold used in this analysis was to find local frequent patterns. Specifically, only records having down regulation were mined for frequent patterns.

To prove that this method could minimize the minimum support threshold, let n is the number of records having down regulation and N is the number of the total records. If we set the local minimum support threshold to s, the actual minimum support threshold is $(s \times n) / N$. Since, in this study, the number of records with down regulation (n) was much smaller than a half of the total number of records (N) in every dataset, the actual minimum support threshold could be dramatically minimized. With this advantage, only down regulation records were mined for generating local frequent patterns in this stage. In addition, “gene expression level” column would be dropped before mining local frequent patterns, since every record had the similar gene expression value (Down). Nonetheless, it is noted that not every local frequent patterns is frequent with respect to the entire data. Therefore, the whole data, both down and not down regulation records, were considered together when computing actual support and confidence of rules in rules generation stage.

3.3.4.5 Rules generation

After all of the frequent patterns were acquired, every possible classification association rule with LINE-1 characteristics on the left-hand side and only a gene expression level that was down regulation (Down) on the right-hand side of the rule were generated (see Figure 3.4).

$$C_1 = V_1 \; and \; C_2 = V_2 \; and \; C_3 = V_3 \; and \; \ldots \; and \; C_n = V_n \rightarrow Down$$

where C_i is the i^{th} LINE-1 characteristic's name

V_i is the i^{th} LINE-1 characteristic's value

Figure 3.4 The structure of generated rules

Next, actual support and confidence of each rule were calculated from a given entire dataset. In addition, the 2×2 contingency table showing the association between antecedence and consequence of rule was built (see Table 3.11). Each entries were the number of the corresponding records in a given final table. With the 2×2 contingency table, chi-square test was applied on every rule. Finally, p-value of each rule from chi-square test was adjusted by Bonferroni and Yate's correction method. Besides, odds ratio of each rule was computed from the 2×2 contingency table.

Table 3.11 The structure of a 2×2 contingency table of the association the antecedence and the consequence of a rule

	Consistent to LHS	Contradictory to LHS
Down		
Not down		

3.3.4.6 Rules filtering

Rules with p-values from chi-square test less than 0.05 or rules with confidence lower than 50% were removed in this stage. It is noted that if there is at least one

expected value in the 2×2 contingency table is less than 5, we would adjust p-value by Yate's correction. In addition, some rules generated from the previous section were regarded as redundant rules. By definition, a rule $X_i \rightarrow Y$ is a redundant rule if and only if there is a rule $X_j \rightarrow Y$, where X_j is a subset of X_i and the confidence of the rule $X_j \rightarrow Y$ is no less than the confidence of the rule $X_i \rightarrow Y$. Therefore, these redundant rules were detected and deleted in this stage.

Yet, the number of the remaining rules was too large to be explored rule by rule. Therefore, we would select certain rules based on genes they support, considering three factors, high confidence, high support and small size of LINE-1 characteristics used in the left hand side of rule. By this manner, we can better understand the biology represented by rules without the loss of the overall support.

To begin with, we categorized rules based on genes that they supported. In case of gene with a unique rule, we keep such rule.

Turning to consider genes with multiple rules, we would select the rules with the highest confidence percentage. Next, we focused on both support percentage and the number of terms in the left hand side of rule. For each rule A in the same group, if there is a rule B which satisfies either of the following two conditions (see Figure 3.5), then we leave rule A out of our consideration. Finally, we would have a smaller group of representative rules from each gene to analyze in terms of biology.

If there is rule B in the same group such that

- support(rule B) > support(rule A) and LHS(rule B) ≤ LHS(rule A)

or

- support(rule B) = support(rule A) and LHS(rule B) < LHS(rule A)

where

support(rule) = the percentage of support of rule

LHS(rule) = the number of terms or L1 characteristics in the left hand side of rule

Figure 3.5 The conditions of rule selecting for biological purpose

3.3.4.7 Biological evaluation

After we had obtained all classification association rules of every GSE dataset and clusters of each set of rules had been created, we interpreted these rules from biological point of view on how they supported the literature or introduced new hypotheses.

CHAPTER IV

Results and Discussion

After analysis on LINE-1 characteristics that mediate gene expression in many kinds of cancer, chi-square test pointed that some LINE-1 characteristics were significantly down-regulated gene expression at p-value = 0.05. Moreover, C4.5 produced such a large tree of each type of cancer that was hard to interpreted. Finally, classification association rules mined from five datasets, prostate, bladder, head and neck, and liver cancer, including 5-AZA data supported the mechanism how LINE-1s regulate gene expression in cancer cells by LINE-1 transcripts [3]. In addition, some rules could be used as new hypotheses of down regulation.

In this chapter, the results from three parts of data analysis, bivariate statistical testing, decision tree mining, and classification association rules mining, are demonstrated and discussed in details. Besides, the interesting LINE-1 characteristics in rules are explored in terms of basic biological point of view.

4.1 The statistical results from bivariate data analysis

Before bivariate data analysis, we observed each data set by creating the 2×2 contingency table to roughly understand the association between the existence of LINE-1 in genes and gene expression (see Table 4.1). Even though some datasets whose the association between having LINE-1 in genes and gene expression are not statistically significant at p-value 0.05, we still performed data mining on those datasets. In other words, having LINE-1 might not enough to indicate the associated gene expression but LINE-1 characteristics were still likely to be statistically associated with gene expression. However, GSE14054 (si-AGO2IP) had not been analyzed since the number of down-regulated genes was not enough for statistical test or mining.

For bivariate data analysis, after performing the statistical tests on each LINE-1 characteristic described in the previous chapter, we found that certain LINE-1

characteristics were statistically significant (p-value < 0.05). Different data sets had different significant characteristics (see Table 4.2 and Table 4.4)

Table 4.1 The association between the existence of LINE-1 in genes and gene expression in each dataset on the 2×2 contingency table

			Gene Expression				
	Dataset		Down	Not Down	Total	p-value	Odds
Cancers	GSE6919 Prostate	L1	94	566	660	0.9853	0.9979
		No L1	1,182	7,102	8,284		
		Total	1,276	7,668	8,944		
	GSE3167 Bladder	L1	377	572	949	**1.08E-14**	1.7007
		No L1	3,382	8,727	12,109		
		Total	3,759	9,299	13,058		
	GSE5816 Lung	L1	121	1,366	1,487	**0.0002**	1.4517
		No L1	1,086	17,798	18,884		
		Total	1,207	19,164	20,371		
	GSE6631 Head & neck	L1	46	614	687	0.1027	1.2981
		No L1	452	7,832	8,284		
		Total	498	8,473	8,971		
	GSE13911 Stomach	L1	359	1,128	1,487	**3.56E-12**	1.5517
		No L1	3,214	15,670	18,884		
		Total	3,573	16,798	20,371		
	GSE14811 Liver	L1	51	303	354	0.5326	1.1020
		No L1	857	5,611	6,468		
		Total	908	5,914	6,822		
	GSE1299 Breast	L1	66	773	839	**4.95E-05**	1.7376
		No L1	427	8,240	8,667		
		Total	493	9,013	9,506		

Table 4.1 (Continued)

			Gene Expression				
	Dataset		Down	Not Down	Total	p-value	Odds
Cancer	GSE9750 Cervical	L1	257	692	949	**8.14E-16**	1.8439
		No L1	2,030	10,079	12,109		
		Total	2,287	10,771	13,058		
	GSE5764 Breast	L1	12	1,475	1,487	0.3652	1.3163
		No L1	116	18,768	18,884		
		Total	128	20,243	20,371		
5-Aza	GSE9764 5-Aza	L1	65	1,422	1,487	**9.18E-06**	1.8066
		No L1	466	18,418	18,884		
		Total	531	19,840	20,371		
	GSE5816 hBEC (Lung)	L1	15	1,472	1,487	0.6393	1.1352
		No L1	168	18,716	18,884		
		Total	183	20,188	20,371		
AGO2	GSE4246 AGO2sh	L1	90	760	850	0.5229	0.9289
		No L1	1,222	9,585	10,807		
		Total	1,312	10,345	11,657		
	GSE14537 AGO2IP	L1	67	1,420	1,487	0.6133	1.0680
		No L1	799	18,085	18,884		
		Total	866	19,505	20,371		
	GSE14054 si-AGO2IP	L1	2	1,485	1,487	0.4237	1.8153
		No L1	14	18,870	18,884		
		Total	16	20,355	20,371		

Table 4.2 The p-value from chi-square test with the odds ratio

(The bold entry indicates that the characteristic in the consistent row is significant at p-value = 0.05)

LINE-1 PART	LINE-1 Characteristics	Cancers																	
		GSE6919 prostate		GSE3167 bladder		GSE5816 lung		GSE6631 Head&neck		GSE13911 stomach		GSE14811 liver		GSE1299 breast		GSE5764 breast		GSE9750 cervical	
		Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value
5' UTR	Runx3 Site	0.915	0.659	1.112	0.363	0.987	0.940	0.979	0.941	1.099	0.377	0.436	**0.017**	1.473	0.098	1.976	0.181	0.978	0.863
	Runx3 ASP	0.950	0.783	0.960	0.711	0.727	0.074	0.784	0.397	1.201	0.067	0.475	**0.015**	1.365	0.165	2.556	0.053	1.078	0.540
	SRY Site1	1.153	0.417	1.020	0.847	0.910	0.536	0.844	0.502	1.051	0.596	0.795	0.350	0.905	0.644	1.469	0.474	0.890	0.304
	SRY Site2	0.813	0.237	1.227	**0.046**	1.043	0.788	0.982	0.942	0.928	0.432	0.691	0.156	1.418	0.103	1.671	0.301	1.036	0.753
	YY1 BoxA+BoxA	1.013	0.952	1.035	0.780	1.209	0.289	0.947	0.863	0.964	0.748	0.436	**0.017**	1.350	0.222	1.321	0.630	1.098	0.494
	TF nkx-2.5	1.229	0.253	1.135	0.222	1.075	0.642	1.255	0.400	1.020	0.838	0.647	0.076	1.021	0.924	1.836	0.286	1.089	0.459
	TF nkx-2.5B	1.061	0.841	1.284	0.166	0.547	0.080	0.821	0.678	0.888	0.491	0.566	0.285	2.212	**0.007**	1.633	0.514	1.189	0.376
ORF1	REKG235	1.103	0.595	0.961	0.711	1.243	0.185	1.048	0.863	0.986	0.883	0.756	0.273	0.911	0.678	1.578	0.427	0.970	0.795
	ARR260	1.629	**0.043**	0.998	0.989	1.530	**0.039**	1.011	0.973	1.035	0.758	0.771	0.364	1.737	0.078	1.163	0.814	1.095	0.514
	YPAKLS282	1.406	0.065	0.976	0.812	1.152	0.371	0.934	0.791	0.959	0.654	0.877	0.599	1.728	**0.023**	1.785	0.310	1.041	0.725

Table 4.2 (Continued)

LINE-1 PART	LINE-1 Characteristics	Cancers																	
		GSE6919 prostate		GSE3167 bladder		GSE5816 lung		GSE6631 head&neck		GSE13911 stomach		GSE14811 liver		GSE1299 breast		GSE5764 breast		GSE9750 cervical	
		Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value
ORF2	N14	1.201	0.584	0.938	0.740	1.385	0.307	0.717	0.421	1.174	0.369	1.750	0.359	0.819	0.584	NA	0.247	1.210	0.387
	E43	0.943	0.841	1.042	0.822	0.864	0.571	0.528	0.072	0.808	0.176	1.244	0.657	3.990	**0.038**	-1.000	0.225	0.665	**0.030**
	Y115	1.017	0.951	0.947	0.740	1.117	0.665	0.872	0.724	1.187	0.271	0.443	**0.026**	1.523	0.292	0.500	0.272	0.657	**0.013**
	D145	1.201	0.493	0.940	0.688	1.441	0.172	1.874	0.178	1.195	0.232	0.944	0.893	0.676	0.208	0.559	0.359	0.800	0.173
	N147	1.592	0.195	0.848	0.370	1.599	0.137	1.422	0.502	1.370	0.070	0.614	0.238	1.784	0.211	1.436	0.725	1.323	0.198
	T192	1.997	0.063	1.304	0.145	1.085	0.763	0.976	0.956	0.954	0.767	0.403	**0.006**	1.377	0.458	1.491	0.698	1.524	0.053
	D205	1.500	0.168	1.108	0.520	0.770	0.220	1.398	0.443	1.210	0.193	1.050	0.909	1.151	0.684	0.950	0.947	1.079	0.668
	SDH228	1.212	0.363	0.993	0.950	0.962	0.826	1.109	0.738	1.147	0.217	1.446	0.251	0.918	0.736	4.517	0.110	0.973	0.837
	R363	1.038	0.854	0.864	0.217	1.400	0.074	0.794	0.417	1.067	0.545	0.553	**0.027**	1.279	0.351	2.326	0.251	1.006	0.964
	FADD700	1.069	0.723	0.843	0.118	1.499	**0.022**	1.363	0.290	1.121	0.258	0.953	0.860	2.116	**0.006**	1.321	0.629	0.923	0.508
	HMKK1091	0.889	0.494	1.079	0.463	1.000	0.999	0.894	0.661	1.015	0.876	0.657	0.086	2.140	**0.002**	1.836	0.286	1.179	0.155
	SSS1096	1.103	0.568	0.903	0.313	1.070	0.657	0.947	0.829	0.977	0.799	0.803	0.370	1.178	0.450	1.321	0.590	0.979	0.850
	I1220	1.243	0.314	0.882	0.295	1.277	0.188	0.870	0.636	1.106	0.351	0.656	0.131	1.233	0.429	2.279	0.263	1.100	0.477
	S1259	1.061	0.806	1.126	0.390	1.124	0.577	0.888	0.728	1.072	0.573	0.796	0.455	1.547	0.181	3.025	0.260	1.241	0.170

Table 4.2 (Continued)

LINE-1 PART	LINE-1 Characteristics	Cancers																	
		GSE6919 prostate		GSE3167 bladder		GSE5816 lung		GSE6631 head&neck		GSE13911 stomach		GSE14811 liver		GSE1299 breast		GSE5764 breast		GSE9750 cervical	
		Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value
3' UTR	PolyA Signal	1.119	0.518	0.916	0.405	0.790	0.155	1.054	0.840	0.976	0.804	0.864	0.574	1.100	0.671	0.919	0.875	**0.779**	**0.037**
ALL	ORF StartStop	-	**0.017**	-	0.818	-	0.125	-	**0.036**	-	0.971	-	0.132	-	0.387	-	0.054	-	0.801
	Type	-	0.867	-	0.973	-	0.158	-	0.374	-	0.415	-	0.650	-	0.808	-	0.355	-	**0.029**
	Strand	0.886	0.473	1.044	0.669	1.100	0.527	1.468	0.131	0.981	0.837	1.424	0.148	1.143	0.531	0.588	0.300	0.895	0.316
	L1M/L1PA Discrimination	0.409	**0.050**	1.331	0.155	0.360	**0.021**	0.646	0.466	1.178	0.351	2.009	0.181	0.953	0.912	0.000	0.275	1.292	0.234
	orientation	1.192	0.299	1.247	**0.028**	1.001	0.995	3.113	**0.000**	1.223	**0.028**	0.946	0.819	0.798	0.292	1.373	0.529	0.848	0.139

Table 4.2 (Continued)

LINE-1 PART	LINE-1 Characteristics	5-aza				AGO2			
		GSE9764 5-aza		GSE5816 hBEC (Lung)		GSE4246 AGO2sh		GSE14537 AGO2IP	
		Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value
5' UTR	Runx3 Site	1.123	0.592	1.647	0.247	0.755	0.201	0.962	0.868
	Runx3 ASP	0.782	0.264	1.273	0.579	0.856	0.436	1.194	0.397
	SRY Site1	0.753	0.133	2.550	0.054	0.647	**0.013**	1.388	0.112
	SRY Site2	1.193	0.357	1.414	0.398	0.860	0.410	1.030	0.882
	YY1 BoxA+BoxA	1.124	0.608	1.041	0.936	0.670	0.089	1.171	0.497
	TF nkx-2.5	0.953	0.802	1.222	0.644	0.773	0.151	1.116	0.591
	TF nkx-2.5B	0.709	0.382	1.036	0.963	0.850	0.622	0.889	0.753
ORF1	REKG235	1.051	0.803	1.277	0.587	0.656	**0.019**	1.130	0.560
	ARR260	1.315	0.273	1.885	0.298	0.893	0.610	2.766	**0.002**
	YPAKLS282	0.816	0.287	2.269	0.095	0.789	0.185	1.342	0.161
ORF2	N14	1.153	0.703	1.925	0.515	0.807	0.504	1.464	0.368
	E43	1.113	0.763	0.637	0.464	0.594	0.068	2.489	0.067
	Y115	0.587	**0.038**	2.685	0.315	0.502	**0.005**	1.167	0.647
	D145	1.150	0.652	1.428	0.629	0.977	0.937	2.302	**0.043**
	N147	1.165	0.667	2.209	0.427	1.270	0.506	1.424	0.371
	T192	1.894	0.126	1.092	0.905	1.533	0.230	1.277	0.512
	D205	1.009	0.975	3.145	0.237	0.559	**0.022**	1.232	0.518
	SDH228	0.873	0.529	0.897	0.819	0.734	0.124	1.796	**0.031**
	R363	0.846	0.425	1.260	0.646	0.938	0.759	1.422	0.152
	FADD700	0.886	0.545	1.069	0.883	0.765	0.153	1.797	**0.014**
	HMKK1091	0.953	0.802	0.853	0.702	0.942	0.743	1.322	0.180
	SSS1096	0.871	0.462	1.322	0.508	0.881	0.473	0.984	0.936
	I1220	1.283	0.284	0.973	0.953	0.793	0.257	1.311	0.264
	S1259	0.924	0.748	1.409	0.578	0.954	0.843	1.669	0.095
3' UTR	PolyA Signal	0.846	0.418	0.831	0.681	1.127	0.517	0.861	0.481

Table 4.2 (Continued)

LINE-1 PART	LINE-1 Characteristics	5-aza				AGO2			
		GSE9764 5-aza		GSE5816 hBEC		GSE4246 AGO2sh		GSE14537 AGO2IP	
		Odds	p-value	Odds	p-value	Odds	p-value	Odds	p-value
ALL	ORF StartStop	-	0.242	-	0.267	-	0.075	-	0.432
	Type	-	0.214	-	0.739	-	0.507	-	0.524
	Strand	1.018	0.924	0.489	0.093	0.764	0.126	0.752	0.146
	L1M/L1PA Discrimination	0.476	0.140	0.613	0.629	1.442	0.277	0.640	0.333
	orientation	0.775	0.180	0.436	0.058	0.963	0.832	0.673	**0.046**

The bold values are significant at p-value = 0.05. In addition, we provides odds ratio for any LINE-1 characteristic with two values (conserved or mutated). If odds ratio is more than one, it means that conserved sequence have more effect on down-regulations of genes than mutated ones; on the contrary, if odds ratio is less than one, it means that mutated sequence have more effect on down-regulations of genes than conserved ones (see Table 4.3).

