The Repertoire of N-termini in the *Saccharomyces cerevisiae* Proteome: An Integrative Study

By

Claire T. Fournier

Faculty Advisor: Dr. Michael P. Weir

A dissertation submitted to the Faculty of Wesleyan University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Middletown, Connecticut May 2013
Chapter Summary

Chapter 1: General Introduction
1.0 Introduction 2
1.1 Eukaryotic protein translation 2
1.2 Recoding 6
1.3 Genome annotation 13
1.4 Peptide mass spectrometry 14
1.5 Review of non-standard protein translation 17

Chapter 2: Analysis of Upstream Protein Translation Initiation
2.0 Introduction 21
2.2 Results 30
2.3 Conclusions and future directions 31
2.4 Methods 32
2.5 Figures 39

Chapter 3: Amino Termini of Many Yeast Proteins Map to Downstream Start Codons
3.0 Introduction 64
3.1 Results and conclusions 66
3.2 Methods 74
3.3 Figures 80

Chapter 3S: Partial Purification of TAP Tagged dnAUG Candidates
3S.0 Introduction 93
3S.1 Results 94
3S.2 Conclusions and future directions 95
3S.3 Methods 96
3S.4 Figures 97

Chapter 4: Gel Slice Approach for Algorithm Assessment
4.0 Introduction 110
4.2 Results 115
4.3 Conclusions and future directions 118
4.4 Methods 119
4.5 Figures 123

Chapter 5: Conclusions and Future Directions 134
Acknowledgements

I would like to acknowledge Dr. Michael Weir for his work towards the completion of my thesis and thank him for all that I have learned during my time in his lab. He contributed a great deal of time and effort when it came to my thesis work and the supportive lab environment that he has created has allowed me to grow as a scientist. He has encouraged me to think creatively about scientific problems and as a result his guidance has been invaluable. I have been fortunate to know Dr. Weir in the lab and in the classroom, and I have learned so much from him.

I am also very grateful to many other members of the Weir lab. Laurel F. Appel was a wonderful guide to the world of genetics and I am very appreciative for her patience and guidance. Michelle Orsulak was an amazing mentor who helped me to learn many new techniques. I have been fortunate to work with many talented graduate students including Justin Cherny, Kris Truncali, Miin Lin, Adam Robbins-Pianka, and Will Gladstone. I am very grateful for all that I have learned from them and all of the help they have given me.

I have learned a great deal from the interactions I have had with the members of the Wesleyan Biology and MB&B departments. This includes the members of the Cell & Development journal club who have provided me with interesting discussions over the years. I would also like to thank the members of my thesis committee, Dr. Laura Grabel, Dr. Fred Cohan, and Dr. Scott Holmes for all of their work.

I would like to thank my husband Robert as he has helped me in so many ways during my time at Wesleyan and has been my strongest supporter. I would also
like to thank my parents, Bernard and Nancy Pelletier, and my sister Caroline for their support. My family has always been a great source of encouragement and humor. Finally I would like to thank my daughter Elizabeth Vera for making every day more memorable.
Abstract

A detailed and evolving description of the Saccharomyces cerevisiae proteome is crucial for understanding many aspects of this model organism. Growing evidence suggests that some genes have functional translation initiation start sites upstream or downstream of the annotated start codon. This means a fraction of genes are likely under-annotated.

Ribosome profile data has identified 80 S translating ribosomes in the 5’ UTR of some genes and peptides have been identified that map to protein sequence preceding the annotated methionine of certain genes. To assess upstream initiation in three genes, candidates were chosen if they had identified upstream peptides, conservation across species in the 5’ UTR, or putative low functioning annotated AUG translation start sites. Mutagenesis of the annotated AUG was performed to test the functionality of the upstream site. Western blots revealed that protein was produced from the mutant candidates, strongly suggesting that an upstream site is able to initiate translation.

To assess downstream initiation, a large-scale approach was taken to identify N-terminal peptides through peptide mass spectrometry on glutaraldehyde-treated cell lysates as well as a parallel assessment of publicly deposited spectra. A large fraction of identified N-terminal peptides (35%) mapped to translation initiation at AUG codons found downstream of the annAUG. A fraction of downPeptides were confirmed after partial purification of the C-terminally tagged candidate genes while analysis of ribosome profile data and sequence context strongly indicates that
identified downstream AUG sites are functional. This screen suggests that downstream initiation is more common than previously thought and may contribute to protein function and localization.

Mass spectrometry followed by spectra analysis with peptide identification algorithms is a powerful tool in the assessment of the budding yeast proteome. To assess the algorithms Sequest and OMSSA and their ability to make accurate peptide matches, the algorithms were used to analyze mass spectra from proteins of a known molecular mass range without their prior knowledge of that range. Of the peptides identified, the majority of identifications were from the correct size range (84.4%-88.8%). This suggests that although different algorithms identify non-overlapping peptides, the non-overlapping set contains real matches. This assessment provides confidence in the algorithms, including their ability to accurately identify peptides from non-standard protein translation event.
Chapter 1
General Introduction
1.0 Introduction

With the sequencing of the *Saccharomyces cerevisiae* genome, extensive work has brought about the annotation of nearly 6,000 genes. The annotated start of an open reading frame (ORF), the presumed site of translation initiation, is crucial in defining the N-terminus of the resulting protein. A comprehensive description of translation initiation start sites greatly contributes to an accurate annotation of an organism’s proteome. This is of importance for understanding protein function, localization, and regulation. In addition, identification of protein translation initiation start sites yields valuable information about how the ribosome interacts with mRNA during translation. This thesis will discuss successful efforts to identify cases of non-standard translation initiation using bioinformatic, molecular genetic, and proteomic means, and in doing so analyze the canonical model of translation initiation.

1.1 Eukaryotic Protein Translation

The process of protein synthesis in eukaryotes begins with an AUG within a mRNA molecule. The AUG sequence is recognized by ribosomes as the start of protein translation. The identification of a translation start site is crucial to the identity of a protein as the N-terminus may contribute to the functionality of the protein. Once initiation occurs the polypeptide chain continues to elongate until a stop codon is reached and the ribosome disassociates from the mRNA. The process by which the AUG sequence is selected for initiation is called the scanning model. This widely accepted model is as follows:
I. The 40 S subunit of the ribosome will bind the ternary complex composed of Met-tRNA, eukaryotic elongation factor 2 (eIF2), and GTP forming the 43 S subunit. This is stabilized by eIF1A and eIF3 (Hinnebusch 1997, Dever 2002).

II. The 43 S subunit begins scanning at the 5’ end of the mRNA as eIF4E mediates the interaction with the m7G cap (Marcotrigiano 1997). eIF4 and eIF3 allow the RNA secondary structure to be unwound allowing the 43 S subunit to bind at the 5’ end of the mRNA. The 43 S subunit moves 5’to 3’ as it seeks to recognize an AUG (Dever 2002) as eIF1 and eIF1A allow the subunit to maintain an open conformation when scanning non-AUG codons (Passmore 2007).

III. The AUG is recognized when the anticodon binds the Met-tRNA in association with eIF1, eIF1A, eIF2 and eIF5, forming the 48 S subunit.

IV. After successful recognition, the GTP bound to eIF2 is hydrolyzed to GDP leaving the Met-tRNA in the P site of the ribosome as many of the initiation factors dissociate from the subunit. In association with eIF5B and GTP, the 60 S subunit is joined to the 43 S forming the 80 S complex and translation will continue with methionine as the first amino acid.

The scanning model has been supported by a number of experiments showing that the 40 S subunit (convention allows 40 S to describe subunits in stages 40 S, 43 S, and 48 S) scans linearly across the mRNA to find the first AUG where it will
initiate. There are three main concepts in this model. The first is that the 40 S subunit is the subunit that scans, independent of the 60 S subunit. Experiments demonstrate that when AUG initiation is chemically inhibited there is an accumulation of 40 S subunits bound to mRNA. Once the inhibition is released, the increase in 40 S subunits disappears, suggesting that the inhibition of initiation caused the 40 S subunits to scan but not initiate (Kozak 1987, 2002).

The second important idea in the model is that the ribosome scans linearly in one direction, instead of jumping to the initiation AUG. To support this idea, experiments show that when stable mRNA secondary structure is added between the m7G cap and the AUG codon, translation is aborted and the aborted complexes measure at about 40 S. This experiment provides evidence that if the 40 S subunit is unable to reach an AUG to initiate it must mean that the ribosome must pass through that sequence suggesting the 40 S subunit is unable to reach an AUG without scanning (Kozak 1987, 2002). Since the ribosome is only able to move forward it cannot survey the entire mRNA and then return to a start site. The model suggests that a ribosome will frequently choose the first AUG if it has acceptable sequence context.

Experiments have shown that if an AUG is introduced before the original AUG, the introduced AUG will be chosen (Cigan 1987). This is even the case when both AUGs are surrounded by the same nucleotide sequence context. In addition if the Met-tRNA anticodon sequence is mutated to 3’-UCC-5’, the ribosome would initiate at the first AGG in the mRNA sequence (Donahue 1988).
The scanning model also proposes that there is pressure to maintain certain nucleotides that surround functional start codons, known as the Kozak sequence. The typical flanking sequence in eukaryotes is GCCRCCaugG. Position -3 (AUG being +1, +2,+3), a purine (often A), is thought to be the most crucial and when mutated strongly discourages initiation (Kozak 1986). The +4 position, often G, is also of importance for initiation (Kozak 1987). If these two nucleotides are accurately represented then there is a high likelihood of initiation in a first position AUG.

The scanning model proposes that the AUG closest to the 5’ end is chosen as the AUG to begin the open reading frame (ORF). In yeast this is the case for most mRNAs. However some mechanisms exist where the AUG used to initiate may not be the first. An AUG found very close to the 5’ end of the mRNA can cause a downstream AUG to be used (Peterson 1986, Kelley 1982). Or, if the AUG lies in poor sequence context (mainly if position -3 is a C or U, or position +4 is not G) this can cause the small subunit to scan past to a downstream AUG.

Translation predominantly follows the scanning model but when the canonical decoding is over-ridden, the resulting events are referred to as recoding events (Baranov 2001). These instances can be programmed or spontaneous but they refer to cases where translation does not adhere to the scanning model. In addition, mRNA production during transcription, as well as mRNA sequence elements, can also have an impact on non-canonical translation. These events can contribute to models where more than one AUG is used in translation. Translation of the yeast gene GCN4 is often used as a model for a description of recoding events.
1.2.1 Recoding: Translation of \textit{GCN4}

Gcn4p is a transcriptional activator of more than 30 genes involved in biosynthesis of amino acids. In fed, unstressed cells the Gcn4p is expressed at low levels but once cells are starved for amino acids the repression is released and the expression increases. The repression is regulated by four small upstream ORFs (uORF) (labeled 1, closest to 5’, through 4) found before the main annotated ORF (annORF) (Grant 1994). If the start codons of the four uORFs are removed, Gcn4p is abundant (Hinnebusch 1997) and therefore production of Gcn4p is thought to be regulated by the uORFs. In both fed and starved conditions the majority of ribosomes will initiate at uORF1. In fed conditions the ribosomes will terminate at the stop codon of ORF1 and reinitiate at uORF4 far more frequently than the \textit{GCN4} annotated ORF, thus keeping the levels low. However in starved conditions \textit{GCN2} is activated, causing phosphorylation of the alpha subunit of eIF2. This prevents eIF2 from associating with GTP and Met-tRNA, which are needed to bind with the 40 S for initiation. This reduction in initiation complexes means that the ribosome will have to scan longer to become competent and will bypass uORF2, uORF3, and uORF4 to initiate at the annAUG of the main \textit{GCN4} ORF (Grant 1994).

1.2.2 Recoding: Reinitiation

Reinitiation occurs when a 80 S scanning ribosome dissociates after translation and the 40 S subunit continues scanning and initiates at a downstream site. For this to occur, the 40 S subunit needs to acquire another Met-tRNA, GTP, and eIF2 complex before the ribosome can reinitiate.
One contributing factor to whether or not reinitiation can occur is the distance between the end of the uORF and the annORF or dnORF. In order for a ribosome to reinitiate it needs to be competent with eIF2, GTP, and Met-tRNA and after translation of a uORF it is lacking this complex. Therefore it needs a minimum distance simply to rebuild the 40 S subunit. As an example, translation of \textit{GCN4} is suppressed because ribosomes initiate at uORF1 and then reinitiate at uORF4. Reinitiation at uORF4 prevents initiation at the main \textit{GCN4} ORF. Translation of uORF4 has been shown to suppress translation of \textit{GCN4} more successfully than uORF1 and uORF2 because it is found 200 nucleotides closer to the start of \textit{GCN4}’s initiation codon (Abastado 1991, Hinnebusch 1997). Even with the abundance of initiation complex factors in fed conditions, time is still necessary for the 40 S to become competent. In starved conditions a competent 40 S takes even longer (between uORF4 and \textit{GCN4}), and therefore \textit{GCN4} is translated. However in experiments when the distance between uORF1 and uORF4 is increased, translation of \textit{GCN4} is decreased and translation of uORF4 is increased (Grant 1994). This confirms that distance between ORFs is crucial for reinitiation and can have a regulatory role.

Another important factor in reinitiation after translation of a uORF is the context of the stop codon. In the example of \textit{GCN4}, the uORF4 stop codon was found to be a more effective stop codon than others. Although both uORF 1 and uORF4 have stop codons, they do not seem to be equal in terms of their ability to suppress downstream reinitiation. In experiments where the uORF1 stop codon was replaced
with the stop codon of uORF4 as well as 10 nucleotides 3’ of the codon, the ribosome’s ability to reinitiate at the downstream \textit{GCN4} ORF was decreased 10 fold (Abastado 1991).

Finally, reinitiation is also dependent on the length of the uORF itself as reinitiation occurs with few exceptions only after translation of short ORFs. Kozak has shown that reinitiation can reliably occur after a uORF of 10 to 12 codons although an exact sequence length is not known as it will be influenced by secondary structure and elongation time. Kozak does discuss an experiment where a uORF was extended from 13 codons to 33 codons and this manipulation suppressed reinitiation (Kozak 1991). Short uORFs may be more permissive of reinitiation because necessary factors may remain bound after a brief bout of translation versus a longer elongation for a larger uORF.

