DERIVATION OF THE COMPLETE TRANSCRIPTOME OF *ESCHERICHIA COLI* FROM MICROARRAY DATA

by

KAN BAO

(Under the Direction of Ying Xu)

ABSTRACT

The availability of complete genomic sequences and microarray expression data calls for computational methods for characterizing transcriptome, the complete collection of alternative transcription units (ATU) and complete transcription units (CTU).

Though numerous computational methods have been developed for prediction of operons (CTU), none of existing computational methods can deal with ATU. We present a new computational method for ATU prediction. The ATU was predicted based on variance of fold changes of expression level and intergenic distance. Then, CTU and ATU were combined to form the transcriptome of Escherichia coli.

The alternative TU predictor achieves 93% prediction accuracy in estimating presence of ATU. The percentage of known ATUs correctly predicted and known single-gene CTU correctly predicted are 84.3% and 80.43% respectively. About 91.94% of transcriptome (include CTU and ATU) from multiple-genes operons are correctly predicted.

INDEX WORDS: Transcriptome, Alternative transcription units, *Escherichia coli*. Complete transcription units, Microarray data, operon, indergenic distance, variance of fold changes

DERIVATION OF THE COMPLETE TRANSCRIPTOME OF *ESCHERICHIA COLI* FROM MICROARRAY DATA

by

KAN BAO

B.S., Shanghai Normal University, P.R. China, 2004

M.S., Shanghai Normal University, P.R. China, 2007

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2009

DERIVATION OF THE COMPLETE TRANSCRIPTOME OF *ESCHERICHIA COLI* FROM MICROARRAY DATA

by

KAN BAO

Major Professor: Ying XU

Committee: Lily Wang
Jien Chen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2009

DEDICATION

To my parents and my friends.

ACKNOWLEDGEMENTS

I would first like to convey my full appreciation to my major professor, Dr. Ying Xu for his guidance throughout this project. His enthusiasm and continuous quest for new ideas inspire me. He also creates an environment where researchers with different backgrounds cooperate together and I could feel the team spirit.

I would also like to thank my instructor, Dr. Victor Olman for his continuous direction throughout this project. Without his direction, completion of this project would be impossible.

I also want to express my gratitude to Dr. Lily Wang and Dr. Jien Chen for serving on my thesis committee. I deeply appreciate their comments and directions for my project.

Next, I would like to take this unique opportunity to thank Dr. Phuongan Dam and Dr. Fenglou Mao for their DOOR databse. Besides, I would like to thank all members of the Computational System Biology Lab at the University of Georgia.

Lastly, I would also like to thank all professors and staffs in the Department of Statistics at the University of Georgia for their help throughout my years in the program.

To each of them above, I extend my deepest appreciation.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

2 LITERATURE REVIEW 3

2.1 OPERON 3

2.2 TRANSCRIPTION UNIT 11

2.3 MICROARRAY DATA ANALYSIS 14

3 ALTERNATIVE TRANSCRIPTION UNITS PREDICTION 19

3.1 DATA DESCRIPTION 19

3.2 PREDICTION MODEL BUILDING PROCESS 24

3.3 RESULTS 36

4 DERIVATION OF THE COMPLETE TRANSCRIPTOME 39

4.1 COMPLETE TRANSCRIPTION UNIT 39

4.2 ALTERNATIVE TRANSCRIPTION UNIT 40

4.3 TRANSCRIPTOME 41

4.4 RESULT 42

5 CONCLUSION 45

BIBLIOGRAPHY 47

APPENDICES 52

A The performance of presence of alternative TU in operons with different sizes 52

B The performance of prediction of alternative TU in operons with different sizes 56

C The evaluation of tus in operons with different sizes 59

LIST OF TABLES

Page

Table 1.1: Reported performances of TU prediction methods 13

Table 2.1: Performances of presence of alternative TU in operon 37

Table 2.2: Performances of prediction of alternative TU in operon 38

Table 3.1: Evaluation of single-gene CTU 43

Table 3.2: Evaluation of transcriptome from multiple-genes operons 43

LIST OF FIGURES

Page

Figure 1.1: Operon structures 5

Figure 1.2: Transcription Unit structures.....11

Figure 2.1: The frequency distribution of operon sizes in *Escherichia coli* strain K-12.....20

Figure 2.2: The frequency distribution of gene expression level of thrB 22

Figure 2.3: The frequency distribution of average gene expression level for 2543 genes 22

Figure 2.4: The frequency distribution of standard deviation of gene expression level.....23

Figure 2.5: The expression level of three genes in operon 3863 under 380 conditions 25

Figure 2.6: The first 380 fold changes of three genes in operon 3863 26

Figure 2.7: The frequency distribution of variance of fold changes in operon 3863.....28

Figure 2.8: The frequency distribution of 90% quantile of variance of fold changes for all three-gene operons.....29

Figure 2.9: The frequency distribution of 99% quantile of variance of fold changes for all three-gene operons.....31

Figure 2.10: The frequency distribution of conditions in operon 4124 32

Figure 2.11: The frequency distribution of intergenic distances within operon 35

Figure 3.1: The frequency distribution of complete TU size.....40

Figure 3.2: The frequency distribution of alternative TU size.....41

Figure 3.3: The frequency distribution of TUs size.....42

CHAPTER 1

INTRODUCTION

The availability of complete genomic sequences and microarray expression data calls for computational methods for reveal the regulatory of cell in prokaryotes. The prediction of operons, a set of genes that co-transcribed into a single mRNA, is the first step in reconstruction of a regulatory network at the genome-wide level [5].

Numerous operons prediction methods have been proposed since 1990s. These methods include log-likelihood model [25], Bayesian network [1], logistic regression [21], neural network [4]], genetic algorithm [17], Hidden Markov model [35], decision tree [3] and graph-theoretic model [11]. Based on these prediction methods, a number of operon database have been developed, including Database of Prokaryotic Operons (DOOR) [13], MicrobesOnline [22], OperonDB [19], Operon Database (ODB) [26], RegulonDB [15] and DBTBS [27].

The prediction of transcriptions units (TUs), the smallest unit of transcription in prokaryotes, is the second step in reconstruction of a regulatory network at the genome-wide level. The prediction of transcriptions units has been implemented using some approaches, including hidden Markov models [30], multiple methods [6], probability of functional clusters of genes [12] and log- likelihood model [20]. However, these transcription unit prediction methods can not deal with alternative transcription unit. All the transcription units predicted from these TU prediction methods only include complete transcription unit.

Transcriptome is defined as the all transcribed regions encoded in the genome. Experimentally defining the complete transcriptome of prokaryotic organisms has been a challenging task, involving large, costly and labor-intensive experiments for sequencing of expressed sequence tag and full-length cDNA libraries. Hence, despite the fact that numerous species have been sequenced, only few transcriptomes have been extensively identified [36].

Unlike the genome, which is essentially a static entity, the transcriptome can be modulated by regulatory factors under different experimental conditions. A multiple-promoters operon transcribes as a complete TU in most conditions, whereas it also may transcribes an alternative TU in some conditions. TUs transcribe together to give rise to an mRNA, which will be transported to cellular ribosomes to guide translation and protein synthesis. The transcriptome thereby serves as a link between an organism's genome and its proteome [32].

In prokaryotes such as *Escherichia coli*, operons are described adjacent genes that transcribed into a single mRNA [2] .For an operon including multiple promoters, a fraction of its genes can be present in several different alternative TUs in different conditions [7]. None of the existing operon predictors are able to deal with alternative TUs.

We have presented a new computational method for TU prediction, which is able to predict alternative transcription units (ATU). Since the first model organism for molecular biology is *Escherichia coli*, we implemented the new TU predictor to produce the alternative transcription units of *Escherichia coli.* Then we combined alternative transcription units with complete transcription units to form the transcriptome of *Escherichia coli.*

CHAPTER 2

LITERATURE REVIEW

Transcriptome is defined as the all transcribed regions encoded in the genome. Unlike the genome, which is essentially a static entity, the transcriptome can be modulated by regulatory factors under different experimental conditions. Hence, transcriptome contain both complete transcriptions units (CTU) and alternative transcriptions units (ATU).

Operons are described adjacent genes than transcribed into a single mRNA. A multiple-promoters operon transcribes as a CTU in most conditions, whereas it also may transcribes as an ATU in some conditions. Transcriptome transcribe together to give rise to an mRNA, which will be transported to cellular ribosomes to guide translation and protein synthesis. The transcriptome thereby serves as a link between an organism's genome and its proteome.

2.1 OPERON

An operon represents a functioning unit of adjacent genes in the complex hierarchical structure of biological processes in a cell of prokaryotes.

2.1.1 TRANSCRIPTION

DNA molecules are responsible for encoding the information necessary to build each protein or RNA molecular found in an organism [33]. The information flow DNA via RNA and thus to the

protein is described as central dogma of molecular biology, which includes the following three major stages. The information contained in DNA is duplicated by replication process. DNA directs the production of encoded messenger RNA through transcription. In the last stage of the information-transfer process, messenger RNA carries the encoded information to protein-synthesizing structures called ribosomes. Through a process named as translation, the ribosomes use this coded information to direct protein synthesis [37].

Transcription is the process of synthesizing RNA copy using DNA as templates. To initiate a transcription process, then DNA double helix is unzipped, starting at the promoter site of a gene. After unzipping DNA double helix, one DNA strand serves as a template strand. RNA molecules are constituted by binding together ribonucleotides complementary to the template strand. Messenger RNA (mRNA) is synthesized from 5' end to the 3' end, whereas the template strand is read from 3' end to 5' end. After the transcription process, the mRNA will be transported to cellular ribosomes to guide translation and protein synthesis [37].

2.1.2 Definition Of Operon

The operon concept was first proposed in the *Proceedings of the French Academy of Science* by the French microbiologists Francois Jacob and Jacques Monod in 1961. They described the regulatory mechanism of the *lac* operon of *Escherichia coli*, a system that allows the bacterium to repress the production of enzymes involved in lactose metabolism [18]. They were awarded Nobel Prize in Medicine in 1965 because of their distinguished research that gave impetus to the development of molecular biology.

Francois Jacob and Jacques Monod defined an operon as a cluster of two or more contiguous genes transcribed from one common promoter that gives rise to a message RNA,

which is known as classical definition of operon [18]. According to RegulonDB database, they extend the definition to include the possibility of operons with only one gene for database purposes [25].

2.1.3 Operon Structures

An operon usually includes an active promoter, several structural genes and a terminator (Figure 1.1). The promoter is a segment of DNA usually occurring upstream from a gene coding region and acting as a controlling element in the expression of that gene. The structural gene is a gene that codes for any RNA or protein product rather than a regulatory factor. The terminator is a DNA sequence that results in termination of transcription.

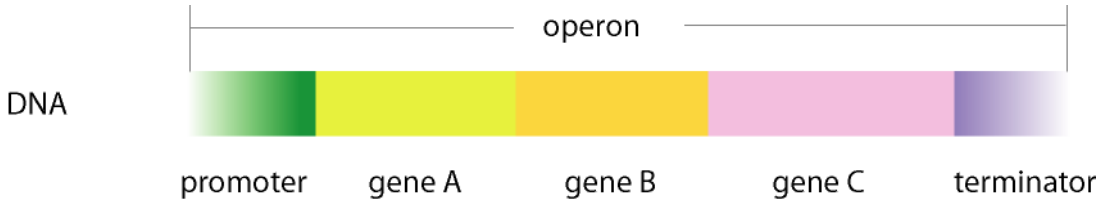

Figure 1.1: Operon structures

2.1.4 Operon Prediction

With more and more prokaryotic genome sequences and microarray data available, the determination of operon structures at a genome-wide level has become the main focus.

