



uOttawa

L'Université canadienne
Canada's university

**FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES**



**FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES**

Pinchao Ma

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Biology)

GRADE / DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Gene Expression and Splicing Efficiency

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. X. Xia

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Dr. M. Ekker

Dr. G. Drouin

Dr. G. Carmody

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Gene Expression and Splicing Efficiency

by
Pinchao Ma

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for the Masters degree
In the Ottawa-Carleton Institute of Biology

Thèse soumise à
Faculté des études supérieures et postdoctorales
Université d'Ottawa
En vue de l'obtention de la maîtrise ès sciences

L'Institut de biologie d'Ottawa-Carleton



Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 978-0-494-48639-9

Our file Notre référence

ISBN: 978-0-494-48639-9

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Acknowledgements

First of all, I am grateful to my supervisor, Dr. Xuhua Xia, for introducing me to this interesting thesis topic and guiding me in detail throughout this thesis project. In particular, I would like to thank him for his advice on how to do research more efficiently and effectively.

I also thank the members in my advisory committee, Drs. Stephane Aris-Brosou and George Carmody for their very helpful advice on my thesis project and detailed comments on my progress report and manuscript.

Moreover, I would like to thank all the members in Dr. Xia's lab for their helpful suggestions and collaboration.

Finally, I am grateful to my wife for her support and encouragement, so I can finish this work.

Abstract

Most eukaryotic protein-coding genes are split into exons and introns, and introns need to be spliced for the production of mature mRNA by pre-mRNA splicing. Pre-mRNA splicing is very important for eukaryotic gene expression, because it is not only a key step in producing mature mRNA, but can also affect transcription and translation. The purpose of this study is to investigate the relationship between gene expression and splicing efficiency since the relationship has not been studied systematically from a bioinformatic approach. In this thesis research, we focus on the question of how gene expression would constrain the evolution of three principal splicing signals: the donor splice site, the acceptor splice site, and the branchpoint sequence (BPS).

We chose yeast, *Saccharomyces cerevisiae*, as the model organism in this study due to its many research advantages such as relatively simple splicing mechanism and extensive genome-wide characterization of gene expression at both transcript and protein levels. We first studied the relationship between gene expression and the strength of the donor and acceptor splice sites in the yeast, with the latter being characterized by position weight matrix scores. We found that donor and acceptor splice sites in highly expressed genes have significantly higher mean, but smaller variance, of splicing strength than that in lowly expressed genes. In addition, genes with extremely low splice site strength tend to be spliced by non-spliceosomal mechanisms, and the last nucleotide on the exon side in the donor splice site (immediately upstream of the 5'GU dinucleotide) is important for splicing. We further studied the relationship between gene expression and BPS in the yeast. The results revealed that the BPS strength of highly expressed yeast intron-containing genes (ICGs) is significantly higher than that of lowly expressed yeast ICGs. Moreover, highly expressed yeast ICGs have significantly longer distance between the donor splice site and BPS (S1 distance) and slightly longer distance between BPS and the acceptor splice site (S2 distance) than lowly expressed yeast ICGs. The long S1 distance of highly expressed yeast ICG does not indicate the potential of enhancing splicing efficiency through forming secondary structure in the region between the donor splice site and BPS.

Résumé

La plupart des gènes codants pour des protéines sont segmentés en exons et introns, ces derniers devant être épissés par le système d'épissage ARNm pour la production d'ARNm matures. L'épissage pré-ARNm en soit est très important pour l'expression génique chez les eucaryotes, et peut aussi influencer la plupart des étapes d'expression génique. Le but du travail présenté dans cette thèse est d'étudier la relation entre l'expression génique et l'efficacité de l'épissage, comme aucune d'étude systématique de la sorte existe. Dans cette étude, l'efficacité de l'épissage est contrainte par principalement trois signaux: le site donneur, le site accepteur et la sequence de point-de-branchement (branchpoint sequence, BPS).

Nous avons choisi la levure, *Saccharomyces cerevisiae*, comme organisme modèle pour notre étude, due à ses multiples avantages de recherche, tel que son mécanisme d'épissage relativement simple, et la caractérisation à l'échelle génomique de son expression génique, tant au niveau de la transcription qu'au niveau protéique. Nous avons d'abord étudié la relation entre l'expression génique et la force, caractérisée par le score de matrices de pondération, des sites donneurs et accepteurs d'épissage chez la levure. Nous trouvons que les sites donneurs et accepteurs de gènes fortement exprimés ont une force plus élevée, avec une variance plus faible, que ceux de gènes faiblement exprimés. De plus, les gènes avec une faible force au site d'épissage sont épissés par des mécanismes qui ne dépendent pas du spliceosome, et le dernier nucléotide du côté de l'exon du site donneur (immédiatement en amont du dinucléotide 5'GU) est très important pour l'épissage. Nous avons aussi étudié la relation entre l'expression génique et le BPS chez la levure. Les résultats révèlent que la force du BPS de gènes fortement exprimés avec introns (intron-containing genes, ICGs) est significativement plus élevée que celle des ICGs faiblement exprimés. Aussi, les ICGs fortement exprimés ont une distance (S1) entre site donneur d'épissage et le BPS significativement plus élevée que pour les ICGs faiblement exprimés (distance S2). La distance S1 élevée des ICGs fortement exprimés n'indique pas de possibilité d'un rehaussement de l'efficacité de l'épissage par l'entremise de la formation de structures secondaires dans la région entre le site donneur et le site BPS.

Abbreviations

ICG	Intron-containing Gene
CAI	Codon Adaptation Index
PWMS	Position Weight Matrix Score
DAMBE	Data Analysis in Molecular Biology and Evolution
BPS	Branchpoint Sequence
S1 distance	The distance between the donor splice site and the branchpoint sequence
S2 distance	The distance between the branchpoint sequence and the acceptor splice site
MFE	Minimum Folding Energy
NCBI	National Centre for Biotechnology Information

LIST OF CHAPTERS

CHAPTER 1	INTRODUCTION	1
1.1	Gene Expression	1
1.2	Pre-mRNA Splicing	1
1.3	Splicing Signals	4
1.3.1	Splice Sites	4
1.3.2	Branchpoint Sequence	5
1.4	The Importance of Splicing in Eukaryotic Gene Expression.....	6
1.5	Research Motivation	8
1.6	Overview of Subsequent Chapters	9
CHAPTER 2	HIGHLY EXPRESSED YEAST GENES HAVE MORE EFFICIENT SPLICE SITES	11
2.1	Abstract	11
2.2	Introduction	12
2.3	Materials and Methods	17
2.4	Results and Discussions.....	22
CHAPTER 3	GENE EXPRESSION AND BRANCHPOINT SEQUENCE IN YEAST.....	36
3.1	Abstract	36
3.2	Introduction	36
3.3	Materials and Methods	39
3.4	Results and Discussions.....	41
CHAPTER 4	INVESTIGATIONS ON EUKARYOTES WITH COMPLEX SPLICING MECHANISM.....	52
REFERENCE		55

LIST OF TABLES

Table 2.1 Yeast genes whose first exon (i.e., the coding part of the first exon) is shorter than five nucleotides.....	20
Table 2.2 Site-specific frequencies and position weight matrix (PWM) for 246 donor sites. The consensus sequence (UAAAG GUAUGUUUAAUU) can be obtained from those large site-specific PWM entries, with the most important sites in <i>bold italics</i> . The χ^2 test is performed for each site against the expected background frequencies. The sites are labeled with first intron site as 1.	26
Table 2.3 Site-specific frequencies and position weight matrix (PWM) for 278 acceptor sites. The consensus sequence (UUUUUUUUAYAG GCUUC) can be obtained from those large site-specific PWM entries, with the most important sites in <i>bold italics</i> . The χ^2 test is performed for each site against the expected background frequencies. The sites are labeled with first intron site as 1.	28
Table 2.4 Testing the predictions that introns in highly expressed genes have higher PWMS and smaller variance in PWMS than in lowly expressed genes, with gene expression measured by CAI, mRNA and protein abundance. Introns spliced by non-spliceosomal mechanisms are excluded. Mann-Whitney tests generate similar results. All tests are two-tailed.....	30
Table 3.1 Frequency distribution of genes containing the ten different BPS sequences.	41
Table 3.2 Highly expressed ICGs are more likely to harbor a highly conserved UACUAAC BPS.	43
Table 3.3 Testing the prediction that the BPS strength measured by PWMS for the group of highly expressed genes is higher than that for the group of lowly expressed genes, with gene expression measured by CAI, mRNA and protein abundance. All statistical tests were performed by the Mann-Whitney test.	44
Table 3.4 Testing the differences of the S1 and S2 distances between highly and lowly expressed yeast ICGs, with gene expression measured by CAI, mRNA and protein abundance. All statistical tests were performed by the Mann-Whitney test.....	45
Table 3.5 Fitting regression model in Eq. (3.1), with gene expression measured by CAI: (a) Results of analysis of variance, (b) parameter estimation and associated significance tests.	51

LIST OF FIGURES

Figure 1.1 The seven steps in pre-mRNA splicing.	3
Figure 1.2 The 5' and 3' intron-exon boundaries of mammal spliceosome introns.	5
Figure 2.1 Expected non-linear relationship between splicing efficiency and gene expression.	15
Figure 2.2 Relationship between splicing efficiency measured by position weight matrix score (PWMS) and gene expression measured by codon adaptation index. Genes known to be spliced by non-spliceosomal mechanisms are excluded.	32
Figure 2.3 Relationship between splicing efficiency measured by position weight matrix score (PWMS) and gene expression measured by mRNA abundance. Genes known to be spliced by non-spliceosomal mechanisms are excluded.	33
Figure 2.4 Relationship between splicing efficiency measured by position weight matrix score (PWMS) and gene expression measured by protein abundance. Genes known to be spliced by non-spliceosomal mechanisms are excluded.	34
Figure 3.1 Correlation between S1 distance and intron length in yeast intron-containing genes.	46
Figure 3.2 Relationship between minimum folding energy (MFE) and S1 distance in the yeast.	50

CHAPTER 1 INTRODUCTION

1.1 Gene Expression

Gene expression is a process by which the functional information encoded in a gene is converted into protein. Studying gene expression is of crucial importance for biology and health science research because it is gene expression that brings the genome to life (Orphanides & Reinberg, 2002). Eukaryotic gene expression consists of several steps, which basically include transcription, pre-mRNA processing, mRNA export, translation in the cytoplasm, and finally decay of mRNA (Orphanides & Reinberg, 2002). This study focuses on the relationship between gene expression and pre-mRNA splicing, which is a key step in pre-mRNA processing.

1.2 Pre-mRNA Splicing

Pre-mRNA splicing is a process in which introns are removed and exons are joined by a series of biochemical reactions (Rogozin et al., 2005; Dewey et al., 2006). It is carried out by spliceosome, which consists of five small nuclear ribonucleoproteins (snRNPs) that are named U1, U2, U4, U5, and U6, and a range of non-snRNP associated protein factors (Lamond, 1993; Madhani & Guthrie, 1994; Nilsen, 1994; Reed, 1996; Rogozin et al., 2005).

The process of pre-mRNA splicing in most eukaryotic species can be divided into seven steps (Figure 1.1). First, U1 snRNP recognizes and binds to the donor splice site of pre-mRNA involving base pairing between the donor site sequence and a complementary sequence at the 5' end of U1 snRNA (Reed, 1996; Hastings & Krainer, 2001), and a protein named U2AF (U2 associated factor) binds to a pyrimidine tract close to the acceptor splice site (Reed, 1996; Hastings & Krainer, 2001). Second, U2 snRNP identifies and binds to the

branch site with the help of U2AF (Reed, 1996; Hastings & Krainer, 2001). Third, the spliceosome is formed when a trimer containing the U4, U5 and U6 snRNPs binds to the previous complex containing U1 and U2 snRNPs. Fourth, U1 is released and U6 is allowed to come into juxtaposition with the 5' splice site. Meanwhile, U5 shifts its position from exon to intron (Reed, 1996; Hastings & Krainer, 2001). Fifth, a reaction that is triggered by the release of U4 and catalyzed by U6 and U2 occurs and leads to the 5' site cleavage and the formation of a lariat; U5 now interacts with the beginning of the second exon while it still remains attached to the end of the first exon (Reed, 1996; Hastings & Krainer, 2001). Sixth, the two exons are ligated when the 3' site is cleaved. In the last step, the spliceosome disassembles, the lariat is degraded and the snRNPs are recycled (Reed, 1996; Hastings & Krainer, 2001).

Given the importance of donor and acceptor splice sites and branchpoint site during the splicing process, it is natural to expect these sites to be under strong stabilizing selection. In particular, we expect highly expressed genes to experience stronger stabilizing selection than lowly expressed genes.

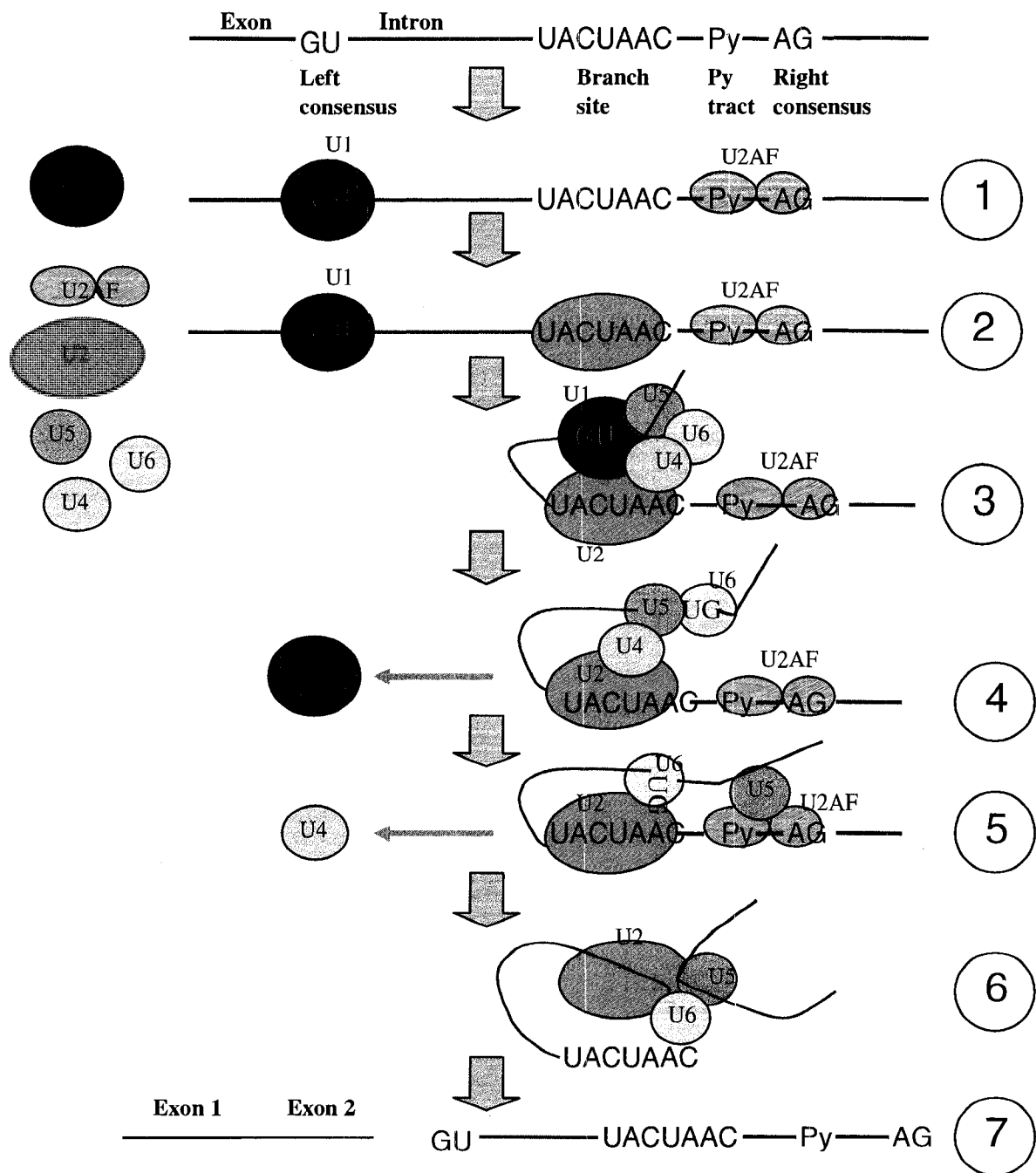


Figure 1.1 The seven steps in pre-mRNA splicing.

