

Spring 2007

Prediction of mRNA polyadenylation sites in the human genome and Mathematical modeling of alternative polyadenylation

Yiming Cheng

New Jersey Institute of Technology

Follow this and additional works at: <https://digitalcommons.njit.edu/dissertations>



Part of the [Mathematics Commons](#)

Recommended Citation

Cheng, Yiming, "Prediction of mRNA polyadenylation sites in the human genome and Mathematical modeling of alternative polyadenylation" (2007). *Dissertations*. 813.

<https://digitalcommons.njit.edu/dissertations/813>

This Dissertation is brought to you for free and open access by the Theses and Dissertations at Digital Commons @ NJIT. It has been accepted for inclusion in Dissertations by an authorized administrator of Digital Commons @ NJIT. For more information, please contact digitalcommons@njit.edu.

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be “used for any purpose other than private study, scholarship, or research.” If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of “fair use” that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select “Pages from: first page # to: last page #” on the print dialog screen



The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

PREDICTION OF MRNA POLYADENYLATION SITES IN THE HUMAN GENOME AND MATHEMATICAL MODELING OF ALTERNATIVE POLYADENYLATION

by
Yiming Cheng

Messenger RNA (mRNA) polyadenylation plays many important roles in the cell, such as transcription termination, mRNA stability and transportation, and mRNA translation in eukaryotic cells. A large number of human and mouse genes have multiple polyadenylation sites (referred to as poly(A) sites) that lead to variable transcripts, some of which are translated into various protein products with different functions. However, the details about when and where the polyadenylation occurs, and how pre-mRNA switches from one poly(A) site to another are still unknown. This kind of 3'-end processing can be regulated by the cell environment, cell cycle stage, and tissue type.

It is generally accepted that the cleavage of pre-mRNA is based on the sequence of nucleotides around the poly(A) sites. So it is possible to predict the poly(A) sites accurately based on the pre-mRNA sequence. To accomplish the supervised prediction of a poly(A) site, a set of statistical models has been used, such as linear discriminant analysis, quadratic discriminant analysis, and support vector machine (SVM). Among these, SVM was chosen as the classification algorithm for the prediction of poly(A) sites in this work. A program called `polya_svm` has been developed using PERL. The true positive and accuracy results obtained using this method are better than the results obtained using other commonly used algorithms.

Compared with the microarray technique, serial analysis of gene expression (SAGE) is another powerful technology for measuring the mRNA expression levels. Our study is the first investigation of the regulation of the transcripts from the same gene by analyzing the SAGE data. By filtering the noise data from the database and calculating the correlation between transcripts from the same unigene cluster, some significant genes are found to have multiple transcripts with opposite expression levels. These genes might be very interesting to biologists and they are worth being verified by biological experiments.

Alternative polyadenylation has been found to be very common in human and mouse genes recently. It has been believed that the selection of different poly(A) sites is related to biological factors such as the developmental stages, cell conditions, and the availability and abundance of some protein factors. However, it is not clear how these factors affect alternative polyadenylation. Mathematical modeling is applied to understand the dynamical selection of poly(A) sites. Cleavage stimulation Factor (CstF) is a very important protein complex required for efficient cleavage, containing subunits of 77, 64, and 50 kD (CstF-77, CstF-64, CstF-50). It has been found that human *cstf-77* gene has several different transcripts due to the alternative polyadenylation and the expression levels of these transcripts display some auto-regulation. A mathematical model with a time delay is constructed to simulate the dynamical gene expression levels of gene *cstf-77*. Experimental data are compared with the model. This kind of mathematical model can also be extended to some other polyadenylation factors that have similar alternative polyadenylation patterns.

**PREDICTION OF MRNA POLYADENYLATION SITES
IN THE HUMAN GENOME AND MATHEMATICAL MODELING
OF ALTERNATIVE POLYADENYLATION**

by
Yiming Cheng

**A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey - Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences**

**Department of Mathematical Sciences, NJIT
Department of Mathematics and Computer Science, Rutgers-Newark**

May 2007

Copyright © 2007 by Yiming Cheng

ALL RIGHTS RESERVED

APPROVAL PAGE

**PREDICTION OF MRNA POLYADENYLATION SITES
IN THE HUMAN GENOME AND MATHEMATICAL MODELING
OF ALTERNATIVE POLYADENYLATION**

Yiming Cheng

Dr. Robert M. Miura, Dissertation Advisor
Professor of Mathematical Sciences, NJIT

Date

Dr. Bin Tian, Dissertation Co-Advisor
Assistant Professor of Biochemistry and Molecular Biology, UMDNJ

Date

Dr. Bruce Byrne, Committee Member
Adjunct Professor of Medicine and Molecular Genetics & Microbiology, and
Associate Director for Education, the Information Institute, UMDNJ

Date

Dr. Sunil K. Dhar, Committee Member
Associate Professor of Mathematical Sciences, NJIT

Date

Dr. Jorge P. Golowasch, Committee Member
Associate Professor of Mathematical Sciences, NJIT and
Biological Sciences, Rutgers-Newark

Date

BIOGRAPHICAL SKETCH

Author: Yiming Cheng
Degree: Doctor of Philosophy
Date: May 2007

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, 2007
- Master of Science in Computational Mathematics, Fudan University, Shanghai, P. R. China, 2002
- Bachelor of Science in Pure Mathematics, Fudan University, Shanghai, P. R. China, 1999

Major: Mathematical Sciences

Presentations and Publications:

Y. Cheng, R. M. Miura, B. Tian,
“Highly accurate prediction of mRNA polyadenylation sites using a support vector machine”, *Bioinformatics* 22(19):2320-5, 2006.

Y. Cheng, Z. Kang, W. Wang,
“Minimal Residual Polynomial Method on the Asymmetric System”. *Journal of Fudan University (Natural Science)*, Vol. 42:160-4, 2003.

Y. Cheng, R. M. Miura, B. Tian,
“Prediction of Polyadenylation Sites Using Support Vector Machines”, *SIAM Conference on the Life Science*, Raleigh, NC, July 31-Aug. 4, 2006.

Y. Cheng, R. M. Miura, B. Tian,
“Prediction of Polyadenylation Sites Using Support Vector Machines”, *SIAM Annual Meeting Boston*, Boston, MA, July 10-14, 2006 (Invited talk by SIAM Chapters).

- C. Cheng, Y. Cheng, P. Dubovski etc.,
“Backward diffusion methods for digital halftoning”, Mathematical Problem
(IBM) in Industry Workshop, Olin College, Needham, MA, June 12-16, 2006.
- Alwehebi, A. Chakraborty, Y. Cheng, M. Franklin, T. Kiehl, N. Zeev,
“Extracellular Matrix Accumulation & Nutrient Diffusion in Hydrogel”, the 3rd
Mathematical Modeling Camp, RPI, Troy, NY, June 6-9, 2006.
- Y. Cheng, R. M. Miura, B. Tian,
“Prediction of Polyadenylation Sites Using Support Vector Machines”, Frontiers
in Applied and Computational Mathematics, NJIT, Newark, NJ, May 15, 2006.
- Y. Cheng, R. M. Miura, B. Tian,
“Prediction of Polyadenylation Sites Using Support Vector Machines”, NJIT the
2nd Annual Provost’s Student Research Showcase, Newark, NJ, Apr. 12, 2006.
- Y. Cheng, R. M. Miura, B. Tian,
“Prediction of Polyadenylation Sites Using Support Vector Machines”, NJIT
Graduate Student Research Day, Newark, NJ, Mar. 6, 2006.
- Y. Cheng, R. Jin, C. Wang, J. Wu, P. Yao,
“Whole Genome Mining Transcriptional Factors Partners of ER α in Breast
Cancer using ChIP-Chip”, MBI program at Ohio State University, Columbus, OH,
Aug. 1-19, 2005.
- Y. Cheng, R. M. Miura,
“Analysis of Equivalent Distorted Ratchet Potentials”, Einstein in the City: A
Student Research Conference at The City College of New York, New York, NY,
Apr. 11-12, 2005.
- Y. Cheng, R. M. Miura,
“Analysis of Equivalent Distorted Ratchet Potentials”, 2005 Eastern Sectional
Meeting of the American Mathematical Society, Newark, DE, Apr. 2-3, 2005

To the Memory of Nianrun Cheng

ACKNOWLEDGMENT

I would like to express my deep and sincere gratitude to my advisor, Dr. Robert M. Miura, for his insightful guidance, full support, and continuous encouragement. His way of advising is very effective for me: he guided me to the right direction without confining my thoughts and ideas. Dr. Miura has been always available when I had questions or concerns about my research, my English, or even my future career plan, for that I feel greatly thankful. I am very fortunate to have this kindly and knowledgeable advisor who helped me overcome many obstacles and guided me into many new research areas.

I would also like to thank my co-advisor, Dr. Bin Tian from University of Medicine and Dentistry of New Jersey. Without his supervision, I would never know what is bioinformatics, what happens to the polyadenylation process inside the cell; without his help, I would never get the opportunity to apply mathematics to the specific biological problems. His advice on my writing and documentation is very important for my future. Special thanks to Dr. Bruce Byrne, Dr. Sunil K. Dhar and Dr. Jorge P. Golowasch for actively serving as my dissertation committee. Their suggestions and comments are very helpful. Many thanks to Dr. Amitabha K. Bose, Dr. Demetrios T. Papageorgiou, Dr. Yuan N. Young, Dr. Shidong Jiang, and Dr. Sylvester Thompson from Radford University for all kinds of help towards to the completion of the thesis.

The chairperson of department of mathematical sciences, Dr. Daljit S. Ahluwalia, is thanked for the financial support and continuous encouragement. Many thanks to the former and current office staff: Susan Sutton, Padma Gulati, Alba Henderson, Sherri M. Brown, Liliana Boland, and Evette Ma. Their help with the non-academic issues made my life easier at NJIT. I would like to express my sincere gratitude to my roommates:

Hao Wang, Hua Ren, and Bin Cheng. With their accompanying, I have a very memorable and peaceful three-year Harrison life. I would also thank Xinlin Wang, Sipeng Gu, Kuan Xu, Rulan Gong, Yu Zhang, Chunsheng Yang, Zhenhua Pan, and Michael Tsai. After meeting them, life becomes more lively and entertaining.

I am very indebted to my wife, Zhaoxia Ji, for always being there for me and giving me great support and encouragement. I don't even know how I would get the thesis done without her constant love and belief in me. I would also like to thank my friends at Boston: Guandong Zhang, Zhuhua Cai, and Lin Wang. Without them, I can't image what kind of weekend life I would have. My heartfelt thanks go to my parents and sisters for their kindness, love, patience, understanding, and support over the years.

TABLE OF CONTENTS

Chapter	Page
1 INTRODUCTION.....	1
1.1 Background Information	1
1.2 Summary of Results	2
2 BIOLOGICAL BACKGROUND.....	5
2.1 Human Genome	5
2.2 Gene Expression	6
2.3 mRNA Polyadenylation and <i>Cis</i> -element	9
3 PREDICTION OF POLYADENYLATION SITES USING SUPPORT VECTOR MACHINES	11
3.1 Fifteen <i>Cis</i> -elements	12
3.2 Datasets and Position Specificity Score Matrix	14
3.3 Correlation between Fifteen <i>Cis</i> -elements	15
3.4 Training, Testing, and Prediction	17
3.4.1 Comparison of Different Methods	18
3.4.2 Leave-One-Out Method	19
3.4.3 Poly_a_svm	21
3.5 Prediction Poly(A) Sites in the Human Genome	25
3.6 Further Improvements of Poly_a_svm	28
4 SAGE DATA ANALYSIS	32
4.1 Introduction to SAGE	32

TABLE OF CONTENTS
(Continued)

Chapter	Page
4.2 Analysis Pipeline	34
4.2.1 Reliable Libraries Selection	34
4.2.2 Is SAGE Data Tissue-Specific?	36
4.2.3 The Reliable Tags	38
4.2.4 Significant Unigenes	39
4.3 Results and Conclusions	41
5 ALTERNATIVE POLYADENYLATION	43
5.1 Introduction to Alternative Polyadenylation	43
5.2 Protein Factors Involved in Polyadenylation	45
5.3 Proposed Mechanisms for Alternative Polyadenylation	48
5.3.1 Type II Gene Poly(A) Site Switching	48
5.3.2 Type III Gene Poly(A) Site Switching between 3U and 3D	52
5.4 Proposed Model for Alternative Polyadenylation of Type III Genes	54
6 MATHEMATICAL MODELING OF CSTF-77 ALTERNATIVE POLYADENYLATION	60
6.1 Introduction to CstF-77	60
6.2 Proposed Model of Alternative Polyadenylation for CstF-77	62
6.3 Mathematical Description	65
6.4 Theoretical Analysis of Differential-Delay Equation	67
6.5 Numerical Simulations	78
6.6 Parameter Estimation	79

TABLE OF CONTENTS
(Continued)

Chapter	Page
6.7 Results	83
6.7.1 Experimental Data	83
6.7.2 Optimal Parameter Set	83
6.7.3 Different Initial Functions	85
6.7.4 Basin of Attraction and Basin Boundary	89
6.8 Conclusions and Discussion	99
7 SUMMARY AND FUTURE WORK	102
APPENDIX A STATISTICAL ALGORITHMS AND SIGNIFICANCE	106
APPENDIX B PROGRAMMING LANGUAGES	112
APPENDIX C SAGE MATERIALS	113
REFERENCES	121

LIST OF TABLES

Table		Page
3.1	Comparison of LDA, QDA, and SVM	19
3.2	Effects of Leaving Out Some <i>Cis</i> -elements	20
3.3	Comparison of PolyA_svm with Polyadq for Different Types of Positive Poly(A) Sites	26
3.4	Comparison of PolyA_svm with Polyadq for Different Types of Negative Poly(A) Sites	28
3.5	Comparison of Different Methods for Replacing Infinite Values	29
4.1	SAGE Data Example	35
4.2	Nine Significant Unigenes with Negatively Regulated Transcripts	41
A.1	First Order Markov Chain Transition Matrix	106
C.1	Sixty-one Significant Unigenes with Negatively Regulated Transcripts	114

LIST OF FIGURES

Figure	Page
2.1 Central dogma of biology	7
2.2 Central dogma of biology: feedback network	7
2.3 RNA synthesis starting at the promoter and ending at the terminator	8
2.4 Diagram of gene expression	9
3.1 Fifteen <i>cis</i> -elements	13
3.2 One example sequence from polyA_DB	14
3.3 The PSSM of AUE1	15
3.4 Clustering of 15 <i>cis</i> -elements using poly(A) sites from polyA_DB	16
3.5 Heatmap plot of predicted probability results of 100 test sequences	22
3.6 Schematic of 15 <i>cis</i> -elements in the poly(A) region and the search algorithm for polyA_svm	23
3.7 Prediction of positive and negative sequences	24
3.8 HPR vs relative maximum accuracy	30
3.9 Boxplot of polyA_svm scores for different types of poly(A) sites	31
4.1 An outline of SAGE	32
4.2 The pipeline of SAGE data analysis	35
4.3 Correlation between the number of unique tags and the total number of tags	37
4.4 The gene expression levels for different tissues	37
4.5 T-value distribution of random tags	40
4.6 The transcripts structure of Unigene Hs.334885	41

**LIST OF FIGURES
(Continued)**

Figure	Page
5.1 Schematic representation of poly(A) sites	44
5.2 Schematic representation of the steps involved in the mammalian pre-mRNA 3' process	46
5.3 Protein factors involved in polyadenylation	47
5.4 Type II gene poly(A) site switching	48
5.5 Type III gene poly(A) site switching	53
5.6 Competition between a SS and a PS	56
5.7 Scatter plot of SC5 and PAs vs DIS	58
5.8 Polyadenylation assembly time + RNAP traveling time vs spliceosome assembly time	59
6.1 SAGE library data for CstF-77.L and CstF-77.S	61
6.2 The transcript structure of gene <i>cstf-77</i>	61
6.3 Alternative polyadenylation of CstF-77	64
6.4 Parameter phase plane for the linear DDE	71
6.5 The phase plane of the long and short form transcripts	79
6.6 Relationship between the delay time and the period of Equation (6.3)	81
6.7 Experimental data at five time points	83
6.8 Solutions fitted with optimal parameter set	84
6.9 Distribution of parameters with error less than 0.5	85
6.10 Solutions of Equation (6.3) with long form CIF set to be 0.0652	87
6.11 Solutions of Equation (6.3) with long form CIF set to be 0.0653	87

LIST OF FIGURES
(Continued)

Figure	Page
6.12 Solutions of Equation (6.3) with long form CIF set to be 5	88
6.13 Solutions of Equation (6.3) with short form CIF set to be 0.01	88
6.14 Solutions of Equation (6.3) with short form CIF set to be 5	89
6.15 Solutions of Equation (6.3) with different CIFs	90
6.16 Basin boundary of CIF	91
6.17 Plot of Equation (6.17)	93
6.18 The bifurcation solutions with short form CIF equal to 0.01	94
6.19 The bifurcation solutions with short form CIF equal to 1.00	94
6.20 The bifurcation solutions with short form CIF equal to 100	95
6.21 Solutions with different long and short form CIFs	96
6.22 Solutions with CIF starting near the fixed point	97
6.23 Phase plane of the limit cycle solution	98
6.24 Dynamical behavior of Equation (6.3) with optimal parameters	99
A.1 Discriminant problem	108
A.2 LDA and QDA	109
A.3 Support vector machine algorithm	109
C.1 The transcripts structure of eight significant unigenes	118

LIST OF ABBREVIATIONS

Abbreviations	Full Descriptions
A	Adenine
ADE	Auxiliary Downstream Element
AUE	Auxiliary Upstream Element
C	Cytosine
CC	Correlation Coefficient
CDE	Core Downstream Element
cDNA	complementary DNA
CDS	Coding Sequences
CF	Cleavage Factor
CIF	Constant Initial Function
CPSF	Cleavage and Polyadenylation Specificity Factor
CstF	Cleavage stimulation Factor
CstF-77.L	CstF-77 Long Form Transcript
CstF-77.S	CstF-77 Short Form Transcript
CTD	C-Terminus Domain
CUE	Core Upstream Element
DDE	Differential-Delay Equation
DNA	Deoxyribonucleic Acid
DPA	Downstream Polyadenylation Site
DSE	Downstream Element

Abbreviations	Full Descriptions
EST	Expressed Sequence Tag
FN	False Negative
FP	False Positive
G	Guanine
HMM	Hidden Markov Model
HPR	High Probability Region
IF	Initial Function
LDA	Linear Discriminant Analysis
MC(M)	Markov Chain (Model)
mRNA	messenger Ribonucleic Acid
nt	nucleotides
PAP	Polyadenylation Polymerase
PAs	Polyadenylation Site Score
PAS	Polyadenylation Signal
PCR	Polymerase Chain Reaction
Poly(A) site	Polyadenylation Site
PSSM	Position Specific Score Matrix
QDA	Quadratic Discriminant Analysis
RNA	Ribonucleic Acid
RNAP	RNA Polymerase II
SAGE	Serial Analysis of Gene Expression
SC5	5' Splicing Site Score

Abbreviations	Full Descriptions
SN	Sensitivity
SP	Specificity
SS	Splicing Site
SVM	Support Vector Machine
T	Thymine
TBP	TATA Binding Protein
TF	True Positive
TIS	Transcription Initiation Site
TN	True Negative
TPM	Tag Per Million
UPA	Upstream Polyadenylation Site
USE	Upstream Element
UTR	Untranslated Region

CHAPTER 1

INTRODUCTION

1.1 Background Information

The amount of information in the life sciences is expanding dramatically. Computer technologies and mathematical modeling are changing the way we look at living cells. Since living things are very complex systems, it is very difficult to understand these systems using traditional biological disciplines. Interdisciplinary research is increasingly important to solve such complex systems, and interdisciplinary education has been proposed (Sung et al. 2003). In this work, several specific biological problems have been investigated from various viewpoints, including molecular biology, bioinformatics, and mathematical modeling. There are a huge number of complex processes occurring in human cells every second, such as transcription and translation. Human cells contain millions of proteins, but contain less than 30,000 genes (Stein 2004). Several important factors contribute to the complexity of how different proteins are made from the same gene. Alternative polyadenylation processing is one of them and the detailed mechanism is still unknown.

The messenger RNA (mRNA) polyadenylation is a post-transcriptional process, including two tightly coupled steps (Colgan and Manley 1997): an endonucleolytic cleavage of pre-mRNA followed by the polymerization of an adenosine tail. In this investigation, three problems associated with mRNA polyadenylation in human cells have been explored. One problem is the prediction of

the polyadenylation sites (poly(A) sites) using a support vector machine (SVM), one of the best classification algorithms. A major improvement in the prediction accuracy is achieved compared with polyadq (Tabaska and Zhang 1999), a very commonly used poly(A) site prediction algorithm. Another problem is to use the serial analysis of gene expression (SAGE) (Velculescu et al. 1995) data to discover genes that have more than one mRNA transcript due to alternative polyadenylation, and these transcripts have shown some negative regulation in gene expression. This is a novel study and some of the results may be very useful for biologists and provide some insight of alternative polyadenylation. The third problem is the mathematical modeling of alternative polyadenylation of gene *cstf-77*, whose protein products have been shown to be very important in regulating the polyadenylation of other genes. A mathematical model with a time delay is proposed for explaining the mechanism of alternative polyadenylation. Some preliminary experimental data have been used to validate the model and fit the model well.

1.2 Summary of Results

Predictions of poly(A) sites have been attempted by several groups during the last several years (Salamov and Solovyev 1997; Tabaska and Zhang 1999; Legendre and Gautheret 2003; Hajarnavis et al. 2004). They have used different methods, including linear discriminant analysis, quadratic discriminant analysis, and hidden Markov model. Several program packages have been developed, among which, polyadq (Tabaska and Zhang 1999) is a commonly used tool for poly(A) site prediction. In the present work, a stand-alone program called *polya_svm* (Cheng et al. 2006) has been developed for poly(A) site prediction using the 15 *cis*-elements (Hu et al. 2005) and SVM. For the overall performance, *polya_svm* is 33.7% more sensitive than polyadq

(52.8% vs. 39.5% in sensitivity (SN)). Some improvements in accuracy have been made since the release of the first version.

Serial analysis of gene expression (SAGE) (Velculescu et al. 1995) is a powerful technology to determine gene expression by simultaneous measurement of the frequencies of a large number of sequence tags derived from mRNA/EST transcripts. Unlike microarray gene expression data, SAGE data are useful to detect the different RNA transcripts from the same gene. Recently, more and more genes have been found to generate more than one mRNA transcript and thus they may produce more than one protein. Little is known about the relative expression levels among these transcripts. Not all the transcripts are protein-coding RNAs and some genes may make use of this to control the protein product indirectly. If that is the case, the expression levels of the coding transcripts and non-coding transcripts for the same gene should be opposite or they should be negatively regulated. This may reveal an unknown mechanism of alternative polyadenylation. Here, by using SAGE data, we discovered a number of genes which have two or more transcripts due to alternative polyadenylation with opposite expression levels.

Like alternative initiation and alternative splicing, alternative polyadenylation contributes to the complexity of the overall pool of mRNA transcripts in human cells. However, it is still not clear how the cell utilizes different poly(A) sites in response to different cell conditions and developmental stages. Several protein factors have been identified as being necessary for *in vitro* cleavage and polyadenylation. If the gene of the polyadenylation factor has multiple poly(A) sites, producing multiple protein products with different functions for polyadenylation, then it could control the overall polyadenylation activity by using alternative poly(A) sites. It would also affect its own productivity and thus form a simple feed-back loop.

The dynamics of the expression levels between the two transcripts could be better understood by applying mathematical modeling to study the utilization of different poly(A) sites. Cleavage and stimulation Factor, CstF, is a very important polyadenylation factor composed of three subunits: CstF-50, CstF-64, and CstF-77. Recently, gene *cstf-77* has been found to contain an intronic poly(A) site and thus has more than one different mRNA transcript (Pan et al. 2006). The use of a downstream poly(A) site generates a functional protein factor, which is known to be one subunit of CstF. The protein generated from the intronic poly(A) site is still unknown, and we hope to discover the function by the simulation. In this investigation, we have proposed a novel mathematical model consisting of differential equations with a time delay to study the dynamical behavior of the gene expression levels. Theoretical analysis has been provided along with some numerical simulations of the equations. Some experimental data have verified the numerical results and further discussion has been given. To our knowledge, this is the first time that mathematical modeling has been applied to study alternative polyadenylation and this may provide some explanation for the observed biological data.

CHAPTER 2

BIOLOGICAL BACKGROUND

The phenomenon of heredity is a central feature in the definition of life. All living cells store their hereditary information in the form of double-stranded deoxyribonucleic acid (DNA). To carry out its information-storage function, DNA must express its information to guide the synthesis of other molecules in the cell, such as RNAs and proteins. Humans are multi-cellular organisms, made up of billions of individual cells. Within each human body, every cell contains exactly the same DNA except when mutations occur in some cells. However, each cell expresses its DNA information differently in different tissues, and this accounts for the differences in gene expression.

2.1 Human Genome

The human genome project (HGP) has been regarded as the most important step in science in recent years. Upon the completion of the sequencing of the human genome in 2003, this wealth of data has been the object of many new intensive scientific research areas, such as genomics, data analysis, and mathematical modeling. If the genome can be interpreted correctly, we will then have a much better understanding of human beings and the biological processes occurring inside us.

A genome is an organism's complete set of DNA, which is made from four nucleotide bases: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). These four nucleotides form two base pairs, namely, A always pairs with T and G always pairs with C. The HGP was first articulated in 1988 by a special committee of the U.S. National Academy of Sciences and later adopted by the National Institutes of Health

(NIH) and the Department of Energy (DOE). It took about 15 years to accomplish, during which there are some milestones in mapping the human genome:

- 1995: The *Haemophilus influenzae* genome (1.8 Mb) (Fleischmann et al. 1995). This is the first complete genome of a free-living organism.
- 1996: The *Saccharomyces cerevisiae* genome (12.1 Mb) (Goffeau et al. 1996). This is the first complete eukaryotic genome.
- 1998: The *Caenorhabditis elegans* genome (97 Mb) (1998). This is the first genome for a multi-cellular species.
- 2000: The *D. melanogaster* genome (180Mb) (Myers et al. 2000). This is the first genome sequenced by the whole-genome shotgun method.
- 2001: The draft of human genome (3Gb) (Venter et al. 2001), (Lander et al. 2001). About 94% of the human genome has been sequenced.
- 2003: Completion of the human genome.

The human genome contains about three billion base pairs, which reside in 23 pairs of chromosomes within the nucleus of all human cells. Each chromosome contains hundreds to thousands of genes, which carry the information for making proteins. There are ~20,000-25,000 genes in the human genome (Stein 2004) and each makes three proteins on average.

2.2 Gene Expression

The sequence information of human genome is analogous to a manual of the human body. Then the question remains as to how this information is translated into making the human body? How does a gene express its sequence? The central dogma of molecular biology was first enunciated to answer this question in 1958 and restated in 1970 by Francis Crick (Crick 1970), namely, “The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred directly from protein to either

protein or nucleic acid.” The idea can be represented in Figure 2.1. The process from DNA to RNA is called transcription and from RNA to protein is called translation.



Figure 2.1 Central dogma of biology. Genetic information is transferred from DNA to RNA, then to protein.

In reality, the processes occurring in a cell are much more complex. A lot of factors are involved to control and regulate the transcription and translation. These factors include the environmental conditions, proteins, and some intermediate products. Also, some processes are tightly coupled with each other. These could form a complex feedback network (see Figure 2.2).

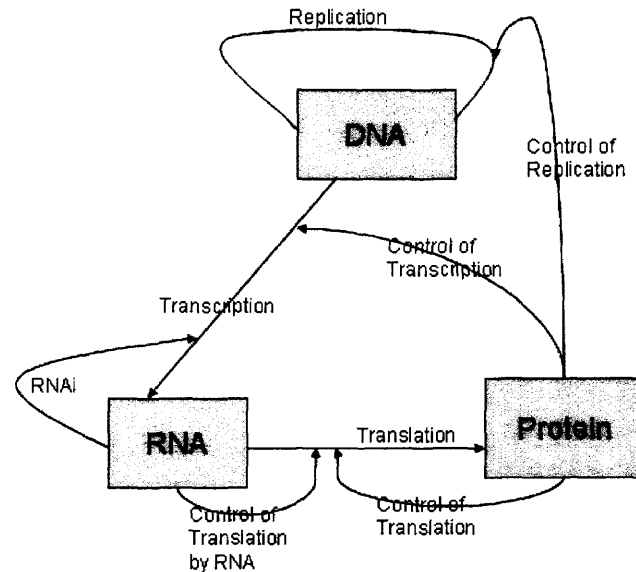


Figure 2.2 Central dogma of biology: feedback network. The proteins generated from RNA would affect the transcription, translation, and DNA replication. The microRNA would also affect transcription by RNA interference (RNAi). The interactions form a feedback network.

When the universal expression of a gene is examined in detail (Orphanides and Reinberg 2002), the gene expression pathway in eukaryotic cells can be written as the following sequential events:

Initiation of transcription → transcription start → capping at 5' end → splicing coupled with transcription → cleavage of pre-mRNA → polymerization of an adenine tail → mature mRNA formation → transportation from nucleus to cytoplasm → initiation of translation → translation start → protein synthesis → protein folding → functional protein → localization of protein to different compartment to participate in cell activity.

mRNA synthesis is catalyzed by RNA polymerase II (RNAP). Transcription is initiated when RNAP binds to a special region, the promoter, at the start of the gene. From this point, RNAP moves along the DNA template, synthesizing RNA until it reaches a terminator sequence (Lewin 2003). Sequences are conventionally written so that transcription processes from left (upstream) to right (downstream) (see Figure 2.3). This corresponds to writing the mRNA in the usual 5' → 3' direction.

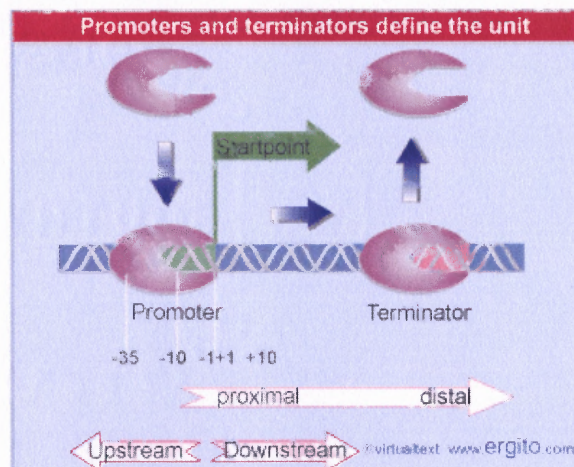


Figure 2.3 RNA synthesis starting at the promoter and ending at the terminator. The promoter is a short DNA sequence with length ~35 nt, which serves as the binding region for RNA polymerase II. The terminator is the region where RNA polymerase II is detached from the DNA template. Figure taken from (Lewin 2003).

In fact, not all genes are coded into proteins. There are two types of genes: one of them is a non-coding RNA gene that represents about 2-5% of the total number of genes and encodes functional RNA molecules, such as tRNA and rRNA; the other is a protein-coding gene that represents the majority of the genes. For the protein-coding genes, not all the nucleotides of the gene are decoded into proteins. Only ~5% of its sequence (exon) is decoded, and the remainder of its sequence (introns) is spliced out during the transcription process (see Figure 2.4). Before the formation of a mature transcript (mRNA) from a protein-coding gene, an oligonucleotide adenine is added after the cleavage of the last exon, the so-called mRNA polyadenylation.

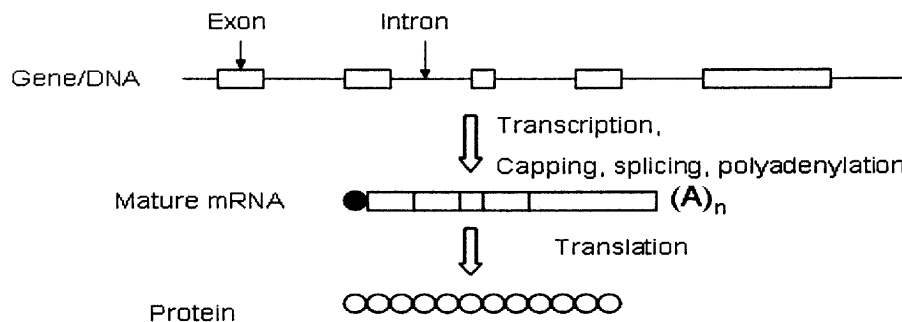


Figure 2.4 Diagram of gene expression. Intron is the region spliced out during transcription, and exon is the region not spliced out during transcription. Capping, splicing, and polyadenylation are coupled with transcription.

2.3 mRNA Polyadenylation and *Cis*-elements

The mRNA polyadenylation is composed of two tightly coupled steps (Colgan and Manley 1997): the first step is an endonucleolytic cleavage that takes place at the site determined by the surrounding RNA sequence and its binding proteins; and the second step involves the polymerization of an adenosine tail (referred to as poly(A) tail) with length ranging over 80-500 nucleotides (nt) at the 3' end of the cleaved RNA. The poly(A) tail is found in nearly all mRNAs in eukaryotes (with the exception of most histone genes), and it is involved in every aspect of mRNA

metabolism, including mRNA stability, mRNA translation (Jacobson and Peltz 1996; Sachs et al. 1997), and mRNA relocalization (Wickens et al. 1997). Short poly(A) tails (20-50 nt) are always correlated with translation repression (de Moor et al. 2005). The kinetics of polyadenylation can have a direct impact on mRNA production. Both enhanced polyadenylation and repressed polyadenylation have been shown to cause human diseases, including Thrombophilia (Gehring et al. 2001) and immunodysregulation (Bennett et al. 2001). The detailed mechanism of polyadenylation will be discussed in a later section.

A *cis*-element is a nucleotide sequence that has regulatory functions in some cellular processes and usually serves as a binding site for some proteins. It is generally accepted that signals are required for triggering polyadenylation near the cleavage site (Keller et al. 1991). A number of auxiliary upstream elements (USEs) or downstream elements (DSEs) have been identified to influence polyadenylation in viral and cellular systems, including SV40 (Carswell and Alwine 1989), HIV (Brown et al. 1991; Valsamakis et al. 1992), and human C2 complement (Moreira et al. 1998). Fifteen candidate *cis*-regulatory elements in humans have been identified (Hu et al. 2005), among which some are conserved in lower species, such as yeast and plants, and some are specific to humans. It is believed that these 15 *cis*-elements would cooperate with each other and then give information about the poly(A) site. This leads to the first part of the work.

CHAPTER 3

PREDICTION OF POLYADENYLATION SITES USING SUPPORT VECTOR MACHINES

Messenger RNA polyadenylation is a post-transcriptional process that adds a poly(A) tail to the cleaved pre-mRNA. Polyadenylation is directly linked to the termination of transcription (Buratowski 2005). Malfunction of polyadenylation has been implicated in several human diseases (Thein et al. 1988; Bennett et al. 2001; Gehring et al. 2001).

Signals around a poly(A) site are required for promoting polyadenylation. The genomic sequence surrounding a poly(A) site is referred to as a poly(A) region. The nucleotide composition of human poly(A) regions is generally U-rich (Legendre and Gautheret 2003; Tian et al. 2005). A hexamer AAUAAA or a close variant, which is located 10-35 nt upstream of most human poly(A) sites, is usually called the polyadenylation signal (PAS) (Tian et al. 2005). U/GU-rich sequences are located within ~40 nt downstream of the poly(A) sites (Zarudnaya et al. 2003; Hu et al. 2005). In addition, a number of auxiliary elements have been identified in viral and cellular systems (Carswell and Alwine 1989; Brown et al. 1991; Valsamakis et al. 1992; Moreira et al. 1998; Hu et al. 2005). Yeast and plant genes utilize a distinct set of *cis*-elements for polyadenylation (Graber et al. 1999; Zhao et al. 1999). While AAUAAA is a prominent hexamer located upstream of poly(A) sites in these species, it occurs to a much lesser extent than in mammalian systems. In addition, UAU and UGUA elements are the efficiency elements located 30-70 nt upstream of yeast poly(A) sites (Graber 2003), which also have been found to be functional elements in human cells (Venkataraman et al. 2005).