Table 4.3 The example of 2×2 contingency table of LINE-1 characteristics which are two-value sequences

	A LINE-1 characteristic (Sequences)	
	Conserved	Mutated
Down	A	B
Not Down	C	D
	Odds = AD/BC	

Table 4.4 The p-value from logistic regression

(The bold entry indicates that the characteristic in the corresponding row is significant at p-value = 0.05)

LINE-1 Part	LINE-1 Characteristics	Cancers									5-aza		AGO2	
		GSE6919 prostate	GSE3167 bladder	GSE5816 lung	GSE6631 Head&neck	GSE13911 Stomach	GSE14811 Liver	GSE1299 breast	GSE5764 breast	GSE9750 cervical	GSE9764 5-aza	GSE5816 hBEC Lung	GSE4246 AGO2sh	GSE14537 AGO2IP
ORF1	ORF1 Gaps	**0.006**	0.592	0.115	0.789	0.227	0.972	0.943	0.221	0.867	0.427	0.911	0.625	0.868
	ORF1 Frameshifts	0.259	0.234	0.337	0.578	0.358	0.247	0.098	**0.034**	0.679	0.146	0.981	0.646	**0.048**
	ORF1 Stops	0.192	0.285	0.089	0.558	0.754	0.639	0.231	0.165	0.434	0.339	0.309	0.068	0.402
	ORF1 %A	0.150	**0.031**	0.101	0.438	0.778	0.893	0.541	0.199	0.564	0.287	0.524	0.155	0.582
	ORF1 %T	0.073	0.127	**0.015**	0.500	0.435	0.080	0.298	0.071	0.530	0.664	0.440	**0.041**	**0.014**
	ORF1 CAI	0.053	0.484	0.145	0.602	0.643	0.912	0.234	0.035	0.096	0.957	0.291	0.762	0.774
ORF2	ORF2 Gaps	0.459	0.724	1.000	0.594	0.773	0.831	0.494	0.081	0.816	0.514	0.592	0.429	0.690
	ORF2 Frameshifts	0.127	0.267	0.282	0.944	0.722	0.105	0.489	0.186	0.257	0.479	0.369	0.335	0.235
	ORF2 Stops	0.101	0.260	**0.047**	0.878	0.387	0.186	0.338	**0.035**	0.410	0.324	0.444	0.087	**0.031**
	ORF2 %A	0.079	**0.047**	0.089	0.821	0.571	0.242	0.449	0.106	0.894	0.519	0.544	0.068	0.125
	ORF2 %T	**0.046**	0.108	**0.008**	0.719	0.670	0.337	0.459	0.122	0.101	0.249	0.618	0.014	**0.003**
	ORF2 CAI	0.241	0.121	**0.030**	0.823	0.888	0.700	0.757	0.122	0.269	0.354	0.085	0.442	0.172

Table 4.4 (Continued)

LINE-1 Part	LINE-1 Characteristics	Cancers									5-aza		AGO2	
		GSE6919 prostate	GSE3167 bladder	GSE5816 lung	GSE6631 head&neck	GSE13911 stomach	GSE14811 Liver	GSE1299 breast	GSE5764 breast	GSE9750 cervical	GSE9764 5-aza	GSE5816 hBEC Lung	GSE4246 AGO2sh	GSE14537 AGO2IP
ORF	ORF1&2 %A	0.059	**0.046**	0.083	0.997	0.546	0.340	0.320	0.146	0.982	0.415	0.581	0.066	0.198
	ORF1&2 %T	**0.016**	0.090	**0.011**	0.767	0.554	0.868	0.235	0.087	0.338	0.511	0.464	**0.007**	**0.006**
3' UTR	Poly-A pure	0.383	0.101	0.298	0.507	0.217	0.705	0.062	0.760	0.267	0.602	0.973	0.624	0.260
	Poly-A est	0.759	0.531	0.200	0.845	0.230	0.077	0.785	0.703	0.646	0.095	0.363	0.946	0.316
ALL	Find TSDs	0.138	0.857	0.235	0.636	0.786	0.779	0.335	0.805	**0.031**	0.181	0.420	0.404	0.147
	G-C Content	0.123	0.217	**0.002**	0.945	0.854	0.815	0.308	0.060	0.486	0.356	0.717	**0.011**	0.076
	CPG Islands	0.835	0.806	0.525	0.357	0.777	**0.012**	0.248	0.086	0.313	0.995	0.358	0.484	0.460
	Intactness score	0.256	0.865	0.311	0.945	0.421	**0.026**	0.050	0.126	0.996	0.785	0.260	**0.045**	**0.032**
	Number of L1	0.590	**0.000**	**0.001**	**0.018**	**0.030**	0.202	**0.000**	0.080	0.220	0.656	0.181	0.331	**0.019**

To discuss the results from statistical tests of each LINE-1 characteristic, we categorize those datasets into three main groups – cancers, 5-AZA, and AGO2.

For cancers group (see Table 4.2 and Table 4.4), the most frequent significant LINE-1 characteristics are the number of LINE-1s in a hosting gene which is significant in five datasets out of nine datasets and the orientation of gene which is significant in three datasets out of nine datasets. According to biological literature, the result that the number of LINE-1 elements is a significant characteristic supports the hypothesis of the earlier work [3], asserting that LINE-1 may repress gene expression by LINE-1 transcript. Therefore, the number of LINE-1 is obviously related to the opportunity of transcription.

For 5-AZA group (see Table 4.2 and Table 4.4), it is found that only one LINE-1 characteristics are significant, that is, the sequence “Y115” in ORF2 related to retrotranspositional activity of LINE-1. The odds ratio of this characteristic is 0.5872, indicating that genes possessing LINE-1 with mutated Y115 will be down-regulated approximately 1.7 times as much as genes with conserved Y115 LINE-1.

For AGO2 group (see Table 4.2 and Table 4.4), the most frequent significant LINE-1 characteristics are the percentage of base “T” only in ORF1 (“ORF1 %T”), the percentage of base T in both ORF1 and ORF2 (“ORF1&2 %T”), and intactness score of LINE-1 element (“Intactness score”).

However, we also tried to perform data mining to find the association between multiple LINE-1 characteristics and down-regulated gene expression. The next two sections are the results from data mining techniques.

4.2 The results from decision tree mining

This section demonstrates the result from C4.5 algorithm to create a tree model of LINE-1 characteristics for gene expression classification (see Table 4.5).

Table 4.5 The summary table of each dataset resulted from C4.5.

	Data set	% Overall Accuracy		% Recall Class [Gene expression = 'Down']	
		Information Gain	Gain Ratio	Information Gain	Gain Ratio
Cancers	GSE6919 Prostate	73.35%	72.55%	15.44%	18.46%
	GSE3167 Bladder	50.96%	51.57%	46.70%	47.69%
	GSE5816 Lung	86.02%	85.60%	13.38%	12.91%
	GSE6631 Head & neck	87.82%	88.27%	8.84%	10.61%
	GSE13911 Stomach	61.69%	63.15%	23.26%	27.81%
	GSE14811 Liver	76.36%	75.45%	15.84%	26.55%
	GSE1299 Breast	87.69%	86.99%	11.18%	9.16%
	GSE5764 Breast	98.63%	98.86%	0.00%	0.00%
	GSE9750 Cervical	62.62%	59.53%	31.68%	34.01%
5-AZA	GSE9764 5-Aza	90.52%	89.91%	4.22%	6.50%
	GSE5816 hBEC (Lung)	97.79%	97.56%	3.33%	3.33%
AGO2	GSE4246 AGO2sh	81.20%	79.68%	19.11%	17.95%
	GSE14537 AGO2IP	91.58%	91.05%	4.52%	8.13%

It is noticed that most of the datasets are imbalanced data, that is, [Gene expression = 'Not down'] class are always larger than [Gene expression = 'Down'] class. This difference in a greater degree of the number of samples in both class explains why the overall accuracy is high but the recall gene expression = 'Down' class percentage is quite low (see Table 4.5). Therefore, this method could not obviously classify down-regulated genes.

In addition, all of the output trees are too large to understand (see

Figure 4.1). Specifically, more than 15 LINE-1 characteristics were often used in the tree path to indicate a few down-regulated genes. In short, the more LINE-1 characteristics, the harder the interpretation. Therefore, we tried to apply association rules mining to find rules with less LINE-1 characteristics.

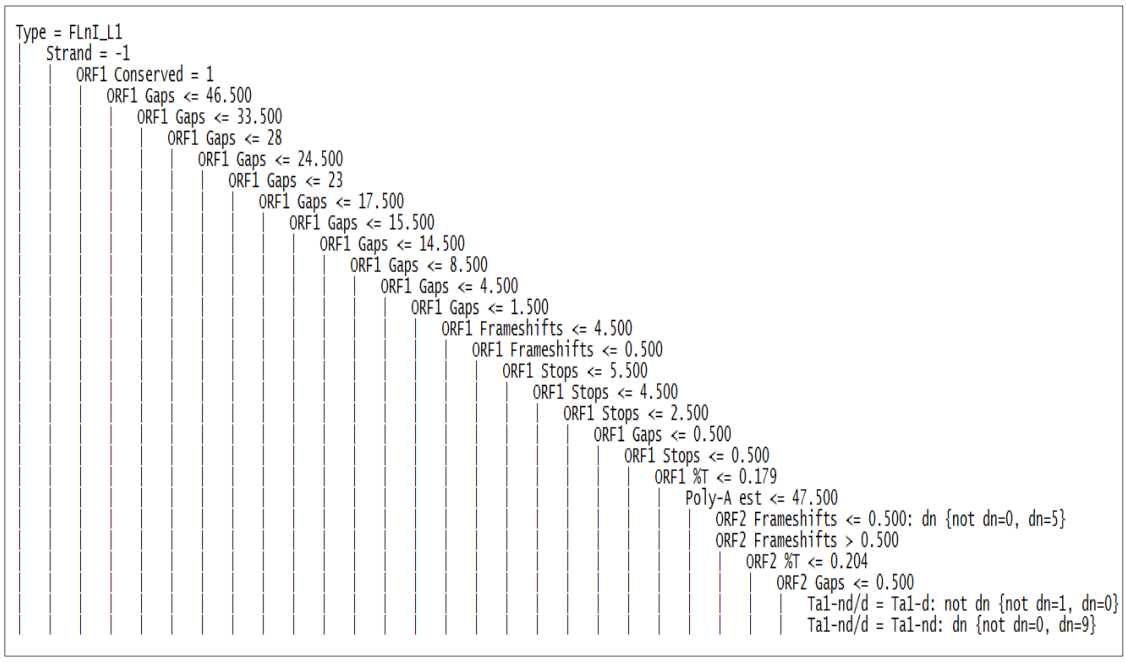


Figure 4.1 The part of the derived tree of GSE3167 (Bladder cancer) from C4.5

4.3 The results from rules mining

In this section, we summarize the data after preprocessing as the final input for rules mining. Moreover, we demonstrate the parameter settings which has been used in the experiment. In addition, we show and discuss mined classification association rules of each dataset including descriptive biological evaluation.

4.3.1 Data after preprocessing

Due to memory space limitation in classification association rules mining for rare events, we not only mined frequent patterns of "Down" to lower minimum support threshold for obtaining as many patterns as possible but also applied logistic regression method to decrease the number of LINE-1 characteristics to be performed in rules mining. Executing logistic regression, we set p-value threshold at significance level 0.5 of each LINE-1 characteristic's coefficient for every data set. Those insignificant coefficients of LINE-1 characteristics at such level were dropped. Approximately, a half of the total number of LINE-1 characteristics was remained in the rules mining. Consequently, we could set minimum support threshold much lower.

Table 4.6 exhibits the details of the final input of each dataset before mining rules: the number of LINE-1 characteristics after performing logistic regression, the total number of records, the number of records containing down regulation in gene expression level ("Down") including its percentage, and the actual minimum support threshold that was set before and after performing feature selection by logistic regression. It is noticed that GSE14054 (Importin8si-AGO2IP vs control-AGO2IP) was not applied by rules mining analysis because there was too few records containing down regulation in gene expression level (only 2 "Down" records out of the total 2626 records); thus, logistic regression was not performed on this dataset as well. In addition, GSE5764 (ductal and lobular breast cancer vs normal breast) still had all LINE-1 characteristics for rules mining because it was not converged when logistic regression was executing.

Table 4.6 The summary table of LINE1 characteristics in each dataset before mining association rules

	Datasets	The number of L1 characteristics using in rules mining (out of 51)	The number of "Down" records	The number of total records
Cancers	GSE6919 Prostate	27	164 (14.47%)	1133
	GSE3167 Bladder	24	719 (44.30%)	1623
	GSE5816_1 Lung	28	190 (7.24%)	2626
	GSE6631 Head & neck	28	68 (6.00%)	1134
	GSE13911 Stomach	26	630 (24.11%)	2613
	GSE14811 Liver	21	80 (14.55%)	550
	GSE1299 Breast	28	94 (6.61%)	1422
	GSE5764 Breast	56	16 (0.61%)	2626
	GSE9750 Cervical	31	447 (27.12%)	1648
5-AZA	GSE9764 5-Aza	26	118 (4.49%)	2626
	GSE5816_2 hBEC (Lung)	25	24 (0.91%)	2626
AGO2	GSE4246 AGO2sh	26	145 (10.48%)	1383
	GSE14537 AGO2IP	24	110 (4.19%)	2626
	GSE14054 si-AGO2I	-	2 (0.08%)	2626

4.3.2 Generated classification association rules

After preprocessing, local frequent patterns or only the frequent patterns in “Down” regulation are mined. We set minimum local support threshold parameter as low as possible for each dataset. Therefore, different datasets were set by different actual minimum support thresholds. Next, classification association rules were generated by considering only “Down” class. These outcome rules were filtered by the confidence threshold at 50% and adjusted p-value from chi-square test at significance level 0.05.

Since we designated low support threshold (50%), we obtained a great number of rules at the first stage. After we filtered redundant rules out, we could manually study the rules in GSE14811 (Liver) dataset since it had only six rules (see Table 4.7). However, other datasets still had too many rules to observe completely. Therefore, we put some rules out of our consideration by pruning some rules which does not effect the overall supports and maximum confidence. The number of rules in each stage of filtering rules is shown in Table 4.7.

Table 4.8 summarizes the important details of the set of rules for each dataset: the actual minimum support threshold, the number of rules which were generated and pruned, the number of LINE-1 characteristics before mining and after mining and maximum-minimum support and confidence percentage. Only five datasets : GSE3167 (bladder carcinoma situ vs normal bladder epithelium), GSE5812 (hBEC high dose vs human bronchial epithelium), GSE6631 (head and neck squamous cell carcinoma vs normal oral epithelium), GSE6919 (metastasis prostate cancer), and GSE14811 (liver cancer vs normal liver), have been discovered classification association rules under the constraints mentioned above.

