uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Xiaoquan Yao

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Biology)

GRADE / DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Sequence Features Affecting Translation Initiation in Eukaryotes: a bioinformatic approach

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. Xuhua 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. Myron Smith

Dr. Stéphane Aris-Brosou

Dr. Guy Drouin

Gary W. Slater

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

Sequence Features Affecting Translation Initiation in Eukaryotes: a bioinformatic approach

Xiaoquan Yao

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 la
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

Your file Votre référence
ISBN: 978-0-494-46506-6
Our file Notre référence
ISBN: 978-0-494-46506-6

Abstract

Sequence features play an important role in the regulation of translation initiation. This thesis focuses on the sequence features affecting eukaryotic initiation. The characteristics of 5' untranslated region in *Saccharomyces cerevisiae* were explored. It is found that the 40 nucleotides upstream of the start codon is the critical region for translation initiation in yeast. Moreover, this thesis attempted to solve some controversies related to the start codon context. Two key nucleotides in the start codon context are the third nucleotide upstream of the start codon (-3 site) and the nucleotide immediately following the start codon (+4 site). Two hypotheses regarding +4G (G at +4 site) in Kozak consensus, the translation initiation hypothesis and the amino acid constraint hypothesis, were tested. The relationship between the -3 and +4 sites in seven eukaryotic species does not support the translation initiation hypothesis. The amino acid usage at the position after the initiator (P1' position) compared to other positions in the coding sequences of seven eukaryotic species was examined. The result is consistent with the amino acid constraint hypothesis. In addition, this thesis explored the relationship between +4 nucleotide and translation efficiency in yeast. The result shows that +4 nucleotide is not important for translation efficiency, which does not support the translation initiation hypothesis. This work improves our current understanding of eukaryotic translation initiation process.

Résumé

Les caractéristiques des séquences jouent un rôle important dans la régulation de l'initiation de la traduction. Cette thèse se concentre sur les caractéristiques des séquences qui affectent ce processus chez les eucaryotes. Les caractéristiques des terminaisons non traduites en 5' (5'-UTR) chez *Saccharomyces cerevisiae* ont été explorées. Il est trouvé que les 40 nucléotides en amont du codon de démarrage de la traduction représentent la région critique pour l'initiation chez la levure. De plus, cette thèse a tenté de résoudre quelques controverses reliées au contexte nucléotide autour du codon de démarrage. Les deux nucléotides clés sont le troisième nucléotide en amont du codon de démarrage (le site -3) et le nucléotide suivant immédiatement le codon de démarrage (le site +4). Les deux hypothèses qui concernent le +4G de la séquence de Kozak, l'hypothèse d'initiation de traduction et l'hypothèse de contrainte d'acide aminé, ont été testées. La relation entre les sites -3 et +4 n'est pas en accord avec l'hypothèse d'initiation de traduction. L'usage d'acide aminé à la position après l'initiateur (la position P1') relatif aux autres positions dans les séquences codantes chez sept espèces d'eucaryotes a été examiné. Le résultat est in accord avec à l'hypothèse de contrainte d'acide aminé. En outre, cette thèse a exploré la relation entre le site +4 et l'efficacité de traduction chez la levure. Le résultat montre que le site +4 n'est pas important pour l'efficacité de traduction et ne soutient pas l'hypothèse d'initiation de traduction. Ainsi, cette thèse approfondit notre compréhension actuelle du processus d'initiation de la traduction chez les eucaryotes étudiés ici.

Acknowledgements

This thesis is a demonstration of my work as a bioinformatician in the recent two years. Two years ago, "bioinformatics" was yet a brand new term for me. The transformation has been done with many people's help.

First of all, I thank my supervisor, Dr. Xuhua Xia, for guiding me into this exciting field. His insights and visions have impressed me all the time. I am grateful to my advisory committee members, Dr. Linda Bonen and Dr. Myron Smith, for providing me with valuable feedback throughout my thesis project. I also want to thank Dr. Stéphane Aris-Brosou for his constructive suggestions and kind help.

I am thankful to the members of Dr. Xia's and Dr. Aris-Brosou's labs for interesting discussions in the lab meetings. Some work of this thesis was finished in a collaborative manner.

I am grateful to the University of Ottawa and the Department of Biology for providing me with financial assistance, and the granting agencies for their contribution to my research.

A number of friends helped me get through the difficult time in my life during these two years' study. I will never forget their friendship.

I would not be able to finish this thesis without the support from my patient husband and my two adorable kids. Their love is the determinant power. This thesis is also a gift to my parents, especially to my father who passed away sixteen years ago. Dad, now I really know, "learning is fun".

Table of Contents

List of Figures

List of Tables

List of Abbreviations

5'-UTR (3'-UTR)	5' untranslated region (3' untranslated region)
AAAI	amino acid adaptation index
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
A. thaliana	*Arabidopsis thaliana*
ATP	adenosine triphosphate
CAI	codon adaptation index
CDS	coding sequence
C-terminus(terminal)	carboxyl terminus(terminal)
DAMBE	Data Analysis in Molecular Biology and Evolution
DNA	deoxyribonucleic acid
EF	elongation factor
eIF	eukaryotic initiation factor
EMBOSS	European Molecular Biology Open Software Suite
eRF	eukaryotic release factor
GTP	guanosine triphosphate
IRES	internal ribosome entry site
LOWESS	locally weighted scatterplot smoothing
MAP	methionine aminopeptidase
MFE	minimum folding energy
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NME	N-terminal methionine excision
nt	nucleotides
N-terminus(terminal)	amino terminus(terminal)

ORF	open reading frame
O. sativa	*Oryza sativa*
R^2	determination coefficient
RF	release factor
RNA	ribonucleic acid
rRNA	ribosome RNA
SAGE	Serial Analysis of Gene Expression
S. bayanus	*Saccharomyces bayanus*
S. cerevisiae	*Saccharomyces cerevisiae*
SD sequence	Shine-Dalgarno sequence
SGD	*Saccharomyces* Genome Database
S. mikatae	*Saccharomyces mikatae*
S. paradoxus	*Saccharomyces paradoxus*
S. pombe	*Schizosaccharomyces pombe*
TAP	tandem affinity purification
tRNA	transfer RNA
uAUG	upstream AUG
uORF	upstream open reading frame

IUPAC Code Table

(The International Union of Pure and Applied Chemistry)

Amino Acids

One Letter Code	Three Letter Code	Name
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartate
E	Glu	Glutamate
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

Nucleotides

Code	Nucleotide	Code	Nucleotide
A	Adenine	R	G A (Purine)
C	Cytosine	Y	T C (Pyrimidine)
G	Guanine	N	A G C T (any)
T	Thymine		
U	Uracil		

Chapter 1 Introduction

1.1 Research context

Transcription and translation are two fundamental biological processes in a cell, during which genetic information is expressed from gene to protein. After messenger RNA (mRNA) is synthesized and ribosome binds to it, the process of translation begins. Just as transcription, translation is composed of three steps, initiation, elongation, and termination. Translation initiation involves interactions between ribosome and mRNA with the participation of translation related factors, initiating transfer RNA (tRNA), and energy produced from guanosine triphosphate (GTP) and adenosine triphosphate (ATP) hydrolyses. Recognition of the start codon is an important step in translation initiation regulation. Prokaryotes and eukaryotes have very similar initiation procedures, but they differ in the composition of ribosome, the way of ribosomal binding, the initiation factors, the initiating tRNA, and the mRNA sequence features.

1.2 Research problems

This thesis focuses on translation initiation in eukaryotes and investigates the mRNA sequence features affecting translation initiation. It has been recognized that secondary structure and length of 5'untranslated region (5'-UTR), secondary structure of immediate 3' side of start codon, start codon identity and start codon context are important factors in translation initiation regulation (Cigan et al., 1988; Kozak, 1999; Kozak, 2002; Kozak, 2005; Slusher et al., 1991). This thesis will explore the

characteristics of 5'-UTR and the start codon context, evaluate recent controversies related to the start codon context, and discuss relevant mechanisms that may explain the observed features. All of these will improve our current understanding of the translation initiation process in eukaryotes.

1.3 Outline of this thesis

Chapter 2 gives a literature review, which provides background information for translation initiation in eukaryotes, presents current researches and controversies in the mRNA sequence features related to eukaryotic initiation, and introduces *Saccharomyces cerevisiae* (*S. cerevisiae*) as a model organism for studying translation initiation which paves the way for studies in Chapter 3 and Chapter 5.

Chapter 3 explores the sequence features of the 5'-UTR related to translation initiation in *S. cerevisiae.* Chapter 4 presents a comparative analysis of start codon context in seven eukaryotic species. Chapter 5 explores the relationship between the nucleotide immediately following the start codon and translation efficiency in *S. cerevisiae*. Chapter 6 is a conclusion for the whole thesis.

1.4 The use of "T" and "U" and significance level in this thesis

In this thesis, "T" and "U" will be exchangeably used in the nucleotide sequences. When "T" is used, it means the nucleotide in deoxyribonucleic acid (DNA); if "U" is used, it implies the nucleotide in ribonucleic acid (RNA).

The significance level is at $\alpha = 0.05$ in all significance tests throughout this thesis.

Chapter 2 Literature Review

2.1 Translation and the relevant mRNA sequence features affecting translation

Similar to transcription, the complex process of translation can be divided into three steps, initiation, elongation, and termination. The following sections will outline these three steps in a eukaryotic cell. Moreover, they will discuss the relevant mRNA sequence features related to the three stages of translation in eukaryotes, but emphasis will be given to initiation. Some differences between eukaryotes and prokaryotes will be provided to help capture the characteristics of eukaryotic initiation. The reason for including elongation and termination in this context is that the sequence features influencing these two stages may become confounding factors in the study of the sequence features related to initiation.

2.1.1 Initiation

2.1.1.1 Initiation process in eukaryotes

In eukaryotes, large (60S) and small (40S) ribosomal subunits are separate before starting initiation. A translation preinitiation complex is formed by eIF1A, 40S subunit-eIF3 complex, and a ternary complex of $tRNA_i^{Met}$, eIF2, and GTP (Alberts et al., 2002; Lodish et al., 2004; Pisarev et al., 2006). eIF stands for eukaryotic initiation factor.

A 7-methyl guanosine cap is attached to the 5' end of an mRNA during mRNA modification (Kozak, 1978a; Kozak, 1978b; Kozak, 1999). At the beginning of

translation initiation, eIF4 binds to the 5' cap of an mRNA. Then an initiation complex is formed by the interaction of mRNA-eIF4 complex with the preinitiation complex. The initiation complex linearly scans the mRNA while ATP hydrolysis provides energy for the helicase activity of eIF4 to melt the mRNA secondary structure (Alberts et al., 2002; Lodish et al., 2004). Scanning stops when the $tRNA_i^{Met}$ anticodon recognizes the start codon, which usually is the first AUG downstream from the 5' end in most eukaryotic mRNAs (Kozak, 1978a; Kozak, 1980; Kozak, 2002). Recognition of the start codon leads to hydrolysis of GTP associated with eIF2. Once the small ribosomal subunit with its bound $tRNA_i^{Met}$ is correctly positioned at the start codon, the small ribosomal subunit combines with the large ribosomal subunit, forming an 80S ribosome. This requires eIF5 and hydrolysis of a GTP associated with it (Alberts et al., 2002; Lodish et al., 2004).

A different initiation mechanism called internal ribosome entry site (IRES) (Jackson and Kaminski, 1995) has been proposed, by which the 40S ribosome is believed to internally bind to the mRNA. Translation of some viral mRNAs, which lack a 5' cap, is thought to be initiated at IRESs in the infected eukaryotic cells (Lodish et al., 2004). How an IRES is recognized is an ongoing research topic (Kozak, 2002; Lodish et al., 2004).

2.1.1.2 Sequence features affecting translation initiation

The cap-dependant scanning model is the dominant hypothesis for translation initiation in eukaryotes as described in the last section 2.1.1.1. AUG is the almost unique start codon used in eukaryotic translation initiation, although there are some exceptions, in which the initiation efficiency greatly decreases (Chen et al., 2007;

Kozak, 1999). The scanning process involves finding the correct start codon. During this process, the secondary structure of initiation region, the length of 5'-UTR, and the start codon context are the major features in mRNA which have been recognized to affect translation initiation (Cigan et al., 1988; Kozak, 1999; Kozak, 2002; Kozak, 2005; Slusher et al., 1991). In addition, the presence of upstream AUG (uAUG) and upstream open reading frame (uORF) may strongly affect the efficiency of translation initiation (Kozak, 1999). uAUG is the AUG in the 5'-UTR and uORF refers to the open reading frame in the 5'-UTR of another open reading frame. uORF is a special case of uAUG.

The secondary structure of the initiation region can influence the recognition of the start codon AUG. But the secondary structures of two different segments in the initiation region have different effects. These two segments are the 5'-UTR and the immediate 3' region of the start codon. The 5'-UTR is also known as the leader sequence. The secondary structure of this region would prevent the 40S ribosome subunit from advancing, thus decreasing initiation efficiency (Kozak, 2005). Introducing GC-rich sequences prone to forming hairpins in the leader sequence of the yeast *HIS4* gene substantially reduced translation (Cigan et al., 1988). In contrast, the immediate 3' region of the start codon is the downstream part of the initiation region, the secondary structure of which would cause the 40S ribosome to stay longer at the start codon position and facilitate the recognition of the start codon (Kozak, 1990; Kozak, 2002).

The effect of the length of 5'-UTR is not known. One experiment with yeast *HIS4* gene showed that altering the 5'-UTR length from 115nt (nucleotides) to 39nt had

no effect on gene expression (Cigan et al., 1988), but the other one with yeast *MOD5* gene revealed that adding a sequence segment of at least 36nt long into the 5'-UTR greatly influenced the alternative AUG selection (Slusher et al., 1991). Review papers from Kozak (1999; 2005) state that the length of a 5'-UTR sequence generally does not affect initiation, but a long 5'-UTR sequence having considerable secondary structure can reduce initiation efficiency. Cigan and Donahue (1987) proposed that ribosomes may require a minimum 5'-UTR length for efficient recognition of an AUG.

The start codon context is another important feature in mRNA to regulate translation initiation, which is highly related to the existence of uAUGs and uORFs. In the literature dealing with the start codon context, position numbers are used to indicate the mRNA nucleotides. The nucleotides in the start codon take the numbers +1, +2, and +3, so the position number for the third nucleotide upstream of the start codon is -3 and the position number for the nucleotide immediately following the start codon is +4. When codon and amino acid are discussed, this thesis will denote the codon immediately following the start codon as P1' codon and the amino acid immediately following the initiator Met as P1' amino acid. This P1' annotation comes from the literature on proteins (Frottin et al., 2006), in which the amino acid immediately after the initiator Met is labelled as P1' amino acid and the follwing one as P2', etc. In many other studies, P1' amino acid is called as penultimate amino acid (Flinta et al., 1986; Hirel et al., 1989). The real meaning of "penultimate" is "next to the last", so this thesis will use P1' to indicate the second codon and amino acid to prevent confusion.

The identity and span of the start codon context is an active research topic. Kozak (1981) first proposed that selection of the initiating AUG was facilitated by the

specific surrounding nucleotides and she identified RCCaugG ("aug" in "RCCaugG" is the start codon) as the optimal context for vertebrates and -3R and +4G as the most important nucleotides in the context (Kozak, 1981; Kozak, 1984a; Kozak, 1984b; Kozak, 1986; Kozak, 1987; Kozak, 1997). Therefore, RCCaugG is called Kozak consensus sequence. Since Kozak identified RCCaugG, research has extended to a variety of other eukaryotic species. It has been found that the preferred nucleotide sequences around the start codon are very diverse across different eukaryotic taxonomic groups, but strong biases of -3R and +5C are commonly observed in many eukaryotic species (Cavener and Ray, 1991; Kochetov, 2005; Nakagawa et al., 2008; Niimura et al., 2003; Pesole et al., 2000) including animals (Cavener, 1987), plants (Joshi et al., 1997; Lukaszewicz et al., 2000; Sawant et al., 2001), fungi (Cigan and Donahue, 1987; Hamilton et al., 1987), and protists (Yamauchi, 1991). The role of +4G in Kozak consensus is controversial and will be discussed in the section 2.2.2.

uAUG and uORF are related to initiation regulation. Some uAUGs can allow two proteins to be produced from one mRNA (Kozak, 1991), and the presence of uORF can decrease the protein production from the downstream open reading frame (ORF) (Kozak, 1999). The question of whether the uAUG is used as the start codon is related to the requirement of a minimum length of 5'-UTR and start codon context (Cigan and Donahue, 1987; Kozak, 1999; Kozak, 2002; Kozak, 2005). If there is a requirement of minimum length of 5'-UTR and the 5'-UTR of an uAUG has a shorter length, this uAUG is unlikely to be used as the start codon; if an uAUG resides in a very unfavorable context, this uAUG is again unlikely to be used as the start codon.

2.1.1.3 *Differences between prokaryotic and eukaryotic initiation regarding mRNA sequence features*

In both prokaryotes and eukaryotes, the key for initiation is the selection of the start codon. There are some differences between prokaryotic and eukaryotic initiations regarding the mRNA sequence features. First, prokaryotic ribosomes can enter and initiate at multiple sites within an mRNA, but eukaryotic ribosomes routinely enter only at the 5' end except IRES (Lodish et al., 2004). Therefore, a eukaryotic mRNA has a 5' cap to facilitate the binding of the preinitiation complex. Second, in addition to AUG, the proportion of using other start codons in prokaryotes is much higher than in eukaryotes (Kozak, 2005). Third, in prokaryotes, an important element in the initiation region is the Shine-Dalgarno (SD) sequence which base pairs with 16S ribosome RNA (rRNA) in small (30S) ribosome subunit to help ribosome binding (Shine and Dalgarno, 1974; Shine and Dalgarno, 1975a; Shine and Dalgarno, 1975b). Eukaryotic mRNAs do not seem to contain elements analogous to the SD sequence although there is more and more evidence showing that mRNA-rRNA base-pairing may affect the translation of some eukaryotic mRNAs, which may explain the activity of some particular IRESes (Dresios et al., 2006).

2.1.2 Elongation

2.1.2.1 *Elongation process in eukaryotes*

The correctly positioned eukaryotic 80S ribosome-$tRNA_i^{Met}$ complex carries out elongation according to the coding sequence of the mRNA. Elongation factors (EFs) are involved in this process. The elongation consists of entry of each succeeding aminoacyl-tRNA, formation of a peptide bond, and translocation of the ribosome one

codon at a time along the mRNA (Alberts et al., 2002; Lodish et al., 2004). The site in the ribosome where $tRNA_i^{Met}$ resides is called the P site. There is a second tRNA binding site in the ribosome which is the A site and a third one which is the E site (Wilson and Nierhaus, 2006). When a new tRNA molecule recognizes the next codon sequence on the mRNA, it attaches to the open A site. A peptide bond forms connecting the amino acid of the tRNA in the P site to the amino acid of the tRNA in the A site. As the ribosome moves along the mRNA, the tRNA in the P site moves to the E site, then released from the E site. And the tRNA in the A site is translocated to the P site. The A binding site becomes empty again until another tRNA that recognizes the new mRNA codon takes the open position. This process keeps going and the amino acid chain grows (Alberts et al., 2002; Lodish et al., 2004; Wilson and Nierhaus, 2006).

2.1.2.2 Sequence features affecting translation elongation

Codon usage bias, amino acid usage bias, and coding sequence length are the important factors influencing translation elongation efficiency. Highly expressed genes tend to use those codons matching the most abundant cognate tRNAs (Xia, 1998; Xia, 2007b) and those cost-effective amino acids (Akashi and Gojobori, 2002), which causes codon usage bias and amino acid usage bias. In addition, the length of the coding sequence would have an effect on translation efficiency. The shorter the sequence is, the faster the protein is synthesized.

2.1.3 Termination

2.1.3.1 Termination process in eukaryotes

Termination requires the involvement of release factors (RFs). Two types of eukaryotic RFs, eRF1 and eRF3, have been discovered (Frolova et al., 2000). The shape of eRF1 is similar to that of tRNA, so it binds to the ribosomal A site and directly recognizes the stop codons (Lodish et al., 2004). The second eukaryotic release factor, eRF3, is a GTP-binding protein. The eRF3 and eRF1 together promote cleavage of the peptidyl-tRNA, thus releasing the completed protein chain (Frolova et al., 2000; Lodish et al., 2004; Rocha et al., 1999).