Prediction of poly(A) sites has been attempted by several groups during the last several years. An early approach by Salamov and Solovyev (Salamov and Solovyev 1997) used linear discriminant function. A number of variables were used, including position weight matrices for the upstream AAUAAA element and downstream U/GU-rich element, distance between AAUAAA and U/GU-rich elements, and hexamer and triplet compositions in both upstream and downstream regions. Tabaska and Zhang (Tabaska and Zhang 1999) developed polyadq, which employed two quadratic discriminant functions for sequences containing AAUAAA and AUUAAA. The program also uses a position weight matrix for the downstream sequence, a weighted average of hit positions for downstream elements, and downstream dimer preferences. In addition, weight-matrix-only (Legendre and Gautheret 2003) and hidden Markov model (HMM) approaches (Graber et al. 2002; Hajarnavis et al. 2004) also have been employed for poly(A) site prediction.

3.1 Fifteen *Cis*-elements

Fifteen *cis*-elements (see Figure 3.1) in poly(A) region surrounding human poly(A) sites were identified by using a hexamer enrichment method called PROBE (Hu et al. 2005). These *cis*-elements were suggested to play enhancing roles in mRNA polyadenylation because

- They are occurring more often in human poly(A) regions compared with random sequences.
- They are used more frequently for strong poly(A) sites than for weak ones.


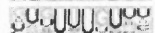




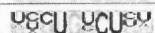





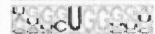


Region	Cis element	Name ^a	Number of Hex ^b	Top 3 hexamers ^c	Percent of Hits ^d
-100/-41		AUE.1	9	GGGGAG, GUGGGG, GGGUGG	48%
		AUE.2	16	UUUGUA, GUAUUU, CUGUGU	93%
		AUE.3	8	UAGUAU, AUAUAU, UUUUAU	51%
		AUE.4	21	UGUAUA, AUGUAU, UGUAAU	82%
-40/-1		CUE.1	17	UAAUUU, UGUUUU, UUUUUU	82%
		CUE.2	23	AAUAAA, AUAUAA, AAUAUA	90%
+1/+40		CDE.1	25	GUGUCU, CUGCCU, UGUCUC	87%
		CDE.2	14	UUUUUU, UUUUUU, UGUUUU	90%
		CDE.3	21	UGUGUG, GUGUGU, CUGUGU	66%
		CDE.4	23	CUGGGG, UGUCUG, GUCUGU	68%
+41/+100		ADE.1	5	CCUCCC, CUCCCC, CACCCC	21%
		ADE.2	6	CCCGCC, CCCCGC, CCCCGG	29%
		ADE.3	15	GGUGGG, GGCUGG, GGGUGG	74%
		ADE.4	6	GGCCAG, GGCCAG, GGGGCC	27%
		ADE.5	18	GGGAGG, GGAGGG, GGGGAG	85%

Figure 3.1 Fifteen *cis*-elements. AUE, auxiliary upstream elements; CUE, core upstream elements; CDE, core downstream elements; ADE, auxiliary downstream elements. Region in the first column indicates the location of the *cis*-elements in the poly(A) region with the poly(A) site set at 0 and the downstream direction set as positive. Figure taken from (Hu et al. 2005).

If the poly(A) site is set to 0, and the downstream direction is referred to as the positive direction, then elements are distributed as following based on their locations: four auxiliary upstream elements (AUEs) in the -100 to -41 nt region; two core upstream elements (CUEs) in the -40 to -1 nt region; four core downstream elements (CDEs) in the +1 to +40 nt region; and five auxiliary downstream elements (ADEs) in the +41 to +100 nt region (Hu et al. 2005).

Naturally, these 15 *cis*-elements can be considered as 15 variables and used in machine-learning tools for poly(A) site prediction. In particular, methods that take into account interactions between the variables are most suitable for predicting poly(A) sites. This is because *cis*-elements are recognized by RNA-binding proteins during mRNA polyadenylation, such as CPSF-160 binding to AAUAAA (Keller et al. 1991), CstF-64 binding to the U-rich and UG-rich elements (Perez Canadillas and Varani

Varani 2003), and hnRNP H family proteins binding to the G-rich element (Arhin et al. 2002). These proteins have been reported to have extensive interactions in the polyadenylation machinery (Proudfoot 2004).

3.2 Datasets and Position Specificity Score Matrix

A comprehensive database of polyadenylation sites, called polyA_DB (Zhang et al. 2005; Lee et al. 2007), has been constructed. The database contains 29,283 human genomic sequences with length 600 nt surrounding the poly(A) sites (-300 to +300 nt), which correspond to 13,942 genes. One example sequence is shown in Figure 3.2 and the sequence always starts from 5' end to 3' end. Poly(A) sites in the polyA_DB database have been identified by aligning the public complementary DNA (cDNA) and expressed sequence tags (ESTs) with genome sequences using a method described in (Tian et al. 2005).

```

>p.2987.1
TCCCTTTCTTCTCACTGGCAGGAAATCAAGAAAGCTCAAAGGACCGGCCCTGAGGCTTGTCTGTCTGTTCTCGGCACCC
CGGGCCCATACAGGACCAGGGCAGCAGCATTGAGCCACCCCTTGGCAGGCGATACGGCAGCTCTGTGCCCTTGGCCAGC
ATGTGGAGTGGAGGAGATGCTGCCCCCTGTGGTTGGAACATCCTGGGGTGACCCCGACCCAGCCTCGCTGGGCTGTCCCC
TGTCCCTATCTCTCACTCTGGACCCAGGGCTGACATCCTAATAAAAATAACTGTTGGATTAGAACTCCATAAATGAGTGG
AATGTGGCCCCAAGGTTGGTGGGGCCCCATCATCCCGATCGGGCCCTGAGCTCCGGCAGCCACCCTAACCACCAGCCCC
AAGGAGGGCCACAAGATGGCCTCTGCCTAGGCATCTGCTGCCTGCCGGCTCGTGGCTGCTGCCCCAGGGCAGCATGATGC
TTGACTGGCAGGCAGGGAAGGTGGCAGGGCTGGTCCCAACCCACCTGGCAGGCTGGCAAGTGGGGAGCAGGAAGCGGCTG
CATGGGCAGCCTGAGGCTGCAGGGGTGGGCCCTGAGTGC

```

Figure 3.2 One example sequence from polyA_DB. The poly(A) site is located in the middle of the sequence, the gene id is indicated after “>p.”, and the poly(A) site is indicated by the number after the last period.

Since the genomic sequences can not be direct input for the classification algorithm, position-specific scoring matrices (PSSM) of 15 *cis*-elements (Hu et al. 2005) were used to search sequences and convert the sequences to numerical values. The PSSM of AUE1 *cis*-element is shown in Figure 3.3. Each score in the matrix reflects how frequently the nucleotide occurs at the given location of the *cis*-element.

Inf means the nucleotide could not be at that location based on the observed sequences.

#AUE1.MAT				
#pos	A	C	G	U
1	-0.73	0.41	0.80	-0.54
2	-0.33	-1.05	-0.98	0.80
3	-2.59	-2.70	2.09	-2.16
4	-2.00	Inf	2.22	Inf
5	Inf	Inf	2.32	Inf
6	Inf	Inf	2.32	Inf
7	0.10	-0.29	0.12	-0.01
8	-0.97	-0.94	1.39	-0.49
9	-0.18	0.14	0.45	-0.29

Figure 3.3 The PSSM of AUE1. The number at the beginning of the row is the location of the nucleotide in the *cis*-element. The scores reflect the relative frequencies of each nucleotide occurring at the given location. Inf means the nucleotide could not be observed at this location based on the sequence data.

Since the 15 *cis*-elements span 200 nt, for a sequence with length 200 nt, the score for every *cis*-element was calculated by the following formula:

$$S = \max \sum_{j=1}^n m_{i,j}$$

where $m_{i,j}$ is the score of nucleotide i at position j in the PSSM, n is the length of the *cis*-element, and the max function is for the sequences with length n in the *cis*-element region. For example, for AUE1, the score is calculated in -100 to -41 nt region. If there exists an Inf value, S is set to be -15, as it is the lowest value observed in the training data, which will be discussed in a later section. That is, for a sequence with length equal to 200 nt, it is converted to a vector with 15 values corresponding to 15 *cis*-elements.

3.3 Correlation Between Fifteen *Cis*-elements

To understand the relationship among *cis*-elements, we applied a hierarchical clustering method to group the 15 *cis*-elements based on their occurrence. For all the sequences in the polyA_DB, scores of the 15 *cis*-elements were calculated based on

the middle 200 nt of the sequences and they formed a matrix with dimensions 29,283×15. Using Pearson correlation as the metric and average linkage for tree building (Eisen et al. 1998), the 15 *cis*-elements can be largely divided into two groups (see Figure 3.4). One group consisted of CUE1, CUE2, CDE2, AUE2, AUE3, and AUE4, which are all upstream elements except CDE2, and the other consisted of downstream elements except AUE1.

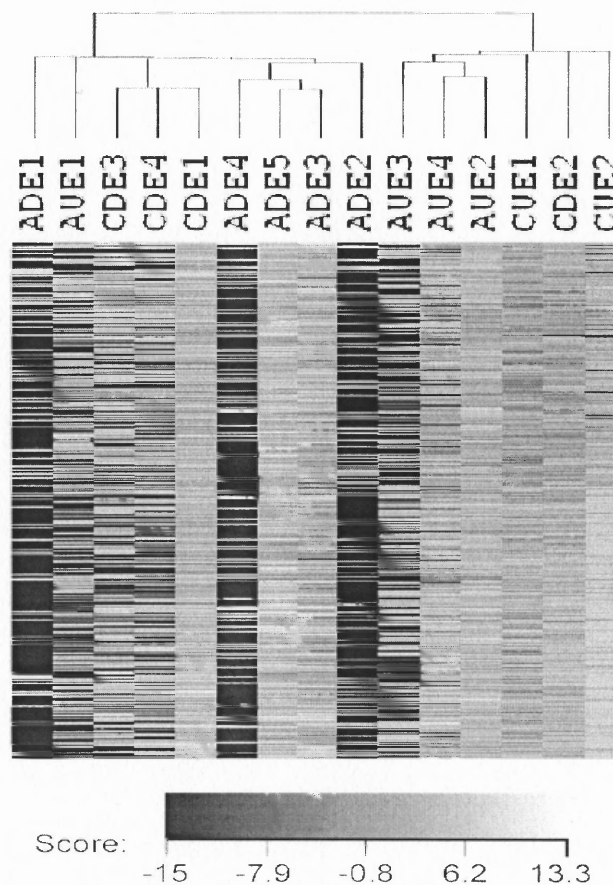


Figure 3.4 Clustering of 15 *cis*-elements using poly(A) sites from polyA_DB. PSSMs of 15 *cis*-elements were used to search human poly(A) sites from polyA_DB in their respective regions. Each row is a poly(A) site (29,283 in total), and each column is a *cis*-element. Hierarchical clustering using Pearson correlation was employed to cluster *cis*-elements, and the resulting tree is shown on top of the heatmap. The gray scale is indicated at the bottom of the plot.

This grouping is robust, as clustering using other parameters, such as Kendall's Tau correlation and complete linkage also resulted in similar groupings.

Thus, upstream elements and downstream elements in general have different profiles, indicating that they may compensate for each other during mRNA polyadenylation. In biochemical terms, this result suggests that weak upstream signals can be “helped” by strong downstream signals, and vice versa. However, further experiments are needed to confirm this model.

3.4 Training, Testing, and Prediction

For classification, there are two steps. First, train the known datasets and generate the trained model. Second, use the trained model to predict the testing dataset and analyze the accuracy. The training dataset used in this study consisted of 2,000 sequences, with 1,000 positive sequences randomly selected from polyA_DB and 1,000 negative sequence generated by first-order Markov chain model (MCM) (see Appendix A.1) from the positive sequences. The testing set is generated the same way with the training set as the number of sequences varying for different purposes. All the scores of training data and testing data were scaled by $(S-s)/\sigma S$, where s is the mean value of S for all positive and negative sequences in the training data and σS is the standard deviation of S for all positive and negative sequences in the training data.

For training, only the middle 200 nt were used because we already know that the poly(A) site is located at the middle of the positive sequence. The true negative sequence is hard to get and it is assumed that random sequences contain very few poly(A) sites. Therefore, the MCM was used to generate the negative sequences of 200 nt in length. By using PSSMs, every sequence of 200 nt generates a vector with 15 values. Therefore, two groups of data with 15 dimensions were generated. The data from positive sequence are labeled with 1, and the data from negative sequence are

labeled with -1. These were the input of the classification algorithms. After training, the model was generated and ready to accept testing data for prediction.

For testing a sequence from the polyA_DB database, we pretended not to know the location of the poly(A) site and used the trained model to predict if the sequence contains a poly(A) site or not. The sequence length is 600 nt, and every 200 nt subsequence was used to generate the values. Multiple-hit issue arises and is discussed in the later section.

3.4.1 Comparison of Different Methods

There are many classification methods. To address which one would be used for our prediction, a one-hit model was generated to test for simplicity. That is, for testing, we only use the middle 200 nt to generate the vector scores. So for each sequence, we only need to predict once, and the result is only to predict if the middle site of the sequence is a poly(A) site or not.

In light of this, we tested three discriminant analysis methods, namely linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) (see Appendix A.2), and support vector machine (SVM) (see Appendix A.3). LDA finds a hyper-plane to separate two or more classes with linear combination of variables (Zhang 2000). It assumes that the data distribution for each class is normal and that all classes have the same covariance. QDA uses a quadratic surface to separate classes (Zhang 2000), and also makes the assumption of a normal distribution, but relaxes the requirement of covariance. SVM employs kernel functions to separate data by a hyper-plane that is supported by vectors lying at the boundaries of classes (Cortes and Vapnik 1995). All these methods have been used on biological sequences for identification of signals such as splice sites (Zhang 2000; Zhang et al. 2003; Yeo et al. 2004).

To compare these methods, we randomly selected 2,000 poly(A) sites from the polyA_DB database, retrieved the $-100/+100$ nt genomic region surrounding each poly(A) site, and used them as a positive dataset. We then randomized the positive dataset by a first-order MCM to obtain 2,000 negative sequences, each with 200 nt in length. Using LDA, QDA, and SVM functions in program R (see Appendix B.2), we compared their performance for prediction of poly(A) sites with respect to sensitivity (SN), specificity (SP), and correlation coefficient (CC) (see Appendix A.4). As summarized in Table 3.1, of these three methods, QDA achieved the best sensitivity, and SVM achieved the best specificity based on 100 random tests. P-values are calculated based on two-tailed t-tests comparing 100 values from LDA or QDA with those from SVM. Overall, SVM has the best performance judged by CC. Thus, we have selected SVM as the prediction method for further studies.

Table 3.1 Comparison of LDA, QDA, and SVM

Method	SN		SP		CC	
	Mean	P-value	Mean	P-value	Mean	P-value
LDA	0.830	3.0E-12	0.785	1.2E-99	0.603	1.6E-87
QDA	0.863	1.3E-24	0.784	1.1E-102	0.628	3.8E-66
SVM	0.843	-	0.848	-	0.693	-

LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; SVM, support vector machine; SN, sensitivity; SP, specificity; CC, accuracy. The p-value is calculated based on t-test comparing 100 values from LDA/QDA with those from SVM.

3.4.2 Leave-One-Out Method

To examine how each element contributes to the prediction and whether or not we can reduce the number of variables, we conducted a leave-one-out experiment, where we left out one element at a time and calculated its effect on SN, SP, and CC in poly(A) site prediction. We reasoned that omission of important elements would significantly

lower the performance of prediction, whereas omission of non-essential ones would not make much difference.

As shown in Table 3.2, we found that CUE2 is the most important one, as omission of CUE2 led to substantial drop of both sensitivity and specificity, which is consistent with the notion that the AAUAAA element is critically important for mRNA polyadenylation.

Table 3.2 Effects of Leaving Out Some *Cis*-elements

Element(s) left out	SN		SP		CC	
	Mean (%)	P-value	Mean (%)	P-value	Mean	P-value
None	84.3	-	84.8	-	0.693	-
AUE1*	84.2	0.33	84.8	0.21	0.691	0.22
AUE2*	84.2	0.35	84.7	0.11	0.692	0.29
AUE3	84.0	0.03	84.7	0.16	0.689	0.04
AUE4*	84.0	0.07	84.9	0.09	0.693	0.44
CUE1	83.9	0.01	84.6	0.03	0.687	8.5E-3
CUE2	65.3	9.9E-161	71.5	2.7E-153	0.395	2.8E-175
CDE1*	84.1	0.14	84.8	0.25	0.691	0.16
CDE2	82.6	9.1E-17	83.8	3.2E-10	0.668	1.2E-19
CDE3*	84.1	0.08	85.0	0.07	0.693	0.44
CDE4*	84.2	0.31	84.8	0.20	0.692	0.35
ADE1*	84.2	0.22	84.7	0.18	0.692	0.26
ADE2*	84.0	0.07	84.8	0.24	0.690	0.10
ADE3	83.7	4.5E-4	84.8	0.22	0.688	0.02
ADE4*	84.3	0.45	84.7	0.17	0.692	0.28
ADE5	83.9	6.6E-3	84.6	0.06	0.687	0.01
9 elements	82.1	1.1E-24	84.5	0.02	0.671	6.3E-17

Element(s) were left out at both the training and testing stages. None, no elements deleted; 9 elements, leaving out 9 elements marked with asterisk in the table; AUE, auxiliary upstream element; CUE, core upstream element; CDE, core downstream element; ADE, auxiliary downstream element. P-values are calculated based on 30 tests comparing with none. Elements marked with an asterisk indicate that their individual omission would not lead to significant change (p-value > 0.05) of SN, SP, or CC.

Omissions of CUE1 or CDE2 had similar effects, albeit to a lesser extent, indicating that U-rich elements surrounding the poly(A) sites are important determinants. In addition, omissions of AUE3, ADE3, or ADE5 caused drops in

sensitivity, indicating their important roles in poly(A) site selection. For the rest of the 9 elements (indicated by an asterisk in Table 3.2), leaving out any single element caused some decrease in prediction performance, while none of them appeared to be significant based on a t-test ($p\text{-value} > 0.05$, Table 3.2). However, leaving out all 9 elements made both sensitivity and specificity drop significantly. Thus, these 9 elements may contribute to poly(A) site recognition coordinately and some elements may be important for only a small subset of poly(A) sites.

Taken together, these data indicate that the 15 *cis*-elements are necessary for poly(A) site prediction, validating the functional importance of these elements for polyadenylation. In addition, the fact that multiple variables are required for poly(A) site prediction suggests a combinatorial mechanism for poly(A) site recognition in human cells.

3.4.3 PolyA_svm

For sequence with length greater than 200 nt, the one-hit model is not enough since we do not know where the poly(A) site is. So the prediction has to be carried out at each possible position of a given sequence, and a multiple testing issue arises when predicting the likelihood of a sequence containing a poly(A) site. That is, for each location, we could predict if that location is a poly(A) site or not. However, the problem arises that, even if the prediction is good, for a long negative sequence, we could predict many poly(A) sites inside the sequence. For example, if the true negative rate is 95%, for a 300 nt long sequence with 100 sites, it is possible to have 5 locations to be predicted as poly(A) sites. Some other methods need to be introduced to eliminate this kind of noise. It has been found that polyadenylation cleavage is usually heterogeneous and occurs in a window rather than at a defined position (Tian et al. 2005). A simple test was done based on 100 sequences, 50 positive sequences

randomly selected from polyA_DB and 50 negative sequences generated from the first-order MCM. The result is shown in the Figure 3.5.

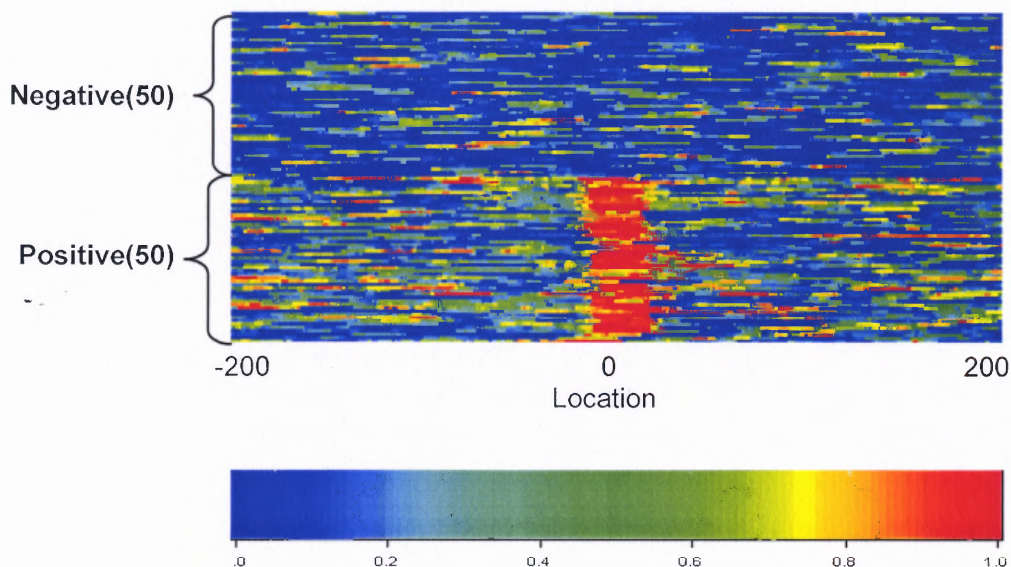


Figure 3.5 Heatmap plot of predicted probability results of 100 test sequences. The poly(A) site is set to be 0, and the downstream direction is referred to as the positive direction. The positive sequences containing poly(A) sites are randomly chosen from polyA_DB, and negative sequences not containing poly(A) sites are generated by a first order MCM with length 600 nt. The scale is indicated at the bottom of the plot.

Based on this, we designed a window-based scoring scheme to address the multiple-hit issue: we required a region M of m nt to have probability greater than 0.5 at each position in the region and another region N of n nt within M to have high probability values. The region M is called a positive region, and the region N is called a high probability region (HPR) (see Figure 3.6). Using different combinations of m and n, we found that m=30 and n=10 achieved the best performance when the product of the 10 probabilities for HPR was set to be greater than 0.5. Thus, on average, the probability of being positive for each position is greater than 0.933 in HPR.

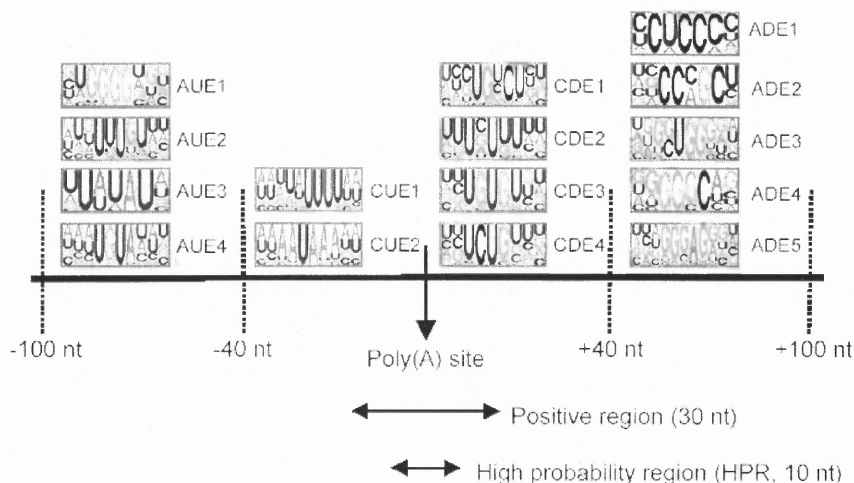


Figure 3.6 Schematic of 15 *cis*-elements in the poly(A) region and the search algorithm for `polya_svm`. A poly(A) site is indicated by a vertical arrow. The predicted probability in the positive region is greater than 0.5, and the product of the predicted probabilities in the high probability region is greater than 0.5.

Motivated by our initial results, we developed a stand-alone program called `polya_svm` using the PERL language (see Appendix B.1). This program uses the 15 *cis*-elements and SVM to predict the poly(A) sites. For SVM predictions, the SVM library LIBSVM (see Appendix A.3) was used with C-support vector classification (C-SVC) method and the radial basis kernel function (RBF). The default settings, i.e. $C=1$ and $\gamma=1/15$, were applied. In fact, we tried different parameter settings, and we found there was no big difference. For calculating probabilities, we applied a window-based adjustment method using probability values generated by LIBSVM: the probability of having a poly(A) site at position i is determined by its E-value, calculated by $E_i = 1 - \prod_j \Pr(j)$, where j is a position relative to i , $\Pr(j)$ is the probability for position j , and $j \in \{-4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$. Thus, the E-value is based on a 10 nt HPR centered at i , and the higher the probability, the lower the E-value. In addition, we also required that the region $(-15/+15)$ adjacent to a positive site to have $\Pr(i) > 0.5$ at every position of its sequence.

E-values of 1,000 positive and 1,000 negative sequences are shown in a heatmap (see Figure 3.7) according to the gray scale shown at the bottom. Each sequence is 600 nt in length. All positive sequences have a poly(A) site in the middle, and some may have multiple poly(A) sites. Poly(A) site prediction was carried out in the 101 to 500 nt region. Thus, each row contains 400 E-values. The x-axis is the position in the sequence. As shown in Figure 3.7, using this method, `polya_svm` can effectively eliminate false positive sites and accurately locate real poly(A) sites.

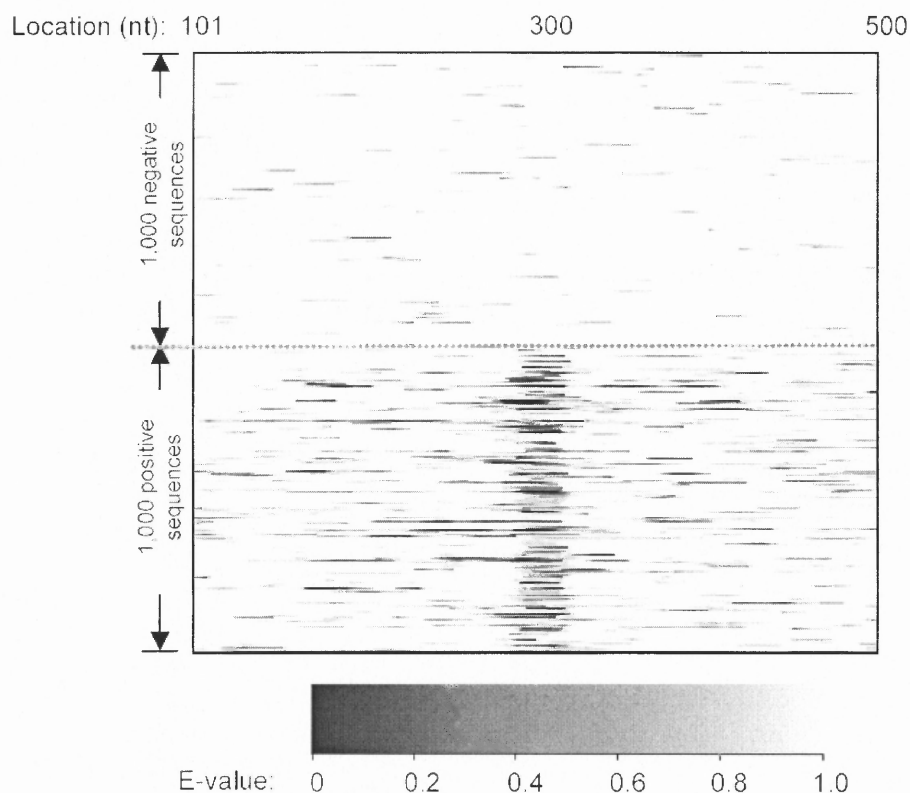


Figure 3.7 Prediction of positive and negative sequences. E-values of 1000 positive and 1000 negative sequences are shown respectively. The gray scale is shown at the bottom of the figure. Each sequence is 600 nt in length, and the poly(A) site is set at location 300. Poly(A) site prediction was carried out for the 101 to 500 nt region.

3.5 Prediction of Poly(A) Sites in the Human Genome

We tested `polya_svm` using all human poly(A) regions in the `polyA_DB` (29,283 in total) and compared its performance with `polyadq`, a commonly used tool for poly(A) site prediction (Tabaska and Zhang 1999). For `polya_svm`, if the predicted location (middle of HPR) is within 24 nt from a real poly(A) site, the prediction was considered as true positive (TP), and otherwise as false negative (FN). For `polyadq`, since it uses the PAS location for poly(A) site prediction, we considered a prediction to be TP if a PAS is within 48 nt upstream of a real poly(A) site. As shown in Table 3.3, `polya_svm` is 33.8% more sensitive than `polyadq` (52.8% SN vs. 39.5% SN).

We then divided poly(A) sites into different groups based on two criteria: their usage and location. For poly(A) site usage, we used the number of supporting cDNA/ESTs for a poly(A) site to determine its frequency of usage. Poly(A) sites in genes with only one poly(A) site are called constitutive sites. Poly(A) sites in genes with multiple sites were grouped into strong, weak, and medium. A strong site is used more than 75% of the time based on supporting cDNA/ESTs. If a gene has a strong site, other sites are called weak sites. If a gene does not have a strong site, all sites are called medium sites. For poly(A) site location, we first separated poly(A) sites located in introns and internal exons (called upstream sites) from those in 3'-most exons, and then divided poly(A) sites in the 3'-most exons into three groups depending upon their location (see Table 3.3). The 5'-most site is called the first site, the 3'-most site is called the last site, and sites in between the 5'-most and the 3'-most sites are called middle sites. In addition, if a 3'-most exon contains only one poly(A) site, the site is called a single site. `Polya_svm` was over 50% more sensitive than `polyadq` for detecting medium and weak poly(A) sites, and about 19.5% and 7.2% more sensitive than `polyadq` for strong and constitutive poly(A) sites, respectively (see Table 3.3).

Table 3.3 Comparison of PolyA_svm with Polyadq for Different Types of Positive Poly(A) Sites

Poly(A) site type		polya_svm			polyadq			SN Diff (%)
		TP	FN	SN (%)	TP	FN	SN (%)	
Total		15,469	13,814	52.8	11,563	17,720	39.5	+33.8
poly(A) site usage	Strong	1,602	655	71.0	1,341	916	59.4	+19.5
	Medium	8,301	8,221	50.2	5,488	11,034	33.2	+51.3
	Weak	1,521	2,565	37.2	961	3,125	23.5	+58.3
	Constitutive	4,045	2,373	63.0	3,773	2,645	58.8	+7.2
poly(A) site location	Single in 3'-most exons	4,941	2,853	63.4	4,516	3,278	57.9	+9.4
	First in 3'-most exons	2,469	3,583	40.8	1,522	4,530	25.2	+62.2
	Middle in 3'-most exons	2,256	2,479	46.7	1,212	3,523	25.6	+86.1
	Last in 3'-most exons	3,763	2,289	62.2	2,839	3,213	46.9	+32.6
	Intron and internal exons	2,040	2,610	43.9	1,474	3,176	31.7	+38.4

For polya_svm, a sequence is predicted to be true positive (TP) if a predicted poly(A) site (the middle position of HPR) is within 24 nt from a real poly(A) site. For polyadq, a sequence is considered TP, if the sequence is predicted to be positive and the real poly(A) site is within 48 nt downstream of a poly(A) signal AAUAAA or AUUAAA. See text for description of different types of poly(A) sites.

As for poly(A) sites located in different regions of a gene, polya_svm also made more sensitive predictions than polyadq in all categories, particularly for poly(A) sites located in the upstream of the 3'-most poly(A) sites (62.2% and 86.1% more sensitive for the first and middle poly(A) sites in 3'-most exons). A more detailed analysis revealed that the high sensitivity of polya_svm is mainly ascribed to its

capability to predict poly(A) sites without AAUAAA or AUUAAA, and sites with weak downstream signals. Taken together, these data demonstrate that `polya_svm` is highly sensitive for poly(A) site prediction.

`Polya_svm`'s performance for sequences without poly(A) sites, or negative sequences was also examined. A true negative sequence is difficult to obtain, as there is no extensive experimental evidence for negative sequences. Thus, we tested several types of sequences that were presumed to have very few poly(A) sites, including randomized poly(A) regions (-300/+300 nt), randomized genome sequences, mRNA coding sequences (CDS), and 5'untranslated regions (UTRs).

As shown in Table 3.4, comparable false positives (FP) were predicted by `polya_svm` and `polyadq` for randomized sequences. However, `polya_svm` predicts significantly more sites than `polyadq` in CDS and 5'UTR sequences (more than 2 fold). Interestingly, the difference was not significant when randomized CDS and 5'UTR sequences were used, suggesting that some of the false positives in CDS and 5'UTRs predicted by `polya_svm` may actually be true positives. Accordingly, it is tempting to speculate that there may, in fact, exist a large number of poly(A) sites in CDS and 5'UTRs. This would be consistent with previous findings of poly(A) sites in internal exons (Tian et al. 2005; Yan and Marr 2005). However, this hypothesis has yet to be tested by wet lab experiments. On the other hand, a highly sensitive method would aid in the identification of these sites, which are difficult to detect by cDNA/EST-based approaches, as many of these poly(A) sites would result in aberrant transcripts that may be rapidly degraded by cellular surveillance mechanisms, such as those without an in-frame stop codon (Frischmeyer et al. 2002).

Table 3.4 Comparison of PolyA_svm with Polyadq for Different Types of Negative Poly(A) Sites

Negative Set	polya_svm				Polyadq				SP Diff (%)	CC Diff (%)
	TN	FP	SP (%)	CC	TN	FP	SP (%)	CC		
Poly(A) region first-order MC	420	80	76.7	0.387	426	74	72.8	0.279	+5.1	+27.9
Genome first-order MC	446	54	83.0	0.451	447	53	78.9	0.334	+4.9	+25.9
CDS	410	90	74.6	0.364	458	42	82.5	0.365	-10.6	-0.3
CDS first-order MC	487	13	95.3	0.561	485	15	93.0	0.447	+2.4	+20.3
5' UTR	445	55	82.8	0.448	482	18	91.7	0.437	-10.7	+2.5
5' UTR first-order MC	491	9	96.7	0.572	492	8	96.1	0.470	+0.6	+17.8

MC, Markov chain. Poly(A) region first-order MC, randomized -300 to +300 nt sequences surrounding poly(A) sites; genome first-order MC, randomized human chromosome 1 sequence; CDS, coding region sequences of human RefSeq sequences; CDS first-order MC, randomized CDS; 5' UTR, 5' untranslated region of human RefSeq sequences; 5' UTR first-order MC, randomized 5' UTRs. For each negative set, 500 sequences were generated and predicted by polya_svm and polyadq. The process was repeated 10 times and the mean values are presented in the table. TP and FN of Table 3.3 for all poly(A) sites in the polyA_DB database were scaled and used to calculate CC.

3.6 Further Improvements of PolyA_svm

We have made several changes to the polya_svm program and improved the accuracy of prediction by 12% over its first release polya_svm 1.0 (Cheng et al. 2006). Infinite values are generated when a *cis*-element does not exist in a given sequence. However, for the practical use of SVM, each variable requires a numerical value. In polya_svm 1.0, infinite values were replaced by the lowest score of all *cis*-elements in the training data, i.e., -15. To examine if this aspect can be improved, we have tried three other options (see Table 3.5). Method 1 is the original method, where all infinite values were set to -15. In Method 2, the infinite values were set to zero. In Method 3,

the infinite values were set to be the lowest value that a given *cis*-element can achieve. In Method 4, the infinite values were set to be a value of -6.28, which is the lowest value for all *cis*-elements. Prediction was performed 30 times for each of the four different settings and the relative CC results are shown in the Table 3.5. All values were normalized to Method 1, the original method. As shown in Table 3.5, *polya_svm* had the best performance when the Method 4 was used, which is 8% gain in accuracy over the original method. This improvement indicates that the hyper-plane in classification is not defined well by the support vectors in the areas with low negative values.