1.3 Splicing Signals

Splicing signals are of crucial importance for accurate and efficient pre-mRNA splicing because the interactions between the spliceosomal components and the splicing signals located in both intron and its flanking exons are required for high-precision recognition of intron (Hastings & Krainer, 2001; Black, 2003). The principal splicing signals include the donor splice site, the acceptor splice site, and the branchpoint sequence (Hastings & Krainer, 2001; Black, 2003).

1.3.1 Splice Sites

Splice sites are the sequences that immediately flank either 5' or 3' exon-intron boundaries and are involved in the interactions with spliceosome for correct splicing. There are two splice sites in an intron (Figure 1.2). The 5' (donor) splice site surrounds the exon-intron boundary at the 5' end of the intron and includes the highly conserved GU dinucleotide as the first two positions of the intron; the 3' (acceptor) splice site surrounds the exon-intron boundary at the 3' end of the intron and includes the highly conserved AG dinucleotide as the last two positions of the intron.

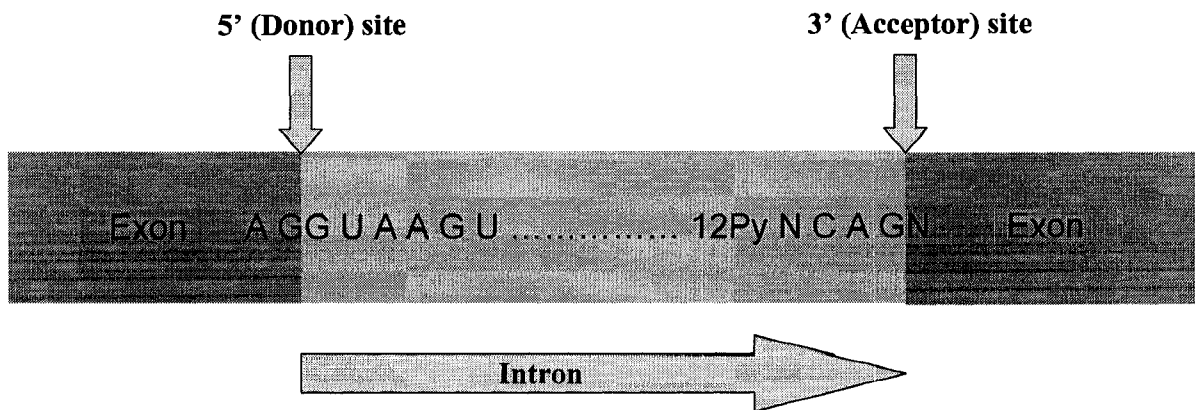


Figure 1.2 The 5' and 3' intron-exon boundaries of mammal spliceosome introns.

Previous studies suggest that the donor and acceptor splice sites tend to be conserved (Rogozin et al., 2005; Dewey et al., 2006). In vertebrates, (A/C)AG|GU(A/G)AGU and CAG|G (the underlined GU and AG are the first two and last two nucleotides of an intron) were identified as the consensus sequences for the donor and acceptor splice sites, respectively (Dewey et al., 2006). It is worth noting that the consensus at the donor and acceptor splice sites can vary significantly among different species (Irimia et al., 2007). For instance, 84% of *Cryptosporidium parvum* introns start with |GUAAGU but only 14% of introns in human begin with the same consensus sequence (Irimia et al., 2007).

1.3.2 Branchpoint Sequence

In the process of splicing, the intron is not removed by directly cutting its two ends at the 5' and 3' splice sites. Actually a lariat is formed as an obligatory intermediate after 5' intron cleavage occurs and before 3' intron cleavage and ligation of exons occur (Hastings & Krainer, 2001). It is formed as a consequence of the 5' end of intron joining an adenosine at the branch point (Cellini et al., 1986). Branchpoint sequence is a short conserved sequence

that contains an adenosine as the branch point and functions in lariat structure formation (Harris & Senapathy, 1990). In the yeast, the branch point is the last adenosine of the branchpoint sequence, which is usually located 18-40 nucleotides upstream of the acceptor splice site (Luukkonen & Seraphin, 1997; McPheeters & Muhlenkamp, 2003). Unlike the branchpoint sequences of higher eukaryotes which are not well conserved and are very degenerate, yeast branchpoint sequence is highly conserved and has the consensus sequence UACUAAC (Bon et al., 2003).

1.4 The Importance of Splicing in Eukaryotic Gene Expression

First of all, splicing is an essential step for producing mature mRNA in eukaryotic gene expression. Accurate splicing is required for all ICGs, otherwise aberrant splicing may produce harmful proteins leading to reduced fitness. Mutations at splicing sites are considered to be a major cause of human diseases (Wang & Cooper, 2007). Efficient and accurate splicing is required for highly expressed ICGs to produce the protein products rapidly (Dewey et al., 2006).

Second, recent advances revealed that splicing can further contribute to gene expression regulation by affecting most of other essential steps in eukaryotic gene expression (Le Hir et al., 2003). It was revealed that splicing can affect transcription efficiency. For instance, transcription can be stimulated by the splicing signals in an intron by enhancing RNA polymerase II (Pol II) initiation (Fong & Zhou, 2001; Furger et al., 2002; Kwek et al., 2002). The components of the spliceosome also can increase phosphorylation for the C-terminal domain (CTD) of the Pol II large subunit, which is necessary for both promoter clearance and efficient transcription elongation (Fong & Zhou, 2001). Moreover, splicing can influence the export of mRNA. For example, in *Xenopus*

oocyte, the export efficiency for spliced mRNAs is much higher than that for unspliced identical mRNAs transcribed from cDNA (Luo & Reed, 1999). This effect is explained by the recruitment of a nucleoprotein complex, which targets mRNA for export as a part of the exon junction complex (EJC) (Luo & Reed, 1999; Le Hir et al., 2000; Le Hir et al., 2001; Reed & Hurt, 2002). Furthermore, translational efficiency can be affected by intron removal from the primary transcripts (Braddock et al., 1994; Matsumoto et al., 1998). It was found that in *Xenopus oocytes*, the translation of a mature mRNA is repressed when it is injected directly into the oocyte nuclei (Braddock et al., 1994). However, including a spliceable intron in the 3' untranslated region (UTR) can overcome this repression of translation (Braddock et al., 1994). Though the molecular mechanisms of the effects of splicing on translation remain unknown, it seems that splicing can enable a mRNA to actively engage ribosomes and avoid the translation repression (Braddock et al., 1994). In addition, splicing can trigger nonsense-mediated decay (NMD). Eukaryotic cells can eliminate premature-termination codons contained in mRNAs by NMD as the quality control mechanism (Maquat & Carmichael, 2001; Wilusz et al., 2001; Isken & Maquat, 2007). Without NMD, harmful proteins can be produced by aberrant mRNAs, which may arise from mutations in DNA. In NMD, distinguishing premature and normal termination codons is a prime step. For instance, in mammalian cells, if stop codons are not located in the terminal exon, they are generally considered as premature (Maquat and Carmichael 2001; Wilusz, Wang et al. 2001). This observation implies that splicing somehow signs the exon-exon junctions in mRNAs, and these signs can be identified by the translation-termination machinery after they accompany the mRNA to the cytoplasm (Maquat & Carmichael, 2001; Wilusz et al., 2001; Isken & Maquat, 2007).

1.5 Research Motivation

The relationship between gene expression and splicing efficiency outlined above naturally leads to the prediction that highly expressed genes should have introns with more efficient splicing sites (i.e., more efficient donor and acceptor sites as well as branchpoint sequence). In this study, I not only quantified the relationship between gene expression and the strength of splicing signals, but also studied the relationship between gene expression and the distances between BPS and the two splice sites.

Several related questions have already been addressed, and interesting patterns revealed, by previous studies. For instance, it was revealed by several studies that highly and widely expressed genes tend to have short introns due to selection pressure whereas lowly expressed genes tend to have longer introns (Castillo-Davis et al., 2002; Eisenberg & Levanon, 2003; Li et al., 2007). Moreover, since more and more evidence indicates that nascent RNA can pair with one strand of DNA and thus form a R-loop structure which can perturb the stability of the genome, Niu proposes that longer introns may be required for the genes with continuously low-level transcription than fast transcribed genes in order to avoid the deleterious R-loop structure (Niu, 2007). Recent studies also revealed some relationships between intron length and splice site strength. In the human genome, Dewey and colleges showed that the strengths of the donor and acceptor splice sites increase with increasing intron length when intron length is greater than 1.5 kb. The proposed explanation is that long introns might hinder the spliceosome from finding the proper position of the splice site and thus correct splicing requires more conserved splice sites (Dewey et al., 2006).

However, there was no systematic research on the direct relationship between gene expression and the splice site strength. Moreover, there were no extensive studies focusing on the relationship between gene expression and BPS.

1.6 Overview of Subsequent Chapters

This chapter is an overview of gene expression, the pre-mRNA splicing process and signals, the importance of splicing in gene expression, and our motivation for this study.

Chapter 2 investigates the relationship between gene expression and splice site strength in the yeast, *Saccharomyces cerevisiae*. Our research focuses on *S. cerevisiae* due to its many research advantages. The strength of the donor and acceptor splices for highly expressed ICGs are expected to be higher than that for lowly expressed ICGs due to strong selection towards high splicing efficiency required for rapid growth and reproduction. This expectation is strongly supported by empirical evidence.

Chapter 3 further examines if there is a relationship between gene expression and the BPS strength as well as the relationships between gene expression and the two distances between BPS and the splice sites in yeast. Yeast BPS is involved in the two steps of splicing and thus is of crucial importance for efficient and accurate splicing. I found that the BPS strength for highly expressed yeast ICGs is higher than that for lowly expressed yeast ICGs. Also, highly expressed yeast ICGs have a significantly longer distance between the donor splice site and BPS (S1 distance) than lowly expressed yeast ICGs. The long S1 distance does not indicate the potential of enhancing splicing efficiency through forming secondary structure in this region.

Chapter 4 extends the research to other model organisms such as human, *Caenorhabditis elegans* and *Arabidopsis thaliana*. It also discusses the implications of the findings and summarizes major conclusions.

CHAPTER 2 HIGHLY EXPRESSED YEAST GENES HAVE MORE EFFICIENT SPLICE SITES

2.1 Abstract

Accurate and efficient splicing is of crucial importance for highly expressed intron-containing genes (ICGs). Introns in highly expressed ICGs are expected to be under strong selection pressure to evolve towards high splicing efficiency. In contrast, the selection for high splicing efficiency should be relatively weak for lowly expressed genes. While introns in some lowly expressed ICGs may also have high splicing efficiency, some introns in such genes may drift to have low splice efficiency through mutation. This has two implications. First, the splicing efficiency of lowly expressed genes should, on average, be lower than that of highly expressed genes. Second, the variance of splicing efficiency should be smaller for highly expressed genes (whose intron splicing efficiency should all be high) than that for lowly expressed genes (whose intron splicing efficiency could be high but may also drift to low values through mutation). We studied the relationship between splicing efficiency and gene expression in the yeast, *Saccharomyces cerevisiae*, with splicing efficiency measured by the position weight matrix scores (PWMSs) of the splice sites and gene expression by three independent indices: codon adaptation index, mRNA abundance and protein abundance. Genes with extremely low PWMSs are spliced by non-spliceosomal mechanisms. The last nucleotide on the exon side in the donor splice site (immediately upstream of the donor GU dinucleotide) is important for splicing. Highly expressed genes have significantly higher splicing efficiency and smaller variance in splicing efficiency than lowly expressed genes.

2.2 Introduction

Most eukaryotic protein-coding genes are split into exons and introns, and introns need to be spliced for the production of mature mRNA (Sharp, 2005). The process of intron removal is known as pre-mRNA splicing and carried out by spliceosome through a multitude of RNA–RNA, RNA–protein and protein–protein interactions involving special sequences at the donor site, acceptor site and branch point of the intron (Padgett et al., 1986; Sharp, 1994; Du & Rosbash, 2002; Lund & Kjems, 2002; Simpson et al., 2002; Jurica & Moore, 2003; Freund et al., 2005; Kent et al., 2005; Nilsen, 2005). The donor and acceptor splice sites are recognized by the base pairing between the sequences at the donor and acceptor splice sites and the complementary sequences at the U1 snRNA and U2AF, respectively, which is then followed by intron excision by the cleavage-ligation reactions after the two ends of intron are brought together.

In multicellular eukaryotes, most introns (>98%) are U2-type introns spliced by U2-type spliceosomes, while U12-type introns (Jackson, 1991; Hall & Padgett, 1994) spliced by the U12-type spliceosomes account for only about 0.1% of all introns (Hall & Padgett, 1994, 1996). Almost all U2-type introns have |GU at the donor splice site and AG| at the acceptor splice site, with very few having |AU and AC| dinucleotides at the donor and splice sites, respectively (Sharp & Burge, 1997). Similarly, the U12-type introns also include introns with |GU...AG| and |AU...AC| at the intron boundary, with the former amounting to 2/3 and the latter to 1/3 of all U12-type introns (Dietrich et al., 2005). The yeast, *Saccharomyces cerevisiae*, appears to have lost all U12-type introns (Sharp & Burge, 1997; Burge et al., 1998), if such introns ever existed in its ancestor. For U2-type introns, the sequences flanking |GU...AG| dinucleotides are important for efficient splicing, with consensus of the

donor and acceptor sites being donor-AG|GUAAGU and YnNAG|G-acceptor, respectively, in multicellular eukaryotes, and being donor-|GUAUGU and YAG|-acceptor in *S. cerevisiae* (Jackson, 1991; Sharp & Burge, 1997; Weaver, 2005, p. 428).

As the principle splicing signals, splice sites are considered to be the major determinant of splicing efficiency, especially for the eukaryotes with relatively simple splicing mechanism such as yeast, *S. cerevisiae*. Thus, the splice site strength should be strongly positively correlated with the splicing efficiency. In general, if an ICG has high splicing efficiency, it also should have high splice site strength; if an ICG has high splice site strength, we will predict its splicing efficiency may be high. The previously experimental studies have demonstrated the importance of splice site strength for splicing efficiency. A mutant 12S adenovirus cannot be spliced when the GG residues at positions 5-6 in the intron are mutated to AU, but the normal splicing is regained when another mutation is introduced into U1 snRNA that restores base pairing at position 5. In addition, splicing efficiency also can be dramatically decreased by mutation of branchpoint sequence even at only one site from UACUAAC to UACUACC in the yeast (Jacquier & Rosbash, 1986).

When a gene is highly expressed, accurate and efficient splicing is of crucial importance, with splicing being a major component of the quality control process in mRNA production in eukaryotes (Maquat & Carmichael, 2001). Thus, all highly expressed genes should have efficient donor and acceptor splice sites to avoid missplicing and aberrant splicing which is not only wasteful but can also overwhelm the quality control machinery and potentially produce wrong proteins that perturb the normal cellular processes. In contrast, the selection for high splicing efficiency should be relatively weak for lowly expressed genes. This does not mean that lowly expressed genes will never have introns of

high splicing efficiency. However, selection for high splice efficiency should be weak for such genes and some splice sites of their introns may drift to have low splice efficiency through mutation. This has two implications. First, the splicing efficiency of lowly expressed genes should, on average, be lower than that of highly expressed genes. Second, the variance of splicing efficiency should be smaller for highly expressed genes (whose intron splicing efficiency should all be high) than that for lowly expressed genes (whose intron splicing efficiency could be high but may also drift to low values through mutation).

It is important to keep in mind that splicing efficiency will not increase with gene expression in a linear fashion. Splicing efficiency is likely determined by only a few nucleotides at the splice motifs (donor and acceptor sites, branchpoint sequence). Once these few sites have been optimized for splicing efficiency, the splicing efficiency will not increase any further no matter how highly expressed a gene is.