2.1.4.1 Experimental Determination Of Operon

The presence of several genes in the same operon can be experimentally detected using several experiment techniques including northern blot, reverse transcription polymerase chain reaction, polar mutation and DNA microarray [34].

Northern blotting involves the use of electrophoresis to separate RNA samples by size and detection with a hybridization probe complementary to part of or the entire target sequence.

Reverse transcription polymerase chain reaction (RT-PCR) is a method of polymerase-chain-reaction amplification of nucleic acid sequences that uses RNA as the template for transcribing the corresponding DNA using reverse transcriptase.

DNA microarray is a technique to monitor gene expression in thousands of genes. Thousands of probe DNAs are spotted or synthesized on microscope slides. Sample RNAs are labeled with fluorescent dyes. Gene expression levels in the sample are detected by hybridization of the labeled RNAs to the probes on the slide.

Polar mutation is a mutation that affects the transcription of part of the gene or operon downstream of the mutant site. These mutations tend to occur early within the sequence of genes and can be nonsense, insertion mutations, which affects the rate of expression of downstream genes.

2.1.4.2 Prediction Features

Though experimental techniques mentioned above can determine whether genes belong to the same operon, they are too labor-intersive for genome-wide level application. To solve this problem, numerous computational methods have been developed for prediction of operons.

Most operon prediction methods divide adjacent gene pairs into two groups: operonic gene pairs and boundary gene pairs. Various features have been examined to distinguish between such gene pairs. These features include intergenic distance, genes functional classes, correlations between adjacent genes, codon usage, length ratio between a pair of gene, transcription signal, biological pathway, conservation of gene pairs and so on [34].

One of the most effective features for operon prediction is intergenic distance proposed by Salgado. They found that adjacent gene pairs within an operon (also known as operonic gene pairs) tend to have shorter intergenic distance, while gene pairs from two consecutive operons (also known as non-operonic gene pairs) tend to have longer distances [25]. The formula for intergenic distance is

$$D_i = G_d start - (G_u end + 1), \tag{1.1}$$

where $G_d start$ is the start position of downstream gene and $(G_u end)$ is the end position of upstream gene.

Another effective feature is codon usage [1]. Bockhorst associated with each gene g_k a set of codon bias vectors $\{\vec{b}_a^k\}$, one for each amino acid. The elements of the bias vector are

$$b_{a,uvw}^k = \hat{f}_{(uvw\,|a)} - \bar{f}_{(uvw\,|a)}, \tag{1.2}$$

where uvw is a codon that codes for a, $\bar{f}_{(\mathrm{uvw\,|a})}$ is the frequency with which a is encoded by uvw over then whole genome.

The smoothed frequency with which *a* is coded for by uvw in then gene is

$$\hat{f}_{(\text{uvw}\,|\text{a})}=\frac{\text{n}_{\text{uvw}}+\bar{f}_{(\text{uvw}\,|\text{a})}}{\sum_{\text{xyz}\,\in\text{condons}\;\;(\text{a})}\text{n}_{\text{xyz}}+1}\;, \tag{1.3}$$

where n_{uvw} is the number of times codon uvw appears in g_k.

The codon usage similarity between two genes is defined as

$$Sim(g_k, g_l) = \sum_a \vec{b}_a^k \cdot \vec{b}_a^l \tag{1.4}$$

The widely used feature in gene expression data analysis is Pearson's correlation coefficient, which measures the extent to which two gene expression patterns are similar with each other [37]. Given two data sets O_i and O_j from two genes respectively, Pearson's correlation coefficient is defined as

$$Pearson(O_i, O_j) = \frac{\sum_{d=1}^{p}(o_{id}-\mu_i)(o_{jd}-\mu_j)}{\sqrt{\sum_{d=1}^{p}(o_{id}-\mu_i)^2}\sqrt{\sum_{d=1}^{p}(o_{jd}-\mu_j)^2}}\;, \tag{1.5}$$

where μ_i and μ_j are the means for $\vec{O}_i$ and $\vec{O}_j$, respectively. The value of Pearson's correlation coefficient ranges between -1 and 1 with a higher value indicating stronger similarity.

2.1.4.3 PREDICTION METHODS

A wide range of computational methods have been used in operon prediction using various prediction features. These methods generally include log-likelihood model [25], Bayesian

network [1], logistic regression [21], neural network [4]], genetic algorithm [17], Hidden Markov model [35], decision tree [3] and graph-theoretic model [11].

Log-likelihood model is a model calculating distance log-likelihoods ratio for adjacent pairs of genes to be in the same operon. Salgado generated the following formula for the log-likelihood model

$$LL(dist) = log \frac{N_{op}\,(dist)/TN_{op}}{N_{nop}\,(dist)/TN_{nop}} \ , \tag{1.6}$$

where N_{op} and N_{nop} are pairs of genes in operons and at transcriptional boundaries respectively, at a distance [$dist$], whereas TN_{op} and TN_{nop} are the total number of pairs of genes in operons and at the transcriptional boundaries respectively [25].

A Bayesian network is way of representing the joint probability distribution of a set of random variables that exploits the conditional independence relationships among the variables. Bockhorst formulated the chain rule

$$Pr(X_1, \cdots\cdots, X_n) = \prod_i^n Pr(X_i|X_1, \cdots\cdots, X_{i-1}), \tag{1.7}$$

where X_i is the random variable [1].

Logistic regression is used for prediction of the probability that gene pair i is an operon. Roback proposed the model

$$\ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 \times cor_{i,1} + \cdots\cdots + \beta_p \times cor_{i,p} + \beta_{p+1} \times dist_i \ , \tag{1.8}$$

where $dist$ is intergenic distance and cor_l is the correlation of expression among experiments in subset l (l=1,2, ⋯⋯, p) [21].

2.1.5 OPERON DATABASES

Based on numerous computational methods in operon prediction and experiment techniques, a lot of operon databases have been developed. Database of Prokaryotic Operons (DOOR) contains operons for 675 prokaryotic genomes [13], MicrobesOnline provides operons for 620 genomes [22], OperonDB contains operons for 550 genomes [19], Operon Database (ODB) provides operons for 203 genomes [26], RegulonDB provides operons in *E. coli* only and DBTBS contains operons in *B. subtilis* only [15]. All operons in DOOR, OperonDB and MicrobesOnline are predicted by computational methods, while ODB, RegulonDB and DBTBS operons are based on experiments, literature and computational methods.

Among these databases, DOOR developed by Computational Systems Biology Lab, ODB developed by Human Genome Center at University of Tokyo and RegulonDB developed by Program of Computational Genomics at Universidad Nacional Autónoma de México are widely used in genome studies.

DOOR contains predicted operons of all sequenced prokaryotic genomes. All the operons in DOOR are predicted using computational methods. The operon database covers 675 complete archeal and bacterial genomes that include both chromosomal and plasmid gene pairs. This database also enables users to search desired operons and predict cis-regulatory [13].

ODB contains known operons of 203 genomes from prokaryotes and eukaryotes curated from the literature. Putative operons are also determined by orthologous gene prediction. This

database also supports operons prediction in 194 organisms using several features [26].

RegulonDB is a comprehensive database consisting of data from transcription regulation for *E.coli* K12, including operons, terminators, promoters, transcription units and regulatory pathways. This database is well curated from experimental data and literature [15].

2.2 Transcription Unit

The stretch of DNA transcribed into an RNA molecule is called transcription units, which can be grouped into two categories: complete transcription units and alternative transcription units (Figure 1.2). The complete set of transcription units is defined as transcriptome.

2.2.1 Definition Of Transcription Unit

A transcription unit is a set of one or more genes transcribed from a common promoter to produce a single messenger RNA. A transcription unit should include one or more genes, one active promoter, and one terminator. Transcription factor binding sites need not be components of a transcription unit. There is a one to one correspondence between transcription units and promoters (Figure 1.2).

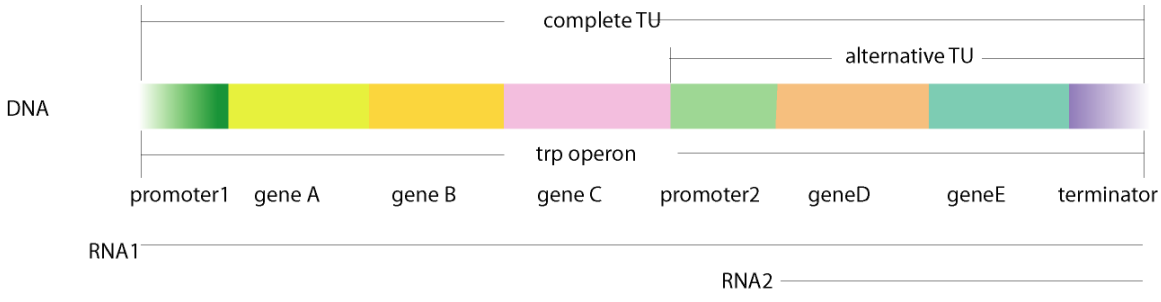

Figure 1.2: Transcription Unit structures

Although the definitions of transcription unit and operon generally contain promoters, terminators, we only deal with the structural genes of transcription unit and operon in this thesis.

2.2.2 Complete Transcription Unit

According to the definition of operon and transcription unit, an operon contains at least one transcription unit called complete transcription unit that includes all the genes in that operon. There is no difference between complete transcription unit and operon.

2.2.3 Alternative Transcription Unit

For operons that include multiple promoters, a transcription unit is defined for each promoter. A complex operon with several promoters contains several transcription units.

An alternative transcription unit is a fraction of operon genes transcribed from a single promoter to produce a single messenger RNA. An alternative transcription unit should include one or more genes, one promoter, and one terminator (Figure 1.2).

The differences between an alternative transcription unit and an operon are (1) an alternative transcription unit must include one promoter. (2) an operon may include more than one promoter.

The *E. coli* operon for galactose utilization (gal) contains a glucose-dependent and a glucose-independent promoter. The *E. coli* tryptophan (trp) (Figure 1.2) and isoleucine-valine (ilv) operons have internal promoters leading to the expression of a fraction of genes in the operons [34].

2.2.4 TRANSCRIPTION UNIT PREDICTION

With more and more microarray expression data and complete genomic sequences available, the determination of properties at a genome-wide level has become an important issue. Numerous genome-wide operon prediction methods have developed since a number of prokaryotic genomes had been sequenced. However, only a few methods for transcription unit prediction methods have been proposed [23]. Among these transcription unit prediction methods, all the transcription units only include complete transcription unit. When an operon contains multiple promoters, several alternative transcription units can be transcribed under different conditions. None of existing transcription unit prediction methods are able to deal with this issue, and only limited attempts have been made for this problem [8].

Among these transcription unit prediction methods, Tjaden proposed hidden Markov models to predict transcription unit [30], Craven used both sequence information and gene expression data to predict transcription unit [6], Ermolaeva found functional clusters of genes based on conservation of gene positions across different genomes and Salgado adopted intergenic distances as the main feature in transcription unit prediction [12].