Table 4.7 The number of rules in each stage of pruning rules

	Datasets	The number of significant rules	The number of rules after pruning redundant rules	The number of rules after pruning rules in the final stage
Cancers	GSE6919 Prostate	828	23	16
	GSE3167 Bladder	133	79	15
	GSE5816 Lung	0	0	0
	GSE6631 Head & neck	8524	374	23
	GSE13911 Stomach	0	0	0
	GSE14811 Liver	156	6	6
	GSE1299 Breast	0	0	0
	GSE5764 Breast	0	0	0
	GSE9750 Cervical	0	0	0
5-AZA	GSE9764 5-Aza	0	0	0
	GSE5816 hBEC (Lung)	48,560	495	24
AGO2	GSE4246 AGO2sh	0	0	0
	GSE14537 AGO2IP	0	0	0

Almost all of the rules with high confidence have low support percentage (see Table 4.8). For example, GSE6919 (Prostate), GSE14811 (Liver), and GSE5816 hBEC (Lung) have rules with 100% confidence but these rules have very low support, less than 1%.

Table 4.8 The summary of the output rules in each dataset

	Datasets	Actual minimum support threshold (%)	# L1 characteristics for mining (out of 51)	The number of final rules	#L1 characteristics used in each rule	Max-min support (%)	Max-min confidence (%)
Cancers	GSE6919 Prostate	0.50	27	16	7-9	0.53 - 0.97	100 - 64.71
	GSE3167 Bladder	3.99	24	15	3-7	11.15 - 4.74	71.96 – 59.54
	GSE5816 Lung	0.94	28	0	-	-	-
	GSE6631 Head & neck	0.40	28	23	7-12	0.62 – 0.44	75 - 53.85
	GSE13911 Stomach	2.17	26	0	-	-	-
	GSE14811 Liver	0.44	21	6	5-9	1.45 - 1.09	100 - 80
	GSE1299 Breast	1.19	28	0	-	-	-
	GSE5764 Breast	0.43	51	0	-	-	-
	GSE9750 Cervical	3.25	31	0	-	-	-
5-AZA	GSE9764 5-Aza	0.36	26	0	-	-	-
	GSE5816 hBEC (Lung)	0.07	25	24	7-11	0.11 - 0.08	100 – 66.67
AGO2	GSE4246 AGO2sh	0.94	26	0	-	-	-
	GSE14537 AGO2IP	0.29	24	0	-	-	-

4.3.3 Results and discussion on classification association rules

Each rule generated from the association rules mining technique may be used as a biological hypothesis that LINE-1 characteristics in the rule are associated with gene expression in cancers; however, there are too many rules to explore manually. To analyze or interpret such a number of classification association rules, filtering approach played an important role. Furthermore, in this study, we choose only rules in a group of the highest confidence to discuss in terms of biology.

Before biological discussion in each dataset, here is basic knowledge of the function of LINE-1 characteristics. As mentioned in the literature review, the structure of LINE-1 is divided into four parts (see Figure 4.2). The characteristics in each part has different function as the following:

5' UTR is involved with transcriptional activities like normal genes.

ORF1 is not exactly known about its function; nevertheless, it is about assisting retrotranspositional activities.

ORF2 has the function related to retrotranspositional activities.

3' UTR may be engaged in transcriptional activities of the next LINE-1 element.

Figure 4.2 The structure of LINE-1

In the following subsections, the classification association rules in each dataset are demonstrated and discussed. According to the biological literature, so far, the association of LINE-1 and gene expression in cancers are focused on four parts of LINE-1 element instead of specific characteristics in these parts, the number of LINE-1, and the orientation of hosting gene; therefore, we will mainly discuss on these factors and leave complete rules as possible hypotheses for future discussions or experiments.

GSE3167 (Bladder carcinoma situ vs normal bladder epithelium)

This dataset represents bladder carcinoma situ vs normal bladder epithelium dataset. In this data set, there are, after pruning, 15 final rules (see Appendix B). Here, top four rules are discussed. All LINE-1 characteristics used in these four rules and their measurements are shown in Figure 4.3 and Table 4.9. However, It is noted that the number of LINE-1 is used in every of the 15 rules, supporting my previous work [31].

Table 4.9 The summary of the top four rules of GSE3167 (Bladder)

Rule No.	Confidence	Support	Consistent to rule		Not consistent to rule		Odds	p-value
			Down	Not down	Down	Not down		
1	71.96%	4.74%	77	30	642	874	3.49	2.52E-09
2	71.17%	4.87%	79	32	640	872	3.36	3.54E-09
3	70.59%	5.18%	84	35	635	869	3.28	2.01E-09
4	70.49%	5.30%	86	36	633	868	3.28	1.40E-09

It is noted that rule 2 is very similar to rule 3 but [Runx3 ASP] = 'mut' added in rule 2, but the confidence percentage of rule 2 is trivially increased when compared to rule 3. So, it can be deduced that [Runx3 ASP] = 'mut' does not play an important role to down regulate hosting gene if the other common four LINE-1 characteristics in rule 2 and rule 3 are concerned. Thus, we will not consider rule 2 in later discussion. Furthermore, in this bladder cancer dataset, all three rules (rule 1,3, and 4) have no LINE-1 characteristics on 5' UTR. In addition, the rules in the highest confidence of this dataset have the most support percentage compared to the rest of datasets. In other words, the rules would explain what LINE-1 characteristics are associated gene expression in more hosting genes.

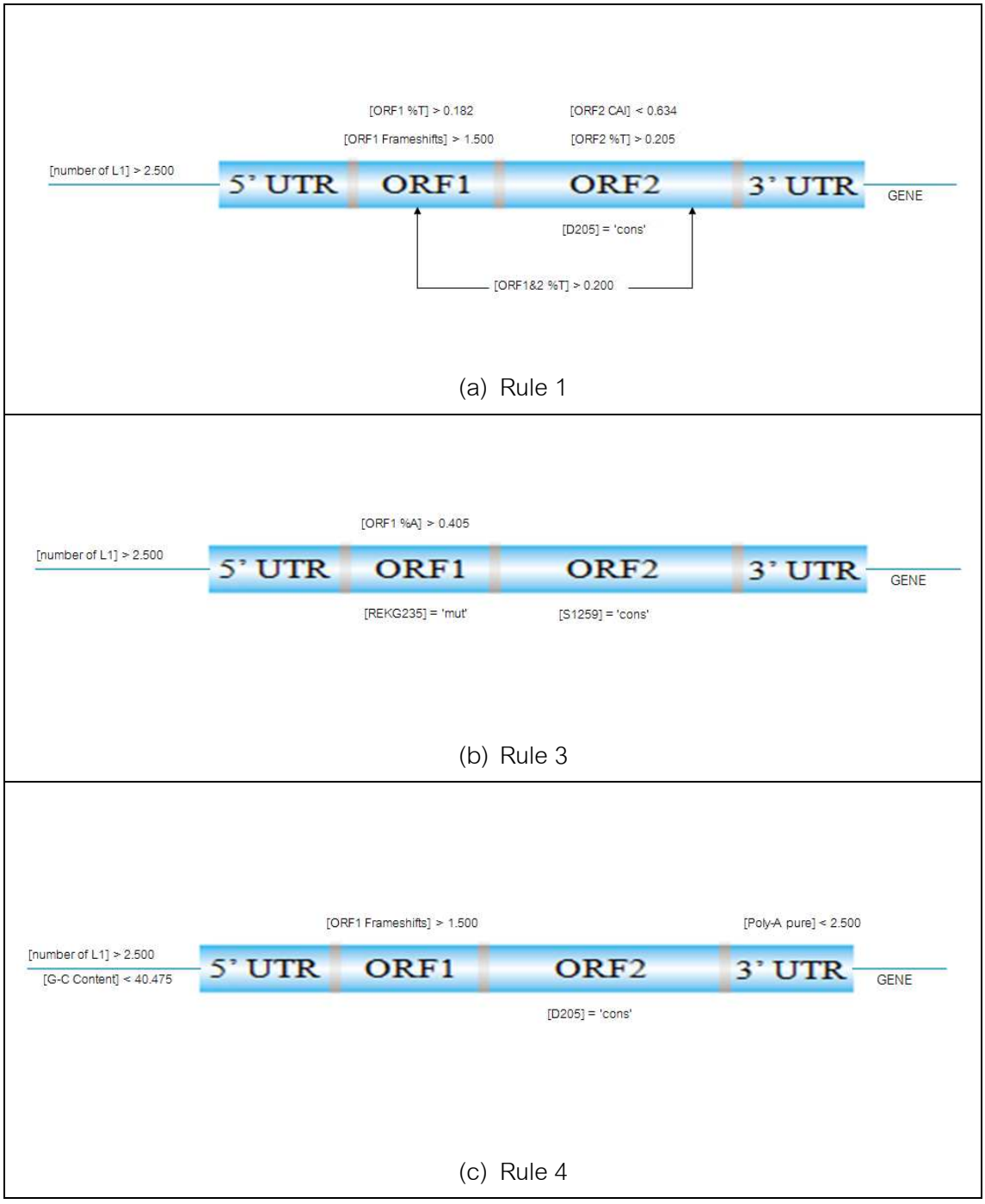


Figure 4.3 LINE-1 characteristics used in each rule of GSE3167 (Bladder), mapped to each position of LINE-1

Discussion the rules of GSE3167 (Bladder) in terms of biology

The conspicuous remark is that every rule contains the characteristics namely “[number of L1] > 2.500” similar to our previous work [31], suggesting that the more the LINE-1 elements in genes, the more risky for those genes to be repressed in bladder cancer. This result strongly supports the hypothesis of the recent work [3], asserting that genes may be lowly expressed in cancers because of LINE-1 transcripts. According to the mechanism proposed in this research [3], when LINE-1 elements are hypomethylated, found in many types of cancer cells, they would be frequently transcribed and LINE-1 RNA level is finally increased. Generally, retrotransposon RNAs or transcripts including LINE-1 elements form dsRNA (double-strand RNA) structures which trigger RISC (RNA-induced silencing complex) assembly to curtail these dsRNAs to siRNA (small interference RNA) elements which are then combined with RISC, called RISC complex. Next, the RISC complex elements are recognized by Argonaute 2 (AGO2) proteins. Finally, RISC containging LINE-1 transcript and attached to RNA of its hosting genes will often be destroyed, resulting in down regulation. In short, intragenic LINE-1 can produce LINE-1 RNAs or transcripts when being hypomethylated in cancer cells, and the hosting gene is consequently down-regulated when AGO2 exists [3]. Therefore, the outcome rules suggest that if a gene contains many LINE-1 elements (more than two LINE-1s), it may be, as explained above, down-regulated with high possibility to the max of 71%.

GSE14811 (Liver cancer vs normal liver)

This dataset represents liver cancer vs normal liver dataset. In this data set, there are, after pruning, 6 final rules (see Appendix B). Moreover, there are three rules with 100% confidence. All LINE-1 characteristics used in these three rules and their measurements are shown in Figure 4.4 and Table 4.10.

Table 4.10 The summary of the top four rules of GSE14811 (Liver)

Rule No.	Confidence	Support	Consistent to rule		Not consistent to rule		Odds	Adjusted p-value
			Down	Not down	Down	Not down		
1	100.00%	1.09%	6	0	7	470	NA	7.14E-8
2	100.00%	1.09%	6	0	7	470	NA	7.14E-8
3	100.00%	1.09%	6	0	7	470	NA	7.14E-8

It is noticed that while rule 3 uses only LINE-1 characteristics on ORF (both ORF1 and ORF2), rule 1 and rule 2 use LINE-1 characteristics in 5' UTR, 3' UTR, and ORF2. However, all of the three rules have the same confidence and support percentage.

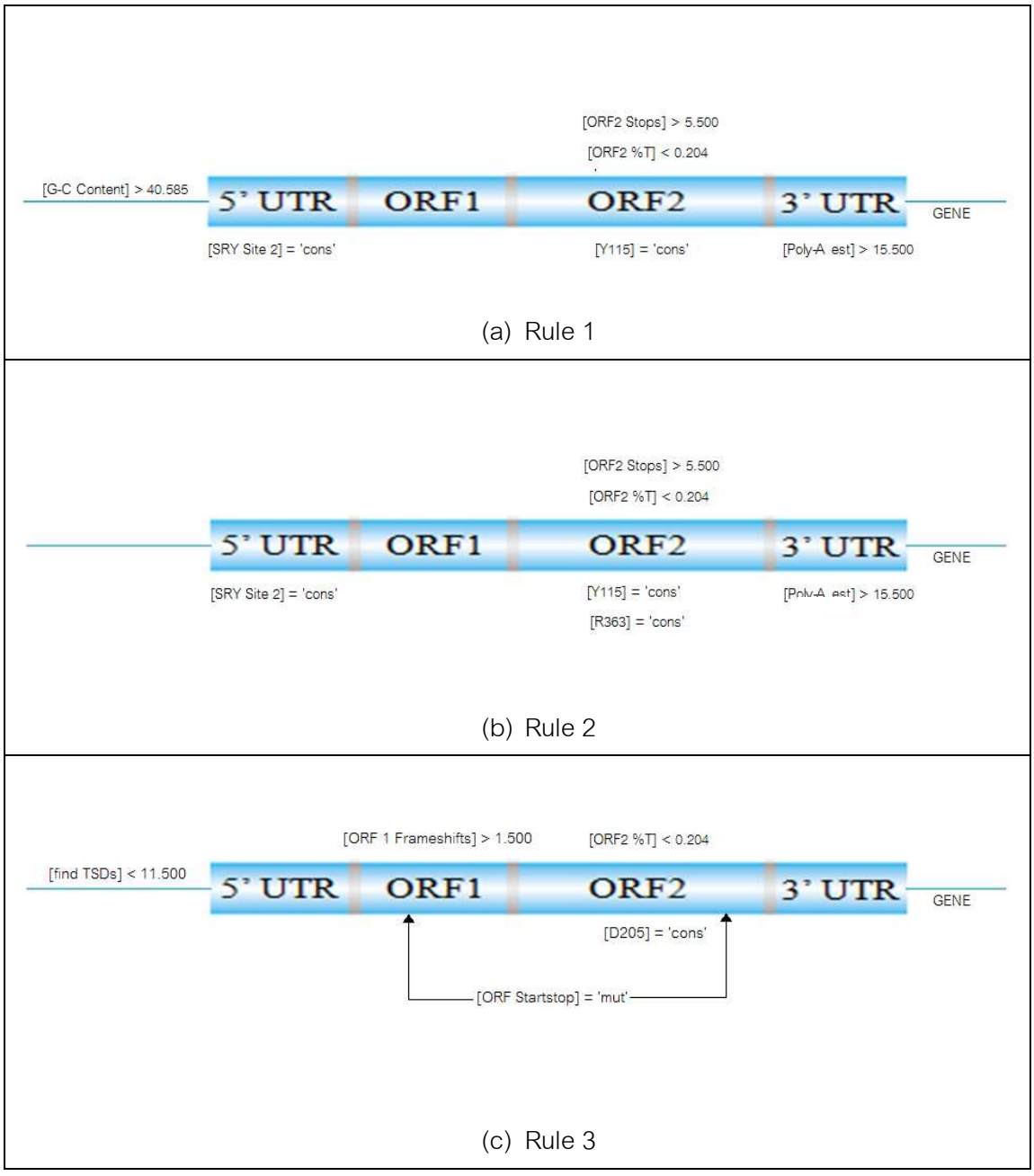


Figure 4.4 LINE-1 characteristics used in each rule of GSE14811 (Liver), mapped to each position of LINE-1

Discussion the rules of GSE14811 (Liver) in terms of biology

This liver cancer dataset is another one to support the hypothesis of the recent study [3]. Like the outcome rules in the prostate cancer dataset (GSE6919) both rule 1 and rule 2, two out of three rules with 100% confidence, consist of the characteristic SRY Site2 in 5' UTR. However, instead of mutated, the sequences of SRY Site2 are conserved ([SRY Site2] = 'cons'), resulting the counter consequence. Again, both of SRY binding sites (SRY Site1 and SRY Site2) have the sequence AACAAA and interact with the DNA-binding domain of SRY [34]. Binding of the SRY sites in the LINE-1 5' UTR can drive transcription of the LINE-1 promoter. Briefly, if the sequences of SRY Site2 are not mutated, they will probably function normally, promoting LINE-1 transcriptional activity. By this manner, the LINE-1 RNA level is increased as well as the mechanism upon the number of LINE-1 in bladder cancer dataset (GSE3167). Consequently, the hosting genes with conserved SRY Site2 can also be down-regulated according to the hypothesis proposed by the recent work [3]. In addition, SRY Sites are the sequences on chromosome Y; thus, the derived rules support the fact that liver cancer is found mostly in males. In other words, this type of cancer may partly be controlled through SRY Site2.

As for rule 3, instead of [SRY Site2] = 'cons', there is [find TSDs] < 11.500. TSD is target site duplications and [find TSDs] is the number of TSDs. According to the previous study [32], these target site duplications are caused by the host DNA repairing mechanism when LINE-1 insertion occurs. Specifically, the host DNA repairs itself by filling the cleavage break, caused by Endonuclease Enzyme, with short direct repeats known as target site duplications. Therefore, if less TSDs exist, then it can be deduced that the host DNA repairing mechanism may not be successful. As a result, the hosting gene inclines to be repressed because of its broken part.