Table 3.5 Comparison of Different Methods for Replacing Infinite Values

Type	Relative SN	Relative SP	Relative CC
Method 1	1.00	1	1.00
Method 2	1.04	0.96	0.95
Method 3	1.02	1.00	1.03
Method 4	1.08	1.01	1.08

Method 1 is the original method, where all infinite values were set to -15. Method 2 sets infinite values to 0. Method 3 sets infinite values to the lowest value a given *cis*- element can achieve. Method 4 sets infinite values to an arbitrary value of -6.28, which is the lowest value for all *cis*-elements. Prediction was performed 30 times for each of the four different settings and the relative CC results are shown in the table. All values are normalized to Method 1, the original method.

We have used two regions for poly(A) site prediction to ensure high specificity, i.e., a 30 nt positive region and a 10 nt HPR (Cheng et al. 2006). To examine whether this step can be simplified without losing performance, we have tested methods only using HPR with different sizes (1 to 99 nt). In addition, we used the formula $s = -\log_2(\prod_{i=1}^n p_i)$ to derive a score for an HPR, where p_i is the probability value for position i , which is provided by LIBSVM, and n is the size of the HPR. We also selected the optimal *polya_svm* cutoff score based on the highest CC value for each HPR setting. As shown in Figure 3.8, we found that HPR of 32 nt and cutoff

score of 6 appear to be the optimal for achieving highest CC value. Thus we used HPR of 32 nt and cutoff score of 6 to determine poly(A) sites in the new program (version 2.0). In addition, when several HRP regions overlapped, we selected the HPR with the best score to represent their combined region.

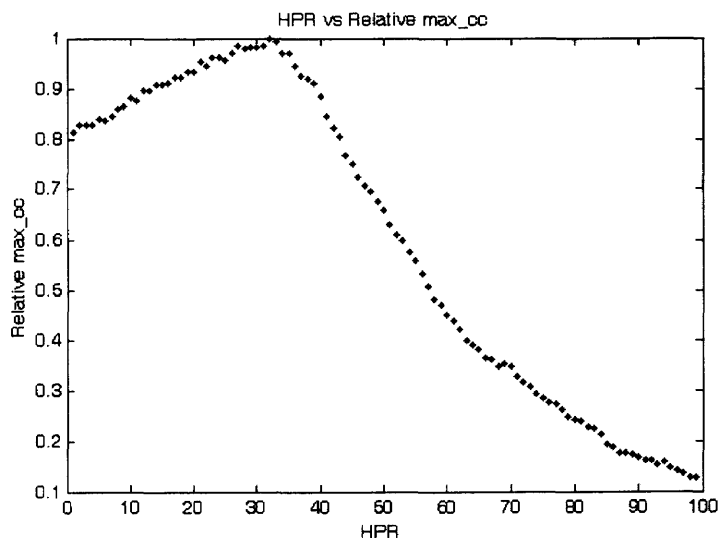


Figure 3.8 HPR vs relative maximum accuracy. The HPR varies from 1 to 99 nt and the cutoff values are [0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.20, 0.25..., 0.90, 0.95, 0.96, 0.97, 0.98, 0.99, 0.991, 0.992 ... 0.999]. For any given HPR value, the best CC can be calculated for a given cutoff value. The x-axis is HPR and the y-axis is the normalized CC for different HPR using the optimized HPR size (32).

To examine if there is correlation between `polya_svm` prediction score and the strength of a poly(A) site, we divided 29,283 poly(A) sites obtained from `polyA_DB` database into several groups based on the usage of a poly(A) site: strong site (S), constitutive site (O), medium site (M), and weak site (W) as described in section 3.5. The usage is based on the number of supporting ESTs from non-normalized cDNA libraries as described in (Hu et al. 2005). As shown in Figure 3.9, a correlation is discernable for the `polya_svm` prediction score and poly(A) site usage, indicating that the score can be used as a guide to determine the strength of a poly(A) site. We

applied the t-test on S vs M and W, and the p-values are less than $1e-16$. That is, the stronger the poly(A) site, the smaller the polya_svm score.

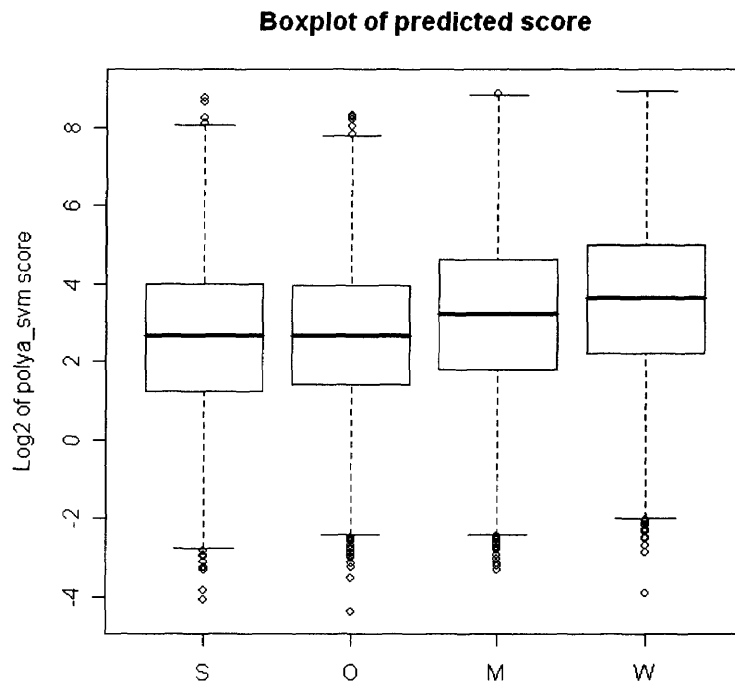


Figure 3.9 Boxplot of polya_svm scores for different types of poly(A) sites. S, the strong poly(A) site; O, the constitutive poly(A) site; M, the median poly(A) site; W, the weak poly(A) site. The y-axis is the log₂ value of the predicted polya_svm scores.

CHAPTER 4

SAGE DATA ANALYSIS

4.1 Introduction to SAGE

Serial Analysis of Gene Expression (SAGE) (Velculescu et al. 1995) is a high-throughput and high-efficient method to simultaneously detect and measure the expression levels of genes expressed in a cell at a given time. Compared with microarray experiment, the SAGE is expensive to perform and it requires higher techniques. The brief steps for SAGE are shown in Figure 4.1 and the detailed protocol can be found from the website: <http://www.sagenet.org/protocol/index.htm> (visited on 07/02/2006).

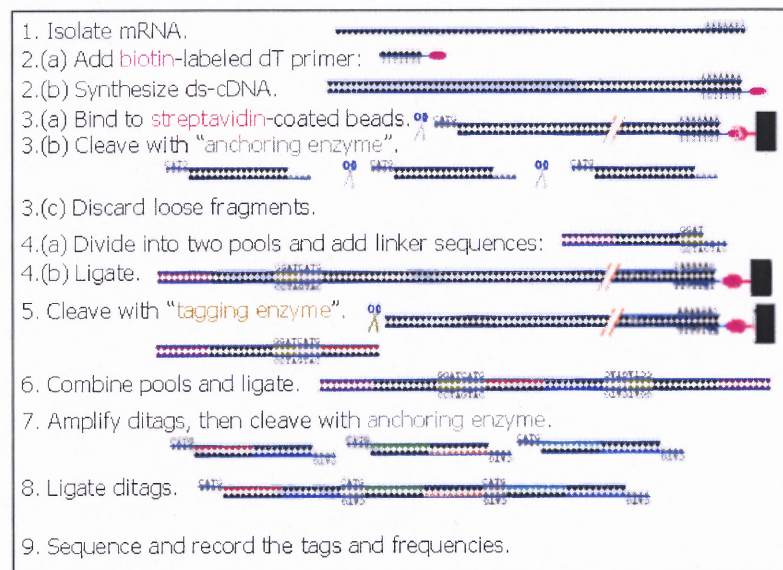


Figure 4.1 An outline of SAGE. SAGE sequences and counts a short tag with 10-17 nt located at the 3' most of the mRNA after the tag CATG using two restriction enzymes. By comparing the number of the tags, SAGE can estimate the relative expression levels of different transcripts in the cell. Figure taken from PowerPoint presentation by Kevin R. Coombes, Section of Bioinformatics, Department of Biostatistics, MD Anderson Cancer Center, University of Texas, 2006.

The original SAGE approach (Velculescu et al. 1995) generated a 10 base pair (bp) tag (referred to as a short SAGE tag) derived from the 3'-end of the transcript by using the tagging enzyme NlaIII. Later on, using a new tagging enzyme MmeI (Saha et al. 2002), a 17 bp tag (referred to as a long SAGE tag) was generated and enhanced the specificity of the mapping from the SAGE tag to the transcriptome, the collection of all RNA transcripts. Due to the earlier introduction of short SAGE tag, the public SAGE database has more short SAGE libraries than long SAGE libraries for the human dataset. Most of the tags are uniquely-mapped to some gene. But, there still exists a small portion of tags having more than one gene being mapped to them. For one tag, it is necessary to find the best gene mapped to this tag based on the evidence of EST/cDNA. Fortunately, this kind of work has been done by SAGE Genie (Boon et al. 2002) and the database is updated regularly. Nonetheless, errors occur in the experimental data due to a lot of factors, such as sequencing and PCR.

Positive and negative regulation of different genes and proteins are very common phenomena in biological systems such as activators or inhibitors and have been studied widely (Ma et al. 1998; Koretzky and Myung 2001; Ku et al. 2005). However, the regulation of alternative transcripts from the same gene has seldom been investigated.

With the progress in genome annotation and emergence of large amount of information with regard to cDNA/ESTs, studying the regulation of different transcripts from the same gene becomes possible and may identify some unknown mechanisms for gene expression. Difficulties still lie in the accurate measurement of expression levels of the transcripts. Large-scale expression tools such as microarray, have been used to analyze the co-expression levels of known complementary transcripts (Nikaido et al. 2003). However, it is not detectable for those mRNA

transcripts that are not imprinted in the chips. Serial Analysis of Gene Expression (Velculescu et al. 1995) is an alternative technology to systematically detect the global gene expression levels. There are a lot of advantages in using SAGE (Tuteja and Tuteja 2004) and most importantly, SAGE can detect the expression levels of different transcripts without prior knowledge of the genes, while microarray can only detect the imprinted transcript expression levels.

4.2 Analysis Pipeline

Motivated by the observation from the SAGE data that gene *cstf-77* has two transcripts with opposite expression levels (Pan et al. 2006), we designed a pipeline to systematically discover the genes with similar expression patterns. The data flow for such discoveries can be represented by Figure 4.2, and they are discussed in detail in the following sections.

4.2.1 Reliable Libraries Selection

The SAGE data were downloaded from the SAGE Genie (Boon et al. 2002) FTP server (see Appendix C.1). Due to the noise and experimental variation of different libraries, not all the libraries could be used in this study. Each library contains the following information: the sequence tags, the frequency for each tag, the total number of tags, the number of unique tags, the tissue type for the library, etc. An example of the SAGE data is shown in Table 4.1.

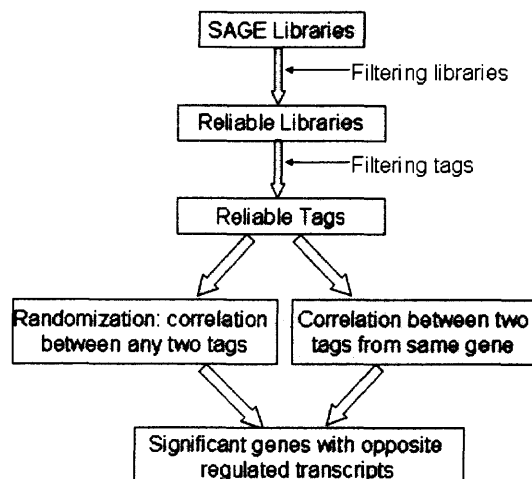


Figure 4.2 The pipeline of SAGE data analysis. The noise exists in the libraries and the mapping from tags to the unigenes. By filtering the unreliable libraries and tags, the significant genes can be selected from the observed correlation between two tags from the same gene. The significance is based on the randomized correlation between any two tags, not necessary from the same gene.

Table 4.1 SAGE Data Example

tag	lib_id	freq
AAAAAAAAAAAAAAAAAAAA	643	4
AAAAAAAAAAAAAAAAAAAC	643	2
AAAAAAAAAAAAAAAAAAAG	643	2
AAAAAAAAAAAAAAAAAAAT	643	1
AAAAAAAAAAAAAAAAAACAG	643	1
AAAAAAAAAGAAAAAAG	643	1
AAAAAAGCAAATTTCA	643	1
AAAAAAGCTTTTGAGA	643	1
AAAAAATCTAGCTCTT	643	1
AAAAAATGCATATGAA	643	1
...

The third column indicates the number of tags from the first column found in the libraries indicated in the second column. For any given library, the number of unique tags for this library is the number of rows corresponding to this library and the total number of tags is the summation of the third column corresponding to this library.

A long or short library is considered to be reliable if the following two conditions are satisfied:

- The total number of tags is greater than 50,000, of which about 5% of the tags are missing (Becquet et al. 2002).

- The ratio (\log_2 of the total number of tags divided by \log_2 of the number of unique tags) is located in the interval $(\text{mean}-2\times\text{sd}, \text{mean}+2\times\text{sd})$, which corresponds to a 95.5% confidence interval, where the mean is the average ratio of all long SAGE libraries or all short SAGE libraries, and sd is the standard deviation of the ratio of all long SAGE libraries or all short SAGE libraries.

We can see that these two conditions are reasonable from the plot of long and short libraries (see Figure 4.3). In Figure 4.3, the vertical line represent the cutoff line of first condition and the three skew lines from top to bottom represents $(\text{mean}+2\times\text{sd})$ line, the mean value line, and $(\text{mean}-2\times\text{sd})$ line. The slope of the lines is the mean value of the ratios. The second condition was made based on the assumption that the average expression level for the whole library should not be too different between the libraries.

The total dataset contains 289 human short SAGE tag libraries and 59 long SAGE tag libraries. After applying these two conditions to the long and short libraries respectively, 201 short libraries and 49 long libraries were selected. The selected libraries are located in the region enclosed by the vertical cutoff line, $(\text{mean}-2\times\text{sd})$ line, and $(\text{mean}+2\times\text{sd})$ line (see Figure 4.3). The long libraries should be more reliable than the short ones and it was confirmed by the relation $(49/59 > 201/289)$.

4.2.2 Is SAGE Data Tissue-Specific?

To investigate if the SAGE data have some bias for different tissues, we plotted the ratio (\log_2 of the total number of tags divided by \log_2 of the number of unique tags) versus the tissue order from reliable short and long SAGE libraries of human dataset. From the Figure 4.4, the average gene expression level for stem cells is higher than for other tissues in the figure, while the gene expression level for the cerebellum is relatively lower. The biological explanation is unclear to us.

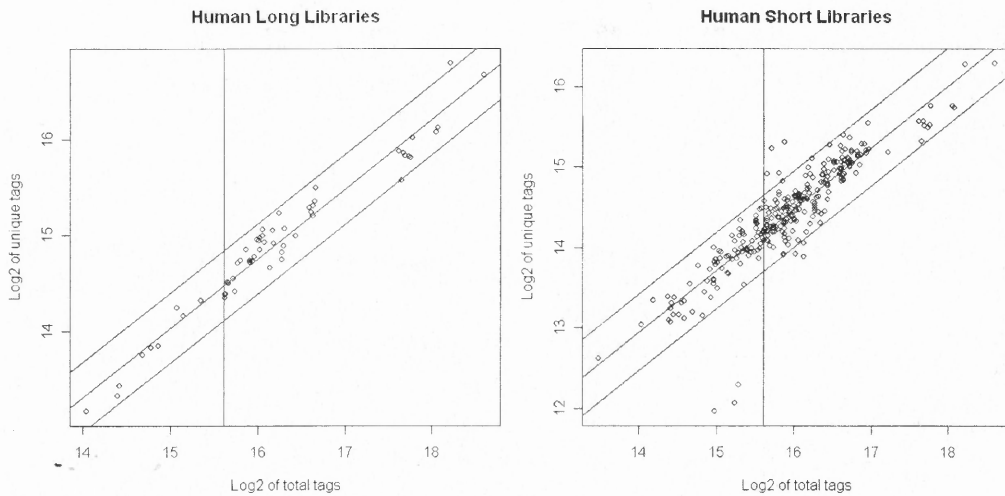


Figure 4.3 Correlation between the number of unique tags and the total number of tags. The data are from human SAGE libraries. The vertical lines represent the first condition and the three skew lines correspond to the second condition. They are used for the selection of the reliable libraries.

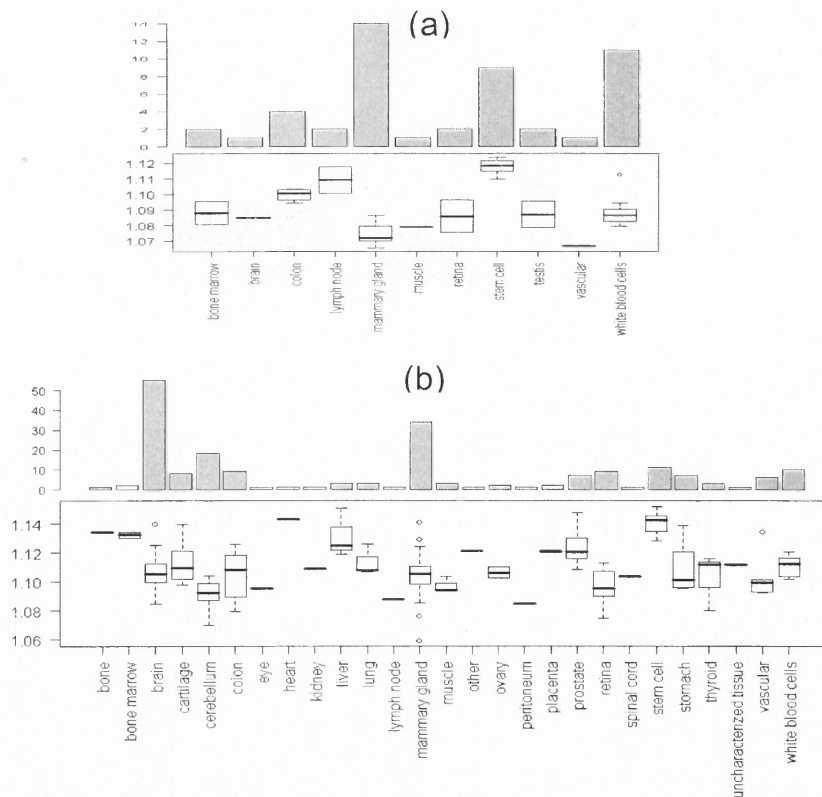


Figure 4.4 The gene expression levels for different tissues. (a) Reliable long tag SAGE libraries for humans. (b) Reliable short tag SAGE libraries for humans. Each column represents one tissue. Each bar on the top represents the number of libraries for that tissue, and the figure on the bottom is the boxplot reflecting the average expression level for that tissue.

4.2.3 The Reliable Tags

Since the mapping from a SAGE tag to a unigene, a transcript cluster is not unique, we need to generate the reliable tags for each unigene. One unigene may contain more than one SAGE tag due to different transcripts belonging to the same unigene. Using the database files `Hs_long.best_gene`, mapping the long tag to the best unigene, `Hs_short.best_gene`, mapping the short tag to the best unigene, `Hs_long.frequencies`, including the long tag frequency data, and `Hs_short.frequencies`, including the short tag frequency data, we generated the reliable tags based on the following two conditions:

- If a short tag was mapped to a unigene by file `Hs_short.best_gene`, and long tag, whose first 10 nt is the same as the short tag, was also mapped to the same unigene by file `Hs_long.best_gene`, then the long tag was considered to be a candidate for the reliable tag of this unigene. It should be noted that the reliable tag is a long tag.
- For a given tag, some libraries may not contain the frequency data for this tag. The tag is considered to be a reliable tag if its ratio (the number of libraries containing the tag divided by the number of total libraries) is greater than 25% (arbitrary percentage, adjustable for more stringent conditions). This can eliminate a lot of tags only expressed in a few libraries.

For example, `GTTCTTGAGAAAAACA` was mapped to gene *cstf-77* by file `Hs_long.best_gene` and expressed in 22 libraries from `Hs_long.frequencies`; `GTTCTTGAGA` was also mapped to gene *cstf-77* by file `Hs_short.best_gene` and expressed in 104 libraries from `Hs_short.frequencies`. `GTTCTTGAGAAAAACA` satisfies the above two conditions, so it is considered to be one reliable tag for gene *cstf-77*. There are a total of 22923 unigenes expressed in these reliable libraries. After applying the first condition, there are 17525 unigenes left. Furthermore, applying the second condition, there remain 9054 unigenes.

4.2.4 Significant Unigenes

Once we generate the reliable tags, the reliable long SAGE libraries and short SAGE libraries can be merged since the short tag frequency data can be treated as the corresponding reliable long tag frequency data. All the following calculations are based on the 250 (201 short + 49 long reliable libraries) merged libraries. To identify the negatively regulated tags, we need at least two reliable tags for each unigene. Among the 9054 unigenes that have at least one reliable tag, only 1957 unigenes have more than one reliable tag. For each unigene, the Pearson correlation of the tag per million (TPM) data (see Appendix C.2) between any two tags was calculated. That is, if the unigene has m reliable tags, we need to calculate $m(m-1)/2$ correlations between these m tags. For each correlation, we calculate the T-value based on the test for independence (see Appendix C.2). If the expression levels of the two tags are opposite to each other, then the T-value should be negative. If we apply a standard t-test to these negative T-values using a p-value = 0.01, then 418 unigenes are selected to contain at least two tags that are negatively regulated.

To get the significant negative regulation, we need to know the T-value distribution of random tags. To obtain this, we randomly selected two tags from all reliable tags 10,000 times. Each time, we shuffled the data 10 times based on a Fisher-Shuffle (see Appendix C.2) and calculated the Pearson correlations and the T-values. We found that the random T-values are linearly correlated with the degrees of freedom (see Figure 4.5), which are the expressed library numbers shared by at least one of the two tags. Sixty-one significant unigenes were selected if we set the p-value cutoff as 0.01. The p-value is calculated based on the random distribution of specified degree of freedom. For example, if the T-value is t_0 for some tag pair with degree of freedom q , from the random simulation, we have the random distribution

for degree of freedom q . The p-value for t_0 is calculated by the number of random data less than t_0 divided by the total number of random data. The detailed information for these 61 unigenes is shown in Appendix C.3.

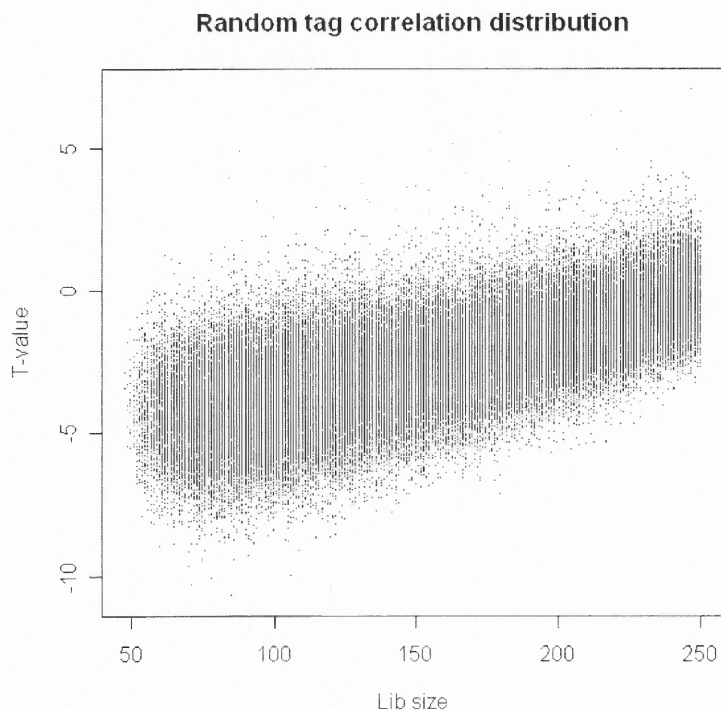


Figure 4.5 T-value distribution of random tags. Each point represents one tag pair. The x-axis is the number of libraries in which at least one tag is expressed. The y-axis is the t-value of the tag pair calculated from the Pearson correlation.

We picked up nine interesting and well-known unigenes (see Table 4.2), trying to understand the mechanism of the negative regulation. Most of them are the DNA/RNA binding proteins. Through generating the alternative transcripts, they may regulate the binding activity and cell function. To see this clearly, we generated the transcript structures corresponding to these unigenes using Bioperl (Stajich et al. 2002). The transcripts were aligned to the genome and the reliable SAGE tags were located. One example, for gene *gtpbp3*, is shown in Figure 4.6. The plots for the other eight unigenes are shown in Appendix C.4.

Table 4.2 Nine Significant Unigenes with Negatively Regulated Transcripts

>Hs.334885	GTPBP3 GTP binding protein 3 (mitochondrial)
>Hs.342307	FLJ10330 PRP38 pre-mRNA processing factor 38 (yeast) domain containing B
>Hs.480073	HNRPD Heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)
>Hs.480073	HNRPD Heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)
>Hs.506759	ATP2A2 ATPase, Ca ⁺⁺ transporting, cardiac muscle, slow twitch 2
>Hs.516539	HNRPA3 Heterogeneous nuclear ribonucleoprotein A3
>Hs.517262	SON SON DNA binding protein
>Hs.529798	BTF3 Basic transcription factor 3
>Hs.531106	RBM25 RNA binding motif protein 25
>Hs.546361	ATP2C1 ATPase, Ca ⁺⁺ transporting, type 2C, member 1

A unigene is the transcript cluster. First column is the unigene ID and the second is the description of the unigene, including gene name and gene function. The information of the tags and correlations is given in Table C.1 of Appendix C.

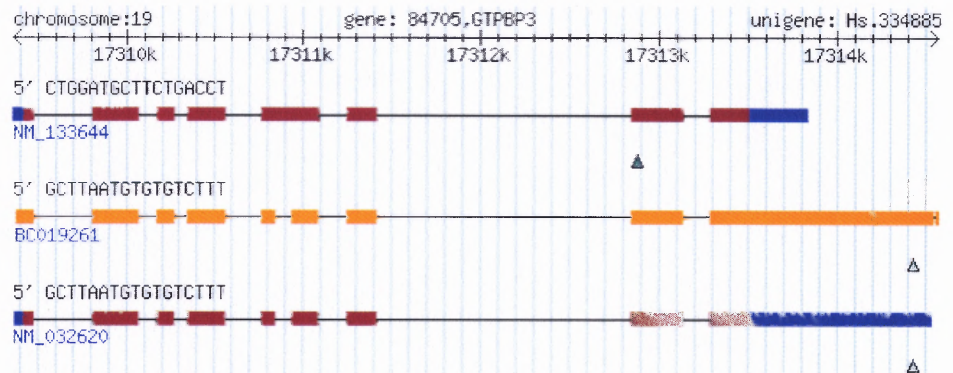


Figure 4.6 The transcript structure of Unigene Hs.334885. The first line represents the chromosome information of the gene, and the following lines represent the transcript structure generated from this gene. The SAGE tag is indicated by the up-triangle and the accession number of the transcript is indicated under the first exon of the transcript.

4.3 Results and Conclusions

Among the 1957 unigenes that have more than one reliable SAGE tag, 54 unigenes have tags that are located in different strands of the chromosome. Among these 54 unigenes, 3 unigenes (indicated in bold typeface in Appendix C.3) are found in the 61

significant unigenes. Most of the transcripts on different strands have co-expression, not negative regulation. Similar results also were found in the literature (Quere et al. 2004). That is, the negative regulation is not mainly due to the orientation of the transcripts, but instead, it must be regulated by some other unknown mechanisms.

From the data of mapping the SAGE tags to the poly(A) sites (private data from Dr Bin Tian's lab), among the 61 significant unigenes, 15 unigenes were found to contain different poly(A) sites, 4 unigenes were found to use the same poly(A) sites, and the poly(A) site information of the remaining 42 unigenes is unknown due to some reasons, such as the different unigene version between SAGE Genie and our database. However, it suggested that the alternative polyadenylation may be one factor that generated the opposite expression pattern of different transcripts.

To our knowledge, this is the first systematic study of the negative regulation between two different transcripts from the same gene, and this may provide some insight into alternative polyadenylation. Here we are mainly focused on the human data. The mouse SAGE data are also available and have larger size than the human SAGE dataset. We can apply the same methods to the mouse SAGE data and find the significant genes. Furthermore, by a homolog analysis between human and mouse genes, negative regulation for some genes may be very significant.

CHAPTER 5

ALTERNATIVE POLYADENYLATION

5.1 Introduction to Alternative Polyadenylation

A large number of human genes have been found to contain more than one poly(A) site, leading to multiple transcripts, even though some of them may not have protein products. It was reported recently that about 54% of human genes and 32% of mouse genes have multiple poly(A) sites (Tian et al. 2005).

Genes were classified into three groups according to the locations of their poly(A) sites (Tian et al. 2005) (see Figure 5.1). A type I gene contains only one poly(A) site; a type II gene has multiple poly(A) sites, all located after the stop codon in the 3'-most exon; a type III gene has poly(A) sites located in regions upstream of the 3'-most exon. It is possible that multiple poly(A) sites are located in the 3' UTRs of the last exon for type III gene. Thus, alternative polyadenylation of a type II gene can lead to variable 3' UTRs with the same protein products. Since 3' UTRs contain RNA *cis*-elements, which are important for mRNA stability, alternative polyadenylation of type II might play a critical role in gene regulation. Alternative polyadenylation of type III genes leads to various protein products.

For each gene type, a further classification was made based on the site locations (Tian et al. 2005). 1S represents a Type I gene poly(A) site; 2F represents the 5'-most poly(A) site in a Type II gene; 2L represents the 3'-most poly(A) site in a Type II gene; 2M represents a middle poly(A) site between 2F and 2L in a Type II gene; 3U represents the poly(A) site located upstream of the 3'-most exon in a Type III gene; 3S represents a single site in the 3'-most exon of a Type III gene; 3F, 3M, 3L,

similar to 2F, 2M, 2L, represent poly(A) sites in a Type III gene. For convenience, 3D represents 3S/3F/3M/3L.

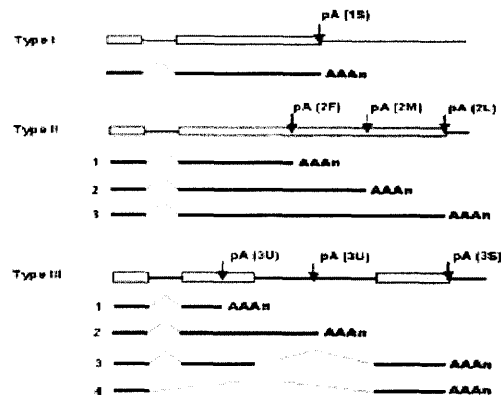


Figure 5.1 Schematic representation of poly(A) sites. Types I, II, and III represent three groups of genes based on the locations of the poly(A) sites. Type I genes contain only one poly(A) site; Type II genes contain multiple poly(A) sites located in the last exon after the stop codon; Type III genes contain multiple poly(A) sites with at least one poly(A) site located in the upstream region of the last exon. Figure taken from (Tian et al. 2005).

The selection of alternative poly(A) sites is believed to be related to non-molecular biological factors, such as developmental stages and cell conditions (Edwards-Gilbert et al. 1997). For example, the IgM heavy-chain gene switches from using one poly(A) site to another during B lymphocyte maturation (Takagaki et al. 1996). This results in a shift in protein production from a membrane-bound form to a secreted form due to the deletion of a C-terminal hydrophobic region responsible for membrane interactions. This switch is an essential step in the immune response. However, the detailed mechanisms involved in alternative polyadenylation are still unknown. What are the molecular factors that determine or regulate the selection of poly(A) sites? Are they polyadenylation protein factors? Is a splicing factor involved?

5.2 Protein Factors Involved in Polyadenylation

Six different *trans*-acting protein factors have been identified as necessary for *in vitro* cleavage and polyadenylation (Keller 1995) (see Figure 5.2). They are the Cleavage and Polyadenylation Specificity Factor (CPSF) (Bienroth et al. 1991; Gilmartin and Nevins 1991; Murthy and Manley 1992), Cleavage stimulation Factor (CstF) (Takagaki et al. 1990; Gilmartin and Nevins 1991), Cleavage Factor Im and IIm (CFIm, CFIIIm) (Takagaki et al. 1989; Ruegsegger et al. 1996), Poly(A) Polymerase (PAP) (Takagaki et al. 1989), and Poly(A) Binding protein II (PAB II) (Wahle 1991).

CPSF is required for cleavage and polyadenylation, and it contains four subunits (30, 73, 100, 160 kDa). Among these, CPSF-160 binds to the poly(A) signal AAUAAA or its variants and interacts with the subunit of CstF, and CPSF-73 is involved in histone-pre-mRNA processing (Dominski et al. 2005). PAP is also required for both cleavage and polyadenylation, initiating the addition of a poly(A) tail to the 3' end of pre-mRNA (Murthy and Manley 1992). PABII binds to the growing poly(A) tail and acts as an elongation factor, regulating its ultimate length and stimulating maturation of the mRNA (Mangus et al. 2003).

CstF is composed of three subunits (50, 64, 77 kDa), namely, CstF-50 that interacts with RNA polymerase II C-terminus Domain (CTD) (McCracken et al. 1997), CstF-64 that directly binds to U/GU-rich elements of transcripts containing poly(A) signals (Perez Canadillas and Varani 2003), and CstF-77 that has been shown to interact with several factors involved in cleavage and polyadenylation. For example, CstF-77 interacts with CstF-50 (Takagaki and Manley 1994), CstF-64 (Hatton et al. 2000), CPSF-160 (Murthy and Manley 1995), and RNA polymerase II CTD (McCracken et al. 1997). It also can interact with itself (Takagaki and Manley 2000). A schematic description of the poly(A) region including *cis*-elements and some

protein factors such as CPSF, CstF and RNA Poly II (Polymerase II) is shown in Figure 5.3.

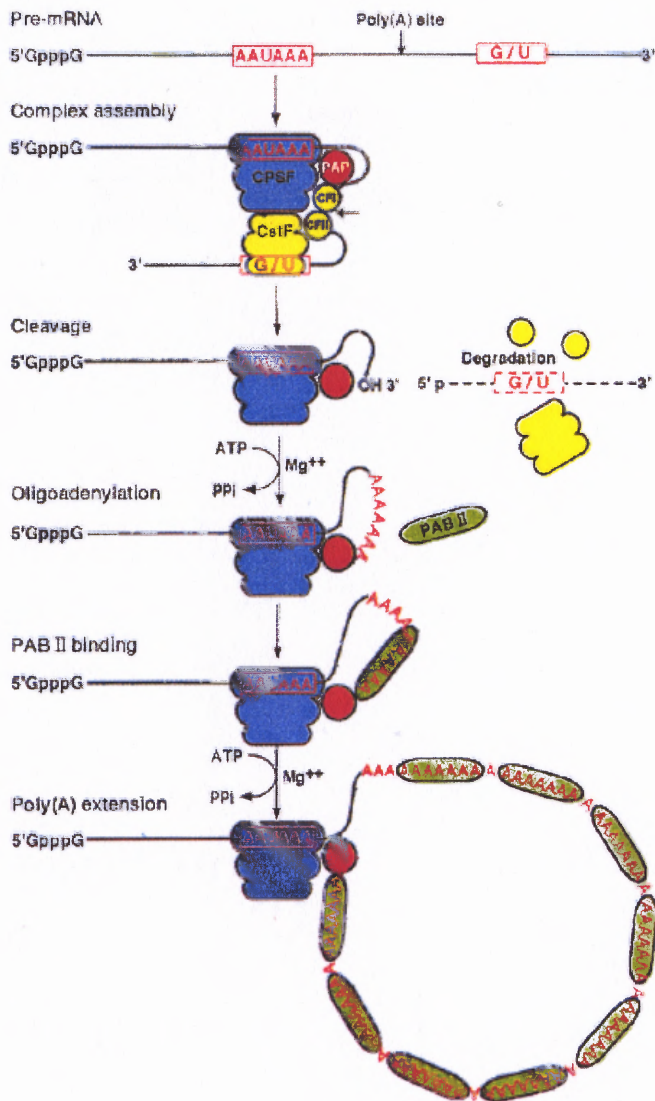


Figure 5.2 Schematic representation of the steps involved in the mammalian pre-mRNA 3' process. Five protein complexes are involved in the polyadenylation process: cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factor I, II (CFI, CFII), and polyadenylation binding protein II (PAB II). Figure taken from (Keller 1995)

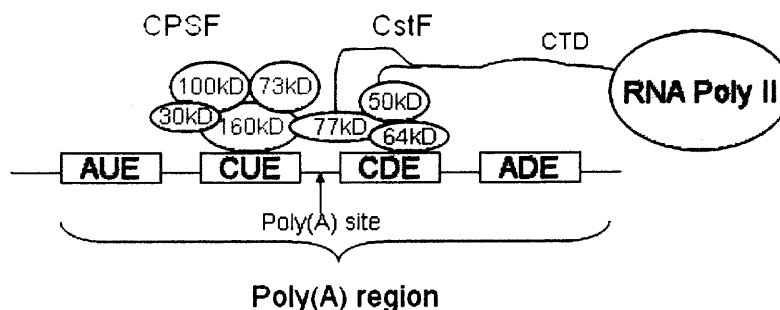


Figure 5.3 Protein factors involved in polyadenylation. CTD, the c-terminus domain. CPSF, the cleavage and polyadenylation specificity factor, includes the four subunits: 100kD, 73kD, 30kD, and 160kD, and CstF, cleavage stimulation factor, includes the three subunits: 50kD, 77kD, and 64kD.