Table 1.1: Reported performances of TU prediction methods

Prediction Methods	Features	TUs sensitivity
HMM	Sequence	59%
Multiple	Sequence and gene expression data	68%
Probability	Conserved gene clusters	50%
Loglikelihood	Intergenic distance, functional class	75%

Table 1.1 reveals the performances of these transcription unit (complete TUs) prediction methods [23]. All the prediction methods were tested on known transcription unit data of E. Coli, from the RegulonDB database. Most prediction methods predict transcription units with 50-75% sensitivity (percentage of known transcription units correctly predicted) [23]. The loglikelihood method achieves 75% sensitivity; HMM method achieves only 59% sensitivity; Multiple method achieves 68% of known transcription units correctly predicted; Probability achieves only 50% sensitivity.

2.3 Microarray Data Analysis

The development of DNA microarray technology enables scientists to capture the gene expression on a genome-wide level. The expression levels of thousands of genes can be monitored using a single microarray chip. DNA microarray has generated a large number of gene expression data over the past several years. Numerous methods have been proposed to analyze gene expression microarray data for different purposes.

2.3.1 K-Fold Change

Fold change method is used to find genes that are differentially expressed. The ratio for a gene is calculated as the average expression over all samples in a condition divided by the average expression over all samples in another condition.

$$ratio = \frac{\mu_1}{\mu_2}, \tag{1.9}$$

where μ_1 represents average of expression value over all samples in the first condition and μ_2 represents average of expression value over all samples in the second condition.

Since most genes express in the same biological pathway in different conditions, the ratios between two conditions should be around one. Genes that demonstrate a significant change between two conditions are considered as differentially expressed.

To facilitate the selection process, the ratio between the two expression levels for several genes is first calculated. In general, an arbitrary threshold such as two-fold ($log_2 x$) or three-fold ($log_3 x$) change is selected and the ratio is considered to be significant if it is large than the threshold [9].

2.3.2 *t*-TEST

The classical method for performing a hypothesis test on two groups observations data is the *t*-test, which was originally named as "student's *t*-test" developed by William Sealy Gosset.

2.3.2.1 PAIRED *t*-TEST

The paired *t*-test is used in paired data, which has a pair of observations for each gene. The null hypothesis is that the gene is not differentially expressed, denoted by H_0: $\mu = 0$. The alternative hypothesis is that the gene is differentially expressed, denoted by H_α: $\mu \neq 0$.

$$x = log_2 \frac{x_1}{x_2} \ , \tag{1.10}$$

$$t = \frac{\bar{x}}{s/\sqrt{n}} \ , \tag{1.11}$$

where x_1 and x_2 are a pair of expression value for each gene, $\bar{x}$ is the mean of the log ratios, s is the standard deviation of log ratios and n is the number of samples. The null hypothesis is rejected or not rejected depends on p-value and a significance level.

The significance of differentially-expressed genes depends not only on the average log ratio but also on both the population variability and the number of samples. The accuracy of the determination of differentially-expressed genes increases with the number of samples [28].

2.3.2.2 UNPAIRED *t*-TEST

The unpaired *t*-test is applicable to unpaired data which contains two unrelated groups of observations. The null hypothesis states that the means of the expression levels of a given gene in the two groups will be equal, denoted by H_0: $\mu_1 = \mu_2$. The alternative hypothesis is that the means of the expression levels of a given gene in the two groups will be unequal, denoted by H_α: $\mu_1 \neq \mu_2$.

Both the equal-variance and unequal-variance unpaired *t*-test use the formula as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{1.12}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of expression levels of a given gene, s_1^2 and s_2^2 are the variances, and n_1 and n_2 are the sizes of the two groups. The null hypothesis is rejected or not rejected depends on p-value and a significance level.

2.3.3 NEIGHBORHOOD ANALYSIS

Neighborhood analysis is also available for the identification of differentially-expressed genes. Given the expression levels of gene g over all the conditions, the following score is defined as:

$$P(g) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)} \ , \tag{1.13}$$

where $\mu_1(g)$ and $\mu_2(g)$ are the mean of the expression levels of gene g in class 1 and class 2 respectively, and $\sigma_1(g)$ and $\sigma_2(g)$ are the standard deviations of g in class 1 and class 2 respectively.

Large absolute values of $P(g)$ indicate a strong correlation between gene expression and class distinction, while a positive value indicates that g is more highly expressed in class 1 and a negative value indicates that g is more highly expressed in class 2 [29].

2.3.4 EUCLIDEAN DISTANCE

Euclidean distance is one of the most widely-used methods to measure the distance between two data. The distance between data D_i and D_j in p-dimensional space is calculated:

$$Euclidean\,(D_i, D_j) = \sqrt{\sum_{d=1}^{p} (D_{id} - D_{jd})^2} \tag{1.14}$$

However, the overall shapes of gene expression profiles are often of greater interest than the individual magnitudes of each feature. To solve this problem, a standardization process is usually performed before calculating the *Euclidean* distance [37].

$$D_{kj}' = \frac{D_{kj} - \mu_k}{\sigma_k} \quad (1 \leq j \leq p) \tag{1.15}$$

$$\mu_k = \frac{\sum_{d=1}^{p} D_{kd}}{p} \tag{1.16}$$

$$\sigma_k = \sqrt{\frac{1}{p}\sum_{d=1}^{p}(D_{kd} - \mu_k)^2} \tag{1.17}$$

2.3.5 *K*-MEANS

The *K*-means algorithm is an extensively used partition-based clustering method. Given a per-specified parameter *K*, the algorithm partitions the data set into *K* disjoint subsets which optimize the following function:

$$V = \sum_{i=1}^{K} \sum_{O \in C_i} |O - \mu_i|^2 \ , \tag{1.18}$$

where O is a data in the cluster C_i and μ_i is the average of C_i.

Therefore, the purpose of function V is to minimize the sum of the squared distances of objects from their cluster centers [24].

CHAPTER 3

ALTERNATIVE TRANSCRIPTION UNITS PREDICTION

According to definition, alternative transcription units only present in multiple-promoters operons rather than single-promoter operons. However, we lack the *Escherichia coli* promoter information at a genome-wide level. Therefore, the first step to predict alternative transcription units is to test whether an operon has an alternative transcription unit based on Database of Prokaryotic Operons (DOOR) and Many Microbe Microarrays Database (M3D). Based on variance of fold changes of within operon genes and intergenic distance, we predicted alternative transcription units of *Escherichia coli*.

3.1 DATA DESCRIPTION

Operons of *Escherichia coli* from DOOR and microarray gene expression data of *Escherichia coli* from M3D were downloaded for the prediction of alternative transcription units. Known transcription units data of *Escherichia coli* were also downloaded from RegulonDB to evaluate the prediction.

3.1.1 OPERON DATA

DOOR (Database of prOkaryotic OpeRons) is an operon database developed by Computational Systems Biology Lab (CSBL). The operons in the database are based on operon-prediction

program. The prediction algorithm is a data-mining classifier, which include Intergenic distance, Neighborhood conservation, Phylogenetic distance, information from short DNA motifs, Similarity score between GO terms of gene pairs and Length ratio between a pair of genes [8].

The complete operons of *Escherichia coli* strain K-12 substrain MG1655 were downloaded from DOOR (http://csbl1.bmb.uga.edu/OperonDB/displayNC.php?id=215). A number of operons which include at least one unexpressed gene based on M3D were eliminated.

Among 827 operons in *Escherichia coli* strain K-12, 436 operons contain two genes, 170 operons include three genes, 98 operons contain four genes, 55 operons include five genes, 37 operons contain six genes, 14 operons include seven genes, 6 operons contain eight genes, 3 operons include nine genes, 3 operons contain eleven genes, 3 operons include twelve genes and 2 operons contain fifteen genes (Figure 2.1).

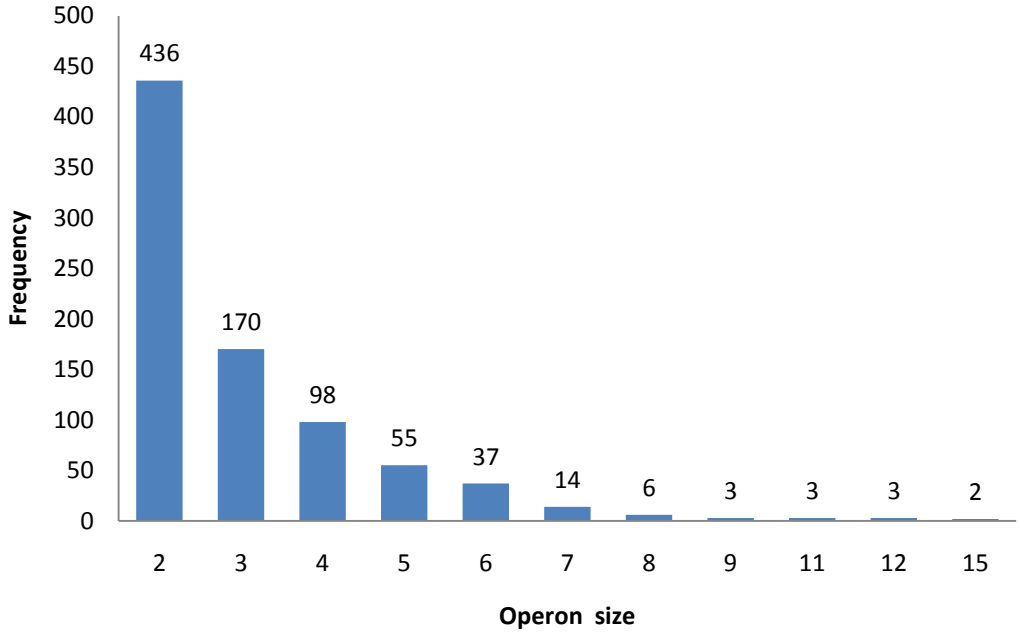

Figure 2.1: The frequency distribution of operon sizes in *Escherichia coli* strain K-12

3.1.2 MICROARRAY DATA

Many Microbe Microarrays Database (M3D) is designed to facilitate the analysis and visualization of expression data in compendia compiled from multiple laboratories. M3D contains over a thousand Affymetrix microarrays for *Escherichia coli*, *Saccharomyces cerevisiae* and *Shewanella oneidensis.* The expression data is uniformly RMA normalized to make the data generated by different laboratories and researchers more comparable. The experimental condition metadata in M3D is curated with each chemical and growth attribute stored as a structured and computable set of experimental features. All versions of the RMA normalized compendia constructed for each species are maintained and accessible in perpetuity to facilitate the future interpretation and comparison of results published on M3D data [14].

We also downloaded RMA normalized microarray gene expression data of *Escherichia coli* from M3D (http://m3d.bu.edu/norm). A number of genes which are not contained in *Escherichia coli* K12 operons based on DOOR were eliminated. Those eliminated genes are considered as single-gene CTU, which are unable to induce any alternative transcription units.

The microarray data consists of expression level of 2543 genes. Each gene has 380 gene expression levels in 380 different conditions respectively. Figure 2.2 shows the frequency distribution of gene expression level of thrB under 380 conditions. The rang of gene expression level for thrB is from 7 to 13. Under 234 conditions, the gene expression level for thrB is around 9. The mean and standard deviation of gene expression level for thrB under 380 conditions are 9.23 and 0.95 respectively.