GSE6919 (Metastasis prostate cancer)

This dataset represents metastasis prostate cancer dataset. In this data set, there are, after pruning, 16 final rules (see Appendix B). Moreover, there are four rules with 100% confidence. All LINE-1 characteristics used in these four rules and their measurements are shown in Figure 4.5 and Table 4.11.

Table 4.11 The summary of the top four rules of GSE6919 (Prostate)

Rule No.	Confidence	Support	Consistent to rule		Not consistent to rule		Odds	Adjusted p-value
			Down	Not down	Down	Not down		
1	100.00%	0.53%	6	0	158	969	NA	2.37E-09
2	100.00%	0.53%	6	0	158	969	NA	2.37E-09
3	100.00%	0.53%	6	0	158	969	NA	2.37E-09
4	100.00%	0.53%	6	0	158	969	NA	2.37E-09

The common characteristics used in all of the four rules are the positive orientation of hosting gene is positive, the mutated sequence of “SSS1096” in ORF2, and the number of stops in ORF1 which is less than 2.5 ([ORF1 Stops] < 2.500). In addition, it is noticed that all four rules are involved with ORF1 and ORF2 of LINE-1 elements.

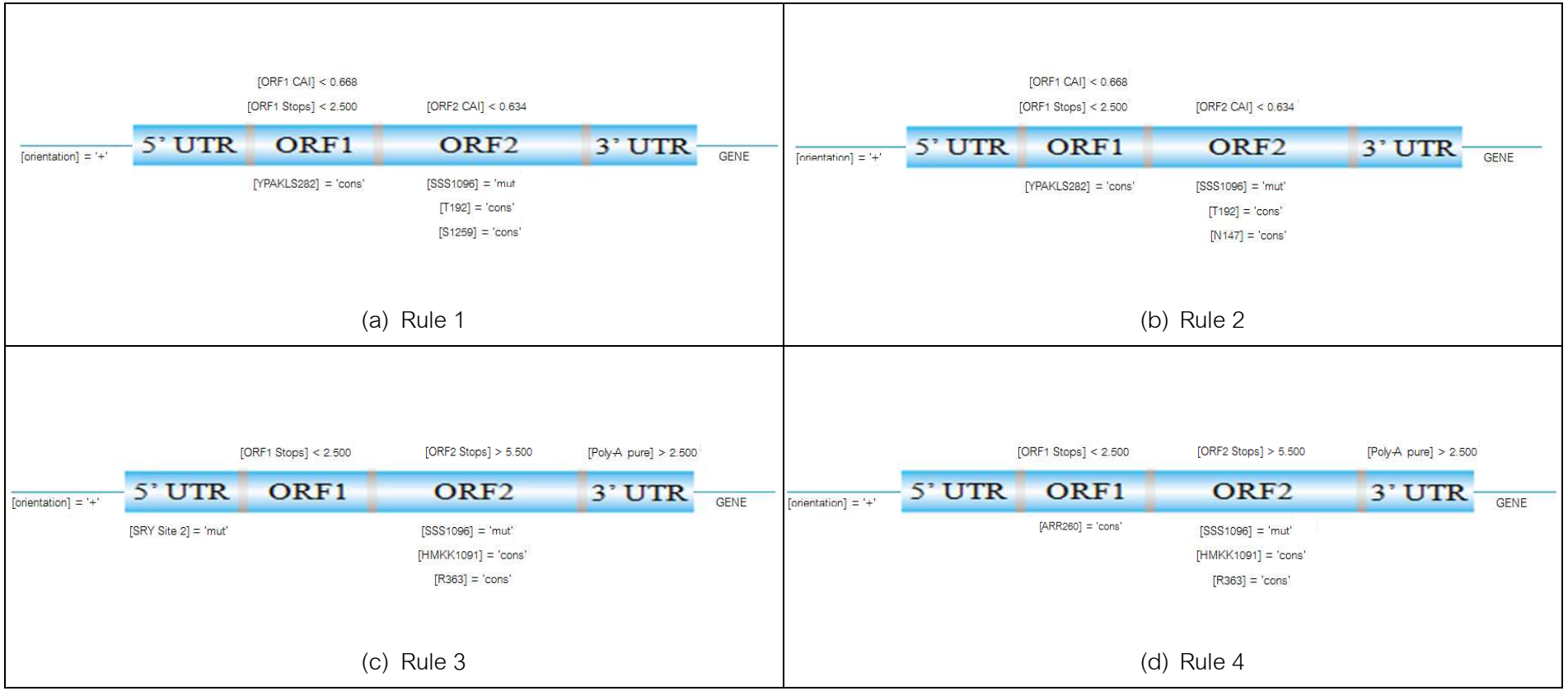


Figure 4.5 LINE-1 characteristics used in each rule of GSE6919 (Prostate), mapped to each position of LINE-1

Discussion the rules of GSE6919 (Prostate) in terms of biology

All of the four rules are involved the conserved sequence in ORF2 and ORF1, representing that LINE-1 retrotranspositional activities. According to the fundamental molecular knowledge [4], the retrotranspositional activities include transcription, RNA stability and processing, translation, DNA restriction, reverse transcription and insertion. However, it has not been reported how retrotranspositional activities of LINE-1 are associated with gene expressions in cancer cells. Thus, we hypothesized that these characteristics may be also related to LINE-1 transcriptional process like other cancers but different method.

The obvious difference of rule 3 from the rest of the rules is that rule 3 has mutated sequences of SRY (Sex-determining Region Y) Site 2 in 5' UTR pertaining to transcriptional activity. Specifically, SRY Site2 mutations are able to prevent transcription by eliminating SOX protein binding [32]. Normally, human L1s contain two functional binding sites for transcription factors of the SRY family (SRY Site1 and SRY Site2), namely SOX factors. These binding sites are responsible for efficient trans-activation of the L1 promoter by the SOX family [34]. Principally, SRY Sites in the LINE-1 5' UTR can drive transcription of the LINE-1 promoter. On the contrary, when the sequences of SRY Site2 are mutated, resulting in SOX protein binding is abolished. So, this characteristic is able to impede LINE-1 transcriptional process, indicating that lowly expressed genes in prostate cancer may be caused by different mechanisms from bladder and liver cancer whose down-regulated genes are influenced by LINE-1 transcript. Besides, prostate cancer only occurs in males, the SRY Sites existing on chromosome Y is not responsible for down regulating genes. Unlike liver cancer mostly found in males, therefore, the conserved sequences of SRY Site1 may not be required to suppress genes but the conserved sequences on ORF1 and/or ORF2, instead.

GSE6631 (Head and neck squamous cell carcinoma vs normal oral epithelium)

This dataset represents head and neck squamous cell carcinoma vs normal oral epithel dataset. In this data set, there are, after pruning, 23 final rules (see Appendix B). Here, three top rules are discussed. All LINE-1 characteristics used in these three rules and their measurements are shown in Figure 4.6 and Table 4.12.

Table 4.12 The summary of the top three rules of GSE6631 (Head & Neck)

Rule No.	Confidence	Support	Consistent to rule		Not consistent to rule		Odds	Adjusted p-value
			Down	Not down	Down	Not down		
1	75.00%	0.53%	6	2	62	1064	51.48	1.59E-16
2	75.00%	0.53%	6	2	62	1064	51.48	1.59E-16
3	75.00%	0.53%	6	2	62	1064	51.48	1.59E-16

From all three rules, it is noticed that all three rules have no LINE-1 characteristics on 3' UTR. Besides, LINE-1 characteristics used in these three rules in this dataset, especially the sequences, are on 5' UTR. Furthermore, The characteristic [number of L1] < 2.500 in rule 1 and rule 2 conflicts with the rules in GSE3167 (Bladder), using [number of L1] > 2.500 and The characteristic [orientation] = '-' in all three rules contradicts to the rules in GSE6919 (Prostate), using [orientation] = '+'.

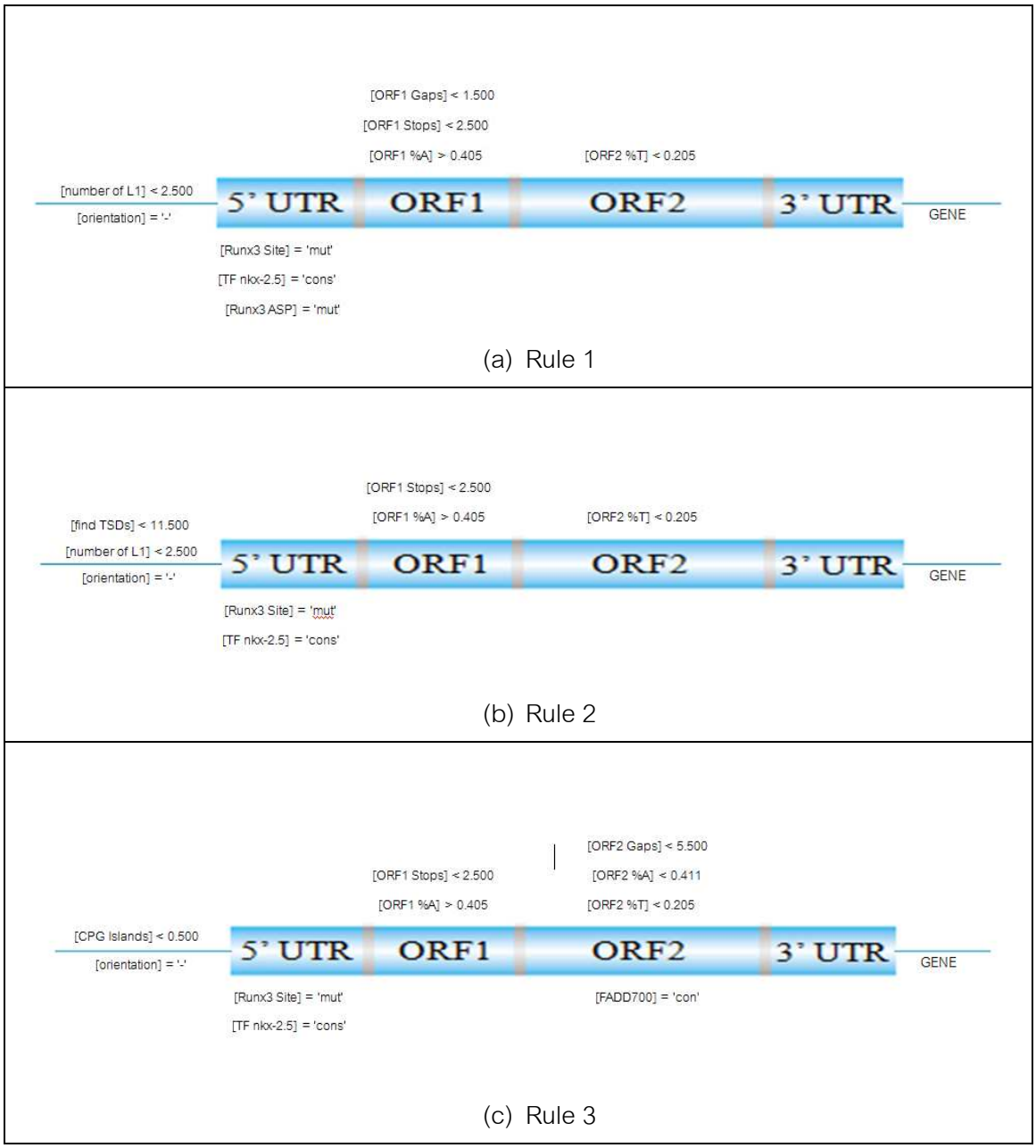


Figure 4.6 LINE-1 characteristics used in each rule of GSE6631 (Head and Neck), mapped to each position of LINE-1

Discussion the rules of GSE6631 (Head and Neck) in terms of biology

It should be noticed that all three rules used in head and neck cancer cells contain either [Runx3 Site] = 'mut' or [Runx3 ASP] = 'mut' in 5' UTR. While Runx3 Site is responsible for the sense strand LINE-1, Runx3 ASP is involved with the other (the antisense strand LINE-1). Sense or antisense of LINE-1 is relied on the orientation of LINE-1 and its hosting gene. If the strand of LINE-1 is similar to the orientation of its hosting gene, then it is known as sense strand LINE-1. On the other hand, if the orientation of LINE-1 conflicts with the transcriptional direction of its hosting gene, then it is so-called antisense strand LINE-1. Having reviewed the literature [32], mutations in Runx sites, no matter in the sense or antisense strand, are capable to decrease retrotranspositional activity where a single nucleotide mutation (84G>A) decreases retrotranspositional activities by 85% [35]. Besides, the description of the retrotranspositional obstruction due to these mutated sequences of Runx3 binding sites is supported by the fact that conserved sequences in ORF2 are hardly discovered in the rules. This evidence points out that the possible mechanism to regulate gene expression by LINE-1 interference may not be involved with retrotranspostional activity. Rather, it is plausible to repress genes by the presence of the conserved sequences at transcription factor binding sites in 5' UTR. Noticeably, every rule with the highest confidence percentage (75%) contains the LINE-1 characteristic called "[TF nkx-2.5] = 'cons'". Therefore, we hypothesized that the conserved sequences of transcription factor nkx2.5 can related to LINE-1 transcriptional activity to repress the expression of the hosting gene.

GSE5816 (hBEC high dose vs human bronchial epithelium)

This dataset represents hBEC high dose vs human bronchial epithelium dataset which is in the 5-AZA group dataset. In this data set, there are, after pruning, 24 final rules (see Appendix B). Moreover, there are 11 rules with 100% confidence. All LINE-1 characteristics used in these 11 rules and their measurements are shown in and Table 4.13.

Table 4.13 The summary of the top 11 rules of GSE5816 (hBEC Lung)

Rule No.	Confidence	Support	Consistent to rule		Not consistent to rule		Odds	Adjusted p-value
			Down	Not down	Down	Not down		
1	100.00%	0.11%	3	0	21	2602	NA	6.34E-51
2	100.00%	0.11%	3	0	21	2602	NA	6.34E-51
3	100.00%	0.11%	3	0	21	2602	NA	6.34E-51
4	100.00%	0.11%	3	0	21	2602	NA	6.34E-51
5	100.00%	0.11%	3	0	21	2602	NA	6.34E-51
6	100.00%	0.11%	3	0	21	2602	NA	6.34E-51
7	100.00%	0.11%	3	0	21	2602	NA	6.34E-51
8	100.00%	0.11%	3	0	21	2602	NA	6.34E-51
9	100.00%	0.08%	2	0	22	2602	NA	3.26E-28
10	100.00%	0.08%	2	0	22	2602	NA	3.26E-28
11	100.00%	0.08%	2	0	22	2602	NA	3.26E-28