2.1.3.2 Sequence features affecting translation termination

The identity of the stop codon and the stop codon context are the major sequence factors influencing termination efficiency. The release factors can be divided into two classes. The first class includes RF1 and RF2 in prokaryotes and eRF1 in eukaryotes and the second class consists of RF3 in prokaryotes and eRF3 in eukaryotes (Frolova et al., 2000). The first class is stop codon specific, so the stop codon could affect termination efficiency. There are three stop codons in the Universal Genetic Code. Selection of one specific codon would be a strategy to regulate translation termination. Moreover, the stop codon context could also have an effect on termination. It has been observed that the nucleotide immediately following the stop codon is very biased in both prokaryotes and eukaryotes (Brown et al., 1990a; Brown et al., 1990b; Nie et al., 2006; Rocha et al., 1999; Shabalina et al., 2004). In addition, strong bias was found at the second last codon preceding the termination codon, which implies the bases at this position could affect translation termination too (Niimura et al., 2003).

2.2 Controversies in eukaryotic translation initiation

2.2.1 Scanning model and Internal Ribosome Entry Site

As mentioned in the above section 2.1.1.1, in eukaryotes, there are two mechanisms proposed in eukaryotic translation initiation, scanning model and internal ribosome entry site. Compared to the scanning model, IRES is still a very immature theory and tends to occur under unusual conditions. No common sequence has been identified in IRESs (Kozak, 2002; Kozak, 2005). The studies presented in this thesis are under the framework of the scanning model.

2.2.2 Two hypotheses regarding +4G in Kozak consensus

The cap-dependant scanning model is the dominant hypothesis for translation initiation in eukaryotes and the start codon context is an important component of this model (Kozak, 1999; Kozak, 2002; Kozak, 2005). Kozak consensus, RCCaugG, is the optimal context in vertebrates (Kozak, 1981; Kozak, 1984a; Kozak, 1984b; Kozak, 1986; Kozak, 1987; Kozak, 1997).

The significance of -3R, A or G, in the consensus is in agreement with mutagenesis experiments, and -3A tends to increase translation initiation more than -3G (Kozak, 1986; Kozak, 2002). In the comparisons of 5'-UTR sequences, it was also found that the -3 site is strongly biased to A (Kozak, 1981; Kozak, 1984a; Kozak, 1987), and this nucleotide is conserved across different eukaryotic taxonomic groups (Cavener, 1987; Cavener and Ray, 1991; Cigan and Donahue, 1987; Hamilton et al., 1987; Joshi et al., 1997; Kochetov, 2005; Lukaszewicz et al., 2000; Nakagawa et al., 2008; Niimura et al., 2003; Pesole et al., 2000; Sawant et al., 2001; Yamauchi, 1991).

However, the importance of +4G in the consensus is not so consistent. First, it is not conserved across different taxonomic groups (Cavener and Ray, 1991; Nakagawa et al., 2008; Niimura et al., 2003; Pesole et al., 2000). +4G was found to be conserved in vertebrates and plants (Cavener and Ray, 1991; Joshi et al., 1997; Kozak, 1991; Nakagawa et al., 2008; Niimura et al., 2003), +4A in protozoa (Yamauchi, 1991), and +4U in highly expressed yeast genes (Cigan and Donahue, 1987; Hamilton et al., 1987). Second, in Kozak's experiments (1986; 1997), the effect of +4G on initiation efficiency was dependant on the -3 site and the initiation efficiency difference with the change of +4G was much less compared to the change of -3R. A or G at the position -3 had a dominant effect. Only when C or U replaced A or G at the position -3, did translation become more sensitive to changes at the position +4. +4G increased translation initiation efficiency very little when -3R was present; the effect of +4G in influencing translation initiation efficiency became significant only when -3R was absent. Third, even in Kozak's experiment (1997), the presence of +4G, followed by U, in the mRNA did not increase translation initiation relative to the control mRNA without +4G.

There are two hypotheses regarding the role of +4G in Kozak consensus: translation initiation hypothesis and amino acid constraint hypothesis (Xia, 2007a). The translation initiation hypothesis states that +4G is a translation initiation signal in vertebrates, especially when -3R is absent. In contrast, the amino acid constraint hypothesis proposes that the +4G is not a translation initiation signal and the presence of +4G is a consequence of the biased P1' amino acid due to the constraint of N-terminal methionine excision (NME).

The experiments conducted by Kozak (1986; 1997) provide support for the translation initiation hypothesis. After compiling and analyzing the mRNA sequences mostly from vertebrates, Kozak (1981; 1984a) identified CCRCCaugG as the consensus sequence in vertebrates. To further explore the effect of single nucleotide changes around the start codon on translation efficiency, Kozak (1986) constructed a series of mutant plasmids containing preproinsulin coding sequence, which were transfected into monkey cells. The proteins were quantified by polyacrylamide gel electrophoresis. Because the preproinsulin was cleaved in the cells, the product that accumulated and was measured was proinsulin. In this experiment, the change of proinsulin abundance was used to evaluate the effect of the single nucleotide changes around the start codon on the translation efficiency. ACCaugG was identified as the optimal sequence for initiation. Mutations within that sequence resulted in a change of proinsulin over a 20-fold range. -3R had a dominant effect; when a U replaced the R at the -3 site, translation became more sensitive to changes at the -1, -2, and +4 sites.

However, in this experiment, the preproinsulin had a signal peptide at its amino terminus, which was cleaved during translation. The amino acid sequence of the signal peptide was altered due to changing +4G into other nucleotides, so the removal of the signal peptide could have been affected. Thus, the production of proinsulin was not an appropriate measure of the effect of mutations around the start codon on translation efficiency (Xia, 2007a). In addition, other studies found that +5 and +6 sites also had effects on translation efficiency by measuring protein abundance (Boeck and Kolakofsky, 1994; Grunert and Jackson, 1994; Kozak, 1997). There were similar limitations in these experiments. The effect of the nucleotide change on the 3' side of

the start codon on translation might be confounded by the change of the amino acid sequence, which can influence protein synthesis and degradation. To get around this problem, Kozak (1997) used a primer extension assay to directly monitor the ribosome-mRNA initiation complexes. A rabbit reticulocyte translation system was employed in the experiment and the inhibitors of elongation were added to hold the 80S ribosome initiation complexes at the AUG codon position. The effects of the single nucleotide changes around the start codon were evaluated by the abundance of the 80S ribosome initiation complexes. The result showed that +4G increased the recognition of AUG and alternative initiation codons and this effect was not generally influenced by +5 and +6 nucleotides. There was one exception: the mRNA with +4G followed by +5U did not increase translation initiation relative to the control mRNA without +4G. No similar experiments were done after that. Since then, RCCaugG has been widely accepted as the optimal context of the selection of start codon.

The main point of the translation initiation hypothesis are that +4G in vertebrates is a translation initiation signal and it can increase translation initiation efficiency especially when -3R is absent. However, other observations have doubt on this conclusion (Cavener and Ray, 1991; Flinta et al., 1986; Nakagawa et al., 2008; Niimura et al., 2003; Pesole et al., 2000; Xia, 2007a).

Flinta et al. (1986) studied the sequence characteristics of cytosolic amino terminal (N-terminal) protein processing. They observed that Ala, Gly, and Val occupied the P1' position in most eukaryotic protein sequences. They proposed that the prevalence of Ala, Gly and Val at the P1' position brought about the biased +4G, since Ala, Gly and Val codons all start with a G. Asp and Glu, also coded by G-starting

codons, were not likewise predominant. Thus, +4G most likely was a consequence of the constraint on the P1' amino acid rather than the requirement for translation initiation. Ala, Gly, and Val are small and uncharged amino acids which are good candidates for NME, a universal protein modification process (Giglione et al., 2004).

NME means that the initiator Met is removed from newly synthesized polypeptide, which occurs soon after the N-terminus of the growing polypeptide chain emerges from the ribosome. Only the specialized tRNA carrying Met can be used to start translation, so the first amino acid in all nascent proteins should be Met (Giglione et al., 2004). The N-terminal residues of most mature proteins are not Mets is the result of NME. NME is a universal protein modification process, conserved from bacteria to eukaryotes (Giglione et al., 2004). NME is not only an important N-terminal modification itself, but also required for further N-terminal modifications. For example, it is required for myristoylation where Gly at the N-terminus, after the removal of the initiator Met, is attached a myristoyl ($C_{14}H_{28}O_2$) fatty acid side chain (Farazi et al., 2001). The proportion of the proteins subject to NME is large as 55 ~ 70% (Giglione et al., 2004). The amino acid requirement for NME is similar in all organisms, small and uncharged amino acids are good candidates for NME (Giglione et al., 2004; Meinnel et al., 1993). There are five amino acids coded by G-starting codons, Ala, Gly, Val, Glu, and Asp. Ala, Gly, and Val are small and uncharged amino acids. Glu and Asp are charged amino acids, so they are not good candidates for NME. Methionine aminopeptidase (MAP) is the major enzyme involved in the NME process (Li and Chang, 1995; Moerschell et al., 1990; Ross et al., 2005).

The observation of Flinta et al. (1986) that Ala, Gly and Val were the prevalent P1' amino acids in 700 eukaryotic protein sequences is the result of NME. This raised a question about the role of +4G in Kozak consensus. +4G may not be related to facilitate the recognition of start codon AUG, but a result of special amino acid requirement, because Ala, Gly, and Val are all good candidates for NME and coded by G-starting codons. This is the basic idea of the amino acid constraint hypothesis formally formulated later (Xia, 2007a).

Niimura et al. (2003) conducted a comparative analysis of seven eukaryotic genomes, including *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Oryza sativa*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. The results showed that the P1' codon was the most biased one among all codons. +5C, instead of +4G, was preferentially used for all seven organisms. +4G and +6G were preferred for human, *A. thaliana,* and *O. sativa*. In agreement with these findings, the most frequent and statistically biased P1' codon for these three species was GCG encoding an Ala. For *S. cerevisiae*, and *S. pombe,* TCT coding for serine was favored as the P1' codon, although the TCT bias was not so obvious compared with GCG bias in human and plants. As for the biases at the first and second positions in the P1' codon, the authors (Niimura et al., 2003) did not tell whether the biases reflects the requirement for translational efficiency or a constraint for specific amino acid. But they also stated that the requirement for NME could make a bias for +5C, because four of seven smallest amino acids, Ala, Cys, Gly, Pro, Ser, Thr or Val, are encoded by NCN codons. In addition, the authors detected that +6 nucleotide was

also highly biased within every species but not across all species, so they suggested that +6 site was also contributed to translation initiation.

Nakagawa et al. (2008) investigated the nucleotide sequences around the start codon in more eukaryotic species including animals, plants, fungi, and protists. They also revealed that +5C was common among various eukaryote species in addition to -3R. The +4 position was highly biased, but a preferred nucleotide was not common among all eukaryotes. In vertebrates and plants, G was preferred at this position, whereas in invertebrates, fungi, and protists, T was generally preferred.

Xia (2007a) first formally characterized the two hypotheses and tested them by using human data. Xia tested the predictions from the two hypotheses as follows. The translation initiation hypothesis predicts that, selection for efficient translation initiation should favour +4G, which should drive the increased usage of amino acids coded by GNN codons at the P1' site and +4G should be prevalent in highly expressed genes than in lowly expressed genes. However, the amino acid constraint hypothesis predicts that not all GNN codons should have increased usage, but only the codons coding small and uncharged amino acids like Ala and Gly should have increased usage. The P1' codons from 34169 human coding sequences were analyzed. It was found that, among all G-starting codons, only Ala (GCN) codons showed strong bias at the P1' position compared to other positions in the coding sequences. Moreover, there was no evidence indicating that highly expressed genes were more likely to have +4G than lowly expressed genes. Among GCN, GGN, GUN, GAN, and non-G-starting codons, only GCN and GGN codons showed insignificant minor differences in their frequencies at the P1' site between high-CAI and low-CAI genes (CAI means codon adaptation index,

which is a measure of codon usage bias and related to gene expression). All of these supported the amino acid constraint hypothesis.

The essence of these two hypotheses is whether the nucleotide at +4 site is a translation initiation signal. If +4 nucleotide is a translation initiation signal, one can reason that there is some association between 40S ribosome subunit (including the rRNAs and the initiator $tRNA^{Met}$) and +4 nucleotide in mRNA to enhance translation initiation (recognition of the start codon). This association likely exists before the codon-anticodon interaction between 80S ribosome and mRNA at the P1' position in the elongation process. For example, it has been suggested that AUGG in mRNA might form a 4 base pair interaction with CCAU in the anticodon loop of initiator $tRNA^{Met}$, but no experiment has been done to test it (Kozak, 1986). Recently, Pisarev et al. (2006) detected that +4G interacted with an exclusive location (5'-$A_{1818}A_{1819}$-3') of 18S rRNA in a rabbit translation system and proposed that the role of +4G was to stabilize the conformational changes which occurred in the ribosomal complex upon the first codon-anticodon base-pairing (Pisarev et al., 2006).

2.3 Different approaches to study sequence features related to translation initiation

There are two different approaches to study sequence features related to translation initiation: an experimental approach and a bioinformatic approach. Kozak employed both approaches. Her finding of consensus comes from compilations of eukaryotic mRNA sequences (1981; 1984a; 1987). Then, she did experiments on these

consensus nucleotides (1984b; 1986; 1997). Cigan et al. (1987; 1988) also used both approaches in their work with yeast. Some other researchers represent users of a bioinformatic approach (Cavener, 1987; Cavener and Ray, 1991; Joshi et al., 1997; Nakagawa et al., 2008; Niimura et al., 2003; Pesole et al., 2000; Xia, 2007a; Yamauchi, 1991). This thesis will also employ a bioinformatic approach.

2.4 The role of *Saccharomyces cerevisiae* in studying eukaryotic translation initiation

2.4.1 The role of *Saccharomyces cerevisiae* in molecular and cell biology

Saccharomyces cerevisiae is a species of budding yeast. It is one of the most intensively studied eukaryotic model organisms in molecular and cell biology. It is a unicellular organism that has a single nucleus and reproduces either asexually by budding and transverse division or sexually through spore formation. As a eukaryote, it has complex internal cell structure as plants and animals, so it is an important research tool in the study of fundamental eukaryotic biological processes (Prescott et al., 2005). The genome of *S. cerevisiae* is the first eukaryotic genome that was completely sequenced. The *Saccharomyces* Genome Database (SGD) (http://www.yeastgenome.org) is well annotated and a very important tool for studying the function and organization of eukaryotic cells.

2.4.2 Translation initiation based on *Saccharomyces cerevisiae*

S. cerevisiae as a model eukaryotic organism is also widely used in studying eukaryotic translation initiation. The genetic and bioinformatic studies in *S. cerevisiae* generally support the scanning model (Chen et al., 2007; Cigan and Donahue, 1987; Cigan et al., 1988; Donahue and Cigan, 1988; Hamilton et al., 1987; Looman and Kuivenhoven, 1993; Looman et al., 1991; Slusher et al., 1991). They focus on the effects of the following sequence features on translation initiation: 1) the start codon identity; 2) the secondary structure of 5'-UTR; 3) the length of 5'-UTR; 4) the start codon context.

The studies with *CYC1* and *HIS4* genes (Cigan and Donahue, 1987) showed that yeast translation initiation was heavily influenced by the first AUG codon nearest the 5' end of mRNA, which provided strong evidence for the scanning model. After Kozak revealed the consensus sequence RCCaugG in vertebrates (1981; 1984a; 1984b; 1987), researchers began to compile the yeast consensus. It turned out that yeast had a totally different consensus. Cigan and Donahue (1987) reported a consensus of (A/Y)A(A/U)AaugUCU in 131 genes. Hamilton et al. (1987) compiled 96 genes and reported the consensus in 18 highly expressed genes as: (A/U)A(A/C)A(A/C)AaugUCY. From the above results, -3A, +4U, and +5C are the most outstanding common nucleotides. Cigan et al. (1988) did a mutational study of the *HIS4* initiation region, suggesting the start codon context had some but not considerable effect on translation initiation. To explore the role of 3' side of the start codon context (+4, +5 and +6 positions), Looman et al. (1991) tested 32 different P1'codons out of 61 possible codons by working on expression of the *Escherichia coli lacZ* gene in *S.*

cerevisiae, finding that there was no preference for a certain nucleotide to be at a particular position in the P1' codon and an increase in homology to $U_{+4}C_{+5}U_{+6}$ did not increase expression. Using the same method, Looman and Kuivenhoven (1993) showed that the -3 and -2 positions had a slight influence on translation. In addition, the secondary structure of the 5'-UTR had a dominant effect, but the length of the 5'-UTR had no influence. Donahue and Cigan (1988) also did a mutational study of the *HIS4* initiation region and proposed that the identity of AUG codon was more important than the context. The context had influence in the recognition of the start codon only when the AUG start codon was changed. Slusher et al. (1991) used the *MOD5* gene and demonstrated that both the 5'-UTR length and the start condon context were important in translation initiation. Chen et al. (2007) studied the most and least favourable sequence context at the -3, -2, and -1 positions for the UUG initiator of *GRS1* gene. The result revealed that AUG was much more insensitive to the start codon context than UUG.

In summary, the previous studies in the yeast translation initiation generally support the scanning model. AUG is the strongest signal in translation initiation. The secondary structure of the 5'-UTR has an effect on translation initiation, but the role of the 5'-UTR length is not consistent. The start codon context can affect the initiation and its role becomes important only when the start codon is changed from AUG to the other start codons.

Chapter 3 The 40 nucleotides upstream of the start codon are critical for translation initiation in *Saccharomyces cerevisiae*

3.1 Abstract

Early compilations of the initiation regions in yeast came from a very limited number of genes and more recent studies have no thorough analysis of 5'-UTRs. The main objective of this work is to perform a very detailed analysis of 5'-UTRs from yeast verified nuclear protein-coding genes to identify sequence features that are related to translation initiation. The result shows that the 40nt upstream of the start codon have significantly different nucleotide composition from the further upstream sequence and this is related to reducing the secondary structure. The 40nt upstream of the start codon is a critical region of 5'-UTR for translation initiation in yeast.

3.2 Introduction

According to the cap-dependent scanning model, the small ribosome subunit scans mRNA and correctly recognizes the start codon during eukaryotic translation initiation (Kozak 1999, 2002, 2005). *Saccharomyces cerevisiae* as a model eukaryotic organism has greatly enhanced the knowledge of eukaryotic translation initiation. In this chapter, yeast will be used to refer to *S. cerevisiae*. The genetic and bioinformatic studies in yeast generally support the scanning model (Chen et al., 2007; Cigan and Donahue, 1987; Cigan et al., 1988; Donahue and Cigan, 1988; Hamilton et al., 1987;

Looman and Kuivenhoven, 1993; Looman et al., 1991; Slusher et al., 1991). AUG is the strongest signal in translation initiation (Donahue and Cigan, 1988). The secondary structure of 5'-UTR is related to translation initiation, but the role of the 5'-UTR length is not known (Cigan et al., 1988; Slusher et al., 1991). -3A dominates in the start codon context (Cigan and Donahue, 1987; Hamilton et al., 1987).

5'-UTR is an important player in yeast translation initiation. However, there are several limitations in the previous studies of 5'-UTRs in yeast. First, early compilations of the initiation regions in yeast covered a limited set of genes. The consensuses reported from Cigan et al. (1987) and Hamilton et al. (1987) are the results of 131 and 96 genes, respectively. Second, more recent studies (Nakagawa et al., 2008; Niimura et al., 2003) involve a large number of yeast genes, but there are no thorough analysis of the 5'-UTRs.

The main objective of this work is to perform a very detailed analysis of 5'-UTRs from 4681 verified nuclear protein-coding genes in yeast. The nucleotide composition and secondary structure are the two major components in the analysis, which help identify the sequence features of 5'-UTR that are related to translation initiation.

3.3 Materials and Methods

The genomic sequence files of *S. cerevisiae*, orf_genomic_1000.fasta.gz and orf_coding.fasta.gz, were downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/orf_dna dated Sep. 6, 2007.

This web site is a component of SGD. The file orf_genomic_1000.fasta.gz contains 5883 annotated nuclear and mitochondrial protein-coding genes (excluding dubious ones) with introns and untranslated region 1000nt upstream of the start codon and downstream of the stop codon. The file orf_coding.fasta.gz only has the coding sequences (CDSs) of all 5883 genes, without 5'-UTR, 3'-UTR, intron sequences, or bases not translated due to translational frameshifting. The present work aims to study the translation initiation occurring in the cytoplasm, so the mitochondrial genes were excluded. To enhance the reliability of the study, uncharacterized genes were also removed. The majority of the uncharacterized genes are putative ones. The gene names of all nuclear verified genes were obtained from http://db.yeastgenome.org/cgi-bin/search/featureSearch residing at SGD. There are a total of 4681 verified nuclear protein-coding genes all of which have AUG as the start codon.