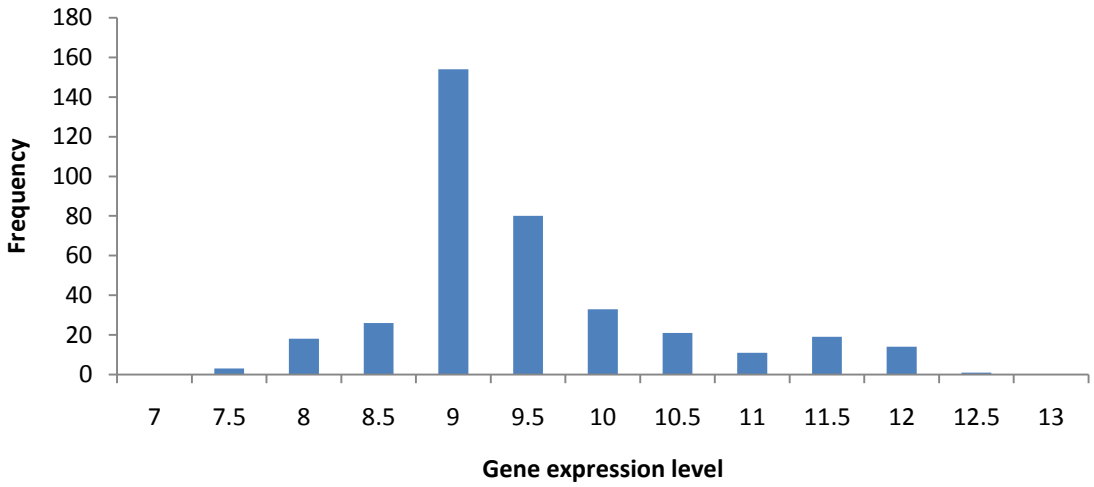

Figure 2.2: The frequency distribution of gene expression level of thrB

Figure 2.3 shows the frequency distribution of average gene expression level of 380 conditions for 2543 genes. The range of average gene expression level is from 3 to 15. About 53.99% average gene expression level fall into the interval between 8 and 10. The histogram follows a Normal curve, with the peak at 8.5.

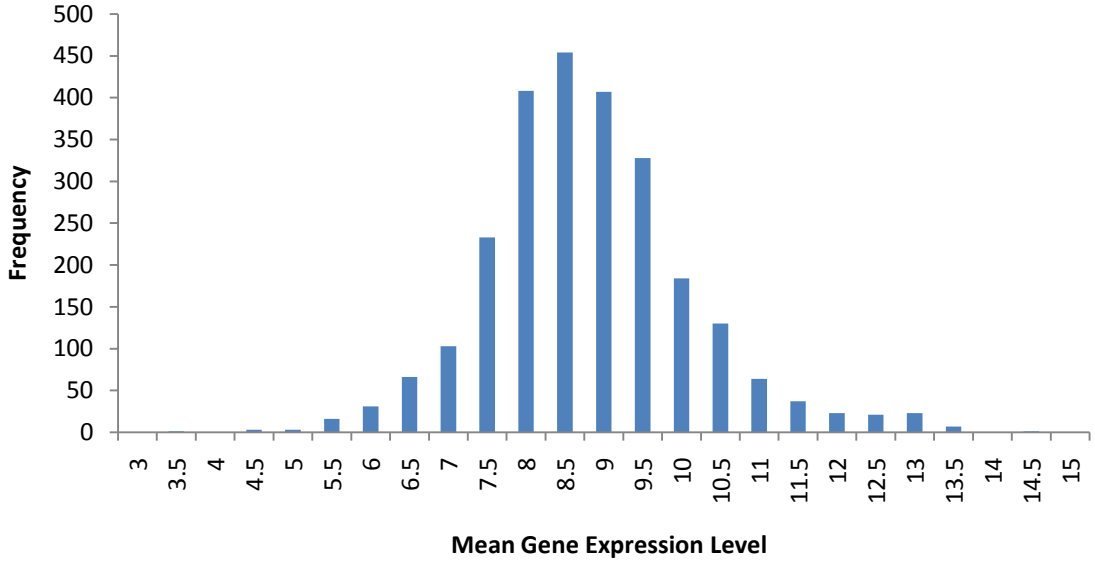

Figure 2.3: The frequency distribution of average gene expression level for 2543 genes

The frequency distribution of standard deviation of gene expression level under 380 conditions for 2543 genes is shown in Figure 2.4. The range of standard deviation of gene expression level is from 0.1 to 2.3. About 50.33% of standard deviation of gene expression level falls into the interval between 3 and 6. The histogram follows a right skewed curve, with the peak at 0.4. With the increase of standard deviation, the frequency decrease tardily.

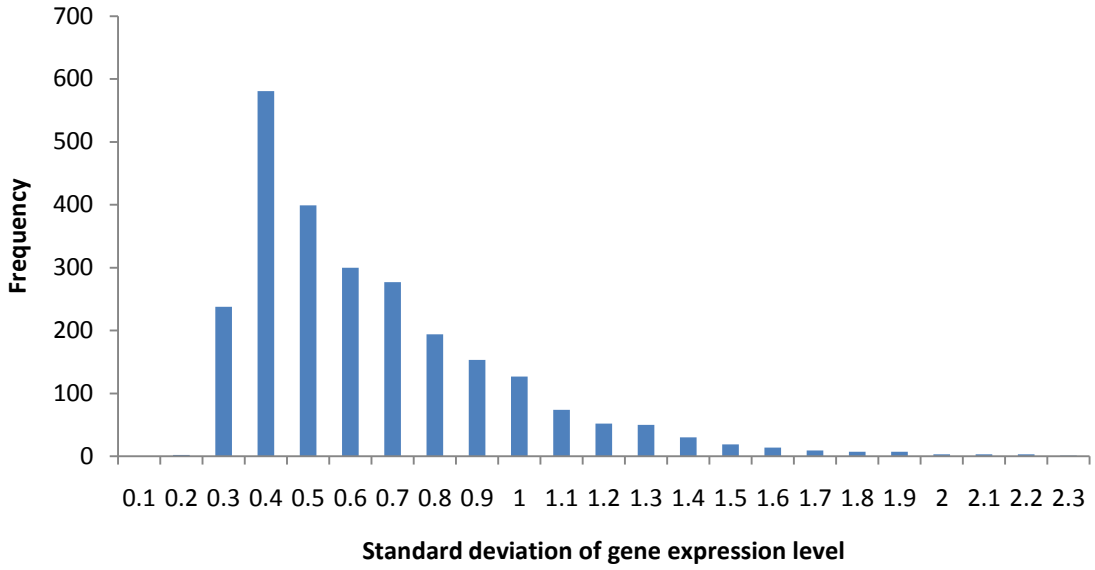

Figure 2.4: The frequency distribution of standard deviation of gene expression level for 2543 genes

3.1.3 TRANSCRIPTION UNIT DATA

RegulonDB is a model of the complex regulation of transcription initiation or regulatory network of the cell. In other words, it is also a model of the organization of the genes in transcription units, operons and regulons. In this regard, RegulonDB is a computational model of mechanisms

of transcriptional regulation.

Experimentally confirmed and computational predicted transcription units data and operon data were downloaded from the RegulonDB (http://regulondb.ccg.unam.mx/) to evaluate the prediction of alternative TU.

3.2 Prediction Model Building Process

None of existing transcription unit prediction methods deals with alternative transcription unit [8]. All the transcription units predicted from these TU prediction methods only include complete transcription unit. We present a new computational method for alternative transcription unit prediction.

3.2.1 Fold Change

The intricacy of the microarray experimentation process generally introduces bias into gene expression level measurements. Bias can be caused by the concentration and amount of DNA placed on the microarrays, lack of spatial homogeneity of the slides, the quantities of mRNA extracted from samples, scanner settings, saturation effects, background fluorescence and linearity of detection response [10].

All the genes in the same operon transcribe from one common promoter to gives rise to a message RNA. The expression levels of genes in the same operon should be close to each other. However, the expression levels of genes in the same operon are totally different because of bias mentioned above. Figure 2.5 reveals the expression level of three genes in operon 3863 under 380 conditions. Operon 3863 is a three-genes operon, including b0190, b0191 and b0192. The expression level of b0190 and b0191 are close to each other under most conditions, but the

expression level of b0192 is significantly different from other two genes. This is also revealed by the average of expression level of b0190, b0191 and b0192, which is 9.37, 9.53 and 8.23 respectively.

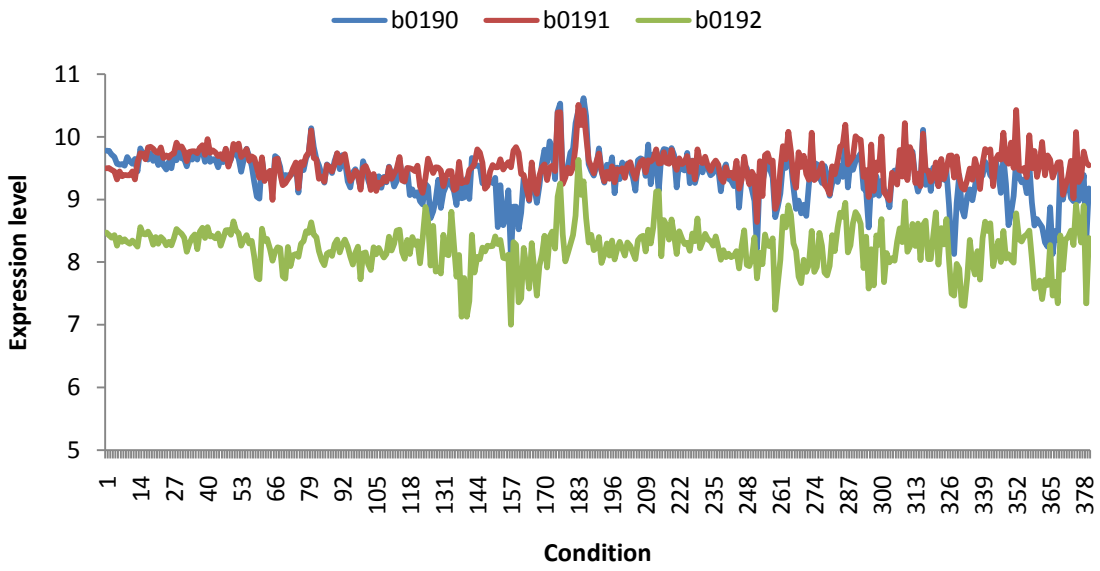

Figure 2.5: The expression level of three genes in operon 3863 under 380 conditions

We use the fold change, the ratio of the measured value for an experimental sample to the value for another sample, to solve the problem that expression levels of genes in the same operon are not close to each other. The fold change for a gene is calculated as the expression level in a condition divided by the expression level in another condition. The microarray data consists of expression level of 2543 genes. Each gene has 380 gene expression levels in 380 different conditions respectively. For any two conditions, each gene has a fold change. Therefore, each gene has 72010 fold change ratios. The formula of fold change is defined as follows:

$$f_{nij} = \frac{e_{ni}}{e_{nj}} \tag{2.1}$$

where f_{nij} represents the fold change of gene n of condition i and condition j, e_{ni} represents the expression level of gene n in condition i, e_{nj} represents the expression level of gene n in condition j, for $n = 1, \cdots\cdots 2543$, $i = 1, \cdots\cdots 380$, $j = 1, \cdots\cdots 380$ and $j > i$.