1.2.3 Recoding: Leaky Scanning

As mentioned in Section 1.1 there is an optimal sequence surrounding an AUG for it to be considered a functional translation start site. It is GCCRCCaugG (with A in position -3 in budding yeast). This is shown in experiments from Kozak (1986) that showed strong sequence context such as ANNaugN or GNNaugG, correlate to consistent selection of an AUG. However, when there are changes to the consensus sequence in the form of no purines in position -3 and no G in position +4, a weaker start site is produced. These are the cases where the recoding event referred to as leaky scanning is seen. This means some ribosomes will initiate at the AUG closest to the 5’end of the mRNA, while others will continue scanning and initiate at
an AUG further downstream. Leaky scanning is not only reserved for cases where an
AUG is poor. It also is seen in genes with intermediate-strength translation start sites
and may also be influenced by secondary structure in some cases. An AUG with
moderate to poor context can initiate translation with the addition of secondary
structure 13-15 nucleotides downstream. The ribosomal pause created by the
secondary structure is thought to be enough to increase the likelihood that the AUG
will be chosen (Kozak 1990). Kozak has shown that if the sequence context is
improved at the first AUG of an mRNA, leaky scanning is reduced. Leaky scanning is
also seen in cases where an annAUG is not in poor context but is seen close to the 5’

A question that is relevant here is how much is leaky scanning an adaptive
mechanism to allow for variable N-termini and modulation of protein levels and how
much is it based on poor efficiency of the ribosome. For example, certain viruses use
leaky scanning to create proteins with different N-termini (Bridgen 2001, Chen 2001)
suggesting that this translational mechanism can be used to alter the function of a
protein depending on its initiation site. This can be seen in the production of c-Myc.
c-Myc, involved in many cellular processes, has been found to have a small isoform,
c-Myc S in addition to the two larger proteins c-Myc 1 and c-Myc 2. c-Myc S
localizes to the nucleus and is often highly expressed in some hematopoietic tumor
cell lines. c-Myc S is generated through leaky scanning shown by experiments that
demonstrate that when the two 5’ AUG’s sequence context is improved, production
of c-myc S is dramatically reduced. When cells are in rapid growth phase the
shortened protein is expressed at similarly high levels to the longer form. However once growth slows the levels drop in relation to the longer forms (Spotts 1997).

Conversely different N-termini, resulting from leaky scanning, may be the result of a lack of pressure to maintain one start site if a protein’s function is not dependent on the N-terminal region. In this case a second AUG codon could provide redundancy in some genes.

1.2.4 Recoding: Frameshift

This recoding event occurs when a translating ribosome moves one nucleotide in either the 5’ direction (-1) or the 3’ direction (+1) (Brierly 1995, Pande 1995). This shift in reading frame does not stop the ribosome from translating but does alter the rest of the protein sequence produced as it is now in a new frame. In the majority of cases it is crucial that the frame be maintained during elongation for a functional protein. However programmed ribosomal frameshifts are used to increase the product possible from an mRNA. Ribosomes are induced to frameshift by mRNA elements found to be ‘slippery’ meaning they induce bound tRNAs to pair non-wobbles bases to out of frame codons. The slippery site (X XXY YYZ, X = any nucleotide, Y + A or U, and Z = A, U, C) is often followed by secondary structure within 8 nucleotides that causes the ribosome to pause. Therefore the sequence is then read as XXX YYY Z in the case of -1 frameshifts (Harger 2002). Frameshift is common especially in viruses for the generation of fusion proteins but is seen in eukaryotes as well. It can be used to bypass stop codons such as in the case of mammalian antizyme, a decarboxylase inhibitor. Translation initiates at an out of frame codon but before the
stop codon UGA is reached, a +1 frameshift causes the ribosome to moved from UCC UGA to UCC GAU. This allows the ribosome to continue elongation to code for antizyme (Hayashi 1996). Predicted slippery sites have been found in several budding yeast genes (Hammell 1999).

1.2.5 Truncated mRNA

In some organisms such as viruses and in some cases eukaryotes, it is possible to have dicistronic and polycistronic mRNAs. In one isoform of the mRNA an AUG is used for initiation. However a downstream AUG may be used in cases where a splicing event causes the first AUG to be removed. In addition, an internal promoter may allow for a truncated mRNA to be produced (Kozak 2002). These three examples could result in three different proteins. The GCN4 transcript is not considered polycistronic as the uORFs do not produce distinct functioning proteins, but instead regulate translation of the main ORF (Kaufman 1987, Blumenthal 1998).

1.2.6 Internal Ribosome Entry Sites (IRES)

Certain viral mRNAs use Internal Ribosome Entry Sites (IRES) to recruit the 40 S subunit of the ribosome directly to an initiation site, without the need for the subunit to bind at the 5’ end and scan. Several viral mRNAs have characteristics that would make AUG recognition via the scanning model very inefficient such as stable secondary structure, upstream non-conserved AUG codons, long 5’ UTRs, and uncapped mRNAs.

When identifying IRESs, Pelletier et al. (1998) working with Poliovirus showed that the mutation of several upstream AUGs had no effect on translation.
efficiency. In Picornaviruses translation was not effected if eIF4F was sequestered even though this factor is needed to mediate interactions with the 5’ cap and its absence would inhibit translation within the scanning model (Nomoto 1977). The IRES sequences of both the EMCV and poliovirus, when inserted in between two cistrons, promoted translation of the downstream cistron even when translation of the upstream sequences was prevented (Jang 1988, Pelletier 1988). There does not appear to be one consensus sequence that describes IRESs but instead it varies by organism. It is unknown if the 40 S subunit binds directly to the IRES sequence or the surrounding region.

There are great differences between viruses and their requirements for IRES initiation. EMCV IRESs require canonical eIFs to recruit the 40 S while the HCV family of IRESs only requires the ternary complex for tRNA positioning. In further contrast CrPV IRESs require no eIFs to assemble the 48 S or even the 80 S. In addition IRESs use different IRES trans-acting factors (ITAFs). These factors are necessary for efficient site identification and recruitment (Hellen 2001). Although more commonly described in viruses, there are known genes in budding yeast with putative IRES sites: URE2, HAP4, and YAP1 (Komar 2003, Seino 2005).

1.2.7 Non-AUG Initiation

In certain circumstances ribosomes can initiate at non-AUG (noncanonical) start sites. Unlike IRES, the ribosome scans from the 5’ end but instead selects a codon other than AUG for initiation. The most common noncanonical sites vary from AUG by one base (Peabody 1989, Kozak 1989, Dasso 1989) and it is thought that a
strong consensus sequence (Kozak 1997, Chen 2008) and downstream secondary structure (Kozak 1990) contribute to this occurrence. Translation initiation factors appear to have a role in non-AUG selection also as eIF1 mutants are more permissive in non-AUG codon selection (Cheung 2007).

This introductory discussion of protein translation serves as a framework for considering the process of canonical and noncanonical AUG selection and initiation. With this we will now move to the concept of how the translation start sites in yeast genes were annotated which is important if we argue that some genes may be incorrectly under annotated.

1.3 *Saccharomyces cerevisiae* Genome Annotation

The budding yeast genome was one of the first genomes to be sequenced (Goffeau 1996). This work identified 6,275 ORFs that theoretically could produce protein greater than 99 amino acids and of those defined 5,885 ORFs that are likely functional. The genome appears to be more compact than other eukaryotes with 70% of sequence corresponding to an ORF (Dujon 1996). Only about 4% of identified genes were found to contain introns, a number that is smaller than other yeast species (Goffeau 1996).

To further refine the initial sequencing effort, Kellis et al. (2003) used sequence conservation as a heuristic for ORF definition. DNA conservation across three yeast species from the *Saccharomyces sensu stricto* group was analyzed with the thinking that there is selective pressure to conserve sequence if it is found within a functional ORF. Species *S. paradoxus, S. mikatae, and S. bayanus* were used because
they are similar enough that there are highly similar orthologues between species; however the species are different enough that areas lacking homology will be distinguishable. Each ORF’s sequence was compared to all three of the other species to identify homologous sequences.

Kellis et al. (2003) began with 6,025 ORFs that could encode a protein greater than 99 amino acids and did not physically overlap another ORF by more than half. Then, ORFs with a certain threshold of conservation (after comparisons with the three yeast species) were deemed biologically meaningful while those below the score were biologically meaningless. 5,538 genes were validated as functional ORFs. The majority of these annotated genes have one identified start site.

In the work presented in this document the AUG that is identified for a gene as the start of protein translation is referred to as the annotated AUG (annAUG). AUGs found upstream in the 5′UTR will be called upAUGs and AUGs downstream in the ORF are dnAUGs. To identify cases of upAUG and dnAUG initiation, peptide mass spectrometry was used to identify N-terminal peptides that result from these non-standard initiation events.

1.4.1 Tandem Mass Spectrometry: Overview

Mass spectrometry has influenced the study of proteomics in great ways and tandem mass spectrometry (MS/MS), also called peptide mass spectrometry, is of special interest to the field of proteomics. Mass spectrometry has the ability to inform the user about a protein’s presence, sequence, and relative abundance and that is
invaluable for understanding the proteome of a tissue or organism. A mass spectrometer consists of three components:

I. The ion source takes peptides and ionizes them so they are measured by their mass and charge ($m/z$)

II. The mass analyzer measures the $m/z$ ratio of each ion.

III. The detector measures and records the abundance of ions at each $m/z$ value

For Chapters 3 and 4 an LCQ DECA XP (Thermo) was used for analysis of both whole-cell lysates and partially purified candidate proteins. Both protocols used the same general methodology:

I. Protein is isolated from yeast cells

II. Isolated protein is then digested with the enzyme trypsin, which will cleave at the C-terminal end of lysines and arginines.

II. The peptides, in liquid, are bound to an immobile resin and eluted off into the mass spectrometer based on their hydrophobicity during liquid chromatography.

III. Once entered into the machine the $m/z$ value is recorded for the ionized peptides (MS).

IV. The most abundant ions are then further fragmented (collision induced dissociation) along the peptide backbone with an inert gas (MS/MS). The most common ions produced are b and y ions. These ions are formed based on where along the peptide bond the fragmentation occurs. The $m/z$ values are recorded as spectra, which are interpreted by peptide
identification algorithms to identify the proteins in a sample (Aebersold 2003).

1.4.2 Tandem Mass Spectrometry: Peptide Identification Algorithms

Identification of peptides from large-scale sets of mass spectra requires the use of peptide identification algorithms. There are two main types of peptide identification algorithms; *de novo* searching and database searching. The two algorithms discussed below are the latter and require the input of a protein sequence database to identify peptides. Generally they function by correlating the experimental spectra with theoretical spectra (generated from the database). Parameter files allow the user to specify parameters such as mass tolerances, modifications to amino acids, and the protease used to cleave the proteins. What parameters are specified and which databases are searched are important factors for accurate identifications.

The algorithms used in this work are Sequest and Open Mass Spectra Search Algorithm (OMSSA). The algorithm Sequest (Thermo) converts protein sequences into theoretical spectra. These differ from experimental spectra because there is no measure of abundance (peak height) in the theoretical set (Eng 1994). The peptide identifications are made when the mass of an experimental spectrum matches to a theoretical spectrum, whose mass falls within a user specified mass tolerance. Sequest provides a number of scores to assess the validity of a match including the probability the match is random and the rank of a peptide match. In Chapter 3, Sequest was run through the program Bioworks 3.2. This platform was upgraded to Proteome Discoverer 1.2 to run Sequest and was used for the analysis in Chapter 4.
Another peptide search algorithm used is OMSSA (Geer 2004), which also requires a database of protein sequences called a sequence search library. OMSSA first takes experimental mass spectra and extracts the m/z values and compares those values to the theoretical spectra m/z values. The measure of match quality, E-value, is the number of random hits that would have an equal or better score than the identified match. A low E-value therefore represents a higher confidence match. OMSSA runs on a cluster and therefore there is a significant reduction in computation time. For the work in this Chapter 4 OMSSA was run with five different parameter settings called parameter sets.

1.5 Review of Non-Standard Translation Initiation Evidence

In addition to GCN4, genes with multiple start sites have been described in eukaryotes including human, mouse, and budding yeast. Human basic fibroblast growth factor (bFGF) has been shown to have two start sites (AUG and CUG) that result in alternate protein N-termini. The longer isoform of the protein is transported to the nucleus while the truncated version resides in the cytoplasm (Bugler 1990, Prats 1989). In mice, the β-1,4-galactosyltransferase gene has two possible start sites that dictate whether the protein localizes to either the Golgi or the plasma membrane (Shaper 1988). The localization of the budding yeast protein Glr1p to either the mitochondria or the cytoplasm is also dependent on translation initiation start site selection (Outten 2004). Finally, a downstream transcription start site results in a truncated mRNA containing a CUC codon that can support translation initiation in the human WT1 gene (Bruening 1996).
The supplemental annotation of these genes is the result of work to describe a protein’s function, localization, or the presence of multiple bands after detection. Conversely, more recent analyses of mRNA molecules have described AUG codons surrounding the annotated start site in an effort to understand how the ribosome selects one or more start codons. UpAUGs have been observed in the 5’UTRs of many genes, and are especially present in genes containing an annAUG with poor sequence context (Rozogin 2001). It is especially interesting that upAUGs are highly enriched in genes that require tight regulation such as proto-onco genes, transcription factors, and genes involved in the cell cycle (Kozak 1987). Looking downstream in the ORF, Kochetov (2005) found mRNA sequences that contained annAUGs with poor sequence context had a considerably higher frequency of in-frame AUGs directly downstream with suppression of out of frame dnAUGs. UpAUGs and dnAUGs in genes with reduced annAUG sequence context strongly suggest the functionality of these upstream and downstream sites as a means to regulate translation and allow functional differences within proteins.

The conclusions from these studies were supported with findings from ribosome profile data (Ingolia 2009, 2011). This method immobilizes 80 S ribosomes as they are translating. The ribosome bound mRNA fragments, corresponding to the ribosome’s footprint (about 30 nucleotides), are sequenced with deep sequencing. The ribosome footprints show which regions of mRNA are bound by the 80 S, suggesting translation of those areas.
Ribosome profile data suggests that a quarter of budding yeast 5’UTRs have abundant 80 S translating ribosome occupancy, partially due to the presence of upstream ORFs. Also a larger than expected number of 5’UTR initiation sites were found to be non-AUG codons. Translation initiation increases in the 5’UTR when cells are starved (Ingolia 2009), suggesting some role for translational regulation in stress conditions. In addition, ribosome profile data from mouse embryonic stem cells (Ingolia 2011) suggests that nearly half of the initiation sites identified through ribosome profiling were not annotated and the majority were found downstream of a gene’s annAUG. In some cases both the annotated site as well as a downstream site were used, resulting in gene products with different N-termini.

Initiation at alternate start sites during the translation of a gene has been documented and described in a number of ways. Molecular genetic approaches have shown that certain gene products have different localizations and functions through initiation at alternate start sites. Large-scale mRNA sequence analyses have shown that upAUGs and dnAUG are found in many mRNA sequences and likely have some functional role in translational regulation (Rozogin 2001, Kochetov 2004, Kozak 1987). Ribosome profile data (Ingolia 2009, 2011) strongly supports the functionality of alternate initiation start sites and has elucidated the high frequency of non-standard translation initiation.