The experimentally verified mRNA 5'-UTR sequence would be the ideal candidate to investigate. Unfortunately, no large-scale experimentally verified mRNA 5'-UTR sequences are available in yeast. In an early study of the yeast mRNA 5'-UTR sequences (Cigan and Donahue, 1987), 86 genes with known transcriptional start points have an average 5'-UTR length of 52nt. A large-scale experiment of mapping transcription start sites in yeast reveals that most transcripts start within the region 15-75nt upstream of the start codon (Zhang and Dietrich, 2005). Therefore, the sequences of 100nt upstream of the start codon should be able to recover the majority of the upstream part of the translation initiation region. Another aspect worth noting is that the genomic upstream sequences of some coding sequences contain introns so they do not correspond to the mature 5'-UTR sequences. Most yeast genes do not contain introns

(Brown, 2007) and fewer protein-coding genes contain introns in the premature 5'-UTR region. Among 4681 verified nuclear protein-coding genes, only 24 have introns in the premature upstream region, all having the ending positions of the introns within the region 100nt upstream of the start codon. For these 24 genes, the upstream 100nt sequences excluding the introns were manually extracted. Table 3.1 lists the names and intron positions of these 24 genes. After excluding the intron sequences, the 100nt upstream of the start codon of 4681 verified nuclear protein-coding genes were analysed. Data Analysis in Molecular Biology and Evolution software package (DAMBE) (Xia, 2000; Xia and Xie, 2001) was used to extract the 100nt upstream of the start codon of these 4681 genes.

This work analysed the nucleotide composition of the 100nt upstream of the start codon. The site-specific nucleotide percent frequencies were used to measure the nucleotide composition. Let N_{ij} as the frequency of the nucleotide i at the position j, so the site-specific percent frequency F_{ij} is N_{ij} /4681.

To measure the potential of forming secondary structure of the different segments in the 5'-UTR, this work chose a window size of 40nt with a step size of 10nt to calculate the minimum folding energy (MFE) for each window in the 100nt upstream of the start codon. DAMBE (Xia, 2000; Xia and Xie, 2001) was used to compute MFE. Because temperature can affect the secondary structure, this work tried two folding temperatures, 37°C and 30°C. No helices of length 1 and no GU pairs at the end of helices were specified. DAMBE uses the same function library as the RNAfold program of the Vienna RNA package (Hofacker, 2003) available at http://www.tbi.univie.ac.at/~ivo/RNA/. The maximum value of MFE is 0, associated

with the least secondary structure. A smaller negative MFE value means more stable secondary structure. Zuker's energy minimization algorithm (Zuker and Stiegler, 1981) is the essence of the MFE computation, which is based on a dynamic programming algorithm. Zuker's algorithm has been widely used to predict the RNA secondary structure (Jia and Li, 2005).

Table 3. 1 – The names and intron positions of 24 yeast protein-coding genes which have introns in their 5'-UTRs

The position number is relative to the start codon.

Systematic Name	Standard Name	Chromosome	Intron position
YBL072C	RPS8A	2	-315..-8
YBL092W	RPL32	2	-333..-1
YBR089C-A	NHP6B	2	-384..-28
YDL061C	RPS29B	4	-421..-13
YDL137W	ARF2	4	-371..-40
YDL189W	RBS1	4	-138..-40
YDR099W	BMH2	4	-826..-84
YER102W	RPS8B	5	-367..-8
YER131W	RPS26B	5	-361..-1
YFR032C-A	RPL29	6	-334..-4
YGL031C	RPL24A	7	-463..-8
YGL187C	COX4	7	-354..-13
YGL189C	RPS26A	7	-378..-11
YGR027C	RPS25A	7	-327..-16
YGR148C	RPL24B	7	-399..-8
YIL123W	SIM1	9	-489..-3
YJL130C	URA2	10	-385..-66
YKL150W	MCR1	11	-144..-57
YKL186C	MTR2	11	-167..-14
YLR333C	RPS25B	12	-436..-14
YLR367W	RPS22B	12	-564..-8
YLR388W	RPS29A	12	-493..-6
YNL066W	SUN4	14	-358..-13
YPL230W	USV1	16	-93..-19

The chi-square goodness of fit test (Zar, 1984) was used to test the significance of nucleotide frequency difference between two sequence segments. The Wilcoxon signed-rank test (Dowdy et al., 2004) was used to test the significance of the MFE difference between two windows. All tests were run in R (http://www.r-project.org/).

3.4 Results and Discussion

3.4.1 The nucleotide composition of the upstream 100nt

Figure 3.1 shows the site-specific nucleotide percent frequency in the 100nt upstream of the start codon in 4681 verified nuclear protein-coding genes in yeast.

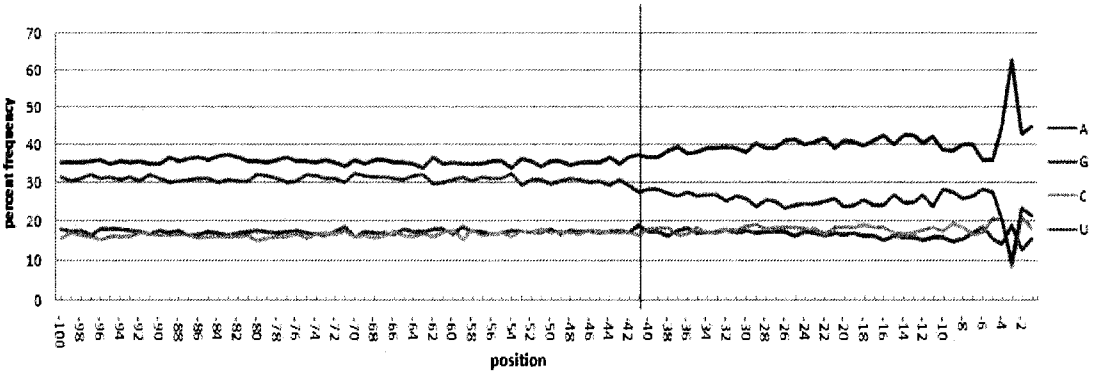


Figure 3. 1 - The site-specific nucleotide percent frequencies of the 100nt upstream of the start codon

X-axis refers to the position number relative to the start codon, and Y-axis indicates the site-specific nucleotide percent frequency

For the all 100 upstream positions, A is the most often used nucleotide with the frequency being 33.7% - 62.8%, and U is the second frequently used one except at the

-3 position, ranging from 9.7% - 32.4%. The frequencies of G and C are fluctuating around 12.8% - 19.1% and 8.4% - 21.0%, respectively. The yeast genome is AT rich and these upstream sequences reflect this feature. Noticeably, starting from the position -41, A has an increased usage. In contrast, almost beginning at the same position, U has a notable decreased frequency (see the red line in Figure 3.1). A previous study (Shabalina et al., 2004) showed a similar observation in 2066 orthologous genes from four yeast species, *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*. In the segment of -40 to -8, the frequency difference between A and U (designated as DF_{A-U}) is greater than in the segment of -100 to -41. At the -6 and -5 sites, the difference decreases with a sharp increase at the position -3. At the -2 and -1 sites, the difference is large too, but it is much less than the difference at the position -3. At the sites -5, -4, -3, -2, the site-specific frequency difference between G and C (designated as DF_{G-C}) is greater than the difference at all other sites. DF_{G-C} has a different direction at the -3 site from the -5, -4, and -2 sites. Across the upstream 100 positions, DF_{G-C} is much smaller than DF_{A-U}. The average of the absolute value of DF_{A-U} is 8.8%, but of DF_{G-C} is 1.45%.

In summary, three segments can be identified with different features among the 100 upstream positions, -100 to -41, -40 to -6, and -5 to -1. For the segment of -100 to -41, DF_{A-U} is fluctuating around 1.3% - 9.4%, and the frequency of G is a little higher than C at the majority of sites. For the segment of -40 to -6, DF_{A-U} increases, and the frequency of C is a little higher than G at most sites, which is opposite to the segment of -100 to -41. Except G, the average frequencies of A, U, and C are significantly different from their average frequencies in the segment of -100 to -41 (Table 3.2). From the -5 site to the -1 site, the frequencies of the four nucleotides have dramatic changes. At the

-3 site, there is a sharp increase of A with a great decrease of C and U. The frequency of G does not change very much, but shows higher than C and U due to the obvious decrease of C and U. In the segment of -5 to -1, the average frequencies of the four nucleotides are significantly different from in the segment of -100 to -41 (Table 3.2).

Table 3. 2 – The average, minimum and maximum site-specific nucleotide frequencies of the 100nt upstream of the start codon

The P value is under the null hypothesis that the average frequency is the same as the average frequency of the segment -100 to -41. The -3 site frequencies are listed alone in the last column on the "Average" rows. All the χ^2 values are associated with a degree of freedom of 1. On the "P" rows, "*" indicates that the P value shows significance.

		Sites -100 to -41	Sites -40 to -6	Sites -5 to -1	Site -3
A	Average	0.354	0.396	0.462	0.628
	Minimum	0.337	0.357	0.359	
	Maximum	0.371	0.426	0.628	
	χ^2		36.108	238.754	1536.753
	P		0.000 *	0.000 *	0.000 *
U	Average	0.308	0.258	0.206	0.097
	Minimum	0.276	0.234	0.097	
	Maximum	0.324	0.284	0.277	
	χ^2		54.906	228.498	977.793
	P		0.000 *	0.000 *	0.000 *
G	Average	0.173	0.167	0.155	0.191
	Minimum	0.161	0.150	0.128	
	Maximum	0.191	0.189	0.191	
	χ^2		1.178	10.601	10.601
	P		0.278	0.001 *	0.001 *
C	Average	0.164	0.178	0.177	0.084
	Minimum	0.148	0.164	0.084	
	Maximum	0.180	0.192	0.210	
	χ^2		6.692	5.770	218.509
	P		0.010 *	0.016 *	0.000 *

This result shows that the upstream 40nt of the start codons have a salient discrepancy in nucleotide composition from the further upstream sequences. Before the site -40, the frequencies of the four nucleotides are maintained at a stable level with AU being larger than GC, and even the frequency differences between A and U and between G and C are somewhat uniform. Starting from the site -40, the frequency difference between A and U increases and the frequency difference between G and C changes direction. The previous studies (Cigan and Donahue, 1987; Hamilton et al., 1987) showed that G was the least used nucleotide in the yeast 5'-UTR. From the present work, that conclusion is true for the upstream 40nt, but before the position -40, C is the least used nucleotide. In the investigation of Cigan et al. (1987), C and U were equally represented from the positions -25 to -1, which is not consistent with the present work. The frequency of U is greater than C, not only from the positions -100 to -26, but also from -25 to -1. In contrast, C and G are roughly equally represented from the positions -100 to -1 except the -3 site. More interestingly, the frequency difference between G and C changes direction around the position -40. The result from the present work is based on much more yeast genes than the previous studies (Cigan and Donahue, 1987; Hamilton et al., 1987), so it more accurately reflects the nucleotide composition of the yeast 5'-UTR.

3.4.2 The secondary structure of the upstream 100nt

The secondary structure of 5'-UTR is related to translation initiation efficiency (Cigan et al., 1988; Slusher et al., 1991). Less secondary structure can facilitate ribosome scanning process (Kozak, 1999; Kozak, 2002; Kozak, 2005), so the increased

A and decreased U in the 40nt upstream sequences could be a strategy for yeast to reduce the secondary structure near the start codon. To measure the potential of forming secondary structure of different segments in the 5'-UTR, the upstream 100nt were separated into seven windows. The size of each window is 40nt and the window slides from 5' to 3' with a step size of 10nt. The MFE values at 37°C and 30°C for each window of each gene were calculated (see Materials and Methods). Figure 3.2 shows the boxplots of MFE of each window at 37°C and 30°C.

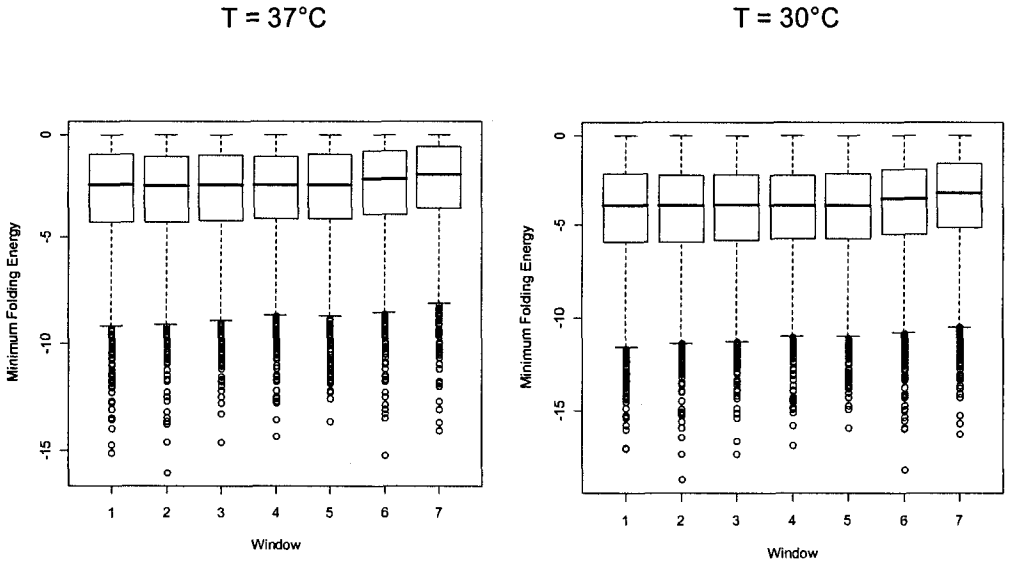


Figure 3. 2 - The boxplot of the minimum folding energy of seven windows in the 100nt upstream of the start codon

X-axis refers to the different windows. Y-axis refers to the minimum folding energy. The segment (indicated by the position relative to the start codon) represented by each window is as follows, Window 1: -100 to -61; Window 2: -90 to -51; Window 3: -80 to -41; Window 4: -70 to -31; Window 5: -60 to -21; Window 6: -50 to -11; Window 7: -40 to -1. The median MFE at 37°C for each window is as follows, Window 1: -2.50; Window 2: -2.52; Window 3: -2.50; Window 4: -2.50; Window 5: -2.50; Window 6: -2.20; Window 7: -2.00. The median MFE at 30°C for each window is as follows, Window 1: -3.94; Window 2: -3.96; Window 3: -3.91; Window 4: -3.95; Window 5: -3.93; Window 6: -3.56; Window 7: -3.26.

The result of the Wilcoxon signed-rank test for the MFE differences between any two windows is shown in Table 3.3.

Table 3. 3 – The result of Wilcoxon signed-rank test for the MFE differences of seven windows in the 100nt upstream of the start codon

The P value is under the null hypothesis that the minimum folding energy has no difference between the two windows compared. The sequence segments represented by the windows are the same as used in Figure 3.2. "*" indicates that the P value shows significance.

Compared Windows	T = 37°C		T=30°C	
	P value	Significance	P value	Significance
Window 1...Window 2	0.433		0.349	
Window 1...Window 3	0.744		0.889	
Window 1...Window 4	0.350		0.508	
Window 1...Window 5	0.084		0.144	
Window 1...Window 6	0.000	*	0.000	*
Window 1...Window 7	0.000	*	0.000	*
Window 2...Window 3	0.647		0.675	
Window 2...Window 4	0.175		0.235	
Window 2...Window 5	0.023	*	0.028	*
Window 2...Window 6	0.000	*	0.000	*
Window 2...Window 7	0.000	*	0.000	*
Window 3...Window 4	0.475		0.493	
Window 3...Window 5	0.048	*	0.033	*
Window 3...Window 6	0.000	*	0.000	*
Window 3...Window 7	0.000	*	0.000	*
Window 4...Window 5	0.242		0.147	
Window 4...Window 6	0.000	*	0.000	*
Window 4...Window 7	0.000	*	0.000	*
Window 5...Window 6	0.000	*	0.000	*
Window 5...Window 7	0.000	*	0.000	*
Window 6...Window 7	0.000	*	0.000	*

The patterns of the MFE values at both folding temperatures are similar. The median MFE values of Window 6 or Window 7 which is closer to the start codon is higher than the values of the other windows (Figure 3.2). The difference between Window 6 or Window 7 and any other window is significant (Table 3.3). Window 7 has the highest median MFE values (Figure 3.2). This shows that the segment of -40 to -1 has the least potential of forming secondary structure. Given the observation that the 40nt upstream of the start codon have an obvious discrepancy in nucleotide composition from the further upstream sequences, it could be reasoned that increasing A and decreasing U in this segment is related to reducing the secondary structure near the start codon for enhancing the recognition of the start codon. It seems that the 40nt upstream of the start codon have a special role in translation initiation.

3.4.3 Association with the previous studies

An experiment with the yeast *PGK* gene showed that translation could occur at the AUG whose 5'-UTR length was reduced to 7nt (van den Heuvel et al., 1989). Completely deleting the 5'-UTR of the yeast *tcm1* gene did not prevent the initiation from the first AUG (Maicas et al., 1990). However, the initiation efficiency in the above experiments was considerably decreased. Given the observations from the present work, it could be speculated that there is a critical region in the 5'-UTR for efficient recognition of AUG, and the 40nt upstream of the start codon may be this region. This may also explain the results from two other experiments. The results about the effect of the 5'-UTR length on translation efficiency are inconsistent in the following two experiments. The 5'-UTR length change from 115nt to 39nt had no effect on gene

expression in the yeast *HIS4* gene (Cigan et al., 1988). In contrast, a study with the yeast *MOD5* gene (Slusher et al., 1991) revealed that a sequence segment of at least 36nt long added into the 5'-UTR greatly influenced the alternative AUG selection. But, if the upstream 40nt is the critical region for translation initiation, the inconsistency in the above two experiments becomes consistency. In the former experiment, because the original 5'-UTR was 39nt long and increasing the 5'-UTR length did not change the critical region of the original sequence, the translation initiation efficiency did not change. In the latter experiment, the wild type gene had a first AUG located 10 ~ 11nt from the 5' end and a second AUG 43 ~ 44nt from the 5' end. About 5-10% of translation was initiated from the first AUG and 90-95% was from the second AUG. But after a sequence segment of at least 36nt long was inserted before the original 5' end, almost all initiation began at the first AUG. This can be explained by the critical upstream 40nt too. When the short 5'-UTR of 10 ~ 11nt for the first AUG were longer than 40nt after the insertion, the recognition of this AUG became much more efficient.

Another study in the evolution of eukaryotic 5'-UTR (Lynch et al., 2005) shows that all eukaryotic genes have very few 5'-UTRs shorter than 50nt. This adds another support for the proposal that the upstream 40nt is a critical region for translation initiation in yeast.

It can be speculated that the nucleotide composition pattern found in the upstream 40nt may be lost in pseudogenes. In the near future, the comparison of the upstream 40nt between functional genes and pseudogenes will be done.

3.4.4 -3 site

The dramatic nucleotide frequency changes at the -3 site, combined with the highly conserved -3R across different taxonomic groups with genomes having different GC content (Nakagawa et al., 2008; Niimura et al., 2003), show strong selection power at the -3 site. Up to now, the mechanism behind this is still a mystery. With the early discovery of consensus CCRCCaugG in vertebrate mRNAs (Kozak, 1981; Kozak, 1984a), it was proposed that a sequence in 18S RNA, 3' -GUGG- 5', was base-paired with 5' –CACC- 3' in mRNA (Kozak, 1986). Two regions were of particular interest: one at 8nt from the 3' end of 18S rRNA (Sargan et al., 1982) and the other near the P site where initiator $tRNA^{Met}$ attached (Kozak, 1986). However, because CCRCCaugG is not conserved among different eukaryotic taxonomic groups, the significance of the 3' -GUGG- 5' in 18S rRNA in translation initiation is doubtful (Cigan and Donahue, 1987). To consider different consensus sequences at the -2 and -1 sites across different lineages, Cavener et al. (1991) suggested another model based on the model of Sargan (1982). It postulated that the -3, -2, -1 sites in mRNA may pair with one of three triplets in a five nucleotide segment, 3'-UUUGG-5', of the 18S rRNA. Site-I (3'-UUU-5') would perfectly base pair with AAA, Site-II (3'-UUG-5') would do so with AAC, and Site-III (3'-UGG-5') would bind to ACC. Site-III was part of the sequence Sargan proposed. But, why and how the 18S rRNA shifts to pair with mRNA in different organisms is a big question for this model. Recently, it was reported that -3G in an mRNA interacts with a eukaryotic initiation factor eIF2α (Pisarev et al., 2006). The experiment compared the interactions of -3G and -3U with the components of the translation initiation complex, but it did not test -3A. Therefore, it can not explain the

mechanism of -3 site, because the most conserved nucleotide at this position is A, not G, among all eukaryotes.