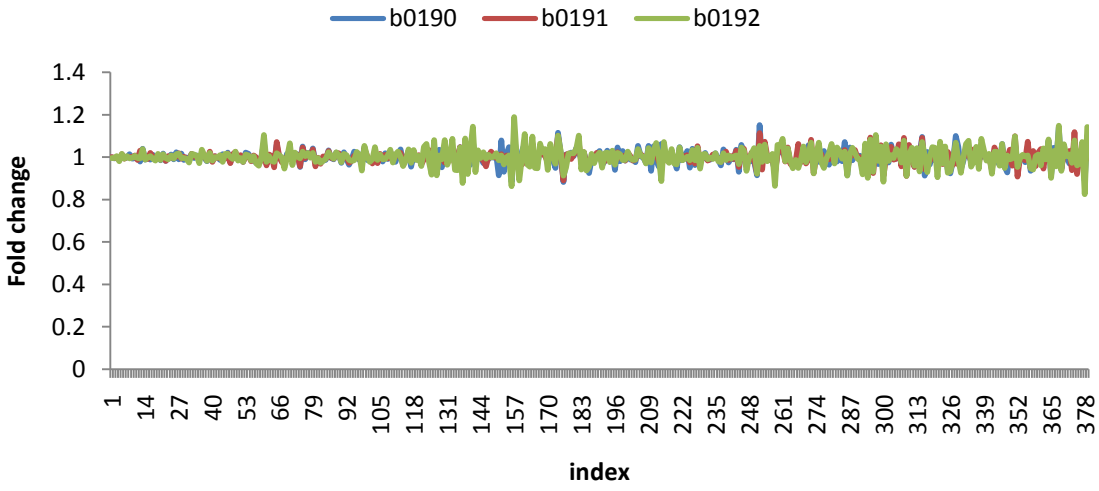

Figure 2.6: The first 380 fold changes of three genes in operon 3863

Though the expression levels of genes in the same operon are different, the fold change of genes in the same operon should be close to each other because they transcribe together. 72010 fold changes of b0190, b0191 and b0192 were calculated based on formula 2.1. The first 380 fold changes are shown in Figure 2.6. The fold changes of three genes are around one. There are no significant difference among the fold changes of b0190, b0191 and b0192.

3.2.2 VARIANCE OF FOLD CHANGE

For any two conditions, the fold change of genes in the same operon should be close to each other. Hence, the variance of fold changes of within operon genes should near zero. On the contrary, if fold change of genes in the same operon are totally different, the variance of fold changes of within operon genes would be far from zero. The formula for the variance of fold changes of within operon genes is generated as follows:

$$var_{mij} = \frac{\sum_{n=1}^{p_m}[\frac{e_{ni}}{e_{nj}} - \frac{(\sum_{n=1}^{p_m}\frac{e_{ni}}{e_{nj}})}{p_m}]^2}{p_m - 1}, \quad (2.2)$$

where var_{mij} represents variance of fold changes of genes in the operon m under condition i and condition j, e_{ni} represents the expression level of gene n in condition i, e_{nj} represents the expression level of gene n in condition j, p_m is the size of operon m for $m = 1, \cdots\cdots 827, n = 1, \cdots\cdots p,\ i = 1, \cdots\cdots 380, j = 1, \cdots\cdots 380\ and\ j > i$.

In order to predict whether an operon include an alternative transcription unit, 827 operons' variance of fold changes of within operon genes were calculated based on formula 2.2.

If the variance of fold changes of within operon genes far from zero, the genes in the operon would not transcribe together. This presents the probability of existence of alternative transcription unit.

Figure 2.7 shows the frequency distribution of variance of fold changes in operon 3863. The range of variance of fold changes is from 0 to 0.021. About 67.13% of variance of fold changes are less than 0.001 and 10.99% of variance of fold changes are larger than 0.003. The histogram

follows a right skewed curve, with the peak at 0.001. With the increase variance, the frequency decrease sharply.

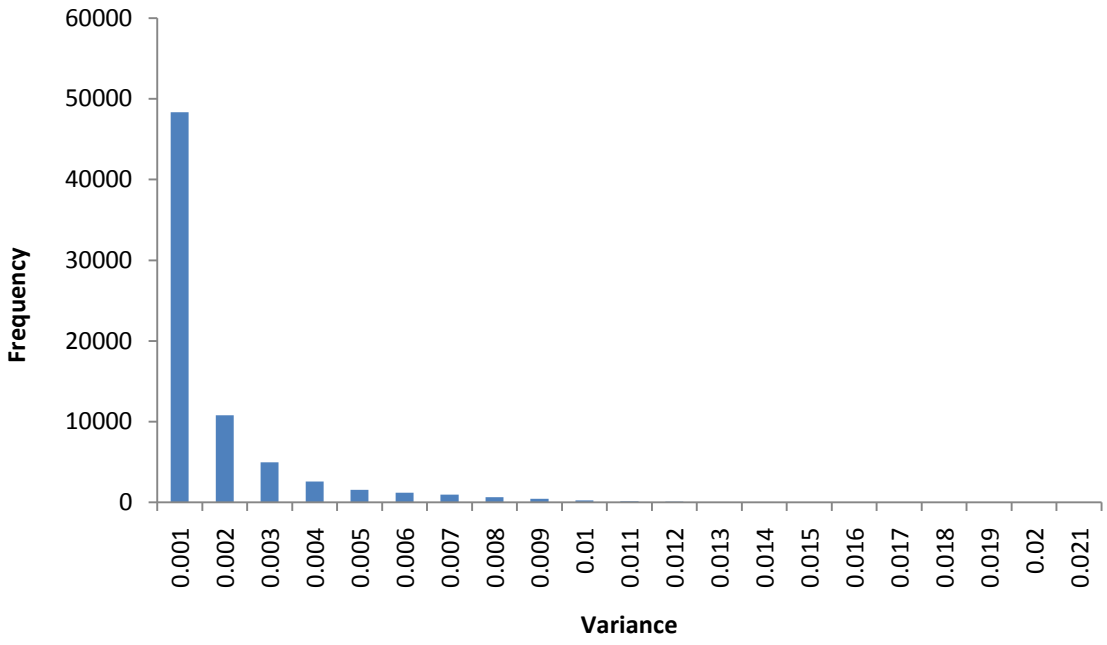

Figure 2.7: The frequency distribution of variance of fold changes in operon 3863

3.2.3 90% Quantile Of Variance Of Fold Changes

What we are interested in Figure 2.7 is 10.99% of variance of fold changes is larger than 0.003, which presents the probability of existence of alternative transcription unit. So we calculated the 90% quantile of variance of fold changes for all operons. The function of 90% quantile of variance of fold changes is defined as follows:

$$Pr(Var_i < Var_{90\%}) = 0.9 \ , \qquad (2.3)$$

where Var_i represents the i-th variance of fold changes , $Var_{90\%}$ represents 90% quantile of variance of fold changes, for $i = 1, \cdots\cdots 72010$.

According to distribution of 90% quantile of variance of fold changes, all operons are grouped into three categories: operons without alternative transcription units, operons may contain alternative transcription units and operons with alternative transcription units. Operons without alternative transcription units were eliminated from the data. The other two groups of operons were kept for further analysis.

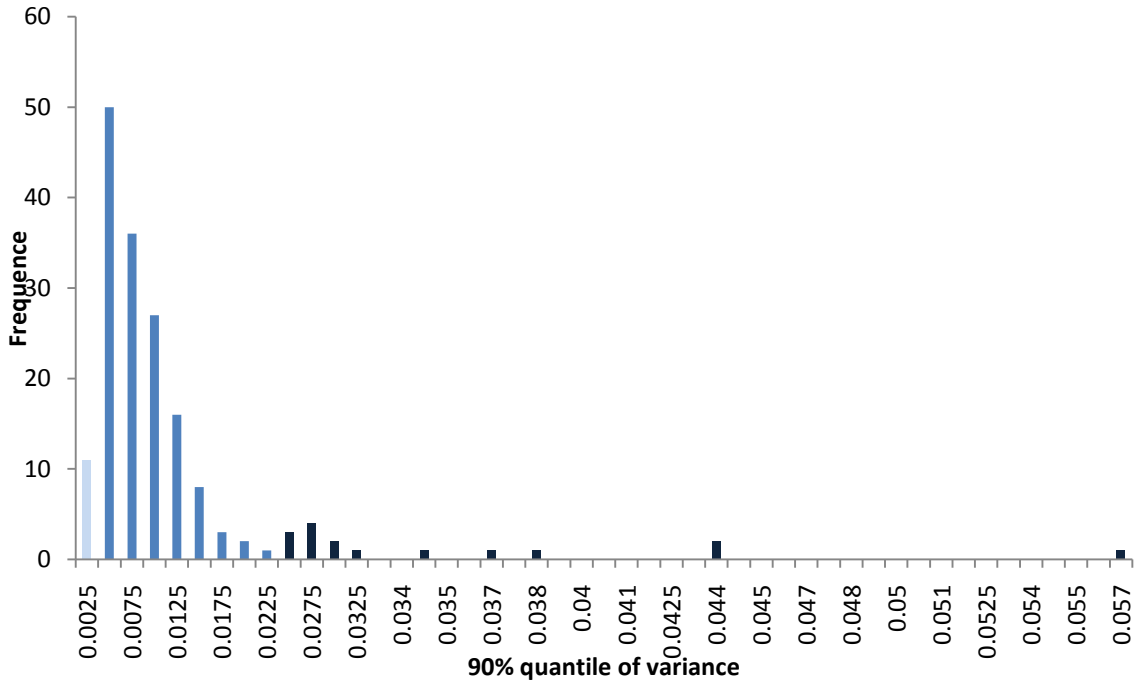

Figure 2.8 The frequency distribution of 90% quantile of variance of fold changes for all three-gene operons.

All the operons' 90% quantile of variance of fold changes were calculated based on formula 2.3. Figure 2.8 shows the frequency distribution of 90% quantile of variance of fold changes for all three-gene operons. The three-gene operons are grouped into three categories: operons without alternative transcription units (90% quantile of variance less than 0.0025), operons may contain alternative transcription units (90% quantile of variance larger than 0.0025 and less than 0.025) and operons with alternative transcription units (90% quantile of variance larger than 0.025). Hence, we eliminated three-gene operons without alternative transcription units.

3.2.4 99% Quantile Of Variance Of Fold Changes

Multi-promoters operons generally transcribe as a complete transcription unit in most conditions, but they will transcribe as an alternative transcription unit under some conditions. The *trp* operon involved in tryptophan biosynthesis has two promoters, which give rise to a complete transcription unit and an alternative transcription unit. Transcription of the alternative transcription unit in vivo is approximately 15% of the complete transcription unit [16].

Among 72010 variances of fold changes for each operon, most of them are close to zero, representing operons transcribe as a complete transcription unit. Hence, the largest portion of variances becomes the major concern because operons may transcribe as an alternative transcription unit under these conditions.

The largest 1% of variances (720 variances) was outputted if the variance larger than 99% quantile of variance. The function of 90% quantile of variance of fold changes is defined as follows:

$$Pr(Var_i < Var_{99\%}) = 0.99 \ , \qquad (2.4)$$

where Var_{mi} represents the i-th variance of fold changes , $Var_{99\%}$ represents 99% quantile of variance of fold changes, for $i = 1, \cdots \cdots 72010$.