The work described in this thesis, which began before the influential ribosome profile data was made available, will support the claim that a portion of the budding yeast proteome is under-annotated. I will discuss how, in addition to
molecular genetic approaches, peptide mass spectrometry was used as a novel and efficient means to identify and screen functional translation initiation start sites. I sought to identify N-terminal peptide sequences to characterize what sites ribosomes have selected for initiation. A thorough characterization of protein N-termini, similar in scope and impact to an analysis of 80 S ribosome occupancy, will greatly contribute to the current understanding of translation initiation. Detection of N-terminal peptides that must result from initiation downstream of the annotated AUG is a convincing way to identify examples of non-standard initiation.
Chapter 2

Analysis of Upstream Protein Translation Initiation
Abstract

Bioinformatic evidence suggests that some budding yeast genes have functional translation start sites found upstream of the annAUG. Ribosome profile data has shown 80 S ribosome occupancy in the 5’ UTR of certain genes (Ingolia 2009) and peptides have been identified that map to protein sequence preceding the annotated methionine of certain genes. Candidates for upAUG analysis were chosen if they had detected upPeptides, sequence conservation in the 5’ UTR, or an annAUG predicted to be sub-optimal. A mutagenesis experiment was done to perturb the functionality of a candidate’s annAUG to assess if an upAUG could be sufficient to initiate translation. Strikingly both upAUG candidates CDC9 and APP1 showed protein expression with mutant annAUGs suggesting that an upAUG or other upstream noncanonical site is able to initiate translation. Candidate AIM29 lacked protein expression with a mutant annAUG suggesting that the site is necessary for initiation.
2.0 Introduction

The majority of annotated genes in budding yeast follow the canonical model whereby one AUG initiates translation (Section 1.1). A gene’s annAUG was described first as the AUG that could initiate the longest ORF possible. This was then refined with the homology and conservation studies discussed in Section 1.3. In this chapter I will discuss evidence that in some cases an AUG upstream of the annAUG is used to initiate translation. Since a ribosome must select an AUG without knowledge of available downstream AUGs, upstream AUGs that act as functional start sites are likely to be found in good sequence context. In these cases an upAUG would initiate translation, but a putative recoding event could shield this upAUG initiation event from detection and make the annAUG appear to be the only AUG used. At the time of this work I did not consider the use of non-AUG sites (Section 1.2.7) by the ribosome as a means to explain the bioinformatic evidence for upAUG initiation. The models of initiation discussed in this chapter do not include the use of non-AUG initiation, except for a few brief comments. However, noncanonical initiation is an alternate explanation to upAUG initiation for these candidate genes. Either the use of an upAUG or upstream noncanonical site suggests under-annotation.

The use of upAUGs and upORFs is found in genes across species including in budding yeast to regulate protein production and allow for variable N-termini. There are three different models for upstream initiation:

I. An upAUG initiates an upORF and the stop codon is found before the annORF
II. An upAUG initiates an upORF, and the stop codon, which is out of frame with the annotated ORF, is found in the annORF sequence because the upORF and annORF are overlapping.

III. An upAUG initiates an ORF but shares the same stop codon as the annORF (Morris 2000)

In situation I, the upORF can regulate translation of the annORF by competing for competent ribosomes (Grant 1994). In the case of II, the upORF could terminate at the in-frame stop codon and act as a regulatory upORF as with I. However a frameshift into frame 1 would allow the ribosome to bypass the stop codon and continue elongation resulting in an alternate N-terminus for the annORF. This is similar to model III, where depending on a cell’s needs, an alternate AUG (upAUG or annAUG) could be used to yield a variable N-terminus. In this chapter’s work I sought to identify and test candidate genes that, based on convincing bioinformatic evidence, use an upAUG to initiate translation.

2.1 Molecular Evidence for upAUG Initiation

2.1.1 Peptide Mass Spectrometry

Work from Eddes et al. (unpublished) found high quality mass spectra from budding yeast that mapped to protein sequence not present in annotated ORFs. This is partially due to posttranslational modifications not considered in the search parameters that prevent peptide identifications. However some spectra may be unmatched because the protein sequence was not found in the sequence database due to an unpredicted upAUG initiation event. In fact, some of these unassigned peptide
matches mapped to sequence before the defined ORF of certain genes. An upstream site would need to initiate translation for these peptides to exist.

### 2.1.2 Assessment of Sequence Sets Containing upAUGs

Unpublished work from Michael Weir analyzed the information (used as a measure of start site functionality described in Section 2.1.7) at nucleotide positions surrounding annAUGs in mRNA sequence sets containing upAUGs. When the sequences were aligned at the annAUG those sequence sets with an upAUG had reduced information at position -3 and +5 than those without upAUGs. When an upAUG is found closer to the annAUG the information at position -3 and +5 is progressively reduced (further) (Figure 2.1). In addition, the frequency of nucleotides at position -3 was used to compare sequences with upAUGs to those with no upAUGs in proximity to the annAUG. In sequence sets where an upAUG is >200 bases from the annAUG, the frequency of A is much greater than in a sequence set where there is an upAUG < 25 bases from the annAUG. In addition, the frequency of A in the sequence set with an upAUG < 25 bases is similar to the frequency of A in the 5’ UTR (Figure 2.2). This analysis of information and nucleotide frequency suggests that some yeast genes may contain functional upAUGs and annAUGs with reduced consensus sequences.

### 2.1.3 Ribosome Profiles

It has been shown with ribosome profile data that there is 80 S ribosome occupancy in about a quarter of the budding yeast 5’UTRs (Ingolia 2009). The 80 S ribosome is formed once the ribosome initiates at an AUG and its presence strongly
suggests translation. This ribosome profile data further supports the roughly 40 genes that molecular genetic experiments have shown to have functioning upORFs in budding yeast (Zhang 2005). In addition, stressed conditions such as starvation increase ribosome occupancy in the 5'UTRs, which may indicate a programmed response (Ingolia 2009).

2.1.4 Experimental Design: upAUG Candidate Selection

To assess individual genes for upAUG initiation, candidates were selected based on several indicators of the presence of a functioning upAUG. The three upAUG candidates were part of an annAUG mutagenesis experiment to assay if initiation at the identified upAUG of each gene was visible when the annAUG was disabled through mutation. The presence of protein from the genes with mutant annAUGs suggests that the upAUG is used, as bioinformatic testing has determined that it has a high likelihood of functionality. The three candidates selected were \textit{CDC9}/YDL164C, \textit{APP1}/YNL094W, and \textit{AIM29}/YKR074W based on the criteria in section 2.1.5-2.1.7. mRNA transcript of the correct size (detected via RT-PCR and gel electrophoresis) was crucial in selecting the final candidates for analysis. To support the hypothesis that upAUGs are used to initiate translation it was important to show, as a preliminary step, that mRNA was produced containing the predicted upAUG for each candidate gene. RT-PCR (Figure 2.3) was carried out with primers (Figure 2.8) that amplified the area between the upAUG and annAUG. The three candidates \textit{CDC9}, \textit{AIM29}, and \textit{APP1} showed amplified cDNA from primers designed to flank the upstream sequence.
2.1.5 Unassigned Spectra

To first identify candidates for upAUG initiation, mass spectra from Eddes et al. (unpublished) was searched for evidence of upstream peptides. A fasta file containing upAUGs found before the annAUG was used to search the unassigned spectra using the Sequest search algorithm (Eng 1994) implemented through Transproteomic Pipeline (Deutsch 2010). Peptide matches were assessed with PeptideProphet, which assigned f-val scores as a measure of quality. The existence of high quality upPeptides necessitates translation initiation upstream of the annAUG.

2.1.6 Sequence Conservation

If translation was to begin at an upstream AUG we might expect high sequence conservation in the region of the upAUG across yeast species as similarly discussed in Section 1.3. Using the fungal alignment viewer from SGD (www.yeastgenome.org), our candidate genes were analyzed for sequence conservation. Upstream sequences were submitted to the fungal alignment viewer and the alignment of *S. cerevisiae*, *S. bayanus*, *S. paradoxus*, and *S. mikatae* was used to determine if the sequence of and around the upAUG was conserved.

2.1.7 TRII Score

To assess not only upAUGs but also the annAUG of sequences with upAUGs, Translation Relative Individual Information scores (TRII) (Weir and Rice 2010) were used as a means to assess the functionality of a start site. As mentioned in Section 1.1, there is pressure on a translation start site to conform to a consensus sequence for successful recognition by the ribosome. TRII scores acts as a measure of
how well an individual start site matches to a consensus sequence weight matrix by measuring the relative individual information of a sequence. A weight matrix is created by aligning sequences at known start sites to generate a consensus sequence. Individual information weight scores at specific positions can be calculated within the weight matrix and then individual mRNA sequences can be evaluated based on those scores. For example in yeast it is common to have an A at position -3. If a sequence did not have an A, this would negatively impact the overall score.

### 2.1.8 Models for Upstream Initiation

After choosing upAUG candidates it was important to identify a model in which the implicated upAUG could sometimes be used to initiate translation in place of the annAUG. This involved an understanding of recoding events, discussed in Sections 1.2.2-1.2.7, and how translation could begin at an upstream AUG for each candidate (Table 2.1).

**CDC9** was shown to have a peptide match to protein sequence upstream of the annAUG (Figure 2.4). **CDC9** had a annAUG with a low TRII score (2.3 bits) and a high scoring upAUG (7.2 bits). The candidate upAUG was in frame 2 and the proposed method of translation would allow initiation at the upAUG, followed by a frameshift into frame 1 at a predicted slippery site at nucleotide position 889 (the annAUG is position 1001 (Table 2.1).

**APP1** showed two peptide matches, one overlapping the protein sequence of the annAUG and the second further upstream (Figure 2.4). The annAUG had a higher TRII score (5.5 bits) compared to the upAUG (2.6 bits) but the two spectra matches
made this a strong candidate. In this model the ribosome would initiate at the frame 2 upAUG, and then would slip into frame 1 at the predicted slippery site at nucleotide position 776 (Table 2.1).

*AIM29*, though it had no peptide matches to upstream protein sequence it was included in the analysis because it had a predicted low scoring annAUG (.21 bits) compared to the upAUG (5.8 bits). We predicted that translation could commence at the frame 2 upAUG and slip into frame 1 for the remainder of elongation at one of the three predicted slippery sites (Table 2.1).

**2.1.9 Experimental Design: Molecular Genetic Analysis**

The three upAUG candidate genes were first epitope-tagged with a HA C-terminal tag (Figure 2.6). Budding yeast are ideal for epitope-tagging given their ability to use homologous recombination, which allowed the HA sequence to be chromosomally inserted at the C-terminal end of the sequence directly before the endogenous stop codon.

Once epitope-tagged, the candidates were cloned into a pRS416 shuttle vector for mutagenesis of the annAUG. The annAUG of each candidate was mutated so the genomic DNA read TTG instead of ATG. This was done to eliminate initiation at the annotated start codon. *AIM29* was mutated to ATT to disrupt an Eco RI site. The mutated candidate plasmid was then transformed into a strain of yeast that was a candidate-specific homozygous knockout (Thermo). Through homologous recombination a selection marker is inserted and the ORF of the gene of interest
deleted (Winzeler 1999). *CDC9* was a heterozygous knockout as the homozygous version was lethal.

### 2.2 Results

A Western blot (Figure 2.7) was done to assay protein expression of all three upAUG candidate proteins in the annAUG mutant strains. Both *APP1* and *CDC9* showed protein produced from the annAUG mutant strain while *AIM29* showed no protein from the annAUG mutant. This suggests that *CDC9* and *APP1* are able to produce protein even when the annAUG is mutated. There are different ways to interpret these results (Figure 2.9).

**I.** The candidate upAUG is able to initiate translation instead of the annAUG. In this case however it would be expected to see an increase in size from the annotated protein. Initiation at the upAUG would yield a predicted 5 kDa increase with *CDC9* and a 9.3 kDa increase with *APP1*. These size shifts were not visible with the Western blot, however the gel was not optimized for separation of small molecular weight differences and a protein’s position on a gel does not always correlate exactly to molecular weight.

**II.** An upstream noncanonical or alternate upAUG is the site that initiates translation. Noncanonical sites are seen in yeast (Ingolia 2009) and could be found closer to the annAUG allowing for the similar size of protein product between the mutant annAUG and the wildtype. Noncanonical sites require strong Kozak consensus sequences surrounding the initiations codon (Chen 2008).
III. Despite the mutation of the annAUG from AUG to UUG, ribosomes could initiate at the noncanonical site created with the mutant annAUG start site. Noncanonical start sites are especially viable if there is only one base pair change from AUG (Ingolia 2009). *APP1* has a high scoring annAUG and therefore the consensus sequence is strong making the mutant annAUG a noncanonical candidate. *AIM29* showed no protein produced when the annAUG was mutated. This suggests that the annAUG is crucial for production of Aim29p. This is especially interesting since this candidate had a annAUG predicted to have low function.

### 2.3 Conclusions and Future Directions

Evidence from ribosome profile data (Ingolia 2009), peptide mass spectrometry (Eddes et al unpublished) and mRNA Information analysis (Michael Weir unpublished) supported the presence of functioning upAUGs in certain genes. Three candidate genes were chosen for mutagenesis of their annAUGs to assess the ability of a putative upstream start site to initiate translation. Western blot analysis showed that both *CDC9* and *APP1* produced protein despite the annAUG mutation. The results were striking but not conclusive. The data instead allowed for two new interpretations in addition to the original; the presence of an upstream noncanonical site or noncanonical initiation at the mutant annAUG.

To confirm that the mutant annAUG was nonfunctional, a mutagenesis of the surrounding sequence would help to confirm that the annAUG was not being used as a noncanonical initiation site. If protein was seen, after sequence mutation, as it was
in the original experiment, it would suggest that it was the result of an upstream initiation and not the mutant annAUG because it would be unlikely that the annAUG was still used with several introduced mutations.

To test the candidate upAUG’s ability to initiate translation it would be informative to mutate the implicated upAUGs in the existing annAUG mutant strains. Lack of protein would support the requirement of the upAUG. If protein was seen that would suggest that another AUG or upstream noncanonical start site is used.

In addition, a screen of the glutaraldehyde-modified spectra from Chapter 3 would allow confirmation of the high confidence peptide matches. The screen would search the spectra with theoretical protein sequences from initiation at upAUGs. To test for noncanonical upstream sites a screen of noncanonical sites such as UUG could be done.

2.4 Methods

*Verification of mRNA Transcript*

RT-PCR Primers were designed to amplify sequence before the upAUG and sequence after the annAUG. I began by doing a phenol RNA extraction with mid-log phase cells. Cells were resuspended in TES solution (Tris pH 7.5, EDTA, SDS) and phenol and vortexed followed by incubation at 65°C. After centrifugation the aqueous phase was combined with NaOAc and ethanol and pelleted by centrifugation. The RNA was then be resuspended in water after an ethanol wash. The RETROscript kit from Ambion was used to carryout a two-step RT-PCR experiment. Random decamer primers were used with reverse transcriptase to amplify the mRNA template. After
ruling out genomic DNA contamination with gel electrophoresis, PCR was carried out with candidate specific-primers. Presence of PCR product was confirmed with gel electrophoresis (Figure 2.3).

**TRII Score Calculation**

A set of mRNA sequences were submitted to the Information Theoretic Alignment Analysis tool on the Wesleyan IGS homepage (http://igs.wesleyan.edu/) developed by Rob Stewart, Michael Weir, and Michael Rice. These sequences had annAUGs thought to be functional, and were therefore ideal for comparison. The background frequencies were adjusted to match those of the 5’ UTR. The analysis tool then took the alignment and created a weight matrix. After this the individual sequences of an annAUG or upAUG (with 20 nucleotides upstream and 17 nucleotides downstream) were entered into the analysis tool. The program calculated a weighted score for G, A, T, C at each position (log₂ (frequency of observed/frequency of expected)) and assigned an overall score for the sequence. Poor start sites were defined as <3.3 bits and strong scores were a score > 7.4 bits based on comparisons with random sequences and high confidence sequences (Michael Weir).