CFIm and CFII_m are required for the cleavage reaction only, but they are less well characterized than the other cleavage-polyadenylation factors (Ruegsegger et al. 1996; Ruegsegger et al. 1998). CFIm has been found to be the factor that recruits the CPSF and PAP to bind with the pre-mRNA if the sequence does not contain a canonical PAS (Venkataraman et al. 2005). CFII_m has been shown to bridge two other cleavage factors (de Vries et al. 2000).

In addition to the six main factors above, other proteins also have been shown to be involved in cleavage and polyadenylation, including CTD of RNA polymerase II (Park et al. 2004; Kaneko and Manley 2005), Symplekin (Takagaki and Manley 2000), PC4 (Calvo and Manley 2001), hnRNP F (Veraldi et al. 2001), hnRNP H/H' (Arhin et al. 2002), Ssu72 (Steinmetz and Brow 2003), U2AF65 (Millevoi et al. 2002), U1 snRNP-A (Lutz et al. 1996; Phillips et al. 2004), SRp20 (Lou et al. 1998), and Fip1 (Forbes et al. 2006). Some of these factors are involved in both polyadenylation and transcription, which supports the notion that these processes are tightly coupled together (Adamson et al. 2005). All of these evidences suggest that polyadenylation can be regulated by many factors and is tightly coupled with other processes.

5.3 Proposed Mechanisms for Alternative Polyadenylation

We know that Type II genes and Type III genes have multiple poly(A) sites. The switch between alternative poly(A) sites may be different for these two different types. For Type II genes, poly(A) sites are all located in the 3'-most exon, usually in 3' UTR, therefore, sometimes they are called tandem poly(A) sites. There is no additional splicing upon reaching the last exon. Also, the distance between these different poly(A) sites is usually not long, at most the length of the last exon. For Type III genes, the switch between 3U and 3D must skip some exon(s) or part of an exon. Thus, splicing at some location is very important for this kind of switch. However, the detailed relationship between splicing and polyadenylation is not well understood. In the following, some hypothetical mechanisms are proposed and investigated for the switching of poly(A) sites for type II and III genes.

5.3.1 Type II Gene Poly(A) Site Switching

For switching between poly(A) sites in type II genes, the alternative transcripts have the same upstream exons except the last exon (see Figure 5.4). The last exons are different only at the 3' end due to the usage of different poly(A) sites. Here, two poly(A) sites PA1 and PA2 are investigated. Since the distance between these two sites is not long, the two sites have an equal chance to be selected, i.e., distance will not affect the selection of different poly(A) sites. Thus, PA1 has no advantage to be selected even though the region around PA1 has been transcribed first.

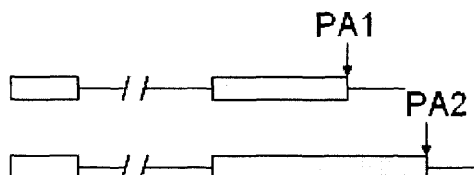


Figure 5.4 Type II gene poly(A) site switching. PA1 and PA2 represent different poly(A) sites in the last exon.

The following four mechanisms M1-M4 are proposed hypothetically and theoretically, based on the existing literature.

M1: *An upstream poly(A) site has a higher priority than a downstream poly(A) site, and the stronger the poly(A) site is, it has a higher probability of being selected.*

The strength of a poly(A) site is generally determined by the usage of different poly(A) signals, the GU/U-rich element of the downstream region of a poly(A) site, and some other unknown *cis*-elements. It also can be estimated from the *polya_svm* program that was discussed in Chapter 3. The smaller the *polya_svm* score is, the greater chance that the poly(A) site will be used.

In herpes simplex virus type 1 (HSV), the *tk* gene preferred to use the upstream poly(A) site (Denome and Cole 1988) according to the following two facts: increasing the number of poly(A) signals 3' to the *tk*-coding region did not affect the total amount of polyadenylated RNA produced; increasing the distance between two signals caused an increase in the use of the 5' signal and a decrease in the use of the 3' signal. This indicates that increasing the number of poly(A) signals may not increase the strength of the poly(A) site and so PAS, the main poly(A) site determinant, is not the unique factor to determine the strength of the poly(A) site. The viral mRNA BPV in mouse cell uses the upstream poly(A) sites 4 times as often as the downstream poly(A) sites (Andrews and DiMaio 1993). When identical tandem viral signals are separated by fewer than 400 nt, they competed for polyadenylation and the upstream site is always chosen preferentially (Batt et al. 1994).

In human acute lymphocytic leukemia cell, the *c-myc* gene uses the downstream poly(A) site 6 times more often than the upstream poly(A) site (Laird-Offringa et al. 1989). To test if the downstream poly(A) site is stronger than the upstream poly(A) sites for gene *c-myc*, the 250 nt poly(A) regions were obtained

from polyA_DB2 (Lee et al. 2007) and predicted by the polya_svm program. The polya_svm score is 5.644 for the upstream poly(A) site and 4.133 for the downstream poly(A) site. This confirmed that the downstream poly(A) site is stronger than the upstream poly(A) site. Gene *cox-2* uses the distal stronger poly(A) site more often than the proximal weaker poly(A) site, and the alternative polyadenylation is also tissue-specific (Hall-Pogar et al. 2005). Using the same method applied to *c-myc*, we obtained a polya_svm score of greater than 6 for the upstream poly(A) site, and a score of 3.943 for the downstream poly(A) site of gene *cox-2*. These two examples confirmed that the strengths of poly(A) sites would contribute to the selection of those poly(A) sites. This also agrees with Figure 3.9 in the Chapter 3.

M2: *The upstream or downstream regions of the poly(A) site may contain some protein binding sequences, excluding the sequences necessary for polyadenylation factors.*

If the upstream or downstream regions of the poly(A) site contains some specific sequences, which are the binding sites for some proteins, not polyadenylation factors, the selection of this poly(A) site would be more complicated. Once the sequence around the poly(A) site is bound by some protein, it may prevent the binding of the polyadenylation factors and thus lower the usage of the poly(A) site. For example, if the sequence is very near PAS, the binding of the protein may prevent CPSF from binding to the PAS. This means this poly(A) site may be skipped and the following site may be used.

Polypyrimidine Tract Binding protein (PTB) was shown to modulate the efficiency of polyadenylation by binding to a downstream element of the poly(A) site (Castelo-Branco et al. 2004), thereby inhibiting the polyadenylation at that site and making the usage of the following poly(A) site possible. The drosophila sex-lethal

protein mediates the polyadenylation switching in the female germ-line (Gawande et al. 2006) by binding to the multiple SXL-binding sites, which include the GU-rich poly(A) enhancer, thus competing for the binding of CstF-64 in vitro. In this case, the selection of a poly(A) site not only depends on the sequence in the poly(A) region, but also on the availability of some proteins and the RNA-protein interactions. The adenovirus MLTU encodes five collinear mRNA families, where the upstream poly(A) site is predominantly used in the early stage, and downstream poly(A) site is used in the later stage (DeZazzo and Imperiale 1989). This may be caused by the absence of some proteins binding to the sequence around the upstream poly(A) site in the early stage.

On the other hand, if some other proteins bind to the region near to PAS or to the downstream U/GU-rich region, it may recruit the polyadenylation factors such as CPSF and CstF more efficiently and enhance the usage of the poly(A) site near the binding sites of the other factors. The U3 sequence has been demonstrated to promote the interaction of CPSF with the core poly(A) site in lentiviruses (Graveley and Gilmartin 1996).

M3: Secondary structure conformation of pre-mRNA around the poly(A) site may influence the selection of poly(A) sites.

It has been found that repression of the 5' poly(A) signal and utilization of the 3' poly(A) signal occurs in HIV-1. They may be caused by the hairpin loop structure around the PAS, and therefore, inhibiting the binding of the polyadenylation factors CPSF (Klasens et al. 1999; Gee et al. 2006). The secondary structure near the downstream U/GU-rich region also might affect the binding of polyadenylation factor CstF. To our knowledge, there is no evidence reported for that.

M4: *The concentration of some polyadenylation factors will affect the poly(A) site selection.*

It has been shown that the alternative polyadenylation is strongly influenced by increasing the concentration of CstF-64, resulting in an increase in the selection of a weaker promoter-proximal poly(A) site over a stronger promoter-distal poly(A) site (Takagaki and Manley 1998; Shell et al. 2005). We do not know for any other polyadenylation factors that have been observed to affect the site selection by changing the concentration.

5.3.2 Type III Gene Poly(A) Sites Switching between 3U and 3D?

For the type III gene poly(A) site switching, the mechanism may be a little different compared with that for type II genes since there is no splicing involved in the type II gene poly(A) site selection. However, some mechanisms described for type II genes may also be true for type III genes. For instance, if some other proteins bind to the splicing protein binding sites and block the binding of splicing factor, slowing down the formation of the splicing complex, the following poly(A) site PA1 (see Figure 5.5) may be used. If some other proteins bind to the upstream or downstream of polyadenylation sites, improve the binding of polyadenylation factors by protein-protein interaction, accelerating the formation of the polyadenylation complex, the poly(A) site PA1 (see Figure 5.5) may be used.

If some other proteins bind to the upstream or downstream of the splicing binding sites and facilitate the binding of the splicing factor, therefore accelerating the formation of the splicing complex, the splicing site (see in Figure 5.5) may be used. Also, if some other proteins bind to CPSF or CstF binding site, slow down the binding of polyadenylation factors, the splicing site may be used.

If the two poly(A) sites are very close in type III genes, the mechanisms described in type II genes may also be true for type III genes. If the two sites are separated far apart, then the splicing process may compete with the cleavage and polyadenylation processes and so affect the determination of the poly(A) site. If the splicing site (SS) has been processed first, the first transcript (the top one in Figure 5.5) will be generated; if the splice site was skipped, no splicing (NS) happens, then the second transcript (the bottom one in Figure 5.5) will be generated. The two mechanisms M5-6 are proposed.

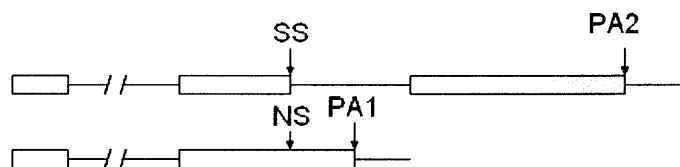


Figure 5.5 Type III gene poly(A) site switching. SS, splicing sites; NS, no splicing; PA1 and PA2, the two poly(A) sites for the gene. PA1 is an intronic poly(A) site.

M5: *The strength of the splicing site (SS) and the poly(A) site determine which one will be used first.*

The strength of the 5' splicing site is determined by the 5' boundary or donor site of introns containing the canonical di-nucleotide GU and the sequence composition around the donor site, such as the exonic splicing enhancers (Cartegni et al. 2002). The alternative splicing is another difficult topic nowadays, and the detailed mechanisms are less known. It is hypothesized that if the SS is very strong, the first transcript will be generated. Otherwise, the second transcript will be generated (See Figure 5.5).

The time difference for the formation of splicing complex and polyadenylation complex may lead to a different poly(A) site. If a splicing complex forms very fast, the sequence (intron) containing poly(A) site will be spliced out and so the first

transcript will be generated. If a splicing complex forms very slowly such that the polyadenylation complex completes the formation before the completion of splicing complex, then polyadenylation will occur, the splicing complex will degrade, and the second transcript will be generated. The formation time is assumed to be correlated with the strength of the site.

M6: *RNAP's moving speed may be related to polyadenylation, such as transcription elongation, pausing, arrest, and termination.*

Transcription termination has been believed to be coupled with polyadenylation (Buratowski 2005). But the order of transcription termination and polyadenylation is still unknown. If termination occurs first, then the poly(A) site close to the termination site may be used. If polyadenylation occurs first, it may facilitate the termination, without contributing to the selection of a poly(A) site. Also, the defects in the transcriptional elongation factor have been shown to enhance the utilization of an upstream poly(A) site (Cui and Denis 2003). It is believed that the defects cause the increase of transcription pausing and arrest, facilitate the transcription termination, and hence elicit the polyadenylation.

5.4 Proposed Model for Alternative Polyadenylation of Type III Genes

Based on the above proposed mechanisms, it is obvious that the selection of poly(A) sites is a very complex process, and differences exist among different genes and different species. For the alternative polyadenylation of type III genes, there are three protein complexes involved: RNA polymerase II complex, which is the transcriptional machinery; the spliceosome complex, which is the splicing machinery; and the polyadenylation complex, which is the cleavage and polyadenylation machinery for most messenger RNA 3' end formation. A simple linear model can be proposed for

Type III gene with composite exon (see Figure 5.5) based on the idea that the order of protein complex formation determines the usage of the poly(A) site: if the spliceosome complex forms first, the first transcript will be generated; if the polyadenylation complex forms first, the second transcript will be generated.

After the transcriptional initiation, RNAP moves along the DNA template sequence and pre-mRNA transcript is produced. When the RNAP passes the 5' SS, the splicing complex begins to form. The time needed for the completion of the complex may depend on the strength of the SS, the availability of splicing factors, and other unknown factors. When the RNAP passes the poly(A) site, the polyadenylation complex begins to form. The time needed for the completion of the complex may depend on the strength of poly(A) site, the availability of polyadenylation factors, and other unknown factors. The elongation speed of RNAP will also affect the timing order. To set up the model, some parameters need to be derived, such as the average time of splicing and polyadenylation complex formation, and the RNAP moving speed.

The overall reaction rate of transcription for bacterial RNA polymerase is ~ 40 nt/second at 37°C , and this is about the same as the rate of translation (15 amino acid/second), but slower than the rate of DNA replication rate (800 nt/sec) (Lewin 2003). In vitro, the speed may be lower, i.e., about 2~5 nt/s (Korzheva et al. 2000).

The time for assembly of the cleavage and polyadenylation apparatus is about 10 seconds in vivo and is faster for a strong poly(A) site than for a weak one (Chao et al. 1999). The reason why polyadenylation assembly forms so fast may be that the cleavage and polyadenylation processes need RNAP CTD (Park et al. 2004; Kaneko and Manley 2005), and the RNAP can only stay around the poly(A) site for about 10 seconds (within 50 nt for the speed 5 nt/s). However, the assembly time may depend

on the polyadenylation factors such as CPSF, CstF, and CF, etc., since it has been shown that the concentration of CstF-64 will affect the selection of poly(A) sites (Shell et al. 2005). The time is then a function of many factors, that is, $T_p = p(p_1, p_2, \dots)$, where p_i are the polyadenylation factors and the strength of the poly(A) site.

The formation of the spliceosome assembly will take several minutes in yeast (Ruby 1997). Also, the time may depend on the splicing factors, that is, $T_s = s(s_1, s_2, \dots)$, where s_i are the splicing factors and the strength of the splicing sites. Assume the time for the RNAP traveling n nucleotides is T_n , which is a function of the distances from the splicing sites and the poly(A) sites, and other factors that affect the speed of RNAP. $T_n = r(n, f_1, f_2, \dots)$, where f_i are some unknown factors, including the concentration of NTPs (Liu and Alberts 1995; Herbert et al. 2006). Also the speed of RNAP varies with different genes due to the different compositions of the sequence. Highly active genes (e.g., rRNA or heat shock genes) may have higher elongation rates (Giardina and Lis 1993; Condon et al. 1995). There is some evidence shown that the distance between the SS and poly(A) site (PS) will affect the selection of different poly(A) sites (Levitt et al. 1989; Qiu and Pintel 2004).

Based on all these evidence and the hypothesis, we proposed that if $T_n + T_p > T_s$, then splicing occurs first, otherwise, cleavage and polyadenylation occur first (see Figure 5.6).

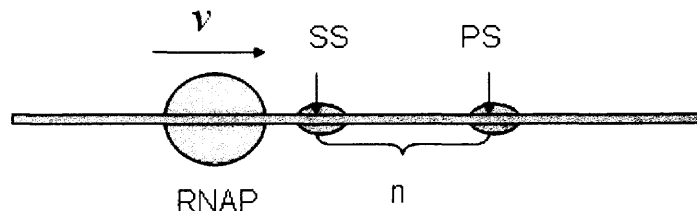


Figure 5.6 Competition between a SS and a PS. SS, splicing site; PS, poly(A) site; v , the speed of the RNA polymerase II (RNAP); n , the distance between SS and PS.

This hypothesis is very difficult to test due to many unknown factors, such as the RNAP elongation speed, the spliceosome, polyadenylation complex formation times, and the availability and concentration of protein factors that would affect the splicing and polyadenylation.

Is there any correlation among the distances (DIS) between splicing sites and poly(A) sites, the strengths of the 5' SS, and the strengths of PS? The scores reflecting the strengths of these sites were generated (private data, obtained from Dr Tian's lab), and they were assumed to be correlated with the complex formation time. The 5' splicing site scores (SC5) range in (-12.87, 12.1), and polyadenylation site scores (PAs) range in (-7.25, 9.04). The PAs here are generated based on PSSM only and they are a little different from `polya_svm`. It is assumed that the smaller the score, the longer the time needed for the formation of the complex. We notice that there is no obvious correlation between them from the plot of DIS vs the scores (see Figure 5.7). This is not what we had expected that if the distances between SS and PS are large, the PA score should also be large, and the 5' splicing site score should be small such that there is some competition between the splicing and polyadenylation. There should exist some other factors.

Next, we took the RNAP moving speed into account. Generally, splicing takes longer than polyadenylation. An assumption was made about the time based on the existing literatures: the SS with the minimal SC5 needs 200 seconds to finish the assembly of a spliceosome complex, and the SS with the maximum SC5 needs 50 seconds to finish the assembly; the PS with the minimal PAs needs 100 seconds to finish the assembly of the polyadenylation complex, and the maximum PAs needs 10 seconds to finish the assembly. We also assumed that the relationship between the scores and the times were linear and the RNAP speed was 50 nt/s.

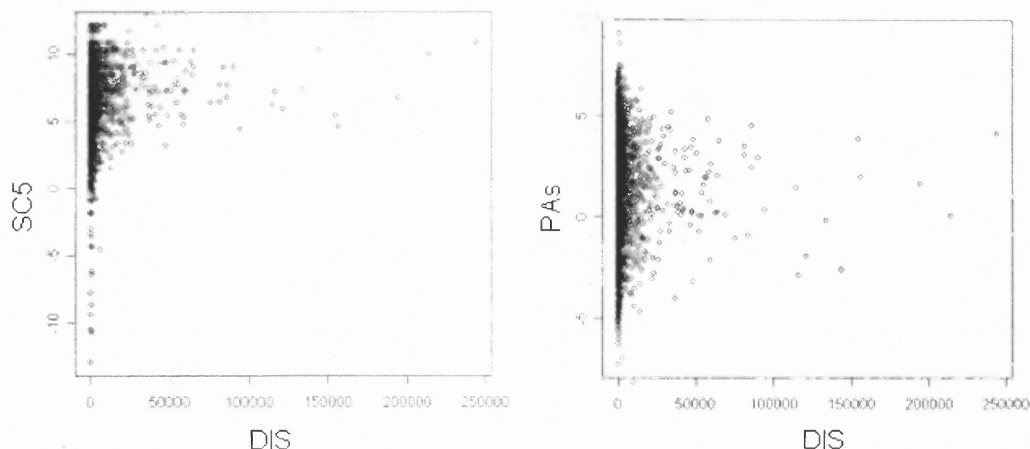


Figure 5.7 Scatter plots of SC5 and PAs vs DIS. SC5, splicing sites score; PAs, polyadenylation site score; DIS, distance between a splicing site and a poly(A) site.

Based on the above assumptions, a plot is made and shown in Figure 5.8 (a). Each point in the plot represents one gene. Based on the scores for splicing sites and polyadenylation sites, the x-axis is the estimated time for the formation of a splicing complex, and the y-axis is the estimated time for the formation of a polyadenylation complex plus the time for RNAP traveling from SS to PS. If they are distributed along the diagonal line, that means both poly(A) sites have chances to be used. The majorities are located around the one center, which is the same as expected. But some are still located vertically (see Figure 5.8 (a)), which means that even if the distance is very large or the poly(A) site is very weak, the downstream weak poly(A) site is still selected. This implies that the selection of a poly(A) site is not only determined by the strength of the site, but also many other unknown factors. The RNAP speed might be very high for some large introns, or the sequence around the 5' splicing site and the poly(A) site contains some binding sequences for other factors. If we delete genes with long distances or we limit the time for polyadenylation and RNAP traveling to be within 200 s, the plot shown in Figure 5.8 (b) reflects the competition.

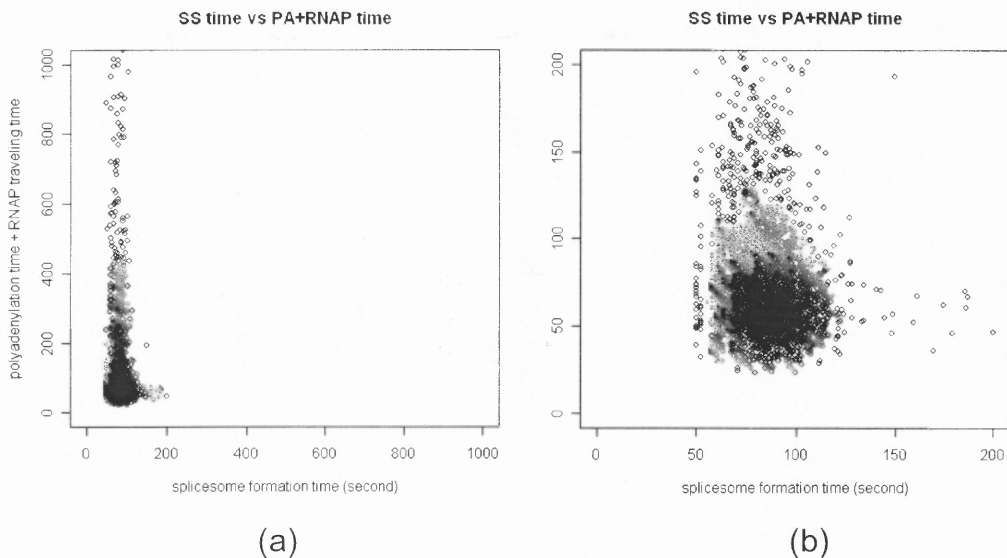


Figure 5.8 Polyadenylation assembly time + RNAP traveling time vs spliceosome assembly time. Each point represents one gene with an intronic poly(A) site. The units on the x- and y-axes are in seconds. (a). All possible genes. (b). Genes with times for polyadenylation assembly and RNAP traveling less than 200 seconds.

Due to the large number of factors contributing to the selection of different poly(A) sites, it is very necessary to focus on some specific gene to understand alternative polyadenylation. Mathematical modeling is a strong tool to study the dynamics of the selection of different polyadenylation sites and this leads the last part of this work.

CHAPTER 6

MATHEMATICAL MODELING OF CSTF-77 ALTERNATIVE POLYADENYLATION

6.1 Introduction to CstF-77

To understand alternative polyadenylation, some possible mechanisms being proposed have been described in the previous section. The mechanisms concerning the selection of different poly(A) sites are very complex, in general, and different genes may use different pathways. Therefore, we want to focus on some special genes to study the regulation of two different transcripts produced by alternative polyadenylation. Gene *cstf-77* (sometimes called *cstf3*), whose protein product is one of the subunits of CstF was chosen for this study. This gene was selected for two reasons: the gene product is a polyadenylation factor and this gene can produce more than one transcript due to alternative polyadenylation. We believe that this kind of gene not only can control the efficiency of polyadenylation by alternative polyadenylation, but also can form a feedback loop by auto-regulation.

The alternative transcript of human *cstf-77* due to alternative polyadenylation has been found recently (Pan et al. 2006). Moreover, it has been found that the expression levels of the two transcripts are opposite to each other based on the observation on the SAGE data (see Figure 6.1). An intronic poly(A) site has been found for *cstf-77* and this leads to the relatively short CstF-77 transcript (referred as CstF-77.S) (see Figure 6.2). The short form protein lacks the function of a normal form (the normal CstF-77 transcript is referred to as CstF-77.L, the long transcript) as a subunit of CstF (Pan et al. 2006) and its detailed function is still unknown.

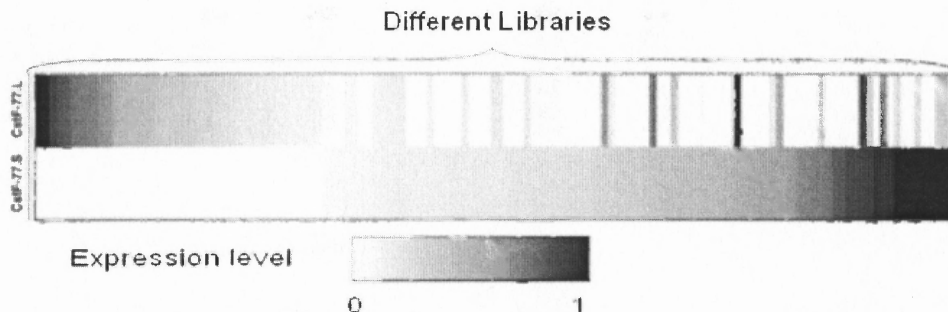


Figure 6.1 SAGE library data for CstF-77.L and CstF-77.S. Each of the 289 vertical lines represents a different library. The top row represents the expression levels of the long form of CstF-77, and the bottom row represents the expression levels of the short form of CstF-77. Figure taken from (Pan et al. 2006).

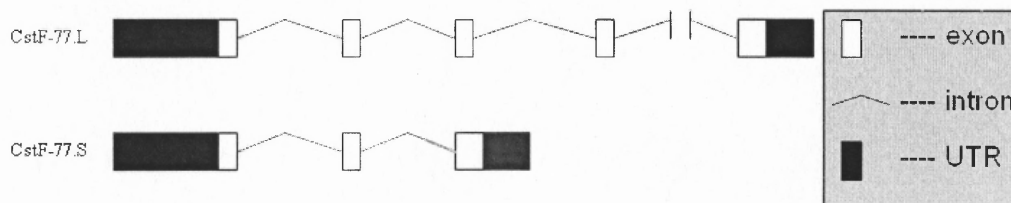


Figure 6.2 Transcript structure of gene *cstf-77*. The use of an upstream poly(A) site generates the CstF-77.S and the use of a downstream poly(A) site generates the CstF-77.L. Figure taken from (Pan et al. 2006).

A number of genes have been found to contain multiple transcripts with opposite expression levels from Chapter 4. Some of the transcripts are generated due to alternative polyadenylation. Others are due to alternative splicing. Is it really true that the long and short form transcript of CstF-77 have opposite expression levels? If this is true, then how does this happen? What controls this type of dynamical expression in a tissue-specific way? If these were understood, it would help us to understand the mechanism of alternative polyadenylation.

There is no reason to assume that they interact directly with each other by RNA-RNA interaction. One possible way for the regulation is that the protein products of the long and short form have different effects on the transcription

efficiency for generating the long and short form transcripts, and thus form an auto-regulation network. We know that the long form protein is a general polyadenylation factor and is required for polyadenylation. However, the function of the short form protein is unknown.

The concentration of CstF-64 has been shown to regulate gene expression (Takagaki and Manley 1998; Shell et al. 2005), but to our knowledge, there is no evidence for CstF-77 and other polyadenylation factors to regulate gene expression. To set up the model, some assumptions will be made for the function of the short form protein.

6.2 Proposed Model of Alternative Polyadenylation for CstF-77

The human CstF-77 short form transcript was first described in (Pan et al. 2006) and no additional literature has been found related to it. This transcript is generated by the intronic polyadenylation site, which has been studied widely (Lou et al. 1998; Bruce et al. 2003; Tian et al. 2007). However, the regulation between the two transcripts from the same gene has seldom been studied. There are a number of reasons for the lack of this kind of study. First, not all of the protein products from different transcripts of the same gene are well understood. Second, a large number of alternative transcripts are expressed at very low levels, and the current technology can not detect these transcripts. Third, the regulation between two different transcripts from the same gene seldom draws too much attention. Fourth, potential products from different transcripts may not regulate the expression of the same gene.

Here, we introduce a new mathematical model to discover and understand the regulation of two different transcripts generated from the same gene due to alternative polyadenylation. We then determine the consequences of the hypotheses in the model.

These two transcripts have the same transcription initiation site at the 5' end, and they have the same 5' un-translated region (UTR). However, the polyadenylation sites at the 3' end are different, with one located upstream of the other. The two poly(A) sites are called the upstream poly(A) site (UPA) and the downstream poly(A) site (DPA). If the UPA is used, then the DPA would not be used for this transcription initiation. On the other hand, if the UPA was skipped for some reason, then the DPA would be used. The detailed mechanism of when and how the upstream poly(A) site is used is not clear and some possible mechanisms were discussed in Chapter 5.

The efficient usage of the poly(A) site is assumed to depend on the protein concentrations of the polyadenylation factors. Some polyadenylation factors, such as CstF-64 (Shell et al. 2005) have been found to affect the polyadenylation efficiency. However, in this study, we assume the concentration of CstF-64 is the same for all the time. That is, the CstF-77 long form protein is assumed to be the only factor that would affect the polyadenylation efficiency. The CstF-77 long form protein is characterized as one subunit of CstF, which is necessary for the cleavage of the pre-mRNA. The function of CstF-77 short form protein is not known. Here, the short form protein either has no function or it may have an inhibitory effect on the cleavage based on the structure analysis in (Pan et al. 2006). Moreover, if it has the same function as the long form protein, then it can be treated as a long form protein and the system would become a positive feedback system in one variable, thus there is no regulation. By setting up the model and using simulations, it is possible to deduce a putative function of the short form protein.

Some work has been done to study the transcription initiation rate for some specific genes. For example, the TATA box containing the promoter region will affect the binding of the TATA binding protein (TBP), therefore, affecting the transcription

initiation rate (Antoniou et al. 1995; Hoopes et al. 1998). The structure of RNAP and non-coding RNA also has been implied in transcription initiation (Young et al. 2002; O'Gorman et al. 2006). For our study, we assume the transcription initiation rate is constant for gene *cstf-77*.

The regulation network can be described as a series of events (see Figure 6.3): when the transcription starts at the transcription initiation site (TIS), the RNAP travels along the DNA template sequence and the pre-mRNA is formed; when the UPA is exposed, cleavage and polyadenylation may happen at that time and the CstF-77.S is formed; if the UPA is not used for some reason, the RNAP continues to move and the DPA will be exposed; the CstF-77.L will be generated by using the DPA; the mRNA transcripts will be exported to cytoplasm, where they are translated into proteins; after the folding and some modification processes, the proteins will then be transported back into the nucleus to act as polyadenylation factors, and then they would affect the transcription efficiency by polyadenylation.

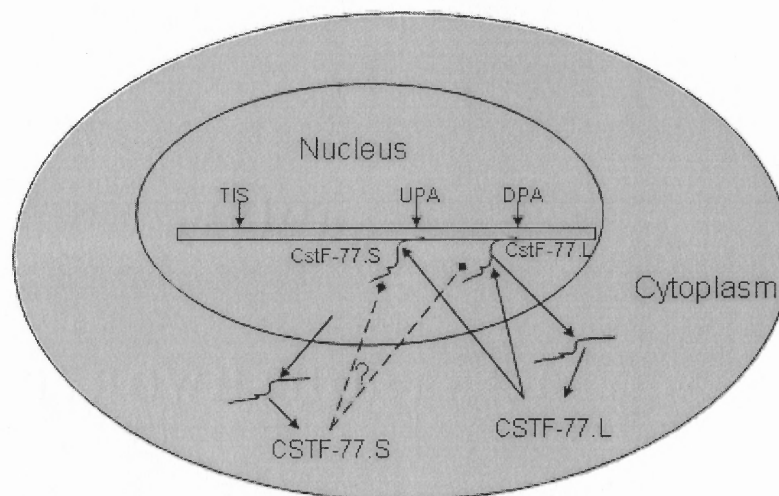


Figure 6.3 Alternative polyadenylation of CstF-77. TIS, transcription initiation site; UPA, upstream poly(A) site; DPA, downstream poly(A) site. CstF-77.S and CstF-77.L represent the short and long form mRNA transcripts. CstF-77.S and CstF-77.L represent the short and long form proteins. CstF-77.S is generated by using UPA, and CstF-77.L is generated by using DPA.

6.3 Mathematical Description

RNA-protein interaction can be viewed as a substrate-enzyme interaction and the Michaelis-Menten equation has been used to describe this kind of kinetics (Fall et al. 2003). Goodwin (Goodwin 1965; Goodwin 1966) postulated an oscillatory model to simulate the dynamics of mRNA and protein interactions, and the dynamical behavior of this model was analyzed by J.S. Griffith (Griffith 1968(a); Griffith 1968(b)). Since the poly(A) site of the short form is located upstream of the long form, this increases the opportunity for the short form to be formed first. This also can be seen from the proposed mechanism M1 from Chapter 5. Thus, it is natural to assume that the transcription rates of the long and short forms have the following representations using the Goodwin oscillator:

$$\begin{aligned} P_s &= \alpha \frac{L^n}{K^n + \epsilon S^n + L^n}, \\ P_l &= I_0 - \alpha \frac{L^n}{K^n + \epsilon S^n + L^n}, \end{aligned} \tag{6.1}$$

where P_l and P_s are the transcription rates of the long and short forms, respectively, α is the maximum transcription rate of the short form, L and S are the relative protein concentrations for the long and short forms, respectively, I_0 is the transcription initiation rate, which is assumed to be constant, K is the equilibrium constant, n is the Hill coefficient, and ϵ is the relative inhibition constant for the short form. Here we assumed that $I_0 = P_l + P_s$, which means that for any transcription initiation, either the long form or short form is produced. If $\epsilon = 0$, then the short form protein has no effect on polyadenylation.

As we know, in eukaryotes, when a gene is turned on, a messenger RNA will be made, processed, and transported from the nucleus to the cytoplasm before the protein is generated. On the other hand, when a gene is turned off, the protein

production will not be affected immediately until the RNA that already has been transcribed decays. The delay time for the formation of a functional protein from a mature mRNA molecular has received some attention (Scheper et al. 1999; Covert et al. 2005). The time delays are different for different genes. The correlation between mRNA and protein abundance has been widely studied in other species (Greenbaum et al. 2003; Nie et al. 2006), however, the correlation does not appear to be well defined because of noise and other factors.