All the operons' 99% quantile of variance of fold changes were calculated based on formula 2.4. Figure 2.9 shows the frequency distribution of 99% quantile of variance of fold changes for all three-gene operons. For each three-gene operon, if the variance larger than its 99% quantile of variance, it would be kept for further analysis. After this process, each three-gene operon has the largest 720 variance of fold changes.

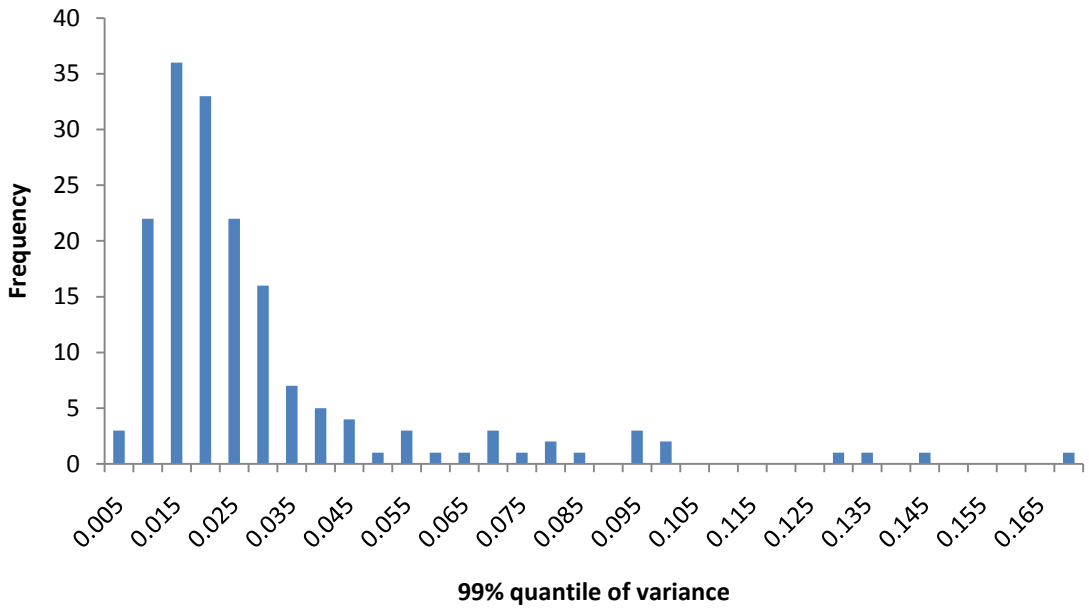

Figure 2.9 The frequency distribution of 99% quantile of variance of fold changes for all three-gene operons.

3.2.5 CONDITIONS IDENTIFICATION

Each operon has the largest 720 variance of fold changes. Identifying conditions in which operons transcribed as an alternative transcription unit is the next step. Since fold change is the ratio of expression level for any two conditions, there are totally 1440 conditions involved in 720 variances for each operon. The frequency distribution of conditions from 1440 conditions for each operon is calculated. The three highest frequency conditions were recorded for further analysis. The probability of presence of alternative transcription units in these three conditions is relatively higher than other three conditions.

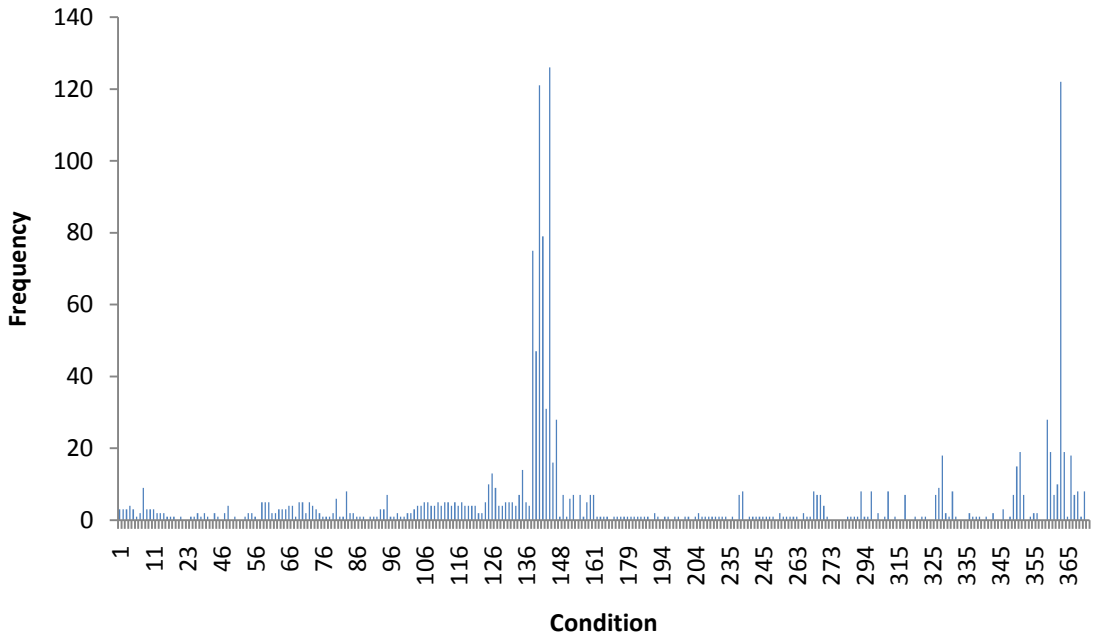

Figure 2.10 The frequency distribution of conditions in operon 4124

The frequency distribution of conditions in 720 highest variances for operon 4124 is displayed in Figure 2.10. The frequency of condition 140, condition 143 and condition 363 are 121, 126 and 122 respectively. These three conditions are recorded for further analysis.

3.2.6 RATIO

After conditions identification, each operon has three conditions in which the operon may transcribe as an alternative TU. The ratio between expression level and average expression level becomes the major concern. The function of ratio is defined as follows:

$$r_{ij} = \frac{e_{ij}}{\sum_{j=1}^{380} e_{ij} / 380} \ , \tag{2.5}$$

where r_{ij} is the ratio of gene *i* under condition *j* and e_{ij} represents the expression level of gene *i* in condition *j*.

The ratio of three identified conditions for each gene was calculated. If the ratio of contiguous genes within operon is close to each other, they probably transcribe together to give rise to an alternative TU.

3.2.7 INTERGENIC DISTANCE

Since the intergenic distance is one of the most effective features for predicting operon, we also used this feature in alternative TU prediction. If the intergenic distance between two adjacent genes is less than or equal to zero, we could not separate the two genes into two alternative TUs because there is no promoters or terminators between two genes. The formula of intergenic distance is defined as follows:

$$d = \begin{cases} G_{ds} - G_{ue} \ , strand = " + " \\ G_{us} - G_{de} \ , strand = " - " \end{cases} , \tag{2.6}$$

where G_{ds} is the start position of downstream gene, G_{ue} is the end position of upstream gene, G_{us} is the start position of upstream gene, G_{de} is the end position of downstream gene, $" + "$ represents that the direction of transcription is forward and $" - "$ represents that the direction of transcription is reverse.

The modified formula of intergenic distance is generated as:

$$DI = \begin{cases} d - 1, \ d > 0 \\ d, \ d = 0 \\ d + 1, d < 0 \end{cases} , \tag{2.7}$$

where *d* is calculated from the formula 2.6.

All the intergenic distances between two adjacent genes in the same operon were calculated based on modified formula of intergenic distance. Figure 2.11 shows the frequency distribution of intergenic distance within operon. 33.99% of intergenic distances are less than or equal to zero, which means there is no promoter or terminator between two genes. In other words, we could not separate the two genes into two alternative TUs if the intergenic distance between two adjacent genes is less than or equal to zero.

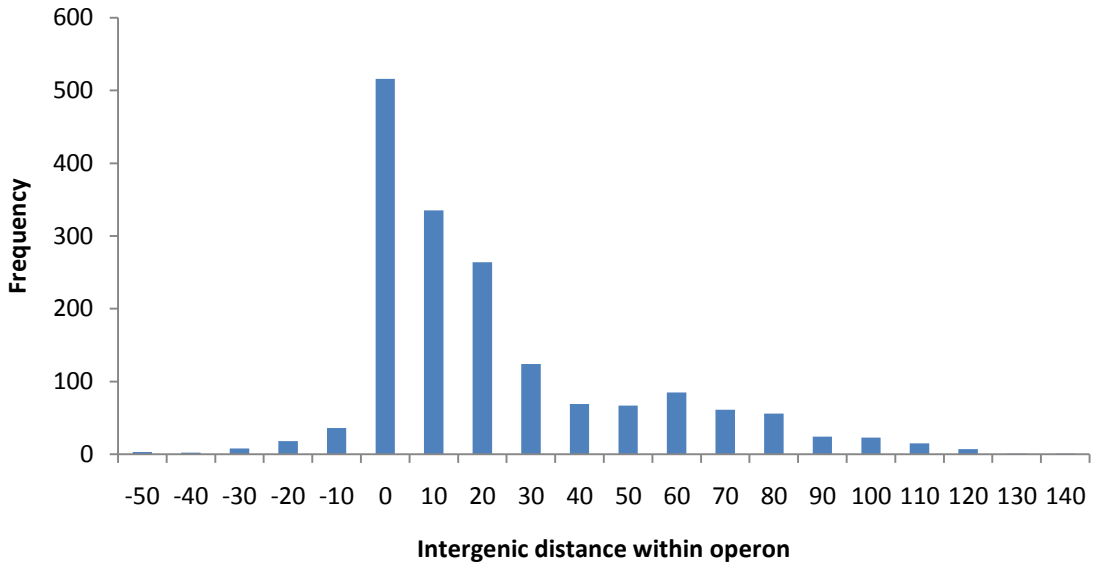

Figure 2.11 The frequency distribution of intergenic distances within operon

3.2.8 *K*-MEANS ALGORITHM

The *K*-means algorithm is a widely used partition-based clustering method. Given the ratio of three identified conditions for each gene and a per-specified parameter *K*=2, the algorithm partitions the data set into 2 disjoint subsets which minimize the following function:

$$V_t = \sum_{k=1}^{2} \sum_{G_i \in S_k} |r_{ij} - \mu_{kj}|^2 \qquad (2.8)$$

where V_t is the variance of operon *t*, G_i is gene *i* within operon *t*, r_{ij} is the ratio of G_i under condition *j* and S_k is subset of operon *t* and μ_{kj} represent the average ratio of subset S_k under condition *j*.

Suppose the size of operon is n, there are totally (n-1) ways to divide the operon into 2 subsets. Therefore, the purpose of function V_t is to minimize the sum of the squared distances of genes from their subsets.

If the intergenic distance between the last gene in first subset and the first gene in second subset is larger than zero, the second subset is generally considered as an alternative TU because it share the same terminator with the complete TU (Figure 1.2).

3.3 RESULTS

We implemented alternative TU predictor based on operon data from DOOR and evaluated alternative TU predictor based on TU and operon data from RegulonDB. However, there are some differences between predicted operons from DOOR and predicted operons from RegulonDB. So we evaluated alternative TUs in operons, which are contained in both DOOR and RegulonDB. Since RegulonDB is a incomplete TU database, we eliminated alternative TUs that are not included in RegulonDB.