Several lines of evidence in yeast allow making a few tentative inferences. First, -3A is much more frequent than -3G which in turn is much more frequent than -3C and -3U. Given that the yeast genome is AT-rich and that U is the second most used nucleotide in the whole upstream region, the rarity of -3U must be maintained by strong selection against it. Second, from -40 to -1, most sites have more C than G, but C is sharply decreased at -3 site, so there must be strong selection against -3C, too. Third, U:A base pairing is the typical Watson-Crick pairing. Among the non canonical base pairings U:G, U:C, and U:U, U:G is the most stable one, and U:C and U:U are generally weak (Curran, 1998). All of the above information points to the possibility of a U somewhere interacting with the -3 site in the mRNA. The location of this U should be investigated. The mostly likely place is in the 18S rRNA, but this should not exclude the possibility of the initiator tRNA and other components in the small ribosome subunit.

3.4.5 The relationship between initiation consensus and genome nucleotide composition

In Kozak consensus RCCaugG, in addition to R at -3 site, C appears at -2 and -1 sites. In yeast, A is the consensus and U is the second most often used nucleotides at these two positions. Therefore, C in Kozak consensus may not be the translation initiation signal, but a reflection of genome nucleotide composition because Kozak consensus was compiled from vertebrate genes mostly composed of mammalian genes which have GC-rich 5'-UTRs (Nakagawa et al., 2008; Shabalina et al., 2004). Pesole et

al. (2000) investigated human genes belonging to different isochores. It was shown that the preference of A and G at -3 site was maintained for all isochore compartments, but the increase of C at -2 corresponded to an equal decrease of A with the increase of GC content in different isochores. All of these indicate the necessity of studying the relationship between consensus and genome nucleotide composition.

3.5 Conclusion

This work shows that, in yeast, the 40nt upstream of the start codon have significantly different nucleotide composition from the further upstream sequences and this is related to reducing the secondary structure near the start codon. The 40nt upstream of the start codon is a critical region of 5'-UTR for translation initiation in yeast.

Chapter 4 Comparative analysis of start codon context in seven eukaryotic species

4.1 Abstract

Start codon context is an important component of Kozak's scanning model in eukaryotic translation initiation. Kozak identified RCCaugG as the initiation consensus in vertebrates. There are two hypotheses regarding the role of +4G in this consensus, translation initiation hypothesis and amino acid constraint hypothesis. According to the translation initiation hypothesis, +4G is important for translation initiation especially when –3R is absent. This predicts that +4G would appear more in the genes with -3Y than -3R. This work examined the relationship between -3 and +4 sites in seven eukaryotic species and showed that the empirical data contradicts the prediction from the translation initiation hypothesis. According to the amino acid constraint hypothesis, not all amino acids coded by G-starting codons are overused at the P1' position, but only the small and uncharged amino acids are overused at the P1' position. This work explored the amino acid usage at the P1' position compared to the other positions of all coding sequences in the genome for the seven eukaryotic species. The result showed that, Ala is remarkably overused at the P1' position in *Homo sapiens*, *Mus musculus*, *Danio rerio,* and *Arabidopsis thaliana,* Ser is remarkably overused in *Caenorhabditis elegans* and *Saccharomyces cerevisiae,* and both Ala and Ser are remarkably overused in *Drosophila melanogaster.* All of these are consistent with the amino acid constraint hypothesis that proposes +4G in Kozak consensus is a result of biased P1' amino acid in vertebrates.

4.2 Introduction

Kozak (1978a; 1978b) proposed the cap-dependant scanning model for translation initiation in eukaryotes. During the ribosome scanning process, the selection of the initiating AUG is affected by the surrounding nucleotides (Kozak, 1981), which are called start codon context. In the later compilations and experimental verifications, Kozak (1984b; 1986; 1987; 1991; 1997) found the optimal start codon context RCCaugG in vertebrates, which is called the Kozak consensus sequence. In this consensus, it is believed that the most important nucleotides are -3R and +4G.

The significance of -3R on translation initiation is in agreement with mutagenesis experiments, and -3A tends to increase translation initiation more than -3G (Kozak, 1986; Kozak, 2002). The comparisons of 5'-UTR sequences across different eukaryotic taxonomic groups also show -3A is very conservative (Cavener, 1987; Cavener and Ray, 1991; Cigan and Donahue, 1987; Hamilton et al., 1987; Joshi et al., 1997; Nakagawa et al., 2008; Niimura et al., 2003; Pesole et al., 2000; Yamauchi, 1991).

However, the importance of +4G for translation initiation is controversial. There are several lines of evidence indicating the inconsistency of +4G. First, +4G is not conserved across different taxonomic groups. It is conserved only in vertebrates and plants (Cavener and Ray, 1991; Joshi et al., 1997; Kozak, 1991; Nakagawa et al., 2008; Niimura et al., 2003), not in invertebrates (Cavener, 1987; Cavener and Ray, 1991; Nakagawa et al., 2008; Niimura et al., 2003), protists (Yamauchi, 1991), and fungi (Cigan and Donahue, 1987; Hamilton et al., 1987). Second, Kozak's experiment (1997) showed that the presence of +4G, followed by U, in the mRNA did not increase

translation initiation relative to the control mRNA without +4G. Third, the effect of +4G was not as much as that of the -3 site (Kozak, 1986; Kozak, 1997). A or G at the position -3 had a dominant effect. Only when C or U replaced A or G at the position -3, did initiation efficiency become more sensitive to changes at the +4 position and the difference of initiation efficiency with the change of +4G was much less compared to the change of -3R.

On the other hand, Ala is overused at the P1' position in human and Ala is coded by GCN codons (Xia, 2007a). Therefore, whether +4G enhances the recognition of start codon in vertebrates, or simply reflects an amino acid bias, is a question.

The two hypotheses regarding the role of +4G in Kozak consensus (Xia, 2007a) have different interpretation about the presence of +4G. The translation initiation hypothesis states that +4G is a translation initiation signal and important for translation initiation when -3R is absent, so it predicts that +4G would appear more in the genes with -3Y than -3R. The amino acid constraint hypothesis proposes that the presence of +4G is a consequence of biased P1' amino acid due to the constraint of NME, so not all amino acids coded by G-starting codons are overused at the P1' position, but only the small and uncharged amino acids, which are efficient for NME, are overused at the P1' position.

Up to now, most bioinformatic studies in start codon context have focused on finding and comparing the consensus sequences in different species (Cigan and Donahue, 1987; Hamilton et al., 1987; Joshi et al., 1997; Kozak, 1991; Nakagawa et al., 2008; Niimura et al., 2003; Yamauchi, 1991). No studies have compiled data about the

relationship between -3 and +4 sites across different taxonomic groups. This chapter will examine this issue in seven eukaryotic species. Moreover, this chapter will analyze the amino acid usage at the P1' position compared to the rest positions of the CDSs in the genome for the seven eukaryotic species to explore the relationship between the +4 nucleotide and the constraint of P1' amino acid usage. The relationship between -3 and +4 sites and the relationship between the +4 nucleotide and the constraint of P1' amino acid usage will give insights into the start codon context and help test the two hypotheses regarding the role of +4G in Kozak consensus.

4.3 Materials and Methods

For simplicity, the common names, human, mouse, zebrafish, fruitfly, nematode, yeast, and *Arabidopsis*, will be used to refer to *Homo sapiens*, *Mus musculus*, *Danio rerio, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae,* and *Arabidopsis thaliana,* respectively.

The complete genomes of human, mouse, zebrafish, fruitfly, nematode, yeast, and *Arabidopsis*, were retrieved from National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=Genome) on April 1, 2007. The GenBank files of all contigs from reference assembly for human chromosomes 1 to 22, X, and Y were downloaded, so were all contigs from reference assembly for mouse chromosomes 1 to 19, X, and Y; all contigs from reference assembly for zebrafish chromosomes 1 to 25; fruitfly chromosomes 2L, 2R, 3L, 3R, 4, and X; nematode chromosomes I, II, III, IV, V, and X; yeast chromosomes 1 to 16; and

Arabidopsis chromosomes 1 to 5. The nucleotides from the positions -3 to +6 of all CDSs were extracted by using DAMBE (Xia, 2000; Xia and Xie, 2001). Translating the P1' codon into the P1' amino acid was also done in DAMBE. Excluding some CDSs for which the upstream information could not be retrieved and some CDSs for which a stop codon appears at the P1' position, the total numbers of CDSs in this work are 29113, 27402, 29963, 36798, 23209, 5854, and 30464 for human, mouse, zebrafish, fruitfly, nematode, yeast, and *Arabidopsis*, respectively.

In order to investigate the relationship between the -3 and +4 sites, this work grouped the CDSs of each species first by -3 nucleotide, then by +4 nucleotide.

To explore the relationship between the +4 nucleotide and the constraint of P1' amino acid usage, this work investigated the amino acid usage bias at the P1' position within the seven species. First, the percent frequency of each amino acid used at the P1' site from all CDSs within each species was computed. However, the P1' amino acid usage can be affected by the overall usage of amino acids within the proteome, which is controlled by the availability of the amino acid pool, tRNA pool and many other factors. So the overall amino acid usage for each species should be considered when the usage bias of the P1' amino acid is investigated and compared. Therefore, in the second step, the proportion of each amino acid at the P1' position in the whole proteome and the proportion of each amino acid used at the other positions other than the initiator (Met) and P1' positions in the whole proteome were calculated for each species. The latter proportion serves as a background proportion for each amino acid in each species, which is the expected proportion of each amino acid at the P1' position if there is no amino acid usage bias at the P1' site compared to the other sites. The proportion of

amino acid in the background is designated as P_{aa} and the number of the total CDSs examined in the species as M. If the usage of an amino acid at the P1' site is the same as the rest of the CDSs, then the expected number of a given amino acid at the P1' site is $P_{aa} \times M$.

The chi-square contingency table test (Zar, 1984) was used to test the +4 nucleotide frequency difference between genes with -3R and -3Y. The chi-square goodness of fit test (Zar, 1984) was used to test the amino acid frequency difference between P1' position and the rest positions in CDS.

4.4 Results and Discussion

4.4.1 The relationship between -3 and +4 sites

The frequencies of the grouped CDSs for each species are listed in Table 4.1.

Table 4. 1 - The frequencies of CDSs grouped by nucleotides at -3 and +4 sites in seven species

The numbers in parentheses are the proportions (percentage) of CDSs grouped by the nucleotide at +4 site in the corresponding -3 group. The χ^2 values and P values are under the null hypothesis that the proportions of the nucleotide at the +4 site are the same between -3R group and -3Y group. All the χ^2 values are associated with a degree of freedom of 1. On the "P" rows, "*" indicates that the P value shows significance.

Species	-3 Site	+4A	+4G	+4C	+4U	Sum
human	A	3206(24.9)	5838(45.4)	1911(14.9)	1900(14.8)	12855
	G	2046(19.0)	5705(53.0)	1687(15.7)	1331(12.4)	10769
	C	659(19.3)	1697(49.6)	677(19.8)	386(11.3)	3419
	U	523(25.2)	876(42.3)	358(17.3)	313(15.1)	2070
	R	5252(22.2)	11543(48.9)	3598(15.2)	3231(13.7)	23624
	Y	1182(21.5)	2573(46.9)	1035(18.9)	699(12.7)	5489
	overall	6434(22.1)	14116(48.5)	4633(15.9)	3930(13.5)	29113
	χ^2	1.219	6.953	43.484	3.306	
	P	0.270	0.008 *	0.000 *	0.069	
mouse	A	3319(25.8)	5849(45.5)	1849(14.4)	1835(14.3)	12852
	G	1860(19.4)	5063(52.8)	1400(14.6)	1267(13.2)	9590
	C	567(19.7)	1437(50.0)	529(18.4)	340(11.8)	2873
	U	461(22.1)	932(44.7)	336(16.1)	358(17.2)	2087
	R	5179(23.1)	10912(48.6)	3249(14.5)	3102(13.8)	22442
	Y	1028(20.7)	2369(47.8)	865(17.4)	698(14.1)	4960
	overall	6207(22.7)	13281(48.5)	4114(15.0)	3800(13.9)	27402
	χ^2	12.686	1.172	27.704	0.192	
	P	0.000 *	0.279	0.000 *	0.661	
zebrafish	A	4433(26.1)	7424(43.7)	2120(12.5)	3017(17.8)	16994
	G	1843(22.7)	3939(48.6)	1033(12.7)	1293(15.9)	8108
	C	634(24.1)	1211(46.1)	415(15.8)	368(14.0)	2628
	U	601(26.9)	943(42.2)	309(13.8)	380(17.0)	2233
	R	6276(25.0)	11363(45.3)	3153(12.6)	4310(17.2)	25102
	Y	1235(25.4)	2154(44.3)	724(14.9)	748(15.4)	4861
	overall	7511(25.1)	13517(45.1)	3877(12.9)	5058(16.9)	29963
	χ^2	0.333	1.463	19.475	9.092	
	P	0.564	0.226	0.000 *	0.003 *	

(This table is to be continued. Please turn to Page 46.)

(Table 4.1 – continued from Page 45)

Species	-3 Site	+4A	+4G	+4C	+4U	Sum
fruitfly	A	5923(27.1)	7030(32.2)	3832(17.5)	5052(23.1)	21837
	G	2198(24.5)	3502(39.0)	1552(17.3)	1734(19.3)	8986
	C	772(25.6)	1055(35.0)	612(20.3)	572(19.0)	3011
	U	794(26.8)	950(32.1)	618(20.9)	602(20.3)	2964
	R	8121(26.3)	10532(34.2)	5384(17.5)	6786(22.0)	30823
	Y	1566(26.2)	2005(33.6)	1230(20.6)	1174(19.6)	5975
	overall	9687(26.3)	12537(34.1)	6614(18.0)	7960(21.6)	36798
	χ^2	0.042	0.810	32.798	16.408	
	P	0.837	0.368	0.000 *	0.000 *	
nematode	A	4268(32.9)	3498(26.9)	2403(18.5)	2821(21.7)	12990
	G	1551(31.2)	1470(39.5)	914(18.4)	1041(20.9)	4976
	C	919(29.8)	849(27.5)	650(21.1)	669(21.7)	3087
	U	639(29.6)	629(29.2)	403(18.7)	485(22.5)	2156
	R	5819(32.4)	4968(27.7)	3317(18.5)	3862(21.5)	17966
	Y	1558(29.7)	1478(28.2)	1053(20.1)	1154(22.0)	5243
	overall	7377(31.8)	6446(27.8)	4370(18.8)	5016(21.6)	23209
	χ^2	13.253	0.559	6.874	0.603	
	P	0.000 *	0.455	0.009 *	0.437	
yeast	A	1065(29.6)	1137(31.6)	448(12.5)	945(26.3)	3595
	G	359(32.0)	325(29.0)	124(11.1)	313(27.9)	1121
	C	189(37.1)	103(20.2)	86(16.9)	132(25.9)	510
	U	226(36.0)	162(25.8)	75(11.9)	165(26.3)	628
	R	1424(30.2)	1462(31.0)	572(12.1)	1258(26.7)	4716
	Y	415(36.5)	265(23.3)	161(14.1)	297(26.1)	1138
	overall	1839(31.4)	1727(29.5)	733(12.5)	1555(26.6)	5854
	χ^2	16.451	25.863	3.229	0.128	
	P	0.000 *	0.000 *	0.072	0.720	
Arabidopsis	A	3885(25.8)	7847(52.2)	1170(07.8)	2132(14.2)	15034
	G	1691(22.3)	4455(58.9)	502(06.6)	921(12.2)	7569
	C	642(19.8)	1861(57.4)	262(08.1)	477(14.7)	3242
	U	822(17.8)	2985(64.6)	283(06.1)	529(11.5)	4619
	R	5576(24.7)	12302(54.4)	1672(07.4)	3053(13.5)	22603
	Y	1464(18.6)	4846(61.6)	545(06.9)	1006(12.8)	7861
	overall	7040(23.1)	17148(56.3)	2217(07.3)	4059(13.3)	30464
	χ^2	119.635	123.268	1.795	2.483	
	P	0.000 *	0.000 *	0.180	0.115	

Several general patterns are visible in Table 4.1. First, the presence of -3R is a common feature among all compiled eukaryotic genomes. The proportions of genes with -3R are 81.1%, 81.9%, 83.8%, 83.8%, 77.4%, 80.6%, and 74.2% in human, mouse, zebrafish, fruitfly, nematode, yeast, and *Arabidopsis*, respectively. A is the dominant nucleotide at the -3 site. The proportions of genes with -3A are 44.2%, 46.9%, 56.7%, 59.3%, 56.0%, 61.4%, and 49.4% in human, mouse, zebrafish, fruitfly, nematode, yeast, and *Arabidopsis*, respectively. The values reflect an association with genome nucleotide composition. For example, yeast has an AT-rich genome, so it has a very high proportion of -3A. The prevalent -3R is consistent with previous findings (Cavener, 1987; Cavener and Ray, 1991; Joshi et al., 1997; Nakagawa et al., 2008; Niimura et al., 2003; Pesole et al., 2000; Yamauchi, 1991) and further indicates the importance of -3R. Second, +4G is common among genes in the three vertebrates species (human, mouse and zebrafish) and *Arabidopsis*, but much less common in invertebrate species (fruitfly, nematode) and yeast. In nematode and yeast, +4A is the most frequent nucleotide. There is a discrepancy concerning the most frequent nucleotide at the +4 site in yeast between this study and the reported consensus sequence of yeast (Cigan and Donahue, 1987; Hamilton et al., 1987) where the consensus contains +4U. In this compilation, it is +4A that is the most frequently used one. The explanation is that the previous compilations (Cigan and Donahue, 1987; Hamilton et al., 1987) included very few genes.

Given the claim of the translation initiation hypothesis that +4G is a translation initiation signal and important for translation initiation especially when –3R is absent (Kozak, 1986; Kozak, 1997), one would expect +4G to be overrepresented in genes without -3R. However, this expectation is not supported by the empirical data (Table

4.1). In human, mouse, zebrafish, fruitfly, and yeast, +4G is underrepresented in genes without -3R, contrary to the prediction of the translation initiation hypothesis. In nematode, +4G appears more in genes with -3Y than with -3R, but the difference is not significant. The *Arabidopsis* genome is the only one that shows significant overuse of +4G in genes without -3R (Table 4.1).

The advocates of the translation initiation hypothesis may argue that +4G should only be overrepresented in highly expressed genes. The reason for genes with -3Y not to have overuse of +4G may be due to many lowly expressed genes included in -3Y gene group. For this reason there is a need to compile data separately for highly efficiently translated and lowly efficiently translated genes and test whether +4G is overused in highly efficiently translated genes without -3R. Because there is not enough data of translation efficiency for the CDSs in most eukaryotic organisms, this work can not be done at this moment. Chapter 5 will take advantage of transcriptomic and proteomic data in yeast and explore the relationship between +4 nucleotide and translation efficiency in yeast.

4.4.2 Biased amino acid usage at P1' position

Because +4 nucleotide is at the first position of the codon coding for P1' amino acid, +4 nucleotide bias could be a reflection of P1' amino acid usage bias. In order to explore the likely association, this work investigated the amino acid usage bias at the P1' position within the seven species. The relative percent frequency of each amino acid used at the P1' site from all CDSs within each species was shown in Figure 4.1 and Table 4.2.