The expected non-linear relationship between splicing efficiency and gene expression is schematically depicted in Figure 2.1 to illustrate the two predictions concerning the differences in mean and variance of splicing efficiency between highly expressed and lowly expressed genes. Note that studies quantifying protein production often involve relatively highly expressed genes and miss lowly expressed genes. If we use such protein data as an index to study the relationship between splicing efficiency and gene expression, we may get only the right side of the relationship depicted in Figure 2.1 and consequently conclude that no relationship exists between the two variables.

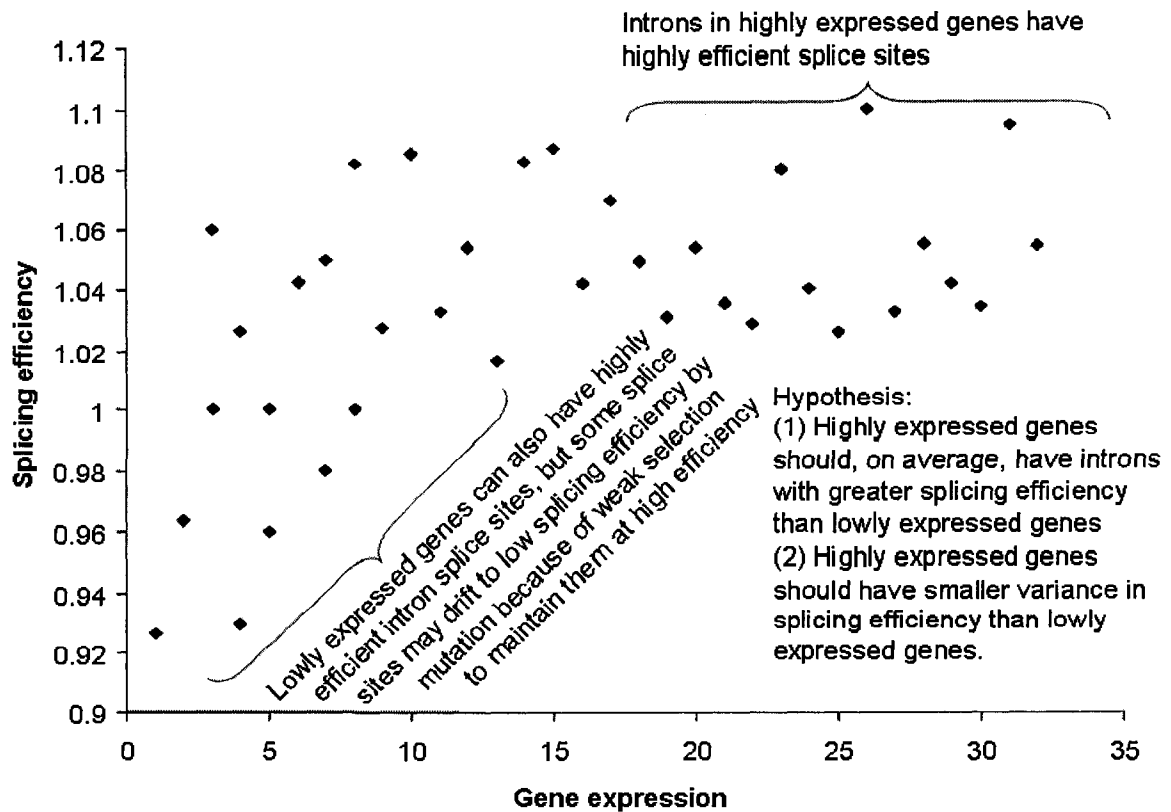


Figure 2.1 Expected non-linear relationship between splicing efficiency and gene expression. The plotted data are fictitious.

Several recent studies have documented the relationship between intron density and length of highly expressed genes, with highly expressed genes having fewer or shorter introns (Comeron, 2004; Fahey & Higgins, 2007; Li et al., 2007), indicating natural selection favoring transcription efficiency. Selection favoring transcription efficiency was also indicated by a previous study showing highly expressed mitochondrial genes exhibit increased use of the most abundant ribonucleotide, i.e., ATP in their mRNA (Xia, 1996). However, the relationship between intron splicing efficiency and gene expression has not been studied.

A comprehensive assessment of the relationship between intron splicing efficiency and gene expression requires accurate characterization of introns and reliable large-scale

measurement of gene expression. The baker's yeast (*S. cerevisiae*) is the only species with accurate characterization of its introns and gene expression at mRNA and protein levels. Two powerful methods have recently been developed to characterize yeast introns. The first is to use high-density yeast tiling arrays in conjunction with a yeast mutant deficient for degradation of processed intron lariats (Juneau et al., 2007). Tiling array that consists of overlapping or closely spaced DNA probes is used to map cellular transcription. When *DBR1*, a gene encoding for a specialized debranching enzyme that functions in degrading the intron lariats, is deleted from the yeast genome, intron lariats build up in the RNA pool (Juneau et al., 2007). The accumulation of lariats is detected by the high-density tiling array which allows not only intron validation but also detection of new introns. The second approach involves designing microarray probes specific for exon-intron junctions and exon-exon junctions to quantitatively characterize unspliced and spliced mRNA (Clark et al., 2002).

In this study, we use the yeast data to investigate the relationship between gene expression and splice site strength. *S. cerevisiae* is one of the most intensively studied eukaryotic model organisms in molecular and cell biology. There are obvious advantages of using *S. cerevisiae* for such a study in addition to what we have mentioned in the previous paragraph. First, the yeast cells replicate rapidly and natural selection should eliminate yeast mutants with inefficient splicing mechanisms. Second, *S. cerevisiae* is the only organism with extensive genome-wide characterization of gene expression at both the transcripts and the protein level. Third, unlike most eukaryotic genomes, the yeast genome has few introns and most of them have been correctly annotated (Spingola et al., 1999; Davis et al., 2000; Clark et al., 2002; Juneau et al., 2007). Fourth, many mRNA processing factors in yeast are not crucial and thus its splicing mechanism is relatively simple compared to higher

eukaryotes (Clark et al., 2002). Fifth, except for a few genes (Irimia et al., 2007), alternative splicing observed in multicellular eukaryotes is rare in the yeast (Davis et al., 2000; Ast, 2004). The splicing mechanism in *S. cerevisiae* appears to be simple even among fungal species, e.g, its genome does not have homologs of U2AF spliceosomal proteins which were found in other fungal species with sequenced genomes (Kupfer et al., 2004). Usually the fewer splicing factors indicate the relatively simple splicing mechanism.

2.3 Materials and Methods

Donor and acceptor sites

The genomic sequences of the 16 chromosomes of *Saccharomyces cerevisiae* were retrieved from ftp.ncbi.nlm.nih.gov/genomes/Fungi/Saccharomyces_cerevisiae. There are 279 annotated introns, with 261 introns from 261 genes each containing a single intron and 18 introns from nine genes (SUS1, VMA9, HMRA1, DYN2, YOS1, RPL7A, AML1, TAD3, and RPL7B) each containing two introns. Some introns from paralogous genes are identical. Genes YBL111C, YHR218W, YLL067C, YLL066C and YML133C are paralogous and contain the same intron. Similarly, genes YIL177C and YJL225C are paralogous and contain the same intron. Genes YRF1-3, YRF1-6 and YRF1-7 are also paralogous and contain the same intron. This creates two problems. The first involves the lack of data independence. The second involves the quantification of mRNA and protein production. Take genes YIL177C and YJL225C for example. It is difficult to know if the mRNA and protein abundance is contributed by only one of the two genes or by both. However, excluding these genes from analysis does not alter the conclusions reached in this thesis.

For each intron, we originally extracted 10 nucleotides (nts) from the exon side and 12 nts from the intron side by using DAMBE (Xia & Xie, 2001). This 10+12 configuration

excluded some splice sites because the first exon in some yeast genes is shorter than 10 nts (Note that, by “first exon”, we refer to the coding part of the first exon in this thesis). For example, the first exon of the two-exon *MUD1* gene is only eight bases long. Because our extraction requires 10 nts on the exon side, the donor site of such genes would be missed. The most extreme is the *RPL20A* and *RPL20B* genes with their first exon being a single nucleotide. The total number of donor sites is 223 in the yeast genome with this requirement of 10+12 configuration, with 205 donor sites from 205 yeast genes containing a single intron, and 18 from the nine yeast genes each containing two introns. Subsequent analysis revealed that only five sites on the exon side of the donor splice site may be important for splicing. For this reason, a donor site in our study consists of 5 nts on the exon site and 12 nts on the intron side. Similarly, an acceptor site consists of 12 nts on the intron side and 5 nts on the exon side. This results in 246 donor sites and 279 acceptor sites.

Some researchers (e.g., Zheng et al., 2005) have taken donor splice site to span from the last 3 nts of the exon to the first 7 nts of the intron. Donor sites defined in this way may produce spurious site patterns in the yeast. For example, as shown in Table 2.1 which lists genes with their donor sites excluded due to too short an upstream exon, 20 *S. cerevisiae* genes have its first exon with exactly three nucleotides (i.e., containing only the initiation codon). Defining the donor site with three nucleotides in the exon side will substantially increase the representation of A, U and G at the three nucleotide sites in the exon side of the splice site. Some yeast introns might have been annotated incorrectly. The annotated intron in the *YJR112W-A* gene is the shortest intron in yeast (49 nts) and does not end with AG. It is possible that the intron is in fact longer with the real acceptor site further downstream. According to SGD annotation, *YJR112W-A* is “Putative protein of unknown function;

identified based on homology to *Ashbya gossypii*". So we did not include its acceptor sites in our analysis. This reduces 279 acceptor sites to 278.

The consensus donor site in the yeast is GUAUGU. Thus, a simple approach to characterization of the donor site strength would be to give donor site a high score if it starts with GUAUGU, but a low score if it is entirely different from this consensus sequence. A more formal approach is to characterize the splice site by position weight matrix (PWM, Hertz & Stormo, 1999; Xia, 2007b, pp. 83-92) and use the PWM score (PWMS) for each splice site as its index of splicing efficiency. This approach has been taken to characterize the splice site strength in human and mouse (Zheng et al., 2005; Dewey et al., 2006). We used DAMBE (Xia & Xie, 2001) to compute PWMS which has been described in detail with numerical illustrations (Xia, 2007b). Two position weight matrices were obtained, one for the donor sites and one for the acceptor sites.

Table 2.1 Yeast genes whose first exon (i.e., the coding part of the first exon) is shorter than five nucleotides.

Gene	1st Exon Len	Sequence
BET4	3	AUG
BOS1	3	AUG
DCN1	3	AUG
MND1	3	AUG
MPT5	3	AUG
PSP2	4	AUGG
QCR9	3	AUG
RPL13A	4	AUGG
RPL13B	4	AUGG
RPL19A	2	AU
RPL19B	2	AU
RPL20A	1	A
RPL20B	1	A
RPL2A	4	AUGG
RPL2B	4	AUGG
RPL30	3	AUG
RPL35A	3	AUG
RPL35B	3	AUG
RPL42A	4	AUGG
RPL42B	4	AUGG
RPL43A	2	AU
RPL43B	2	AU
RPS17A	3	AUG
RPS17B	3	AUG
RPS24A	3	AUG
RPS24B	3	AUG
RPS27A	3	AUG
RPS27B	3	AUG
RPS30A	3	AUG
RPS30B	3	AUG
UBC12	3	AUG
VMA10	3	AUG
YSF3	3	AUG

Characterizing the efficiency of splice sites

The nucleotide frequencies of entire transcripts (i.e., including both exons and introns) were used as background frequencies for computing PWM, with $A = 0.3279$, $C = 0.1915$, $G = 0.2043$, and $U = 0.2763$. Because some site-specific frequencies are 0, a pseudocount with $\alpha = 0.01$ is added to all frequencies to avoid taking Log_2 of 0 (pp. 83-92, Xia, 2007b).

Gene expression

We used three measures of gene expression. The first is codon adaptation index (Sharp & Li, 1987a) with its improved implementation in DAMBE (Xia, 2007c), computed with the reference set of highly expressed yeast genes whose codon usage table is compiled in the Eysc_h.cut file distributed with EMBOSS (Rice et al., 2000). The coding sequences (CDSs) for computing CAI were extracted by using DAMBE. CAI measures the efficiency of translation elongation, but is highly and positively correlated with gene expression at the protein and mRNA level (Duret & Mouchiroud, 1999; Coghlan & Wolfe, 2000).

The second measure of gene expression is the relative mRNA abundance of yeast genes from a previous study using microarray to characterize genome-wide RNA abundance in yeast (Holstege et al., 1998). We downloaded data from http://web.wi.mit.edu/young/pub/data/orf_transcriptome.txt, which is the supplementary data of the paper. The data set includes mRNA levels for 5460 yeast genes. A total of 207 intron donor sites and 231 acceptor sites are in genes with mRNA levels.

The third measure of gene expression is the protein abundance of yeast genes (Ghaemmaghami et al., 2003), downloaded from <http://www.nature.com/nature/journal/v425/n6959/extref/nature02046-s2.xls>. In this dataset, the measured protein levels are reported in terms of molecules/cell. Some of the genes were

tagged with “-”, “%”, or “#”, with “-” indicating no detected expression, “%” indicating a detected band that was unquantifiable because of extremely low signal, and “#” indicating a detected band that was unquantifiable due to experimental problems with the Western blot. We removed all the ORFs flagged with “-”, “%”, or “#”, with the remaining 3868 genes with protein production in the unit of protein molecules/cell. A total of 153 intron donor sites and 174 acceptor sites are in genes with protein levels.

In the mRNA and protein characterization, YAR044W is synonymous to YAR042W in the GenBank file, so is YDR474C to YDR475C, YJL018W to YJL019W, YJL021C to YJL020C, YPR090W to YPR089W, and YFR024C to YFR024C-A. Some genes (YEL068C, YER084W, YHR173C, YIL054W, YJR146W, YLR358C, YNL140C, YNL143C, YNL184C, YOR105W) were annotated in SGD as “Dubious open reading frame unlikely to encode a protein”, and are not annotated at all in the *S. cerevisiae* genome in NCBI. However, they were found to be expressed at both mRNA (Holstege et al., 1998b) and protein levels (Ghaemmaghami et al., 2003). YFL006W and YFL007W have been merged into YFL007W, YJL017W and YJL016W into YJL016W, and YOR087W and YOR088W into YOR087W in the most recent yeast genome annotation.

2.4 Results and Discussions

The position weight matrices (Tables 2.2 and 2.3) confirm previous experimental studies on *S. cerevisiae* that the consensus sequence is GUAUGU for the donor sites and YAG for the acceptor site. The extended consensus sequence is CCCCCN(A/C)AAG|**GUAUGU** for the donor site and UUUUUNU**AYAG**|GC for the acceptor site (Tables 2.2 and 2.3). The rationale for using position weight matrix score (PWMS) as an index of splicing efficiency (Zheng et al., 2005; Dewey et al., 2006) is easy to see with the PWM in Table 2.2. If we

have a sequence whose site-specific nucleotides correspond to the highest PWM_{ij} values in Table 2.2 (where subscript i and j indicate site i and nucleotide j), e.g., a donor sequence such as UAAAGGUAUGUUUAAUU, then it will have the highest possible PWMS (=18.049) given the PWM for the donor site. In contrast, a donor sequence with its site-specific nucleotides corresponding to smallest PWM_{ij} values, e.g., a donor sequence such as CUUGUUGCGCGCGGGCG will have the poorest splice efficiency, with the lowest possible PWMS (= -41.623) given the PWM for the donor site.

Introns with poor PWMSs for their splice sites are spliced by unconventional mechanisms

Donor sites in three genes (*HAC1*, *HFM1*, and *HOP2*) have negative PWMSs (-8.8291, -7.3825, and -7.8898, respectively), and one naturally would wonder how these introns would get recognized and spliced by the spliceosome. It turns out that the splicing of the pre-mRNA of these genes requires either non-spliceosomal mechanisms for intron removal or additional protein factors. *HAC1*, which plays a key role in the unfolded protein response (Chapman & Walter, 1997; Kawahara et al., 1997; Kaufman, 1999), is one of the few yeast genes whose first exon is much longer than the second (661 nts and 56 nts, respectively), and its transcript is processed by an unconventional mechanism (i.e., non-spliceosomal splicing), with the intron cleaved by the protein kinase involved in UPR signaling, Ire1p (Sidrauski & Walter, 1997; Gonzalez et al., 1999; Kaufman, 1999).