3.3.1 PRESENCE EVALUATION OF ALTERNATIVE TU IN OPERON

The first step to predict alternative TU is to estimate whether an operon has an alternative TU. According to K-MEANS variance method, the operons are grouped into two categories: operons with alternative TU and operons without alternative TU. The operons being predicted correctly is coded as 1, whereas the operons with alternative TU being predicted to operons without alternative TU is coded as 0.

Table 2.1: Performances of presence of alternative TU in operon

Correct	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	17	6.14	17	6.14
1	260	93.86	277	100.00

Table 2.1 shows the performances of presence of alternative TU in operon. 260 out of 277 operons were predicted correctly compared with experimental identified TU data from RegulonDB. The alternative TU predictor achieves 93% prediction accuracy in estimating presence of alternative TU in *Escherichia coli* operons. The performance of presence of alternative TU in operons with different sizes is displayed in Appendix A.

3.3.2 Validation Of Alternative Tu Prediction

K-means algorithm and intergenic distance were used in prediction of alternative TU. The alternative TUs being predicted correctly is coded as 1, whereas the alternative TUs being predicted incorrectly is coded as 0.

Table 2.2 shows the performances of prediction of alternative TU in operon. 248 out of 294 operons were predicted correctly compared with experimental identified TU data from RegulonDB. The alternative TU predictor achieves 84.3% prediction accuracy in prediction of alternative TU in *Escherichia coli* operons. The performance of prediction of alternative TU in operons with different sizes is displayed in Appendix B.

Table 2.2: Performances of prediction of alternative TU in operon

Correct	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	46	15.65	46	15.65
1	248	84.35	294	100.00

CHAPTER 4

DERIVATION OF THE COMPLETE TRANSCRIPTOME

In prokaryotes such as *Escherichia coli*, operons are described adjacent genes than transcribed into a single mRNA [2] .For an operon including multiple promoters, a fraction of its genes can be present in several different alternative TUs in different conditions [7]. None of the existing operon predictors are able to deal with alternative TUs.

Since we have predicted the complete alternative TUs in *Escherichia coli*, we combine alternative transcription units with complete transcription units to form the transcriptome of *Escherichia coli.*

4.1 COMPLETE TRANSCRIPTION UNIT

Given the definition of operon and transcription unit, an operon contains at least one transcription unit called complete transcription unit that includes all the genes in that operon.

We derived multiple-gene complete TU of *Escherichia coli* from operon data form DOOR. Since DOOR operon data contain operons with two or more genes, those eliminated genes that are not included in DOOR operon data are considered as single-gene complete transcription units. We combined single-gene complete TU with multiple-gene complete TU as complete TU.

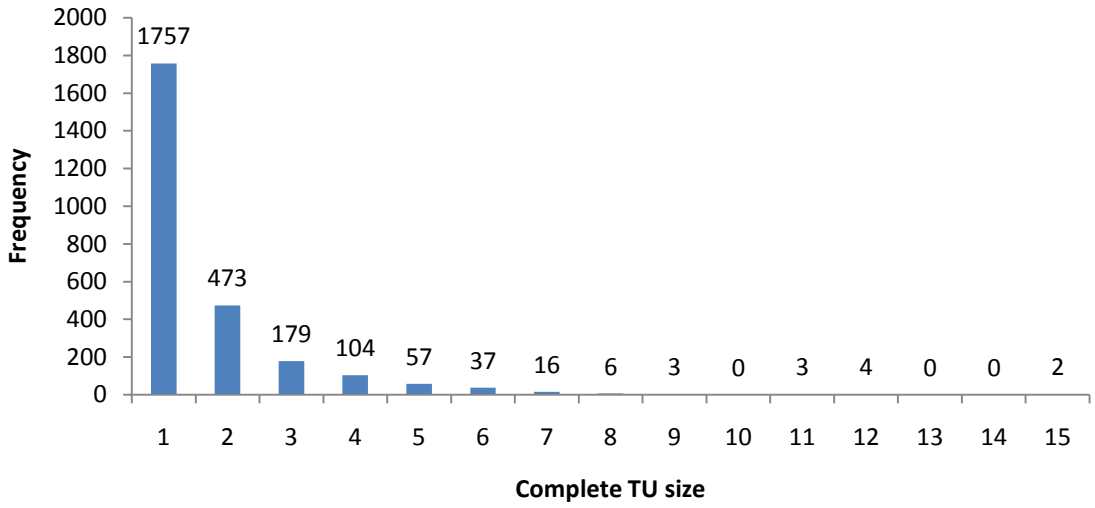

Figure 3.1 The frequency distribution of complete TU size

Figure 3.1 shows the frequency distribution of complete TU size. There are 2641 complete TUs in total. Single-gene complete TUs account for 66.53% of all complete TUs, two-genes complete TUs account for 17.91% and three-genes complete TUs account for 6.77% of all completE TUs. There are no complete TUs consist of ten, thirteen or fourteen genes.

4.2 Alternative Transcription Unit

The first step to predict alternative transcription units is to estimate whether an operon has an alternative transcription unit based on Database of Prokaryotic Operons (DOOR) and Many Microbe Microarrays Database (M3D). Based on variance of fold changes of within operon genes and intergenic distance, we predicted alternative TUs of *Escherichia coli*.

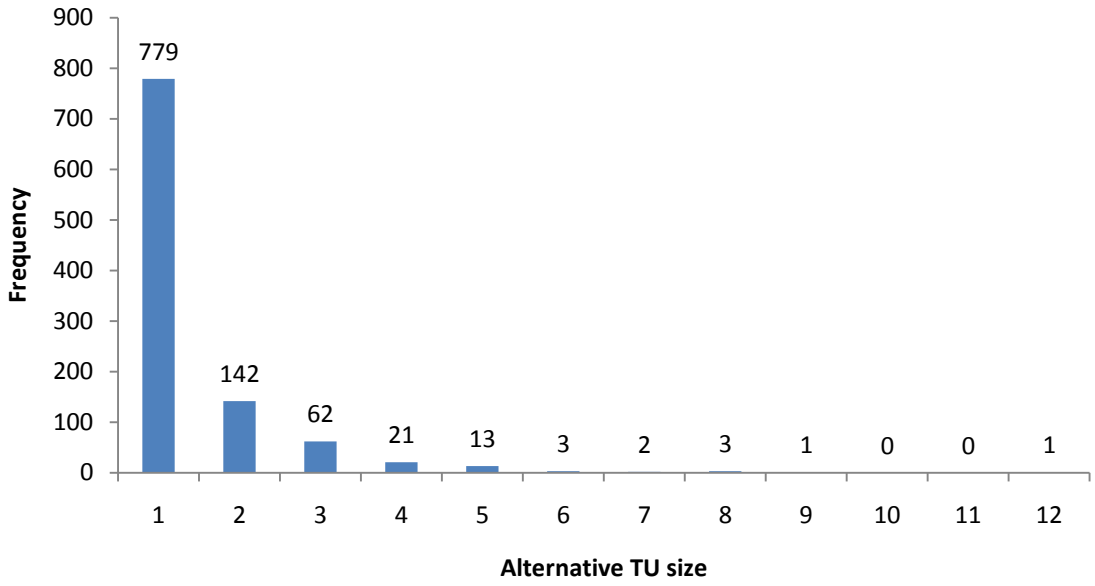

Figure 3.2 The frequency distribution of alternative TU size

Based on prediction model described above, we derived 1027 alternative TUs. Figure 3.2 reveals the frequency distribution of alternative TU size. Among these alternative TUs, 779 alternative TUs contains only one gene, 142 alternative TUs include two genes, 62 alternative TUs consist of three genes.

4.3 TRANSCRIPTOME

The expression level of genome in *Escherichia coli* have been analyzed, using K-means algorithm and intergenic distance. The TU predictor revealed 1027 alternative TUs and 2641 complete TUs. Transcriptome is defined as all transcribed regions encoded in the genome. Hence, transcriptome is a complete collection of alternative TUs and complete TUs.

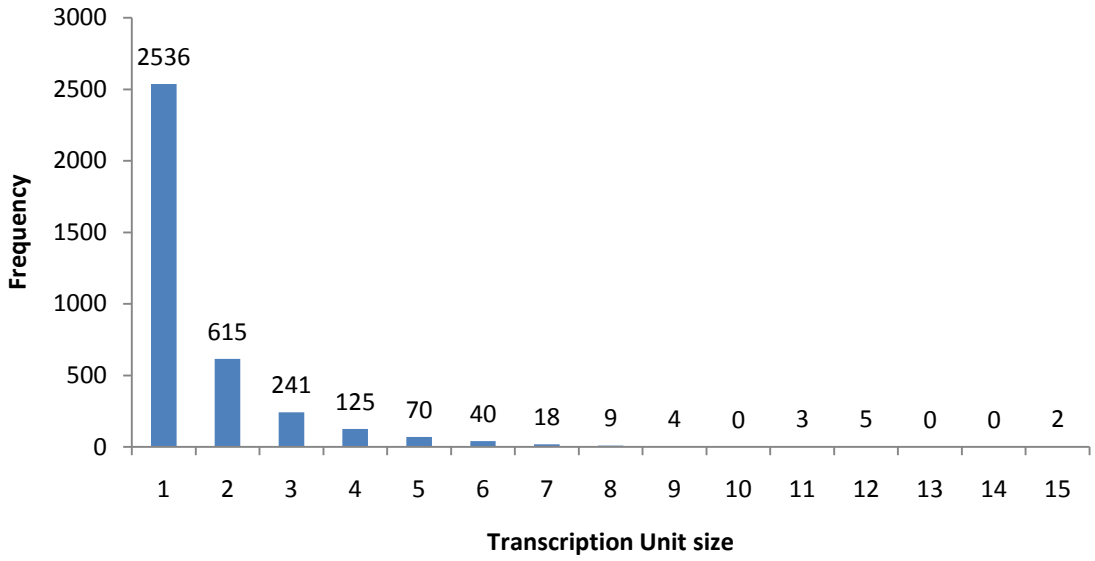

Figure 3.3 The frequency distribution of TUs size

Figure 3.3 shows the frequency distribution of TUs size. Transcriptome consist of 3668 transcription units. 69.13% of TUs contain one gene, 16.77% of TUs include two genes, 6.57% of TUs consist of three genes and 7.52% of TUs contain three or more genes.

4.4 RESULT

4.4.1 EVALUATION OF SINGLE-GENE CTU

We derived single-gene complete transcription units (CTU) from DOOR. Single-gene CTU was verified by single-gene operons from RegulonDB. Since RegulonDB is an incomplete TU database, we eliminated single-gene CTU that is not included in RegulonDB. The single-gene CTU predicted correctly is coded as 1, whereas the single-gene CTU being predicted incorrectly is coded as 0.

Table 3.1: Evaluation of single-gene CTU

Correct	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	351	19.57	351	19.57
1	1443	80.43	1794	100.00

Table 3.1 shows the percentage of known single-gene CTU correctly predicted. 80.43% of single-gene CTU is predicted correctly. All single-gene CTU were tested on single-gene operons from RegulonDB database.

4.4.2 Validation Of Transcriptome

Since the difference of predicted operons between DOOR and RegulonDB, we evaluated transcriptome in operons, which are contained in both DOOR and RegulonDB. TUs that are not included in RegulonDB were eliminated because we are unable to verify them. The TU predicted correctly is coded as 1, whereas the TU being predicted incorrectly is coded as 0.