**Sequence Conservation**

Using the fungal alignment viewer from SGD (http://www.yeastgenome.org/cgi-bin/FUNGI/showAlign) one of the three candidate genes was entered. The species : SGD_Scer, MIT_Sbay, MIT_Spar, and WashU_Sbay were selected for comparison. In picking which area of the gene to analyze, “ ORF DNA +1 kb up/downstream” was chosen.
HA Tag Construction

The HA construct was first amplified with candidate specific primers and this PCR product was transformed into YSH474 cells. Tagging primers were designed for individual candidate genes with 60 bases homologous to endogenous DNA sequence flanking the stop codon to allow for homologous recombination after PCR amplification. The 3’ end of each primer had sequence used to amplify the region of the HA plasmid (Bahler 1998) including a Kanamycin resistance gene (Figure 2.5). The PCR to amplify the construct was done with a 9-minute extension time for 35 cycles. PCR samples were combined and prepared for transformation by adding an equal volume of phenol and centrifuging. This was combined with ethanol and sodium acetate and spun to yield a pellet. The pellet was then resuspended for transformation.

To transform the 3HA construct, YSH474 Cells were grown overnight to early log phase and pelleted. They were resuspended in 100mM LiOAc and then mixed with PEG, LiOAc, salmon sperm, and the PCR amplified DNA from above. After incubation at 30°C and heat shock at 42°C they were plated on YPD and grown at 30°C overnight. The lawn of cells was replica plated on a YPD +G418 plate and cells were allowed to grow for two days.

Colonies were selected and grown overnight in liquid YPD at 30°C and pelleted. After resuspension in “smash and grab” buffer (Triton X-100, SDS, NaCl, Tris 8, EDTA) the cells were lysed with acid washed glass beads and phenol with vortexing for 1 hour. The samples were centrifuged and the top layer removed
and added to ethanol and NaOAc. The DNA was precipitated for 1 hour at -80°C and then spun and pelleted. Once resuspended in water, the sample was run out on a gel to verify the presence of the HA tag (visible by a 2 kb size difference between the tagged and untagged gene (Figure 2.6)).

**Mutagenesis of annAUG**

Verified HA candidate genes were cloned into the shuttle vector pRS416 (cloning done by Michelle Orsulak). Primers for mutagenesis were designed to flank the region overlapping the annAUG of each candidate. They needed to be between 25 and 45 bases long and have the area of mutation flanked by 10-15 bases of homologous sequence on either side. Ideally they should terminate with a pyrimadine and the Tm >= 80°C. PCR reactions were done using the reagents from the Stratagene QuikChange XL site directed mutagenesis kit to amplify pRS416 plasmid from a Qiagen plasmid prep. An initial denaturation step of 30 seconds at 95°C was followed by 16 cycles of 95°C denaturation (30 sec), 55°C annealing (1 minute), and elongation at 68°C. I transformed the PCR product into XL-10 competent cells and the cells were plated onto LB + Ampicillin plates and grown overnight at 37°C.

To confirm the mutation of the ATG to TTG, DNA was sent for sequencing. DNA was extracted using a Qiagen plasmid prep and resuspended in TE. Once confirmed, the DNA was transformed into candidate-specific knockout strains. Cells were plated onto YPD plates and replica plated onto SD-URA plates. Surviving colonies were selected and grown.

**Western Blot for Mutagenesis Results**
25 ml cultures of each candidate strain (wild-type in pRS416, mutant in knockout, HA chromosomal, and knockout) were grown to mid-log phase and lysed with acid washed glass beads and RIPA buffer (150 mM NaCl, 1% Igepal, 0.1% SDS, 50 mM Tris pH 8.0). The supernatant was spun down at 5,000 RPM and incubated at 4°C with HA resin (Roche) to bind the tagged candidates. After elution in sample buffer the protein was run on a 10% SDS-PAGE gel and transferred for an hour onto a PVDF membrane. The membrane was probed with anti HA primary (Roche) and anti-goat secondary antibody (Jackson) followed by detection with an ECL kit (Amersham).

**Primer List**

**RT-PCR Transcript Verification Primers**

**CDC9** Forward:

5’-CTTGAGATGGCTGATGG-3’

**CDC9** Reverse:

5’-ACTTTCTTCCGGTGTTGC-3’

**AIM29** Forward:

5’-GAAGGAAGTTAGAACTTGAAATGC-3’

**AIM29** Reverse:

5’-ACTCCAATTACCGAGAAGGGTT-3’

**APP1** Forward:

5’-GGTTGCACAAGGCTTTCAA-3’

**APP1** Reverse:

5’-CATCCTGGCGATCATTGCATCGCG-3’
**HA Tagging PCR Primers**

*CDC9* Forward:

5’-GATGCAACCTCTTCTGACCAGATCGTGAATTGTATGAAAACCAATCCCACA
TGCAAAAATCGGATCCCCGCGGTAAATTAA-3’

*CDC9* Reverse:

5’-CTGAACAGAAGCCTCTTTTTTTTTTTATGAAATTGAGAAAAATAGTGT
GCACATGAAATCAGCAGTCTTTAAAC-3’

*AIM29* Forward:

5’-ATTTCGCTTTTCAAACAAAGGATTACCTAAGGTTAAAGAAAAACCCTGAAG
AAAAATGGCGGATCCCCGCGGTAAATTAA-3’

*AIM29* Reverse:

5’-ATATATGTAATATATTTTCTAGCAGCTTTTACTTATATGAAAGCCATAAGC
CAAAACACATCGAATCAGCAGCTTTAAAC-3’

*APP1* Forward:

5’-TCG GAT CCG GCA TTA AGT TTG GAA GAC AGT ATT CGC AGA ATT AGG
GAG AAG TAT TCA AAC CGG ATC CCC GGG TTA ATT AA-3’

*APP1* Reverse:

5’-GGT CAC TTT AGT GTA TAC AAT TTT TAA ACT CCC TCC CGA TGT ATA TAA
ATA ACA GTG TAT GAA TTT GAG CTC GTT TAA AC-3’

Note: The blue font indicates plasmid specific sequence.

**Mutagenesis Primers**

*CDC9* Forward:

5’-GTTCATCAATTACTTGCGCAGATTACTGACCAGGTTGCC-3’
CDC9 Reverse:
5’-CAAGTAGTTAATGAACGCGTTCTAATGACTGGCCAACGG-3’

APP1 Forward:
5’-GGAGAAGGAATAAAAAAGTTGAATAGTCAAGGTTACGATGAAAGCTC-3’

APP1 Reverse:
5’-CCTCTTCCTTATTTTTCATACCTATCAGTTCCAATGCAATGCTACTTTTCGAG-3

AIM29 Forward:
5’-GCATTATTTCATAATTCAACTAAAAAGGCTGAAAATGTCAACAACC-3’

AIM29 Reverse:
5’-CGTAATAAAGTATTTAAGTGGATTTTCGACTTTTACAGTGTTGG-3’
2.5 Figures
Figure 2.1: Relative Information Scores for Sequences with upAUGs

Relative Information at Nucleotide Positions Surrounding annAUGs in the Presence of at Least One upAUG Within Varying Distances
Figure 2.1

MRNA sequence sets containing upAUGs were compared after alignment at the annAUG (positions 1,2,3). Relative information decreases at position -3 and position +5 as upAUGs are found in closer proximity to the annAUG (Dr. Michael Weir).
Figure 2.2: Nucleotide Frequencies of Sequences with upAUGs

Frequency of Nucleotides at Position -3 in the Presence of at Least One upAUG Within Varying Distances

Position of upAUG
Figure 2.2

MRNA sequence sets containing upAUGs were compared after alignment at the annAUG. When no upAUGs are present in sequences, the frequency of A at position -3 is much higher than when there is an upAUG within 25 bases. Not only is the frequency of A lower in this set, but the frequency of A is similar to that of the frequency in the 5’ UTR (Dr. Michael Weir).
Figure 2.3: RT-PCR Analysis of mRNA Transcript Presence in upAUG Candidates

APP1

CD9

Genomic cDNA

Genomic cDNA
Figure 2.3

RT-PCR and gel electrophoresis analysis of upAUG candidate mRNA. Primers, flanking sequence from the upAUG to the annAUG, were used to amplify both genomic DNA as well as cDNA. This demonstrates that the implicated upAUGs are present in the mRNA message. The AIM29 gel was unavailable.
Figure 2.4: Unassigned Spectra Map to Peptides in the 5’UTR of CDC9 and APP1

1. **APP1**

   \[ \text{upAUG} \rightarrow \text{...LSINSVICSSSVGLLLAIK...KMNSQGYDESSSSSTAATSGPTSGDPRMGK} \]

   \[ \text{annAUG} \]

2. **CDC9**

   \[ \text{upAUG} \rightarrow \text{...SNDFGGLEQLLTSLFINYM} \]

   \[ \text{annAUG} \]
Figure 2.4

1. *APP1* shows an upAUG followed by a predicted slippery site. The two detected upstream peptides are in red with the second overlapping the annotated methionine.  
2. *CDC9* has an expected slippery site after the upAUG, followed by a detected upstream peptide overlapping the annotated methionine.
Figure 2.5: upAUG Experimental Workflow: HA Tagging Through to Mutagenesis

1. Candidate sequence

2. HA

3. Candidate sequence
Figure 2.5

1. Representation of HA construct from Bahler et al. (1998). The grey boxes represent the KanMX6 module promoter, terminator, and Kanomycin resistance gene. The black boxes represent the tagging modules within the plasmid including the terminator sequence from the ADH1 gene. Arrows inside boxes represent the direction of transcription and the flanking black arrow represent the candidate specific tails of the primers used to amplify the HA construct. 2. The location of the HA tag was directly before the endogenous stop codon of each gene. 3. The change from ATG to TTG after mutagenesis.
Figure 2.6: PCR Verification of C-terminal HA Tag

\[
\begin{array}{cc}
\text{CDC9} & \text{AIM29} \\
\text{HA} & \text{WT} \\
\end{array}
\]

\[
\begin{array}{cc}
\text{HA} & \text{WT} \\
\end{array}
\]
Figure 2.6

Gel electrophoresis of PCR product of two HA-tagged candidate genes as well as the wildtype. The increase in size from the C-terminal HA epitope tag (about 2 KB) is visible. APP1 gel unavailable.
Figure 2.7: Western Blot Analysis of annAUG Mutagenesis
**Figure 2.7**

Western blot results show each of the candidate’s four strains (wildtype-3HA in plasmid, annAUG mutant, wildtype-3HA, and Knockout). Both *CDC9* annAUG mutant and *APP1* annAUG mutant show protein expression while *AIM29* does not. Note the upper band in *AIM29* is a nonspecific band from the anti-HA antibody found at 29 kDa.
Figure 2.8: Sequence Conservation in Area of upAUG

**CDC9**

SGD_Scer_CDC9/YDL164C  837 TGGAGATGCTGGAATTTTACCTAAAGCGGAAAGCCGCAACG  886
MIT_Sbay_c843_3397       843 CAAACAGCCTGGAATTTTACCTAAAGCGGAAAGCCGCAACG  891
MIT_Spar_c429_3168       835 CGGAGATGCTGGAATTTTACCTAAAGCGGAAAGCCGCAACG  884
WashU_Sbay_Contig592.8   843 CAATACGCTGGAATTTTACCTAAAGCGGAAAGCCGCAACG  891

**AIM29**

SGD_Scer_AIM29/YKR074W    985 ATTGTCTAATCAATTTACTGCAATTTACTGAAGTTCTGCTTTTCTTTCA  1033
MIT_Sbay_c283_14951       984 GTCACTTCAACATATGACAGACTCGTGGCTGTTACTTACCCA   1033
MIT_SmiK_c122_13965       984 ATTGTCTAATCAATTTACTGCAATTTACTGAAGTTCTGCTTTTCTTTCA  1033
MIT_Spar_c404_13216       984 GTCACTTCAACATATGACAGACTCGTGGCTGTTACTTACCCA   1033

Symbols:

- SGD_Scer_CDC9/YDL164C
- MIT_Sbay_c843_3397
- MIT_Spar_c429_3168
- WashU_Sbay_Contig592.8

- SGD_Scer_AIM29/YKR074W
- MIT_Sbay_c283_14951
- MIT_SmiK_c122_13965
- MIT_Spar_c404_13216

Symbols:

- SGD_Scer_AIM29/YKR074W
- MIT_Sbay_c283_14951
- MIT_SmiK_c122_13965
- MIT_Spar_c404_13216
Figure 2.8

Nucleotide alignments of four yeast species (*S. cerevisiae*, *S. bayanus*, *S. paradoxus*, and *S. mikatae*). Red boxes label the putative upAUG and black boxes label the annAUG. The arrows indicate the location of primers used for RT-PCR verification of transcript seen in Figure 2.3. The *APP1* forward primer had to be designed to match sequence after the upAUG to achieve successful binding.
Figure 2.9: Models for Translation Initiation in upAUG Candidates CDC9 and APP1

1.

2.

Noncanonical
Figure 2.9

Possible models of translation initiation in upAUG candidates. Model 1 suggests that initiation occurs at the upAUG at least some of the time. Model 2 suggests that translation is initiated at another upAUG or a noncanonical upstream sequence. Both models also suggest the possibility that translation begins at the annAUG.
**Table 2.1: upAUG Candidate Summary**

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Upstream conservation</th>
<th>annAUG TRII score</th>
<th>upAUG TRII score</th>
<th>Detected upPeptide</th>
<th>Predicted slip site</th>
<th>upAUG</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CDC9</strong></td>
<td>Yes</td>
<td>2.3</td>
<td>7.2</td>
<td>SNDFGGLEQLLTSLFINY MIR</td>
<td>889</td>
<td>843</td>
<td>4</td>
</tr>
<tr>
<td><strong>APP1</strong></td>
<td>No</td>
<td>5.5</td>
<td>2.6</td>
<td>LSINSVICSSSVGLLAIK</td>
<td>776</td>
<td>720</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>KMNSQGYDESSSSTAATSGPTSGPRMGKK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AIM29</strong></td>
<td>No</td>
<td>0.21</td>
<td>5.8</td>
<td>N/A</td>
<td>968,971,974</td>
<td>951</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 2.1

Description of upAUG candidates. Upstream conservation was a visual assessment of sequence conservation (Figure 2.8). Both annAUG TRII score and upAUG TRII score were calculated as mentioned in the methods section (2.4). The upPeptides were identified from Eddes et al. available spectra. The predicted slippery sites were identified from a search of our relational databases. The sites were identified with a script that searches the budding yeast genome for heptomeric slip sequences (Justin Cherny, Adam Robbins-Pianka, Kris Truncali). The rank and nucleotide start of each upAUG is also present.
Chapter 3

Amino Termini of Many Yeast Proteins Map to Downstream Start Codons
The work discussed in this chapter was a collaboration done by the authors cited below. I worked on the generation of the glutaraldehyde-modified lysate and the MS/MS analysis, as well as the individual dnAUG TAP purifications, and the parent protein size assessment. Fellow Ph.D. student Justin Cherny worked equally on the analysis of the PeptideAtlas data, downPeptide reading frame assessment, ribosome profiles, CAI and TRII score analysis, and contributed to the figures. Dr. Michael Weir wrote the stored procedures used to assess N-terminal peptides and did the assessment of N-terminal amino acid composition in acetylated peptides (Polevoda 2003). He also wrote the majority of the text and figures. Kris Truncali created the fasta databases used for both the glutaraldehyde and PeptideAtlas peptide identifications and ran the PeptideAtlas spectra through Sequest. Miin Lin ran many of the MS/MS trials for the parent protein assessment. Adam Robbins-Pianka contributed to the sequence database generation.