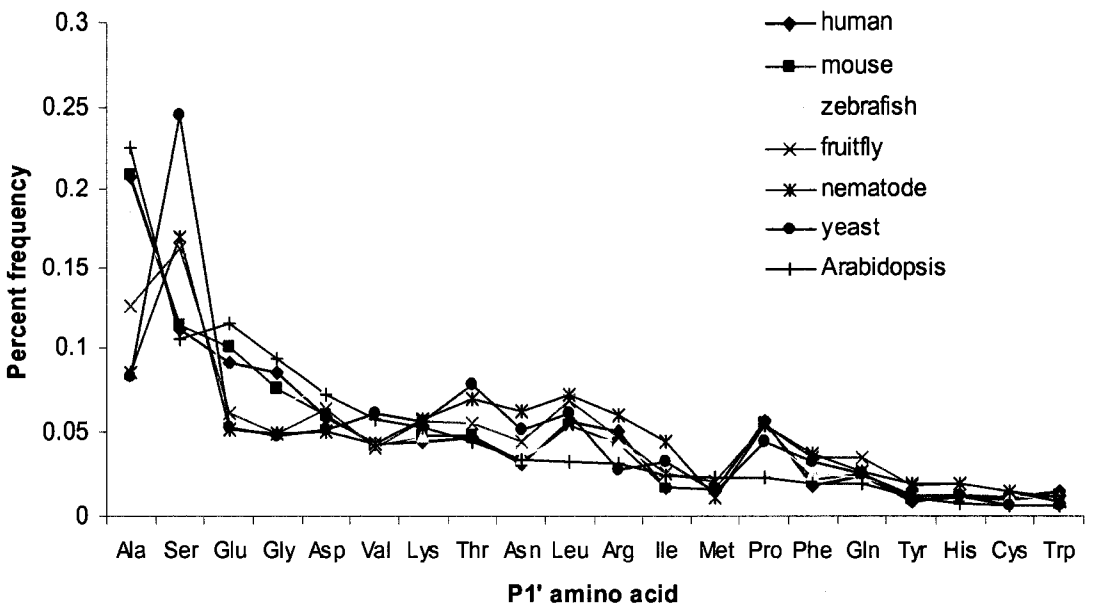


Figure 4. 1- The relative percent frequency of each amino acid used at the P1' site from all CDSs within each species

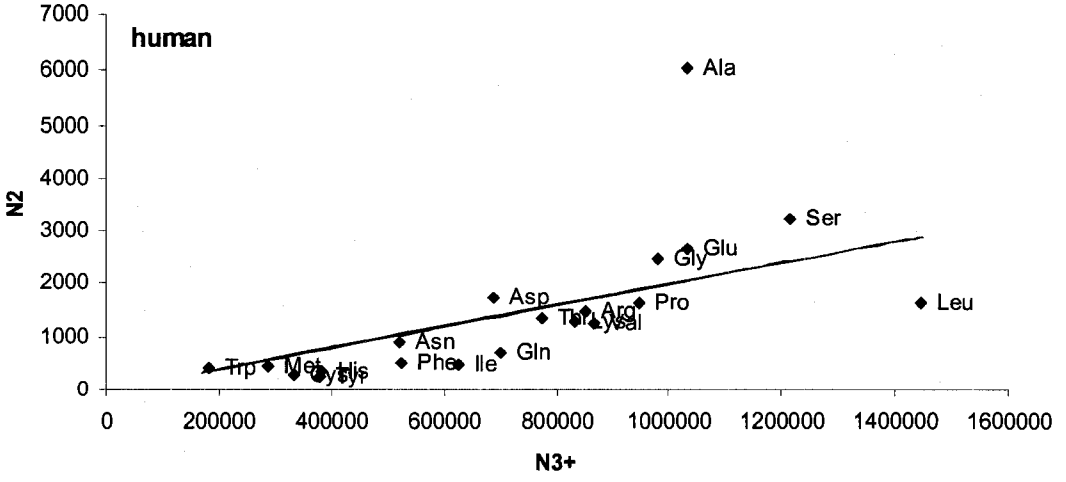
Table 4. 2 - The relative percent frequency of each amino acid used at the P1' site from all CDSs within each species

	human	mouse	zebrafish	fruitfly	nematode	yeast	*Arabidopsis*
Ala	0.207	0.208	0.184	0.126	0.085	0.082	0.225
Ser	0.111	0.114	0.133	0.161	0.168	0.245	0.105
Glu	0.091	0.100	0.083	0.061	0.052	0.053	0.115
Gly	0.085	0.075	0.066	0.050	0.049	0.048	0.093
Asp	0.059	0.060	0.075	0.064	0.050	0.051	0.072
Val	0.043	0.042	0.043	0.041	0.043	0.061	0.058
Lys	0.044	0.048	0.048	0.056	0.057	0.057	0.053
Thr	0.047	0.048	0.059	0.054	0.069	0.078	0.045
Asn	0.031	0.032	0.035	0.044	0.062	0.051	0.033
Leu	0.056	0.054	0.049	0.068	0.072	0.061	0.032
Arg	0.050	0.043	0.044	0.046	0.059	0.028	0.031
Ile	0.016	0.017	0.023	0.025	0.045	0.032	0.024
Met	0.016	0.016	0.015	0.021	0.011	0.014	0.023
Pro	0.056	0.056	0.047	0.053	0.054	0.045	0.023
Phe	0.018	0.021	0.025	0.035	0.037	0.033	0.019
Gln	0.024	0.025	0.023	0.034	0.026	0.025	0.019
Tyr	0.009	0.009	0.014	0.018	0.019	0.012	0.011
His	0.012	0.010	0.013	0.019	0.019	0.012	0.007
Cys	0.010	0.010	0.012	0.014	0.014	0.006	0.006
Trp	0.015	0.012	0.009	0.010	0.009	0.006	0.006
Sum	1.000	1.000	1.000	1.000	1.000	1.000	1.000

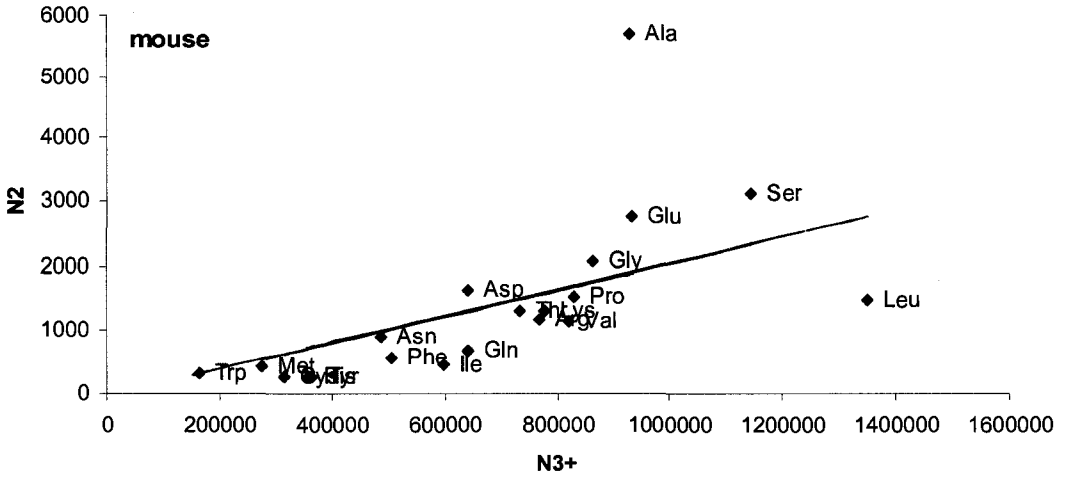
In general, Ala or Ser has the greatest frequency at the P1' site among the seven species (Figure 4.1 and Table 4.2), except in *Arabidopsis* where the frequency of Glu is a little higher than Ser. The frequency of Ala is the greatest in human, mouse, zebrafish, and *Arabidopsis*, but Ser is the most frequent amino acid at the P1' site in fruitfly, nematode, and yeast. Ala is over 20% in human, mouse and Arabidopsis, and Ser is over 20% in yeast.

To consider the effect of the proteome amino acid usage, the numbers of the amino acids at the other positions other than the initiator (Met) and P1' positions in the whole proteome in the seven species were computed. The observed frequencies of P1' amino acids against the expected from the other positions are plotted in Figure 4.2.

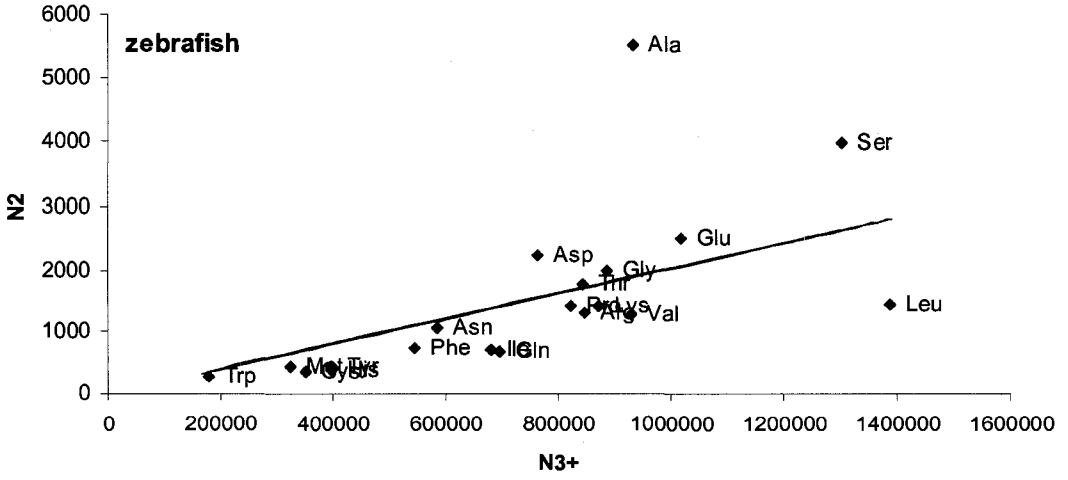
A.



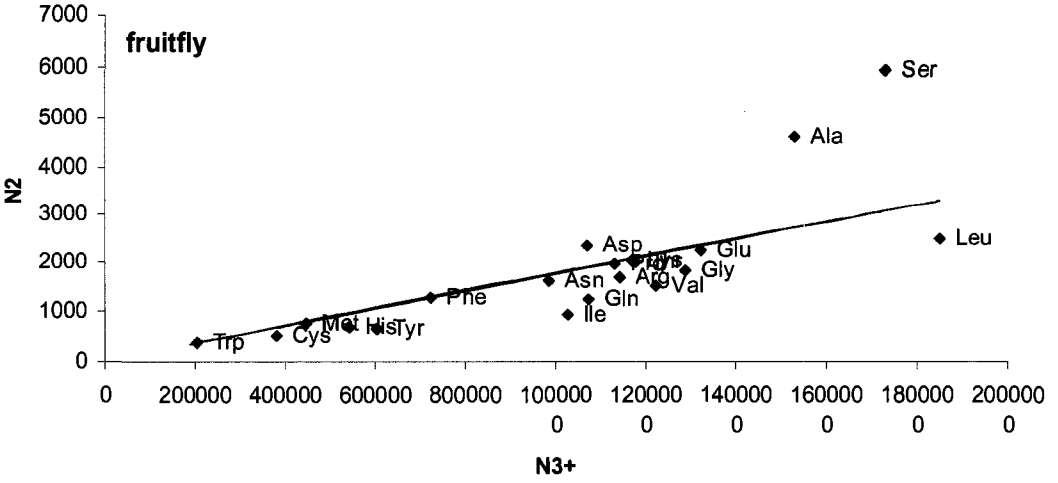
B.


mouse
N2
N3+
0
1000
2000
3000
4000
5000
6000
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000
Ala
Ser
Glu
Gly
Asp
Pro
Leu
Asn
Gln
Phe
Ile
Trp

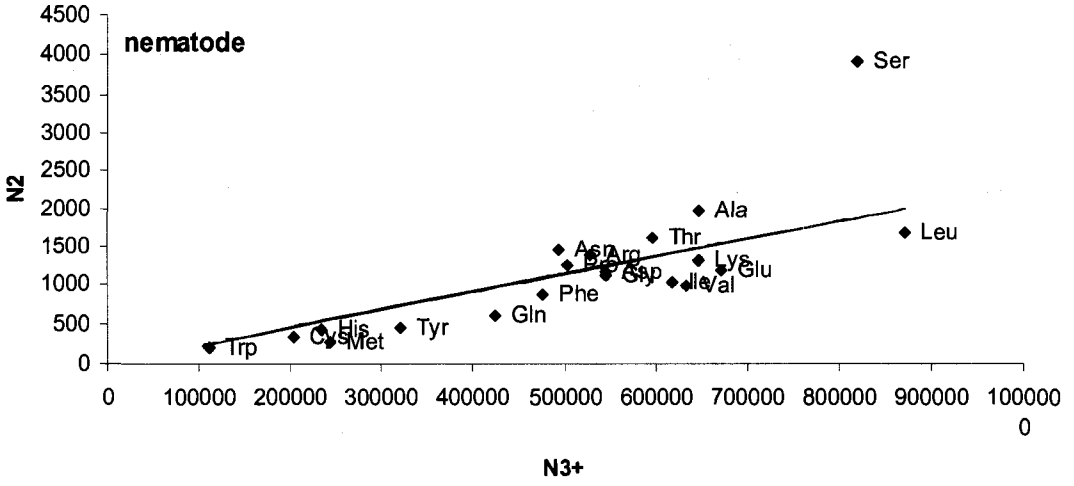
C.


zebrafish
N2
N3+
0
1000
2000
3000
4000
5000
6000
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000
Ala
Ser
Glu
Asp
Gly
Leu
Val
Asn
Phe
Trp

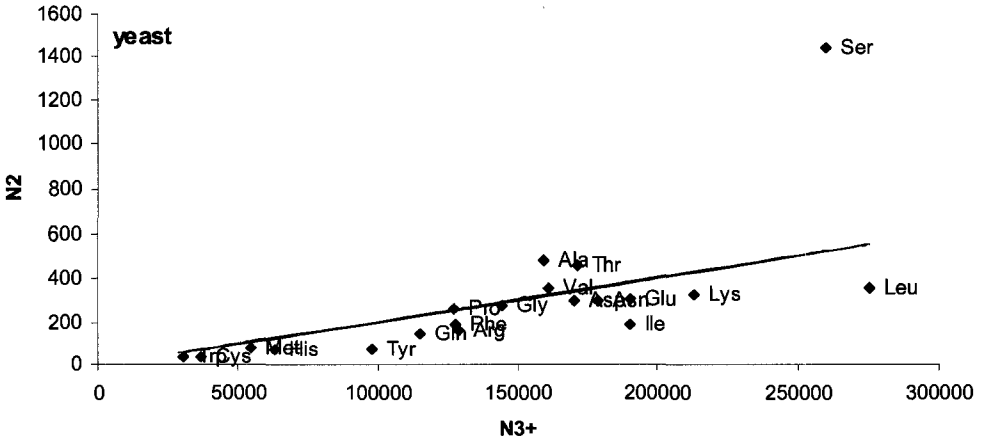
D.


fruitfly
N2
N3+
7000
6000
5000
4000
3000
2000
1000
0
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000
1800000
2000000
Ser
Ala
Leu
Asp
Glu
Gly
Asn
Val
Gln
Phe
Ile
Trp
Cys
Met
His
Tyr

E.


nematode
N2
N3+
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
Ser
Ala
Thr
Leu
Asn
Lys
Glu
Asp
Ile
Val
Phe
Gln
His
Tyr
Trp
Cys
Met

F.



G.

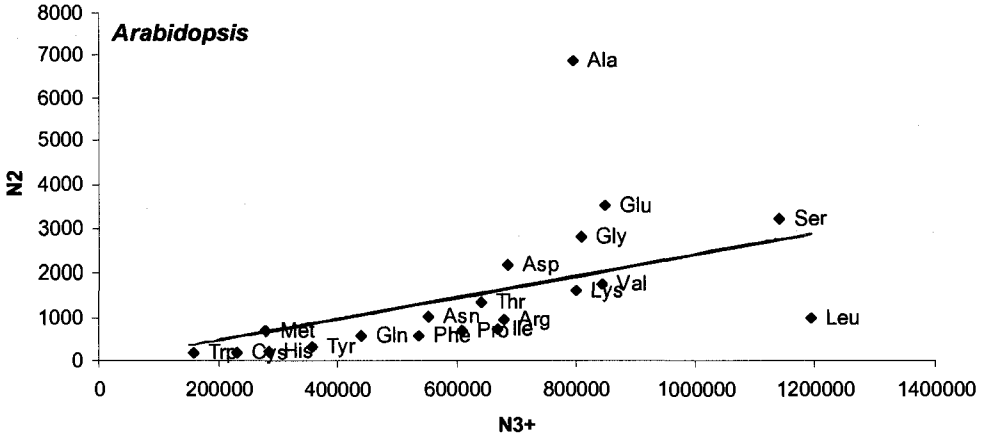


Figure 4. 2 - The frequencies of P1' amino acids (N2) relative to the frequencies of the same amino acids in the rest of the CDSs (N3+) excluding the first Met and P1' site

The pink line indicates the expected frequencies when the P1' amino acids have the same usage as the rest of the CDSs. The species name is labelled at the left top of each graph.

A notable pattern is shown in Figure 4.2. Ala and/or Ser are the two strongly biased used amino acids at the P1' position compared to the rest positions. In human, mouse, zebrafish and *Arabidopsis*, Ala is the amino acid whose usage remarkably deviates from the expected; in nematode and yeast, the frequencies of Ser remarkably departs from the expected. Fruitfly is an intermediate, having both Ala and Ser as the remarkably overused amino acids at the P1' position with the degree of deviations being less than the other organisms.

In order to clearly show the usage bias of Ala and Ser at the Pl' position, the differences between the observed and the expected P1' amino acid percent frequencies within each species were computed (Table 4.3).

Table 4. 3 - The difference between the observed and expected P1' amino acid percent frequencies within each species

"D" means the difference between the observed and expected percent frequencies. Positive values mean the observed frequencies are greater than the expected and negative values mean the opposite. The χ^2 and P values are under the null hypothesis that there is no difference between the observed and expected percent frequencies. All χ^2 values are associated with a degree of freedom of 1.

		human	mouse	zebrafish	fruitfly	nematode	yeast	*Arabidopsis*
	D	0.136	0.139	0.121	0.051	0.021	0.027	0.162
Ala	χ^2	8222.910	8286.722	7410.340	1408.333	172.247	82.843	13423.230
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D	0.028	0.029	0.045	0.078	0.087	0.156	0.015
Ser	χ^2	291.178	294.767	747.153	2877.874	2387.679	1732.696	79.904
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D	0.020	0.031	0.014	-0.003	-0.015	-0.013	0.048
Glu	χ^2	184.003	413.193	96.141	6.712	82.145	15.482	1092.883
	P	0.000	0.000	0.000	0.010	0.000	0.000	0.000
	D	0.017	0.011	0.006	-0.013	-0.005	-0.002	0.028
Gly	χ^2	141.333	57.141	18.336	103.076	12.141	0.598	410.591
	P	0.000	0.000	0.000	0.000	0.000	0.439	0.000
	D	0.012	0.012	0.023	0.012	-0.004	-0.008	0.018
Asp	χ^2	90.100	92.682	336.629	103.317	7.058	6.501	183.219
	P	0.000	0.000	0.000	0.000	0.008	0.011	0.000
	D	-0.016	-0.019	-0.020	-0.018	-0.020	0.006	-0.010
Val	χ^2	136.861	177.918	211.792	220.635	152.682	3.519	44.423
	P	0.000	0.000	0.000	0.000	0.000	0.061	0.000
	D	-0.013	-0.010	-0.011	-0.001	-0.007	-0.017	-0.011
Lys	χ^2	87.394	46.883	70.720	0.358	16.649	25.156	62.232
	P	0.000	0.000	0.000	0.550	0.000	0.000	0.000
	D	-0.006	-0.006	0.002	-0.003	0.010	0.019	-0.006
Thr	χ^2	24.211	22.508	1.710	4.955	42.642	37.162	24.991
	P	0.000	0.000	0.191	0.026	0.000	0.000	0.000
	D	-0.004	-0.004	-0.004	-0.004	0.014	-0.010	-0.011
Asn	χ^2	16.550	10.027	15.870	10.722	91.282	10.676	86.664
	P	0.000	0.002	0.000	0.001	0.000	0.001	0.000
	D	-0.043	-0.046	-0.046	-0.022	-0.014	-0.034	-0.063
Leu	χ^2	598.32	643.204	728.331	214.629	56.708	79.728	1415.994
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000

(This table is to be continued. Please turn to Page 56.)