The *HFM1* (formerly *MER3*) and *HOP2* are both meiosis-specific genes, with *HFM1* coding for a meiosis-specific DNA helicase (Nakagawa & Ogawa, 1999; Nakagawa & Kolodner, 2002b) that participates in crossover control and unwinding of Holliday junctions (Nakagawa & Ogawa, 1997, 1999; Nakagawa & Kolodner, 2002a, 2002b; Mazina et al., 2004) and *HOP2* coding for a protein essential for forming meiotic synapsis between homologous chromosomes (Leu et al., 1998; Tsubouchi & Roeder, 2002). The splicing of

their transcripts is not constitutive but strictly regulated. The splicing of the *HFM1/MER3* transcripts does not recruit spliceosomes but is regulated by the Mer1p and Bud13p proteins (Spingola & Ares, 2000; Spingola et al., 2004; Scherrer & Spingola, 2006). The spliceosome does not bind *HOP2* transcripts (Davis et al., 2000) and the unspliced *HOP2* transcripts accumulate when the cell is not in meiosis (Moore et al., 2006). The splicing of the *HOP2* transcripts depends heavily on the nuclear exosome component *RRP6* protein, with the loss of *RRP6* dramatically decreasing the splicing efficiency of *HOP2* transcripts (Moore et al., 2006).

Other than the three genes above, the gene with the smallest donor site PWMS is *BUD25* with its PWMS equal to 0.4267. The *BUD25* gene is also implicated in chromosome segregation and meiosis. Most yeast introns can be deleted with no effect, but deletion of *BUD25* intron causes defective growth (Parenteau et al., 2008), suggesting that splicing is important for its function and that splice sites may be under additional constraints other than splicing efficiency. In other words, the splice site of *BUD25* may not be free to evolve towards high splice efficiency.

The acceptor sites of several genes also have negative PWMSs. The intron of the *HAC1* gene, which is spliced by a non-spliceosome mechanism (Sidrauski & Walter, 1997; Gonzalez et al., 1999; Kaufman, 1999), has an acceptor site with the smallest PWMS (-4.4039). The intron whose acceptor site has the second smallest PWMS (-3.4464) belongs to *REC102* which is also a meiosis-specific gene, required for chromosome synapsis (Malone et al., 1991; Bhargava et al., 1992; Jiao et al., 2003). The splicing of its intron also makes use of a non-spliceosome mechanism (Davis et al., 2000).

In short, splice sites with extremely poor PWMSs are spliced by non-spliceosomal mechanisms. This substantiates the use of PWMS as a measure of splicing efficiency in

spliceosome-mediated splicing and suggests a general approach for predicting introns spliced by non-spliceosomal mechanisms.

-1G in donor site and splicing efficiency

The position weight matrix for the donor splice site and the χ^2 -tests (Table 2.2) suggest that -1G in the extended consensus of the donor splice site (Table 2.2, where nucleotides on the exon side is labeled -1, -2, -3, etc.) may be important. Not only is the χ^2 -test highly significant, but the associated PWM entry (= 0.6933) is the largest among sites in the exon side (Table 2.2). The association between -1G and the splicing efficiency is substantiated by two additional lines of evidence. First, if -1G is important, then we should expect that the first exon (i.e., the coding part of the first exon) may terminate at G because the last G in the first exon would then be a -1G. However, we should expect the first exon to rarely terminate at A or AU in the initiation codon because these would result in either a -1A or -1U and should reduce splicing efficiency. This prediction is consistent with data in Table 2.1 that shows few first exons terminating at A or AU of the initiation AUG codon, but many terminating at AUG. Second, for first exons terminating one nucleotide after the initiation AUG, i.e., in the configuration of “AUGN|Intron”, we predict that the last N should be G in order to provide a -1G site. This is again consistent with the data as all seven genes (Table 2.1) in this category have their first exon being AUGG.

Table 2.2 Site-specific frequencies and position weight matrix (PWM) for 246 donor sites. The consensus sequence (UAAAG|GUAUGUUAAU) can be obtained from those large site-specific PWM entries, with the most important sites in *bold italics*. The χ^2 test is performed for each site against the expected background frequencies. The sites are labeled with first intron site as 1.

Site	A	C	G	U	χ^2	p	A	C	G	U
-5	83	30	49	84	10.10	0.0177	0.0525	-0.6332	-0.0260	0.3143
-4	103	44	46	53	10.04	0.0182	0.3613	-0.0878	-0.1162	-0.3434
-3	121	36	38	51	30.01	0	0.5920	-0.3739	-0.3886	-0.3981
-2	122	38	33	53	32.16	0	0.6038	-0.2969	-0.5893	-0.3434
-1	81	40	81	44	28.33	0	0.0177	-0.2238	0.6933	-0.6081
1	0	1	245	0	948.34	0	-6.6464	-5.0056	2.2841	-6.6469
2	0	9	0	237	582.23	0	-6.6464	-2.3190	-6.6480	1.8032
3	239	1	2	4	462.46	0	1.5693	-5.0056	-4.3320	-3.8633
4	16	24	1	205	387.81	0	-2.2655	-0.9496	-5.0680	1.5946
5	2	0	243	1	928.96	0	-4.8476	-6.6483	2.2723	-5.3416
6	9	7	2	228	521.06	0	-3.0427	-2.6612	-4.3320	1.7475
7	87	15	34	110	53.66	0	0.1198	-1.6111	-0.5468	0.7006
8	84	49	30	83	11.71	0.0085	0.0696	0.0659	-0.7246	0.2971
9	111	39	33	63	19.09	0.0003	0.4684	-0.2599	-0.5893	-0.0969
10	106	38	31	71	17.24	0.0006	0.4024	-0.2969	-0.6781	0.0738
11	92	30	40	84	13.69	0.0034	0.1997	-0.6332	-0.3155	0.3143
12	80	38	36	92	14.32	0.0025	-0.0001	-0.2969	-0.4655	0.4445

There are, however, two alternative explanations for the uniform AUGG sequence in the four-base first exon. First, the last G in AUGG has also been labeled as +4G site (with A labeled as site 1) as a part of the Kozak consensus and may be important for translation initiation in multicellular organisms (Kozak, 1986, 1997; but see Xia, 2007a). Thus, the last G in AUGG may be required as a +4G site for translation initiation, not as a -1G site for splicing. Second the last G in AUGG may be constrained by the requirement of a small amino acid at the penultimate site, i.e., the site immediately after the initiating methionine (Xia, 2007a) for the following reason. About half of proteins undergo N-terminal methionine excision (NME) which is efficient only with a small amino acid at the penultimate site. The two smallest amino acids, alanine and glycine, happens to be both coded by G-starting codons (GCN for alanine and GGN for glycine) which would contribute

to the prevalence of +4G (Xia, 2007a). However, translation initiation in the yeast does not require +4G. Indeed, among 5861 protein-coding genes in *S. cerevisiae*, only 1732 (29.55%) have +4G. In addition, the two alternative explanations cannot explain the prevalence of -1G when it is not +4G at the same time (Table 2.2). Thus, it is more plausible to interpret the four-nucleotides first exon in the seven genes in Table 2.1 (all being AUGG) as constrained by the need for a -1G in the donor splice site than the need for a +4G in the translation initiation.

In contrast to -1G, -1U is strongly avoided (Table 2.2). The nucleotide frequencies of all yeast transcripts are 0.2043 for G but 0.2763 for U and we should expect more -1U if there is no selection. However, among the 246 donor sites, 81 have -1G and only 44 have -1U.

ICGs (intron-containing genes) with a -1G in the donor site produce more mRNA than those with a -1U in the donor site. Among 141 ICGs with both mRNA and protein levels available, 49 have introns with -1G in the donor site and have a mean mRNA level equal to 1.6345 (log-transformed value). In contrast, 23 genes have -1U in the donor site and have a mean mRNA level equal to 0.6383. The difference is significant ($T = 2.448$, $DF = 70$, $p = 0.0084$, one-tailed test). Introns with -1G at the donor site are also associated with more protein production than introns with -1U at the donor site (8.8629 vs 8.1676, log-transformed values), but the difference is not statistically significant ($T = 1.667$, $DF = 70$, $p = 0.0510$, one-tailed test).

We should mention that there is a striking difference of gene length between yeast and the higher eukaryotes. The average yeast gene is 1.4 kb but the average human gene is 27 kb. Moreover, only a few yeast genes are longer than 5 kb but many genes in flies or mammals have the length greater than 5 kb.

A prominent poly-U upstream of the AG dinucleotide in the acceptor site

It is worth noting that high-efficiency introns in the yeast are characterized by a poly-U tract upstream of the 3' AG (Table 2.3). Such a poly-U tract can increase the efficiency of acceptor splice site has previously been demonstrated in *S. cerevisiae* (Patterson & Guthrie, 1991), especially in introns with a long distance between the branch point and the acceptor splice site (Parker & Patterson, 1987). A recent study of intron splicing of mammalian genes in YACs (yeast artificial chromosome) further suggests the importance of the poly-U tract upstream of the acceptor splice site in *S. cerevisiae* (Kunze et al., 2000). Previous compilations of yeast introns (Jackson, 1991; Sharp & Burge, 1997) have missed the poly-U tract upstream of the acceptor splice site. Thus, the poly-U tract upstream of the acceptor splice site has not been included as a feature of *S. cerevisiae* intron in molecular biology textbooks (e.g., Weaver, 2005, p. 428).

Table 2.3 Site-specific frequencies and position weight matrix (PWM) for 278 acceptor sites. The consensus sequence (UUUUUUUUAYAG|GCUUC) can be obtained from those large site-specific PWM entries, with the most important sites in *bold italics*. The χ^2 test is performed for each site against the expected background frequencies. The sites are labeled with first intron site as 1.

Site	A	C	G	U	χ^2	p	A	C	G	U
-12	61	53	34	130	56.0	0	-0.5608	0.0033	-0.7205	0.7648
-11	70	47	20	141	83.2	0	-0.3649	-0.1682	-1.4696	0.8813
-10	79	42	12	145	99.9	0	-0.1926	-0.3285	-2.1802	0.9214
-9	38	30	23	187	219.4	0	-1.2308	-0.8067	-1.2731	1.2867
-8	51	42	27	158	121.6	0	-0.8149	-0.3285	-1.0470	1.0447
-7	91	33	28	126	53.8	0	0.0093	-0.6715	-0.9956	0.7200
-6	95	42	35	106	22.0	0.0001	0.0707	-0.3285	-0.6794	0.4722
-5	93	33	23	129	63.3	0	0.0403	-0.6715	-1.2731	0.7537
-4	136	25	38	79	43.3	0	0.5842	-1.0647	-0.5626	0.0517
-3	12	121	0	145	272.3	0	-2.8223	1.1862	-6.6480	0.9214
-2	277	1	0	0	563.7	0	1.6056	-5.1232	-6.6480	-6.6469
-1	0	0	278	0	1082.7	0	-6.6464	-6.6483	2.2900	-6.6469
1	93	37	73	75	9.7	0.0217	0.0403	-0.5089	0.3691	-0.0226
2	72	64	54	88	8.0	0.0466	-0.3248	0.2729	-0.0619	0.2059
3	90	54	48	86	2.5	0.4771	-0.0065	0.0300	-0.2299	0.1730
4	83	43	54	98	8.7	0.0337	-0.1221	-0.2950	-0.0619	0.3599
5	90	65	37	86	10.6	0.0140	-0.0065	0.2951	-0.6005	0.1730

The poly-U tract upstream of the yeast acceptor site is different from the polypyrimidine tract often present upstream of the acceptor site in multicellular eukaryotes as well as in *Schizosaccharomyces pombe*. In multicellular eukaryotes and *S. pombe*, the polypyrimidine tract upstream of the acceptor splice site is important for splicing efficiency (Romfo & Wise, 1997) and is recognized by the essential U2AF splicing factor (Sridharan & Singh, 2007). However, there is no homolog of U2AF proteins found in *S. cerevisiae* (Kupfer et al., 2004).

Gene expression and splicing efficiency

We have previously argued that highly expressed genes should be selected to have more efficient splice sites than lowly expressed genes. Although lowly expressed genes may also have highly efficient splice sites, the selection for maintaining the high-efficiency should be weak and some of their splice sites will accumulate mutations and become less efficient (Figure 2.1). We predicted that lowly expressed genes will, on average, have lower splicing efficiency (as measured by PWMS) but have greater variance in PWMS than highly expressed genes (Figure 2.1). The greater variance in lowly expressed genes than highly expressed genes is because lowly expressed genes may have both high and low PWMS but highly expressed genes should have only high PWMS.

Table 2.4 Testing the predictions that introns in highly expressed genes have higher PWMS and smaller variance in PWMS than in lowly expressed genes, with gene expression measured by CAI, mRNA and protein abundance. Introns spliced by non-spliceosomal mechanisms are excluded. Mann-Whitney tests generate similar results. All tests are two-tailed.

	Donor			Acceptor		
	CAI	lnMRNA ⁽¹⁾	lnPROT ⁽²⁾	CAI	lnMRNA	lnPROT
N ⁽³⁾	81	68	51	91	76	58
MeanH ⁽⁴⁾	11.5881	11.5799	11.7856	9.0314	9.0410	8.5511
MeanL ⁽⁵⁾	10.0642	10.5251	10.9747	5.5586	5.7834	6.0942
DF ⁽⁶⁾	126	104	87	147	146	100
T	4.1298	2.6119	1.9077	9.7074	9.1608	5.7687
p	0.0001	0.0103	0.0597	0.0000	0.0000	0.0000
VarH ⁽⁷⁾	2.6699	2.5462	2.7952	3.0599	3.0214	3.3178
VarL ⁽⁸⁾	8.3582	8.5451	6.4200	8.4958	7.4930	8.0335
F	3.1306	3.3560	2.2968	2.7765	2.4800	2.4213
p	0.0000	0.0000	0.0020	0.0000	0.0001	0.0005

(1) Natural logarithm of mRNA abundance

(2) Natural logarithm of protein abundance

(3) Number of splice sites in the highly expressed and lowly expressed groups (Note that $N_1 = N_2 = N$).

(4) Mean PWMS in highly expressed group

(5) Mean PWMS lowly expressed group

(6) The t-test assuming unequal variance is used. So DF is not equal to $(N_1 + N_2 - 2)$.

(7) Variance in the highly expressed group

(8) Variance in the lowly expressed group

The splice efficiency characterized by PWMS exhibited expected relationship with gene expression when the latter is measured by either CAI (Figure 2.2), mRNA abundance (Figure 2.3) or protein abundance (Figure 2.4). In all cases we see graphically the trend that highly expressed genes have, on average, higher PWMS values (indicating higher splicing efficiency) than lowly expressed genes. In addition, lowly expressed genes have greater variation in PWMS values than highly expressed genes. To statistically test the differences in mean and variance, we have ranked genes by gene expression, i.e., ranked separately by CAI, mRNA abundance or protein abundance. For each ranking, we designate one third of the genes with the highest expression values (i.e., highest CAI, mRNA or protein abundance, respectively) as the high expression group and another one third of the genes

with the lowest expression values as the low expression group, and tested the differences in mean PWMS and the variance of PWMS between the two groups. As shown in results summarized in Table 2.4, the two predictions are consistently supported, i.e., (1) the highly expressed genes have significantly greater mean PWMS values than lowly expressed genes and (2) the highly expressed genes have significantly smaller variance in PWMS than the lowly expressed genes (Table 2.4). The t-tests used assume unequal variance between the two groups. The tests for differences in variance between the two groups are regular variance ratio F-test (Zar, 1999, pp. 136-139).