I have added a small supplemental chapter (3S) to discuss the purification of TAP tagged dnAUG candidates as they are only briefly discussed here. The text, figures (Chapter 3), and the appendix are taken from the published paper cited below. I am very grateful to everyone for their contributions.

Abstract

Comprehensive knowledge of proteome complexity is crucial to understanding cell function. Amino termini of yeast proteins were identified through peptide mass spectrometry on glutaraldehyde-treated cell lysates as well as a parallel assessment of publicly deposited spectra. An unexpectedly large fraction of detected amino-terminal peptides (35%) mapped to translation initiation at AUG codons downstream of the annotated start codon. Many of the implicated genes have suboptimal sequence contexts for translation initiation near their annotated AUG, and their ribosome profiles show elevated tag densities consistent with translation initiation at downstream AUGs as well as their annotated AUGs. These data suggest that a significant fraction of the yeast proteome derives from initiation at downstream AUGs, increasing significantly the repertoire of encoded proteins and their potential functions and cellular localizations.
3.0 Introduction

The size and complexity of eukaryotic proteomes has been of considerable interest, especially in light of the unexpected reductions in the estimates of gene number predicted after sequencing of human and other genome (Lander 2001, Venter, 2001) However, recent ribosome profiling data (Ingolia 2009, 2011) and observations of nonstandard translation initiation mechanisms, such as leaky scanning or translation reinitiation (Harrison 2002, Kochetov 2005, Kochetov 2008, Wang 2004, Porras 2006, Kochetov 2005, Kozak 2002) (Figure 3.1D), have led to suggestions that eukaryotic proteomes may be significantly more complex than currently annotated. For example, translation initiation at AUG codons downstream (dnAUGs) of the annotated start codon (annAUG) would lead to truncated proteins, or if out of frame, proteins with different amino acid sequence, with potentially new functions or cellular localizations and targeting.

Translation initiation is thought to depend on a ribosome scanning mechanism (Kozak 1989) in which ribosomes typically initiate translation at the first AUG they encounter when scanning from the 5’ cap of the mRNA; however, AUGs can be skipped if they are too close to the 5’ cap (Kozak 1991) or in a poor sequence context (Kozak 2002). Current annotations of proteomes in genome projects generally rely on the assumption that, in most cases, individual mRNAs each give rise to a single protein, typically encoded by the longest open reading frame (ORF) uninterrupted by stop codons. For example, the initial annotation of the yeast proteome identified protein coding regions as ORFs encoding more than 99 amino acids (Goffeau 1996).
The annotated translation start positions were in some cases subsequently adjusted to more C-terminal start sites when cross-species sequence comparisons revealed poor protein conservation at the site originally annotated (Kellis 2003). But is it correct to assume that scanning ribosomes generally initiate at only one site on an mRNA? By analyzing the amino-termini of proteins on a proteomic scale through peptide mass spectrometry, we show here that for many genes, ribosomes appear to use multiple start sites, including sites downstream of the annotated start codon, giving rise to protein products with different amino termini, and in some cases, different reading frames.

To assess the repertoire of amino termini of proteins in the budding yeast proteome, tandem peptide mass spectrometry (MS/MS) was performed on yeast cell lysates. Proteins in cell lysates were digested in vitro with trypsin endopeptidase, and the resulting peptides were sampled and further fragmented by collision induced dissociation (CID). The Sequest algorithm (TurboSequest (Eng 1994)) was used to determine the sequences of tryptic peptides through computational comparison of experimental spectra with theoretical spectra computed from an exhaustive sequence database of annotated proteins in the yeast proteome. By also including in the database sequences expected from initiation at dnAUGs within 100 nucleotides (nt) of the annAUG, we identified peptides resulting from initiation at both dnAUGs and annAUGs. The peptides originating from the N-termini of proteins, including those initiated at dnAUGs, can be distinguished from internal and C-terminal tryptic fragments because they do not have a trypsin site at their N-terminus (Figure 3.1A).
The sequence database included amino-terminal sequences from signal peptide cleavage predicted by the SignalP algorithm (Bendtsen 2004) and methionine cleavage by methionine aminopeptidase (Chen 2002) (Figure 3.1C).

Output from the TurboSequest algorithm was filtered by decoy analysis (Fitzgibbon 2008, Elias 2007) to ensure a target false identification rate below 5%. Confidence in the TurboSequest matches was confirmed in control experiments that tracked parent protein masses before trypsin digestion (Chapter 4), and separate control experiments using partial purification of C-terminal-epitope-tagged proteins (see below and see Chapter 3S).

The detection efficiency of N-terminal peptides was increased through treatment of proteins in cell lysates with glutaraldehyde which modifies their amino termini by reductive alkylation (Russo 2008). The reductive alkylation results in a molecular weight increase of +68 Da on the amino-terminal amino acid of proteins as well as internal lysines, and increases MS/MS detection of a1 fragment ions. Proteins that had been acetylated in vivo would not be modified at their N-terminus by glutaraldehyde. Therefore, tryptic peptides were assessed for potential molecular weight increases of +68 Da (glutaraldehydation) or +42 Da (acetylation). TurboSequest matches were observed to 69 glutaraldehyde-modified and 54 acetylated N-terminal peptides, including one observed in both forms (Table 3.1).

3.1 Results and Discussion

Of the 122 different N-terminal peptides observed in the glutaraldehyde-treatment experiments, 99 corresponded to N-terminal peptides expected from
translation initiation at annAUGs (Table 3.1). A surprisingly large fraction (19%, 23 peptides) of the detected N-terminal peptides corresponded to peptides expected from translation initiation at dnAUGs (Table 3.1; examples are illustrated in Figure 3.2 and Supplementary Figure S1, Appendix). Given this unexpected result, the analysis was expanded to include an assessment of publicly deposited spectra for yeast cell lysates (http://www.peptideatlas.org/repository). Based on the public data, a total of 818 different N-terminal peptides were detected including 299 peptides corresponding to initiation at dnAUGs (37%; Table 3.1). The data for the glutaraldehyde treatment experiments with those of the publicly deposited spectra, we observed 583 peptides for annAUG initiation (“annPeptides”), and 320 peptides for dnAUG initiation (35% “downPeptides”; Table 3.1, Supplementary Table S2, Appendix). Although for most genes with detected N-termini, we observed either annPeptides (531 genes) or downPeptides (271 genes), but not both, it is likely that some of these genes express both forms even though only one was detected in this sampling of the proteome.

Indeed, 26 genes had different N-terminal peptides that mapped to the annAUG and a dnAUG; and two additional genes had peptides for the annAUG and two different dnAUGs; moreover, 12 genes had peptides for two dnAUGs but not their annAUG. The frequencies of annPeptide and downPeptide detection were similar for many of the 28 genes for which both classes were detected (Supplementary Table S3, Appendix). Detection of five of the frame-1 downPepides (HOM3, YJL171C, EFT2, PRS5, and URA1) was confirmed following partial purification of C-terminal-epitope-tagged versions of the proteins. For four of these (all except URA1), we confirmed
detection of both an annPeptide and a downPeptide (Chapter 3S).

We examined post-translational modifications of the 903 detected N-termini: 393 N-termini showed methionine aminopeptidase cleavage of the first amino acid (339 cases) or signal peptide cleavage (Bendtsen 2004) (Table 3.1), 409 were acetylated, 458 were not acetylated, and 36 were detected in both forms, acetylated and not acetylated. In addition, the amino acids found in the first four positions of the acetylated and nonacetylated (glutaraldehyde modified) N-terminal peptides had similar frequency profiles to those observed by Polevoda and Sherman (Polevoda 2003) who collated published yeast protein amino terminus data from multiple researchers (Supplementary Figure S2, Appendix). For example, both data sets showed that an N-terminal serine is typically acetylated. One-hundred five (22%) of the detected N-termini correspond to amino termini reported by Polevoda and Sherman (Polevoda 2003), and the acetylation characteristics were consistent for 102 of these (97%).

DownPeptides were detected in all three reading frames (Figure 3.3; Supplementary Table S4, Appendix). A significantly elevated percentage (41%) were in frame 1 even though only 32% of the screened dnAUGs were in this frame (chi-square \( p < 0.001 \)). This result, and the observation that the frequencies of frame-2 and -3 dnAUGs are depressed close to the annAUG (Figure 3.3; Kochetov 2005), suggest that there is selection for expression of N-terminal truncations of frame-1 proteins, rather than frame-2 or -3 proteins. Relative instability of shorter frame-2 or -3 proteins may also contribute to the higher detection of frame-1 downPeptides. The
ORF lengths of the implicated frame-2 and -3 proteins are significantly longer than randomly selected downstream ORFs, suggesting there has been selection against stop codons in these ORFs (Figure 3.3D).

High frequencies of frame-1, -2, and -3 translation initiation upstream and downstream of the annAUG have also been suggested by transcriptome-scale ribosome profiling experiments in yeast (Ingolia 2009) and mouse cells (Ingolia 2011) that describe mRNA sequences protected by ribosomes. We analyzed downPeptide genes using ribosome-profile sequence reads of Ingolia et al. (Ingolia 2009) DownPeptide genes aligned at the dnAUG showed ribosome profiles consistent with translation initiation at these sites (Figure 3.4A,B). The characteristic initiation profile was particularly pronounced for downPeptide genes with poorer sequence context surrounding their annAUGs (“Kozak consensus” (Cavener 1987, Kozak 1999) measured by Translation Relative Individual Information (TRII) score (Robbins-Pianka 2010, Weir, 2010), suggesting a trend toward dnAUG initiation for these genes. Indeed, alignment of the first AUG downstream of the annAUG of all genes, not just those identified in this MS/MS analysis, revealed translation initiation ribosome profiles that are particularly pronounced for genes with poorer Kozak context at their annAUG (Figure 3.4C,D). Ribosome tag densities were examined for 81 of the 320 detected downPeptide genes that had tag densities >0.1 tags/nt in the first 200 nt of their annORFs, and pronounced tag densities (≥10 tags/30 nt) in the 30-nt window starting at their dnAUG. All but five of these genes had >3 ribosome tags in the 30-nt window starting at their annAUG, consistent with translation initiation at
the annAUG in addition to the detected dnAUG initiation (Supplementary Table S5, Appendix). This suggests that most downPeptide genes are under-annotated rather than mis-annotated, with translation initiation occurring at both the annotated AUG and the implicated dnAUG. The unexpected prevalence of downPeptides, and parallel ribosome profile data, suggests that translation initiation at dnAUGs is common. As summarized in Figure 3.1D, mechanisms that would lead to dnAUG initiation include:

(i) expression of 5’-truncated mRNAs or spliced mRNAs that do not include the annAUG; (Muir 2006, Nagalakshmi 2008)

(ii) leaky scanning where the ribosome sometimes skips the annAUG and instead initiates at a dnAUG; (Harrison 2002, Kochetov 2005, 2008, Wang 2004)

(iii) translation reinitiation at a dnAUG following translation of an ORF that starts in the 5’UTR; (Porras 2006, Kochetov 2005, Kozak 2002)

(iv) translation initiation at internal ribosome entry sites (IRES (Baird 2006).

Truncated mRNAs (Supplementary Table S6, Appendix) may account for 16% of the detected downPeptides. These genes have high-confidence capped mRNAs (Muir 2006, Nagalakshmi, 2008) with 5’ ends that map downstream of the annAUG or within 20 nt upstream; 5’UTRs less than 20 nt in length are reported to reduce translation initiation efficiency (Kozak 1991) Many of these genes also
express longer mRNAs from more upstream transcription initiation that could allow translation initiation at their annAUG or their dnAUGs.

Expression levels of mRNAs (Muir 2006, Nagalakshmi 2008) and proteins (Ghaemmaghami 2003) are lower in downPeptide genes compared to annPeptide (Figure 3.4E-G). The lower mRNA levels may be explained in part by nonsense mediated decay (Gonzalez 2001) due to frame-2 or -3 translation initiation at dnAUGs or upstream AUGs. The lower protein levels, measured using frame-1 encoded C-terminal epitope tags, may also be explained in part by inefficiency of translation reinitiation (Kozak 2002) or lower protein half-lives. Consistent with their lower expression, fewer internal trypsin fragments were detected for downPeptide genes (average of 5 internal peptides) compared to the annPeptides genes (9 internal peptides). DownPeptide detection might also be biased toward genes with lower mRNA and protein expression due to under-annotation, for example, of undetected truncated mRNAs.

The Kozak sequence contexts surrounding the annAUGs of downPeptide genes are of poorer quality compared to those of annPeptides genes (Figure 3.4H), consistent with the expected underutilization of annAUGs in downPeptide genes. This is emphasized by the prevalence of U at the third nucleotide upstream of the annAUG in downPeptide genes (19%) compared to the strong selection against U at this position in annPeptide genes (10%; chi-square $p < 0.0002$); the identity of the $-3$ nucleotide is key to translation initiation (Cavener 1987, Kozak 1999). The downPeptide genes with poorer annAUG Kozak context also have poorer protein
sequence conservation coded near the annAUG compared to the dnAUG (Supplementary Figure S3, Appendix). The downPeptide genes also exhibit elevated Codon Adaptation Indices (CAI (Sharp 1987)) immediately downstream of their implicated dnAUG, suggesting selection for more commonly used codons in this region (Supplementary Figure S4, Appendix).

The N-terminal peptides that map to annAUGs and dnAUGs are most readily explained by translation initiation at the implicated AUGs. However, it is formally possible that alternative mechanisms, such as protease activities, could account for some instances of these N-terminal peptides, even though they map to AUG codons. For example, Vogtle et al. (Vogtle 2009) characterized the amino-termini of 615 proteins following cleavage of mitochondrial signal peptides, and one of these proteins coincidentally has the same amino terminus as one of the detected downPeptides (for the SDH2 gene). However, the vast majority of our reported downPeptides are most likely explained by translation initiation at dnAUGs, especially given the supporting data from ribosome profiling, TRII score analysis, and CAI assessment.