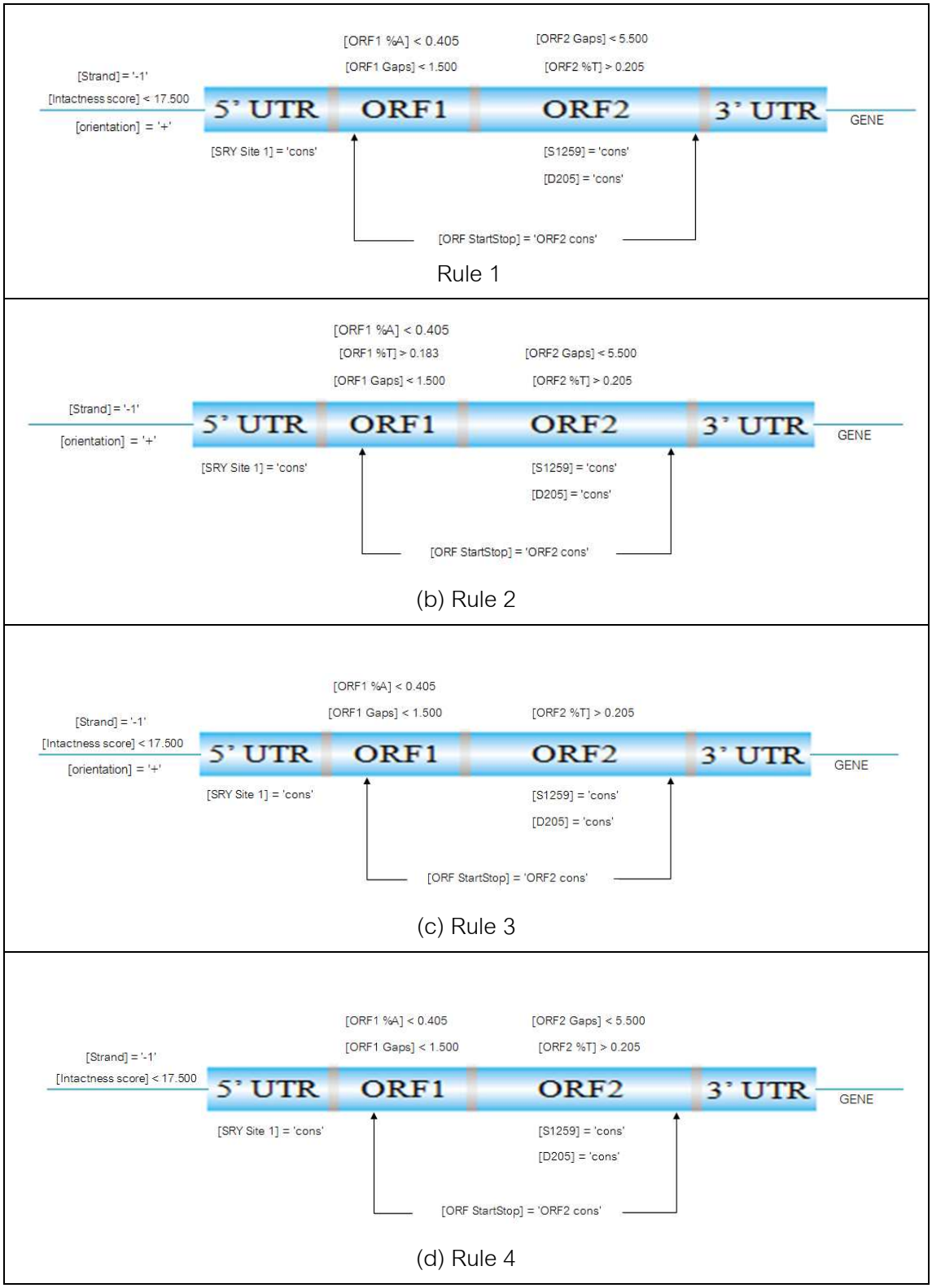
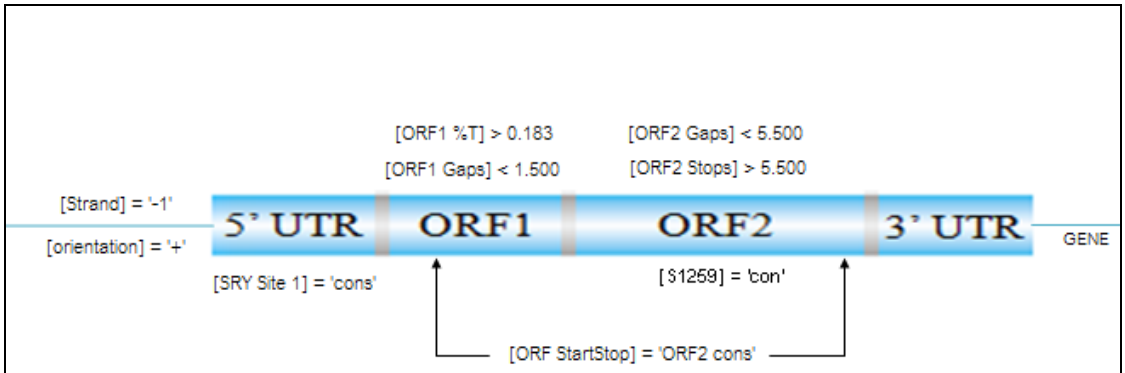
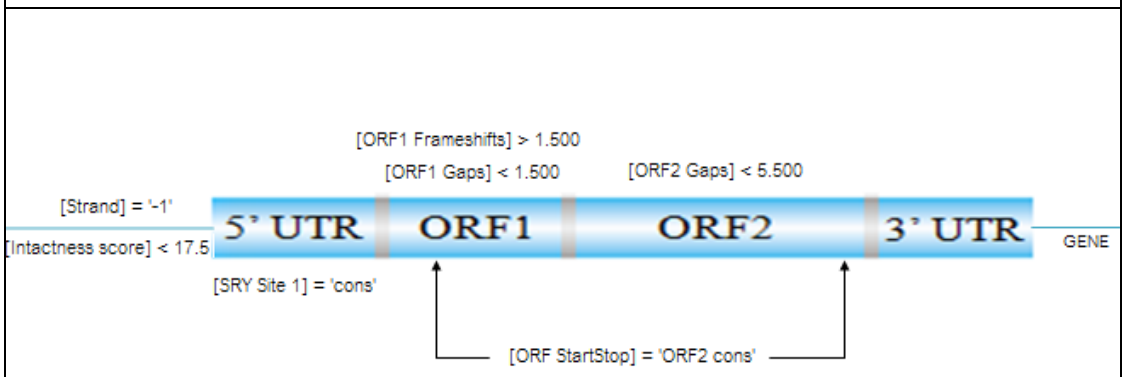


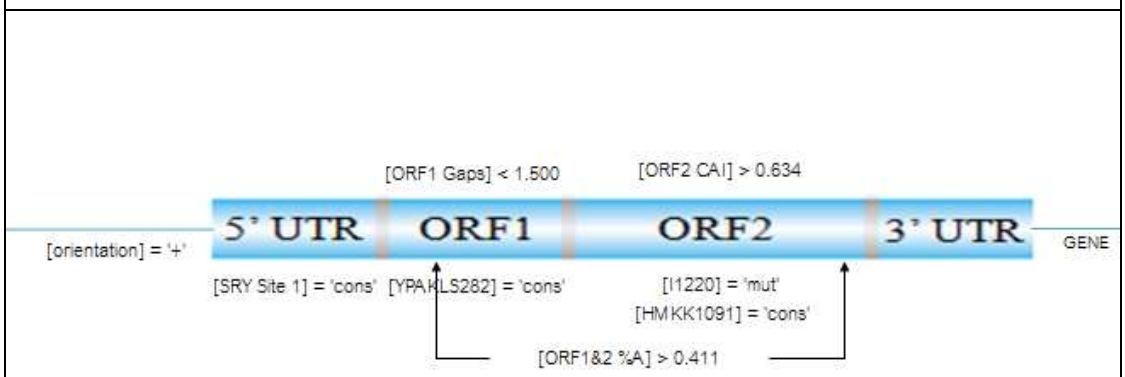
Figure 4.7 LINE-1 characteristics used in each rule of GSE5816 (hBEC Lung), mapped to each position of LINE-1



(e) Rule 5



(f) Rule 6



(g) Rule 7

Figure 4.7 (Continued)

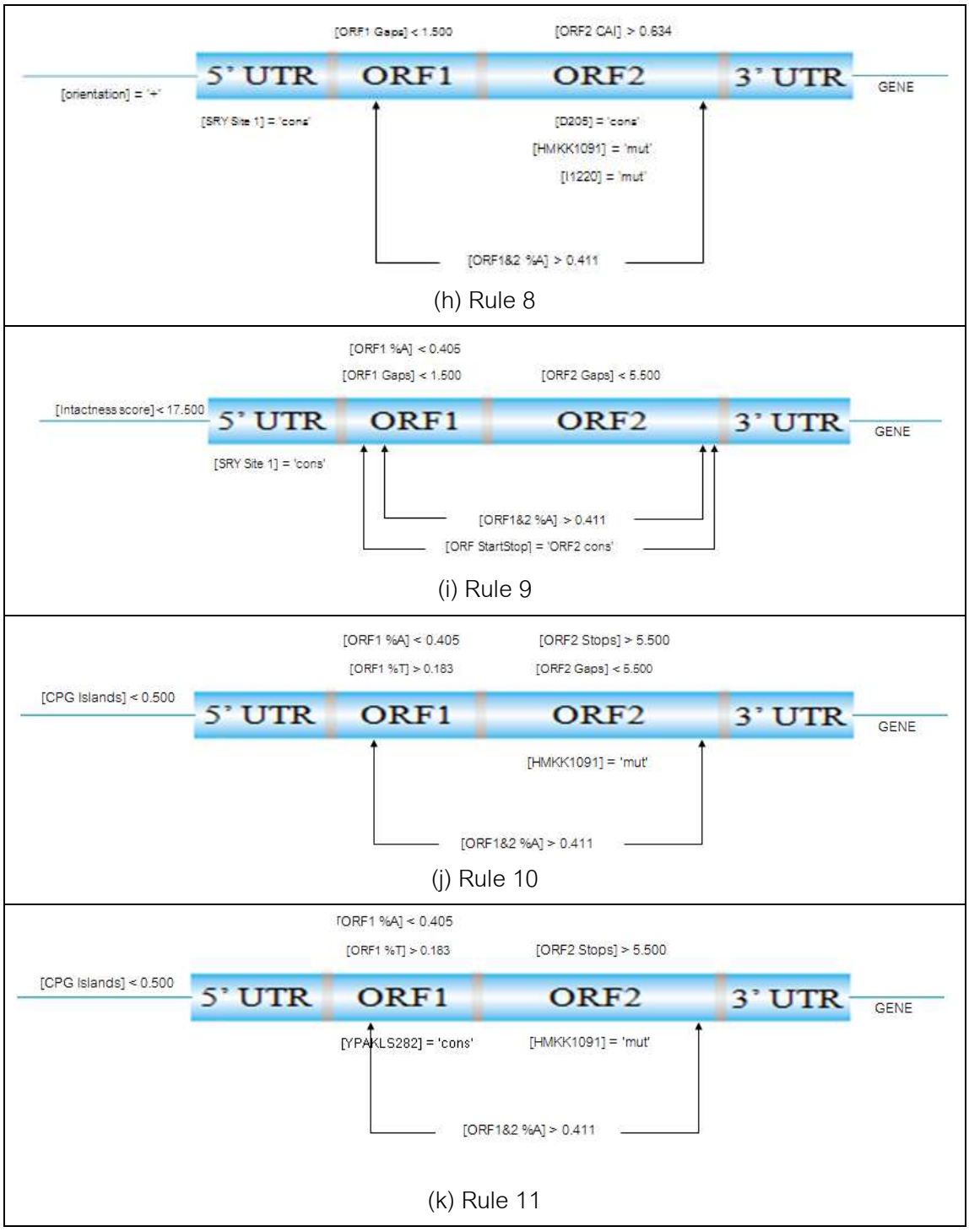


Figure 4.7 (Continued)

Discussion the rules of GSE5816 (hBEC Lung) in terms of biology

It is noticed that almost every rule (9 out of 11 rules) with the highest confidence percentage (100%) contains the characteristic [SRY Site1] = 'cons'. Even though in the hBEC Lung dataset the conserved sequences are on SRY Site1 instead of SRY Site2, the importance of this characteristic is the same as the conserved SRY Site2 used in rules of the liver cancer dataset (GSE14811), promoting LINE-1 transcriptional activity because both sites of SRY are responsible for efficient trans-activation of the LINE-1 promoter. Once more, the hosting genes with this characteristic may also be down-regulated, according to the hypothesis proposed by the prior work [3].

Besides, rule 1, 2, 3, and 5 comprise the characteristics both [orientation] = '+' and [Strand] = '-1'. It was hypothesized that both sense and antisense LINE-1 element can regulate gene expression. According to the assumption that genes in cancer cells may be down-regulated by LINE-1 transcript, whereas sense LINE-1 element suppresses its hosting gene by promoting LINE-1 transcription in the normal direction (5' to 3'), antisense LINE-1 represses the neighboring gene by using the anti-promoter on 3' as the starting point of transcription. To sum up, this derived rules suggest that no matter what kinds of orientation of LINE-1, the expression of genes in cancer cells can probably be controlled by transcriptional activity of LINE-1.

Chapter V

Conclusion

According to the previous work [3], it has been found that some genes with LINE-1 are significantly down regulated in many kinds of cancer. To contribute biologists to understand the function of LINE-1 characteristics associated with gene expression in cancers, therefore, the aim of this study was to find LINE-1 characteristics mediating gene expression in cancers. We applied three main approaches, statistical method for bivariate data analysis, C4.5 tree and association rules mining based on classification purpose for multivariate data analysis. The result from bivariate analysis pointed out the significant LINE-1 characteristics, especially the number of LINE-1, individually associated with gene expression. In addition, the result from rules mining informed the interactions of LINE-1 characteristics associated with down-regulated genes as follows.

In cancer datsets, each type of cancer has distinct rules. This result may be due to different in hypomethylation mechanisms or different in LINE-1 transcription or post transcription mechanisms. In this research, the derived rules support the hypothesis that down-regulated genes in cancer cells may be controlled by LINE-1 transcripts [3]. However, it is found that different cancer may have different transcription factors of LINE-1s. First of all, the derived rules from bladder cancer suggest that if the number of LINE-1s is greater than 2, LINE-1 transcription is likely to be promoted. Consequently, the hosting gene might be repressed. Secondly, LINE-1 with conserved sequences of SRY Site2 in most of the rules in liver cancer mostly found in males might be responsible for down regulation of its hosting gene through LINE-1 transcriptional mechanism. Thirdly, in prostate cancer found only in males, SRY Site1 may not be required to regulate genes. Instead, LINE-1 with conserved sequences on ORF1 and/or ORF2 are possible to suppress genes in prostate cancer cells, even though the mechanism has still been questionable. Finally, the rules from head and neck cancer

imply that the conserved sequences of TF-nkx-2.5 in 5' UTR of LINE-1 are associated with lowly expressed genes, but the mechanism is unknown either.

Likewise, the rules derived from hBEC Lung dataset in 5-AZA group concur with the assumption proposed by C. Aporntewan, *et al* [3]. The LINE-1 transcription is driven by the conserved sequences of SRY Site1. Besides, sense and antisense LINE-1 can probably control the expression of genes by either directions of LINE-1 transcription.

Recommendation

This study should be improved in feature selection stage which is important to effect the outcome rules. For example, we performed logistic regression by putting all LINE-1 characteristics to create the model. This method might be developed by using backward or forward concept to delete or add LINE-1 characteristics in the model to find the better model. However, this approach is based on greedy concept. In other words, it cannot ensure that the result model is the best one. Otherwise, it should be tried to apply other methods for feature selection. Similarly, it is not known that the selected features should be chosen until rules mining is performed.

Future work

- Those rules should be supported by further biological literatures and proved in biological experiments.

- The demonstrated rules should be scrutinized in depth on biological perspectives and other significant rules should be also considered.

- Those rules should be validated in other datasets by applying the same data mining settings and comparing the derived rules. For example, to prove the assumption of down regulation by the protein named AGO2, association between LINE-1 characteristics and up-regulated gene expression should be performed.

References

[1] World Health Organization [online]. 2008. Available from : http://www.who.int/en

[2] K. D. Robertson. DNA methylation and human disease. Nature Review Genetics 6 (2005) : 597-610.

[3] C. Aporntewan, C. Phokaew, J. Piriyapongsa, C. Ngamphiw, C. Ittiwut, S. Tongsima and A. Mutirangura. Hypomethylation of intragenic LINE-1 represses transcription in cancer cells through AGO2. PLoS ONE 6(3) (2011) : doi:10.1371/journal.pone.0017934.

[4] D. Clark. Molecular Biology. China : Elsevier Academic Press, 2005.

[5] O. V. Pidpala, A. P. Yatsishina and L. L. Lukash. Human Mobile Genetic Elements: structure, distribution and functional role . Cytology and Genetics 42(6) (2008) : 71-83.

[6] A. V. Furano. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Nucleic Acid Research and Molecular Biology 64 (2000) : 255-294.

[7] T. Penzkofer, T. Dandekar and T. Zemojtel. L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acid Research 33 (2005) : D498-D500.

[8] J. S. Milton and J. C. Arnold. Introduction to Probability and Statistics. USA : McGraw-Hill, 2003.

[9] J. F. Hair, W. C. Black, B. J. Babin and R. E. Anderson. Multivariate Data Analysis. Seventh Edition. USA : PEARSON, 2010.

[10] J. Han and M. Kamber. Data Mining Concepts and Techniqes. Second Edition. USA : Elsevier, 2006.

[11] J. Han, J. Pei, Y. Yin and R. Mao. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Mining and Knowledge Discovery 8 (2004) : 53-87.

[12] K. R. Seeja, M.A. Alam and S. K. Jain. Identification of co-regulated signature genes in pancreas cancer- a data mining approach. Lecture Notes in Computer Science 5226 (2008) : 138-145.

[13] R. Alves, D. S. Rodrigues-Baena and J. S. Aguilar-Ruiz. Gene association analysis: a survey of frequent pattern mining from gene expression data. Brief Bioinform 11(2) (2009) : 210-224.

[14] M. Anandhavalli. Association rule mining in genomics. International Journal of Computer Theory and Engineering 2(2) (2010): 1793-8201.

[15] U. M. Fayyad and K. B. Irani. Multi-interval discretizing of continuous-valued attributes for classification learning. Proceedings of the International Joint Conference on Uncertainty in AI, Chambery, France, 1993 : pp.1022-1027. France: Morgan Kaufmann, 1993.

[16] Z. Zhang, A. Teo, B. C. Ooi and K. L. Tan. Mining deterministic biclusters in gene expression data. IEEE Computer Society (2004) : 283-290.

[17] C. Creighton and S. Hanash. Mining gene expression databases for association rules. Bioinformatics 19(1) (2003) : 79-86.

[18] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. M. Carazo and A. Pascual-Montano. Integrated analysis of gene expression by association rules discovery. BMC Bioinformatics 7(54) (2006) : doi:10.1186/1471-2105-7-54.