The protein concentration is more difficult to measure than mRNA expression level, therefore, the variables in the model will be chosen to be the expression levels of the mRNA transcripts. We assume that the protein level at the current time is proportional to the mRNA level that has been transcribed at a previous time. Thus we have the relationships: $L(t) \sim l(t-\tau_0)$ and $S(t) \sim s(t-\tau_0)$, where t is the time variable, $l(t)$ and $s(t)$ are the relative mRNA expression levels of the long and short form at time t , τ_0 is the time duration between the formation of the mRNA and the functional protein, which includes the time for mRNA transportation from nucleus to cytoplasm, translation, protein folding, protein relocation, etc. Also, the time delay is assumed to be the same for both the long and short form transcript.

Based on the mass balance: *change of expression level = production rate – degradation rate*, the dynamics of the long and short form transcripts is modeled by the following differential-delay equations:

$$\begin{aligned} \frac{dl}{dt} &= I_0 - \alpha \frac{l^n(t-\tau_0)}{K^n + \varepsilon s^n(t-\tau_0) + l^n(t-\tau_0)} - d_l l, \\ \frac{ds}{dt} &= \alpha \frac{l^n(t-\tau_0)}{K^n + \varepsilon s^n(t-\tau_0) + l^n(t-\tau_0)} - d_s s. \end{aligned} \tag{6.2}$$

where d_l and d_s denote the degradation rates for the long and short form transcripts, respectively. It is also possible that for some transcription initiation, both poly(A)

sites are skipped and that there is no transcript generated, and this sometimes is called transcription abortion (Sousa et al. 1992; Rocha and Danchin 2003). The probability that the DPA is used will be assumed to depend on the concentration of the polyadenylation factors, characterized by the Hill equation with a different equilibrium constant than used by UPA. By slightly modifying the first equation of Equation (6.2), we obtain a more complete, but more complicated system:

$$\begin{aligned} \frac{dl}{dt} &= (I_0 - \alpha \frac{l^n(t-\tau_0)}{K^n + \epsilon s^n(t-\tau_0) + l^n(t-\tau_0)}) \frac{l^n(t-\tau_0)}{K_l^n + \epsilon s^n(t-\tau_0) + l^n(t-\tau_0)} - d_l l, \\ \frac{ds}{dt} &= \alpha \frac{l^n(t-\tau_0)}{K^n + \epsilon s^n(t-\tau_0) + l^n(t-\tau_0)} - d_s s. \end{aligned} \quad (6.3)$$

6.4 Theoretical Analysis of Differential-Delay Equation

The differential-delay equations (DDEs) are difficult to analyze mathematically in general. A DDE (also called a retarded functional differential equation, RFDE) is a special type of functional differential equation. It is very similar to an ordinary differential equation (ODE), but its solution involves past values of the state variable. Thus, the solution of a DDE requires knowledge of the current state, as well as the state at a certain earlier time. Therefore, the DDE is an infinite-dimensional system, so we need to specify an initial function for an initial time interval.

To understand the dynamical behavior of Equation (6.3), a simpler DDE will be considered. $K_l = 0$ and $\epsilon = 0$ mean that the short form protein has no regulatory function and that it generates either a long or short form transcript after the transcription has been initiated. In this case, Equation (6.3) is reduced to the following simpler one:

$$\begin{aligned}\frac{dl}{dt} &= I_0 - \alpha \frac{l^n(t-\tau_0)}{K^n + l^n(t-\tau_0)} - d_l l, \\ \frac{ds}{dt} &= \alpha \frac{l^n(t-\tau_0)}{K^n + l^n(t-\tau_0)} - d_s s.\end{aligned}\tag{6.4}$$

Note that the first equation is independent of the second equation, so the above system can be studied by the following single DDE:

$$\frac{dl}{dt} = I_0 - \alpha \frac{l^n(t-\tau_0)}{K^n + l^n(t-\tau_0)} - d_l l.\tag{6.5}$$

Equation (6.5) has six parameters $(I_0, \alpha, n, K, d_l, \tau_0)$, and it can be reduced to five parameters by the transformation:

$$\hat{l}(t) = \frac{l(t)}{K}.$$

The results is

$$\frac{d\hat{l}}{dt} = I_1 + \frac{\alpha_1}{1 + \hat{l}^n(t-\tau_0)} - d_l \hat{l}, \text{ where } I_1 = \frac{I_0 - \alpha}{K}, \alpha_1 = \frac{\alpha}{K}.$$

Dropping the $\hat{}$ on $l(t)$, we obtain:

$$\frac{dl}{dt} = I_1 + \frac{\alpha_1}{1 + l^n(t-\tau_0)} - d_l l.\tag{6.6}$$

If $I_l = 0$, then this becomes the same equation as that has been studied by Mackey and Glass (Mackey and Glass 1977), for which they obtained some results. In their report, the following two different equations were considered:

$$\frac{dP}{dt} = \frac{\beta_0 \theta^n}{\theta^n + P_\tau^n} - \gamma P,\tag{6.7}$$

$$\frac{dP}{dt} = \frac{\beta_0 \theta^n P_\tau^n}{\theta^n + P_\tau^n} - \gamma P,\tag{6.8}$$

where $P_\tau = P(t - \tau)$, β_0 , θ , γ are positive constants. They claimed that the solution of Equation (6.7) will either approach a fixed point or show a stable limit cycle oscillation, while the solution of Equation (6.8) can be chaotic.

In the following, Equation (6.5) will be studied in more detail and some conclusions will be made (see also, (Diekmann 1995)). Two questions can be addressed about Equation (6.5): does Equation (6.5) have a periodic solution, and what is the relationship between the period and the delay?

Equation (6.5) has only one steady state positive solution l_0 , which is independent of the time delay. If we set the right-hand side of Equation (6.5) to zero with zero delay, we obtain

$$I_0 - \alpha \frac{l^n}{K^n + l^n} - d_1 l = 0.$$

The equation can be written in another form:

$$I_0 - d_1 l = \alpha \frac{l^n}{K^n + l^n}.$$

The above equation can have only one positive solution since the line on the left side and the curve on the right side of the above equation can intersect only at one positive point l_0 . The stability of this steady state solution can be found by linearizing the non-linear term around the steady state. The stability analysis is much more complex where there is nonzero time delay. After linearization of Equation (6.5), the following linear DDE is obtained:

$$\frac{d\hat{x}}{dt} = \hat{\alpha}x(\hat{t}) + \hat{\beta}\hat{x}(\hat{t} - \tau_0) \quad (6.9)$$

where $\hat{x}(\hat{t})$ is the perturbation around the steady state solution l_0 and

$$\hat{\alpha} = -d_l, \quad \hat{\beta} = -\frac{\alpha_1 n l_0^{n-1}}{(1+l_0^n)^2}.$$

After scaling the time by setting $\hat{t} = \tau_0 t$, we get the following linear DDE, dropping the $\hat{\cdot}$ on $x(t)$:

$$\frac{dx}{dt} = \alpha x(t) + \beta x(t-1), \quad (6.10)$$

where $\alpha = \hat{\alpha}\tau_0 = -d_l\tau_0$, $\beta = \hat{\beta}\tau_0 = -\frac{\alpha_1 n l_0^{n-1}\tau_0}{(1+l_0^n)^2}$.

We look for the solution in the form $x(t) = e^{zt}$. After substitution, we obtain the characteristic equation of Equation (6.10):

$$z = \alpha + \beta e^{-z}. \quad (6.11)$$

Since z is an unknown number, write $z = \mu + i\nu$ and substitute this into the above characteristic equation, we obtain the following parameterized equations:

$$\begin{aligned} \alpha &= \mu + \frac{\nu \cos \nu}{\sin \nu}, \\ \beta &= -\frac{\nu}{\sin \nu} e^{\mu}. \end{aligned} \quad (6.12)$$

If $\mu = 0$, then Equation (6.11) has the pure imaginary solution and Equation (6.10) has an oscillatory solution. By setting $\mu = 0$, we get the following equations:

$$\begin{aligned} \alpha &= \frac{\nu \cos \nu}{\sin \nu}, \\ \beta &= -\frac{\nu}{\sin \nu}. \end{aligned} \quad (6.13)$$

The above equations are even in ν , so we can restrict $\nu \geq 0$. The above expression has singularities at $\nu = k\pi, k = 0, 1, 2, \dots$, and we can divide the right half axis of ν into intervals such that sine function has a single sign in each interval. Define the intervals:

$$I_k^- = ((2k-1)\pi, 2k\pi), \quad I_k^+ = (2k\pi, (2k+1)\pi)$$

and define the curves C_k^\pm in the (α, β) -plane parameterized by ν , which is varying within the interval I_k^\pm as:

$$C_k^\pm = \left\{ (\alpha, \beta) = \left(\frac{\nu \cos \nu}{\sin \nu}, -\frac{\nu}{\sin \nu} \right) \mid \nu \in I_k^\pm \right\}. \quad (6.14)$$

For $k=0$, $I_0^- = (-\pi, 0)$, $I_0^+ = (0, \pi)$, so $C_0^- = C_0^+ = C_0$.

We can plot all the curves of Equation (6.13) in the same (α, β) -plane by varying the parameter ν , and we get the parameter phase plane (see Figure 6.4).

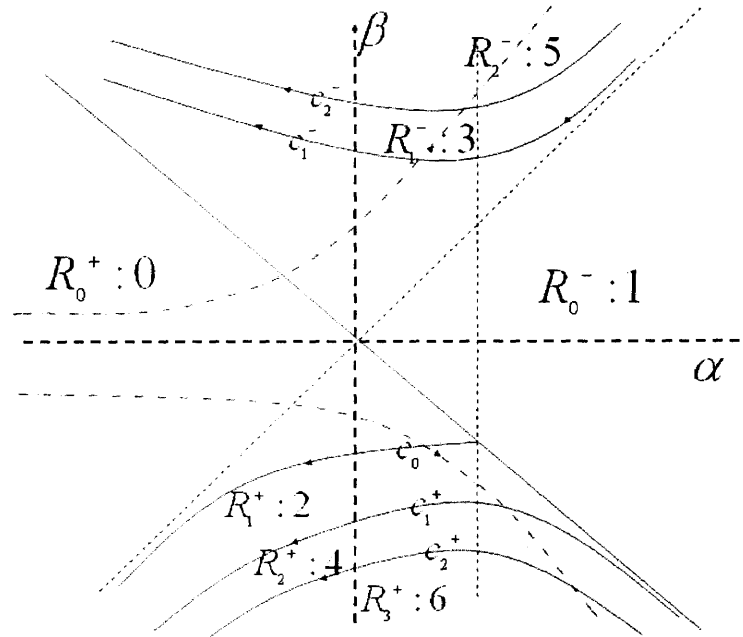


Figure 6.4 Parameter phase plane for the linear DDE. α and β are the coefficients in Equation (6.11). R_i^\pm is the region bounded by solid curves, and c_i^\pm is the solid curve generated by Equation (6.13). The numbers given after the colon represent the number of roots of Equation (6.11) located in the right half plane.

Some properties are hold for the above curves.

Lemma 6.1 For each curve C_k^\pm , as we vary the ν from the left end to the right end of the interval I_k^\pm , α is decreasing, β keeps the same sign, and the extreme value of β is at $\alpha = 1$.

Proof. From $\alpha(\nu) = \frac{\nu \cos \nu}{\sin \nu}$, taking a derivative with respect to ν , we get

$$\alpha'(\nu) = \frac{\frac{1}{2} \sin 2\nu - \nu}{\sin^2 \nu}.$$

Since $\sin 2\nu - 2\nu \leq 0$ for all $\nu \geq 0$, so α is decreasing. $\beta = -\frac{\nu}{\sin \nu}$ does not change the sign when ν changes within the interval I_k^\pm . $\beta > 0$ on the curve C_k^- and $\beta < 0$ on the curve C_k^+ . Also

$$\beta'(\nu) = -\frac{\sin \nu - \nu \cos \nu}{\sin^2 \nu}.$$

When $\sin \nu - \nu \cos \nu = 0$, we have $\alpha = 1$. \square

Lemma 6.2 All the curves C_k^\pm defined in Equation (6.14) do not intersect with each other and they are asymptotic to the lines $\alpha \pm \beta = 0$.

Proof. The curves C_k^+ and C_k^- do not intersect with each other since $\beta > 0$ on the curve C_k^- and $\beta < 0$ on the curve C_k^+ . To prove that there are no intersections among the curves C_k^+ or C_k^- , we can use contradiction. If there exist two distinct curves intersecting with each other, then from Equation (6.14), $\exists \nu_k \in I_k^\pm$, $\nu_j \in I_j^\pm$, $j \neq k$, we have the following:

$$\frac{\nu_k \cos \nu_k}{\sin \nu_k} = \frac{\nu_j \cos \nu_j}{\sin \nu_j},$$

$$\frac{\nu_k}{\sin \nu_k} = \frac{\nu_j}{\sin \nu_j},$$

that is,

$$\cos \nu_k = \cos \nu_j,$$

$$\frac{\nu_k}{\sin \nu_k} = \frac{\nu_j}{\sin \nu_j}.$$

Then we have $\nu_k = \nu_j$, which contradicted with the assumption that ν_k and ν_j belong to different intervals and these intervals never intersect. Since when $\nu \rightarrow k\pi^-$,

$\frac{\nu \cos \nu}{\sin \nu} \rightarrow \pm \frac{\nu}{\sin \nu}$, so the curves C_k^\pm are asymptotic to the lines $\alpha \pm \beta = 0$. \square

From Lemmas 6.1-2, we know that for each point (α, β) lying on the curve C_k^\pm , there exists a unique $\nu_0 \in I_k^\pm$, and the solution of Equation (6.11) would be $z = \pm \nu_0 i$

In order to determine the number of roots in the right-half plane in other regions other than on the curves, we need to fix ν at ν_0 and we have μ -curve:

$$\alpha = \mu + \frac{\nu_0 \cos \nu_0}{\sin \nu_0},$$

$$\beta = -\frac{\nu_0}{\sin \nu_0} e^\mu. \tag{6.15}$$

The tangent vector of the μ -curve is given by $v_2 = \begin{pmatrix} 1 \\ -\frac{\nu_0}{\sin \nu_0} \end{pmatrix}$ at ν_0 .

We have to compute the inner product of v_2 and a designated normal vector for C_k^\pm . The arrows along the solid curve C_k^\pm (see Figure 6.4) indicated the direction

of ν increasing. If $v_1 = (p, q)$ is a tangent vector to C_k^\pm , then $v_1^\perp = (-q, p)$ is the normal vector. A vector $w = (r, s)$ points to the left of the curve compared with the arrow direction if and only if $w \cdot v_1^\perp > 0$. The tangent vector of the curve C_k^\pm is given by:

$$v_1 = \left(\frac{\cos \nu_0 \sin \nu_0 - \nu_0}{\sin^2 \nu_0}, \frac{\nu_0 \cos \nu_0 - \sin \nu_0}{\sin^2 \nu_0} \right).$$

The normal vector is given by:

$$v_1^\perp = \left(-\frac{\nu_0 \cos \nu_0 - \sin \nu_0}{\sin^2 \nu_0}, \frac{\cos \nu_0 \sin \nu_0 - \nu_0}{\sin^2 \nu_0} \right).$$

So whether the vector v_2 points to the left or the right depends on the sign of

$$v_2 \cdot v_1^\perp = \frac{\sin^2 \nu_0 - 2\nu_0 \cos \nu_0 \sin \nu_0 + \nu_0^2}{\sin^3 \nu_0}.$$

It's true that

$$h(\nu_0) = \sin^2 \nu_0 - 2\nu_0 \cos \nu_0 \sin \nu_0 + \nu_0^2 > 0$$

for $\nu_0 > 0$, since $h(0) = 0$ and $h'(\nu) = 2\nu(1 - \cos 2\nu) > 0$ for $\nu \in I_k^\pm$. The sign of $v_2 \cdot v_1^\perp$ doesn't change along the curve C_k^\pm since $\sin \nu_0$ doesn't change the sign along the curve. That is, if we fix ν at ν_0 on the curve C_k^+ or C_0 , and increase the μ , from above we know that $v_2 \cdot v_1^\perp > 0$, and so the μ -curve will go to the left of the curve. Using the same arguments, if we fix ν at ν_0 on the curve C_k^- and increase the μ , from above we know that $v_2 \cdot v_1^\perp < 0$ and so the μ -curve will go to the right of the curve.

We can conclude that, when moving away from C_k^+ or C_0 to the left or moving away from C_k^- to the right, the critical roots located on the pure imaginary axis move into the right half-plane.

Lemma 6.3 The μ -curve will intersect each of the C_k^- or C_k^+ curves in at least one point when we vary μ from $-\infty$ to ∞ .

Proof. From Equation (6.15), the μ -curve is given by the following exponential function:

$$\beta = \left(-\frac{\nu_0}{\sin \nu_0} e^{-\frac{\nu_0 \cos \nu_0}{\sin \nu_0}} \right) e^\alpha \quad (6.16)$$

If we set

$$f(\nu) = -\frac{\nu}{\sin \nu} e^{-\frac{\nu \cos \nu}{\sin \nu}},$$

it can be seen that $f(\nu)$ is a continuous function in the interval I_k^\pm . Now look at the limit of the function $f(\nu)$ as ν approaches the end points of the intervals:

$$\lim_{\nu \rightarrow 2k\pi^+} f(\nu) = \lim_{\nu \rightarrow 2k\pi^+} \left(-\frac{\nu}{\sin \nu} e^{-\frac{\nu \cos \nu}{\sin \nu}} \right) = \lim_{\nu \rightarrow 2k\pi^+} \left(-\frac{\nu}{\sin \nu} e^{-\frac{\nu}{\sin \nu}} \right) = \lim_{x \rightarrow -\infty} (xe^x) = 0$$

$$\lim_{\nu \rightarrow 2k\pi^-} f(\nu) = \lim_{\nu \rightarrow 2k\pi^-} \left(-\frac{\nu}{\sin \nu} e^{-\frac{\nu \cos \nu}{\sin \nu}} \right) = \lim_{\nu \rightarrow 2k\pi^-} \left(-\frac{\nu}{\sin \nu} e^{-\frac{\nu}{\sin \nu}} \right) = \lim_{x \rightarrow \infty} (xe^x) = \infty$$

$$\lim_{\nu \rightarrow (2k-1)\pi^+} f(\nu) = \lim_{\nu \rightarrow (2k-1)\pi^+} \left(-\frac{\nu}{\sin \nu} e^{-\frac{\nu \cos \nu}{\sin \nu}} \right) = \lim_{\nu \rightarrow (2k-1)\pi^+} \left(-\frac{\nu}{\sin \nu} e^{-\frac{\nu}{\sin \nu}} \right) = \lim_{x \rightarrow \infty} (xe^{-x}) = 0$$

$$\lim_{\nu \rightarrow (2k-1)\pi^-} f(\nu) = \lim_{\nu \rightarrow (2k-1)\pi^-} \left(-\frac{\nu}{\sin \nu} e^{-\frac{\nu \cos \nu}{\sin \nu}} \right) = \lim_{\nu \rightarrow (2k-1)\pi^-} \left(-\frac{\nu}{\sin \nu} e^{-\frac{\nu}{\sin \nu}} \right) = \lim_{x \rightarrow -\infty} (xe^{-x}) = -\infty$$

That is, when $\nu \in I_k^- = ((2k-1)\pi, 2k\pi)$, $f(\nu)$ takes the range from 0 to ∞ continuously; when $\nu \in I_k^+ = (2k\pi, (2k+1)\pi)$, $f(\nu)$ takes the range from $-\infty$ to 0

continuously. Since the curve C_k^\pm approach the diagonal line, the exponential function must intersect with C_k^\pm . \square

From above, we can see that, for some value $v_k \in I_k^-$, we can find the μ -curve from Equation (6.16). There must exist a different value $v_j \in I_j^-$ such that the μ -curve corresponding to v_j is exactly the same with the μ -curve corresponding to v_k because of the properties of the function $f(v)$ described in Lemma 6.3. The number of roots in the right-half plane changes when the parameter pair crosses C_k^\pm or R as you follow the μ -curve. For a parameter pair (α, β) in the open region R_k^\pm bounded by the bold lines and curves, the number of the roots located in the right-half plane is given by the numbers following the colons in the Figure 6.4. If we trace back along the μ -curve, we get an infinite number of roots located in the left-half plane except along the α -axis. Since the μ -curve has no intersection with the α -axis, the number of roots on the α -axis can not be counted by this procedure. The α -axis corresponds to $\beta = 0$ and the characteristic equation has only one root at $z = \alpha$.

In conclusion, for Equation (6.11), there are two pure imaginary roots for parameters on the curves C_k^\pm ; the number of roots in the right-half plane for parameters on the curves C_k^- is the same as in the open region R_{k-1}^- ; and the number of roots in the right-half plane for parameters on the curves C_k^+ is the same as in the open region R_k^+ . The number of roots in the left-half plane is infinite except along the α -axis.

As to why the number of roots in the right half-plane is constant in the region bounded by the curves C_k^+ and the line R, a theorem described in the book (Diekmann 1995) says the following:

Theorem 6.1 (Continuity of the roots of an equation as a function of parameters) Let Ω be an open set in \mathbb{C} , F a continuous complex-valued function on $\mathbb{R} \times \mathbb{R} \times \Omega$ such that, for each $(\alpha, \beta), z \mapsto F(\alpha, \beta, z)$ is analytic in Ω . Let ω be an open subset of Ω whose closure $\bar{\omega}$ in \mathbb{C} is compact and contained in Ω . Let α_0, β_0 be such that no zero of $F(\alpha_0, \beta_0, z)$ is on the boundary of ω . Then there exists a neighbourhood U of (α_0, β_0) in $\mathbb{R} \times \mathbb{R}$ such that:

- for any $(\alpha, \beta) \in U, F(\alpha, \beta, z)$ has no zeros on the boundary of ω ;
- the number of zeros of $F(\alpha, \beta, z)$ in ω , taking multiplicities into account, is constant for $(\alpha, \beta) \in U$. \square

Thus the parameters that make the linear DDE have periodic solutions must be located on the curves C_k^\pm . From the definition of the parameters from Equation (6.10):

$$\alpha = \hat{\alpha}\tau_0 = -d_l\tau_0, \quad \beta = \hat{\beta}\tau_0 = -\frac{\alpha_1 n l_0^{n-1} \tau_0}{(1+l_0^n)^2},$$

we know that $\alpha < 0, \beta < 0$. That is, for Equation (6.5), the parameter point (α, β) is located on the III quadrant, in which the ν ranges in the interval

$(\pi + 2k\pi) > \nu > (\frac{\pi}{2} + 2k\pi)$. Thus, we have the following conclusion:

Lemma 6.4 The period T of the periodic solution of Equation (6.5) must satisfy $2\tau_0 < T < 4\tau_0$.

Proof. From above, we know that the periodic solution has the form $e^{i\nu t}$ and

$(\pi + 2k\pi) > \nu > (\frac{\pi}{2} + 2k\pi)$. The period of the solution is $T' = \frac{2\pi}{\nu'}$, where $\pi > \nu > \frac{\pi}{2}$,

so we have $2 < T' < 4$. Since we scale the time by the transformation: $\hat{t} = \tau_0 t$, the period T of the original system with delay τ_0 will satisfy $2\tau_0 < T < 4\tau_0$. \square

Lemma 6.4 gives us a way to estimate the parameter τ_0 .

6.5 Numerical Simulations

Differential-delay equations are difficult to solve analytically (Murray 1989), and many results have been obtained about the linear and nonlinear DDE, such as stability analysis, the existence of periodic solutions, the bifurcation analysis, etc (Heiden and Mackey 1982; Losson Jm et al. 1993; Mackey and Nechaeva 1994). If all the nine parameters $I_0, n, \tau_0, \alpha, K, d_l, d_s, \varepsilon$ and K_l are given, we can solve the DDE equation (6.3) by a numerical method.

A lot of software packages have been developed to solve the DDE, such as DDE-BIFTOOL written in Matlab (Engelborghs et al. 2002), DDEFIT written in C/C++ (Wood 2003), DKLAG6 written in Fortran 77 (Corwin et al. 1997), and dde23 (Shampine 2001) written in Matlab. The latter two are used most often and dde23 is one part of Matlab (version 6.0 or later), so they are used in this study. The package dde23 in Matlab (running in windows, Pentium M 1.7G CPU, 1G RAM) is slower than DKLAG6 (running in Linux) with the same computer settings. Dde23 took 0.037s to compute the solution with the time duration of 100 hr for Equation (6.3), while DKLAG6 only took 0.0011s, 30 times faster than dde23.

Since the initial function (IF) is very difficult to determine biologically, we use a constant initial function (CIF) in this work. If not stated otherwise, the CIF of the DDE is always: $l(t) = s(t) = 1, -\tau_0 \leq t \leq 0$.

The solutions of Equation (6.3) could be very complex for some set of parameters. Different types of solutions are possible, such as steady-state solutions, periodic solutions, and chaotic solutions. One solution is shown in Figure 6.5 for randomly selected parameter set.

$$I_0 = 0.3987, \alpha = 0.3584, K = 0.2853, K_I = 0.1426,$$

$$d_l = 0.6264, d_s = 0.2412, \varepsilon = 0, \tau_0 = 8.2579, n = 8.6864$$

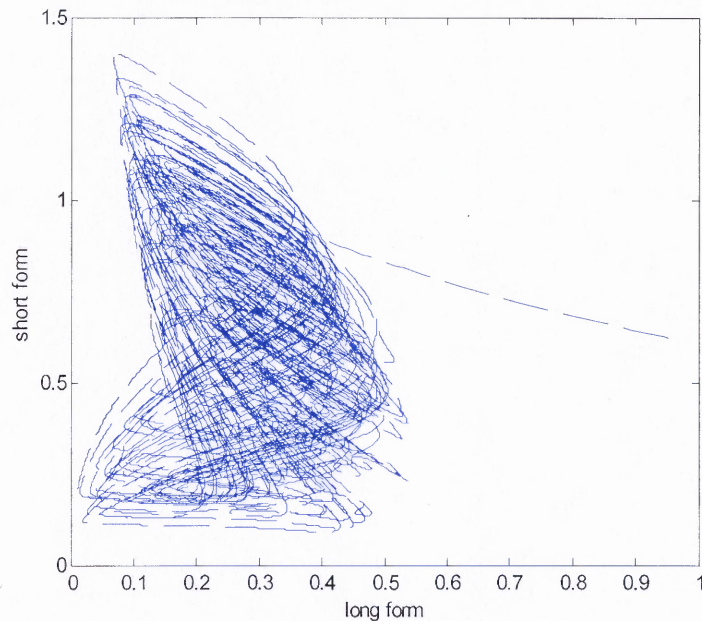


Figure 6.5 The phase plane of the long and short form transcripts. The solutions of Equation (6.3) are obtained for randomly selected parameter values given at the top of the plot. The constant initial functions for the long and short form are set equal to 1.

6.6 Parameter Estimation

Some of the nine parameters in Equation (6.3) have been chosen according to the following considerations. For the transcription initiation rate, without loss of generality, we set $I_0 = 1$. The parameter α , the maximum transcription rate of the short form, needs to be less than the transcription initiation rate 1, and it is to be determined. The Hill coefficient representing the cooperative binding process is difficult to determine. In the modified Goodwin oscillator, Griffith (Griffith 1968(b))

found stable limit cycles only with a Hill coefficient of eight or more. In Scheper et al. (Scheper et al. 1999), the Hill coefficient varies from 1 to 8; in Xiong et al. (Xiong and Ferrell 2003), they use 5 as the Hill coefficient; in Goldbeter (Goldbeter 1995), the Hill coefficient is 4; and in Elowitz et al. (Elowitz and Leibler 2000), the Hill coefficient is 2. Since for different activator and inhibitor substrates, the Hill coefficients are different, we chose a Hill coefficient of 4 in this study.

A discrete delay has been introduced to study gene regulation. Monk (Monk 2003) used a transcriptional delay around 15~20 minutes to study the oscillatory expression of the transcription factors Hes1, p53, and NF- κ B. Lewis et al. (Lewis et al. 1997) included a delay of 8 hr of the period protein (PER) to study the circadian rhythms. From the previous theoretical analysis, we know that if the periodic solution exists, the period T of the periodic solution of the linear DDE equation (6.10) is between two to four times the delay time. This is also true for Equation (6.3) if the system has a periodic solution resulting from a random simulation (see Figure 6.6). It has been found that the protein production rates fluctuate over a time scale of about one cell cycle, and cells about to divide produced on average twice as much protein per unit of time as newly divided cells (Rosenfeld et al. 2005). This motivates us to estimate that the period of the protein production oscillation is directly related to the cell cycle. For a typical mammalian cell, the cell cycle lasts about 24 hr. Thus, it is natural to estimate the time delay to be 8 hr, one third of the cell cycle period.

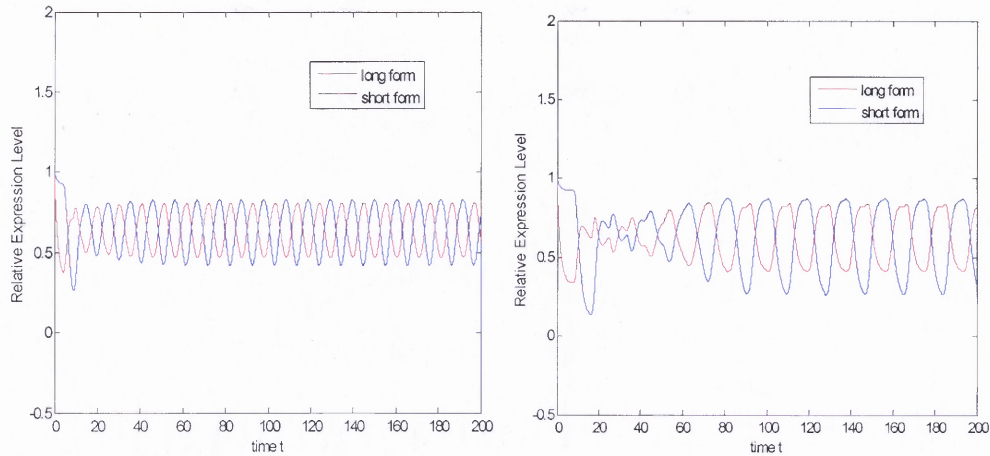


Figure 6.6 Relationship between the delay time and the period of Equation (6.3). The parameters for these two plots are $I_0 = 1, n = 4, \alpha = 0.8462, K = 0.5252, d_l = 0.6721, d_s = 0.8381, \varepsilon = 0.019, K_l = 0.2026$. The delay time is 4 for the left plot and 8 for the right plot. The period is ~ 11 for the left plot and ~ 23 for the right plot. The constant initial functions for the long and short form are set equal to 1.

To solve Equation (6.3), we have to estimate the remaining six parameters $\alpha, K, d_l, d_s, \varepsilon, K_l$. Generally, these parameters are difficult to estimate by biological experiments, thus, in-silicon estimation comes into play based on the experimental data. An optimization method will facilitate the estimation procedure. Very little information is known about the magnitude of these parameters, so we assume that all of the six parameters are located between 0 and 1. To find an optimal parameter set in the six dimensional parameter space, we use a global searching method.

The parameters in the model equations can be estimated by minimizing the error between the experimental data and the simulation data. Assume that the experimental data contains m time points at $[t_1, \dots, t_m]$ for both long and short form mRNA expression levels, i.e., $E = [l_1, \dots, l_m; s_1, \dots, s_m]$. For any given set of parameters in the parameter space, the DDE system (6.3) can be solved and evaluated at these m

time points with a CIF, i.e., $N(\alpha, K, d_l, d_s, \varepsilon, K_l) = [\tilde{l}_1, \dots, \tilde{l}_m; \tilde{s}_1, \dots, \tilde{s}_m]$. The mean

square error is calculated by: $err = \sqrt{\frac{\sum_{i=1}^m (l_i - \tilde{l}_i)^2 + (s_i - \tilde{s}_i)^2}{m}}$. We want to find an

optimum parameter set to minimize the error. If the error is less than some tolerance, say 0.01, we say that the parameter set is a good approximation. But, in fact, we can never achieve this tolerance because of the noise in the experimental data. Since the parameter space is continuous, it is impossible to evaluate a solution at every single point. So instead, we can evaluate the error at uniformly distributed points in the space to approximate the error.

If we uniformly divide each parameter space into p points, by estimating all possible combinations of the parameter values, we can find the global minimal error by solving the DDE equations p^6 times. By using the DKLAGE6 package, it will take $0.0011 \times p^6$ seconds. For example, if $p=50$, i.e., we search the space with a step of 0.02, it will take half a year to finish the search. To accelerate the computation, the message passing interface (MPI) for parallel computing is adopted. We use the Hydra cluster in the Department of Mathematical Sciences at NJIT, which was funded by a grant from NSF to do the computation. Each node contains 2 AMD Opteron 250 CPUs and 4GB of RAM. If we use 20 nodes from Hydra, then it will take one week to finish the search with a step of 0.02 in every parameter space. To further accelerate the computation, we also assume the degradation rates for the long and short form are the same, i.e., $d_l = d_s$. In this case, it will take 2 hours to complete the job for $p=50$, with 20 nodes. We found that this was acceptable, and we can use this framework in the future if we have some experimental data.

6.7 Results

6.7.1 Experimental Data

The following paragraph is provided by Zhenhua Pan from Dr. Bin Tian's lab.

HeLa cells were seeded at 70% confluence in 12-well plate in DMEM medium supplemented with 10% FBS. Total cellular RNAs were extracted using the RNeasy kit (Qiagen) according to manufacturer's protocols. mRNAs were reverse-transcribed using oligo-dT primers (Promega). Real-time quantitative PCR was carried out using the 7500 Real time PCR system (Applied Biosystems) with Syber-Green I as dye. The following primers were used for detecting various regions of the CstF-77 gene: F0 (5'-GAGGCCATGTCAGGAGAC), R1 (5'-GCAACTCCAAAATGCAACAA), R3 (5'-CATAAATCAATGTGCAAACC). Primers for Cyclophilin A (CYPH) were 5'-ATGGTCAACCCCACCGTGT and 5'-TTCCTGCTGTCTTTGGAACCTTGTC. Five time point data were obtained at 16, 24, 36, 48, 56 hr, and the relative expression level is shown in Figure 6.7.

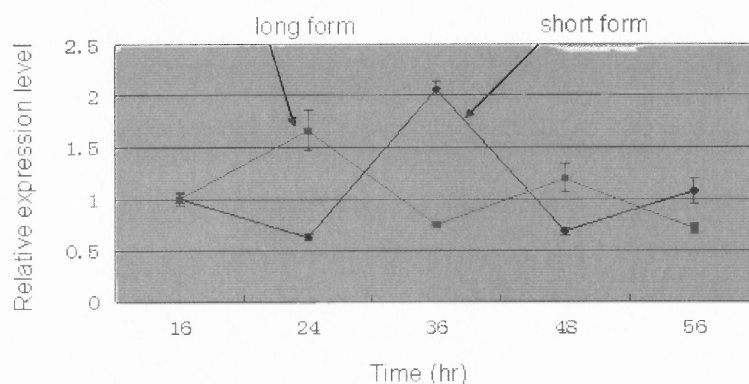


Figure 6.7 Experimental data at five time points. The x-axis is the time in hours and the y-axis is the relative expression level normalized to the first time point at 16 hr.

6.7.2 Optimal Parameter Set

By using the method described in Section 6.5 and the experiment data provided by Dr. Bin Tian's lab, the optimal parameter set we obtained was:

$$\alpha = 0.88, K = 0.98, d_l = d_s = 0.44, \varepsilon = 0.02, K_l = 0.10,$$

which led to the smallest error of $err = 0.4319$ for CIF equal to 1. The simulated results and the experimental results are plotted and compared (see Figure 6.8). The simulated oscillation fits the experimental oscillation well, but not perfectly due to the noise in the experimental data and some unknown factors.