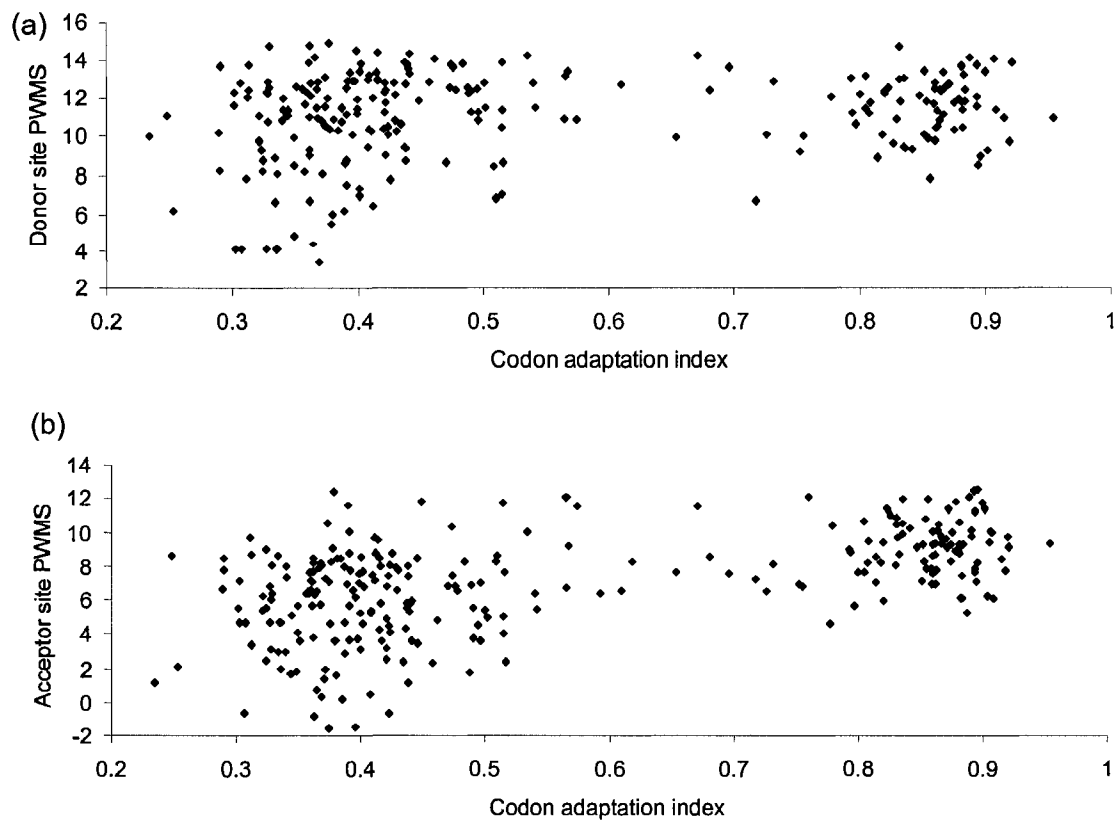


Figure 2.2 Relationship between splicing efficiency measured by position weight matrix score (PWMS) and gene expression measured by codon adaptation index. Genes known to be spliced by non-spliceosomal mechanisms are excluded.

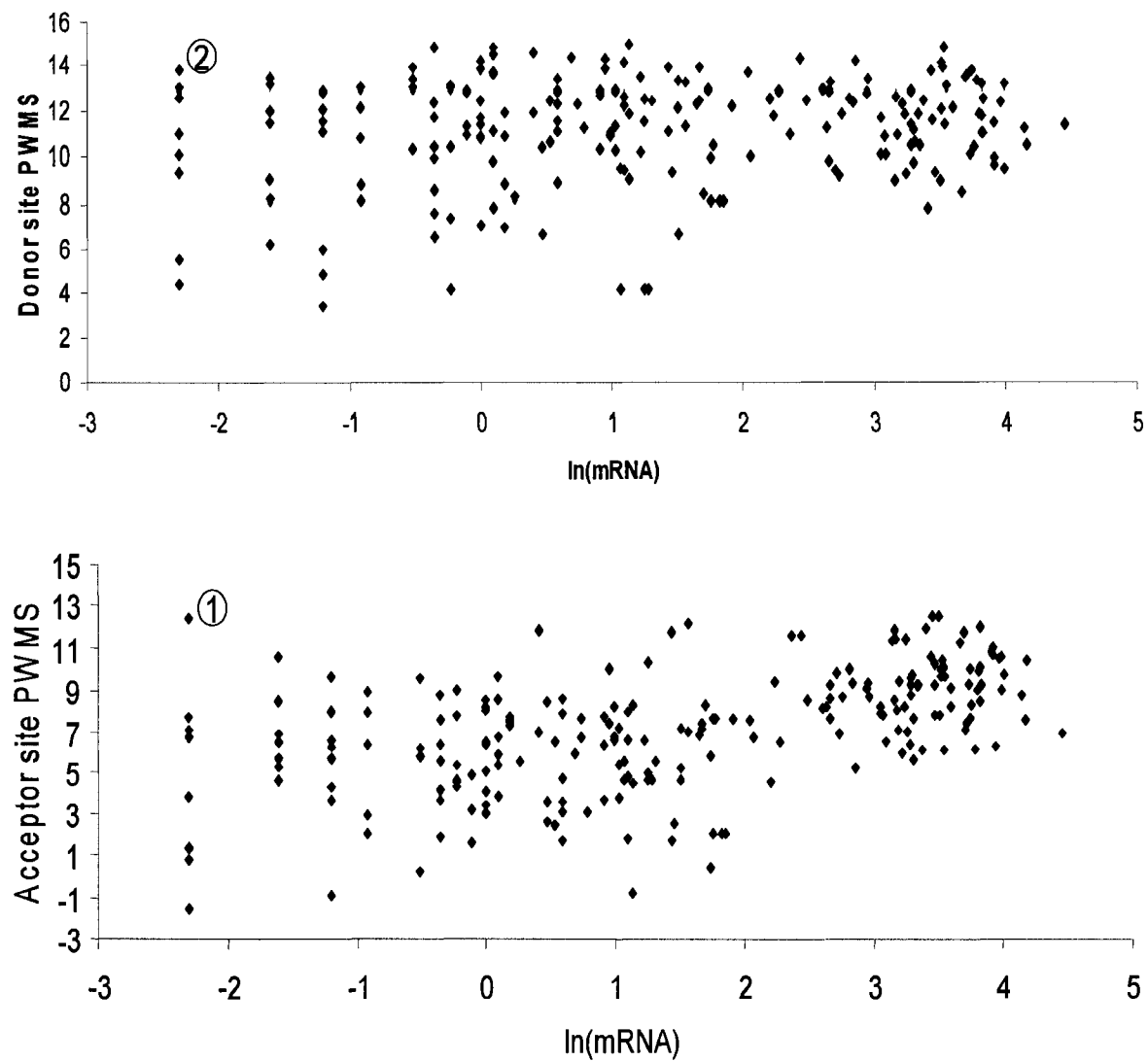


Figure 2.3 Relationship between splicing efficiency measured by position weight matrix score (PWMS) and gene expression measured by mRNA abundance. Genes known to be spliced by non-spliceosomal mechanisms are excluded.

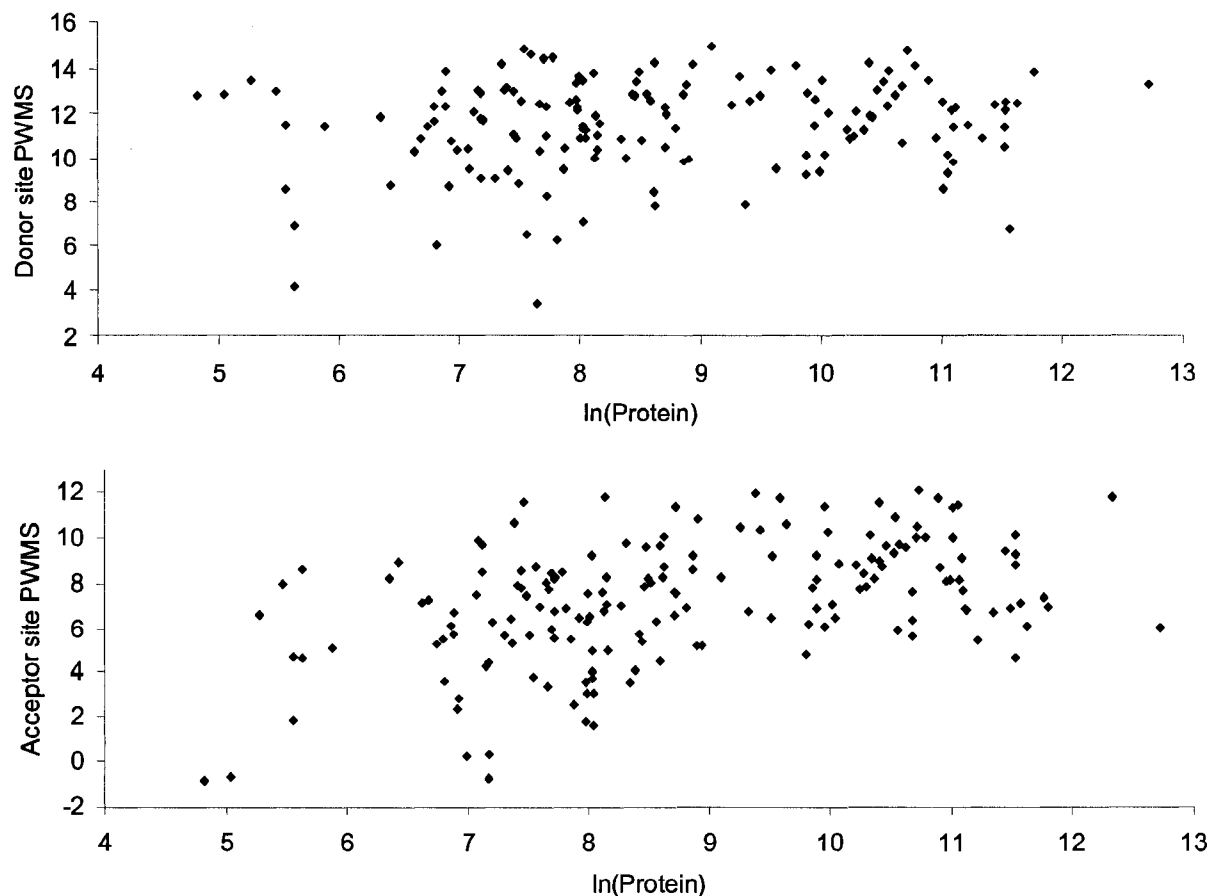


Figure 2.4 Relationship between splicing efficiency measured by position weight matrix score (PWMS) and gene expression measured by protein abundance. Genes known to be spliced by non-spliceosomal mechanisms are excluded.

There are a few obvious outlying points in Figure 2.3. The point labeled by the number 1 circle is the *OSW2* gene. The deletion mutant of this gene is viable (Giaever et al., 2002), but the gene appears to be required for a very late step in spore wall synthesis (Coluccio et al., 2004). This suggests that the gene is not normally expressed, with little mRNA (Holstege et al., 1998) and no detectable protein (Ghaemmaghani et al., 2003). The point would have shifted to the right (making it no longer appear an outlier) had the mRNA and protein abundance determined during spore wall synthesis when the gene is expressed. The point labeled by the number 2 circle in Figure 2.3 is the *DID4* gene. The deletion mutant of this gene is also viable (Giaever et al., 2002), but the gene is needed on a non-

fermentable carbon source (Steinmetz et al., 2002). This again explains why it has little mRNA (Holstege et al., 1998) and no detectable protein (Ghaemmaghami et al., 2003). Had the mRNA and protein abundance been determined with yeast cells on non-fermentable carbon source, the point would have shifted to the right of the graph. In general, genes with a high PWMS but low mRNA and protein abundance are genes expressed only at specific growth conditions or specific time. For this reason, CAI has its advantage as an index of gene expression because it is not affected by when and where the gene is expressed.

Given that yeast has few and short introns, one may not be certain whether yeast can be used as a representative of eukaryotes. However, using yeast to investigate the relationship between gene expression and splicing efficiency is definitely a good starting point since yeast shares many core spliceosome functions with other multicellular eukaryotes such as human and there are many research advantages in the yeast mentioned before. Usually it is wise to do research starting from a relatively simple way rather than from a very complicated way. Our in-depth understanding for the relationship between gene expression and splicing efficiency in the yeast with fewer introns and relatively simpler splicing mechanism can provide a solid foundation for the further investigations on other eukaryotes such as human with many introns and the complex splicing mechanism.

We should mention that splicing efficiency depends not only on the donor and acceptor sites, but also on the branchpoint sequence (BPS, Reed & Maniatis, 1988; Chiara et al., 1996; BPS, Simpson et al., 1996) as well as the spacing between BPS and the donor and acceptor splice site (Cellini et al., 1986; Brys & Schwer, 1996; Luukkonen & Seraphin, 1997). The effect of selection on BPS will be illustrated in the next chapter (Chapter 3).

CHAPTER 3 GENE EXPRESSION AND BRANCHPOINT SEQUENCE IN YEAST

3.1 Abstract

Other than the donor and acceptor splice sites, splicing efficiency also depends on the branchpoint sequence (BPS) as well as the two distances between BPS and the two splice sites. Here we study the relationship between gene expression and BPS in the yeast, *Saccharomyces cerevisiae*. The results showed that the BPS strength (measured by position weight matrix scores) of highly expressed yeast ICGs is significantly higher than that of lowly expressed yeast ICGs. In addition, highly expressed yeast ICGs have significantly longer distance between the donor splice site and BPS (S1 distance) and slightly longer distance between BPS and the acceptor splice site (S2 distance) than lowly expressed yeast ICGs. The long S1 distance of highly expressed yeast ICGs does not indicate the potential of enhancing splicing efficiency through forming secondary structure in the region between the donor splice site and BPS.

3.2 Introduction

In yeast, pre-mRNA splicing is divided into two steps: 1) cleavage at 5' intron-exon boundary and formation of the 2'-5' branched lariat intermediate, and 2) cleavage at 3' intron-exon boundary and ligation of the exons (Brys & Schwer, 1996). Yeast BPS is involved in both steps and thus plays a crucial role in yeast pre-mRNA splicing (Cellini et al., 1986; Parker et al., 1987; Thompson-Jager & Domdey, 1987; McPheeters & Muhlenkamp, 2003). In the first step, base pairing between BPS and U2 snRNA is required for accurate and efficient recognition of the donor splice site and the branch point, and later

cleavage at the donor splice site and the formation of a lariat intermediate (de Roos, 2005; Rogozin et al., 2005). In the second step, selection of the acceptor splice site is accomplished with the recognition of BPS in the previous step and usually the first AG downstream of BPS is selected as the last two nucleotides at the 3' end of intron (McPheeters & Muhlenkamp, 2003).

Most yeast BPSs are conserved with UACUAAC motif due to its crucial importance in pre-mRNA splicing. This importance was also demonstrated by a previous study in which splicing can be almost completely eliminated if BPS was mutated even at only one site from UACUAAC to UACUACC (Jacquier & Rosbash, 1986). Other than the highly conserved BPS motif UACUAAC (>80%), there are at least seven other types of BPS motifs such as GACUAAC accounting for only ~20% (Bon et al., 2003). Here we designate UACUAAC as the conserved BPS and all other BPSs as the non-conserved BPS in yeast. Usually, the conserved BPS and the non-conserved yeast BPSs differ at only one site (Bon et al., 2003).

Despite the fact of prevalence of the conserved BPS in yeast intron, it is very interesting to see if the distribution of the non-conserved BPSs is different between highly and lowly expressed yeast ICGs groups. We expect that just a few introns in highly expressed yeast ICGs have non-conserved BPSs under strong selection pressure to evolve towards high splicing efficiency. On the contrary, selection pressure is relatively weak for lowly expressed yeast ICGs to evolve towards high splicing efficiency and thus a relatively larger proportion of introns in lowly expressed genes may have non-conserved BPS. In the first step of this study, we compared the BPS strength between the two groups of highly and lowly expressed yeast ICGs. We predict that the BPS strength of highly expressed yeast ICG group is higher than that of lowly expressed yeast ICG group.