Table 3.2: Evaluation of transcriptome from multiple-genes operons

Correct	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	46	8.06	46	8.06
1	525	91.94	571	100.00

Table 3.2 shows the evaluation of transcriptome from multiple-genes operons. 91.94% of TUs (include CTU and ATU) from multiple-genes operons are correctly predicted. The evaluation of TUs in operons with different sizes is displayed in Appendix C.

CHAPTER 5

CONCLUSION

Transcriptome is defined as the all transcribed regions encoded in the genome. Experimentally defining the complete transcriptome of prokaryotic organisms has been a challenging task. Hence, despite the fact that numerous species have been sequenced, only few transcriptomes have been extensively identified. The availability of complete genomic sequences and microarray expression data calls for computational methods for characterizing transcriptome, the complete collection of alternative transcription units (ATU) and complete transcription units (CTU).

Though numerous computational methods have been developed for prediction of operons (CTU), none of existing computational methods can deal with ATU. We have presented a new computational method for TU prediction, which is able to predict alternative transcription unit. Since the first model organism for molecular biology is *Escherichia coli*, we implemented the new TU predictor to produce the transcriptome of *Escherichia coli.*

The first step to predict ATU is to test whether an operon has an ATU based on Database of Prokaryotic Operons (DOOR) and Many Microbe Microarrays Database (M3D). Then ATU of *Escherichia coli* was predicted based on variance of fold changes of within operon genes and intergenic distance. Lastly, single-gene CTU, multiple CTU and ATU were combined to form the transcriptome of *Escherichia coli*.

Predicted TUs were tested on known TUs of *Escherichia coli* from RegulonDB. The alternative TU predictor achieves 93% prediction accuracy in estimating presence of ATU in *Escherichia coli* operons. The percentage of known ATUs correctly predicted and known single-gene CTU correctly predicted are 84.3% and 80.43% respectively. 91.94% of TUs (include CTU and ATU) from multiple-genes operons are correctly predicted.

We plan to use this computational prediction method in the prediction of operons (CTU) in the future. Based on the predicted result of both CTU and ATU, the transcriptome can be derived automatically when the microarray data are available.

BIBLIOGRAPHY

[1] Bockhorst, J., Craven, M., Page, D., Shavlik, J. and Glasner, J. (2003) A Bayesian network approach to operon prediction. *Bioinformatics*, 19, 1227-1235.

[2] Brian Tjaden, Rini Mukherjee Saxena1, Sergey Stolyar, David R. Haynor, Eugene Kolker and Carsten Rosenow. (2002) Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Research*, 30, 3732-3738

[3] Che, D, Li. G, Mao. F, Wu H, Xu Y. (2006) Detecting uber-operons in prokaryotic genomes. *Nucleic acids research*, 34, 2418-2427

[4] Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y., Jiang, T. (2004) Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome *Nucleic Acids Research,* 32, 2147-2157.

[5] Chiara Sabatti, Lars Rohlin, Min-Kyu Oh and James C. Liao (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Research*, 30, 2886-2893

[6] Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) A probabilistic learning approach to whole-genome operon prediction. *Proc. Conf. Intell. Syst. Mol. Biol.*, 8: 116–127.

[7] D H Calhoun, J W Wallen, L Traub, J E Gray and H F Kung.(1985) Internal promoter in the ilvGEDA transcription unit of *Escherichia coli* K-12. *Journal of Bacteriology*. 161, 128-132

[8] Dam, P., Olman, V., Harris, K., Su, Z., Xu, Y. (2007) Operon prediction using both genome-specific and general genomic information. *Nucleic acids research*, 35, 288-298

[9] DeRisi JL, Iyer VR, Brown PO. (1997) Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, 278, 680-686.

[10] Dhammika Amaratunga, Javier Cabrera. (2003) *Exploration and analysis of DNA microarray and protein array data*. Wiley, New York.

[11] Edwards, M.T., Rison, S.C., Stoker, N.G. and Wernisch, L. (2005) A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. *Nucleic acids research*, 33, 3253-3262.

[12] Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Research*, 29,1216–1221.

[13] F. Mao, P. Dam, J. Chou, V. Olman, Y. Xu (2009) DOOR: A Database of prOkaryotic OpeRons. *Nucleic acids research.* 37: D459-D463

[14] Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, and Gardner TS. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*, 36, D866–D870

[15] Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, et al. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K- 12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic acids research.* 36, D120-124

[16] Horowitz H, Platt T (1982) Identification of trp-p2, an internal promoter in the tryptophan operon of *Escherichia coli. Journal of Molecular Biology*, 156, 257-267

[17] Hu, J., Li, B., Kihara, D (2005) Limitations and potentials of current motif discovery algorithms.*Nucleic acids research*, 33, 4899-4913

[18] Jacob, F; Perrin, D; Sanchez, C; Monod, J (1960). Operon: a group of genes with the expression coordinated by an operator. *Proceedings of the French Academy of Science* ,250**,** 1727–1729

[19] Mihaela Pertea, Kunmi Ayanbule, Megan Smedinghoff and Steven L. Salzberg. (2009) OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Research*, 37,D479-D482

[20] Moreno-Hagelsieb,G. and Collado-Videe,J. (2002) A powerful nonhomology method for the prediction of operons in prokaryotes. *Bioinformatics*. 18, S329–336.

[21] P. Roback, J. Beard, D. Baumann, C. Gille, K. Henry, S. Krohn, H. Wiste, M.I. Voskuil, C. Rainville and R. Rutherford (2007) A predicted operon map for Mycobacterium tuberculosis. *Nucleic Acids Research,* 35,5085-5095

[22] Price, M.N., Huang, K.H., Alm, E.J., Arkin, A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research,* 33, 880-892

[23] Romero, P.R. and Karp, P.D. (2004) Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway genome databases. *Bioinformatics*, 20, 709-717.

[24] Saeed Tavazoie, Jason D. Hughes, Michael J. Campbell, Raymond J. Cho George M. Church (1999) Systematic determination of genetic network architecture. *Nature Genetics*, 22, 281 - 285

[25] Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides, J. (2000). Operons in *Escherichia coli*: genomic analyses and predictions. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 6652-6657.

[26] Shujiro Okuda, Toshiaki Katayama, Shuichi Kawashima, Susumu Goto and Minoru Kanehisa (2006) ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic acids research.* 34, D358–D362

[27] Sierro N., Makita Y., de Hoon M.J.L. and Nakai K. (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Research*, 36,D93-D96

[28] Stekel, D. (2003) *Microarray bioinformatics.* Cambridge University Press

[29] T.R.Golub, D.K.Slonim, P.Tamayo, C.Huard, M.Gaasenbeek, J.P.Mesirov, H.Coller, M.L.Loh, J.R.Downing, M.A.Caligiuri, C.D.Bloomfield, (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286, 531-537.

[30] Tjaden,B., Haynor,D.R., Stolyar,S., Rosenow,C. and Kolker,E.(2002) Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. *Bioinformatics*, 18, S337–S344.

[31] Tran, T.T., Dam, P., Su, Z., Poole, F.L., 2nd, Adams, M.W., Zhou, G.T. and Xu, Y. (2007) Operon prediction in Pyrococcus furiosus. *Nucleic Acids Research,* 35, 11-20.

[32] Victor E. Velculescu, Lin Zhang, Wei Zhou, Jacob Vogelstein, Munira A. Basrai, Douglas E. Bassett, Jr., Phil Hieter, Bert Vogelstein and Kenneth W. Kinzler (1997) Characterization of the Yeast Transcriptome. *Cell*, 88, 243-251

[33] Xu,Ying (2004) Computational genome annotation. *Microbial Functional Genomics. Wiley-LISS*, Hoboken, NJ.

[34] Xu,Ying; J.Peter Goqarten (2008). *Computational methods for understanding bacterial and archaeal genomes*. Imperial College Press, London

[35] Yada, T., Nakao, M., Totoki, Y. and Nakai, K. (1999) Modeling and predicting transcriptional units of Escherichia coli genes using Hidden Markov models. *Bioinformatics*, 15, 987-993.

[36] Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J. Adiconis X, Schroth G, Luo S,

Khrebtukova I, Gnirke A, Nusbaum C, Thompson DA, Friedman N, Regev A. (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences*, 106:3264-3269.

[37] Zhang,aidong (2006) *Advanced analysis of gene expression microarray data.* World Scientific, NJ

APPENDICES

APPENDIX A.

THE PERFORMANCE OF PRESENCE OF ALTERNATIVE TU IN OPERONS WITH DIFFERENT SIZES

frq=2

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	10	7.75	10	7.75
1	119	92.25	129	100.00

frq=3

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	4	6.25	4	6.25
1	60	93.75	64	100.00

frq=4

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	3	7.89	3	7.89
1	35	92.11	38	100.00

frq=5

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	21	100.00	21	100.00

frq=6

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	11	100.00	11	100.00

frq=7

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	100.00	6	100.00

frq=8

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	100.00	1	100.00

frq=9

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	100.00	1	100.00

frq=11

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	100.00	2	100.00

frq=12

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3	100.00	3	100.00

frq=15

ort	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	100.00	1	100.00

APPENDIX B.

THE PERFORMANCE OF PREDICTION OF ALTERNATIVE TU IN OPERONS WITH DIFFERENT SIZES

frq=2

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**11**	**6.15**	**11**	**6. 15**
1	**168**	**93.85**	**179**	**100.00**

frq=3

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**11**	**26.83**	**11**	**26.83**
1	**30**	**73.17**	**41**	**100.00**

frq=4

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**6**	**46.15**	**6**	**46.15**
1	**7**	**53.85**	**13**	**100.00**

frq=5

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**5**	**20**	**5**	**20**
1	**20**	**80**	**25**	**100.00**

frq=6

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**9**	**39.13**	**9**	**39.13**
1	**14**	**60.87**	**23**	**100.00**

frq=7

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**3**	**30**	**3**	**30**
1	**7**	**70**	**10**	**100.00**

frq=8

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	**1**	**100.00**	**1**	**100.00**

frq=9

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	**1**	**100.00**	**1**	**100.00**

frq=15

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**1**	**50**	**1**	**50**
1	**1**	**50**	**2**	**100.00**

APPENDIX C.

THE EVALUATION OF TUS IN OPERONS WITH DIFFERENT SIZES

frq=2

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11	3.57	11	3.57
1	297	96.43	308	100.00

frq=3

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**11**	**10.48**	**11**	**10.48**
1	**94**	**89.52**	**105**	**100.00**

frq=4

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**6**	**11.76**	**6**	**11.76**
1	**45**	**88.24**	**51**	**100.00**

frq=5

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**5**	**10.87**	**5**	**10.87**
1	**41**	**89.13**	**46**	**100.00**

frq=6

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**9**	**26.47**	**9**	**26.47**
1	**25**	**73.53**	**34**	**100.00**

frq=7

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**3**	**18.75**	**3**	**18.75**
1	**13**	**81.25**	**16**	**100.00**

frq=8

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	**1**	**100.00**	**1**	**100.00**

frq=9

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	**2**	**100.00**	**2**	**100.00**

frq=11

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	**2**	**100.00**	**2**	**100.00**

frq=12

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	**3**	**100.00**	**3**	**100.00**

frq=15

dt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**1**	**33.33**	**1**	**33.33**
1	**2**	**66.67**	**3**	**100.00**