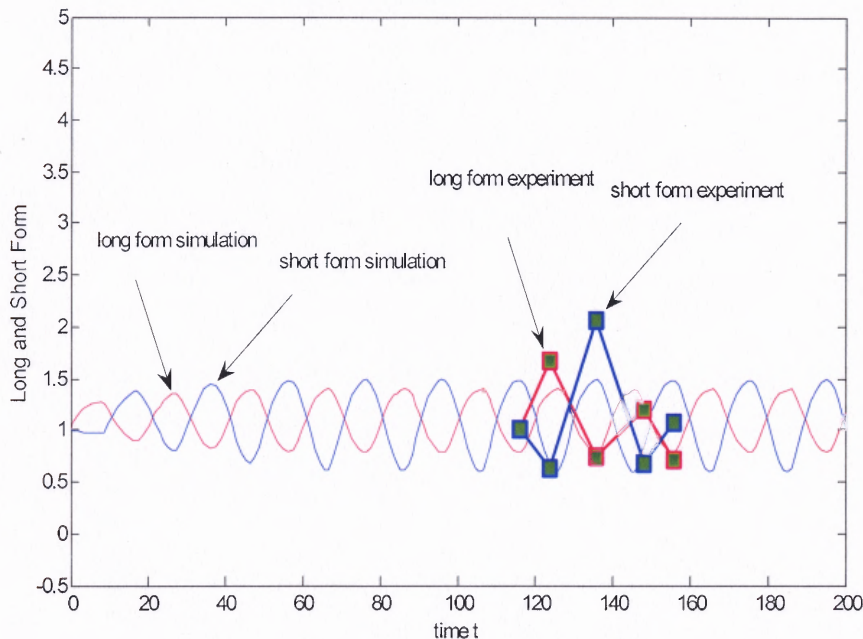


Figure 6.8 Solutions fitted with optimal parameter set. The optimal parameter set is obtained by the minimization of the error between the experimental data and the simulation data. The experimental data is shifted 100 hr to the right for better representation.

Except the obtained optimal parameter set and the smallest error, different combinations of parameters with errors less than 0.5 are saved. The distributions of these parameters are plotted (see Figure 6.9). We can see that parameters α , K , d_l , d_s , ε are more sensitive than K_l for the errors. Several observations can be obtained from the numerical simulations:

- The parameter α is less than 1, but close to 1, which is what we expected since the maximum transcription rate is less than the transcription initiation rate. If the long form protein is over expressed, the cell would generate more short form transcripts, and the transcription rate of the short form approaches

the transcription initiation rate so that the cell can control the long form production very efficiently.

- K_l is less than K . This is because there are two poly(A) sites located in the last exon to generate the long form (Pan et al. 2006). Two poly(A) sites in the last exon are also found in other species, e.g., *Drosophila melanogaster* and *Drosophila Virilis* (Audibert and Simonelig 1998). That is, if the long and short form transcripts have the same opportunity for polyadenylation, the probability for generating the long form is higher than for the short form.
- The parameter ε is very small, which means that the short form has a small inhibitory function or no function at all. The CstF-77.S protein has not been detected yet (private communication with Dr. Bin Tian). If expressed, from the putative protein sequence, it would only contain the first HAT domain, and lack the C-terminal region responsible for interacting with CstF-64, CstF-50, and itself (Takagaki and Manley 2000). It is not clear if it would bind to CPSF-160 or the CTD since the protein regions for these functions have yet to be mapped.

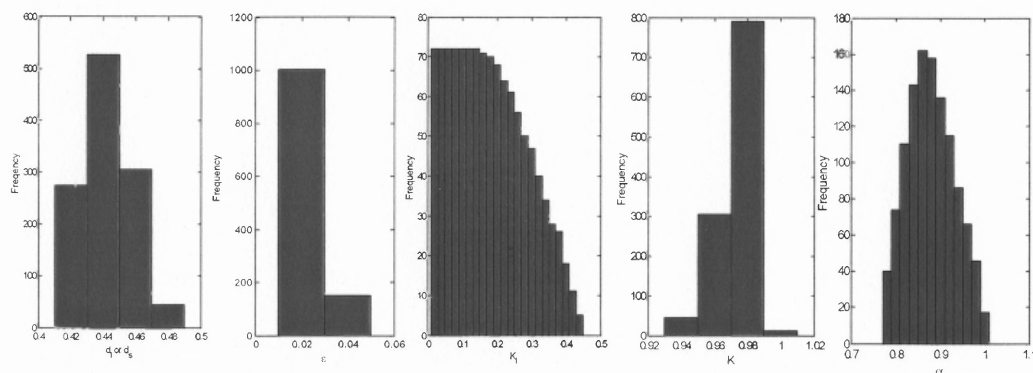


Figure 6.9 Distribution of parameters with error less than 0.5. The parameter values are given on the x-axis, and the frequency data for each parameter are given on the y-axis. d_1 and d_s range from 0.42 to 0.46; ε ranges from 0.02 to 0.04; K_l ranges from 0 to 0.44; K ranges from 0.94 to 1; α ranges from 0.78 to 1.

6.7.3 Different Initial Functions

The CIF $l(t) = s(t) = 1, -\tau_0 \leq t \leq 0$ is used for the unperturbed system. In fact, the initial conditions are arbitrary and are hard to determine based on the biological experiments. Most biological systems are very robust, for example, proteins can tolerate thousands of amino acid changes, metabolic networks continue to sustain life even after removal of importance chemical reactions, gene regulation networks

continue to function after alteration of key gene interactions, etc. (Wagner 2005). In terms of robustness, it means the system will continue to function in the face of perturbations.

Some experiments have been done to knock down or reduce the long and short form gene expression levels (unpublished data from Dr. Bin Tian's lab) and this corresponds to the simulated model with different initial conditions. Knock-down is associated with a technique to reduce the gene expression levels, but not reduce the gene expression levels to a very low level or zero, which is called knock-out. RNA interference (RNAi) is a very common know-down technique. If we knock down the long form, then the initial condition for $l(t)$ needs to be set to a smaller value; on the other hand, if we knock down the short form, then the initial condition for $s(t)$ needs to be set to a smaller value. If we knock out the long form, the initial condition for $l(t)$ needs to be set to a very small value or zero. We tried different CIFs and simulated the consequential "phenotype". If not specified, the CIF would be the constant 1.

If the long form is knocked out, i.e., the long form CIF less than 0.0653, then the solutions go to zero (see Figure 6.10). This implies that without the long form protein, the cell will die since the long form protein is essential for the cell to survive. It has been verified by experiment that if the Suppressor of forked ($Su(f)$), a *Drosophila* homologue of CstF-77 is knocked out, the *drosophila* will die at the larvae stage (Simonelig et al. 1996).

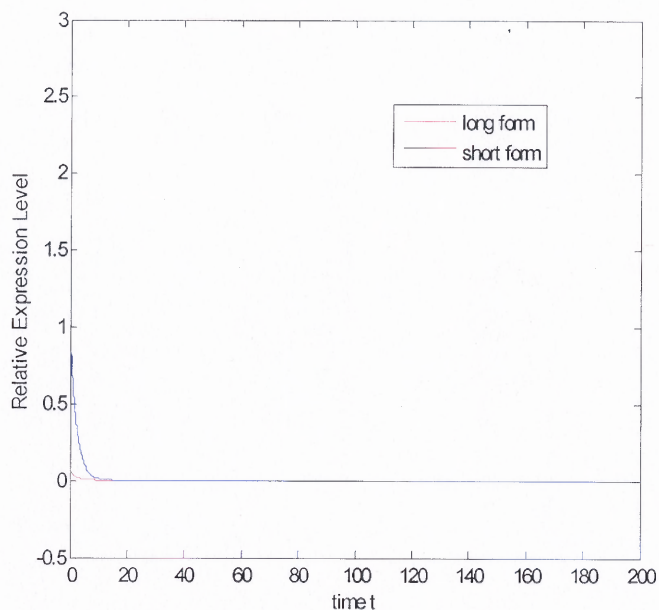


Figure 6.10 Solution of Equation (6.3) with long form CIF set to be 0.0652. CIF, constant initial function. The short form CIF is set to 1.

If we just knock down the long form, i.e., the long form CIF greater than 0.0652, the cell can recover very slowly and then survive (see Figure 6.11). That means, knocking down the long form would not kill the cell, but instead the small amount of long form protein would rescue the cell by itself.

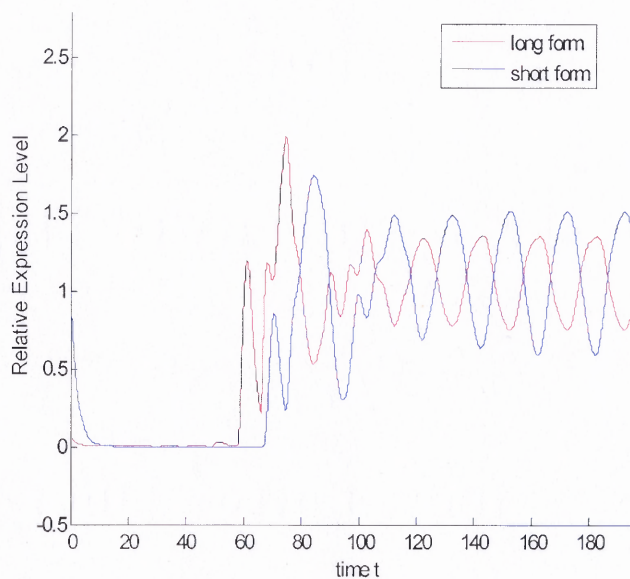


Figure 6.11 Solution of Equation (6.3) with long form CIF set to be 0.0653. CIF, constant initial function. The short form CIF is set to 1.

If the long form is over expressed, i.e., the long form CIF is greater than 5, the system will return to an oscillatory solution (see Figure 6.12). Overexpressing or knocking down the short form does not affect the system (see Figure 6.13-14).

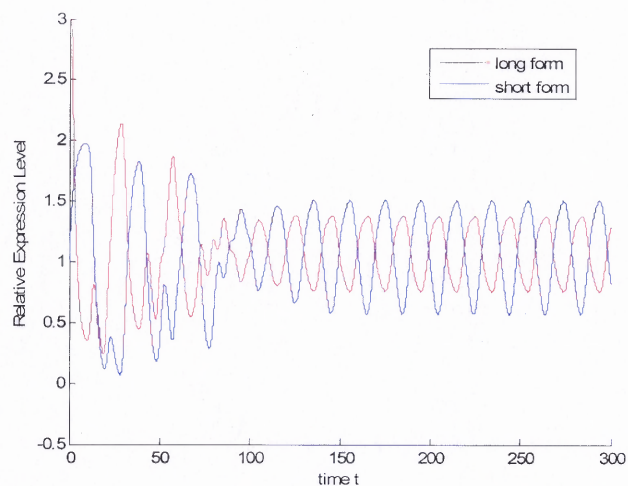


Figure 6.12 Solution of Equation (6.3) with long form CIF set to be 5. CIF, constant initial function. The short form CIF is set to 1.

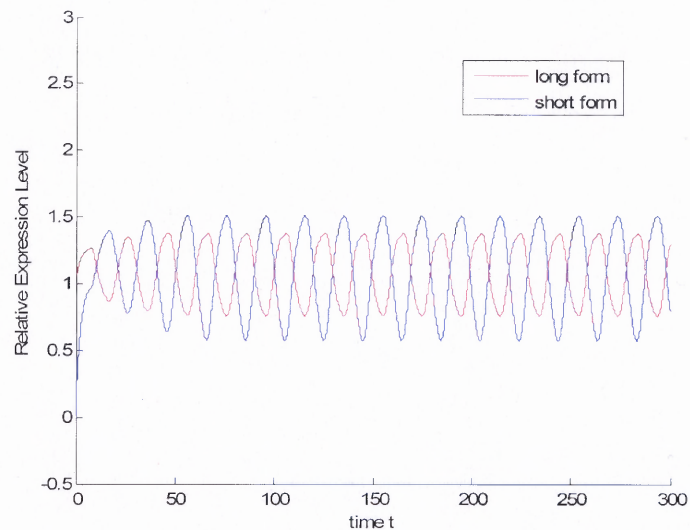


Figure 6.13 Solution of Equation (6.3) with short form CIF set to be 0.01. CIF, constant initial function. The long form CIF is set to 1.

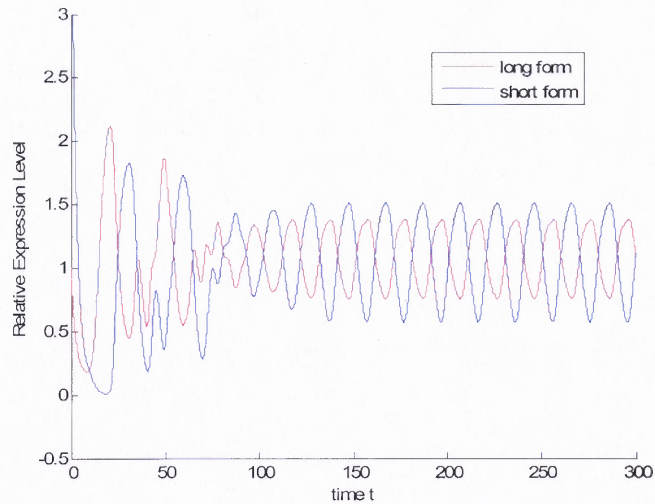


Figure 6.14 Solution of Equation (6.3) with short form CIF set to be 5. CIF, constant initial function. The long form CIF is set to 1.

Also from Figure 6.10-14, we can see that, when the long form expression level decreases, the short form expression level also decreases, and when the long form expression level increases, the short form expression level also increases in the first several hours. This agrees with the results that the lack of Su(f) function is correlated with the disappearance of the short form Su(f) RNA, and accumulation of the short form requires the wild-type Su(f) protein (Audibert and Simonelig 1998).

6.7.4 Basin of Attraction and Basin Boundary

The set formed by all the initial points in the phase space of a dynamical system which is attracted to a given solution (e.g., a fixed point, a limit cycle) is called the basin of attraction for that solution. Many nonlinear dynamical systems possess multi-stable solutions, and different initial conditions, belonging to different basin of attraction will be attracted to the corresponding solutions (Losson Jm et al. 1993). The boundaries of the various basins of attraction are known as basin boundaries. The

dependence of solution behavior on the initial conditions in certain first-order nonlinear DDEs has been investigated (Losson Jm et al. 1993).

We notice that Figure 6.11-14 have the same periodic solutions (referred to as the limit cycle solution) for different initial conditions by numerical simulations (see Figure 6.15). We hypothesize that this is a limit cycle. The limit cycle solution is neutral stable as shown by numerical simulations, with the solution attracted to the limit cycle with different CIFs (see Figure 6.15).

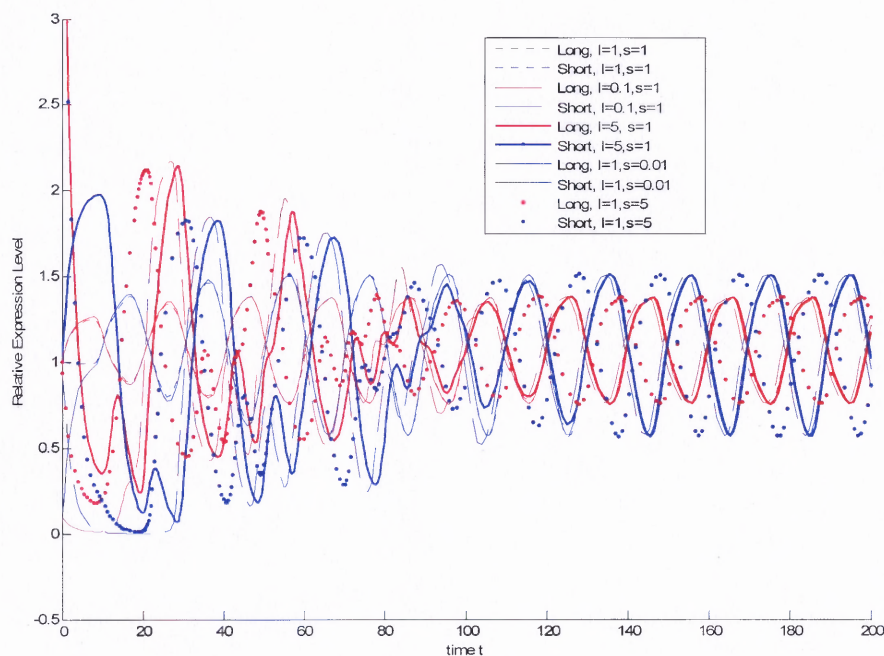


Figure 6.15 Solutions of Equation (6.3) with different CIFs. CIF, constant initial function. For different CIFs, the solutions go to the same limit cycle solution.

From Figure 6.10-11, we found that there is a bifurcation point of the long form CIF if the short form CIF is 1. That is, if the long form CIF is greater than 0.0652, the system returns to the limit cycle solution, otherwise, it will go to zero. For any given short form CIF, different CIFs of long form are tried and the Equation (6.3) is solved based on the CIFs. The CIF for long form was found such that with a little decreasing of this value, the solution goes to zero, while with a little increasing of this

value, the solution goes to the limit cycle solution. For different short form CIF, the bifurcation points of the long form CIF can be generated numerically and they are plotted in Figure 6.16. The curve is called the basin boundary, separating the phase space into two regions: zero region and oscillatory region. If the CIF starts in the zero region, the solutions of the Equation (6.3) are attracted to the origin. If the CIF starts in the oscillatory region, the solutions go to the limit cycle solution.

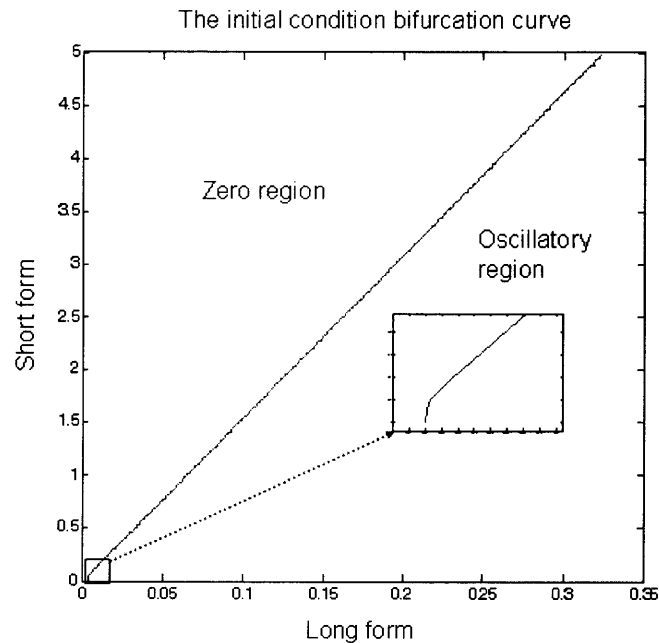


Figure 6.16 Basin boundary of CIF. CIF, constant initial function. The boundary is found numerically such that with any given short form CIF, a little increase of long form CIF would lead to an oscillatory solution in the oscillatory region. A little decrease of long form CIF would lead asymptotically to the zero solution. The basin boundary around zero is a vertical line indicated in the inserted plot.

We also found two interesting phenomena about the basin boundary: when the CIFs are different from zero, the basin boundary is almost a straight line with the ratio (long/short) approaching 0.064 as the CIF increases; when the CIFs are around zero, the basin boundary becomes a vertical line (see the amplification plot in Figure 6.16), which means that, if the short form CIF is very small, then to have the limit cycle solution, the long form constant IF needs to greater than 0.00413. Biologically, this

means that there could be no short form protein, but a small amount of long form protein is essential to rescue the cell from certain death.

In fact, it is easy to see that zero is a fixed point of Equation (6.3). By linearizing the system around zero, we can show that zero is a stable fixed point. In fact, when Equation (6.3) is linearized around zero, the linearized equation has no delay term and becomes the following ODEs:

$$\begin{aligned}\frac{d\hat{l}}{dt} &= -d_l \hat{l}, \\ \frac{d\hat{s}}{dt} &= -d_s \hat{s}.\end{aligned}$$

The equation has steady state solution at zero, and so zero is a stable fixed point of Equation (6.3).

Why is the basin boundary a vertical line near zero?

If the right side of the first equation of Equation (6.3)

$$f(l) = \left(I_0 - \alpha \frac{l^n}{K^n + l^n}\right) \frac{l^n}{K_l^n + l^n} - d_l l \quad (6.17)$$

is negative with $s \rightarrow 0$, the system will go to zero. Set $f(l) = 0$ and solve for l near the zero. It is very difficult to solve the $f(l)=0$ analytically. So we plot Equation (6.17) first (see Figure 6.17). We found that except the zero, Equation (6.17) has another small root around 0.005. By using the function `fsolve` in Matlab with the initial condition 0.005 to solve Equation (6.17), we get $l_0 = 0.0041316256293$ and $f'(l_0) > 0$. This result agrees very well with the numerical finding that the long form CIF needs to be greater than 0.00413 for generating the oscillatory solution.

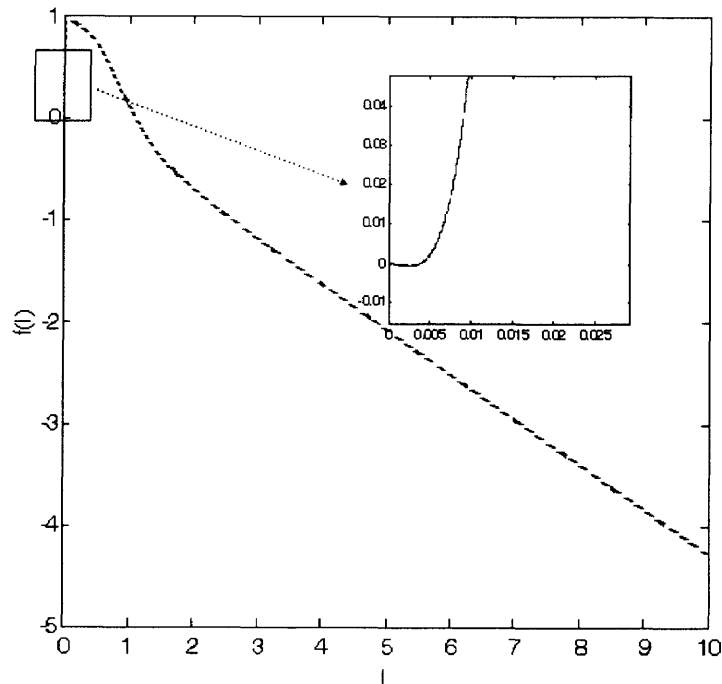


Figure 6.17 Plot of Equation (6.17). Equation (6.17) has two positive roots: one is close to 1 and the other is very close to 0. The small roots can be seen more clearly in the inserted plot.

It is still not clear why the basin boundary approaches the line with ratio (long/short) around 0.064 when the CIFs are not near zero. For three different initial conditions of the short form, $s_0 = 0.01, 1$ and 100 , we calculate the long form bifurcation points of the CIF with accuracy to 11 digits, $l_0 = 0.0041328205677, 0.065228128392$ and 6.4525236006 , respectively (see Figure 6.18-20). With the long and short form CIF (s_0, l_0) , the solutions go to zero. When the last digits of l_0 increase by one, the solutions go to the limit cycle solution.

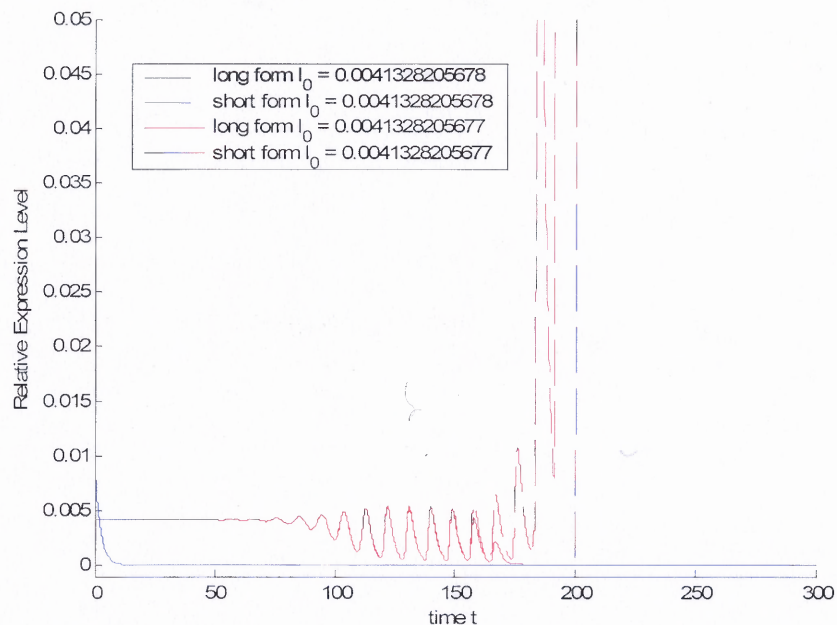


Figure 6.18 The bifurcation solutions with short form CIF equal to 0.01. CIF, constant initial function. The long form CIF bifurcation point with accuracy to 11 digits is generated so that the solutions go to the limit cycle solution when the last digit increases by 1.

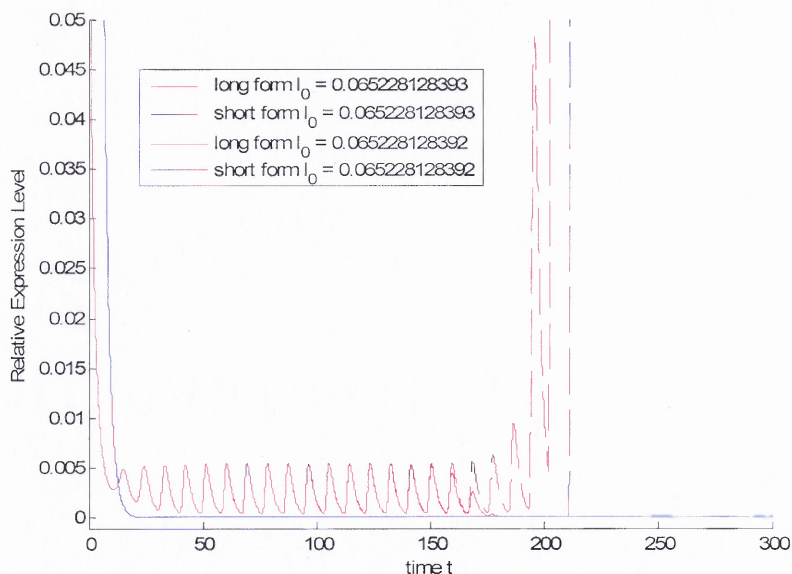


Figure 6.19 The bifurcation solutions with short form CIF equal to 1.00. CIF, constant initial function. The long form CIF bifurcation point with accuracy to 11 digits is generated so that the solutions go to the limit cycle solution when the last digit increases by 1.

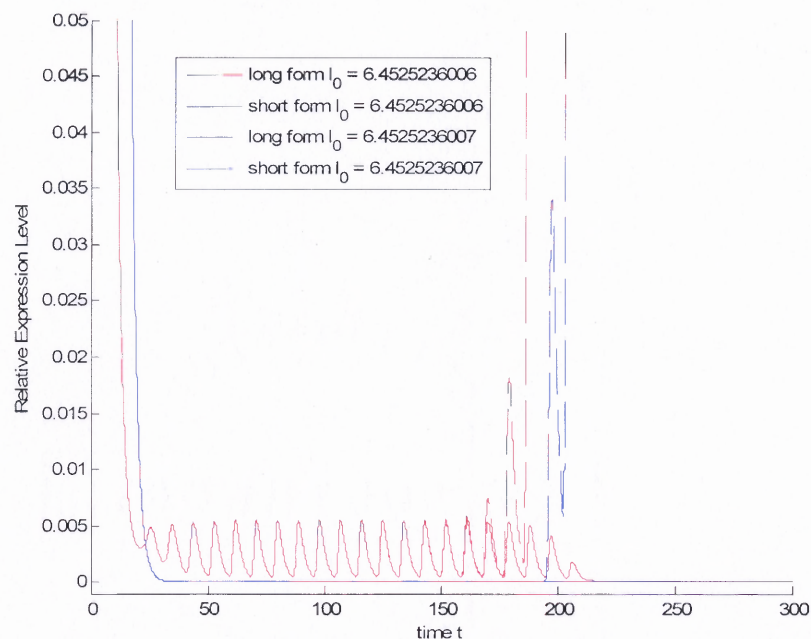


Figure 6.20 The bifurcation solutions with short form CIF equal to 100. CIF, constant initial function. The long form CIF bifurcation point with accuracy to 11 digits is generated so that the solutions go to the limit cycle solution when the last digit increases by 1.

From the observation shown in Figure 6.21 (three solutions in one plot), we hypothesize that for given short form CIF s^* , there exists a long form CIF l^* , such that with the CIF (l^*, s^*) , the solution of Equation (6.3) would be a long form oscillation at very small magnitude and the short form goes to zero exponentially. Furthermore, with different CIF (l^*, s^*) , the oscillatory solution of the long form would be the same except there is a small shift (see Figure 6.21).

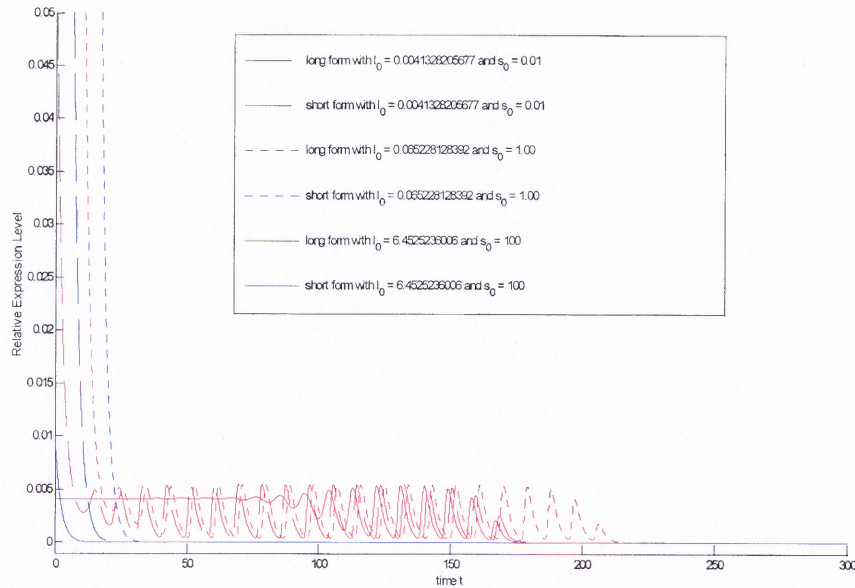


Figure 6.21 Solutions with different long and short form CIFs. Three distinct solutions for different CIFs near the basin boundary. For these CIFs, the solutions for the long form transcript go to zero after a small amplitude transient oscillation. The solutions for the short form transcript go to zero rapidly.

We also notice that the small oscillation has a period of about nine, which is a slightly higher than the delay time of eight. Why does this happen? From Figure 6.21, we know that when the small oscillation occurs, $s(t)$ goes to zero, $l(t)$ is very small with a maximum around 0.005, and K is very large compared with $l(t)$. Thus the first equation of Equation (6.3) will can be approximated by the following equation:

$$\frac{dl}{dt} = I_0 \frac{l^n(t - \tau_0)}{K_l^n + l^n(t - \tau_0)} - d_l l. \quad (6.18)$$

Equation (6.18) has a fixed point at $l_0 = 0.00413167192288$. When we linearize the equation around the stationary point, we obtain the following linear DDE:

$$\frac{d\hat{x}}{d\hat{t}} = \hat{\alpha}x(\hat{t}) + \hat{\beta}\hat{x}(\hat{t} - \tau_0) \quad (6.19)$$

where $\hat{\alpha} = -d_l < 0$ and $\hat{\beta} = \frac{I_0 K_l^n n l_0^{n-1}}{(K_l^n + l_0^n)^2} > 0$.

From Figure 6.4, we know that if Equation (6.18) has periodic solutions around the fixed point, and the linearized DDE (6.19) has the parameter coefficients located on in the second quadrant, which corresponds to the curve parameter in the interval $((2k+1.5)\pi, (2k+2)\pi)$, then the period must be between τ_0 and $\frac{4}{3}\tau_0$.

Because the period of the solution is $T' = \frac{2\pi}{\nu'}$, where $2\pi > \nu' > 1.5\pi$. This agrees with the numerical simulation very well.

By setting the right-hand side of Equation (6.3) to zero, with the function `lsqnonlin` in Matlab, we find another steady state solution numerical solution P1 (1.0906, 1.1547). With this CIF, we should have gotten the constant steady state solution. But due to the numerical errors, the solution went to the same limit cycle after some time (see Figure 6.22-23), which means that the fixed point P1 is not strictly stable.

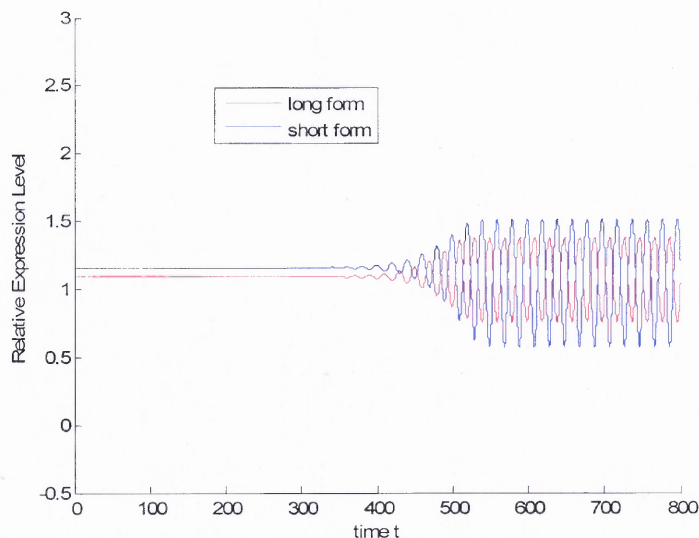


Figure 6.22 Solutions with CIF starting near the fixed point. This fixed point at (1.09059346010033, 1.15472111890551) is an unstable point of Equation (6.3).

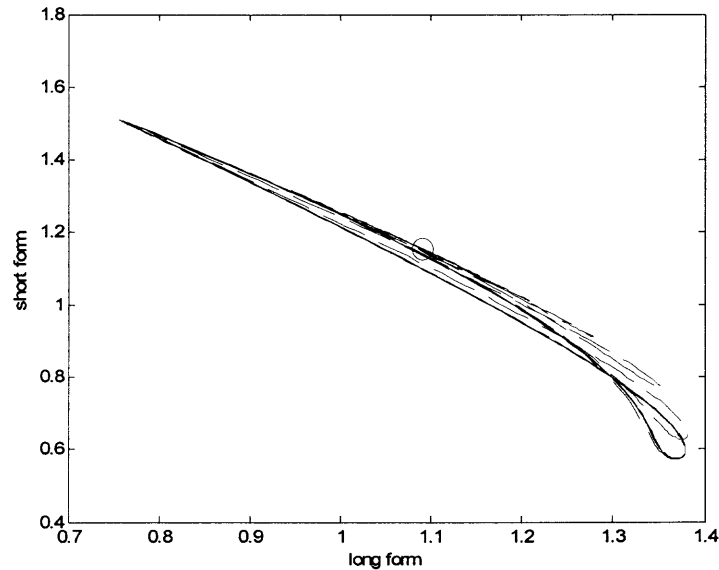


Figure 6.23 Phase plane of the limit cycle solution. The fixed point (1.09059346010033, 1.15472111890551) indicated by the circle is near the limit cycle solution.

Summarize the dynamics of the solutions for the CIF cases (see Figure 6.24): when the CIFs are located to the left of the basin boundary, the solution goes to (0, 0), which is a stable steady state solution of the Equation (6.3); when the initial conditions are located on the boundary condition, the long form goes to a very small oscillation around 0.00413 and the short form goes to zero, the point (0.00413, 5.8279×10^{-10}) is another steady state solution and it is unstable; when the initial conditions are located to the right of the basin boundary, the solution goes to a limit cycle near the fixed point (1.09059346010033, 1.15472111890551), which is neutrally stable.