Other than the splicing signals such as the splice sites and BPS, splicing efficiency also depends on the spacing between the two (donor and acceptor) splice sites and BPS (Cellini et al., 1986; Thompson-Jager & Domdey, 1987). Previous studies revealed that, in yeast, there is a minimum requirement for the S1 distance (Thompson-Jager & Domdey, 1987), while the S2 distance is constrained to be relatively small and splicing can be impaired or even completely abolished if this distance is enlarged by the insertion of a random sequence with a certain length (e.g., 15 nts) (Cellini et al., 1986). The early explanation of these phenomena is physical constraints of the splicing machine (Thompson-Jager & Domdey, 1987). Meanwhile, the S1 and S2 distances can be flexible. Although the suggested minimum S1 distance in *S.cerevisiae* is 40 nts (Thompson-Jager & Domdey, 1987; Kohrer & Domdey, 1988), some yeast ICGs can have short S1 distance such as *TAN1* (30 nts) with normal splicing. Also some yeast ICGs can have long S2 distances such as *RPL40B* (75 nts) with normal splicing. The mechanism for the flexibility of the S1 and S2 distances remains unclear although some clues are emerging. Brys and Schwer revealed that yeast pre-mRNA splicing factor SLU7 is dispensable for splicing of mRNAs in which the acceptor splice site is very close to the branch point, but it is required when the S2 distance is greater than a certain distance (> 7 nts) (Brys & Schwer, 1996).

However, such early findings still did not answer many questions regarding the S1 and S2 distances. In particular, the relationship between gene expression and the two distances (S1 and S2) in yeast ICGs has not been studied. We try to characterize the relationship in this project.

3.3 Materials and Methods

BPS and the S1 and S2 distances

The BPSs, the S1 and S2 distances, and the length for each yeast intron were retrieved from http://mips.gsf.de/cgi-bin/proj/yeast/intron_Sc_list.pl?tab=splice (Guldener et al., 2005). The original data retrieved from the above link contained 275 yeast spliceosomal introns. We deleted those introns annotated as having no experimental evidence. We also deleted those introns mis-annotated or spliced by non-spliceosome mechanism (*HAC1*, *HFMI*, *HOP2*, *BUD25* and *YJR112W-A*). Thus, finally we had a total of 257 yeast introns in our dataset including information on their BPS, the intron length, and the S1 and S2 distances.

Gene expression

We used three measures of gene expression. The first is codon adaptation index (CAI) (Sharp & Li, 1987) with its improved version (Xia, 2007c) implemented in DAMBE (Xia, 2001; Xia & Xie, 2001). CAI value is highly and positively correlated with gene expression at the protein and mRNA level in yeast (Duret & Mouchiroud, 1999; Coghlan & Wolfe, 2000). We retrieved the genomic sequences of the 16 chromosomes of *S.cerevisiae* from ftp.ncbi.nlm.nih.gov/genomes/Fungi/Saccharomyces_cerevisiae. Then we extracted coding sequences (CDSs) and calculated their corresponding CAI values by using DAMBE (Xia & Xie, 2001b).

The relative mRNA abundance of yeast genes is used as the second measure of gene expression. Holstege and colleagues investigated genome-wide RNA abundance in yeast using microarray (Holstege et al., 1998) and their gene expression data were downloaded from http://web.wi.mit.edu/young/pub/data/orf_transcriptome.txt. The dataset includes gene

expression data for 5460 yeast genes but we only can find the gene expression data in mRNA level for 217 out of total 257 yeast ICGs in our dataset.

For the third measure of gene expression, we used the protein abundance of yeast genes (Ghaemmamghami et al., 2003) and downloaded the gene expression data from <http://www.nature.com/nature/journal/v425/n6959/extref/nature02046-s2.xls>. The dataset includes gene expression data in protein level for 3868 yeast genes after we removed genes with unquantified or unquantifiable gene expression. Finally, we have a total of 169 introns in genes with gene expression data in protein level.

BPS strength

Similar to the method in Chapter 2, we used position weight matrix (PWM) to characterize the BPSs and the PWM score (PWMS) for each BPS as its index of strength (PWM, Hertz & Stormo, 1999; Xia, 2007b, pp. 83-92). The BPS motif UACUAAC is considered to be the optimal BPS with the highest splicing efficiency in the yeast, and BPS that deviate more from this optimal motif is assumed to be less efficient in splicing. The PWMSs of the BPSs were computed by DAMBE with the background nucleotide frequencies of all intron sequences, i.e., A=0.3238, C=0.16398, G=0.17372 and U=0.33848.

Minimum folding energy

We used minimum folding energy (MFE) to measure the potential of forming secondary structure for a given sequence. All yeast introns were extracted by DAMBE, and then the intronic sequences between the donor splice site and the BPS (hereafter referred to as the S1 sequence) were further extracted by a Perl script. The MFE values of the S1 sequences were computed by DAMBE which incorporated the Vienna RNA package (Hofacker et al., 1994; Hofacker, 2003).

3.4 Results and Discussions

There are ten different types of BPSs, but most ICGs share the BPS of UACUAAC (Table 3.1). Among 257 yeast ICGs, 212 (82.5%) share a BPS of UACUAAC (Table 3.1). Similar percentages are observed for the 217 ICGs with transcriptomic data and 169 ICGs with proteomic data (Table 3.1). Our subsequent analyses and inferences are based on the assumption that the highly conserved BPS with sequence UACUAAC is the most efficient BPS.

Table 3.1 Frequency distribution of genes containing the ten different BPS sequences.

BPS	N _{CAI} ⁽¹⁾	N _{RNA} ⁽²⁾	N _{Prot} ⁽³⁾	PWMS ⁽⁴⁾
UACUAAC	212	183	141	12.9120
AACUAAC	11	8	7	9.1952
AACUAAU	5	5	4	3.4202
CACUAAC	8	5	5	9.3415
GACUAAC	8	8	5	9.0935
UACUAAU	4	2	3	7.1370
UACUGAC	2	1	0	6.9322
UAUUAAC	3	2	1	5.6185
UGCUAAC	3	3	3	7.8847
CGCUAAC	1	0	0	4.3142
Sum	257	217	169	

(1) Frequency distribution of the ten different BPSs from 257 ICGs with CAI computed.

(2) Frequency distribution of the ten different BPSs from 217 ICGs with RNA data

(3) Frequency distribution of the ten different BPSs from 169 ICGs with proteomic data.

(4) Position weight matrix score.

ICGs with highly conserved BPS are more highly expressed than ICGs with less conserved BPS

Yeast ICGs with BPS sharing the highly conserved UACUAAC are more highly expressed when gene expression is measured by CAI. The mean CAI is 0.6007 for ICGs with the conserved BPS (UACUAAC, N = 212) and 0.4762 for the rest of ICGs (N = 45). The

difference is highly significant based on a two-sample t-test ($t = 3.4104$, $DF = 255$, $p = 0.0004$, one-tailed test).

The result is similar when gene expression is measured by transcript abundance. The mean $\ln\text{RNA}$ (log-transformed RNA abundance) is 1.7504 for ICGs with the conserved BPS (UACUAAC, $N = 183$) and 0.5801 for the rest of ICGs ($N = 34$). The difference is highly significant based on a two-sample t-test ($t = 6121$, $DF = 215$, $p = 0.0002$, one-tailed test).

No significant difference in gene expression is found between the two groups of ICGs when gene expression is measured by protein abundance. While the mean $\ln\text{Protein}$ (log-transformed protein abundance) for ICGs with UACUAAC BPS (8.9542, $N = 141$) is higher than that for the rest of the ICGs (8.5244, $N = 28$), the difference is not significant ($t = 1.2444$, $DF = 167$, $p = 0.1075$, one-tailed test). This partially may be due to the smaller sample size for the protein data relative to the CAI data and RNA data (Table 3.1). In addition, some yeast proteins may have high degradation rates, i.e., although they are produced in large quantities, they are degraded very quickly so we may not see high abundance for these proteins. A small number of such proteins would lead to poor correlation between splicing efficiency and protein production.

An alternative way of checking the association between gene expression and BPS conservation is to check whether highly expressed genes have a high chance of possessing a UACUAAC BPS. We ranked ICGs by gene expression from the highest to the lowest and took the top and bottom one third as the high-expression and low-expression ICG groups. We then counted the number of UACUAAC-containing ICGs in each group. Our results showed that high-expression ICGs indeed are more likely to have a UACUAAC BPS than low-expression ICGs (Table 3.2). For example, when gene expression is measured by CAI, 80 out of the 86 ICGs in the high-expression group have a UACUAAC BPS, whereas only

63 out of the 86 ICGs in the low-expression group have a UACUAAC BPS (Table 3.2). The difference is highly significant when tested with a χ^2 -test ($\chi^2 = 11.9865$, $p = 0.0005$, Table 3.2). The difference is also highly significant when gene expression is measured by lnRNA (Table 3.2). While no significant difference is found for the protein data, the difference is still in the right direction, i.e., the high-expression group has more ICGs containing a UACUAAC BPS than the low-expression group (Table 3.2).

Table 3.2 Highly expressed ICGs are more likely to harbor a highly conserved UACUAAC BPS.

GEI ⁽¹⁾	BPS ⁽²⁾	High ⁽³⁾	Low ⁽³⁾	χ^2	p
CAI	UACUAAC	83	68	11.9865	0.000536
	Other	6	21		
lnRNA	UACUAAC	68	53	11.64211	0.000645
	Other	4	19		
lnProtein	UACUAAC	49	43	2.191304	0.138792
	Other	7	13		

(1) Gene expression index, quantified by CAI, log-transformed RNA abundance (lnRNA) and log-transformed protein abundance (lnProtein).

(2) Branchpoint sequence, either UACUAAC or not.

(3) High: high-expression ICGs (top one third when gene expression is ranked from the highest to the lowest); Low: low-expression ICGs (bottom one third).

Given that we have quantified the splicing strength of BPSs by the position weight matrix scores (PWMSs), we have also checked our expectation that that highly expressed ICGs should have stronger BPSs (greater PWMSs). This expectation is generally confirmed by Mann-Whitney tests between the high-expression and low-expression groups (Table 3.3). The reason for using the non-parametric Mann-Whitney test instead of the parametric t-test is that the frequency distribution of PWMS is not normal. The difference is significant for the data sets with gene expression measured by CAI or RNA abundance (Table 3.3, $p = 0.012$ for the CAI data set and $p = 0.0332$ for the RNA data set). The difference is in the

right direction but not significant for the data set with gene expression measured by protein abundance ($p = 0.1729$). However, this might be due to the relatively small sample size and high degradation rate of a number of proteins. Overall, the result is consistent with our prediction that highly expressed yeast ICGs should have their BPSs more strongly constrained by purifying selection and should maintain more efficient splicing signals.

Table 3.3 Testing the prediction that the BPS strength measured by PWMS for the group of highly expressed genes is higher than that for the group of lowly expressed genes, with gene expression measured by CAI, mRNA and protein abundance. All statistical tests were performed by the Mann-Whitney test.

GEI ⁽¹⁾	N ⁽²⁾	W	p
CAI	89	8743	0.012
lnRNA	72	5680	0.0332
lnProtein	56	3326	0.1729

(1) Gene expression index, quantified by CAI, log-transformed RNA abundance (lnRNA) and log-transformed protein abundance (lnProtein).

(2) Number of BPSs in the highly expressed and lowly expressed groups

Highly expressed yeast ICGs have much longer S1 distance and slightly longer S2 distance than lowly expressed yeast ICGs

We tested the difference in the S1 and S2 distances between the high expression group and the low expression group by using the Mann-Whitney test. The results revealed that highly expressed yeast ICGs have much longer S1 distances and slightly longer S2 distances than lowly expressed yeast ICGs (Table 3.4). We chose the non-parametric test (Mann-Whitney) test because the data of the S1 and S2 distances is much biased from normal distribution. From the results, we can see the Mann-Whitney test uses median instead of mean to measure central tendency of data. If the data is very biased from normal distribution, an extremely high or low value can greatly affect the mean, but the median would not be affected by this

kind of values. This is why median is used to measure central tendency of data in the non-parametric test.

Table 3.4 Testing the differences of the S1 and S2 distances between highly and lowly expressed yeast ICGs, with gene expression measured by CAI, mRNA and protein abundance. All statistical tests were performed by the Mann-Whitney test.

	S1 distance			S2 distance		
	CAI	ln(mRNA) ⁽¹⁾	ln(PROT) ⁽²⁾	CAI	ln(mRNA)	ln(PROT)
N ⁽³⁾	89	72	56	89	72	56
MedianH ⁽⁴⁾	355	348	338.5	33	34	33
MedianL ⁽⁵⁾	49	49	52.5	28	28	26
W	11546	7515	4488.5	8612.5	6039	3593.5
p	<0.0001	<0.0001	<0.0001	0.06	0.0011	0.0125

(1) Natural logarithm of mRNA abundance

(2) Natural logarithm of protein abundance

(3) Number of the S1 or S2 distance in the highly expressed and lowly expressed groups

(4) MedianH: Median PWMS in highly expressed group

(5) MedianL: Median PWMS in lowly expressed group

Note that S1 length is almost perfectly correlated with intron length in yeast ICGs, with $r = 0.9908$ (Figure 3.1), in contrast to the S2 distance which is only weakly correlated with intron length ($r = 0.2038$). Thus, the observation that highly expressed yeast ICGs have significantly longer S1 and S2 distances than lowly expressed ones implies that highly expressed yeast ICGs have longer introns than lowly expressed ones. If we replace S1 distance in Table 3.4 by intron length, the p values from significance tests are similar, i.e., highly expressed yeast ICGs have significantly longer introns than lowly expressed ones. A similar positive association between gene expression and intron length in yeast has been reported before (Juneau et al., 2006). This positive association in yeast is contrary to the pattern documented in other species in previous studies (Comeron, 2004; Fahey & Higgins, 2007; Li et al., 2007) in which highly expressed genes tend to have shorter introns. Because transcribing introns costs time and energy, it is reasonable to expect that highly expressed genes should have shorter introns than lowly expressed genes (Castillo-Davis et al., 2002; Li

et al., 2007). Our observation that highly expressed yeast ICGs have much longer introns than lowly expressed genes suggest that the conventional evolutionary explanation is insufficient. Several previous studies have noted that genes coding for ribosomal proteins tend to be highly expressed and have long introns (Li et al., 1996), and some of these long introns harbor snoRNAs, which are involved in, and promote, pre-mRNA splicing (Bachellerie et al., 1995; Villa et al., 1998). Thus, the association between gene expression and intron length may be attributed to the presence of highly expressed ribosomal mRNA with long introns. However, we still have the question of why highly expressed genes coding for ribosomal proteins should have long introns.

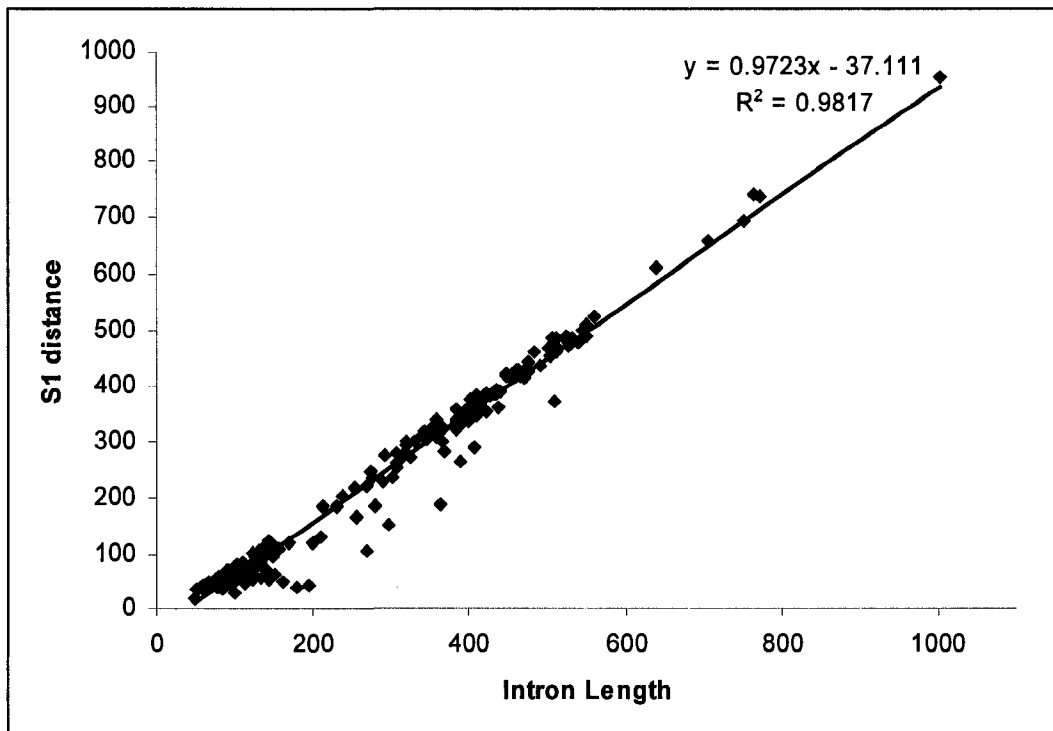


Figure 3.1 Correlation between S1 distance and intron length in yeast intron-containing genes (ICGs).