Translation initiation at dnAUGs expands the proteome’s repertoire of protein functions as well as targeting and localization within the cell (Supplementary Tables S7, S8, Appendix). A large number (Kozak 1999) of the detected frame-1 downPeptide genes encode a predicted signal sequence between the annAUG and dnAUG, which in some cases may target the full-length and truncated versions of the proteins to different cellular compartments (Figure 3.1B). Predicted motifs for several
enzyme activities are also encoded between the annAUG and dnAUG (Supplementary Tables S7, S8, Appendix). In addition, the amino acid composition coded between the annAUG and dnAUG of downPeptide genes has a greater tendency toward predicted disorder than the equivalent regions of randomly selected genes (Supplementary Figure S5, Appendix), which can facilitate association with regulatory cofactors (Dunker 2008). Although translation initiation at annAUGs and dnAUGs may result in proteins with different functions or localizations, it is also likely that for many genes, the alternative frame-1 products are not significantly different in their behaviors. Indeed, over evolutionary time, the diversification of initially similar products could underlie the emergence of new gene functions. Global assessment of the Genome Ontology terms (GO terms; Supplementary Table S9, Appendix) did not reveal significant differences between the terms associated with annotated frame-1 proteins of the downPeptide genes compared to the terms associated with the detected annPeptide genes. However, proteins initiating at frame-2 and -3 dnAUGs add to the yeast proteome complexity and potentially provide significant new functions.

These results suggest that the yeast proteome is more complex than currently annotated. The proteomes of yeast, and by extension other species, are likely under-annotated. In some cases, the amino-termini of proteins may be misannotated, but in many cases, genes probably utilize multiple translation start sites, as also suggested by recent ribosome profile results, (Ingolia 2009, 2011) raising the interesting possibility that translation initiation can be a regulated or probabilistic event.
Translation of alternative protein products with different functions and cellular localizations or targeting, provides a versatile mechanism to regulate and evolve gene expression behaviors in response to environmental and developmental cues (Thomas 2008). Moreover, analyses of individual gene products do not generally assume that proteins have a diversity of amino termini, and this knowledge could have profound implications for the design and interpretation of these studies.

3.2 Methods

**Glutaraldehyde Treatment of Proteins**

To enhance detection, amino termini of proteins were modified by reductive glutaraldehydation based on the protocol of Russo et al. (2008) A pellet of approximately 1000 μg of protein lysate from YSH474 yeast cells was resuspended in 0.1% acetic acid (pH 4) to give a protein concentration of 5 μg/μL. The resuspension was then brought to a concentration of 100 mM sodium cyanoborohydride and 0.5% glutaraldehyde, and incubated at room temperature for 20 min. An equal volume of 1 M Tris-HCl was added and the sample was pelleted with 30% TCA, and digested with 10 μg of trypsin (Promega); some samples were concentrated with ZipTips (Millipore). The trypsin digest was loaded onto a nanospray column packed with C18 resin (Michrom Bioresources) and analyzed using a LCQ Deca XP mass spectrometer (Thermo Scientific).

**TAP-tagged Protein Purifications**

1 or 2 liters of TAP tagged cells (Rigaut 1999) were grown to early log phase and collected by centrifugation. Pelleted cells were resuspended in 5 ml of Hepes...
lysis buffer (100mM Hapes-KOH, pH 8-8.5, 20mM Mg(OAc)$_2$, 10% glycerol, 10mM EGTA, 0.1 mM EDTA, 0.4% NP-40, 100mM PMSF n ETOH, and 2 mini-tablet protease inhibitors (Roche)) and vortexed with glass beads for lysis. The lysate was then spun at 6,000 RPM to remove cell debris and the supernatant was collected and boiled with 50ul sample buffer to elute bound proteins. The sample was run on a 10% SDS-PAGE gel, fixed, and stained with colloidal coomassie (Invitrogen). The bands of interest were excised and digested in-gel according to Shevchenko et al. (2007). The peptides were resuspended in 0.1% TFA and loaded onto a C18 resin packed column for MS/MS analysis.

**TurboSequest Assessment**

Peptides were identified with TurboSequest (Bioworks 3.2 package) using a Dell XPS server. For analysis of our glutaraldehyde-treated cell lysates, we included optional mass increases to peptides: (i) N-terminal amino acid +68 Da for glutaraldehyde modification; (ii) N-terminal amino acid +42 Da for acetylation modification (requiring a second TurboSequest run); (iii) any lysine +68; (iv) N-terminal proline +86 Da; and (v) N-terminal methionine oxidation +16 Da. Each TurboSequest parameter file included a precursor mass tolerance of 3.0 Da and a fragment mass tolerance of 1.0 Da. Precursor peptides for MS/MS analysis were prepared by trypsin digestion, which cuts after R or K, except when flanked by P. For the TurboSequest analysis, in addition to requiring trypsin-cleavage sites at both ends of the precursor peptides (or one end if a terminal peptide), an internal trypsin site was allowed in the Glutaraldehyde-treatment analysis (two internal sites were allowed.
in the PeptideAtlas analysis discussed below). TurboSequest matches to trypsin
fragments with internal trypsin sites were subsequently discarded based on our decoy
analysis (see below).

A sequence “database” file of downstream Open Reading Frames (dnORFs) in
FASTA format was constructed as follows. Using an MS-SQL database of yeast
genomic sequences (6718 genes downloaded from SGD March 2008 (Robbins-
Pianka 2010)), translated dnORFs were computed for translation initiation at all AUG
triplets within 100 nt downstream of the annotated translation start codon. DnORFs in
the same reading frame as the annAUG (frame 1) included sequence from the
initiation methionine to the following second trypsin site; frame-2 and -3 dnORFs
included the full dnORF. We also included (i) N-terminal sequences that would result
from amino peptidase cleavage of the initiation methionine if it was followed by
amino acids A, C, S, T, G, V, or P; (Chen 2002) and (ii) N-terminal sequences
following cleavage of a signal peptide predicted by the SignalP algorithm version 3.0
(Bendtsen 2004) (these sequences extended to the second trypsin site after their N-
terminus).

The TurboSequest algorithm computes a probability score for each peptide
match, and for each set of equivalent MS/MS experiments (e.g., in a single
PeptideAtlas accession; Supplementary Table S1, Appendix), spectrum matches were
considered that were below a probability score threshold chosen to ensure that the
target false identification rate was below 5%. False identification rates for detection
of N-terminal peptides were calculated as follows. For each MS/MS run, we
calculated the number of forward and reverse decoy N-terminal peptides matched. Individual peptides with more than one match were only counted once per MS/MS run, and these counts were then summed over all MS/MS runs. The target false identification rate was calculated as (decoy peptides/forward peptides) as described (Fitzgibbon 2008). The false identification rate was 5% ± 0.5% (±s.d. based on bootstrap analysis). In our glutaraldehyde treatment experiments, false identification rates <5% were obtained for probability score thresholds of 0.28 for glutaraldehyde-modified and 0.06 for acetylated peptides (Supplementary Table S1, Appendix). Matches to 35 unmodified N-terminal peptides were disregarded because the false identification rate for these matches was very high (30%), suggesting that many of these were false matches and that the glutaraldehyde treatment is efficient.

In addition to our glutaraldehyde-treatment experiments, we also analyzed publicly deposited spectra from yeast cell lysates downloaded from http://www.peptideatlas.org/repository/ and analyzed as follows. TurboSequest was run using the Sequest parameter files posted for each experimental series (Supplementary Table S1, Appendix). The parameter files were modified to include optional acetylation (+42 Da) at the N-terminus of peptides. TurboSequest was run using a FASTA file sequence “database” composed of all annORFs, as well as all dnORFs mapping within 100 nt of the annAUG, and with the methionine aminopeptidase and signal peptide cleavage modifications as described above. We analyzed 20 experimental series uploaded by multiple research groups (Supplementary Table S1, Appendix). For each experimental series, we computed a
TurboSequest probability threshold that gave a false identification rate <5% and only considered TurboSequest matches below this probability threshold (Supplementary Table S1, Appendix). Matches to N-terminal, internal, and C-terminal peptides were used to calculate the probability threshold.

In this analysis, we excluded all matches with a TurboSequest initial ranking (RSp rank) > 1 because in our glutaraldehyde-treatment experiments, matches with RSp rank >1 had a false identification rate of 18% instead of 5% suggesting that they were less reliable (also see ref (Elias 2007)). We also excluded all matches to peptides with internal trypsin sites since these matches were unreliable; for example, N-terminal peptides with a single internal trypsin site had a false identification rate of 28% consistent with reports (Olsen 2004) that trypsin digestion can be highly efficient. The TurboSequest algorithm also reported N-terminal modifications (+68 or +42 Da) to internal peptides, but these matches were expected to be false since N-terminal glutaraldehydation prior to trypsin treatment or in vivo acetylation should be restricted to the N-termini of proteins, and correspondingly, the decoy detection rate was very high (749 decoy internal peptides; 760 forward internal peptides). We also performed control experiments (not shown) in which cell lysates were partitioned into gel slices of known molecular weight size ranges (25–37, 37–50 kDa). Although the TurboSequest algorithm had no knowledge of the parent protein sizes before trypsin digestion, peptide matches conformed to the correct parent size ranges, the frequency of incorrect matches being a little higher than the decoy frequencies in these experiments. (This analysis took into account the portion of the proteome in each size
range that would be detected randomly.) These controls provide confidence in the observed matches to N-terminal peptides and their association with translation initiation at annAUGs and dnAUGs.

All outputs from TurboSequest runs were uploaded into a relational database and analyzed using stored procedures written in MS-SQL to compute false identification rates and integrate the MS/MS data with other data sets including the yeast genome, ribosome profile, mRNA and protein expression, and Gene Ontology data.
3.3 Figures
Figure 3.1

A

tryptic sites ()

dnAUG(fr1)

dnAUG(fr2or3)

stop

B

signal sequence

//

Full-length Yji171cp

//

N-truncated Yji171cp

C

M

Met N-term

M_(A,C,S,T,G,V,P)_

Met cleavage

M

Signal cleavage

Ac

Acetylation (endogenous)

Glut

Glutaraldehyde (experimental)

D

5' uAUG annAUG dnAUG

Leaky Scanning

5' N N

Reinitiation

5' N

5' N

Truncated mRNA

5' IRES N

Internal Ribosome Entry Site
Figure 3.1

Protein products from initiation at dnAUGs. (A) Amino-terminal derived peptides from translation initiation at dnAUGs can be distinguished from annAUG initiation by the absence of an amino-terminal trypsin site. (B) Example of dnAUG initiation. Yjl171cp is annotated as a GPI-anchored cell wall protein, and the detected full-length Yjl171cp protein has an ER signal sequence; however, the signal sequence is not predicted in the detected N-terminal truncated protein from initiation at a dnAUG. N-terminal peptides for both the full-length and truncated proteins were also detected following partial purification of C-terminal TAP-tagged Yjl171cp protein. (C) Sequence databases for peptide mass spectrometry included products for terminal methionine and signal peptide cleavage. Optional mass increases for acetylation (+42 Da) or glutaraldehydation (+68 Da) were also assessed. (D) Potential mechanisms for translation initiation at dnAUGs.
Figure 3.2

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MJ</td>
<td>172.05</td>
<td>200.05</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>243.03</td>
<td>371.03</td>
<td>13.165</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>K</td>
<td>450.75</td>
<td>180.75</td>
<td>952.75</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>L</td>
<td>562.27</td>
<td>580.00</td>
<td>756.47</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>N</td>
<td>779.20</td>
<td>807.39</td>
<td>550.33</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Q</td>
<td>307.45</td>
<td>658.45</td>
<td>416.29</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>L</td>
<td>303.54</td>
<td>1048.53</td>
<td>388.23</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>R</td>
<td>175.12</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MJ</td>
<td>88.53</td>
<td>100.53</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>172.05</td>
<td>159.05</td>
<td>512.33</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>K</td>
<td>728.10</td>
<td>234.69</td>
<td>410.29</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>L</td>
<td>278.64</td>
<td>260.64</td>
<td>378.24</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>N</td>
<td>332.16</td>
<td>154.16</td>
<td>502.20</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>Q</td>
<td>394.20</td>
<td>450.20</td>
<td>260.66</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>L</td>
<td>543.23</td>
<td>483.23</td>
<td>206.63</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>R</td>
<td>518.77</td>
<td>528.77</td>
<td>144.63</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td>88.06</td>
<td>1</td>
</tr>
</tbody>
</table>

**CC78 id 3545**

coord 1074 to 1128
frame 2
ions 20/24

+1 ions labeled
Figure 3.2

MS/MS spectrum of *CCT8* downPeptide. MS/MS spectrum of glutaraldehyde modified downPeptide fragment MAKLLNQLR from gene *CCT8* (id3545); the dnAUG maps 24 codons downstream of the annAUG. The methionine and lysine are both modified by glutaraldehydation (+68 Da). This spectrum has matches to many of the possible CID fragments (19 resolvable +1 ions out of 24; in addition, the b6 fragment was detected among the +2 charged ions *). Matched +1 ion peaks are labeled with designations (a, b and y ions labeled with green, red and blue respectively); the a1 ion (arrowhead) is sometimes enhanced with glutaraldehyde treatment (Russo 2008).
Figure 3.3
Figure 3.3

Frame analysis of dnAUGs. (A–C) Distances (in nucleotides) between dnAUGs and annAUGs in downPeptide genes are graphed for all frames pooled (A) or partitioned by frame (C). Compared to the frequencies of all dnAUGs within 100 nt of the annAUG in the full genome (B), frame-1 downPeptides are over-represented (chi-square $p < 0.001$, all bins $>10$). (D) Lengths of frame-2 or -3 ORFs, defined by dnAUGs of detected downPeptides (188 ORFs, mean 74 nt), are compared with the full set of theoretical ORFs of frame-2 or -3 dnAUGs that start and finish within the annotated frame-1 ORFs (86 088 ORFs, mean 60 nt). ORFs of detected downPeptides are significantly longer based on chi-square analysis of the full distributions (chi-square $p < 2 \times 10^{-7}$), as well as bootstrap analysis of means ($p < 0.002$).
Figure 3.4

A) An xkcd comic showing a flawed scientific study.

B) A webpage displaying a related comic.

C) A letter from the comic's author explaining the flaw.

D) A graph showing the flawed data.

E) A chart illustrating the correct data distribution.

F) A bar chart comparing the two distributions.

G) A graph depicting the correlation between the two sets of data.