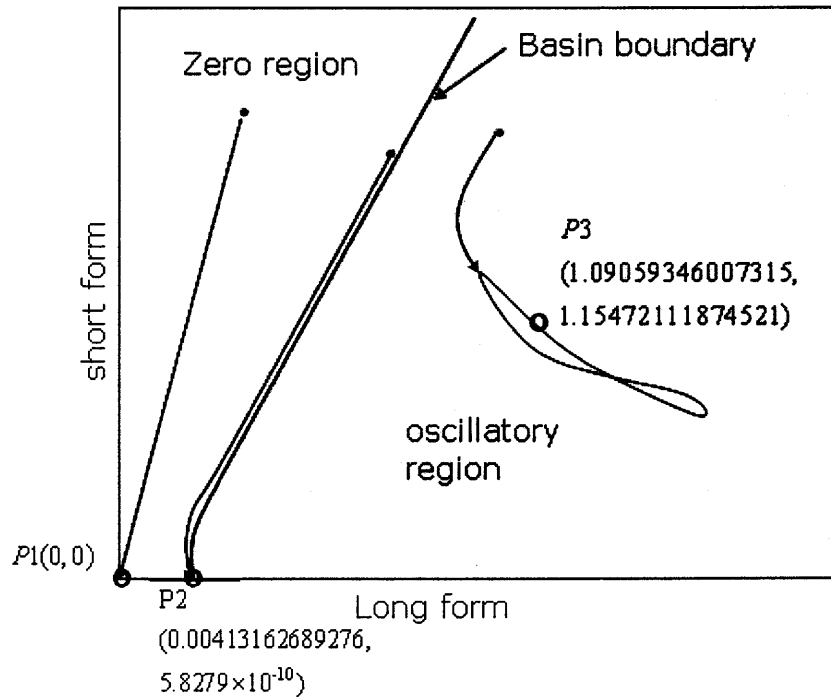


Figure 6.24 Dynamical behavior of Equation (6.3) with optimal parameters. The zero region is to the left of the basin boundary, and the oscillatory region is to the right of the basin boundary. Three fixed points of Equation (6.3) are indicated by the small circles. If the CIF starts in the zero region, the solution is attracted to the origin; if the CIF starts in the oscillatory region, the solution is attracted to the limit cycle solution; if the CIF starts on the basin boundary, the solution for the long form will oscillate with a small amplitude, and the solution for the short form transcript goes to zero rapidly.

There still remain two questions to be asked theoretically based on the above observation of the system with the optimized parameters: why should the system have the same limit cycle or zero solution no matter where the initial conditions located in the oscillatory region, basin boundary, or zero region; how should we find the long form bifurcation CIF theoretically with the given short form CIF.

6.8 Conclusions and Discussion

In summary, a mathematical model has been proposed to simulate the regulation between two different transcripts from gene *cstf-77*. The model is able to produce

solutions that approximate the experimental data. The capability of the model to describe the observed oscillations has been proven. These oscillations can also explain why some long form mRNA cannot be detected even if the mRNA is necessary for the cell from the SAGE data. This model can also be applied to other genes with similar polyadenylation patterns such as poly(A) polymerase (PAP). Multiple forms of poly(A) polymerase (PAPs I,II, III) cDNA have been found in human and mouse and auto-regulation may also exist in the gene expression regulation between different isoforms (Zhao and Manley 1996; Lee et al. 2000).

The system shows a robust and stable oscillation. Robust means that the precise initial conditions are not critical for oscillation to occur if the long form is over some threshold. It implies that the polyadenylation process is very stable, suggesting its importance in the cell cycle and the CstF-77 long form protein is essential for the cell to function.

The dynamical system with the optimized parameter has been studied in detail. The constant initial functions of the model led to the bifurcation of the stable solutions, which agrees well with the fact that knocking down the polyadenylation factors would kill the cell, and a certain amount of long form is essential to rescue the cell from death. The basin boundary has been found numerically, and it showed a very interesting pattern. Three basins of attraction have been found: limit cycle, zero and zero-limit cycle with the short form approaching zero and long form approaching a very small amplitude oscillation.

Three fixed points have been found for Equation (6.3). If the delay is zero, the system becomes ordinary differential equations (ODEs) and the stability can be analyzed. By calculating the Jacob matrix at these points, the eigenvalues can be found. For the ODEs, the point P1 (0, 0) is a stable fixed point, the point P2 (0.00413,

5.8279×10^{-10}) is a saddle point, and the point P3 (1.0906, 1.1547) is a stable fixed point. It implies that the time delay is required for the system to generate the oscillatory solutions.

By adding some more factors, such as CstF-64, the model can be extended to more complex biosynthetic pathways. The identification of the complete alternative polyadenylation mechanism is a complex task due the complexity of the system and the limited knowledge that we have about the model.

The experimental data is very limited at this point. However, once the data is available, this work would provide a comprehensive and consistent mathematical framework for understanding the alternative polyadenylation process.

CHAPTER 7

SUMMARY AND FUTURE WORK

Various problems related to the polyadenylation process have been investigated in this work. When RNA polymerase II binds to the gene promoter region and synthesizes the pre-mRNA, how does the cell know if the transcribed unit is enough for the cell to function? Theoretically, if cleavage can occur at any location of the pre-mRNA, at least as many transcripts as the length of the gene can be generated. But why can only a few transcripts be found in the cell for one gene based on the cDNA/EST? One reason could be that some transcripts are not stable, and they are degraded very soon after being transcribed such that the current technology can not even detect them. There may be some other reasons.

It has been found that the cleavage cannot occur anywhere along a pre-mRNA transcript and signals are required for the cleavage and polyadenylation. By an analysis of the sequence composition around the cleavage sites, some short sequences, the *cis*-elements, could be found to occur more often than other arbitrary combinations of nucleotides. If these *cis*-elements are the determinant factors for the cleavage and polyadenylation (CP), then the occurrence of CP can be treated as a function of these *cis*-elements. The classification can be applied and this leads to the first part of this work (Chapter 3). A lot of similar work has been done to predict the poly(A) sites, and our methods achieved better accuracy and sensitivity than a commonly used one. Since the release of the first version of the work, some improvements in accuracy have been made. However, the false positives and false negatives still exist, which means that for some sequences, it can be cleaved and polyadenylated based on cDNA/EST, but they are predicted not to contain any poly(A)

sites based on the *cis*-elements. Therefore, CP must depend on some other factors, and the one-dimensional sequence information is not enough for the prediction. Thus, for the future, work can be done on the prediction of poly(A) sites, but there will need to be added some more biological factors, such as the two-dimensional pre-mRNA structure of the sequence.

Another natural question may be asked: if there are multiple cleavage sites within the pre-mRNA, how does the cell know which poly(A) site would be used for the current cell condition? Do they generate equal numbers of transcripts for each poly(A) site or do they prefer some sites more than others? Some evidence has been found that the usage of poly(A) sites is tissue-specific. That is, in one tissue, a poly(A) site may be used more often than other poly(A) sites, but in another tissue, some poly(A) sites may be seldom used. The availability of large amount of gene expression data, such as microarray and SAGE data, makes it possible to analyze the relative expression levels of different transcripts from the same gene. This leads to the second part of this work (Chapter 4). Since most of the different transcripts use the same promoter region of the gene, it is hypothesized that if the expression level of one transcript goes up, other transcripts from the same gene may go down, and vice versa. By analyzing the SAGE data, we found that this kind of negative regulation is not common in human. Some significant genes are picked out for containing multiple transcripts, among which some of them have opposite expression levels. Fifteen of the 61 significant unigenes are found to contain alternative poly(A) sites corresponding to the SAGE tags. This may suggest that the alternative polyadenylation may be one factor that generates the negatively regulated transcripts. These genes may be very useful for biologists, and they need to be verified by biological experiments further. Also, some genes have positive regulation between different transcripts, and most of

genes have random correlation among the transcripts. All these suggest that different genes may have different pathways for generating the alternative transcripts. In the future, it is necessary to study the pathway of transcript regulation within the same gene to understand the alternative transcripts generated from alternative splicing or alternative polyadenylation. To achieve this, one gene was selected for this study, which leads to the last part of this work (the Chapter 6).

Gene *cstf3* was selected because the protein product of this gene is a cleavage and polyadenylation factor, and its gene expression could form an auto-regulation by generating different transcripts. The gene regulation network has been studied widely and the transcript regulation network has received little attention. The reasons could be that some transcripts are not stable and can not be detected, the RNA transcripts are not important compared to the protein, which is the functional unit of the cell, and some protein products from these transcripts have not been found to have any function. Mathematical modelling is a very important tool to understand the complex dynamics. Differential-delay equations have been applied to study the biological systems more often recently and they will be widely used in the future.

A novel mechanism has been proposed to explain the alternative polyadenylation and a two-dimensional DDE has been set up to simulate the process. Some interesting results have been observed for the optimal parameters estimated by minimization of the errors of preliminary experimental data and the simulation data. We found that the long form protein is essential for the cell to survive and the short form protein is not that important. This agrees with the current understanding of the *cstf-77* protein products. Also, from the simulation, the long and short form mRNA transcripts oscillate, and sometimes they have very low expression levels. This may explain why in some SAGE libraries, the long form is not detectable even though it is

essential for cells. There is a basin boundary of the long and short form constant initial functions. The theoretical analysis of the basin boundary would be the future work. The experimental data are not sufficient at this time and errors may exist. If the data were complete, this model may provide more insight into the mechanism. The model assumed that only the one protein factor would affect the polyadenylation. Some other factors, such as CstF-64, could be added into the model in the future if the data were available.

Mathematics is complex, and biology is even more complex in some sense. We believe that biology can be better understood by applying mathematics and mathematics can be pushed forward by solving problems coming from biology.

APPENDIX A

STATISTICAL ALGORITHMS AND SIGNIFICANCE

A.1 Markov Chain Model

A Markov chain is a sequence of random variables with the properties that, the present state is only dependent on the past several states. Formally,

$$\Pr(X_{n+1} = x | X_n = x_n, \dots, X_1 = x_1, X_0 = x_0) = \Pr(X_{n+1} = x | X_n = x_n, \dots, X_{n-j+1} = x_{n-j+1})$$

Table A.1 First Order Markov Chain Transition Matrix

Past\Current	A	T	G	C
A	P_{AA}	P_{AT}	P_{AG}	P_{AC}
T	P_{TA}	P_{TT}	P_{TG}	P_{TC}
G	P_{GA}	P_{GT}	P_{GG}	P_{GC}
C	P_{CA}	P_{CT}	P_{CG}	P_{CC}

A, T, G, and C represent the four nucleotides. p_{MN} is the probability that the current nucleotide is N, given that the previous nucleotide was M.

The possible values of X_i form a countable set S called the state space of the chain. j (in the above formula) is the order of the Markov chain. The probability is called the transition probability. If the state space is finite, the transition probability distribution can be represented by a matrix, called the transition matrix. For example, consider the first-order Markov chain of the nucleotide sequence, the transition matrix can be represented by a 4×4 matrix (see Table A.1), where p_{MN} is the probability that the

current nucleotide is N, given that the previous nucleotide was M. To be noticed that:

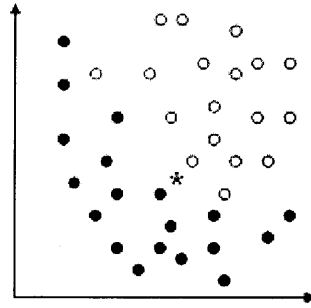
$$\sum_N p_{MN} = 1 \text{ for any given M.}$$

A.2 Linear Discriminant Analysis and Quadratic Discriminant Analysis

Discriminant analysis is a statistical procedure in which an individual unlabeled item (test data set) is put into a group based on previously labeled items (training data set). For example, if there are two groups of labeled data (observation data): one represented by the black dot and the other represented by the white dot (Figure A.1), then the question of how to separate them and to which group a new unlabelled item marked by asterisk in Figure A.1 is assigned have to be addressed. These questions were answered by R.A. Fisher (Fisher 1936) about 70 years ago, who showed that a quadratic decision function given by the following:

$$F(x) = \text{sgn}\left(\frac{1}{2}(x-m_1)^T \Sigma_1^{-1}(x-m_1) - \frac{1}{2}(x-m_2)^T \Sigma_2^{-1}(x-m_2) + \ln \frac{|\Sigma_2|}{|\Sigma_1|}\right) \quad (\text{A.1})$$

was the optimal solution that could discriminate between two populations if the two populations had normal distributions $N(m_1, \Sigma_1)$, $N(m_2, \Sigma_2)$. The attributes of the two population are n dimensional vectors x with mean vectors m_1 , m_2 and co-variance matrices Σ_1 , Σ_2 . It says that when $F(x)$ is equal to 1, the vector x is belonged to one population; otherwise, it is belonged to another population.



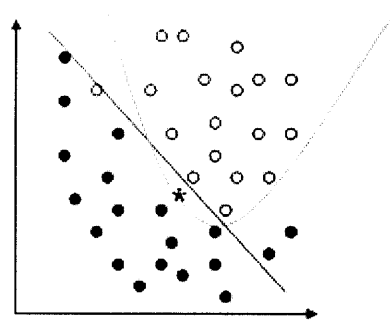
Discriminant Problem

Figure A.1 Discriminant problem. The black dots and white dots represent two groups of data. The asterisk represents the unlabelled data.

If $\Sigma_1 = \Sigma_2 = \Sigma$, Equation (A.1) degenerates to the following linear function:

$$F(x) = \text{sign}((m_1 - m_2)^T \Sigma^{-1} x - \frac{1}{2}(m_1^T \Sigma_1^{-1} m_1 - m_2^T \Sigma_2^{-1} m_2)) \quad (\text{A.2})$$

In fact, Equation (A.2) is a hyper-plane that separates the two populations, and thus this method is called Linear Discriminant Analysis (LDA). The function in Equation (A.1) is a quadratic function, and thus this method is called Quadratic Discriminant Analysis (QDA). A two-dimension LDA and QDA are simply represented in Figure A.2. It can be seen that using different discriminant function, the same data may be assigned to different groups. For example, the unlabeled data marked by asterisk in Figure A.2 would be labeled as black dot by LDA and white dot by QDA.



LDA & QDA

Figure A.2 LDA and QDA. LDA, linear discriminant analysis; QDA, quadratic discriminant analysis. LDA uses a linear function to separate the two groups of data and QDA uses a quadratic function to separate the two groups of data.

A.3 Support Vector Machine

A support vector machine (SVM) is a set of algorithms used for classification and regression (Cortes and Vapnik 1995). The general algorithm includes two steps: the first step is to map the input space to a higher dimensional feature space through some mapping (called a kernel function). Then, in the feature space, find a hyper-plane that separates the data into different groups (see Figure A.3).

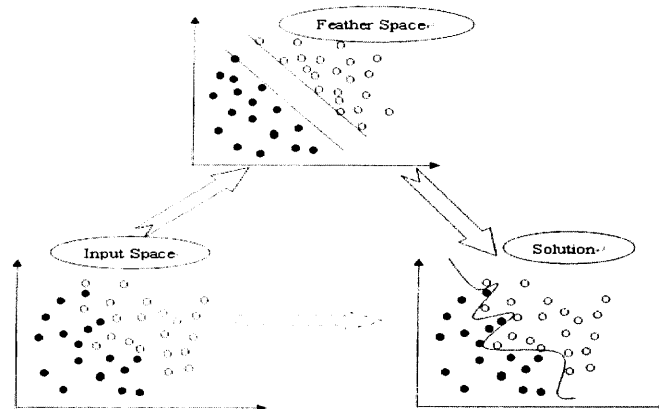


Figure A.3 Support vector machine algorithm. Input space, the original data space; feature space, a higher dimensional space mapped from the input space through some kernel functions.

The optimal separating hyper-plane is selected by maximizing the margin between the classes' closest points. The points lying on the boundaries are called support vectors. Using a kernel function, an SVM is an alternative training method for polynomial functions, radial basis functions (RBF), and sigmoid functions. There are several formulations and kernels when using an SVM (Burges and Smola 1998). Here, a C-classification with an RBF kernel is briefly described, which can be mathematically formulated by:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ & \alpha = (\alpha_1 \dots \alpha_n)^T, \quad 0 \leq \alpha_i \leq C, \quad i=1, \dots, n \\ \text{subject to} \quad & y^T \alpha = 0. \end{aligned} \tag{A.3}$$

Where α is a vector of coefficient to be optimized, e is the vector of all ones, $C > 0$ is a predefined constant, Q is an $n \times n$ positive semi-definite matrix with the component defined by

$$Q_{i,j} = y_i y_j K(x_i, x_j),$$

where

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

is the kernel function, γ is another predefined constant. The decision function is given by the following:

$$D(x) = \text{sgn} \left(\sum_{i=1}^n y_i \alpha_i K(x_i, x) + b \right).$$

The support vector machine algorithm has been implemented into software libraries by many groups. The most popular package is LIBSVM (Chang and Lin 2001) written in C++ and Java and it is used in this work.

A.4 Statistical Significance

True Positive (TP): the number of data predicted positive and the data itself is positive (the prediction is true).

False Negative (FN): the number of data predicted negative and the data itself is positive (the prediction is false), also called Type II error.

False Positive (FP): the number of data predicted positive and the data itself is negative (the prediction is false), also called Type I error.

True Negative (TN): the number of data predicted negative and the data itself is negative (the prediction is true)

$$\text{Sensitivity (SN): } SN = \frac{TP}{TP + FN}, \text{ Specificity (SP): } SP = \frac{TN}{TN + FP}$$

$$\text{Accuracy (CC): } CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

APPENDIX B

PROGRAMMING LANGUAGES

B.1 PERL

Practical Extraction and Report Language (PERL) is a script language and is very famous for its strong regular expression ability. Originally, it was developed for text manipulation. But now, it has wide usage including system administration, web development, network programming. It is very useful in comparing gene sequence, finding the gene patterns and retrieving sequences. Also, there is a biological module in PERL called Bioperl (Stajich et al. 2002). You can retrieve sequence from database such as GenBank and SwissProt just by writing several lines scripts. Also, via Bioperl, we can execute analysis such as BLAST, ClustalW, etc.

B.2 R Language

R is a language and environment for statistical computing and graphics. It is similar to the award-winning S system, which was developed at Bell Laboratories by John Chambers et al (Chambers 1998). It provides a wide variety of statistical and graphical techniques (linear and nonlinear modeling, statistical tests, time series analysis, classification, and clustering, etc.). Most importantly, it's free to download (<http://www.r-project.org>) and install and so it is widely used by university and research institute. Also, there are a lot of packages free to install and easy to use. Support Vector Machines package e1071 is also included in R (Dimitriadou et al. 2006).

APPENDIX C

SAGE MATERIALS

C.1 SAGE Genie

The SAGE data were downloaded from SAGE Genie (Boon et al. 2002) FTP server: <ftp://ftp1.nci.nih.gov/pub/SAGE/HUMAN>. The database version for this study is updated to 07/11/2006. The following files are downloaded from the server: “Hs.libraries”, which is the human SAGE libraries information file, including both long and short libraries; “Hs_long.best_gene”, which is the mapping file of long SAGE tag to the best unigene; “Hs_long.frequencies”, which is frequency data file for long SAGE libraries; “Hs_long.map”, which is the mapping file of long tag to accession number and the rank of the mapping; “Hs_short.best_gene”, which is the mapping file of short SAGE tag to the best unigene; “Hs_short.frequencies”, which is the frequency data file for short SAGE libraries; “Hs_short.map”, which is the mapping file of short tag to accession number and the rank of the mapping. The unigene build number is #194.

C.2 Methods in SAGE Analysis

TPM. The transcript abundance is normalized to transcripts per million (TPM) for comparison between samples. Pearson correlation is based on TPM.

Test of independence. The test statistic (T.Le 2003) is given by:

$$T = r \sqrt{\frac{n-2}{1-r^2}}$$

where r is the Pearson correlation coefficient between two tags, and n is the degree of freedom, the number of libraries that at least one of the two tags has expressed.

Fisher-Shuffle. For an array with m items, this method can generate a random array based on this array. The steps are the following: scan the array from the first element to the last element; for each location i of the array, generate a random integer j less than or equal to the length of the array, and exchange the item i and j . After the scanning, a new array is generated with m items. This method is used for the randomization of SAGE data.

C.3 Sixty-one Significant Unigenes with Negatively Regulated Transcripts

Sixty-one significant unigenes are generated based on the p -value less than 0.01 and the detailed methods are described in Chapter 4.

Table C.1 Sixty-one Significant Unigenes with Negatively Regulated Transcripts

>Hs.114033	SSR1 Signal sequence receptor, alpha			
TCACCATAGAAGGCAAC	AAATTTCTGCTGACTTT	-0.446	158(110,100,52)	-6.2165
>Hs.13885	MGC5309 Hypothetical protein MGC5309			
GGACCGAGGGGCTGGAG	CCTATAATAAACTAAGT	-0.332	223(118,193,88)	-5.2346
>Hs.146585	LEPROTL1 Leptin receptor overlapping transcript-like 1			
AGTTAGAGAGCTGGTGA	GCAGGCATCAAAGCTTT	-0.326	189(159,103,73)	-4.7121
>Hs.153022	TAF1C TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa			
CCTAAGGGAGACATTTA	GCCAAGCTACCACCCCA	-0.428	183(152,77,46)	-6.3783
>Hs.155218	HNRPUL1 Heterogeneous nuclear ribonucleoprotein U-like 1			
CGGTCTTAGAAATGAAA	GCACCTCCTAGCAGGAA	-0.227	223(77,204,58)	-3.4705
>Hs.160211	THRAP3 Thyroid hormone receptor associated protein 3			
TATCACGTGGAGTTGCT	TTTAGGGGGAAAATGAA	-0.23	227(92,212,77)	-3.5444
TGAATAAACTTTGAAGT	GTTCTGGGTCCCTGAGT	-0.398	179(138,79,38)	-5.7649
>Hs.161008	KPNA1 Karyopherin alpha 1 (importin alpha 5)			
ATGTTATTTAAGCAGCC	ACCTGCTTAACCCAAAT	-0.462	153(72,115,34)	-6.397
>Hs.16130	E2-230K Likely ortholog of mouse ubiquitin-conjugating enzyme			
GTCACACTGGGACAGGC	TGGAGATGTGAATGCCT	-0.255	216(127,190,101)	-3.859
>Hs.162877	PACSIN2 Protein kinase C and casein kinase substrate in neurons 2			
TCTCAGTGTCTATCTGT	TGTAGTATTTGAGGAAA	-0.362	207(102,162,57)	-5.5689
>Hs.199561	RANBP2 RAN binding protein 2			
GTGTGAAATAAAAGTTT	GCGCGGGCGAGTGTAGG	-0.165	245(217,222,194)	-2.6122
GCGCGGGCGAGTGTAGG	TTCTTTCGTAAAGATT	-0.226	229(222,87,80)	-3.502

>Hs.233458	NFYC Nuclear transcription factor Y, gamma			
CCTGGGGGCCGAGATTC	AAATGCAATAAATCTCA	-0.287	227(155,184,112)	-4.4961
>Hs.241558	ARIH2 Ariadne homolog 2 (Drosophila)			
TTGAACTGGCCTCTTTT	AATTTACCATATATCTT	-0.278	212(188,135,111)	-4.1925
>Hs.264482	APG12L APG12 autophagy 12-like (S. cerevisiae)			
GTAAAAGTTAATATACT	GTGGCTTACACCTGTAA	-0.376	183(83,152,52)	-5.4552
>Hs.288940	TMEM8 Transmembrane protein 8			
ATGACTAGCGACAATA	TATATATGGGGTTTTTT	-0.39	165(100,117,52)	-5.4069
>Hs.302903	UBE2I Ubiquitin-conjugating enzyme E2I (UBC9 homolog, yeast)			
CTTCTCACCGTGCAGAG	AGGTGCCTCGGAATTAG	-0.177	244(230,193,179)	-2.7955
>Hs.334885	GTPBP3 GTP binding protein 3 (mitochondrial)			
GCTTAATGTGTGTCTTT	CTGGATGCTTCTGACCT	-0.383	173(139,83,49)	-5.4188
>Hs.342307	FLJ10330 PRP38 pre-mRNA processing factor 38 (yeast)			
AGTGTGGAATAAATACT	GGTCTTCTAGTTTTGAC	-0.326	197(171,78,52)	-4.8137
>Hs.344812	TREX1 Three prime repair exonuclease 1			
GTGGCACCAGCACCCT	TATGGGGTTCACAGCCTC	-0.39	188(138,112,62)	-5.7696
>Hs.368304	RPS2 Ribosomal protein S2			
ATGGCTGGTATCGATGA	ATAACTGTCAGAGCTTT	-0.316	246(246,112,112)	-5.208
ATGGCTGGTATCGATGA	GAGAGCCTCAGAATGGG	-0.28	246(246,181,181)	-4.563
ATAACTGTCAGAGCTTT	AAGATCAAGTCCCTGGA	-0.312	196(112,156,72)	-4.5664
AAGATCAAGTCCCTGGA	GAGAGCCTCAGAATGGG	-0.276	222(156,181,115)	-4.2663
>Hs.374127	CDC16 CDC16 cell division cycle 16 homolog (S. cerevisiae)			
GCTTAAGAATGTCCCAC	ATGTTAGAGACATCTAT	-0.386	180(120,117,57)	-5.5886
>Hs.385986	UBE2B Ubiquitin-conjugating enzyme E2B (RAD6 homolog)			
CCCCAGTATTAGCAATG	ACAGAATATACACATTT	-0.442	164(89,126,51)	-6.2774
>Hs.387755	C6orf149 Hypothetical protein LOC285778			
CTGGGCAGCACCAGGTGC	ACCATTGTGTATAGCAT	-0.346	184(103,147,66)	-4.9734
>Hs.407604	CRSP2 Cofactor required for Sp1 transcriptional activation, subunit 2, 150kDa			
TCTTTTGCCTCTTTTGT	AAATGTTCTGGGAGTTG	-0.408	161(84,126,49)	-5.6341
>Hs.413812	RAC1 Ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)			
TATGACTTAATAAATCC	GCTAAGGAGATTGGTGC	-0.169	246(239,242,235)	-2.6717
>Hs.431367	C6orf55 Chromosome 6 open reading frame 55			
GATTGGCCACCTGTTAC	TATTAGAGAATGAAAAG	-0.468	147(86,102,41)	-6.3706
>Hs.432898	RPL4 Mitogen-activated protein kinase kinase kinase 13			
CGCCGGAACACCATTCT	AAAGAACATAGAATATT	-0.199	246(246,143,143)	-3.1768
>Hs.433750	EIF4G1 Eukaryotic translation initiation factor 4 gamma, 1			
GACACACAGATGGCCCG	AATTCAATTAATAAAAAA	-0.157	241(158,219,136)	-2.4573
>Hs.435255	UBXD1 UBX domain containing 1			
GGTCCTGTTCCCGTGTG	GGTCCCGTCCCGTGTG	-0.346	229(143,169,83)	-5.5564

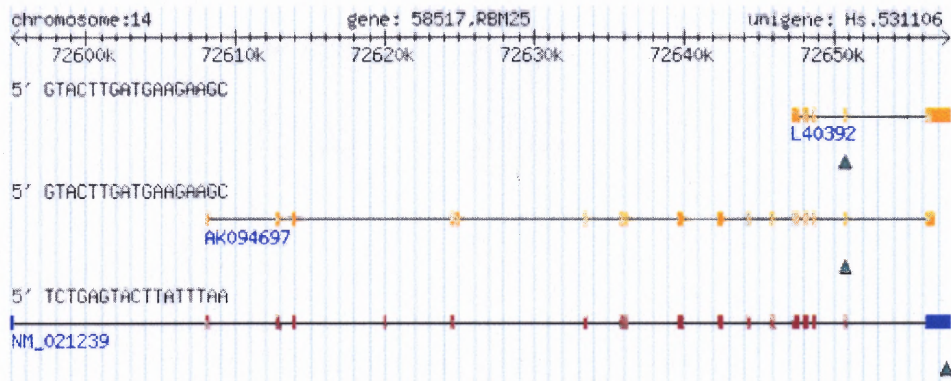
>Hs.437894	BRD7 Bromodomain containing 7			
GCTGACTTGCAGAAAAC	GAAAACAACCTTTGTGGA	-0.365	183(95,150,62)	-5.2757
>Hs.442592	CSNK1A1 Casein kinase 1, alpha 1			
TCAGGATAAATTTAAAAG	ATTATCACATTCTGCCA	-0.348	182(84,146,48)	-4.983
>Hs.463010	SMARCE1 Hypothetical protein MGC45562			
TATGAGTATGTATTTGT	AGCCACCGCACCCAGCC	-0.213	235(152,223,140)	-3.3267
AGCCACCGCACCCAGCC	AGCATTACAGCTGCTGA	-0.213	234(223,121,110)	-3.3257
>Hs.46700	ING1 Inhibitor of growth family, member 1			
GACAAAGCCCTGGAGAA	CTTTTGCTATCACCAAT	-0.497	143(88,93,38)	-6.7979
>Hs.467637	CDC42 Cell division cycle 42 (GTP binding protein, 25kDa)			
CACTCGTGTGAGACAAG	TCTCAATTCTTTGTATA	-0.186	246(213,235,202)	-2.953
CACTCGTGTGAGACAAG	GAAAGACTCTTAATGCA	-0.275	234(213,166,145)	-4.3605
>Hs.469022	DGUOK Deoxyguanosine kinase			
AAGTTCAGGCTGTGATC	ATGTTTCAGGCTGTGATC	-0.285	209(119,155,65)	-4.28
>Hs.470544	PIIG Peptidyl-prolyl isomerase G (cyclophilin G)			
TTGTTTTTGGACAAGTA	AGAAAAACAATAAAAAA	-0.362	178(109,123,54)	-5.1486
>Hs.471785	NCE2 NEDD8-conjugating enzyme			
ACTGCTCATTGTAGATG	TTTACTTTTTTGAAGCTT	-0.279	203(171,128,96)	-4.116
>Hs.480073	HNRPD Heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)			
GACTCTACTGTTTCAGCT	ATTAAAATTGCAGTGAA	-0.524	122(78,71,27)	-6.7375
>Hs.488478	FLJ10099 Hypothetical protein FLJ10099			
ATGGTCAGTAGCTGATC	CACACCAGTTACTTCCT	-0.229	238(203,167,132)	-3.6078
>Hs.502829	SF1 Splicing factor 1			
AAGTGATTCTGTTGACA	CCGCCCTTCGGGATGCC	-0.156	243(234,159,150)	-2.4595
>Hs.506759	ATP2A2 ATPase, Ca ⁺⁺ transporting, cardiac muscle, slow twitch 2			
ATGATCCGGATTTAATT	GTTTCAGGTAAATAAAT	-0.203	243(188,216,161)	-3.2137
>Hs.507087	SPPL3 Signal peptide peptidase 3			
GTTTGTTCCTCTTAGA	CCCTCACTCCTTTAAGA	-0.305	197(100,168,71)	-4.4783
>Hs.515243	LOC93343 Hypothetical protein BC011840			
TTTTTACTTCTGTAGAA	GATGGGGTTCCTTCAC	-0.403	162(88,120,46)	-5.573
>Hs.515515	KDELRI KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 1			
TTTTTGTACAGAAGCTGA	GTGAGCAAGACCGGCGA	-0.204	246(241,151,146)	-3.2496
>Hs.516539	HNRPA3 Heterogeneous nuclear ribonucleoprotein A3			
GTTTTAGTGAAGCAAAC	AAAGGGGGCAGTTTTGG	-0.204	240(185,227,172)	-3.2185
>Hs.517262	SON SON DNA binding protein			
GAAATTGTATTTAACCA	AAAACAGCAAGACTGTA	-0.432	167(78,129,40)	-6.1485
>Hs.518804	ANKRD17 Ankyrin repeat domain 17			
TGAAGGTGGTGCCTAC	GAAAAATTAATTATATG	-0.369	177(99,127,49)	-5.253
GAAAAATTAATTATATG	AACAGTGTGCATATGAA	-0.35	182(127,110,55)	-5.0163

>Hs.520049	HLA-DRB1 Major histocompatibility complex, class II, DR beta 3			
GAAAAAAAAAAAAAAAAA	GCAGTTCTGACAGTGAC	-0.195	243(239,86,82)	-3.0807
>Hs.525238	C14orf119 Chromosome 14 open reading frame 119			
GTGGTGTGCACCTGTAG	AAATGTGTAAAGTAGAA	-0.203	240(211,191,162)	-3.2036
>Hs.529782	VCP Valosin-containing protein			
TTGTAAAAGGACAATAA	CGCTTTGCGCGCCGTTTC	-0.214	246(235,185,174)	-3.4195
>Hs.529798	BTF3 Basic transcription factor 3			
CTGAGACGAAGCAGCTG	CTGAGACAAAGCAGCTG	-0.25	245(84,233,72)	-4.0312
>Hs.529957	SEC63 SEC63-like (<i>S. cerevisiae</i>)			
CAAAGGAAGCTTTTTTT	AGGCTAAGCCTGTGCCA	-0.306	215(208,86,79)	-4.6981
>Hs.5308	UBA52 Ubiquitin A-52 residue ribosomal protein fusion product 1			
CAGATCTTTGTGAAGAC	TGGAAGCTTTCCTTCG	-0.256	246(243,185,182)	-4.1307
>Hs.531106	RBM25 RNA binding motif protein 25			
TCTGAGTACTTATTTAA	GTACTIONGATGAAGAAGC	-0.341	196(141,128,73)	-5.049
>Hs.531879	RAD1 RAD1 homolog (<i>S. pombe</i>)			
CCTGTAATGCCAGCTAC	TAAATAAATGTTTTCT	-0.255	217(163,142,88)	-3.8733
>Hs.533626	SECP43 TRNA selenocysteine associated protein			
ATGTGAGGGAGATGAGA	GTGGCTCACTTTGGGAG	-0.248	219(167,166,114)	-3.7737
>Hs.546303	ST13 Suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein)			
CAGGATCCAGAAGTTAT	CTTTTTAAATCCTTTAA	-0.183	237(231,175,169)	-2.8484
>Hs.546361	ATP2C1 ATPase, Ca ⁺⁺ transporting, type 2C, member 1			
TTAGTTCGACATCATCA	GAAGCCATTGTCTAATC	-0.43	152(84,108,40)	-5.8381
>Hs.546449	MGC4268 Hypothetical protein MGC4268			
AAAGGTTTGTAGTTGAG	ATGTGTTGACATCTACA	-0.531	127(79,75,27)	-7.0055
>Hs.549706	MGC16037 Hypothetical protein MGC16037			
GGGGCACCCGCTGCCCC	GGGGCAGCCGCTGCCCC	-0.237	226(118,192,84)	-3.643
>Hs.7037	PLDN Similar to LINE-1 reverse transcriptase homolog			
GTACTGGTACCAAAACA	TAAATATGCAAATGTTG	-0.24	226(128,203,105)	-3.696
>Hs.96996	HNRPA0 Heterogeneous nuclear ribonucleoprotein A0			
AAGAGCGGCGGCGGCGG	TCATTATATAAACTGT	-0.441	157(106,91,40)	-6.1102

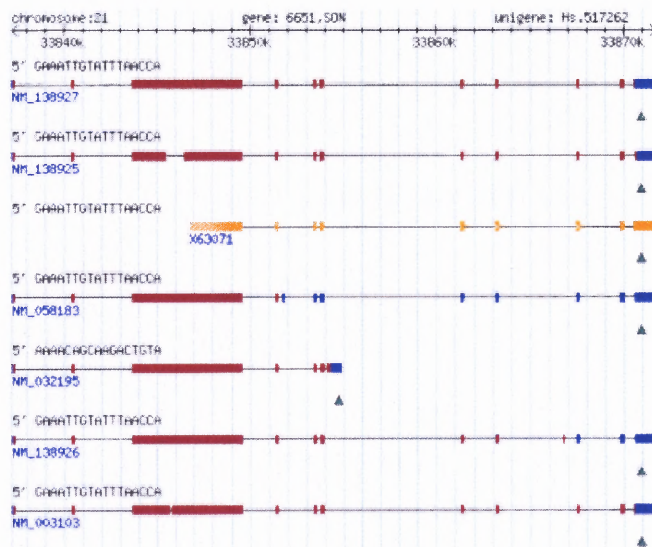
The row beginning with ">" in the first column is the unigene id and the gene description. The rows beginning with tags are the tag information for the unigene id given in the previous line: the first two columns are the negatively regulated tags, the third column is the Pearson correlation of the TPM data for the two tags, the fourth column data is formatted like s(m,n,p), where s is the total library number that at least one tag expressed in, m is the expressed library number for the first column tag, n is the expressed library number for the second column tag, and p is the expressed library number for both tags. The fifth column is the T-values. The lines in bold typeface are the genes with transcripts located in different strands of chromosomes.