(Table 4.3 – continued from Page 55)

		human	mouse	zebrafish	fruitfly	nematode	yeast	*Arabidopsis*
	D	-0.008	-0.014	-0.014	-0.009	0.007	-0.017	-0.022
Arg	χ^2	33.630	98.567	102.535	60.817	25.331	38.190	301.150
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D	-0.026	-0.027	-0.023	-0.025	-0.017	-0.034	-0.029
Ile	χ^2	495.403	481.424	352.327	473.370	110.853	107.540	513.223
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D	-0.004	-0.005	-0.007	-0.001	-0.013	-0.005	0.001
Met	χ^2	22.171	29.510	67.015	1.085	165.491	7.158	0.615
	P	0.000	0.000	0.000	0.298	0.000	0.007	0.433
	D	-0.009	-0.006	-0.008	-0.002	0.005	0.001	-0.026
Pro	χ^2	38.909	16.388	38.919	1.811	10.653	0.070	435.774
	P	0.000	0.000	0.000	0.178	0.001	0.792	0.000
	D	-0.018	-0.016	-0.012	0.000	-0.010	-0.012	-0.024
Phe	χ^2	280.288	199.691	121.008	0.017	48.161	18.647	421.079
	P	0.000	0.000	0.000	0.896	0.000	0.000	0.000
	D	-0.024	-0.023	-0.025	-0.018	-0.015	-0.015	-0.016
Gln	χ^2	357.986	313.606	402.957	239.018	138.249	33.518	242.646
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D	-0.017	-0.018	-0.013	-0.012	-0.013	-0.021	-0.018
Tyr	χ^2	345.018	319.205	181.377	180.692	122.919	81.647	342.155
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D	-0.014	-0.016	-0.014	-0.007	-0.005	-0.010	-0.016
His	χ^2	229.600	278.427	222.645	73.583	21.653	26.238	333.690
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D	-0.013	-0.013	-0.012	-0.005	-0.006	-0.006	-0.012
Cys	χ^2	206.900	212.091	177.949	42.607	41.441	19.661	247.527
	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D	0.002	0.000	-0.003	0.000	-0.002	-0.004	-0.007
Trp	χ^2	13.231	0.029	20.486	0.555	122.919	11.145	44.423
	P	0.000	0.865	0.000	0.456	0.000	0.001	0.000

It can be seen from Table 4.3, Ala shows a very large percent frequency difference in human (13.6%), mouse (13.9%), zebrafish (12.1%), and *Arabidopsis* (16.2%), while Ser shows a very large percent frequency difference in nematode (8.7%)

and yeast (15.6%). Both Ser and Ala show large differences in fruitfly. The difference of Ser is 7.8%, and of Ala is 5.1%.

The proportion of the proteins subject to NME is large as 55 ~ 70% (Giglione et al., 2004). The amino acid requirement for NME is similar in all organisms, small and uncharged amino acids are efficient for NME (Giglione et al., 2004; Meinnel et al., 1993). The amino acid constraint hypothesis proposes that the presence of +4G in vertebrate initiation consensus is a consequence of biased P1' amino acids due to the constraint of NME, so it predicts that not all amino acids coded by G-starting codons are overused at the P1' position, but only the small and uncharged amino acids, which are good candidates for NME, are overused at the P1' position. Among the five amino acids coded by G-starting codons, Ala (GCN), Gly (GGN), and Val (GUN) are small and uncharged amino acids, and Glu (GAR), Asp (GAY) are charged amino acids. Ser is also a small and uncharged amino acid and it is not coded by G-starting codons. So the pattern of the P1' amino acid usage in the seven eukaryotes are consistent with the amino acid constraint hypothesis. First, not all amino acids coded by G-starting codons are equally biased at the P1' position in vertebrates, only Ala is remarkably overused. Second, the overused P1' amino acids are not limited to the ones coded by G-starting codons. Ser is also remarkably overused at the P1' position in fruitfly, nematode, and yeast.

In human, mouse, zebrafish and *Arabidopsis*, Ala is the amino acid whose usage at the P1' position remarkably deviates from the expected, which could cause the prevalence of +4G. On the other hand, in nematode and yeast, the number of P1' Ser remarkably departs from the expected. This should be the reason that +4G is not the

most frequent nucleotide in these two species. The fruitfly has both Ala and Ser as the remarkably overused amino acids at P1' site with the degrees of deviations being less than other organisms, so the proportion of genes with +4G is less than human, mouse, zebrafish and Arabidopsis, but more than nematode and yeast. Ala and Ser seem to be two special amino acids used in the eukaryotic organisms. There is a trend that vertebrates tend to used Ala as the P1' amino acid, but invertebrates are more likely to have Ser as the P1' amino acid. This may be related to the different MAP activities among the eukaryotes. Why the other small and uncharged amino acids are not overused at the P1' position may be related to the MAP activities too.

It can be observed that, compared with human, mouse, and *Arabidopsis*, nucleotides A and U are used more often at the +4 site in yeast. Because Ser is encoded by both AGY and UCN codons, this observation can be explained by the overuse of Ser as the P1' amino acid in yeast (Figure 4.1 and Table 4.2). Yeast proteins with a Ser at their P1' sites can undergo efficient NME (Moerschell et al., 1990) and its overuse at the P1' site is consistent with the amino acid constraint hypothesis (Xia X, 2007). On the other hand, if +4 site is related to translation initiation and if one particular nucleotide can increase translation initiation, then one would expect that either +4A or +4U should be favoured but not both. If +4A increases translation initiation efficiency, then one should expect Ser to be coded more often by AGY at the P1' site than at other sites; if +4U increases translation initiation, then one should expect Ser to be coded more often by UCN codons at the P1' site than at other sites. The proportion of UCN codons is 73.75% at the P1' site and 72.63% at other sites, and the differences is not

significant ($\chi^2 = 0.891$, P = 0.35). Thus, there is no indication that the +4 site is a translation initiation signal, which contradicts the translation initiation hypothesis.

4.5 Conclusion

This work examined the relationship between the -3 and +4 sites and the amino acid usage at the P1' position in seven eukaryotic species. It shows that empirical data do not support that +4G is a translation initiation signal and important for translation initiation especially when –3R is absent. On the other hand, the amino acid usage at the P1' position is consistent with the amino acid constraint hypothesis that proposes +4G in Kozak consensus is a result of biased P1' amino acid at the P1' site due to the constraint of NME.

4.6 Acknowledgement

Part of this work comes from a course project. Downloading the genomes and extracting -3 to +6 nucleotides were completed by Malisa Carullo for mouse, Jan Mennigen for *Arabidopsis*, and Sam Khalouei, Pinchao Ma, and Ziyu Song for human.

Chapter 5 The nucleotide immediately following the start codon is not important for translation efficiency in *Saccharomyces cerevisiae*

5.1 Abstract

The previous studies used transcriptomic data to approximate the translation expression level and did not control for confounding factors related to translation efficiency when testing the two hypotheses regarding the role of G immediately following the start codon (+4G) in Kozak consensus. The work in this chapter took advantage of the transcriptomic and proteomic data in *Saccharomyces cerevisiae* to more accurately evaluate translation efficiency and applied a regression analysis to explore the relationship between +4 site and translation efficiency in yeast. After considering the confounding factors related to translation efficiency, such as the secondary structure of 5' untranslated region, -3 nucleotide, codon usage bias and coding sequence length, it was found that +4 site is not important for translation efficiency. The result from yeast does not support the translation initiation hypothesis.

5.2 Introduction

There are two hypotheses regarding +4G in Kozak consensus (RCCaugG), translation initiation hypothesis and amino acid constraint hypothesis (Kozak, 1986; Kozak, 1997; Xia, 2007a). The translation initiation hypothesis states that +4G is a translation initiation signal in vertebrates, while the amino acid constraint hypothesis proposes that the presence of +4G is a consequence of the biased P1' amino acid due to

the constraint of NME. The strong support for the translation initiation hypothesis was provided by the experiment conducted by Kozak (1997), but other observations casted doubt on this conclusion (Flinta et al., 1986; Xia, 2007a).

Flinta et al. (1986) observed that Ala, Gly and Val were the prevalent P1' amino acids in 700 eukaryotic protein sequences, so they proposed that +4G in Kozak consensus may not be related to the selection of the start codon, but simply caused by the P1' amino acids Ala, Gly, and Val all of which are coded by G-starting codons (Flinta et al., 1986). Then, Kozak conducted an experiment (1997), showing that +4G increased the recognition of AUG and alternative start codons. In this experiment, the effect of +4G on translation efficiency was not consistent. For example, the presence of +4G, when followed by U, in the mRNA did not increase translation initiation relative to the control mRNA without +4G. No similar experiment was done after that.

Compared to the strong conservation of -3R, +4G is not conserved in several major taxonomic groups. +4G is common among genes in human, mouse, zebrafish and *Arabidopsis*, but much less common in invertebrate species (see Chapter 4). The consensus nucleotide at +4 site is A in protozoa (Yamauchi, 1991), nematode and yeast (see Chapter 4). It is difficult to argue that +4 site is important for translation initiation because one then has to argue that the translation machinery makes use of +4G in vertebrates and plant, while +4A in nematode, protozoa and yeast. In addition, +4U is the consensus sequence in yeast highly expressed genes (Hamilton et al., 1987). This would induce another proposal that +4U can increase translation initiation efficiency in yeast.

When the aforementioned facts contradict the translation initiation hypothesis, they somehow favour the amino acid constraint hypothesis. First, the most often used amino acid at P1' position is Ala in human, mouse, zebrafish, and *Arabidopsis*, and Ser in fruitfly, nematode, and yeast (see Chapter 4). Second, Ala is coded by G-starting codons and Ser is coded by U-starting codons and A-starting codons. Third, both Ala and Ser are small amino acids which are good candidates for NME (Giglione et al., 2004).

In 2007, Xia first formally characterized the two hypotheses and tested them by using human data (2007a). Xia's results favoured the amino acid constraint hypothesis (see Chapter 2). But, there are some limitations in Xia's research. First, Xia used two sets of data to evaluate gene expression level. One was Serial Analysis of Gene Expression (SAGE) data, and the other was CAI data. The SAGE data is transcriptomic. Generally, the correlation between mRNA abundance and protein abundance is only moderate (Nie et al., 2006). Therefore, using transcriptomic data to approximate the translation expression level is problematic. While using CAI to predict gene expression is well established in bacterial species and vertebrate mitochondria, where there is selection pressure favouring codon-anticodon adaptation, it is very controversial in multicellular organisms, like human. Second, the observation that highly expressed genes were not more likely to have +4G than lowly expressed genes was taken as inconsistent with the predictions of the translation initiation hypothesis. However, the studied genes were not controlled for other sequence features influencing translation efficiency. The advocates of the translation initiation hypothesis could argue that the highly expressed genes may have higher proportion of -3A (since -3A is more important

than +4G) or higher elongation or termination efficiency, so they do not need to have a higher proportion of +4G to increase translation initiation.

This work took advantage of transcriptomic and proteomic data in *S. cerevisiae* to obtain large-scale mRNA and protein abundance data and applied multiple regression analysis to test the two hypotheses. MRNA and protein abundance data together can provide a more accurate evaluation of translation efficiency. Furthermore, when the multiple regression analysis was conducted, all the sequence features related to translation were taken into account. In addition to +4 nucleotide, these features include the secondary structure of initiation region, -3 nucleotide, +5 nucleotide, +6 nucleotide, codon usage bias, amino acid usage bias, sequence length, termination codon usage bias, and termination codon context. It is a more effective way to study the relationship between +4 nucleotide and translation efficiency while controlling for these confounding factors.

5.3 Materials and Methods

5.3.1 Data

5.3.1.1 Protein abundance data

The protein abundance data of *S. cerevisiae* were downloaded from http://www.nature.com/nature/journal/v425/n6959/extref/nature02046-s2.xls, which is the supplementary information accompanying (Ghaemmaghami et al., 2003). There are 6234 ORFs reported in this dataset, in which the measured protein levels are reported in

terms of molecules/cell. The ORFs flagged with "-", "%", or "#" were removed because protein abundance values can not be assigned to these ORFs. The final protein abundance dataset contains 3868 ORFs.

The global protein abundance data were obtained by using a nascent technology. Ghaemmaghami et al. (2003) created a *S. cerevisiae* fusion library where each ORF was tagged with a high-affinity epitope and expressed from its natural chromosomal location. A tandem affinity purification (TAP) tag was inserted immediately preceding the stop codon of each ORF (Figure 5.1), by which the pattern of protein expression were expected to be minimally disrupted. The TAP tag allowed the immunodetection and immunopurification of the entire yeast proteome using a single antibody.

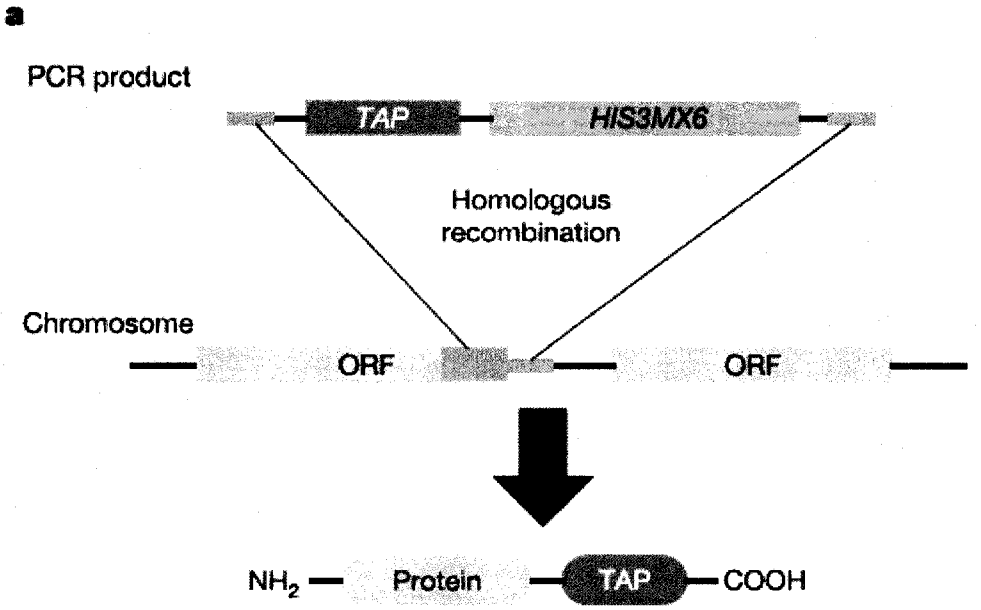


Figure 5. 1 - The schematic diagram of TAP tagging strategy in the method of Ghaemmaghami et al.

Adopted from Figure 1.a of (Ghaemmaghami et al., 2003).

The TAP tag is 558nt long, in which the last three nucleotides are "TGA". In addition to the TAP tag, the sequence of *Schizosaccharomyces pombe his5*$^{+}$ gene was inserted just right after "TGA" of the TAP tag and has a length of 1538nt. Therefore, the tagged ORFs have a common C-terminal sequence, stop codon, and downstream sequence.

5.3.1.2 mRNA abundance data

The mRNA abundance data of *S. cerevisiae* comes from http://web.wi.mit.edu/young/pub/data/orf_transcriptome.txt, which is the supplementary information accompanying (Holstege et al., 1998). The microarray technology was used to characterize genome-wide RNA abundance. The dataset includes 6179 yeast ORFs, of which 5460 ORFs have detectable mRNA abundance. The column under the heading "ExpressionLevel" in the dataset represents the levels of all detectable mRNA species in yeast in terms of molecules/cell.

5.3.1.3 Sequence features

This work aimed at studying the relationship between +4 nucleotide and translation efficiency while controlling for other confounding factors influencing translation efficiency which include the secondary structure of initiation region, -3 nucleotide, +5 nucleotide, +6 nucleotide, codon usage bias, amino acid usage bias, sequence length, termination codon usage bias, and termination codon context.

The genomic sequence files of *S. cerevisiae*, orf_genomic_1000_all.fasta.gz and orf_coding_all.fasta.gz, were downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/orf_dna dated Sep. 6, 2007.

This web site is a component of SGD. The file orf_genomic_1000_all.fasta.gz contains all ORF sequences (including dubious ORFs) with introns and untranslated region 1000 bp upstream of the initial ATG and 1000bp downstream of the stop codon, and the file orf_coding_all.fasta.gz has only the CDSs of all ORFs (including dubious ORFs), without 5'-UTR, 3'-UTR, intron sequences, or bases not translated due to translational frameshifting. DAMBE (Xia, 2000; Xia and Xie, 2001) was used to extract the upstream region and coding sequence of each ORF, calculate minimum folding energy (MFE) of the initiation region, and obtain the codon adaptation index (CAI) value, amino acid adaptation index (AAAI) value, and sequence length for each coding sequence.

The secondary structures of two different segments in the initiation region, the 5'-UTR and the immediate 3' side of the start codon, have a different effect on the recognition of the start codon. The secondary structure of the 5'-UTR would prevent the 40S ribosome subunit from advancing and decrease the initiation efficiency (Kozak, 2005); the secondary structure of the immediate 3' side of the start codon would cause 40S ribosome to stay longer at the start codon position and facilitate the recognition of the start codon (Kozak, 1990; Kozak, 2002). The MFE values of these two segments were calculated separately. As discussed in Chapter 3, the upstream 40nt (excluding intron sequences) is a critical region for translation initiation in yeast, so the MFE values of this region were computed to indicate the secondary structure of the 5'-UTR. Downstream +4 ~ +30 (total 27 nt) sites were used to calculate the MFE values of the 3' immediate side of start codon.

CAI is a measure of synonymous codon usage bias. It was first proposed in 1987 by Sharp and Li (Sharp and Li, 1987). CAI quantifies the codon usage bias in a gene by checking the codon usage of the gene against a reference set of highly expressed genes from a species. A relative adaptiveness is assigned to every codon by using the information contained in the reference set. The relative adaptiveness is calculated by comparing the frequency of that codon with the maximum frequency of its synonymous codons for the same amino acid (Sharp and Li, 1987). The reference set comes from a set of known highly expressed genes, such as ribosome genes. DAMBE has an improved implementation of CAI calculation over that in the European Molecular Biology Open Software Suite (EMBOSS) (Xia, 2007c). There are four reference sets available for yeast in DAMBE to calculate CAI, "Eyeast", "Eyeastcai", "Eysc", and "Eysc_h". "Eysc_h" was chosen as the reference set because it contains the codons used in highly expressed genes in *S. cerevisiae.*

The concept of Amino Acid Adaptation Index (AAAI) is same as CAI. It is a measure of amino acid usage bias (Xia, 2007b). AAAI quantifies the amino acid usage bias in one gene by using a reference set of highly expressed genes from a species to assess the relative merits of each amino acid of another gene in the same species or the host species. Again, "Eysc_h" was chosen as the reference set.

5.3.2 Data Editing

5.3.2.1 Matching the protein and mRNA abundance data

In the protein abundance dataset, some ORFs do not have corresponding mRNA abundance data. The two datasets were matched, so each ORF entering the following regression analysis has both protein and mRNA abundance values.

There was inconsistency in the ORF names among the protein abundance dataset, mRNA abundance dataset, and SGD sequence files. In the mRNA and protein abundance characterization, YAR044W was synonymous to YAR042W in the SGD files, so was YDR474C to YDR475C, YFR024C to YFR024C-A, YJL012C-A to YJL012C, YJL018W to YJL019W, YJL021C to YJL020C, YKL097W-A to YKL096W-A, and YPR090W to YPR089W. Some genes (YEL068C, YER084W, YHR173C, YIL054W, YJR146W, YLR358C, YMR290W-A, YNL140C, YNL143C, YNL184C, and YOR105W) were annotated in SGD as "dubious open reading frame unlikely to encode a protein", and were not annotated at all in the *S. cerevisiae* genome in NCBI. However, they were found to be expressed at both mRNA (Holstege et al., 1998) and protein levels (Ghaemmaghami et al., 2003). YFL006W and YFL007W had been merged into YFL007W, YJL017W and YJL016W into YJL016W, and YOR087W and YOR088W into YOR087W in the SGD. The synonymous pairs, "YAR044W and YAR042W", "YJL018W and YJL019W", "YJL021C and YJL020C", "YFR024C and YFR024C-A", had different mRNA abundance value or protein abundance value for both ORFs in the pairs. So did the merged pairs, "YFL006W and YFL007W", "YJL017W and YJL016W", and "YOR087W and YOR088W". Therefore, these ORFs

were discarded, because the abundance data could not be unambiguously assigned. After these operations, 3708 ORFs remained in the dataset.