Highly expressed yeast ICGs also have significantly longer S2 distances than lowly expressed genes, although the difference is relatively small compared to that involving S1 distances (Table 3.4). A previous study (Cellini et al., 1986) has explored the effect of

lengthening or shortening the S2 distance and found that the lengthening tend to diminish splicing efficiency whereas shortening the S2 distance by 15 nts has no detectable effect on splicing efficiency. Superficially, this would suggest that introns with shorter S2 are more efficient in splicing. However, the study involves only a few experimental intron constructs, and the S2 distance is substantially greater than the median S2 for high-expression ICGs (33-34 nts in Table 3.4). Our results provide a better interpretation of the findings in Cellini et al. (Cellini et al., 1986). If optimal S2 distance is around 33 or 34 nts, then elongating an S2 sequence that is already substantially longer the optimal length will decrease splicing efficiency, but deleting a segment in the S2 sequence to make it closer to the optimal length or slightly below the optimal length will either increase its splicing efficiency or have little effect on splicing efficiency.

Several studies suggest that the first AG downstream of BPS is usually chosen as the acceptor splice site (Chen et al., 2000; McPheeters & Muhlenkamp, 2003). Such a suggestion misses the importance of the S2 distance in intron splicing. Our results show that S2 distance is related to gene expression and should be studied further experimentally.

Highly expressed yeast ICGs are not richer in secondary structure than lowly expressed genes

It is puzzling that highly expressed genes have longer introns than lowly expressed genes (Table 3.4). As we have mentioned before, transcribing introns costs time and energy, and one would expect highly expressed genes to have shorter introns to save time and energy. In several multicellular eukaryotic species, highly expressed genes indeed have shorter introns (Comeron, 2004; Fahey & Higgins, 2007; Li et al., 2007). Why do yeast ICGs defy this simple evolutionary explanation?

One possibility is that intron splicing in yeast requires the formation of certain secondary structure in the S1 sequence region and that the formation of the secondary structure requires the S1 sequence to be above a certain length. Secondary structure in the S1 sequence region has been reported before but its function has been interpreted to enhance splicing efficiency by bringing the donor splice site close BPS (Spingola et al., 1999). This interpretation is substantiated by experimental studies of yeast *rp51b* gene in which intron splicing is enhanced through forming secondary structure by two intronic complementary regions near (downstream) the donor splice site and the (upstream) BPS, with the two regions being designated as UB1 and DB1 respectively (Libri et al., 1995; Charpentier & Rosbash, 1996; Libri et al., 2000). This interpretation, however, has two problems. First, it assumes that splicing efficiency can be enhanced when donor splice site is close to BPS. This assumption does not seem to be true given our finding that highly expressed genes have longer S1 distances in yeast ICGs. Second, if it is important for the donor splice site to be close to BPS, then a simple way to achieve it is to evolve shorter S1 sequences which should be more reliable to bring the donor site close to BPS than having a long S1 sequence with a lot of twists and bends. A more plausible explanation is that secondary structure formed at the S1 sequence region enhance the recognition of the splice sites and hence increase the splicing efficiency. It is known that splicing efficiency is reduced if these potential bases pairing is disrupted in the S1 sequence region (Libri et al., 1995; Charpentier & Rosbash, 1996).

If secondary structure is indeed important in the S1 region, then we expect highly expressed yeast ICGs to have more secondary structure (i.e., be richer in secondary structure) than lowly expressed genes. To substantiate this expectation, we calculated minimum folding energy (MFE) of the intronic sequence in the S1 region to see if highly

expressed genes are richer in secondary structure. Folding energy is conceptually defined as the difference in free energy between the unfolded and folded state (Ringner & Krogh, 2005) and is negatively correlated with the secondary structure stability because the free energy of a base pair is negative (Ringner & Krogh, 2005). Given a mRNA sequence, minimum folding energy can be used to measure the potential of forming a stable secondary structure for this sequence (Juan & Wilson, 1999; Clote et al., 2005). Thus, the expectation that highly expressed yeast ICGs should have more secondary structure than lowly expressed yeast ICGs is reduced to the expectation that highly expressed ICGs have more negative MFE than lowly expressed ICGs.

To properly test the expectation that highly expressed genes are richer in secondary structure than lowly expressed genes in the S1 sequence, it is important to control for S1 distance because MFE increases with sequence length. Note that, although the number of secondary structure is expected to increase exponentially with sequence length, the relationship between MFE and sequence length is roughly linear in the yeast (Figure 3.2). Thus, a simple way of controlling for S1 distance is to carry out a multiple regression with the following model:

$$\text{MFE} = b_0 + b_1 D_{S1} + b_2 \text{GE} \quad (3.1)$$

where GE is gene expression, measured by either CAI, log-transformed RNA abundance or log-transformed protein abundance. From Figure 3.2, we know that b_1 should be negative. If highly expressed yeast ICGs have more negative MFE than lowly expressed yeast ICGs, then b_2 should also be negative.

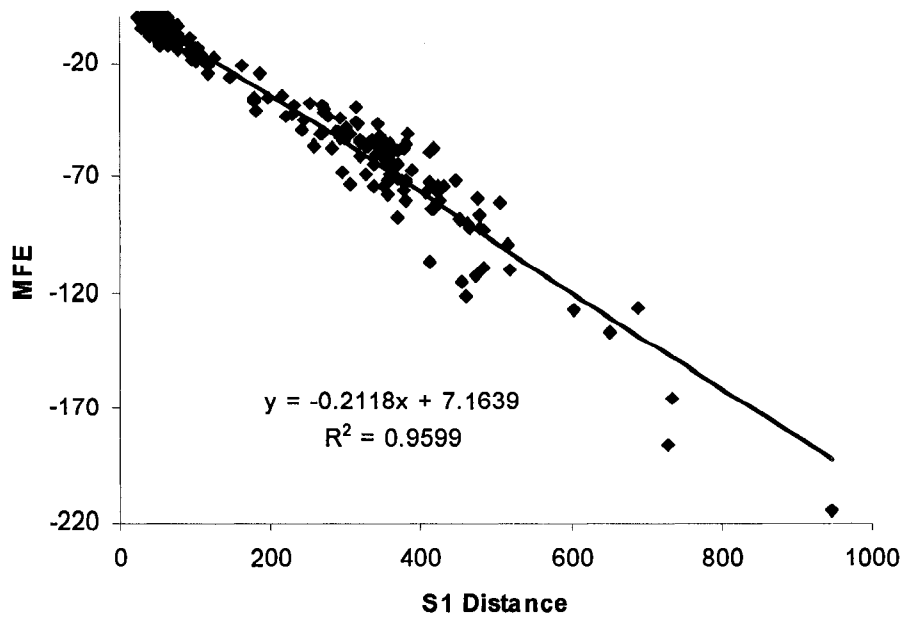


Figure 3.2 Relationship between minimum folding energy (MFE) and S1 distance in the yeast.

The model expressed in Eq. (3.1) fits the data very well when gene expression is measured by CAI (Table 3.5). Both b_1 and b_2 are statistically highly significantly different from zero (Table 3.5). However, while parameter b_1 is negative as expected, parameter b_2 is significantly positive which implies that highly expressed yeast ICGs have greater MFE (i.e., weaker secondary structure) than lowly expressed yeast ICGs. This is contrary to the expectation of a negative b_2 , i.e., the expectation that highly expressed genes should have more secondary structure.

Table 3.5 Fitting regression model in Eq. (3.1), with gene expression measured by CAI: (a) Results of analysis of variance, (b) parameter estimation and associated significance tests.

(a)	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Model	2	351039.7	175519.9	3558.962	<0.0001
Residual	254	12526.7	49.3		
Total	256	363566.4			

(b)	<i>Value</i>	<i>SE</i>	<i>t</i>	<i>p</i>
b_0	0.1880	1.2651	0.1486	0.88201
b_1	-0.2272	0.0035	-65.5122	4.1E-161
b_2	17.1522	2.6639	6.4387	5.98E-10

One may argue that the intercept should be zero in Eq. (3.1) because MFE would approach zero when sequence length approaches 0. However, almost identical results are obtained by forcing $b_0 = 0$, with the estimated $b_1 = -0.2273$ and $b_2 = 17.4912$.

The results are similar for the data set with gene expression measured by log-transformed RNA abundance, with both b_1 and b_2 being highly significantly different from zero, and with b_1 being negative and b_2 positive. For the data set with gene expression measured by log-transformed protein abundance data, b_1 is also negative and b_2 positive, but both are significantly different from zero only when b_0 is forced to be zero. When b_0 is allowed to differ from zero, b_2 is not significant, although still positive.

In short, the expectation that highly expressed yeast ICGs may be richer in secondary structure than lowly expressed yeast ICGs are not supported empirically. This casts doubt on the suggestion that secondary structure in the S1 sequence region would enhance splicing efficiency by bringing the donor splice site close BPS (Spingola et al., 1999). In retrospect, this suggestion does not have much theoretical basis. If efficient splicing in yeast requires the donor splice site to be close to BPS, then highly expressed yeast ICGs should have small S1 than lowly expressed yeast ICGs. However, highly expressed yeast ICGs have highly significantly greater S1 than lowly expressed ICGs (Table 3.4).

CHAPTER 4 INVESTIGATIONS ON EUKARYOTES WITH COMPLEX SPLICING MECHANISM

This study focuses on the relationship between gene expression and splicing efficiency in the yeast. Compared with other higher eukaryotes, the splicing mechanism and machinery in the yeast are considered to be relatively simple, and the donor and acceptor splice sites and BPS are sufficient for accurate and efficient splicing (Clark et al., 2002). Thus, in the yeast, it is very likely that splicing efficiency mainly depends on the donor and acceptor splice sites and BPS as well as the S1 and S2 distances.

However, the splicing mechanism and machinery for many higher eukaryotic ICGs seem to be more complicated than that for the yeast. Unlike the yeast, many higher eukaryotic ICGs feature alternative splicing (Ast, 2004). In human, introns in 35-65% of genes are expected to be spliced by alternative splicing (Ast, 2004). In addition, the canonic splicing signals located at the donor and acceptor splice sites and BPS are insufficient for accurate and efficient splicing in nematodes and flies (Dewey et al., 2006). In mammals and plants, only approximately one-half of the information required for correct splicing of short introns comes from these signals (Lim & Burge, 2001). Furthermore, recent studies have identified many other splicing signals located in exons or introns that are involved in splicing by either enhancing (enhancers) or suppressing (silencers) the accuracy and efficiency of splicing (Black, 2003; Zhang et al., 2005; Dewey et al., 2006). For example, exonic splicing enhancers (ESEs) that appear to exist in most of mammalian exons located in the exonic region close to the donor or acceptor splice site can enhance splicing accuracy and efficiency by promoting recognition of exon (Blencowe, 2000; Dewey et al., 2006). Moreover, unlike the yeast, the sequence consensuses of BPS is quite degenerate in many

higher eukaryotes such as human (Senapathy et al., 1990; Sun & Chasin, 2000; Zhang et al., 2005; Gao et al., 2008). Thus, for higher eukaryotes, we do not expect that the strengths of the donor and acceptor splice sites and BPS for highly expressed ICGs must be significantly higher than that for lowly expressed ICGs, although they might be.

To verify our prediction, we analyzed the relationship between gene expression and splicing efficiency for some other model organisms of higher eukaryotes such as human, *Caenorhabditis elegans* and *Arabidopsis thaliana*, which can be used to represent mammal, invertebrate and plant, respectively. Here we only studied the relationship between gene expression and the strengths of the donor and acceptor splice sites. We did not investigate the relationship between gene expression and BPS since there is no sufficient BPS data for these three organisms (Gao et al., 2008).

The analysis method is almost the same as we did in the yeast. The only modification is that we need to first separate ICGs into different groups based the number of introns contained in genes. Given that intron number is strongly negatively correlated ($r=-0.85$) with the donor splice site strength (Irimia et al., 2007), it is not appropriate to compare the splice site strength between the genes with different number of introns.

Before we performed the statistical analysis, we excluded some genes with a large number of introns because there are too few such genes for reliable statistical analysis. For instance, in human, there is only one gene with 310 introns. We set up 50 as the threshold of group size because it is considered to be the appropriate minimum sample size to perform the statistical tests. Finally, there were 22 ICG groups in human (every gene in the same group contains the same number of introns which is from 1 to 22), 14 ICG groups in *C.elegans*, and 15 ICG groups in *A. thaliana*. These ICG groups already contain 93%, 95% and 94% of total ICGs in human, *C.elegans* and *A.thaliana*, respectively. Thus, the vast

majority of ICGs in these three organisms were analyzed and the conclusions can be considered to apply for all ICGs. For each ICG group with the same number of introns, we then compared the donor and acceptor splice site strengths between highly and lowly expressed ICGs using the same analysis method as we did in the yeast.

Our results (not shown) indicate that in human, the acceptor splice site strength for highly express ICGs is significantly higher than that for lowly expressed genes in all groups whereas in many groups, the donor splice site strength for highly express ICGs is not significantly higher than that for lowly expressed genes. In *C.elegans*, with a few minor exceptions (ICG groups with intron numbers of 10, 11 and 12 account for only 6.9% ICGs in the dataset), for most groups containing the vast majority of ICGs (~92% ICGs in our dataset), highly expressed ICGs have significantly higher donor splice site strengths than those of lowly expressed ICGs. However, for the acceptor splice site, there is no significant difference in splicing strength between highly and lowly expressed ICGs in many groups. In *A. thaliana*, for either donor or acceptor splice site, there is no significant difference in the splicing strength between highly and lowly expressed ICGs for most of groups.

From the results shown above, we realize that the splicing mechanism and machinery are different not only between yeast and higher eukaryotes, but also among various higher eukaryotes. An extensive comparative study should shed light on the evolution and maintenance of splicing mechanisms in different species.

REFERENCE

- Ast G. 2004. How did alternative splicing evolve? *Nat Rev Genet* 5:773-782.
- Bachellerie JP, Michot B, Nicoloso M, Balakin A, Ni J, Fournier MJ. 1995. Antisense snoRNAs: a family of nucleolar RNAs with long complementarities to rRNA. *Trends Biochem Sci* 20:261-264.
- Bhargava J, Engebrecht J, Roeder GS. 1992. The rec102 mutant of yeast is defective in meiotic recombination and chromosome synapsis. *Genetics* 130:59-69.
- Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72:291-336.
- Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 25:106-110.
- Bon E, Casaregola S, Blandin G, Llorente B, Neuveglise C, Munsterkotter M, Guldener U, Mewes HW, Van Helden J, Dujon B, Gaillardin C. 2003. Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res* 31:1121-1135.
- Braddock M, Muckenthaler M, White MR, Thorburn AM, Sommerville J, Kingsman AJ, Kingsman SM. 1994. Intron-less RNA injected into the nucleus of *Xenopus* oocytes accesses a regulated translation control pathway. *Nucleic Acids Res* 22:5255-5264.
- Brys A, Schwer B. 1996. Requirement for SLU7 in yeast pre-mRNA splicing is dictated by the distance between the branchpoint and the 3' splice site. *Rna* 2:707-717.
- Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell* 2:773-785.
- Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet* 31:415-418.
- Cellini A, Felder E, Rossi JJ. 1986. Yeast pre-messenger RNA splicing efficiency depends on critical spacing requirements between the branch point and 3' splice site. *Embo J* 5:1023-1030.
- Chapman RE, Walter P. 1997. Translational attenuation mediated by an mRNA intron. *Curr Biol* 7:850-859.
- Charpentier B, Rosbash M. 1996. Intramolecular structure in yeast introns aids the early steps of in vitro spliceosome assembly. *Rna* 2:509-522.