[19] G. Cong, A. K. H. Tung, X. Xu, F. Pan and J. Yang. FARMER: finding interesting rule groups in microarray datasets. SIGMOD (2004) : 10.1145/1007568.1007587.

[20] G. Cong, K.L. Tan, A. K. H. Tung and X. Xu. Mining top-k covering rule groups for gene expression data. SIGMOD (2005) : 670-681.

[21] T. Mcintosh and S. Chawla. High-confidence rule mining for microarray analysis. IEEE/ACM Transaction on Computational Biology and Bioinformatics 4(4) (2007) : 611-623.

[22] F.J. Lopez, M. Cuadros, A. Blanco and A. Concha. Unveiling fuzzy associations between breast cancer prognostic factors and gene expression data. IEEE (2009) : 338-342.

[23] F. Berzal, J. C. Cubero, N. Marin, D. Sanchez, J. M. Serrano and A. Vila. Association rule evaluation for classification purposes. TAMIDA (2005) : 135-144.

[24] C. Becquet, S. Blachon, B. Jeudy, J. F. Boulicaut and O. Gandrillon. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. Genome Biology 3(12) (2002) : doi:10.1186/gb-2002-3-12-research0067.

[25] H. Nam, K. Lee and D. Lee. Identification of temporal association rule from time-series microarray data sets. BMC Bioinformatics 10(3) (2009) : doi: 10.1186/1471-2105-10-S3-S6.

[26] Gene Ontology : online

[27] National Center for Biotechnology Information : online.

[28] R. Martinez, N. Pasquier and C. Pasquier. Mining association rule bases from integrated genomic data and annotations. Computational Intelligence Methods for Bioinformatics and Biostatistics (2009) : 78-90.

[29] J. Pattamadilok, N. Huapai, P. Rattanatanyong, A. Vasurattana, S. Triratanachat, D. Tresukosol and A. Mutirangura. LINE-1 hypomethylation level as a potential prognostic factor for epithelial ovarian cancer. Int J Gynecol Cancer 18 (2008) : 711-717.

[30] K. Chalitchagorn, S. Shuangshoti, N. Hourpai, N. Kongruttanachok, P. Tangkijvanich, D. Thong-ngam, N. Voravud, V. Sriuranpong and A. Mutirangura. Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. Oncogene 23 (2004) : 8841-8846.

[31] N. Pratanwanich, C. Aporntewan and A. Mutirangura. Mining LINE-1 characteristics that mediate gene expression. CCIS, CSBio 2010 115 (2010) : 83-93.

[32] R. Hastings. Developing informatics resources for LINE-1 retrotransposons. Doctoral dissertation, Department of Genetics University of Leicester, 2009.

[33] J. Han, S. Szak and J. Boeke. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429(6989) (2004) : 268-274.

[34] T. Tchenio, J. F. Casella and T. Heidmann. Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Research 28(2) (2000) : 411-415.

[35] N. Yang, L. Zhang and H. H. Kazazian Junior. An important role for RUNX3 in human L1 transcription and retrotransposition. Nucleic Acids Research 31(16) (2003) : 4929-4940.

Appendices

Appendix A

LINE-1 characteristics description from L1Base [7]

Nominal characteristics

L1 characteristics	L1 part	Description
Type	All	Type of L1 FLnI L1 : Full length non-intact L1 FLI L1 : Full length intact L1 (intact both ORF1&2) ORF2 L1 : intact ORF2 but disruptedORF1
Strand	All	The strand of L1 element -1 : Transcription from 5'UTR to 3' UTR +1 : Transcription from 3'UTR to 5' UTR
L1M/L1PA Discrimination	All	The family of L1 element Mammalian L1M Primate L1PA
PolyA Signal	3' UTR	The intactness of sequence of PolyA Signal
Runx3 Site	5' UTR	The intactness of presence of RUNX3 binding motif in the 5' UTR
Runx3 ASP	5' UTR	The intactness of presence of RUNX3 Anti-Sense-Promoter binding motif in the 5' UTR
SRY Site1	5' UTR	The intactness of presence of first SRY1 binding motif in the 5' UTR
SRY Site2	5' UTR	The intactness of presence of second SRY1 binding motif in the 5' UTR
YY1 BoxA+BoxA	5' UTR	The intactness of presence of YY1 binding motif at the very beginning of the L1 Element
TF nkx-2.5	5' UTR	The intactness of sequence of the first putative nkx2-5 site
TF nkx-2.5B	5' UTR	The intactness of sequence of the second putative nkx2-5 site

L1 characteristics	L1 part	Description
REKG235	ORF1	The intactness of sequence of REKG235
ARR260	ORF1	The intactness of sequence of ARR260
YPAKLS282	ORF1	The intactness of sequence of YPAKLS282
N14	ORF2	The intactness of sequence of N14
E43	ORF2	The intactness of sequence of E43
Y115	ORF2	The intactness of sequence of Y115
D145	ORF2	The intactness of sequence of D145
N147	ORF2	The intactness of sequence of N147
T192	ORF2	The intactness of sequence of T192
D205	ORF2	The intactness of sequence of D205
SDH228	ORF2	The intactness of sequence of SDH228
R363	ORF2	The intactness of sequence of R363
FADD700	ORF2	The intactness of sequence of FADD700
HMKK1091	ORF2	The intactness of sequence of HMKK1091
SSS1096	ORF2	The intactness of sequence of SSS1096
I1220	ORF2	The intactness of sequence of I1220
S1259	ORF2	The intactness of sequence of S1259
ORF StartStop	ORF	The presence of valid methionine start codons and stop codons of the ORFs of the L1 mut : mutate cons : conserved on ORF1 and ORF2 ORF1 cons: conserved on ORF1 ORF2 cons: conserved on ORF2
Orientation	All	The orientation of hosting gene + : Transcription from 5'UTR to 3' UTR - : Transcription from 3'UTR to 5' UTR

The LINE-1 characteristics, that are not specified their values in the table above, have to values, "cons" and "mut", which refer to conserved and mutated sequence, respectively.

Numeric charateristics

L1 characteristics	L1 part	Description
ORF1 Gaps	ORF1	The number of gaps in ORF1
ORF1 Frameshifts	ORF1	The number of frameshifts in ORF1
ORF1 Stops	ORF1	The number of stops in ORF1
ORF2 Gaps	ORF2	The number of gaps in ORF2
ORF2 Frameshifts	ORF2	The number of frameshifts in ORF2
ORF2 Stops	ORF2	The number of stops in ORF2
Poly-A pure	3' UTR	the length of the pure Poly-A tail
Poly-A est	3' UTR	the length of an estimated Poly-A Tail (containing mutations)
Find TSDs	All	The number of Target-Site-Duplications flanking the L1 Element
G.C.Content	All	The percentage of base pair 'G-C' of L1
ORF1 %A	ORF1	The percentage of base 'A' found in ORF1 of L1
ORF1 %T	ORF1	The percentage of base 'T' found in ORF2 of L1
ORF2 %A	ORF2	The percentage of base 'A' found in ORF2 of L1
ORF2 %T	ORF2	The percentage of base 'T' found in ORF2 of L1
ORF1&2 %A	ORF	The percentage of base 'A' found in ORF1 and ORF2 of L1
ORF1&2 %T	ORF	The percentage of base 'T' found in ORF1 and ORF2 of L1
CPG Islands	All	The number of GpG Islands found in L1
ORF1 CAI	ORF1	The codon adaptation index of ORF1
ORF2 CAI	ORF2	The codon adaptation index of ORF2
Intactness score	All	The score determining the intactness of L1 elements
Number of L1	All	The number of L1 elements found in the host gene

Appendix B

Final rules from classification association rules mining

GSE3167 Bladder cancer

Rule 1		Rule 2		Rule 3		Rule 4	
Support	4.74%	Support	4.87%	Support	5.18%	Support	5.30%
Confidence	71.96%	Confidence	71.17%	Confidence	70.59%	Confidence	70.49%
Odds ratio	3.49	Odds ratio	3.36	Odds ratio	3.28	Odds ratio	3.28
p-value	2.52E-09	p-value	3.54E-09	p-value	2.01E-09	p-value	1.40E-09
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[number of L1] > 2.500	ALL	[number of L1] > 2.500	ALL	[number of L1] > 2.500	ALL	[number of L1] > 2.500	ALL
[D205] = 'cons'	ORF2	[S1259] = 'cons'	ORF2	[S1259] = 'cons'	ORF2	[D205] = 'cons'	ORF2
[ORF1 Frameshifts] > 1.500	ORF1	[ORF1 %A] > 0.405	ORF1	[ORF1 %A] > 0.405	ORF1	[ORF1 Frameshifts] > 1.500	ORF1
[ORF1 %T] > 0.182	ORF1	[REKG235] = 'mut'	ORF1	[REKG235] = 'mut'	ORF1	[Poly-A pure] < 2.500	3' UTR
[ORF2 %T] > 0.205	ORF2	[Runx3 ASP] = 'mut'	5' UTR			[G-C Content] < 40.475	ALL
[ORF2 CAI] < 0.634	ORF2						
[ORF1&2 %T] > 0.200	ORF1&2						

GSE3167 Bladder cancer (continued)

Rule 5		Rule 6		Rule 7		Rule 8	
Support	5.42%	Support	5.79%	Support	5.79%	Support	5.98%
Confidence	69.29%	Confidence	68.12%	Confidence	68.12%	Confidence	67.83%
Odds ratio	3.09	Odds ratio	2.94	Odds ratio	2.94	Odds ratio	2.91
p-value	3.52E-09	p-value	3.91E-09	p-value	3.91E-09	p-value	2.99E-09
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[D205] = 'cons'	ORF 2	[D205] = 'cons'	ORF 2	[T192] = 'cons'	ORF 2	[D205] = 'cons'	ORF 2
[Runx3 ASP] = 'mut'	5' UTR	[ORF2 %A] > 0.411	ORF 2	[Runx3 ASP] = 'mut'	5' UTR	[ORF1&2 %A] > 0.411	ORF
[ORF1&2 %A] > 0.411	ORF	[ORF1&2 %T] > 0.200	ORF	[ORF1&2 %A] > 0.411	ORF	[ORF2 %A] > 0.411	ORF 2
[ORF2 %A] > 0.411	ORF 2	[ORF1 %T] > 0.182	ORF 1	[ORF2 %A] > 0.411	ORF 2	[ORF1&2 %T] > 0.200	ORF
[G-C Content] < 40.475	ALL	[number of L1] > 2.500	ALL	[G-C Content] < 40.475	ALL	[number of L1] > 2.500	ALL
[number of L1] > 2.500	ALL	[ORF1 Frameshifts] > 1.500	ORF 1	[number of L1] > 2.500	ALL	[ORF1 Frameshifts] > 1.500	ORF 1
[ORF1 Frameshifts] > 1.500	ORF1			[ORF1 Frameshifts] > 1.500	ORF 1		

GSE3167 Bladder cancer (continued)

Rule 9		Rule 10		Rule 11		Rule 12	
Support	6.04%	Support	7.21%	Support	7.83%	Support	8.44%
Confidence	67.59%	Confidence	65.00%	Confidence	64.14%	Confidence	62.84%
Odds ratio	2.88	Odds ratio	2.59	Odds ratio	2.52	Odds ratio	2.39
p-value	3.32E-09	p-value	3.05E-09	p-value	2.00E-09	p-value	3.14E-09
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[D205] = 'cons'	ORF 2	[D205] = 'cons'	ORF 2	[TF nkx-2.5B] = 'mut'	5' UTR	[T192] = 'cons'	ORF 2
[ORF2 CAI] < 0.634	ORF 2	[Runx3 ASP] = 'mut'	5' UTR	[S1259] = 'cons'	ORF 2	[D205] = 'cons'	ORF 2
[number of L1] > 2.500	ALL	[ORF1 %A] > 0.405	ORF 1	[Runx3 ASP] = 'mut'	5' UTR	[Runx3 ASP] = 'mut'	5' UTR
[ORF1 Frameshifts] > 1.500	ORF 1	[number of L1] > 2.500	ALL	[ORF1&2 %A] > 0.411	ORF	[ORF1&2 %A] > 0.411	ORF
				[number of L1] > 2.500	ALL	[number of L1] > 2.500	ALL

GSE3167 Bladder cancer (continued)

Rule 13		Rule 14		Rule 15	
Support	8.75%	Support	10.91%	Support	11.15%
Confidence	62.56%	Confidence	59.80%	Confidence	59.54%
Odds ratio	2.37	Odds ratio	2.15	Odds ratio	2.14
p-value	2.37E-09	p-value	2.92E-09	p-value	2.97E-09
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[TF nkx-2.5B] = 'mut'	5' UTR	[D205] = 'cons'	ORF 2	[TF nkx-2.5B] = 'mut'	5' UTR
[S1259] = 'cons'	ORF 2	[ORF2 %A] > 0.411	ORF 2	[S1259] = 'cons'	ORF 2
[ORF1&2 %A] > 0.411	ORF	[number of L1] > 2.500	ALL	[Poly-A pure] < 2.500	3' UTR
[number of L1] > 2.500	ALL			[number of L1] > 2.500	ALL

GSE14811 Liver cancer

Rule 1		Rule 2		Rule 3		Rule 4	
Support	1.09%	Support	1.09%	Support	1.09%	Support	1.27%
Confidence	100.00%	Confidence	100.00%	Confidence	100.00%	Confidence	87.50%
Odds ratio	NA	Odds ratio	NA	Odds ratio	NA	Odds ratio	44.9726
p-value	7.14E-08	p-value	7.14E-08	p-value	7.14E-08	p-value	7.02E-08
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[Y115] = 'cons'	ORF 2	[Y115] = 'cons'	ORF 2	[D205] = 'cons'	ORF 2	[Y115] = 'cons'	ORF 2
[ORF2 Stops] > 5.500	ORF 2	[R363] = 'cons'	ORF 2	[ORF1 Frameshifts] > 1.500	ORF 1	[SDH228] = 'cons'	ORF 2
[G-C Content] > 40.585	ALL	[ORF2 Stops] > 5.500	ORF 2	[find TSDs] < 11.500	ALL	[ORF2 Stops] > 5.500	ORF 2
[ORF2 %T] < 0.204	ORF 2	[ORF2 %T] < 0.204	ORF 2	[ORF2 %T] < 0.204	ORF 2	[G-C Content] > 40.585	ALL
[Poly-A est] > 15.500	3' UTR	[Poly-A est] > 15.500	3' UTR	[ORF StartStop] = 'mut'	ORF	[ORF2 %T] < 0.204	ORF 2
[SRY Site 2] = 'cons'	5' UTR	[SRY Site 2] = 'cons'	5' UTR			[SRY Site 2] = 'cons'	5' UTR

GSE14811 Liver cancer (continued)

Rule 5		Rule 6	
Support	1.45%	Support	1.45%
Confidence	80.00%	Confidence	80.00%
Odds ratio	26.00	Odds ratio	26.00
p-value	4.44E-08	p-value	4.44E-08
L1 characteristics	L1 part	L1 characteristics	L1 part
[N14] = 'cons'	ORF 2	[CPG Islands] < 0.500	ALL
[YY1 BoxA+BoxA] = 'mut'	5' UTR	[YY1 BoxA+BoxA] = 'mut'	5' UTR
[Y115] = 'cons'	ORF 2	[Y115] = 'cons'	ORF 2
[SDH228] = 'cons'	ORF 2	[SDH228] = 'cons'	ORF 2
[Runx3 ASP] = 'mut'	5' UTR	[Runx3 ASP] = 'mut'	5' UTR
[ORF1 Frameshifts] > 1.500	ORF 1	[ORF1 Frameshifts] > 1.500	ORF 1
[G-C Content] > 40.585	ALL	[G-C Content] > 40.585	ALL
[find TSDs] > 11.500	ALL	[find TSDs] > 11.500	ALL
[ORF2 Stops] < 5.500	ORF 2	[ORF2 Stops] < 5.500	ORF 2

GSE6919 Prostate cancer

Rule 1		Rule 2		Rule 3		Rule 4	
Support	0.53%	Support	0.53%	Support	0.53%	Support	0.53%
Confidence	100.00%	Confidence	100.00%	Confidence	100.00%	Confidence	100.00%
Odds ratio	NA	p-value	NA	Odds ratio	NA	Odds ratio	NA
p-value	7.12E-08	Odds ratio	7.12E-08	p-value	7.12E-08	p-value	7.12E-08
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[T192] = 'cons'	ORF 2	[T192] = 'cons'	ORF 2	[R363] = 'cons'	ORF 2	[ARR260] = 'cons'	ORF 1
[S1259] = 'cons'	ORF 2	[N147] = 'cons'	ORF 2	[orientation] = '+'	ALL	[R363] = 'cons'	ORF 2
[YPAKLS282] = 'cons'	ORF 1	[YPAKLS282] = 'cons'	ORF 1	[HMKK1091] = 'cons'	ORF 2	[orientation] = '+'	ALL
[orientation] = '+'	ALL	[orientation] = '+'	ALL	[SRY Site 2] = 'mut'	5' UTR	[HMKK1091] = 'cons'	ORF 2
[ORF1 Stops] < 2.500	ORF 1	[ORF1 Stops] < 2.500	ORF 1	[ORF1 Stops] < 2.500	ORF 1	[ORF1 Stops] < 2.500	ORF 1
[ORF2 CAI] < 0.634	ORF 2	[ORF2 CAI] < 0.634	ORF 2	[Poly-A pure] > 2.500	3' UTR	[Poly-A pure] > 2.500	3' UTR
[ORF1 CAI] < 0.668	ORF 1	[ORF1 CAI] < 0.668	ORF 1	[SSS1096] = 'mut'	ORF 2	[SSS1096] = 'mut'	ORF 2
[SSS1096] = 'mut'	ORF 2	[SSS1096] = 'mut'	ORF 2	[ORF2 Stops] > 5.500	ORF 2	[ORF2 Stops] > 5.500	ORF 2