5.3.2.2 Modifying ORF coding sequences

As mentioned in the data source section, the experiment for detecting protein abundance level tagged each expressed ORF at the C-terminus, so the coding sequences of the 3708 ORFs were different from the original sequences. The modified sequences should be used for the analysis in this work. To obtain the modified sequences, the stop codon was first removed from each coding sequence. Then, the sequence of the TAP tag was added to the left sequence. For example, suppose the original sequence is:

ATGGGCGTACGTTACCATGGAACGGCGGCATCACGTACGGCTTGCGTCACT
TAA

For simplicity, a very short ORF is presented here, in which "TAA" is the original stop codon. After modification, the sequence becomes as:

ATGGGCGTACGTTACCATGGAACGGCGGCATCACGTACGGCTTGCGTCACT
GGTCGACGGATCCCCGGGTTAATTAATCCATGGAAGAGAAGATGGAAAAAG
AATTTCATAGCCGTCTCAGCAGCCAACCGCTTTAAGAAAATCTCATCCTCCG
GGGCACTTGATTATGATATTCCAACTACTGCTAGCGAGAATTTGTATTTTCA
GGGAGAATTCGGCCTTGCGCAACACGATGAAGCCGTGGACAACAAATTCAA
CAAAGAACAACAAAACGCGTTCTATGAGATCTTACATTTACCTAACTTAAA
CGAAGAACAACGAAACGCCTTCATCCAAAGTTTAAAAGATGACCCAAGCCA
AAGCGCTAACCTTTTAGCAGAAGCTAAAAAGCTAAATGATGCTCAGGCGCC
GAAAGTAGACAACAAATTCAACAAAGAACAACAAAACGCGTTCTATGAGA

TCTTACATTTACCTAACTTAAACGAAGAACAACGAAACGCCTTCATCCAAA
GTTTAAAAGATGACCCAAGCCAAAGCGCTAACCTTTTAGCAGAAGCTAAAA
AGCTAAATGATGCTCAGGCGCCGAAAGTAGACGCGAATCATCAGTGA

The underlined bases are the nucleotides in the TAP tag. "TGA" is the current stop codon. The length of the tag is 558nt (including the stop codon), so each coding sequence after modification is 555nt longer than its original sequence. All sequence modifications were done by Perl scripts.

The promoter region is the major regulator of mRNA expression level. The TAP tag is added at the C-terminus of coding region, leaving the promoter region and the N-terminus of coding region untouched. The change of mRNA expression caused by this modification for a relatively long transcript (compared to the tag) can be negligible, but the effect of inserting 558nt into a short transcript is very difficult to evaluate, so only the ORFs whose coding sequences are not shorter than 2000nt were considered. The final dataset contains 939 ORFs.

5.3.2.3 Transforming protein abundance and mRNA abundance data

Some of the descriptive statistics of the original protein abundance and mRNA abundance values for the 939 ORFs are listed in Table 5.1.

Table 5. 1 - The descriptive statistics of the original protein abundance and mRNA abundance values for 939 ORFs

Molecular	Number	Min	1st Quartile	Median	3rd Quartile	Max	Mean
protein	939	41.1	477.3	1485.0	4002.0	1256000.0	10080.0
mRNA	939	0.1	0.4	0.7	1.2	43.1	1.257

Table 5.1 shows that the distributions of both protein and mRNA abundance are extremely positively skewed, so the data was $\log_{10}$-transformed. The same statistics of the transformed protein and mRNA abundance values are listed in Table 5.2. The distributions are shown in Figure 5.2. After log transformation, the protein and mRNA abundance are much closer to normal distribution, which were used in the following regression analysis.

Table 5. 2 - The descriptive statistics of the log transformed protein abundance and mRNA abundance values for 939 ORFs

Molecular	Number	Min	1st Quartile	Median	3rd Quartile	Max	Mean
protein	939	1.614	2.679	3.172	3.602	6.099	3.204
mRNA	939	-1.00000	-0.39790	-0.15490	0.07918	1.63400	-0.14400

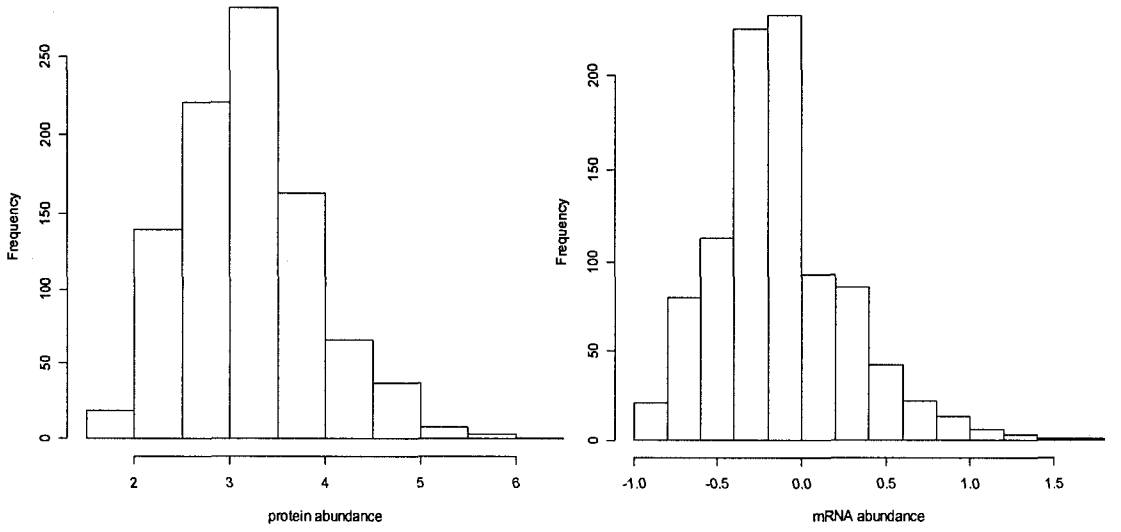


Figure 5. 2 - The distribution of the log transformed protein abundance and mRNA abundance values

5.3.3 Regression

5.3.3.1 Defining translation efficiency

This work aimed at studying the relationship between +4 nucleotide and translation efficiency while controlling for confounding factors influencing translation efficiency, so the translation efficiency is the response variable in the multiple regression model and all sequence features related to translation are the predictor variables.

Due to the nonlinear relationship between the protein and mRNA abundance data, a LOWESS (locally weighted scatterplot smoothing) regression was used to regress the protein abundance against the mRNA abundance. LOWESS fits simple models to localized subsets of the data to build up a function that describes the deterministic part of the variation in the data, point by point (Trexler and Travis, 1993).

The residuals from the LOWESS regression are the protein abundances of those ORFs after removing the effect of mRNA abundance level, so they were used to evaluate translation efficiency. The LOWESS regression was run in R with a span of 0.75.

5.3.3.2 Variables

"efficiency" was assigned as the variable name for translation efficiency, so were "upMFE", "downMFE", "minus3", "plus4", "plus5", "plus6", "CAI", "AAAI", "SeqLen" for the minimum folding energy of the 5'-UTR region (-40 to -1), the minimum folding energy of the immediate 3'-side of start codon (+4 to +30), -3 nucleotide, +4 nucleotide, +5 nucleotide, +6 nucleotide, codon adaptation index, amino acid adaptation index, and the sequence length, respectively. Among all the variables presented here, "MFE", "CAI", and "AAAI" are quantitative and they can directly enter into regression function, and "minus3", "plus4", "plus5", and "plus6" are categorical variables which need dummy coding (Neter et al., 1985).

When the values of "CAI", "AAAI", and "SeqLen" were calculated, the start codons and the stop codons were removed, so these three variables represent the elongation process. Because all ORFs have the same stop codon and stop codon

context, the sequence features for termination are the same. Therefore, this study does not assess the effect of termination on translation efficiency.

5.3.3.3 Regression model and function

The general regression model used in this work is:

$$\text{efficiency}=\beta_0+\beta_1\ \text{upMFE}+\beta_2\ \text{minus3}+\beta_3\ \text{downMFE}+\beta_4\ \text{plus4}+\beta_5\ \text{plus5}+\beta_6\ \text{plus6}+\beta_7\ \text{CAI}+\beta_8\ \text{AAAI}+\beta_9\ \text{SeqLen} \quad (5.1)$$

All the regression analyses were run in R (http://www.r-project.org/).

The highest absolute value of the correlation coefficients between any two of the continuous predictor variables is 0.1670, so the multicollinearity problem does not exist in the regression. The diagnostic plots (Figure 5.3) for the general model (5.1) show the assumptions of the multiple linear regression are not violated.

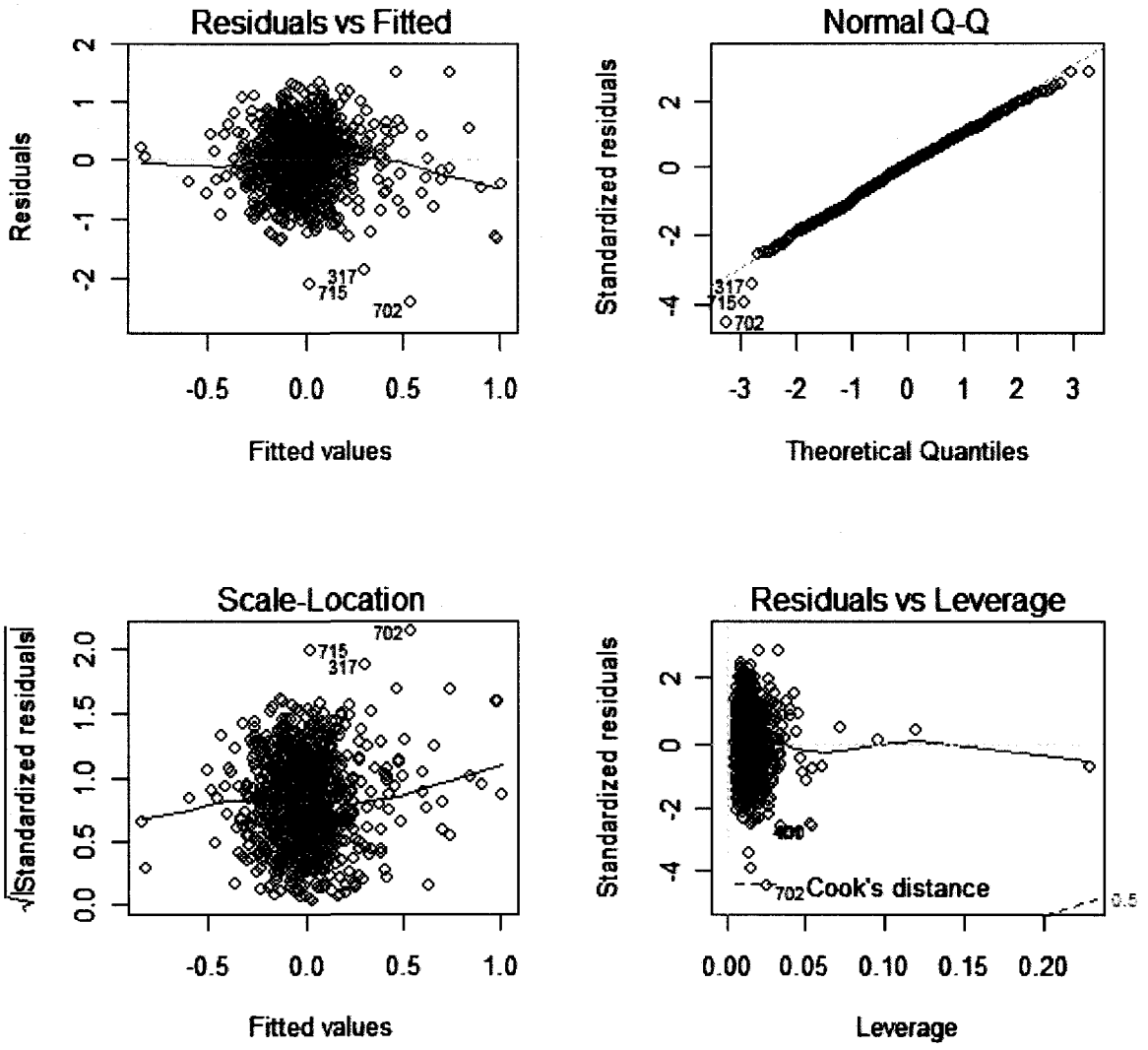


Figure 5. 3 - The diagnostic plots of the multiple linear regression

5.4 Results and Discussion

5.4.1 Regression analysis does not support the translation initiation hypothesis

Regression of the translation efficiency against the multiple mRNA sequence features (Based on Model 5.1) is significant, with multiple R^2 (determination coefficient) and adjusted R^2 being 0.1173 and 0.1170 respectively, which indicates that 11.7% of the translation efficiency variation can be explained by the multiple mRNA sequence features considered.

This work took data from yeast to test the two hypotheses regarding the role of +4G in Kozak consensus. Although +4G in Kozak consensus is for vertebrates, the essence of the two hypotheses regarding +4G in Kozak consensus is whether +4 nucleotide is important for translation initiation, so the relationship between +4 nucleotide and translation efficiency in yeast can help understand this issue.

A forward sequential regression was performed to explore the relationship between the +4 nucleotide and the translation efficiency. From the standpoint of a biological process, among the mRNA sequence features considered except codon usage bias, amino acid usage bias, and sequence length, the time order for them to occur in translation should be the secondary structure of 5'-UTR, -3 nucleotide, the secondary structure of immediate 3' side of start codon, +4 nucleotide, +5 nucleotide, +6 nucleotide. One can not distinguish the time order among codon usage bias, amino acid usage bias, and sequence length. The sequential regression started with "upMFE", and then added one variable at a time into the regression function in order of its occurring

time in translation. Table 5.3 lists the result of analysis of variance (ANOVA) for this sequential regression. F and P values shown in Table 5.3 are the values when all predictors were put into the regression. Table 5.4 lists the regression coefficient for each predictor when all predictors were put into the regression in the same order as in Table 5.3.

Table 5. 3 – The ANOVA result from the sequential regression of translation efficiency against the mRNA sequence features

In the "F value" column, the numbers in parentheses are the degrees of freedom associated with the corresponding F value. In the "P value" column, "*" indicates that the P value shows significance.

	Sum of Squares	Mean Square	F value	P value
upMFE	1.682	1.682	5.854 (1, 921)	0.016 *
minus3	3.821	1.274	4.433 (3, 921)	0.004 *
downMFE	0.028	0.028	0.096 (1, 921)	0.757
plus4	0.947	0.316	1.099 (3, 921)	0.349
plus5	1.764	0.588	2.047 (3, 921)	0.106
plus6	2.193	0.731	2.544 (3, 921)	0.055
CAI	16.961	16.961	59.022 (1, 921)	0.000 *
AAAI	0.243	0.243	0.846 (1, 921)	0.358
SeqLen	7.533	7.533	26.215 (1, 921)	0.000 *

Table 5. 4 – The regression coefficient for each predictor

"minus3C", "minus3G" and "minus3T" are dummy codes for "minus3". The same strategy was applied to "plus4", "plus5" and "plus6".

	Estimated Coefficient	Standard Error	t value	P value
(intercept)	3.71200	6.267	0.592	0.554
upMFE	0.01280	0.008	1.637	0.102
minus3				
minus3C	0.03131	0.061	0.515	0.607
minus3G	0.06042	0.047	1.294	0.196
minus3T	-0.12850	0.056	-2.290	0.022
downMFE	-0.00259	0.009	-0.284	0.777
plus4				
plus4C	0.00706	0.061	0.116	0.907
plus4G	-0.01189	0.046	-0.258	0.797
plus4T	0.07080	0.052	1.353	0.177
plus5				
plus5C	-0.03775	0.048	-0.787	0.432
plus5G	-0.11260	0.058	-1.942	0.052
plus5T	-0.04450	0.055	-0.807	0.420
plus6				
plus6C	-0.01206	0.053	-0.229	0.819
plus6G	0.13990	0.055	2.526	0.012
plus6T	-0.00884	0.046	-0.193	0.847
CAI	2.68100	0.335	8.005	0.000
AAAI	-5.52100	7.689	-0.718	0.473
SeqLen	-0.00008	0.000	-5.120	0.000

In terms of the occurring time, "CAI", "AAAI", and "SeqLen" can not be separated. All 6 possible orders were permutated. The sums of squares are similar for

them in all 6 orders. Only "CAI" and "SeqLen" have significant influence on the translation efficiency variation.

The result from the sequential regression (Table 5.3) shows that only "upMFE", "minus3", "CAI", and "SeqLen" are significant variables in explaining the variation of translation efficiency. It indicates that the secondary structure of the region -40 to -1, -3 nucleotide, codon usage bias, and sequence length are important factors in determining translation efficiency. The secondary structure of immediate 3' side of start codon, +4 nucleotide, +5 nucleotide, +6 nucleotide, and amino acid usage bias are not important determinants.

To further explore the relationship between +4 nucleotide and the translation efficiency by controlling for the other sequence features, another sequential regression was performed in which "plus4" was the last variable into the regression model (see Model 5.2).

$$\text{efficiency}=\beta_0+\beta_1\ \text{upMFE}+\beta_2\ \text{minus3}+\beta_3\ \text{downMFE}+\beta_4\ \text{plus5}+\beta_5\ \text{plus6}+\beta_6\ \text{CAI}+\beta_7\ \text{AAAI}+\beta_8\ \text{SeqLen}+\beta_9\ \text{plus4} \quad (5.2)$$

The ANOVA result from Model 5.2 (not shown) indicates "upMFE", "minus3", "plus6", "CAI", and "SeqLen" are significant variables in explaining the variation of translation efficiency. "plus4" is not significant (P=0.427). This result about the relationship between the +4 nucleotide and the translation efficiency does not support the translation initiation hypothesis, but is consistent with the conclusion from Xia's research (2007a). Moreover, this result is established after controlling for the confounding factors, so it is more reliable.

The key point of the translation initiation hypothesis and the amino acid constraint hypothesis is whether the nucleotide at +4 site is important for translation initiation. If +4 nucleotide is important for translation initiation, one can reason that there is some association between 40S ribosome subunit (including the rRNAs and the initiator $tRNA^{Met}$) and +4 nucleotide in mRNA to enhance translation initiation (recognition of the start codon). This association likely exists before the codon-anticodon interaction between 80S ribosome and mRNA at the P1' position in the elongation process. For example, from the standpoint of the translation initiation hypothesis, it has been suggested that AUGG in mRNA might form a 4 base pair interaction with CCAU in the anticodon loop of initiator $tRNA^{Met}$ (Kozak, 1986), but no experiment has substantiated it. Recently, Pisarev et al. (Pisarev et al., 2006) detected that +4G interacted with an exclusive location ($AA_{1818-1819}$) of 18S rRNA in a rabbit translation system, proposing the role of +4G is to stabilize the conformational changes which occurs in the ribosomal complex upon the first codon-anticodon base-pairing. On the other hand, if +4 nucleotide is not important for translation initiation, it is likely that there is no such association between 40S ribosome subunit and +4 nucleotide in mRNA. From this work, this statement is more likely and the +4 nucleotide bias is just a reflection of amino acid preference at P1' position.

Table 5.3 also shows that, compared with "plus4" and "plus5", "plus6" has a marginal significance level. The ANOVA result from Model 5.2 (not shown) indicates that "plus6" is significant in explaining the variation of translation efficiency. To further explore the relationship between +6 nucleotide and the translation efficiency by

controlling for the other sequence features, a regression with "plus6" as the last variable was performed (see Model 5.3).

$$\text{efficiency}=\beta_0+\beta_1\ \text{upMFE}+\beta_2\ \text{minus3}+\beta_3\ \text{downMFE}+\beta_4\ \text{plus4}+\beta_5\ \text{plus5}+\beta_6\ \text{CAI}+\beta_7\ \text{AAAI}+\beta_8\ \text{SeqLen}+\beta_9\ \text{plus6} \quad (5.3)$$

The same result was obtained as the one from Model 5.2. "upMFE", "minus3", "plus6", "CAI", and "SeqLen" are significant variables in explaining the variation of translation efficiency. +6 site does not subject to amino acid constraint and it can reflect the codon usage bias resulting from mutation and codon-anticodon adaptation. The most likely explanation should be codon-anticodon adaptation. The ribosome should move away from the initiation region as quickly as possible, and a codon at P1' position that matches the most abundant tRNA species clearly would allow the ribosome to move downstream quickly along the mRNA.

A stepwise regression with Model 5.1 at the beginning results in the following model

$$\text{efficiency}=\beta_0+\beta_1\ \text{upMFE}+\beta_2\ \text{minus3}+\beta_3\ \text{plus6}+\beta_4\ \text{CAI}+\beta_5\ \text{SeqLen} \quad (5.4)$$

The standard of selecting this model is AIC (Akaike Information Criterion) (Akaike, 1974; Burnham and Anderson, 2002). The multiple R^2 and adjusted R^2 are 0.1098 and 0.1011 ($P < 2.2e\text{-}16$), respectively, and the AIC value is -1161.12. In this model, "downMFE", "plus4", "plus5", and "AAAI" are cancelled out. This implies that, the secondary structure of the region -40 to -1, -3 site, +6 site, codon usage bias and

sequence length can better explain the translation efficiency. The +4 site is not important for explaining the translation efficiency, contradicting to the translation initiation hypothesis. +6 site is a factor influencing translation efficiency. The result from the stepwise selection is consistent with the above sequential regressions.