C.4 Transcript Structure of Eight Significant Unigenes

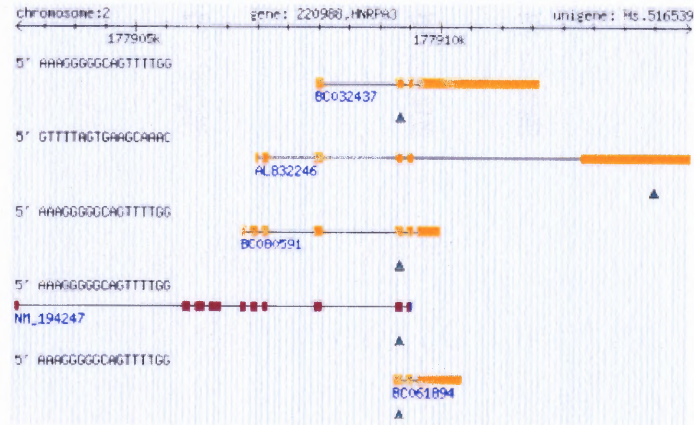
Transcript structures are generated for eight well-known significant unigenes and shown in Figure C.1.



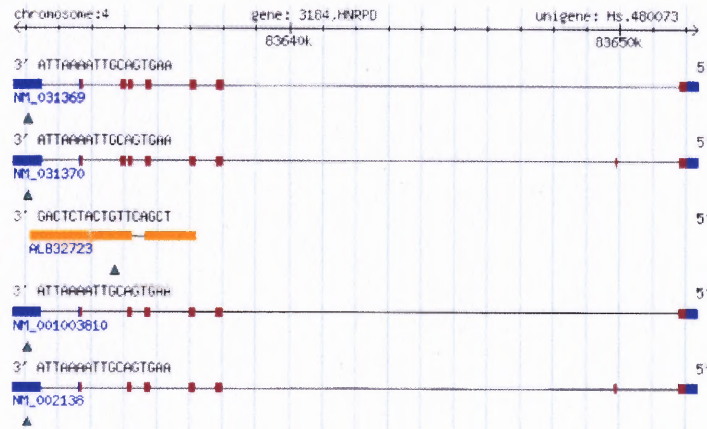
(A)



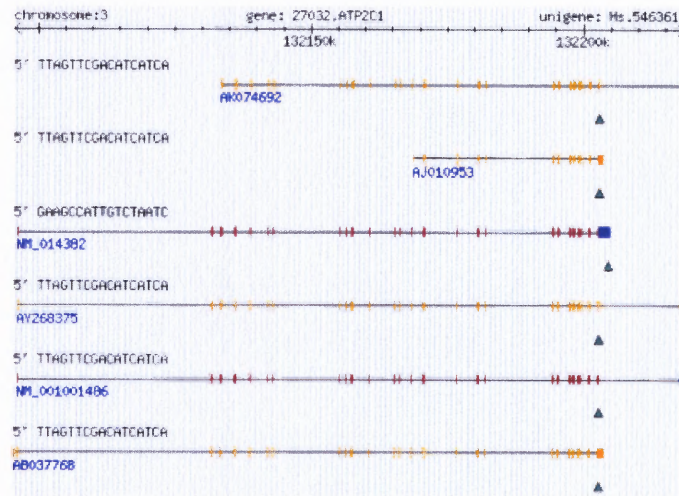
(B)



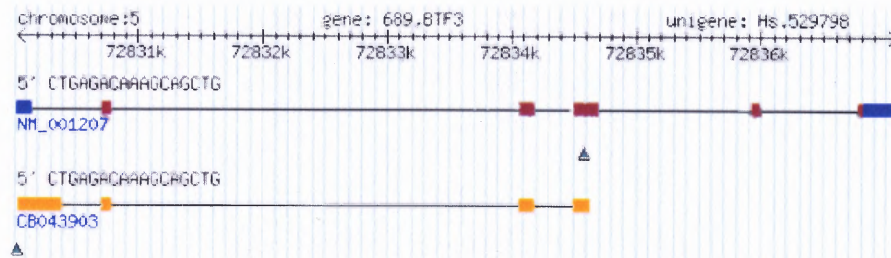
(C)



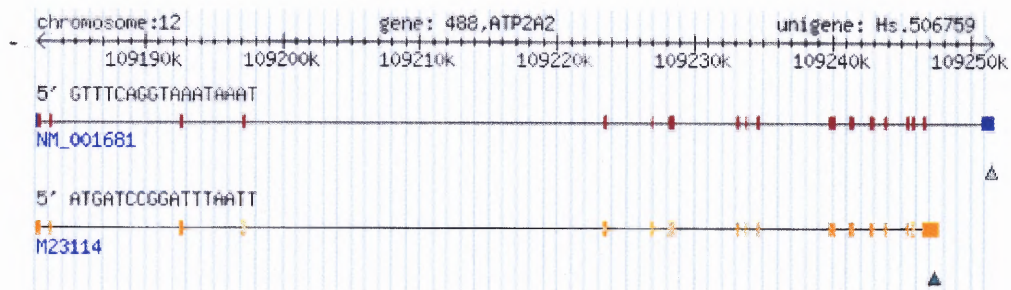
(D)



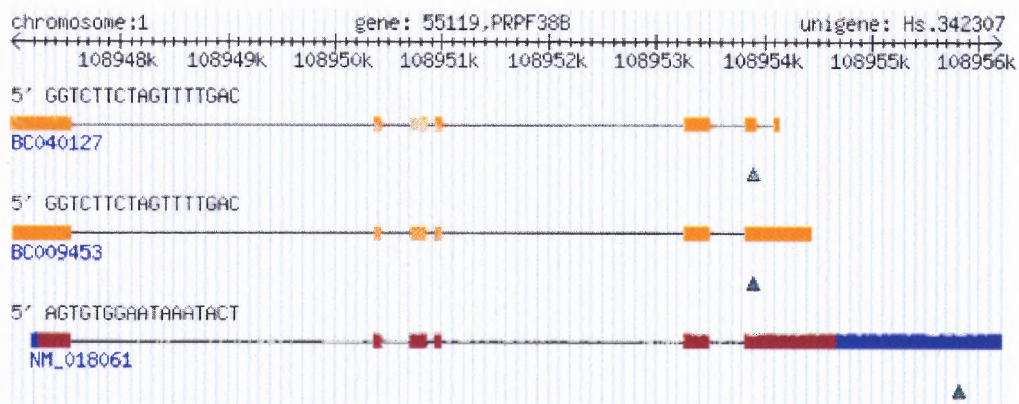
(E)



(F)



(G)



(H)

Figure C.1 The transcript structure of eight significant unigenes. The first line represents the chromosome information of the gene, and the following lines represent the transcript structure generated from this gene. The SAGE tag is indicated by the up-triangle and the accession number of the transcript is indicated under the first exon of the transcript. (A). Hs.531106. (B). Hs.517262. (C). Hs.516539. (D). Hs.480073. (E). Hs.546361. (F). Hs.529798. (G). Hs.506759. (H). Hs.342307.

REFERENCES

- (1998). "Genome sequence of the nematode *C. elegans*: a platform for investigating biology." *Science* **282**(5396): 2012-8.
- Adamson, T. E., D. C. Shutt and D. H. Price (2005). "Functional coupling of cleavage and polyadenylation with transcription of mRNA." *J Biol Chem* **280**(37): 32262-71.
- Andrews, E. M. and D. DiMaio (1993). "Hierarchy of polyadenylation site usage by bovine papillomavirus in transformed mouse cells." *J Virol* **67**(12): 7705-10.
- Antoniou, M., E. de Boer, E. Spanopoulou, A. Imam and F. Grosveld (1995). "TBP binding and the rate of transcription initiation from the human beta-globin gene." *Nucleic Acids Res* **23**(17): 3473-80.
- Arhin, G. K., M. Boots, P. S. Bagga, C. Milcarek and J. Wilusz (2002). "Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals." *Nucleic Acids Res* **30**(8): 1842-50.
- Audibert, A. and M. Simonelig (1998). "Autoregulation at the level of mRNA 3' end formation of the suppressor of forked gene of *Drosophila melanogaster* is conserved in *Drosophila virilis*." *Proc Natl Acad Sci U S A* **95**(24): 14302-7.
- Batt, D. B., Y. Luo and G. G. Carmichael (1994). "Polyadenylation and transcription termination in gene constructs containing multiple tandem polyadenylation signals." *Nucleic Acids Res* **22**(14): 2811-6.
- Becquet, C., S. Blachon, B. Jeudy, J. F. Boulicaut and O. Gandrillon (2002). "Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data." *Genome Biol* **3**(12): RESEARCH0067.
- Bennett, C. L., M. E. Brunkow, F. Ramsdell, K. C. O'Briant, Q. Zhu, R. L. Fuleihan, A. O. Shigeoka, H. D. Ochs and P. F. Chance (2001). "A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome." *Immunogenetics* **53**(6): 435-9.
- Bienroth, S., E. Wahle, C. Suter-Crazzolara and W. Keller (1991). "Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors." *J Biol Chem* **266**(29): 19768-76.
- Boon, K., E. C. Osorio, S. F. Greenhut, C. F. Schaefer, J. Shoemaker, K. Polyak, P. J. Morin, K. H. Buetow, R. L. Strausberg, et al. (2002). "An anatomy of normal and malignant gene expression." *Proc Natl Acad Sci U S A* **99**(17): 11287-92.
- Brown, P. H., L. S. Tiley and B. R. Cullen (1991). "Efficient polyadenylation within the human immunodeficiency virus type 1 long terminal repeat requires flanking U3-specific sequences." *J Virol* **65**(6): 3340-3.

- Bruce, S. R., R. W. Dingle and M. L. Peterson (2003). "B-cell and plasma-cell splicing differences: a potential role in regulated immunoglobulin RNA processing." *Rna* **9**(10): 1264-73.
- Buratowski, S. (2005). "Connections between mRNA 3' end processing and transcription termination." *Curr Opin Cell Biol* **17**(3): 257-61.
- Burges, C. J. C. and A. J. Smola (1998). *Advances in Kernel Methods-Support Vector Learning*. Cambridge, MA, MIT Press.
- Calvo, O. and J. L. Manley (2001). "Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination." *Mol Cell* **7**(5): 1013-23.
- Carswell, S. and J. C. Alwine (1989). "Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences." *Mol Cell Biol* **9**(10): 4248-58.
- Cartegni, L., S. L. Chew and A. R. Krainer (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing." *Nat Rev Genet* **3**(4): 285-98.
- Castelo-Branco, P., A. Furger, M. Wollerton, C. Smith, A. Moreira and N. Proudfoot (2004). "Polypyrimidine tract binding protein modulates efficiency of polyadenylation." *Mol Cell Biol* **24**(10): 4174-83.
- Chambers, J. M. (1998). *Programming with Data: A Guide to the S Language*. New York, Springer-Verlag.
- Chang, C.-C. and C.-J. Lin. (2001). "*LIBSVM* : a library for support vector machines." from <<http://www.csie.ntu.edu.tw/~cjlin/libsvm>>, (visited Feb. 1, 2006).
- Chao, L. C., A. Jamil, S. J. Kim, L. Huang and H. G. Martinson (1999). "Assembly of the cleavage and polyadenylation apparatus requires about 10 seconds in vivo and is faster for strong than for weak poly(A) sites." *Mol Cell Biol* **19**(8): 5588-600.
- Cheng, Y., R. M. Miura and B. Tian (2006). "Prediction of mRNA polyadenylation sites by support vector machine." *Bioinformatics* **22**(19): 2320-5.
- Colgan, D. F. and J. L. Manley (1997). "Mechanism and regulation of mRNA polyadenylation." *Genes Dev* **11**(21): 2755-66.
- Condon, C., C. Squires and C. L. Squires (1995). "Control of rRNA transcription in *Escherichia coli*." *Microbiol Rev* **59**(4): 623-45.
- Cortes, C. and V. Vapnik (1995). "Support-vector Networks." *Machine Learning* **20**: 273-97.
- Corwin, S. P., D. Sarafyan and S. Thompson (1997). "DKLAG6: a code based on continuously imbedded sixth-order Runge-Kutta methods for the solution of

- state-dependent functional differential equations.” *Applied Numerical Mathematics* **24**(2-3): 319-330.
- Covert, M. W., T. H. Leung, J. E. Gaston and D. Baltimore (2005). “Achieving stability of lipopolysaccharide-induced NF-kappaB activation.” *Science* **309**(5742): 1854-7.
- Crick, F. (1970). “Central Dogma of Biology.” *Nature* **227**: 561-3.
- Cui, Y. and C. L. Denis (2003). “In vivo evidence that defects in the transcriptional elongation factors RPB2, TFIIS, and SPT5 enhance upstream poly(A) site utilization.” *Mol Cell Biol* **23**(21): 7887-901.
- de Moor, C. H., H. Meijer and S. Lissenden (2005). “Mechanisms of translational control by the 3' UTR in development and differentiation.” *Semin Cell Dev Biol* **16**(1): 49-58.
- de Vries, H., U. Ruegsegger, W. Hubner, A. Friedlein, H. Langen and W. Keller (2000). “Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors.” *Embo J* **19**(21): 5895-904.
- Denome, R. M. and C. N. Cole (1988). “Patterns of polyadenylation site selection in gene constructs containing multiple polyadenylation signals.” *Mol Cell Biol* **8**(11): 4829-39.
- DeZazzo, J. D. and M. J. Imperiale (1989). “Sequences upstream of AAUAAA influence poly(A) site selection in a complex transcription unit.” *Mol Cell Biol* **9**(11): 4951-61.
- Diekmann, O. (1995). *Delay Equations: Functional, Complex, & Nonlinear Analysis*. New York, Springer.
- Dimitriadou, E., K. Hornik, F. Leisch, D. Meyer and A. Weingessel. (2006). “Support Vector Machines, the interface to libsvm in package e1071.” from <http://cran.r-project.org/src/contrib/Descriptions/e1071.html>, (visited Feb. 1, 2006).
- Dominski, Z., X. C. Yang and W. F. Marzluff (2005). “The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing.” *Cell* **123**(1): 37-48.
- Edwards-Gilbert, G., K. L. Veraldi and C. Milcarek (1997). “Alternative poly(A) site selection in complex transcription units: means to an end?” *Nucleic Acids Res* **25**(13): 2547-61.
- Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein (1998). “Cluster analysis and display of genome-wide expression patterns.” *Proc Natl Acad Sci U S A* **95**(25): 14863-8.
- Elowitz, M. B. and S. Leibler (2000). “A synthetic oscillatory network of transcriptional regulators.” *Nature* **403**(6767): 335-8.

- Engelborghs, K., T. Luzyanina and D. Roose (2002). "Numerical bifurcation analysis of delay differential equations using DDE-BIFTOOL." *ACM Trans. Math. Softw.* 28(1): 1-21.
- Fall, C., E. Marland, J. Wagner and J. Tyson (2003). *Computational Cell Biology*, Springer.
- Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems." *Ann. Eugenics* 7: 111-32.
- Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, et al. (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd." *Science* 269(5223): 496-512.
- Forbes, K. P., B. Addepalli and A. G. Hunt (2006). "An Arabidopsis Fip1 homolog interacts with RNA and provides conceptual links with a number of other polyadenylation factor subunits." *J Biol Chem* 281(1): 176-86.
- Frischmeyer, P. A., A. van Hoof, K. O'Donnell, A. L. Guerrerio, R. Parker and H. C. Dietz (2002). "An mRNA surveillance mechanism that eliminates transcripts lacking termination codons." *Science* 295(5563): 2258-61.
- Gawande, B., M. D. Robida, A. Rahn and R. Singh (2006). "Drosophila Sex-lethal protein mediates polyadenylation switching in the female germline." *Embo J* 25(6): 1263-72.
- Gee, A. H., W. Kasprzak and B. A. Shapiro (2006). "Structural differentiation of the HIV-1 polyA signals." *J Biomol Struct Dyn* 23(4): 417-28.
- Gehring, N. H., U. Frede, G. Neu-Yilik, P. Hundsdoerfer, B. Vetter, M. W. Hentze and A. E. Kulozik (2001). "Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia." *Nat Genet* 28(4): 389-92.
- Giardina, C. and J. T. Lis (1993). "Polymerase processivity and termination on Drosophila heat shock genes." *J Biol Chem* 268(32): 23806-11.
- Gilmartin, G. M. and J. R. Nevins (1991). "Molecular analyses of two poly(A) site-processing factors that determine the recognition and efficiency of cleavage of the pre-mRNA." *Mol Cell Biol* 11(5): 2432-8.
- Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, et al. (1996). "Life with 6000 genes." *Science* 274(5287): 546, 563-7.
- Goldbeter, A. (1995). "A model for circadian oscillations in the Drosophila period protein (PER)." *Proc Biol Sci* 261(1362): 319-24.
- Goodwin, B. C. (1965). "Oscillatory behavior in enzymatic control processes." *Adv Enzyme Regul* 3: 425-38.

- Goodwin, B. C. (1966). "An entrainment model for timed enzyme syntheses in bacteria." *Nature* **209**(22): 479-81.
- Graber, J. H. (2003). "Variations in yeast 3'-processing cis-elements correlate with transcript stability." *Trends Genet* **19**(9): 473-6.
- Graber, J. H., C. R. Cantor, S. C. Mohr and T. F. Smith (1999). "In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species." *Proc Natl Acad Sci U S A* **96**(24): 14055-60.
- Graber, J. H., G. D. McAllister and T. F. Smith (2002). "Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3'-processing sites." *Nucleic Acids Res* **30**(8): 1851-8.
- Graveley, B. R. and G. M. Gilmartin (1996). "A common mechanism for the enhancement of mRNA 3' processing by U3 sequences in two distantly related lentiviruses." *J Virol* **70**(3): 1612-7.
- Greenbaum, D., C. Colangelo, K. Williams and M. Gerstein (2003). "Comparing protein abundance and mRNA expression levels on a genomic scale." *Genome Biol* **4**(9): 117.
- Griffith, J. S. (1968(a)). "Mathematics of cellular control processes. I. Negative feedback to one gene." *J Theor Biol* **20**(2): 202-8.
- Griffith, J. S. (1968(b)). "Mathematics of cellular control processes. II. Positive feedback to one gene." *J Theor Biol* **20**(2): 209-16.
- Hajarnavis, A., I. Korf and R. Durbin (2004). "A probabilistic model of 3' end formation in *Caenorhabditis elegans*." *Nucleic Acids Res* **32**(11): 3392-9.
- Hall-Pogar, T., H. Zhang, B. Tian and C. S. Lutz (2005). "Alternative polyadenylation of cyclooxygenase-2." *Nucleic Acids Res* **33**(8): 2565-79.
- Hatton, L. S., J. J. Eloranta, L. M. Figueiredo, Y. Takagaki, J. L. Manley and K. O'Hare (2000). "The *Drosophila* homologue of the 64 kDa subunit of cleavage stimulation factor interacts with the 77 kDa subunit encoded by the *suppressor of forked* gene." *Nucleic Acids Res* **28**(2): 520-6.
- Heiden, U. a. d. and M. C. Mackey (1982). "The dynamics of production and destruction: analytic insight into complex behavior." *Journal of mathematical biology* **16**: 75-101.
- Herbert, K. M., A. La Porta, B. J. Wong, R. A. Mooney, K. C. Neuman, R. Landick and S. M. Block (2006). "Sequence-resolved detection of pausing by single RNA polymerase molecules." *Cell* **125**(6): 1083-94.
- Hoopes, B. C., J. F. LeBlanc and D. K. Hawley (1998). "Contributions of the TATA box sequence to rate-limiting steps in transcription initiation by RNA polymerase II." *J Mol Biol* **277**(5): 1015-31.

- Hu, J., C. S. Lutz, J. Wilusz and B. Tian (2005). "Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation." *RNA* **11**(10): 1485-93.
- Jacobson, A. and S. W. Peltz (1996). "Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells." *Annu Rev Biochem* **65**: 693-739.
- Kaneko, S. and J. L. Manley (2005). "The mammalian RNA polymerase II C-terminal domain interacts with RNA to suppress transcription-coupled 3' end formation." *Mol Cell* **20**(1): 91-103.
- Keller, W. (1995). "No end yet to messenger RNA 3' processing!" *Cell* **81**(6): 829-32.
- Keller, W., S. Bienroth, K. M. Lang and G. Christofori (1991). "Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA." *Embo J* **10**(13): 4241-9.
- Klasens, B. I., M. Thiesen, A. Virtanen and B. Berkhout (1999). "The ability of the HIV-1 AAUAAA signal to bind polyadenylation factors is controlled by local RNA structure." *Nucleic Acids Res* **27**(2): 446-54.
- Koretzky, G. A. and P. S. Myung (2001). "Positive and negative regulation of T-cell activation by adaptor proteins." *Nat Rev Immunol* **1**(2): 95-107.
- Korzheva, N., A. Mustaev, M. Kozlov, A. Malhotra, V. Nikiforov, A. Goldfarb and S. A. Darst (2000). "A structural model of transcription elongation." *Science* **289**(5479): 619-25.
- Ku, M., S. Y. Sokol, J. Wu, M. I. Tussie-Luna, A. L. Roy and A. Hata (2005). "Positive and negative regulation of the transforming growth factor beta/activin target gene goosecoid by the TFII-I family of transcription factors." *Mol Cell Biol* **25**(16): 7144-57.
- Laird-Offringa, I. A., P. Elfferich, H. J. Knaken, J. de Ruiter and A. J. van der Eb (1989). "Analysis of polyadenylation site usage of the c-myc oncogene." *Nucleic Acids Res* **17**(16): 6499-514.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, et al. (2001). "Initial sequencing and analysis of the human genome." *Nature* **409**(6822): 860-921.
- Lee, J. Y., I. Yeh, J. Y. Park and B. Tian (2007). "PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes." *Nucleic Acids Res* **35**(Database issue): D165-8.
- Lee, Y. J., Y. Lee and J. H. Chung (2000). "An intronless gene encoding a poly(A) polymerase is specifically expressed in testis." *FEBS Lett* **487**(2): 287-92.
- Legendre, M. and D. Gautheret (2003). "Sequence determinants in human polyadenylation site selection." *BMC Genomics* **4**(1): 7.

- Levitt, N., D. Briggs, A. Gil and N. J. Proudfoot (1989). "Definition of an efficient synthetic poly(A) site." *Genes Dev* **3**(7): 1019-25.
- Lewin, B. (2003). *Gene VIII*, Prentice Hall.
- Lewis, R. D., G. R. Warman and D. S. Saunders (1997). "Simulations of Free-running Rhythms, Light Entrainment and the Light-pulse Phase Response Curves for the Locomotor Activity Rhythm in period Mutant of *Drosophila melanogaster*." *Journal of Theoretical Biology* **185**(4): 503-510.
- Liu, B. and B. M. Alberts (1995). "Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex." *Science* **267**(5201): 1131-7.
- Losson Jm, J., M. C. Mackey and A. Longtin (1993). "Solution multistability in first-order nonlinear differential delay equations." *Chaos* **3**(2): 167-176.
- Lou, H., K. M. Neugebauer, R. F. Gagel and S. M. Berget (1998). "Regulation of alternative polyadenylation by U1 snRNPs and SRp20." *Mol Cell Biol* **18**(9): 4977-85.
- Lutz, C. S., K. G. Murthy, N. Schek, J. P. O'Connor, J. L. Manley and J. C. Alwine (1996). "Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro." *Genes Dev* **10**(3): 325-37.
- Ma, X., H. Riemann, G. Gri and G. Trinchieri (1998). "Positive and negative regulation of interleukin-12 gene expression." *Eur Cytokine Netw* **9**(3 Suppl): 54-64.
- Mackey, M. C. and L. Glass (1977). "Oscillation and chaos in physiological control systems." *Science* **197**(4300): 287-9.
- Mackey, M. C. and I. G. Nechaeva (1994). "Noise and stability in differential delay equations." *Journal of Dynamics and Differential Equations* **6**(3): 395-426.
- Mangus, D. A., M. C. Evans and A. Jacobson (2003). "Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression." *Genome Biol* **4**(7): 223.
- McCracken, S., N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S. D. Patterson, M. Wickens and D. L. Bentley (1997). "The C-terminal domain of RNA polymerase II couples mRNA processing to transcription." *Nature* **385**(6614): 357-61.
- Millevoi, S., F. Geraghty, B. Idowu, J. L. Tam, M. Antoniou and S. Vagner (2002). "A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3'-end processing." *EMBO Rep* **3**(9): 869-74.
- Monk, N. A. (2003). "Oscillatory expression of Hes1, p53, and NF-kappaB driven by transcriptional time delays." *Curr Biol* **13**(16): 1409-13.

- Moreira, A., Y. Takagaki, S. Brackenridge, M. Wollerton, J. L. Manley and N. J. Proudfoot (1998). "The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms." *Genes Dev* **12**(16): 2522-34.
- Murray, J. D. (1989). *Mathematical Biology*. Berlin, Springer.
- Murthy, K. G. and J. L. Manley (1992). "Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus." *J Biol Chem* **267**(21): 14804-11.
- Murthy, K. G. and J. L. Manley (1995). "The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation." *Genes Dev* **9**(21): 2672-83.
- Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, et al. (2000). "A whole-genome assembly of *Drosophila*." *Science* **287**(5461): 2196-204.
- Nie, L., G. Wu and W. Zhang (2006). "Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: A quantitative analysis." *Genetics*.
- Nikaido, I., C. Saito, Y. Mizuno, M. Meguro, H. Bono, M. Kadomura, T. Kono, G. A. Morris, P. A. Lyons, et al. (2003). "Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling." *Genome Res* **13**(6B): 1402-9.
- O'Gorman, W., K. Y. Kwek, B. Thomas and A. Akoulitchev (2006). "Non-coding RNA in transcription initiation." *Biochem Soc Symp*(73): 131-40.
- Orphanides, G. and D. Reinberg (2002). "A unified theory of gene expression." *Cell* **108**(4): 439-51.
- Pan, Z., H. Zhang, L. K. Hague, J. Y. Lee, C. S. Lutz and B. Tian (2006). "An intronic polyadenylation site in human and mouse CstF-77 genes suggests an evolutionarily conserved regulatory mechanism." *Gene* **366**(2): 325-34.
- Park, N. J., D. C. Tsao and H. G. Martinson (2004). "The two steps of poly(A)-dependent termination, pausing and release, can be uncoupled by truncation of the RNA polymerase II carboxyl-terminal repeat domain." *Mol Cell Biol* **24**(10): 4092-103.
- Perez Canadillas, J. M. and G. Varani (2003). "Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein." *EMBO J* **22**(11): 2821-30.
- Phillips, C., N. Pachikara and S. I. Gunderson (2004). "U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions." *Mol Cell Biol* **24**(14): 6162-71.

- Proudfoot, N. (2004). "New perspectives on connecting messenger RNA 3' end formation to transcription." *Curr Opin Cell Biol* **16**(3): 272-8.
- Qiu, J. and D. J. Pintel (2004). "Alternative polyadenylation of adeno-associated virus type 5 RNA within an internal intron is governed by the distance between the promoter and the intron and is inhibited by U1 small nuclear RNP binding to the intervening donor." *J Biol Chem* **279**(15): 14889-98.
- Quere, R., L. Manchon, M. Lejeune, O. Clement, F. Pierrat, B. Bonafoux, T. Commes, D. Piquemal and J. Marti (2004). "Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression." *Nucleic Acids Res* **32**(20): e163.
- Rocha, E. P. and A. Danchin (2003). "Essentiality, not expressiveness, drives gene-strand bias in bacteria." *Nat Genet* **34**(4): 377-8.
- Rosenfeld, N., J. W. Young, U. Alon, P. S. Swain and M. B. Elowitz (2005). "Gene regulation at the single-cell level." *Science* **307**(5717): 1962-5.
- Ruby, S. W. (1997). "Dynamics of the U1 small nuclear ribonucleoprotein during yeast spliceosome assembly." *J Biol Chem* **272**(28): 17333-41.
- Ruegsegger, U., K. Beyer and W. Keller (1996). "Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors." *J Biol Chem* **271**(11): 6107-13.
- Ruegsegger, U., D. Blank and W. Keller (1998). "Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits." *Mol Cell* **1**(2): 243-53.
- Sachs, A. B., P. Sarnow and M. W. Hentze (1997). "Starting at the beginning, middle, and end: translation initiation in eukaryotes." *Cell* **89**(6): 831-8.
- Saha, S., A. B. Sparks, C. Rago, V. Akmaev, C. J. Wang, B. Vogelstein, K. W. Kinzler and V. E. Velculescu (2002). "Using the transcriptome to annotate the genome." *Nat Biotechnol* **20**(5): 508-12.
- Salamov, A. A. and V. V. Solovyev (1997). "Recognition of 3'-processing sites of human mRNA precursors." *Comput Appl Biosci* **13**(1): 23-8.
- Scheper, T., D. Klinkenberg, C. Pennartz and J. van Pelt (1999). "A mathematical model for the intracellular circadian rhythm generator." *J Neurosci* **19**(1): 40-7.
- Shampine, L. F. a. S. T. (2001). "Solving DDEs in MATLAB." *Applied Numerical Mathematics*, **37**: 441-458.
- Shell, S. A., C. Hesse, S. M. Morris, Jr. and C. Milcarek (2005). "Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection." *J Biol Chem* **280**(48): 39950-61.

- Simoneg, M., K. Elliott, A. Mitchelson and K. O'Hare (1996). "Interallelic complementation at the suppressor of forked locus of *Drosophila* reveals complementation between suppressor of forked proteins mutated in different regions." *Genetics* **142**(4): 1225-35.
- Sousa, R., D. Patra and E. M. Lafer (1992). "Model for the mechanism of bacteriophage T7 RNAP transcription initiation and termination." *J Mol Biol* **224**(2): 319-34.
- Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, et al. (2002). "The Bioperl toolkit: Perl modules for the life sciences." *Genome Res* **12**(10): 1611-8.
- Stein, L. D. (2004). "Human genome: end of the beginning." *Nature* **431**(7011): 915-6.
- Steinmetz, E. J. and D. A. Brow (2003). "Ssu72 protein mediates both poly(A)-coupled and poly(A)-independent termination of RNA polymerase II transcription." *Mol Cell Biol* **23**(18): 6339-49.
- Sung, N. S., J. I. Gordon, G. D. Rose, E. D. Getzoff, S. J. Kron, D. Mumford, J. N. Onuchic, N. F. Scherer, D. L. Sumners, et al. (2003). "Educating Future Scientists." *Science* **301**: 1485.
- T.Le, C. (2003). *Introductory Biostatistics*. Hoboken, New Jersey, John Wiley & Sons, Inc.
- Tabaska, J. E. and M. Q. Zhang (1999). "Detection of polyadenylation signals in human DNA sequences." *Gene* **231**(1-2): 77-86.
- Takagaki, Y. and J. L. Manley (1994). "A polyadenylation factor subunit is the human homologue of the *Drosophila* suppressor of forked protein." *Nature* **372**(6505): 471-4.
- Takagaki, Y. and J. L. Manley (1998). "Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation." *Mol Cell* **2**(6): 761-71.
- Takagaki, Y. and J. L. Manley (2000). "Complex protein interactions within the human polyadenylation machinery identify a novel component." *Mol Cell Biol* **20**(5): 1515-25.
- Takagaki, Y., J. L. Manley, C. C. MacDonald, J. Wilusz and T. Shenk (1990). "A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs." *Genes Dev* **4**(12A): 2112-20.
- Takagaki, Y., L. C. Ryner and J. L. Manley (1989). "Four factors are required for 3'-end cleavage of pre-mRNAs." *Genes Dev* **3**(11): 1711-24.
- Takagaki, Y., R. L. Seipelt, M. L. Peterson and J. L. Manley (1996). "The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation." *Cell* **87**(5): 941-52.

- Thein, S. L., R. B. Wallace, L. Pressley, J. B. Clegg, D. J. Weatherall and D. R. Higgs (1988). "The polyadenylation site mutation in the alpha-globin gene cluster." *Blood* **71**(2): 313-9.
- Tian, B., J. Hu, H. Zhang and C. S. Lutz (2005). "A large-scale analysis of mRNA polyadenylation of human and mouse genes." *Nucleic Acids Res* **33**(1): 201-212.
- Tian, B., Z. Pan and J. Y. Lee (2007). "Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing." *Genome Res* **17**(2): 156-65.
- Tuteja, R. and N. Tuteja (2004). "Serial analysis of gene expression (SAGE): unraveling the bioinformatics tools." *Bioessays* **26**(8): 916-22.
- Valsamakis, A., N. Schek and J. C. Alwine (1992). "Elements upstream of the AAUAAA within the human immunodeficiency virus polyadenylation signal are required for efficient polyadenylation in vitro." *Mol Cell Biol* **12**(9): 3699-705.
- Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). "Serial analysis of gene expression." *Science* **270**(5235): 484-7.
- Venkataraman, K., K. M. Brown and G. M. Gilmartin (2005). "Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition." *Genes Dev* **19**(11): 1315-27.
- Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, et al. (2001). "The sequence of the human genome." *Science* **291**(5507): 1304-51.
- Veraldi, K. L., G. K. Arhin, K. Martincic, L. H. Chung-Ganster, J. Wilusz and C. Milcarek (2001). "hnRNP F influences binding of a 64-kilodalton subunit of cleavage stimulation factor to mRNA precursors in mouse B cells." *Mol Cell Biol* **21**(4): 1228-38.
- Wagner, A. (2005). *Robustness and Evolvability in Living Systems*, Princeton University Press.
- Wahle, E. (1991). "A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation." *Cell* **66**(4): 759-68.
- Wickens, M., P. Anderson and R. J. Jackson (1997). "Life and death in the cytoplasm: messages from the 3' end." *Curr Opin Genet Dev* **7**(2): 220-32.
- Wood, S. N. (2003). "DDefit 0.504." from <<http://www.maths.bath.ac.uk/~sw283/simon/ddefit.html>>, (visited Mar. 4, 2007).
- Xiong, W. and J. E. Ferrell, Jr. (2003). "A positive-feedback-based bistable 'memory module' that governs a cell fate decision." *Nature* **426**(6965): 460-5.

- Yan, J. and T. G. Marr (2005). "Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat." *Genome Res* **15**(3): 369-75.
- Yeo, G., D. Holste, G. Kreiman and C. B. Burge (2004). "Variation in alternative splicing across human tissues." *Genome Biol* **5**(10): R74.
- Young, B. A., T. M. Gruber and C. A. Gross (2002). "Views of transcription initiation." *Cell* **109**(4): 417-20.
- Zarudnaya, M. I., I. M. Kolomiets, A. L. Potyahaylo and D. M. Hovorun (2003). "Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures." *Nucleic Acids Res* **31**(5): 1375-86.
- Zhang, H., J. Hu, M. Recce and B. Tian (2005). "PolyA_DB: a database for mammalian mRNA polyadenylation." *Nucleic Acids Res* **33 Database Issue**: D116-20.
- Zhang, M. Q. (2000). "Discriminant analysis and its application in DNA sequence motif recognition." *Brief Bioinform* **1**(4): 331-42.
- Zhang, X. H., K. A. Heller, I. Hefter, C. S. Leslie and L. A. Chasin (2003). "Sequence information for the splicing of human pre-mRNA identified by support vector machine classification." *Genome Res* **13**(12): 2637-50.
- Zhao, J., L. Hyman and C. Moore (1999). "Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis." *Microbiol Mol Biol Rev* **63**(2): 405-45.
- Zhao, W. and J. L. Manley (1996). "Complex alternative RNA processing generates an unexpected diversity of poly(A) polymerase isoforms." *Mol Cell Biol* **16**(5): 2378-86.