- Chen S, Anderson K, Moore MJ. 2000. Evidence for a linear search in bimolecular 3' splice site AG selection. *Proc Natl Acad Sci U S A* 97:593-598.
- Chiara MD, Gozani O, Bennett M, Champion-Arnaud P, Palandjian L, Reed R. 1996. Identification of proteins that interact with exon sequences, splice sites, and the branchpoint sequence during each stage of spliceosome assembly. *Mol Cell Biol* 16:3317-3326.
- Clark TA, Sugnet CW, Ares M, Jr. 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* 296:907-910.
- Clote P, Ferre F, Kranakis E, Krizanc D. 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *Rna* 11:578-591.
- Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* 16:1131-1145.
- Coluccio A, Bogengruber E, Conrad MN, Dresser ME, Briza P, Neiman AM. 2004. Morphogenetic pathway of spore wall assembly in *Saccharomyces cerevisiae*. *Eukaryot Cell* 3:1464-1475.
- Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* 167:1293-1304.
- Davis CA, Grate L, Spingola M, Ares M, Jr. 2000. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res* 28:1700-1706.
- de Roos AD. 2005. Origins of introns based on the definition of exon modules and their conserved interfaces. *Bioinformatics* 21:2-9.
- Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 7:311.
- Dietrich RC, Fuller JD, Padgett RA. 2005. A mutational analysis of U12-dependent splice site dinucleotides. *RNA* 11:1430-1440.
- Du H, Rosbash M. 2002. The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. *Nature* 419:86-90.
- Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc Natl Acad Sci U S A* 96:4482-4487.

- Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends Genet* 19:362-365.
- Fahey ME, Higgins DG. 2007. Gene expression, intron density, and splice site strength in *Drosophila* and *Caenorhabditis*. *J Mol Evol* 65:349-357.
- Fong YW, Zhou Q. 2001. Stimulatory effect of splicing factors on transcriptional elongation. *Nature* 414:929-933.
- Freund M, Hicks MJ, Konermann C, Otte M, Hertel KJ, Schaal H. 2005. Extended base pair complementarity between U1 snRNA and the 5' splice site does not inhibit splicing in higher eukaryotes, but rather increases 5' splice site recognition. *Nucleic Acids Res* 33:5112-5119.
- Furger A, O'Sullivan JM, Binnie A, Lee BA, Proudfoot NJ. 2002. Promoter proximal splice sites enhance transcription. *Genes Dev* 16:2792-2799.
- Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res* 36:2257-2267.
- Ghaemmighami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. *Nature* 425:737-741.
- Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418:387-391.
- Gonzalez TN, Sidrauski C, Dorfler S, Walter P. 1999. Mechanism of non-spliceosomal mRNA splicing in the unfolded protein response pathway. *Embo J* 18:3119-3132.
- Guldener U, Munsterkotter M, Kastenmuller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE, Michael H, Kaps A, Talla E, Dujon B, Andre B, Souciet JL, De Montigny J, Bon E, Gaillardin C, Mewes HW. 2005. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res* 33:D364-368.

- Hall SL, Padgett RA. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol* 239:357-365.
- Hall SL, Padgett RA. 1996. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science* 271:1716-1718.
- Harris NL, Senapathy P. 1990. Distribution and consensus of branch point signals in eukaryotic genes: a computerized statistical analysis. *Nucleic Acids Res* 18:3015-3019.
- Hastings ML, Krainer AR. 2001. Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol* 13:302-309.
- Hertz GZ, Stormo GD. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563-577.
- Hofacker IL. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res* 31:3429-3431.
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie* 125:167-188.
- Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998a. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717-728.
- Irimia M, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. *Trends Genet* 23:321-325.
- Isken O, Maquat LE. 2007. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev* 21:1833-1856.
- Jackson IJ. 1991. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res* 19:3795-3798.
- Jacquier A, Rosbash M. 1986. RNA splicing and intron turnover are greatly diminished by a mutant yeast branch point. *Proc Natl Acad Sci U S A* 83:5835-5839.
- Jiao K, Salem L, Malone R. 2003. Support for a meiotic recombination initiation complex: interactions among Rec102p, Rec104p, and Spo11p. *Mol Cell Biol* 23:5928-5938.
- Juan V, Wilson C. 1999. RNA secondary structure prediction based on free energy and phylogenetic analysis. *J Mol Biol* 289:935-947.
- Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW. 2006. Introns regulate RNA and protein abundance in yeast. *Genetics* 174:511-518.

- Juneau K, Palm C, Miranda M, Davis RW. 2007. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proc Natl Acad Sci U S A* 104:1522-1527.
- Jurica MS, Moore MJ. 2003. Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell* 12:5-14.
- Kaufman RJ. 1999. Stress signaling from the lumen of the endoplasmic reticulum: coordination of gene transcriptional and translational controls. *Genes Dev* 13:1211-1233.
- Kawahara T, Yanagi H, Yura T, Mori K. 1997. Endoplasmic reticulum stress-induced mRNA splicing permits synthesis of transcription factor Hac1p/Ern4p that activates the unfolded protein response. *Mol Biol Cell* 8:1845-1862.
- Kent OA, Ritchie DB, Macmillan AM. 2005. Characterization of a U2AF-independent commitment complex (E') in the mammalian spliceosome assembly pathway. *Mol Cell Biol* 25:233-240.
- Kohrer K, Domdey H. 1988. Splicing and spliceosome formation of the yeast MATa1 transcript require a minimum distance from the 5' splice site to the internal branch acceptor site. *Nucleic Acids Res* 16:9457-9475.
- Kozak M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44:283-292.
- Kozak M. 1997. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *Embo J* 16:2482-2492.
- Kunze B, Hellwig-Burgel T, Weichenhan D, Traut W. 2000. Transcription and proper splicing of a mammalian gene in yeast. *Gene* 246:93-102.
- Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW. 2004. Introns and splicing elements of five diverse fungi. *Eukaryot Cell* 3:1088-1100.
- Kwek KY, Murphy S, Furger A, Thomas B, O'Gorman W, Kimura H, Proudfoot NJ, Akoulitchiev A. 2002. U1 snRNA associates with TFIIF and regulates transcriptional initiation. *Nat Struct Biol* 9:800-805.
- Lamond AI. 1993. A glimpse into the spliceosome. *Curr Biol* 3:62-64.
- Le Hir H, Gatfield D, Izaurralde E, Moore MJ. 2001. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *Embo J* 20:4987-4997.

- Le Hir H, Izaurralde E, Maquat LE, Moore MJ. 2000. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *Embo J* 19:6860-6869.
- Le Hir H, Nott A, Moore MJ. 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci* 28:215-220.
- Leu JY, Chua PR, Roeder GS. 1998. The meiosis-specific Hop2 protein of *S. cerevisiae* ensures synapsis between homologous chromosomes. *Cell* 94:375-386.
- Li B, Vilardell J, Warner JR. 1996. An RNA structure involved in feedback regulation of splicing and of translation is critical for biological fitness. *Proc Natl Acad Sci U S A* 93:1596-1600.
- Li SW, Feng L, Niu DK. 2007. Selection for the miniaturization of highly expressed genes. *Biochem Biophys Res Commun* 360:586-592.
- Libri D, Lescure A, Rosbash M. 2000. Splicing enhancement in the yeast rp51b intron. *Rna* 6:352-368.
- Libri D, Stutz F, McCarthy T, Rosbash M. 1995. RNA structural patterns and splicing: molecular basis for an RNA-based enhancer. *Rna* 1:425-436.
- Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* 98:11193-11198.
- Lund M, Kjems J. 2002. Defining a 5' splice site by functional selection in the presence and absence of U1 snRNA 5' end. *RNA* 8:166-179.
- Luo MJ, Reed R. 1999. Splicing is required for rapid and efficient mRNA export in metazoans. *Proc Natl Acad Sci U S A* 96:14937-14942.
- Luukkonen BG, Seraphin B. 1997. The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in *Saccharomyces cerevisiae*. *Embo J* 16:779-792.
- Madhani HD, Guthrie C. 1994. Dynamic RNA-RNA interactions in the spliceosome. *Annu Rev Genet* 28:1-26.
- Malone RE, Bullard S, Hermiston M, Rieger R, Cool M, Galbraith A. 1991. Isolation of mutants defective in early steps of meiotic recombination in the yeast *Saccharomyces cerevisiae*. *Genetics* 128:79-88.
- Maquat LE, Carmichael GG. 2001. Quality control of mRNA function. *Cell* 104:173-176.

- Matsumoto K, Wassarman KM, Wolffe AP. 1998. Nuclear history of a pre-mRNA determines the translational activity of cytoplasmic mRNA. *Embo J* 17:2107-2121.
- Mazina OM, Mazin AV, Nakagawa T, Kolodner RD, Kowalczykowski SC. 2004. *Saccharomyces cerevisiae* Mer3 helicase stimulates 3'-5' heteroduplex extension by Rad51; implications for crossover control in meiotic recombination. *Cell* 117:47-56.
- McPheeters DS, Muhlenkamp P. 2003. Spatial organization of protein-RNA interactions in the branch site-3' splice site region during pre-mRNA splicing in yeast. *Mol Cell Biol* 23:4174-4186.
- Moore MJ, Schwartzfarb EM, Silver PA, Yu MC. 2006. Differential Recruitment of the Splicing Machinery during Transcription Predicts Genome-Wide Patterns of mRNA Splicing. *Mol Cell* 24:903.
- Nakagawa T, Kolodner RD. 2002a. The MER3 DNA helicase catalyzes the unwinding of holliday junctions. *J Biol Chem* 277:28019-28024.
- Nakagawa T, Kolodner RD. 2002b. *Saccharomyces cerevisiae* Mer3 is a DNA helicase involved in meiotic crossing over. *Mol Cell Biol* 22:3281-3291.
- Nakagawa T, Ogawa H. 1997. Involvement of the MRE2 gene of yeast in formation of meiosis-specific double-strand breaks and crossover recombination through RNA splicing. *Genes Cells* 2:65-79.
- Nakagawa T, Ogawa H. 1999. The *Saccharomyces cerevisiae* MER3 gene, encoding a novel helicase-like protein, is required for crossover control in meiosis. *Embo J* 18:5714-5723.
- Nilsen TW. 1994. RNA-RNA interactions in the spliceosome: unraveling the ties that bind. *Cell* 78:1-4.
- Nilsen TW. 2005. Spliceosome assembly in yeast: one ChIP at a time? *Nat Struct Mol Biol* 12:571-573.
- Niu DK. 2007. Protecting exons from deleterious R-loops: a potential advantage of having introns. *Biol Direct* 2:11.
- Orphanides G, Reinberg D. 2002. A unified theory of gene expression. *Cell* 108:439-451.
- Padgett RA, Grabowski PJ, Konarska MM, Seiler S, Sharp PA. 1986. Splicing of messenger RNA precursors. *Annu Rev Biochem* 55:1119-1150.
- Parenteau J, Durand M, Veronneau S, Lacombe AA, Morin G, Guerin V, Cecez B, Gervais-Bird J, Koh CS, Brunelle D, Wellinger RJ, Chabot B, Abou Elela S. 2008. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol Biol Cell* 19:1932-1941.

- Parker R, Patterson B. 1987. Architecture of fungal introns: implications for spliceosome assembly. In: Inouye M, Dudock B, eds. *Molecular Biology of RNA, New Perspectives*. New York, NY.: Academic Press. pp 133–149.
- Parker R, Siliciano PG, Guthrie C. 1987. Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell* 49:229-239.
- Patterson B, Guthrie C. 1991. A U-rich tract enhances usage of an alternative 3' splice site in yeast. *Cell* 64:181-187.
- Reed R. 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr Opin Genet Dev* 6:215-220.
- Reed R, Hurt E. 2002. A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell* 108:523-531.
- Reed R, Maniatis T. 1988. The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes Dev* 2:1268-1276.
- Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16:276-277.
- Ringner M, Krogh M. 2005. Folding free energies of 5'-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput Biol* 1:e72.
- Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV. 2005. Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform* 6:118-134.
- Romfo CM, Wise JA. 1997. Both the polypyrimidine tract and the 3' splice site function prior to the first step of splicing in fission yeast. *Nucleic Acids Res* 25:4658-4665.
- Scherrer FW, Jr., Spingola M. 2006. A subset of Mer1p-dependent introns requires Bud13p for splicing activation and nuclear retention. *RNA* 12:1361-1372.
- Seffens W, Digby D. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res* 27:1578-1584.
- Senapathy P, Shapiro MB, Harris NL. 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol* 183:252-278.
- Sharp PA. 1994. Split genes and RNA splicing. *Cell* 77:805-815.
- Sharp PA. 2005. The discovery of split genes and RNA splicing. *Trends Biochem Sci* 30:279-281.

- Sharp PA, Burge CB. 1997. Classification of introns: U2-type or U12-type. *Cell* 91:875-879.
- Sharp PM, Li WH. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281-1295.
- Sidrauski C, Walter P. 1997. The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response. *Cell* 90:1031-1039.
- Simpson CG, Clark G, Davidson D, Smith P, Brown JW. 1996. Mutation of putative branchpoint consensus sequences in plant introns reduces splicing efficiency. *Plant J* 9:369-380.
- Simpson CG, Thow G, Clark GP, Jennings SN, Watters JA, Brown JW. 2002. Mutational analysis of a plant branchpoint and polypyrimidine tract required for constitutive splicing of a mini-exon. *RNA* 8:47-56.
- Spingola M, Ares M, Jr. 2000. A yeast intronic splicing enhancer and Nam8p are required for Mer1p-activated splicing. *Mol Cell* 6:329-338.
- Spingola M, Armisen J, Ares M, Jr. 2004. Mer1p is a modular splicing factor whose function depends on the conserved U2 snRNP protein Snu17p. *Nucleic Acids Res* 32:1242-1250.
- Spingola M, Grate L, Haussler D, Ares M, Jr. 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *Rna* 5:221-234.
- Sridharan V, Singh R. 2007. A conditional role of U2AF in splicing of introns with unconventional polypyrimidine tracts. *Mol Cell Biol* 27:7334-7344.
- Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW. 2002. Systematic screen for human disease genes in yeast. *Nat Genet* 31:400-404. Epub 2002 Jul 22.
- Sun H, Chasin LA. 2000. Multiple splicing defects in an intronic false exon. *Mol Cell Biol* 20:6414-6425.
- Thompson-Jager S, Domdey H. 1987. Yeast pre-mRNA splicing requires a minimum distance between the 5' splice site and the internal branch acceptor site. *Mol Cell Biol* 7:4010-4016.
- Tsubouchi H, Roeder GS. 2002. The Mnd1 protein forms a complex with hop2 to promote homologous chromosome pairing and meiotic double-strand break repair. *Mol Cell Biol* 22:3078-3088.

- Villa T, Ceradini F, Presutti C, Bozzoni I. 1998. Processing of the intron-encoded U18 small nucleolar RNA in the yeast *Saccharomyces cerevisiae* relies on both exo- and endonucleolytic activities. *Mol Cell Biol* 18:3376-3383.
- Wang GS, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 8:749-761.
- Weaver RF. 2005. *Molecular Biology*. Boston: McGraw-Hill Higher Education.
- Wilusz CJ, Wang W, Peltz SW. 2001. Curbing the nonsense: the activation and regulation of mRNA surveillance. *Genes Dev* 15:2781-2785.
- Xia X. 1996. Maximizing transcription efficiency causes codon usage bias. *Genetics* 144:1309-1320.
- Xia X. 2001. *Data analysis in molecular biology and evolution*. Boston: Kluwer Academic Publishers.
- Xia X. 2007a. The +4G site in Kozak consensus is not related to the efficiency of translation initiation. *PLoS ONE* 2:e188.
- Xia X. 2007b. *Bioinformatics and the cell: Modern computational approaches in genomics, proteomics and transcriptomics*. New York: Springer US.
- Xia X. 2007c. An Improved Implementation of Codon Adaptation Index. *Evolutionary Bioinformatics* 3:53–58.
- Xia X, Xie Z. 2001. DAMBE: Software package for data analysis in molecular biology and evolution. *J Hered* 92:371-373.
- Zar JH. 1999. *Biostatistical analysis*. Upper Saddle River, New Jersey: Prentice Hall.
- Zhang XH, Leslie CS, Chasin LA. 2005. Dichotomous splicing signals in exon flanks. *Genome Res* 15:768-779.
- Zheng CL, Fu XD, Gribskov M. 2005. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* 11:1777-1787.