H) A histogram showing the frequency of the flawed data.
Figure 3.4

Ribosome profiles, gene expression, and TRII score analysis. (A, B) Ribosome profiles (Ingolia 2009) of annAUGs of annPeptide genes have prominent signals at nt positions 2, 3 and 15, and characteristic depression surrounding position 15. Aligned dnAUGs of downPeptide genes show elevated signal at nt 2 and 15 (arrowheads), and depression surrounding nt 15 (arrow). X-axis values show the 14th nt of ribosome tags corresponding to the middle of the ribosome; position 1 is the A of the AUG start codon. Y-axis values show mean relative tag densities for profiles of 259 annPeptide genes, and 17 downPeptide genes with annAUG TRII scores <8 and >15 nt between the annAUG and dnAUG. Tag densities for each gene were computed relative to their mean tags per nt in the annORF (limited to genes with mean density >0.1 tags per nt, 5% false identification rate for glutaraldehyde-treatment data, and 1% for PeptideAtlas data). (C, D) Ribosome profiles of all rank 1 dnAUGs with annAUG TRII < 8 (1267 genes) show elevated tag densities at nt positions 2 and 15 suggesting that many of these genes initiate translation at their dnAUG. Elevations at nt 2 and 15 are less pronounced for annAUG TRII ≥ 10 (1372 genes). (E–G) DownPeptides genes have depressed mRNA expression (E), protein expression (F) and ribosome densities (G) of their annotated ORFs (solid lines) compared to genes with amino-peptides that map to the annAUG (annPeptide genes; broken lines). mRNA expression and ribosome densities are from deep sequence analysis of Ingolia et al. 2009. Protein expression is of C-terminal TAP-tagged proteins as reported by
Ghaemmaghami et al. 2003. The downPeptide and annPeptide distributions are significantly different in (E-G) (chi-square $p < 0.01$). (H) DownPeptide genes (solid line) have poorer sequence context surrounding their annAUG, measured by Translation Relative Individual Information (TRII score(Robbins-Pianka 2010, Weir 2010)), compared to annPeptide genes (broken line). The distributions of TRII scores plotted for the two gene sets are significantly different (chi-square $p < 0.01$).
Table 3.1

<table>
<thead>
<tr>
<th>Amino Peptide Property</th>
<th>GLUT&lt;sup&gt;a&lt;/sup&gt;</th>
<th>PeptideAtlas&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total N-term peptides&lt;sup&gt;c&lt;/sup&gt;</td>
<td>122</td>
<td>818</td>
</tr>
<tr>
<td>annAUG peptides&lt;sup&gt;d&lt;/sup&gt;</td>
<td>99</td>
<td>519</td>
</tr>
<tr>
<td>downAUG peptides&lt;sup&gt;e&lt;/sup&gt;</td>
<td>23</td>
<td>299</td>
</tr>
<tr>
<td>Acetylated peptides</td>
<td>53</td>
<td>376</td>
</tr>
<tr>
<td>Not acetylated peptides</td>
<td>68&lt;sup&gt;f&lt;/sup&gt;</td>
<td>406</td>
</tr>
<tr>
<td>Acetylated and nonacetylated peptides</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>N-term met cleaved</td>
<td>73</td>
<td>294</td>
</tr>
<tr>
<td>Signal peptide cleaved</td>
<td>8</td>
<td>49</td>
</tr>
</tbody>
</table>
Table 3.1

(a) Glutaraldehyde-treated cell lysates; detection efficiency of N-terminal peptides, compared to internal peptides, was elevated due to glutaraldehyde treatment.

(b) PeptideAtlas publicly deposited data at http://www.peptideatlas.org/repository/

(c) A total of 903 N-terminal peptides were detected in 830 genes; 37 N-terminal peptides were detected in both experimental series.

(d) 583 annPeptides were detected in 559 genes; 22 genes had both Met N-terminal-cleaved and uncleaved peptides, 2 genes had signal-peptide-cleaved and uncleaved peptides.

(e) 320 downPeptides were detected in 299 genes; 14 genes had different downPeptides that mapped to different dnAUGs, and 7 to the same dnAUG (with and without cleavage modifications).

(f) Glutaraldehyde modified peptides.
Chapter 3 Supplemental:

Partial Purification of TAP Tagged dnAUG Candidates
3S.0 Introduction

The screen of glutaraldehyde modified whole-cell yeast lysate and PeptideAtlas data identified a large number of downPeptides, 35% of identified N-termini. Although a 5% False Identification Rate (FIR) was imposed on the data, further confirmation was important to support the findings especially since the frequency of dnAUG initiation was higher than expected. DownPeptides identified in the screen were chosen for partial purification. I used C-terminally TAP-tagged (Rigault 1999) dnAUG candidates, which allowed for purification with magnetic beads coupled to whole-molecule IgG. A second MS/MS identification was of importance as most of the downPeptides were detected only once in the original screen. Partial purification increased the abundance of the candidate protein and with that downPeptides were seen confirming the original data.

Of the 320 dnAUG candidates only some were well suited for partial purification. TAP purification candidates were chosen based on a few criteria.

I. The downPeptide trypsin fragment was >5 and <25 amino acids long to increase likelihood of detection. This is the optimal range for trypsin fragment lengths.

II. There was no trypsin site before the dnAUG methionine. If a trypsin site were present there would be no way to distinguish if the downPeptide was the start of the protein or simply the result of a trypsin cleavage.
III. The protein abundance was as high as possible to yield enough protein for analysis.

IV. The DownPeptide was found in frame 1 to allow for translation of the TAP tag.

3S.1 Results

The candidates were partially purified as described in Section 3.2. 16 candidates were tested, 3 from the glutaraldehyde set, 12 from the PeptideAtlas set and one that was common to both. Five (2 purified by Justin Cherny) candidates had successful confirmation of downPeptides (Table 3S.2). The confidence in these identifications was high as there were many theoretical peaks matched (Figure 3S.1), and high peptide coverage (Figure 3S.2). In these cases, four of the five dnAUG candidates also showed annPeptide fragments as well as the downPeptide fragments. This suggests that these genes have two functioning start sites and the resulting N-terminal peptides are detectable through peptide mass spectrometry. As discussed in Section 1.2.3, this could be evidence of leaky scanning or reinitiation, resulting in initiation at both the annAUG and dnAUG.

The downPeptides of the 11 other candidates were not detected (Table 3S.1). In nine of the candidates the overall candidate protein coverage was <10% and therefore this low coverage may have contributed to the lack of N-terminal detection. In some of these instances the band on the gel, after magnetic bead purification, was weak. This could have been due to lower than expected protein levels, or poor to moderate efficiency of IgG binding to the TAP tag. In addition not all of the
candidate protein is detectable after digesting with trypsin. In some cases fragments are too small or large and this contributes to reduced coverage.

**3S.2 Conclusions and Future Directions**

Five downPeptides were confirmed by partial purification using a C-terminal TAP tag. These results provide confidence that the other downPeptides identified in the screen exist. In the case of *PRS5* and *EFT2* the downPeptides were seen more than once in repeated preps (repeated to confirm protocol). There are several experiments that could be done to yield more information from these data and create new data to expand the repertoire of N-termini.

Candidates with both a annPeptide and downPeptide could serve as possible models to further analyze recoding events such as leaky scanning and reinitiation. It would be valuable to mutate the two start sites in different experiments and use a quantitative Western blot to visualize the frequency of initiation at each site. It would be expected that the detected protein level would be reduced when one of the start sites was nonfunctional since both appear to be used.

Confirmation of more dnAUG candidates could be done by carrying out more purifications. Since some did not seem to purify well with the IgG-magnetic bead protocol, another method such as a size exclusion column could be useful. Lysate would pass through the column and the candidate fraction would be taken based the time point a specific molecular weight would elute. The sample would then be trypsin digested and analyzed with MS/MS.
To identify a new set of N-termini a new protease could be used. Some N-terminal trypsin fragments could not be visualized because of their size. Fragments generated by cleavage with a different enzyme would mean new N-termini would become available for analysis. This would require a new set of glutaraldehyde treated lysate be created.

3S.3 Methods

_Coupling of magnetic beads_

The vial of magnetic beads (Invitrogen) was resuspended in 2 ml of Dimethylformamide (DMF) and stored at 4°C. To couple beads, 200ul of bead suspension was removed from the vial after warming to room temperature. All wash and coupling steps were done by first applying a magnet to separate beads from lysate. After the DMF was removed the beads were washed 3 times in 400ul sodium phosphate buffer with brief vortexing. After washing, 133 ul of sodium phosphate was added to the beads and briefly vortexed. Then 133 ul of whole-molecule rabbit IgG (Jackson) (120 ul sodium phosphate buffer, 13 ul of antibody) was added and the beads were briefly vortexed. Finally, 133ul of 3M ammonium sulfate was added and vortexed with the beads. The bead suspension was then incubated at 37°C for 16-24 hours while rocking. After incubation the beads were washed twice with sodium phosphate buffer and stored at 4°C until use.
3S.4 Figures
Figure 3S.1: Matched Ions of DownPeptides

### EFT2

<table>
<thead>
<tr>
<th>AA</th>
<th>A</th>
<th>B</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>102.04</td>
<td>130.04</td>
</tr>
<tr>
<td>2</td>
<td>V</td>
<td>221.11</td>
<td>229.11</td>
</tr>
<tr>
<td>3</td>
<td>I</td>
<td>314.20</td>
<td>342.19</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>385.23</td>
<td>413.23</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>522.28</td>
<td>550.29</td>
</tr>
<tr>
<td>6</td>
<td>V</td>
<td>521.35</td>
<td>549.36</td>
</tr>
<tr>
<td>7</td>
<td>D</td>
<td>736.39</td>
<td>764.38</td>
</tr>
<tr>
<td>8</td>
<td>H</td>
<td>873.45</td>
<td>901.44</td>
</tr>
<tr>
<td>9</td>
<td>G</td>
<td>991.74</td>
<td>959.46</td>
</tr>
<tr>
<td>10</td>
<td>K</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### PRS5

<table>
<thead>
<tr>
<th>AA</th>
<th>A</th>
<th>B</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>112.04</td>
<td>130.04</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>216.08</td>
<td>244.08</td>
</tr>
<tr>
<td>3</td>
<td>I</td>
<td>357.17</td>
<td>357.17</td>
</tr>
<tr>
<td>4</td>
<td>V</td>
<td>423.24</td>
<td>451.23</td>
</tr>
<tr>
<td>5</td>
<td>V</td>
<td>527.31</td>
<td>555.30</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>674.36</td>
<td>702.37</td>
</tr>
<tr>
<td>7</td>
<td>G</td>
<td>731.40</td>
<td>759.39</td>
</tr>
<tr>
<td>8</td>
<td>D</td>
<td>799.42</td>
<td>816.41</td>
</tr>
<tr>
<td>9</td>
<td>D</td>
<td>933.45</td>
<td>931.44</td>
</tr>
<tr>
<td>10</td>
<td>S</td>
<td>990.48</td>
<td>1018.47</td>
</tr>
<tr>
<td>11</td>
<td>H</td>
<td>1127.54</td>
<td>1155.53</td>
</tr>
<tr>
<td>12</td>
<td>P</td>
<td>1224.59</td>
<td>1252.59</td>
</tr>
<tr>
<td>13</td>
<td>E</td>
<td>1353.60</td>
<td>1381.60</td>
</tr>
<tr>
<td>14</td>
<td>L</td>
<td>1485.72</td>
<td>1494.71</td>
</tr>
<tr>
<td>15</td>
<td>V</td>
<td>1555.75</td>
<td>1593.75</td>
</tr>
<tr>
<td>16</td>
<td>T</td>
<td>1555.83</td>
<td>1594.83</td>
</tr>
<tr>
<td>17</td>
<td>K</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### YJL171C

<table>
<thead>
<tr>
<th>AA</th>
<th>A</th>
<th>B</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>146.05</td>
<td>174.05</td>
</tr>
<tr>
<td>2</td>
<td>L</td>
<td>250.14</td>
<td>278.13</td>
</tr>
<tr>
<td>3</td>
<td>H</td>
<td>396.20</td>
<td>424.19</td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>497.24</td>
<td>525.24</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>596.31</td>
<td>624.31</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>667.35</td>
<td>695.34</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>738.30</td>
<td>766.36</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>825.42</td>
<td>853.41</td>
</tr>
<tr>
<td>9</td>
<td>G</td>
<td>882.44</td>
<td>910.43</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>978.49</td>
<td>1007.48</td>
</tr>
<tr>
<td>11</td>
<td>Q</td>
<td>1107.65</td>
<td>1135.65</td>
</tr>
<tr>
<td>12</td>
<td>S</td>
<td>1194.58</td>
<td>1222.58</td>
</tr>
<tr>
<td>13</td>
<td>Y</td>
<td>1357.65</td>
<td>1385.64</td>
</tr>
<tr>
<td>14</td>
<td>Q</td>
<td>1466.70</td>
<td>1513.70</td>
</tr>
<tr>
<td>15</td>
<td>K</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 3S.1

The possible CID fragments for each of the confirmed downPeptide candidates are listed next to the peptide sequence of the trypsin fragment. Matched ions are labeled in color corresponding to ion type (a, b, or y). EFT2 was modified (|) by glutaraldehyde while PRS5 and YJL171C were acetylated.
### Figure 3S.2: Peptide Coverage Across Candidates

**PRS5**

<table>
<thead>
<tr>
<th>6-10</th>
<th>15-20</th>
<th>25-30</th>
<th>35-40</th>
<th>45-50</th>
<th>55-60</th>
<th>65-70</th>
<th>75-80</th>
<th>85-90</th>
<th>95-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
</tr>
<tr>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
</tr>
</tbody>
</table>

**EFT2**

<table>
<thead>
<tr>
<th>6-10</th>
<th>15-20</th>
<th>25-30</th>
<th>35-40</th>
<th>45-50</th>
<th>55-60</th>
<th>65-70</th>
<th>75-80</th>
<th>85-90</th>
<th>95-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
</tr>
<tr>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
<td>SIRYKWSL</td>
</tr>
</tbody>
</table>
Figure 3S.2

The Sequest peptide coverage output for three dnAUG candidates (analyzed more than once) with high coverage. The candidate amino acid sequence is in black and the red sequence represents identified peptides from that analysis. The yellow bars displayed over the sequence represent the overall coverage for that MS/MS trial, for that candidate. A downPeptide was not confirmed for *YLL041C* however this candidate had high peptide coverage.
Table 3S.1: dnAUG Candidates for Partial Purification

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Standard name</th>
<th>Systematic name</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glutaraldehyde</td>
<td><em>BUD6</em></td>
<td>YLR319C</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Glutaraldehyde</td>
<td><em>MDJ2</em></td>
<td>YNL328C</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Glutaraldehyde</td>
<td><em>RPN13</em></td>
<td>YLR421C</td>
<td>37</td>
</tr>
<tr>
<td>Glu/Atlas</td>
<td><em>EFT2</em></td>
<td>YDR385W</td>
<td>80</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>ATP2</em></td>
<td>YJR121W</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>DOT1</em></td>
<td>YDR440W</td>
<td>10</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>MEP1</em></td>
<td>YGR121C</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>QCR10</em></td>
<td>YHR001W-A</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>UTP18</em></td>
<td>YJL069C</td>
<td>12</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>YEF1</em></td>
<td>YEL041W</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>YJL171C</em></td>
<td>YJL171C</td>
<td>26</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>SDH2</em></td>
<td>YLL041C</td>
<td>20</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>BAG7</em></td>
<td>YOR134W</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>SFG1</em></td>
<td>YOR315W</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>YPL162C</em></td>
<td>YPL162C</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Atlas</td>
<td><em>PRS5</em></td>
<td>YOL061W</td>
<td>25</td>
</tr>
</tbody>
</table>
Table 3S.1

A table of the dnAUG candidates chosen for partial purification. Candidates were chosen from both the glutaraldehyde dataset as well as the PeptideAtlas data. The peptide coverage of each candidate ranged from <10%-80%. Some candidates were tested once while others were tested several times.
**Table 3S.2: Successful DownPeptide Confirmations**