GSE6919 Prostate cancer (Continued)

Rule 5		Rule 6		Rule 7		Rule 8	
Support	0.62%	Support	0.62%	Support	0.71%	Support	0.71%
Confidence	87.50%	Confidence	87.50%	Confidence	80.00%	Confidence	80.00%
Odds ratio	43.16	Odds ratio	43.15924	Odds ratio	24.80	Odds ratio	24.80
p-value	7.17E-08	p-value	7.17E-08	p-value	4.66E-08	p-value	4.66E-08
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[T192] = 'cons'	ORF 2	[ARR260] = 'cons'	ORF 1	[S1259] = 'cons'	ORF 2	[ORF1&2 %T] < 0.200	ORF
[N147] = 'cons'	ORF 2	[ORF2 Stops] < 5.500	ORF 2	[ORF1&2 %T] < 0.200	ORF	[HMKK1091] = 'cons'	ORF 2
[S1259] = 'cons'	ORF 2	[Strand] = '-1'	ALL	[Strand] = '-1'	ALL	[Strand] = '-1'	ALL
[ARR260] = 'cons'	ORF 1	[ORF1 CAI] > 0.668	ORF 1	[ORF1 CAI] > 0.668	ORF 1	[ORF1 CAI] > 0.668	ORF 1
[Strand] = '-1'	ALL	[find TSDs] > 11.500	ALL	[find TSDs] > 11.500	ALL	[find TSDs] > 11.500	ALL
[ORF1 Stops] < 2.500	ORF 1	[ORF1 Stops] > 2.500	ORF 1	[ORF1 Stops] > 2.500	ORF 1	[ORF1 Stops] > 2.500	ORF 1
[ORF2 CAI] < 0.634	ORF 2	[SRY Site 2] = 'cons'	5' UTR	[SRY Site 2] = 'cons'	5' UTR	[SRY Site 2] = 'cons'	5' UTR
[ORF1 CAI] < 0.668	ORF 1						
[SSS1096] = 'mut'	ORF 2						

GSE6919 Prostate cancer (Continued)

Rule 9		Rule 10		Rule 11		Rule 12	
Support	0.71%	Support	0.71%	Support	0.71%	Support	0.71%
Confidence	80.00%	Confidence	80.00%	Confidence	80.00%	Confidence	80.00%
Odds ratio	24.80	Odds ratio	24.80	Odds ratio	24.80	Odds ratio	24.80
p-value	4.66E-08	p-value	4.66E-08	p-value	4.66E-08	p-value	4.66E-08
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[ORF1&2 %T] < 0.200	ORF	[ARR260] = 'cons'	ORF 1	[T192] = 'cons'	ORF 2	[ARR260] = 'cons'	ORF 1
[Strand] = '-1'	ALL	[HMKK1091] = 'cons'	ORF 2	[ARR260] = 'cons'	ORF 1	[ORF1 Gaps] < 1.500	ORF 1
[ORF1 CAI] > 0.668	ORF 1	[Strand] = '-1'	ALL	[Strand] = '-1'	ALL	[Strand] = '-1'	ALL
[find TSDs] > 11.500	ALL	[ORF1 CAI] > 0.668	ORF 1	[ORF1 CAI] > 0.668	ORF 1	[ORF1 CAI] > 0.668	ORF 1
[Intactness score] > 17.500	ALL	[find TSDs] > 11.500	ALL	[find TSDs] > 11.500	ALL	[find TSDs] > 11.500	ALL
[ORF1 Stops] > 2.500	ORF 1	[ORF1 Stops] > 2.500	ORF 1	[ORF1 Stops] > 2.500	ORF 1	[ORF1 Stops] > 2.500	ORF 1
[SRY Site 2] = 'cons'	5' UTR	[SRY Site 2] = 'cons'	5' UTR	[SRY Site 2] = 'cons'	5' UTR	[SRY Site 2] = 'cons'	5' UTR

GSE6919 Prostate cancer (Continued)

Rule 13		Rule 14		Rule 15		Rule 16	
Support	0.71%	Support	0.79%	Support	0.79%	Support	0.97%
Confidence	80.00%	Confidence	75.00%	Confidence	75.00%	Confidence	64.71%
Odds ratio	24.79	Odds ratio	18.70	Odds ratio	18.70	Odds ratio	11.54
p-value	4.66E-08	p-value	3.03E-08	p-value	3.03E-08	p-value	3.03E-08
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[ARR260] = 'cons'	ORF 1	[T192] = 'cons'	ORF 2	[YPAKLS282] = 'cons'	ORF 1	[L1M/L1PA Discrimination] = 'Primate L1PA'	ALL
[ORF1&2 %T] < 0.200	ORF	[ARR260] = 'cons'	ORF 1	[R363] = 'cons'	ORF 2	[T192] = 'cons'	ORF 2
[Strand] = '-1'	ALL	[YPAKLS282] = 'cons'	ORF 1	[ORF1 Gaps] < 1.500	ORF 1	[S1259] = 'cons'	ORF 2
[ORF1 CAI] > 0.668	ORF 1	[Strand] = '-1'	ALL	[ORF1&2 %T] < 0.200	ORF	[ARR260] = 'cons'	ORF 1
[find TSDs] > 11.500	ALL	[ORF1 CAI] > 0.668	ORF 1	[ORF2 Stops] < 5.500	ORF 2	[ORF1 Gaps] < 1.500	ORF 1
[ORF1 Stops] > 2.500	ORF 1	[number of L1] < 2.500	ALL	[SRY Site 2] = 'mut'	5' UTR	[ORF1 Stops] > 2.500	ORF 1
[SRY Site 2] = 'cons'	5' UTR	[ORF StartStop] = 'cons'	ORF	[ORF1&2 %A] < 0.411	ORF	[number of L1] > 2.500	ALL
		[ORF1 Stops] > 2.500	ORF 1	[Intactness score] < 17.500	ALL	[SSS1096] = 'mut'	ORF 2
				[Strand] = '1'	ALL		

GSE6631 Head and neck cancer

Rule 1		Rule 2		Rule 3		Rule 4	
Support	0.53%	Support	0.53%	Support	0.53%	Support	0.44%
Confidence	75.00%	Confidence	75.00%	Confidence	75.00%	Confidence	71.43%
Odds ratio	51.48	Odds ratio	51.48	Odds ratio	51.48	Odds ratio	42.22
p-value	6.27E-14	p-value	6.27E-14	p-value	6.27E-14	p-value	7.23E-11
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[Runx3 Site] = 'mut'	5' UTR	[Runx3 Site] = 'mut'	5' UTR	[CPG Islands] < 0.500	ALL	[CPG Islands] < 0.500	ALL
[Runx3 ASP] = 'mut'	5' UTR	[orientation] = '-'	ALL	[Runx3 Site] = 'mut'	5' UTR	[Runx3 Site] = 'mut'	5' UTR
[orientation] = '-'	ALL	[number of L1] < 2.500	ALL	[FADD700] = 'cons'	ORF 2	[orientation] = '-'	ALL
[number of L1] < 2.500	ALL	[TF nkx-2.5] = 'cons'	5' UTR	[orientation] = '-'	ALL	[TF nkx-2.5] = 'cons'	5' UTR
[ORF1 Gaps] < 1.500	ORF 1	[ORF1 %A] > 0.405	ORF 1	[TF nkx-2.5] = 'cons'	5' UTR	[ORF1 %A] > 0.405	ORF 1
[TF nkx-2.5] = 'cons'	5' UTR	[find TSDs] < 11.500	ALL	[ORF1 %A] > 0.405	ORF 1	[find TSDs] < 11.500	ALL
[ORF1 %A] > 0.405	ORF 1	[ORF1 Stops] < 2.500	ORF 1	[ORF1 Stops] < 2.500	ORF 1	[ORF1 Stops] < 2.500	ORF 1
[ORF1 Stops] < 2.500	ORF 1	[ORF2 %T] < 0.205	ORF 2	[ORF2 %T] < 0.205	ORF 2	[ORF2 %T] < 0.205	ORF 2
[ORF2 %T] < 0.205	ORF 2			[ORF2 %A] < 0.411	ORF 2	[ORF2 %A] < 0.411	ORF 2
				[ORF2 Gaps] < 5.500	ORF 2		

GSE6631 Head and neck cancer (Continued)

Rule 5		Rule 6		Rule 7		Rule 8	
Support	0.44%	Support	0.44%	Support	0.44%	Support	0.62%
Confidence	71.43%	Confidence	71.43%	Confidence	71.43%	Confidence	70.00%
Odds ratio	42.22	Odds ratio	42.22	Odds ratio	42.22	Odds ratio	40.66
p-value	7.23E-11	p-value	7.23E-11	p-value	7.23E-11	p-value	2.93E-15
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[Runx3 Site] = 'mut'	5' UTR	[D145] = 'cons'	ORF 2	[E43] = 'cons'	ORF 2	[Runx3 Site] = 'mut'	5' UTR
[orientation] = '-'	ALL	[E43] = 'cons'	ORF 2	[Runx3 Site] = 'mut'	5' UTR	[FADD700] = 'cons'	ORF 2
[TF nkx-2.5] = 'cons'	5' UTR	[Runx3 Site] = 'mut'	5' UTR	[FADD700] = 'cons'	ORF 2	[Runx3 ASP] = 'mut'	5' UTR
[ORF1 %A] > 0.405	ORF 1	[FADD700] = 'cons'	ORF 2	[number of L1] < 2.500	ALL	[R363] = 'cons'	ORF 2
[find TSDs] < 11.500	ALL	[Runx3 ASP] = 'mut'	5' UTR	[ORF1 Gaps] < 1.500	ORF 1	[number of L1] < 2.500	ALL
[ORF1 Stops] < 2.500	ORF 1	[number of L1] < 2.500	ALL	[TF nkx-2.5] = 'cons'	5' UTR	[ORF1 %A] > 0.405	ORF 1
[ORF1 %T] < 0.182	ORF 1	[ORF1 Gaps] < 1.500	ORF 1	[ORF1 %A] > 0.405	ORF 1	[find TSDs] < 11.500	ALL
[ORF2 %T] < 0.205	ORF 2	[TF nkx-2.5] = 'cons'	5' UTR	[find TSDs] < 11.500	ALL	[ORF1 Stops] < 2.500	ORF 1
[ORF2 %A] < 0.411	ORF 2	[ORF1 %A] > 0.405	ORF 1	[ORF1 Stops] < 2.500	ORF 1	[ORF2 %A] < 0.411	ORF 2
		[ORF1 Stops] < 2.500	ORF 1	[ORF2 %T] < 0.205	ORF 2		
		[ORF1 %T] < 0.182	ORF 1				
		[ORF2 %T] < 0.205	ORF 2				

GSE6631 Head and neck cancer (Continued)

Rule 9		Rule 10		Rule 11		Rule 12	
Support	0.62%	Support	0.62%	Support	0.62%	Support	0.62%
Confidence	70.00%	Confidence	70.00%	Confidence	70.00%	Confidence	70.00%
Odds ratio	40.66	Odds ratio	40.66	Odds ratio	40.66	Odds ratio	40.66
p-value	2.93E-15	p-value	2.93E-15	p-value	2.93E-15	p-value	2.93E-15
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[CPG Islands] < 0.500	ALL	[Runx3 Site] = 'mut'	5' UTR	[CPG Islands] < 0.500	ALL	[Runx3 Site] = 'mut'	5' UTR
[Runx3 ASP] = 'mut'	5' UTR	[Runx3 ASP] = 'mut'	5' UTR	[Runx3 Site] = 'mut'	5' UTR	[orientation] = '-'	ALL
[ORF1 Gaps] < 1.500	ORF 1	[orientation] = '-'	ALL	[orientation] = '-'	ALL	[TF nkx-2.5] = 'cons'	5' UTR
[TF nkx-2.5] = 'cons'	5' UTR	[ORF1 Gaps] < 1.500	ORF 1	[TF nkx-2.5] = 'cons'	5' UTR	[ORF1 %A] > 0.405	ORF 1
[ORF1 Stops] < 2.500	ORF 1	[TF nkx-2.5] = 'cons'	5' UTR	[ORF1 %A] > 0.405	ORF 1	[find TSDs] < 11.500	ALL
[ORF1 %T] < 0.182	ORF 1	[ORF1 %A] > 0.405	ORF 1	[find TSDs] < 11.500	ALL	[ORF1 Stops] < 2.500	ORF 1
[ORF StartStop] = 'ORF2 cons'	ORF	[ORF1 Stops] < 2.500	ORF 1	[ORF1 Stops] < 2.500	ORF 1	[ORF1 %T] < 0.182	ORF 1
		[ORF1 %T] < 0.182	ORF 1	[ORF2 %A] < 0.411	ORF 2	[ORF2 %A] < 0.411	ORF 2
		[ORF2 Stops] < 5.500	ORF 2				

6631 (continued)

Rule 13		Rule 14		Rule 15		Rule 16	
Support	0.62%	Support	0.62%	Support	0.62%	Support	0.62%
Confidence	70.00%	Confidence	70.00%	Confidence	70.00%	Confidence	70.00%
Odds ratio	40.66	Odds ratio	40.66	Odds ratio	40.66	Odds ratio	40.66
p-value	2.93E-15	p-value	2.93E-15	p-value	2.93E-15	p-value	2.93E-15
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[N14] = 'cons'	ORF 2	[CPG Islands] < 0.500	ALL	[CPG Islands] < 0.500	ALL	[E43] = 'cons'	ORF 2
[Runx3 Site] = 'mut'	5' UTR	[Runx3 Site] = 'mut'	5' UTR	[Runx3 Site] = 'mut'	5' UTR	[Runx3 Site] = 'mut'	5' UTR
[FADD700] = 'cons'	ORF 2	[Runx3 ASP] = 'mut'	5' UTR	[Runx3 ASP] = 'mut'	5' UTR	[FADD700] = 'cons'	ORF 2
[Runx3 ASP] = 'mut'	5' UTR	[orientation] = '-'	ALL	[orientation] = '-'	ALL	[number of L1] < 2.500	ALL
[number of L1] < 2.500	ALL	[ORF1 Gaps] < 1.500	ORF 1	[TF nkx-2.5] = 'cons'	5' UTR	[TF nkx-2.5] = 'cons'	5' UTR
[ORF1 %A] > 0.405	ORF 1	[TF nkx-2.5] = 'cons'	5' UTR	[ORF1 %A] > 0.405	ORF 1	[ORF1 %A] > 0.405	ORF 1
[find TSDs] < 11.500	ALL	[ORF1 %A] > 0.405	ORF 1	[ORF1 Stops] < 2.500	ORF 1	[find TSDs] < 11.500	ALL
[ORF1 Stops] < 2.500	ORF 1	[ORF1 Stops] < 2.500	ORF 1	[ORF2 %A] < 0.411	ORF 2	[ORF1 Stops] < 2.500	ORF 1
[ORF2 %A] < 0.411	ORF 2	[ORF2 Stops] < 5.500	ORF 2			[ORF2 CAI] > 0.634	ORF 2

GSE6631 Head and neck cancer (Continued)

Rule 17		Rule 18		Rule 19		Rule 20	
Support	0.53%	Support	0.62%	Support	0.62%	Support	0.62%
Confidence	66.67%	Confidence	63.64%	Confidence	63.64%	Confidence	63.64%
Odds ratio	34.29	Odds ratio	30.46721	Odds ratio	30.47	Odds ratio	30.47
p-value	2.78E-12	p-value	9.11E-14	p-value	9.11E-14	p-value	9.11E-14
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[E43] = 'cons'	ORF 2	[E43] = 'cons'	ORF 2	[N147] = 'cons'	ORF 2	[D145] = 'cons'	ORF 2
[FADD700] = 'cons'	ORF 2	[FADD700] = 'cons'	ORF 2	[D145] = 'cons'	ORF 2	[orientation] = '-'	ALL
[number of L1] < 2.500	ALL	[number of L1] < 2.500	ALL	[CPG Islands] < 0.500	ALL	[number of L1] < 2.500	ALL
[ORF1 %A] > 0.405	ORF 1	[ORF1 %A] > 0.405	ORF 1	[orientation] = '-'	ALL	[TF nkx-2.5] = 'cons'	5' UTR
[ORF1 Stops] < 2.500	ORF 1	[find TSDs] < 11.500	ALL	[find TSDs] < 11.500	ALL	[ORF1 %A] > 0.405	ORF 1
[ORF2 %T] < 0.205	ORF 2	[ORF1 Stops] < 2.500	ORF 1	[ORF1 %T] < 0.182	ORF 1	[find TSDs] < 11.500	ALL
[SRY Site 1] = 'mut'	5' UTR	[SRY Site 1] = 'mut'	5' UTR	[ORF1&2 %A] < 0.411	ORF	[ORF1 Stops] < 2.500	ORF 1
				[ORF2 Stops] > 5.500	ORF 2	[ORF2 Gaps] < 5.500	ORF 2