The change of +5 site comes with the change of P1' amino acid. The amino acid constraint hypothesis states that the presence of +4G in Kozak consensus is a consequence of biased P1' amino acid Ala which is coded by GCN codon and a good candidate for NME. There is possibility of more efficient NME brings more efficient translation. In Xia's research (2007a), Xia examined the percent frequencies of genes with GCN, GGN, GAN, GUN, and non-G-starting codons at P1' position in high-CAI and low-CAI groups. He found that the genes with GCN and GGN codons exhibited minor differences ($P > 0.05$) in their frequencies between high-CAI and low-CAI groups. Although CAI is not a reliable gene expression index in human, it implies highly expressed genes tend to use more Ala and Gly at the P1'site. To further investigate the relationship between P1' amino acid and translation efficiency, more analysis was performed. In the first round, "P1" was put into Model 5.4. "P1" is a categorical variable representing the amino acid identity.

$$\text{efficiency}=\beta_0+\beta_1\ \text{upMFE}+\beta_2\ \text{minus3}+\beta_3\ \text{plus6}+\beta_4\ \text{P1}+\beta_5\ \text{CAI}+\beta_6\ \text{SeqLen} \quad (5.5)$$

Then, ANOVA was applied. The result shows no significant difference between Model 5.5 and Model 5.4 ($P=0.0663$), indicating that P1' amino acid has no effect on the translation efficiency.

In the second round, considering the possibility of more efficient NME could cause more efficient translation and NME efficiency is related to the size of the P1' amino acid, another variable "radius" was introduced. The values of "radius" are the radius of gyration (listed in Table 5.5), which is an index of the relative size of the amino acids (Levitt, 1976).

Table 5. 5 - The radius of gyration of 20 amino acids

Adopted from (Levitt, 1976)

AA	Radius(Å)	AA	Radius(Å)	AA	Radius(Å)	AA	Radius(Å)
Gly	0	Pro	1.25	Ile	1.56	Phe	1.9
Ala	0.77	Val	1.29	Gln	1.75	Lys	2.08
Ser	1.08	Asp	1.43	Glu	1.77	Tyr	2.13
Cys	1.22	Asn	1.45	His	1.78	Trp	2.21
Thr	1.24	Leu	1.54	Met	1.8	Arg	2.38

"radius" was put it into Model 5.4.

$$\text{efficiency}=\beta_0+\beta_1\ \text{upMFE}+\beta_2\ \text{minus3}+\beta_3\ \text{plus6}+\beta_4\ \text{radius}+\beta_5\ \text{CAI}+\beta_6\ \text{SeqLen} \quad (5.6)$$

ANOVA shows no significant difference between Model 5.6 and Model 5.4 (P= 0.2214), either.

5.4.2 The necessity of using correct measure to evaluate translation efficiency

In the previous studies, due to technology limitations, inappropriate data were used to evaluate translation efficiency. Xia (2007a) used two sets of data to evaluate the

gene expression level, SAGE data and CAI data. Although the SAGE tags selected were found ubiquitously in human tissues, the gene expression level is a transcription level. In Nakagawa et al.'s study (2008), microarray data were used to evaluate translation initiation efficiency. These are not appropriate. Generally, the correlation between mRNA abundance and protein abundance is only moderate (Nie et al., 2006). Therefore, one can not use the transcriptomic data to approximate the translation expression level. While using CAI to predict gene expression is well established in bacterial species and vertebrate mitochondria where there is selection pressure favouring codon-anticodon adaptation, it is very controversial in multicellular organisms like human.

This work took advantage of transcriptomic and proteomic technology in *S. cerevisiae* to get relatively large-scale mRNA and protein abundance data. The residuals from the LOWESS regression of the protein abundance against the mRNA abundance can provide a more accurate evaluation of translation efficiency.

5.4.3 The necessity of excluding the confounding factors in studying the relationship between +4 site and translation efficiency

From this work, one can see that the most significant feature related to translation efficiency is codon usage bias, followed by the sequence length (Table 5.3). Therefore, when investigating the relationship of +4 site with translation efficiency, one can not overlook these two important factors related to elongation process. Because it is difficult to separate the overall translation efficiency and pure translation initiation efficiency, previous studies (Nakagawa et al., 2008; Xia, 2007a) often assume they are interchangeable. But it should be pointed out that this is not appropriate. One research in a bacterial species *Desulfovibrio vulgaris* studying how much variation in protein

production can be explained by sequence features related to translation shows that elongation accounts for more variation than initiation (Nie et al., 2006). Termination affects the translation efficiency too. In this work, termination efficiency is controlled by the modification of the ORF sequences. The observation that highly expressed genes are not more likely to have +4G than lowly expressed genes was taken as inconsistent with the predictions of the translation initiation hypothesis (Xia, 2007a). The advocates of the translation initiation hypothesis may argue that the highly expressed genes can have higher codon usage bias so they do not need to have a higher proportion of +4G to increase translation efficiency.

-3 site is another important confounding factor. -3 site is strongly biased to A (Kozak, 1987), and this base is very conservative across different eukaryotic taxonomic groups (Nakagawa et al., 2008), so it is considered to have an important function in translation initiation. The result of this work is consistent with the previous findings in -3 site. Therefore, -3 site also needs to be controlled when testing the two hypotheses.

Compared to the secondary structure of 5'-UTR, the secondary structure of immediate 3' side of start codon is less important for translation efficiency.

5.5 Conclusion

This work more accurately characterized translation efficiency and performed regression analysis to control some confounding factors in studying the relationship between +4 nucleotide and translation efficiency. The results show that +4 site is not

important for translation efficiency in *Saccharomyces cerevisiae,* which does not support the translation initiation hypothesis.

Chapter 6 Conclusion

Initiation is the first step of translation. It involves the interactions between ribosome and mRNA with the participation of translation related factors and energy produced from GTP and ATP hydrolyses. A key in the initiation is the correct recognition of the start codon, which plays an important role in the regulation of translation. This thesis focused on the initiation in eukaryotes and investigates the sequence features affecting translation initiation. It has been recognized that the 5'-UTR, the 3' side of start codon, the start codon identity and the start codon context are important players in translation regulation. This thesis explored the characteristics of 5'-UTR and the start codon context, attempted to solve some controversies related to the start codon context, and discussed relevant mechanisms that may explain the observed features. The two hypotheses regarding +4G in Kozak consensus, the translation initiation hypothesis and the amino acid constraint hypothesis, were tested in this thesis. All of these can deepen the current understanding of the translation initiation process.

Chapter 3 studied the 5'-UTRs from the verified nuclear genes in *Saccharomyces cerevisiae*. It has shown that, in yeast, the 40nt upstream of the start codon have significantly different nucleotide composition from the further upstream sequences and this is related to reducing the secondary structure near the start codon. The 40nt upstream of the start codon is a critical region of 5'-UTR for translation initiation in yeast.

Chapter 4 examined the relationship between -3 and +4 sites and the amino acid usage at the P1' position in seven eukaryotic species. It shows that empirical data do not

support that +4G is a translation initiation signal especially when –3R is absent. On the other hand, the amino acid usage at the P1' position is consistent with the amino acid constraint hypothesis that proposes +4G in Kozak consensus is a result of biased P1' amino acid at the P1' site due to the constraint of NME.

Chapter 5 took advantage of the transcriptomic and proteomic data in *S. cerevisiae* to evaluate translation efficiency and demonstrated the benefit of applying regression in analyzing the relationship between +4 site and translation efficiency. After considering the confounding factors in translation efficiency, it was found that there is no correlation between +4 site and translation efficiency. The results do not support the translation initiation hypothesis.

This thesis lends new support for one of the hypotheses explaining the presence of +4G in Kozak consensus, the amino acid constraint hypothesis. The essence of the two hypotheses is whether the nucleotide at +4 site is important for translation initiation. This thesis may give a correct direction to study translation initiation. Kozak's 1997 experiment directly detected abundance of 80S ribosome initiation complex to evaluate the translation initiation efficiency. This should be the most reliable method. But unfortunately, no similar experiment has been done after that. This thesis calls for the necessity to redo this experiment and do similar experiment in yeast translation system.

The new findings from this thesis will contribute to the correct understanding of eukaryotic translation initiation, which is related to accurately setting up the reading frame (Kozak, 1999; Xia, 1998). The correct reading frame is the base of an accurate

translation of a protein, which is done by the selection of the start codon, so the study of how the start codon is selected and what factors function is very important. A good understanding of +4G will contribute to accurately interpreting the experimental results.

Because the translation initiation step is a very rate-limiting procedure, it is directly related to the gene expression level. A review suggested links between the translation initiation and human diseases (Kozak, 2002). Understanding the translation initiation process may shed light on how to treat these human diseases.

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control *19*, 716-723.

Akashi, H., and Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci USA *99*, 3695-3700.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular biology of the cell, 4th edn (New York and Abingdon, Garland Science).

Boeck, R., and Kolakofsky, D. (1994). Positions +5 and +6 can be major determinants of the efficiency of non-AUG initiation codons for protein synthesis. EMBO Journal *13*, 3608-3617.

Brown, C. M., Stockwell, P. A., Trotman, C. N. A., and Tate, W. P. (1990a). Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. Nucleic Acids Research *18*, 6339-6345.

Brown, C. M., Stockwell, P. A., Trotman, C. N. A., and Tate, W. P. (1990b). The signal for the termination of protein synthesis in procaryotes. Nucleic Acids Research *18*, 2079-2086.

Brown, T. A. (2007). Genomes three (New York and London, Garland Science Publishing).

Burnham, K. P., and Anderson, D. R. (2002). Model selection and multimodel inference: a practical-theoretic approach, 2nd edn (New York, Springer-Verlag).

Cavener, D. R. (1987). Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. Nucleic Acids Research *15*, 1353-1361.

Cavener, D. R., and Ray, S. C. (1991). Eukaryotic start and stop translation sites. Nucleic Acids Research *19*, 3185-3192.

Chen, S.-J., Lin, G., Chang, K.-J., Yeh, L.-S., and Wang, C.-C. (2007). Translational efficiency of a non-AUG initiation codon is significantly affected by its sequence context in yeast. The Journal of Biological Chemistry.

Cigan, A. M., and Donahue, T. F. (1987). Sequence and structural features associated with translational initiator regions in yeast - a review. Gene *59*, 1-18.

Cigan, A. M., Pabich, E. K., and Donahue, T. F. (1988). Mutational analysis of the *HIS4* translational initiator region in *Saccharomyces cerevisiae*. Molecular and Cellular Biology *8*, 2964-2975.

Curran, J. F. (1998). Modified nucleotides in translation (Herndon, ASM Press).

Donahue, T. F., and Cigan, A. M. (1988). Genetic selection for mutations that reduce or abolish ribosomal recognition of the *HIS4* translational initiator region. Molecular and Cellular Biology *8*, 2955-2963.

Dowdy, S., Wearden, S., and Chilko, D. (2004). Statistics for research, 3rd edn (Hoboken, John Wiley & Sons, Inc.).

Dresios, J., Chappell, S. A., Zhou, W., and Mauro, V. P. (2006). An mRNA-rRNA base-pairing mechanism for translation initiation in eukaryotes. Nature structural & molecular biology *13*, 30-34.

Farazi, T. A., Waksman, G., and Gordon, J. I. (2001). The biology and enzymology of protein N-myristoylation. J Biol Chem *276*, 39501-39504.

Flinta, C., Persson, B., Jornvall, H., and von Heijne, G. (1986). Sequence determinants of cytosolic N-terminal protein processing. Eur J Biochem *154*, 193-196.

Frolova, L. Y., Merkulova, T. I., and Kisselev, L. L. (2000). Translation termination in eukaryotes: Polypeptide release factor eRF1 is composed of functionally and structurally distinct domains. RNA *6*, 381-390.

Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C., and Meinnel, T. (2006). The proteomics of N-terminal methionine cleavage. Mol Cell Proteomics *5*, 2336-2349.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737-714.

Giglione, C., Boularot, A., and Meinnel, T. (2004). Protein N-terminal methionine excision. Cellular and Molecular Life Sciences *16*, 1455-1474.

Grunert, S., and Jackson, R. J. (1994). The immediate downstream codon strongly influences the efficiency of utilization of eukaryotic translation initiation codons. The EMBO Journal *13*, 3618-3630.

Hamilton, R., Watanabe, C. K., and Boer, H. A. d. (1987). Compilation and comparison of the sequence context around the AUG start codons in *Saccharomyces cerevisiae* mRNAs. Nucleic Acids Research *15*, 3581-3593.

Hirel, P.-H., Schmitter, J.-M., Dessen, P., Fayat, G., and Blanquet, S. (1989). Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. Proc Natl Acad Sci USA *86*, 8247-8251.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. Nucleic Acids Research *31*, 3429-3431.

Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717-728.

Jackson, R. J., and Kaminski, A. (1995). Internal initiation of translation in eukaryotes: the picornavirus paradigm and beyond. RNA *1*, 985-1000.

Jia, M., and Li, Y. (2005). The relationship among gene expression, folding free energy and codon usage bias in *Escherichia coli*. FEBS Letters *579*, 5333-5337.

Joshi, C. P., Zhou, H., Huang, X., and Chiang, V. L. (1997). Context sequences of translation initiation codon in plants. Plant Molecular Biology *35*, 993-1001.

Kochetov, A. V. (2005). AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context. Bioinformatics *21*, 837-840.

Kozak, M. (1978a). How do eucaryotic ribosomes select initiation regions in messenger RNA? Cell *15*, 1109-1123.

Kozak, M. (1978b). Identification of features in 5' terminal fragments from reovirus mRNA which are important for ribosome binding. Cell *13*, 201-212.

Kozak, M. (1980). Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. Cell *22*, 7-8.

Kozak, M. (1981). Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucleic Acids Research *9*, 5233-5252.

Kozak, M. (1984a). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Research *12*, 857-872.

Kozak, M. (1984b). Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin *in vivo*. Nature *308*, 241-246.

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. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Research *15*, 8125-8148.

Kozak, M. (1990). Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proc Natl Acad Sci USA *87*, 8301-8305.

Kozak, M. (1991). Structural features in eukaryotic mRNAs that modulate the initiation of translation. J Biol Chem *266*, 19867-19870.

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 Journal *16*, 2482-2492.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. Gene *234*, 187-208.

Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. Gene *299*, 1-34.

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene *361*, 13-37.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol *104*, 59-107.

Li, X., and Chang, Y.-H. (1995). Amino-terminal protein processing in Saccharomyces cerevisiae is an essential function that requires two distinct methionine aminopeptidases. Proc Natl Acad Sci USA *92*, 12357-12361.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, S. L., and Darnell, J. (2004). Molecular cell biology, 5th edn (New York, W. H. Freeman and Company).

Looman, A. C., and Kuivenhoven, J. A. (1993). Influence of the three nucleotides upstream of the initiation codon on expression of the *Escherichia coli lacZ* gene in *Saccharomyces cerevisiae*. Nucleic Acids Research *21*, 4268-4271.

Looman, A. C., Laude, M., and Stahl, U. (1991). Influence of the codon following the initiation codon on the expression of the *lacZ* gene in *Saccharomyces cerevisiae*. Yeast *7*, 157-165.

Lukaszewicz, M., Feuermann, M., Jerouville, B., Stas, A., and Boutry, M. (2000). In vivo evaluation of the context sequence of the translation initiation codon in plants. Plant Science *154*, 89-98.

Lynch, M., Scofield, D. G., and Hong, X. (2005). The evolution of transcription-initiation sites. Mol Biol and Evol *22*, 1137-1146.

Maicas, E., Shago, M., and Friesen, J. D. (1990). Translation of the *Saccharomyces cerevisiae tcm1* gene in the absence of a 5'-untranslated leader. Nucleic Acids Research *18*, 5823-5828.

Meinnel, T., Mechulam, Y., and Blanquet, S. (1993). Methionine as translation start signal: a review of the enzymes of the pathway in *Escherichia coli*. Biochimie *75*, 1061-1075.

Moerschell, R. P., Hosokawa, Y., Tsunasawa, S., and Sherman, F. (1990). The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine *in vivo*. The Journal of Biological Chemistry *265*, 19638-19643.

Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H., and Miura, K.-i. (2008). Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Research *36*, 861-871.

Neter, J., Wasserman, W., and Kutner, M. H. (1985). Applied linear statistical models: regression, analysis of variance, and experimental designs, 2nd edn (Homewood, Richard D. Irwin, Inc.).

Nie, L., Wu, G., and Zhang, W. (2006). Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis. Genetics *174*, 2229-2243.

Niimura, Y., Terabe, M., Gojobori, T., and Miura, K.-i. (2003). Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. Nucleic Acids Research *31*, 5195-5201.

Pesole, G., Gissi, C., Grillo, G., Licciulli, F., Liuni, S., and Saccone, C. (2000). Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. Gene *261*, 85-91.

Pisarev, A. V., Kolupaeva, V. G., Pisareva, V. P., Merrick, W. C., Hellen, C. U. T., and Pestova, T. V. (2006). Specific functional interactions of nucleotides at key -3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. Genes & Development *20*, 624-636.

Prescott, L. M., Harley, J. P., and Klein, D. A. (2005). Microbiology, 6th edn (New York, The McGraw-Hill Companies, Inc.).

Rocha, E. P. C., Danchin, A., and Viari, A. (1999). Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Research *27*, 3567-3576.

Ross, S., Giglione, C., Pierre, M., Espagne, C., and Meinnel, T. (2005). Functional and developmental impact of cytosolic protein N-terminal methionine excision in Arabidopsis. Plant Physiology *137*, 623-637.

Sargan, D. R., Gregory, S. P., and Butterworth, P. H. W. (1982). A possible novel interaction between the 3'-end of 18S ribosomal RNA and the 5'-leader sequence of many eukaryotic messenger RNAs. FEBS Letters *147*, 133-136.

Sawant, S. V., Kiran, K., Singh, P. K., and Tuli, R. (2001). Sequence architecture downstream of the initiator codon enhances gene expression and protein stability in plants. Plant Physiology *126*, 1630-1636.

Shabalina, S. A., Ogurtsov, A. Y., Rogozin, I. B., Koonin, E. V., and Lipman, D. J. (2004). Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucleic Acids Research *32*, 1774-1782.

Sharp, P. M., and Li, W.-H. (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research *15*, 1281-1295.

Shine, J., and Dalgarno, L. (1974). The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci USA *71*, 1342-1346.

Shine, J., and Dalgarno, L. (1975a). Determinant of cistron specificity in bacterial ribosomes. Nature *254*, 34-38.

Shine, J., and Dalgarno, L. (1975b). Terminal-sequence analysis of bacterial ribosomal RNA: correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. Eur J Biochem *57*, 221-230.

Slusher, L. B., Gillman, E. C., Martin, N. C., and Hopper, A. K. (1991). mRNA leader length and initiation codon context determine alternative AUG selection for the yeast gene *MOD5*. Proc Natl Acad Sci USA *88*, 9789-9793.

The R project. (http://www.r-project.org/).

Trexler, J. C., and Travis, J. (1993). Nontraditional regression analysis. Ecology *74*, 1629-1637.

van den Heuvel, J. J., Planta, R. J., and Raue, H. A. (1989). Effect of deletions in the 5'-noncoding region on the translational efficiency of phosphoglycerate kinase mRNA in yeast. Gene *79*, 83-95.

Wilson, D. N., and Nierhaus, K. H. (2006). The E-site story: the importance of maintaining two tRNAs on the ribosome during protein synthesis. Cell Mol Life Sci *63*, 2725-2737.

Xia, X. (1998). How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? Genetics *149*, 37-44.

Xia, X. (2000). 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, Springer Science+Business Media).

Xia, X. (2007c). An improved implementation of codon adaptation index. Evolutionary Bioinformatics *3*, 53-58.

Xia, X., and Xie, Z. (2001). DAMBE: Software package for data analysis in molecular biology and evolution. Journal of Heredity *92*, 371-373.

Yamauchi, K. (1991). The sequence flanking translational initiation site in protozoa. Nucleic Acids Research *19*, 2715-2720.

Zar, J. H. (1984). Biostatistical analysis, 2 edn (Englewood Cliffs, Prentice-Hall, Inc.).

Zhang, Z., and Dietrich, F. S. (2005). Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE. Nucleic Acids Research *33*, 2838-2851.

Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research *9*, 133-148.