GSE6631 Head and neck cancer (Continued)

Rule 21		Rule 22		Rule 23	
Support	0.62%	Support	0.62%	Support	0.62%
Confidence	58.33%	Confidence	58.33%	Confidence	53.85%
Odds ratio	24.35	Odds ratio	24.35	Odds ratio	20.27
p-value	1.60E-12	p-value	1.60E-12	p-value	1.80E-11
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[CPG Islands] < 0.500	ALL	[D145] = 'cons'	ORF 2	[D145] = 'cons'	ORF 2
[Runx3 Site] = 'mut'	5' UTR	[CPG Islands] < 0.500	ALL	[N14] = 'cons'	ORF 2
[number of L1] < 2.500	ALL	[FADD700] = 'cons'	ORF 2	[CPG Islands] < 0.500	ALL
[ORF1 Gaps] < 1.500	ORF 1	[R363] = 'cons'	ORF 2	[Runx3 Site] = 'mut'	5' UTR
[ORF1 %T] < 0.182	ORF 1	[orientation] = '-'	ALL	[Runx3 ASP] = 'mut'	5' UTR
[ORF2 %A] < 0.411	ORF 2	[number of L1] < 2.500	ALL	[orientation] = '-'	ALL
[ORF2 Stops] > 5.500	ORF 2	[ORF1 Gaps] < 1.500	ORF 1	[ORF1 Gaps] < 1.500	ORF 1
[CPG Islands] < 0.500	ALL	[TF nkx-2.5] = 'cons'	5' UTR	[SRY Site 1] = 'cons'	5' UTR
		[ORF1 %A] > 0.405	ORF 1	[ORF2 %T] < 0.205	ORF 2
		[ORF1 Stops] < 2.500	ORF 1	[ORF2 CAI] < 0.634	ORF 2
		[ORF1 %T] < 0.182	ORF 1		
		[ORF2 Gaps] < 5.500	ORF 2		

GSE5816 hBEC Lung

Rule 1		Rule 2		Rule 3		Rule 4	
Support	0.11%	Support	0.11%	Support	0.11%	Support	0.11%
Confidence	100.00%	Confidence	100.00%	Confidence	100.00%	Confidence	100.00%
Odds ratio	NA	Odds ratio	NA	Odds ratio	NA	Odds ratio	NA
p-value	6.34E-51	p-value	6.34E-51	p-value	6.34E-51	p-value	6.34E-51
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[D205] = 'cons'	ORF 2	[D205] = 'cons'	ORF 2	[D205] = 'cons'	ORF 2	[D205] = 'cons'	ORF 2
[S1259] = 'cons'	ORF 2	[S1259] = 'cons'	ORF 2	[S1259] = 'cons'	ORF 2	[S1259] = 'cons'	ORF 2
[SRY Site 1] = 'cons'	5' UTR	[SRY Site 1] = 'cons'	5' UTR	[SRY Site 1] = 'cons'	5' UTR	[SRY Site 1] = 'cons'	5' UTR
[orientation] = '+'	ALL	[orientation] = '+'	ALL	[orientation] = '+'	ALL	[ORF1 Gaps] < 1.500	ORF 1
[ORF1 Gaps] < 1.500	ORF 1	[ORF1 Gaps] < 1.500	ORF 1	[ORF1 Gaps] < 1.500	ORF 1	[Strand] = '-1'	ALL
[ORF2 Gaps] < 5.500	ORF 2	[Strand] = '-1'	ALL	[Strand] = '-1'	ALL	[ORF2 Gaps] < 5.500	ORF 2
[ORF1 %A] < 0.405	ORF 1	[ORF2 Gaps] < 5.500	ORF 2	[ORF1 %A] < 0.405	ORF 1	[ORF1 %A] < 0.405	ORF 1
[ORF2 %T] > 0.205	ORF 2	[ORF1 %A] < 0.405	ORF 1	[ORF2 %T] > 0.205	ORF 2	[ORF2 %T] > 0.205	ORF 2
[Intactness score] < 17.500	ALL	[ORF2 %T] > 0.205	ORF 2	[Intactness score] < 17.500	ALL	[Intactness score] < 17.500	ALL
[ORF StartStop] = 'ORF2 cons'	ORF	[ORF1 %T] > 0.183	ORF 1	[ORF StartStop] = 'ORF2 cons'	ORF	[ORF StartStop] = 'ORF2 cons'	ORF
		[ORF StartStop] = 'ORF2 cons'	ORF				

GSE5816 hBEC Lung (Continued)

Rule 5		Rule 6		Rule 7		Rule 8	
Support	0.11%	Support	0.11%	Support	0.11%	Support	0.11%
Confidence	100.00%	Confidence	100.00%	Confidence	100.00%	Confidence	100.00%
Odds ratio	NA	Odds ratio	NA	Odds ratio	NA	Odds ratio	NA
p-value	6.34E-51	p-value	6.34E-51	p-value	6.34E-51	p-value	6.34E-51
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[S1259] = 'cons'	ORF 2	[SRY Site 1] = 'cons'	5' UTR	[YPAKLS282] = 'cons'	ORF 1	[D205] = 'cons'	ORF 2
[SRY Site 1] = 'cons'	5' UTR	[ORF1 Gaps] < 1.500	ORF 1	[SRY Site 1] = 'cons'	5' UTR	[SRY Site 1] = 'cons'	5' UTR
[orientation] = '+'	ALL	[Strand] = '-1'	ALL	[orientation] = '+'	ALL	[orientation] = '+'	ALL
[ORF1 Gaps] < 1.500	ORF 1	[ORF2 Gaps] < 5.500	ORF 2	[ORF1 Gaps] < 1.500	ORF 1	[ORF1 Gaps] < 1.500	ORF 1
[Strand] = '-1'	ALL	[ORF1 Frameshifts] > 1.500	ORF 1	[ORF2 CAI] > 0.634	ORF 2	[ORF2 CAI] > 0.634	ORF 2
[ORF2 Gaps] < 5.500	ORF 2	[Intactness score] < 17.500	ALL	[ORF1&2 %A] > 0.411	ORF 1	[ORF1&2 %A] > 0.411	ORF
[ORF1 %T] > 0.183	ORF 1	[ORF StartStop] = 'ORF2 cons'	ORF	[HMKK1091] = 'mut'	ORF 2	[HMKK1091] = 'mut'	ORF 2
[ORF2 Stops] > 5.500	ORF 2			[I1220] = 'mut'	ORF 2	[I1220] = 'mut'	ORF 2
[ORF StartStop] = 'ORF2 cons'	ORF						

GSE5816 hBEC Lung (Continued)

Rule 9		Rule 10		Rule 11		Rule 12	
Support	0.08%	Support	0.08%	Support	0.08%	Support	0.11%
Confidence	100.00%	Confidence	100.00%	Confidence	100.00%	Confidence	75.00%
Odds ratio	NA	Odds ratio	NA	Odds ratio	NA	Odds ratio	371.5714
p-value	3.26E-28	p-value	3.26E-28	p-value	3.26E-28	p-value	2.25E-38
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[SRY Site 1] = 'cons'	5' UTR	[CPG Islands] < 0.500	ALL	[CPG Islands] < 0.500	ALL	[D205] = 'cons'	ORF 2
[ORF1 Gaps] < 1.500	ORF 1	[ORF2 Gaps] < 5.500	ORF 2	[YPAKLS282] = 'cons'	ORF 1	[SRY Site 1] = 'cons'	5' UTR
[ORF2 Gaps] < 5.500	ORF 2	[ORF1&2 %A] > 0.411	ORF	[ORF1&2 %A] > 0.411	ORF	[orientation] = '+'	ALL
[ORF1&2 %A] > 0.411	ORF	[ORF1 %A] < 0.405	ORF 1	[ORF1 %A] < 0.405	ORF 1	[Poly-A pure] < 2.500	3' UTR
[ORF1 %A] < 0.405	ORF 1	[ORF2 Stops] < 5.500	ORF 2	[ORF2 Stops] < 5.500	ORF 2	[ORF1 %A] < 0.405	ORF 1
[Intactness score] < 17.500	ALL	[ORF1 %T] < 0.183	ORF 1	[ORF1 %T] < 0.183	ORF 1	[HMKK1091] = 'cons'	ORF 2
[ORF StartStop] = 'ORF2 cons'	ORF	[HMKK1091] = 'mut'	ORF 2	[HMKK1091] = 'mut'	ORF 2	[ORF2 %T] > 0.205	ORF 2
						[ORF1 Frameshifts] > 1.500	ORF 1
						[ORF StartStop] = 'ORF2 cons'	ORF

GSE5816 hBEC Lung (Continued)

Rule 13		Rule 14		Rule 15		Rule 16	
Support	0.11%	Support	0.11%	Support	0.11%	Support	0.11%
Confidence	75.00%	Confidence	75.00%	Confidence	75.00%	Confidence	75.00%
Odds ratio	371.57	Odds ratio	371.57	Odds ratio	371.57	Odds ratio	371.57
p-value	2.25E-38	p-value	2.25E-38	p-value	2.25E-38	p-value	2.25E-38
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[SRY Site 1] = 'cons'	5' UTR	[SRY Site 1] = 'cons'	5' UTR	[D205] = 'cons'	ORF 2	[SRY Site 1] = 'cons'	5' UTR
[I1220] = 'cons'	ORF 2	[orientation] = '+'	ALL	[SRY Site 1] = 'cons'	5' UTR	[orientation] = '+'	ALL
[orientation] = '+'	ALL	[Strand] = '-1'	ALL	[orientation] = '+'	ALL	[PolyA Signal] = 'mut'	3' UTR
[Strand] = '-1'	ALL	[Poly-A pure] < 2.500	3' UTR	[Poly-A pure] < 2.500	3' UTR	[Strand] = '-1'	ALL
[Poly-A pure] < 2.500	3' UTR	[ORF1 %A] < 0.405	ORF 1	[ORF1 %A] < 0.405	ORF 1	[ORF1 %A] < 0.405	ORF 1
[ORF1 %A] < 0.405	ORF 1	[HMKK1091] = 'cons'	ORF 2	[HMKK1091] = 'cons'	ORF 2	[HMKK1091] = 'cons'	ORF 2
[HMKK1091] = 'cons'	ORF 2	[ORF2 %T] > 0.205	ORF 2	[ORF1&2 %T] > 0.200	ORF	[ORF2 %T] > 0.205	ORF 2
[ORF1 Frameshifts] > 1.500	ORF 1	[ORF1 Frameshifts] > 1.500	ORF 1	[ORF1 Frameshifts] > 1.500	ORF 1	[ORF1 Frameshifts] > 1.500	ORF 1
[ORF StartStop] = 'ORF2 cons'	ORF	[ORF StartStop] = 'ORF2 cons'	ORF	[ORF StartStop] = 'ORF2 cons'	ORF	[ORF StartStop] = 'ORF2 cons'	ORF

GSE5816 hBEC Lung (Continued)

Rule 17		Rule 18		Rule 19		Rule 20	
Support	0.11%	Support	0.11%	Support	0.11%	Support	0.11%
Confidence	75.00%	Confidence	75.00%	Confidence	75.00%	Confidence	75.00%
Odds ratio	371.57	Odds ratio	371.57	Odds ratio	371.57	Odds ratio	371.57
p-value	2.25E-38	p-value	2.25E-38	p-value	2.25E-38	p-value	2.25E-38
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[SRY Site 1] = 'cons'	5' UTR	[SRY Site 1] = 'cons'	5' UTR	[TF nkx-2.5B] = 'mut'	5' UTR	[D205] = 'cons'	ORF 2
[orientation] = '+'	ALL	[orientation] = '+'	ALL	[S1259] = 'cons'	ORF 2	[SRY Site 1] = 'cons'	5' UTR
[Strand] = '-1'	ALL	[PolyA Signal] = 'mut'	3' UTR	[SRY Site 1] = 'cons'	5' UTR	[I1220] = 'cons'	ORF 2
[Poly-A pure] < 2.500	3' UTR	[Strand] = '-1'	ALL	[orientation] = '+'	ALL	[orientation] = '+'	ALL
[ORF1 %A] < 0.405	ORF 1	[ORF1 %A] < 0.405	ORF 1	[ORF1 Gaps] < 1.500	ORF 1	[Poly-A pure] < 2.500	3' UTR
[HMKK1091] = 'cons'	ORF 2	[HMKK1091] = 'cons'	ORF 2	[Strand] = '-1'	ALL	[ORF1 %A] < 0.405	ORF 1
[ORF1&2 %T] > 0.200	ORF	[ORF1&2 %T] > 0.200	ORF	[ORF2 CAI] > 0.634	ORF 2	[HMKK1091] = 'cons'	ORF 2
[ORF1 Frameshifts] > 1.500	ORF 1	[ORF1 Frameshifts] > 1.500	ORF 1	[ORF2 Gaps] < 5.500	ORF 2	[ORF1 Frameshifts] > 1.500	ORF 1
[ORF StartStop] = 'ORF2 cons'	ORF	[ORF StartStop] = 'ORF2 cons'	ORF	[ORF1&2 %A] > 0.411	ORF	[ORF StartStop] = 'ORF2 cons'	ORF
				[ORF StartStop] = 'ORF2 cons'	ORF		

GSE5816 hBEC Lung (Continued)

Rule 21		Rule 22		Rule 23		Rule 24	
Support	0.11%	Support	0.11%	Support	0.077%	Support	0.08%
Confidence	75.00%	Confidence	75.00%	Confidence	66.67%	Confidence	66.67%
Odds ratio	371.57	Odds ratio	371.57	Odds ratio	236.45	Odds ratio	236.45
p-value	2.25E-38	p-value	2.25E-38	p-value	3.92E-19	p-value	3.92E-19
L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part	L1 characteristics	L1 part
[SRY Site 1] = 'cons'	5' UTR	[SRY Site 1] = 'cons'	5' UTR	[CPG Islands] < 0.500	ALL	[SRY Site 1] = 'cons'	5' UTR
[I1220] = 'cons'	ORF 2	[orientation] = '+'	ALL	[SRY Site 1] = 'cons'	5' UTR	[ORF1 Gaps] < 1.500	ORF 1
[orientation] = '+'	ALL	[Strand] = '-1'	ALL	[ORF1&2 %A] > 0.411	ORF	[ORF2 CAI] > 0.634	ORF 2
[PolyA Signal] = 'mut'	3' UTR	[Poly-A pure] < 2.500	3' UTR	[ORF1 %A] < 0.405	ORF 1	[ORF2 Gaps] < 5.500	ORF 2
[Strand] = '-1'	ALL	[ORF1 %A] < 0.405	ORF 1	[ORF1 Frameshifts] < 1.500	ORF 1	[ORF1 %A] < 0.405	ORF 1
[ORF1 %A] < 0.405	ORF 1	[HMKK1091] = 'cons'	ORF 2	[HMKK1091] = 'mut'	ORF 2	[ORF1 %T] > 0.183	ORF 1
[HMKK1091] = 'cons'	ORF 2	[ORF2 %T] > 0.205	ORF 2	[ORF2 CAI] < 0.634	ORF 2	[ORF2 Stops] > 5.500	ORF 2
[ORF1 Frameshifts] > 1.500	ORF 1	[ORF1&2 %T] > 0.200	ORF			[ORF StartStop] = 'ORF2 cons'	ORF
[ORF StartStop] = 'ORF2 cons'	ORF	[ORF StartStop] = 'ORF2 cons'	ORF				

Biography

Miss Naruemon Pratanwanich was born on February 3, 1987 in Trad, Thailand. She had graduated Bachelor of Engineering with first class honors in 2009 from Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand. Then, she had been granted by the ChulalongKorn University Graduate Scholarship to Commemorate the 72nd Anniversary of His Majesty King Bhumibol Adulyadej to study the Degree of Master of Science Program in Biomedical Engineering at Chulalongkorn University, Bangkok, Thailand in 2009. While studying, she published her research with her advisor and co-advisor under the topic "Mining LINE-1 Characteristics That Mediate Gene Expression" in the 1st International Conference of Computational Systems-Biology and Bioinformatics (CSBio 2010) held on 3-5 November 2010. This article was also published in the book series of Communications in Computer and Information Science (CCIS), Volume 115, page 83-95. In 2010, she won Thai Government Science and Technology Scholarship for pursuing PhD abroad.