<table>
<thead>
<tr>
<th>Gene</th>
<th>id</th>
<th>annPeptide Match</th>
<th>d.nAUG start</th>
<th>Frame</th>
<th>Modification</th>
<th>MS/MS detected peptide</th>
<th>Peptide count</th>
<th>Xcorr</th>
<th>DeltaCn</th>
<th>Ions</th>
<th>PAc</th>
</tr>
</thead>
<tbody>
<tr>
<td>YJL171C</td>
<td>3707</td>
<td>Yes</td>
<td>1031</td>
<td>1</td>
<td>no mod</td>
<td>-.MFMLHTVAASGPQSYQK.L</td>
<td>1</td>
<td>1.73</td>
<td>0.21</td>
<td>21/64</td>
<td>88</td>
</tr>
<tr>
<td>PRS5</td>
<td>5422</td>
<td>Yes</td>
<td>1007</td>
<td>1</td>
<td>Acet</td>
<td>-.SNIVVIQGDLYELVLTVK.L</td>
<td>2</td>
<td>3.55</td>
<td>0.53</td>
<td>19/32</td>
<td>145</td>
</tr>
<tr>
<td>HOX3</td>
<td>854</td>
<td>Yes</td>
<td>1007</td>
<td>1</td>
<td>Acet</td>
<td>-.MDFQIPIT88SNWWVQK.L</td>
<td>2</td>
<td>2.90</td>
<td>0.43</td>
<td>18/30</td>
<td>324</td>
</tr>
<tr>
<td>URA1</td>
<td>1699</td>
<td>No</td>
<td>1079</td>
<td>1</td>
<td>no mod</td>
<td>-.TTQELDELANSK.A</td>
<td>1</td>
<td>4.13</td>
<td>0.53</td>
<td>19/22</td>
<td>89</td>
</tr>
<tr>
<td>EFT2</td>
<td>2793</td>
<td>Yes</td>
<td>1064</td>
<td>1</td>
<td>no mod</td>
<td>-.SIVAHVHDHGK.S</td>
<td>13</td>
<td>3.03</td>
<td>0.44</td>
<td>16/18</td>
<td>324</td>
</tr>
</tbody>
</table>
Table 3S.2

Table of dnAUG candidates whose downPeptides were confirmed by partial purification. All except URA1 showed an annPeptide as well as the downPeptide. Two of the five downPeptides begin with Met while three have the N-terminal Met cleaved and all peptides end in K due to a trypsin cleavage event. The downPeptides of URA1 and YJL171C were only seen once in the initial screen while EFT2, HOM3, and PRS5 were seen more than once before purification. All of these successful partial purification candidates came from the PeptideAtlas data (PAe designates trial number) and were analyzed with Sequest.
Chapter 4

Gel Slice Approach for Algorithm Assessment
The work discussed in Chapter 4 was a collaboration with Masters student Miin Lin. I generated the gel slice protocol and carried out the glutaraldehyde-treated trials and a portion of the gel slice trials. Miin Lin carried out the majority of gel slice trials and wrote the stored procedure to analyze conformance scores. She also carried out the analysis of Sequest and OMSSA parameter sets discussed in the results. Dr. Michael Weir and Justin Cherny also contributed to the stored procedure.
Abstract

The availability of many peptide identification search algorithms has been a benefit to the field of proteomics but has added uncertainty in that different algorithms identify different peptides. To compare the algorithms Sequest and OMSSA, MS/MS spectra from proteins of a known molecular mass range were analyzed by both algorithms, neither having knowledge of the size range. 84.4% of the peptides identified by Sequest conformed to the correct molecular weight range while the five parameter sets of OMSSA had conformance between 87.6 to 88.8 %. This conformance scoring method can be used to compare different algorithms as well as different algorithm permutations in the form of parameter settings.
4.0 Introduction

As peptide mass spectrometry (MS/MS) technology advances we are able to analyze more proteins with higher resolution. To identify the dnAUG candidates in Chapter 3 a large set of spectra was generated and an even larger set of spectra was downloaded (PeptideAtlas). With this abundance of mass spectrometry data there is a need for accurate, efficient, and reliable peptide identification algorithms. These algorithms are needed to make peptide identifications from the most abundant peptides to the least abundant while filtering out noise. With many algorithms available, both open source and proprietary, it is important to be able to assess an algorithm’s accuracy in peptide identifications. One way to do this is to compare their ability to successfully search the same dataset. The advances in mass spectra generation are only of use if one can accurately identify peptides from the data and have confidence in the output.

4.1.1 Introduction: Decoy Method

The most accepted method for assessment of assigned peptide matches is the decoy method. For every forward protein sequence a reversed version of that sequence, called the decoy, is added to the database the algorithm searches to identify peptides (Elias 2007, Fitzgibbon 2008). These sequences do not occur in the proteome and therefore they provide a gauge of confidence by the number chosen as real matches by the algorithm. A threshold of a match quality score, in this case a probability score, can be used to achieve a false identification rate (FIR) of 5%. The decoy method can be applied to individual or sets of MS/MS trials and provides a
confidence that the majority (95%) are actual real matches. However this does not
provide complete confidence and leaves one with questions such questions as: Does
this mean we have confidence in peptides identified by one algorithm but not
another? Do we have confidence in peptides that are found only when certain
parameters are applied? As discussed in Section 1.4.2, algorithms have different
approaches to peptide identifications. This includes different parameters, which often
result in different peptide identifications. The FIR is crucial when filtering matches
within an MS/MS experiment but less informative when comparing peptides
identified across algorithms.

4.1.2 Experimental Justification

To begin a mass spectrometry experiment, proteins must first be digested with
a protease such as trypsin. This reduces the collection of proteins to a collection of
peptides, which can be analyzed by the mass spectrometer. After analysis with
Sequest or OMSSA, peptides from this collection are identified as well as their parent
protein.

Analysis of parent proteins was first suggested by an unexpected observation
in our earlier analysis of downPeptides. These parent proteins became of interest
when analyzing partially purified proteins (Chapter 3S). The dnAUG candidate
proteins were excised from a gel (base on size), trypsin digested, and analyzed with
MS/MS. It was expected that the proteins identified along with the dnAUG candidate
would be of a similar molecular weight since they had been removed from the same
area of the gel. In cases when the dnAUG candidate was successfully identified, the
surrounding matches seemed to correspond to proteins of a similar size. In the cases when no dnAUG candidate was found, and the surrounding proteins were from a variety of sizes, the data appeared to be poor by other measures. Conversely if no dnAUG candidate protein was identified but the surrounding matches were of the expected size, it suggested that the data was sound but for another reason (preparation error, etc.) the dnAUG candidate protein was not found.

The size of surrounding proteins became a way of validating the individual mass spectrometry runs in a somewhat simple way. In this chapter I will discuss how this method was scaled up as a means to assess Sequest and OMSSA and compare their ability to make accurate peptide matches.

4.1.3 Experimental Strategy

Whole-cell yeast lysate was separated with SDS-PAGE and sections of gel were excised based on protein size ranges. The samples were trypsin digested and the peptides in each gel slice analyzed with MS/MS. The spectra were analyzed with search algorithms Sequest and OMSSA, which had no knowledge of the input protein size range. The number of identified peptides whose parent proteins fit into the specified size range was used to calculate a conformance score. The conformance score of each algorithm can be used for comparison between algorithms and parameters.

4.1.4 Discussion of N-terminal Modification with Glutaraldehyde

With the success of the glutaraldation of N-termini to enhance detection of al ions from Section 3.1, it was considered for implemented again for this analysis. The
goal would be to generate conformance scores as described above but also to generate a set of dnAUG candidates. The addition of the dnAUG candidates would do two things.

It would increase the total number of dnAUG candidates to further expand our collection of identifications. In the previous analysis of dnAUGs, a large portion of the proteome was examined in each MS/MS experiment. Since this experiment would partition the lysate by size it was expected that this may yield a set of lower abundance candidates, including downPeptides, since there would be a less complex protein mixture analyzed in each experiment.

A set of dnAUGs from this type of experiment would also provide increased confidence in the set since there would be the additional confidence measure from the conformance score. Although the decoy method and partial purification of select dnAUG candidates strengthened the dnAUG identifications, further justification is desirable.

The glutaraldehydation of this lysate varied from the method in Section 3.2 however and was not successful because this process had to occur in-gel. This became problematic as glutaraldehyde crosslinks when exposed to acrylamide (Russo 2008). The peptides needed to be modified in-gel however this property of glutaraldehyde made it impractical to remove them from the gel for MS/MS analysis. The few peptides that were removed successfully were less consistently glutaraldehyde modified and therefore were not useful for this analysis. Therefore the gel slice experiments described below were not glutaraldehyde modified.
4.1.5 Implementation of Gel Slice Method

A total of 22 mass spectrometry runs were done to generate spectra for analysis (Table 4.1). Of the 22 experiments there were three size ranges represented; 25-37 kDa, 37-50 kDa, and 50-75 kDa (Figure 4.1). The peptides identified were labeled as conforming if the parent protein was within the molecular weight range or nonconforming if the parent protein did not fit into the molecular weight range. The conformance score was calculated by taking the number of conforming peptides over the total number of distinct peptides (conforming and nonconforming).

4.1.6 Experimental Uncertainties

It was important to consider that nonconforming proteins would be identified independent of algorithm failure because of these properties of the experiment:

I. Posttranslational modifications that are not specified within the search parameters will cause some proteins to be out of range because the predicted molecular weight of a protein is solely based on sequence. Modifications such as protein splicing (Shao 1997, Kawasaki 1997) will cause a protein to be found in a lower size range despite the predicted non-spliced molecular weight.

II. Proteins may run aberrantly in SDS-PAGE causing some to not settle at the predicted molecular weight. This is due to the charge of amino acids, areas of high disorder, and interactions with certain detergents (Ilakoucheva 2001, Stevens 1999). Therefore an aberrant protein
would register as nonconforming because its position on the gel is not in line with its molecular weight.

III. The gel slice method uses a razor to excise the bands. Although caution is exercised to cut in the same places for each range, there may be some researcher error.

In addition to the physical limitations of the gel slice method we needed to acknowledge that the algorithms will identify false matches, which just like real matches will have a parent protein with a molecular weight. Therefore some of both the conforming and nonconforming peptide matches will be false and contribute to the conformance score.

These four factors were present and more importantly very difficult to quantify. They are shared across the size ranges and different gels and are independent of algorithm choice. The conformance score calculation (Figure 4.2) is not a direct measure of algorithm accuracy but instead it is a measure of comparison between algorithms and parameter sets.

4.2 Results

The 22 gel slice experiments, used together as a set, were analyzed first with Sequest, then OMSSA with the parameters in Section 4.5. A stored procedure was used to calculate the number of peptides that were found to conform to the expected molecular weight size range and the resulting conformance score. The peptides, counted once per mass spectrometry trial, were first filtered for a FIR of (5%) based on probability score. Sequest does not list the molecular weights of the parent
proteins and therefore they needed to be calculated based on the sequence of the identified parent protein. Once the molecular weights were computed within the stored procedure the script identified the conforming peptides from the nonconforming peptides. To account for some variability in gel excision and post-translational modifications, a 10% mass tolerance was applied to each size range. For the 25-37 kDa group the actual size of conformance was 22.5 to 40.7 kDa (Table 4.2).

Sequest had an overall conformance score of 84.4%. OMSSA, with the 5 parameter sets, had a conformance score range of 87.3% to 88.9% (Table 4.2). Both Sequest and OMSSA showed similarly high conformance scores. As discussed in Section 4.1.5, part of the reason for nonconforming proteins may be because of posttranslational modifications that impart a size change to the protein.

We chose to analyze the conforming and nonconforming peptides for possible glycosylation sites as a means to assess if this modification could explain some of the non-conformance seen. In yeast there are N-linked glycosylation and O-linked glycosylation sites (Kung 2009). In the latter, glycans are attached to serines and threonines and by calculating the percent serine and threonine of proteins, a measure of possible glycosylation sites was obtained. The nonconforming set of peptides did have a slightly higher percentage of serine and threonine than the conforming set of peptides (Figure 4.3). However when the nonconforming peptides were separated into two groups, those that were larger than the size range (above) or those who are smaller (below), the slight increase in threonine and serine percentage was found in the above range set. Glycosylation would make sense in the below range
nonconforming set because it would add mass causing smaller proteins to run at a higher molecular weight. However glycosylation in the above range nonconforming set does not help to describe why some higher molecular weight proteins are found running at a smaller size range. Although there was no strong evidence for this one posttranslational modification, I hypothesize that modifications are contributing in some manner to protein location in gel.

As mentioned, Sequest was run with one set of parameters due to slow running time. However OMSSA, run on a cluster, allowed 5 different parameter sets to be tested. While each gave a high conformance score (range 87.3% to 88.9 %), only 39% of peptides were found in all 5 parameter sets. That suggests that each yields novel matches and when the union of all the OMSSA parameter sets is taken, a more complete picture of peptides present in a sample can be achieved. In addition peptides only found in one parameter set tended to have a lower conformance score suggesting that representation across parameter sets can provide another measure of confidence, however overlap between all parameter sets is not necessary.

With the information that different search parameters identified different peptides we used our knowledge of collision induced dissociation (CID) to expand the peptide search. We had included a-type ions in the parameter sets in Chapter 3 to enhance detection of N-termini in addition to the standard b and y ions. We asked what would happen to the peptide pool and resulting conformance score if a ions were calculated as well as b and y. Peptides identified with Sequest had higher conformance scores when they were found in both the b,y, and a,b,y, screen
compared to peptides found in one or the other. The non-overlap peptides showed lower protein expression than the overlap set, possibly contributing to their low conformance scores. Conversely, peptides identified by OMSSA had less overlap between the two ion parameters and those that did not overlap (the majority) had high conformance scores. In addition protein expression was higher for the overlap set unlike the Sequest results. This suggests that both algorithms identify peptides differently, as seen by the overlap of 54% of identified peptides between the two algorithms. This conformance analysis has shown that we can have confidence in both algorithms ability to identify peptides. However they appear to function differently and that means their results should be handled differently.

4.3 Conclusions and Future Directions

The overall conformance scores for Sequest, 84.4 % and OMSSA, 87.3-88.9%, suggest that both algorithms are making successful peptide identifications. Peptides identified by both algorithms are of high confidence but only represent 53.3 % of the detected peptides. The gel slice experiments demonstrate that peptides identified by only one algorithm should also be considered especially within a 5% FIR setting. Algorithms may have only modest overlap because they make matches differently resulting in non-overlapping peptide identifications. In the case of Sequest, running the data with a series of reasonable parameters and selecting the overlap may be the best way to increase confidence in peptides. Multiple parameter analysis was done with OMSSA; however taking the union of matches instead of the overlap gave the highest confidence matches. The gel slice analysis approach is a practical method
for users of peptide identification algorithms to validate their algorithms performance. The spectra files will be made available for download which will allow other research groups to analyze their own parameters and algorithms. The gel slice experiment itself can also be easily repeated to generate spectra for those who use an alternative form of peptide bond dissociation or m/z detection.

To expand the analysis it would be beneficial to run the gel slice spectra set through other algorithms. In addition to Sequest and OMSSA, Mascot (Matrix Science) is a commonly used algorithm whose conformance assessment with the gel slice data would be informative and a valuable comparison. The software PEAKS (Bioinformatic Solutions Inc.) would also be another candidate for comparison. This would be especially interesting as PEAKS can act as either a de novo or database search algorithm. Not only would we be able to compare PEAKS to other algorithms but an analysis would allow a comparison of its two identification modes.

In addition to an expanded investigation of algorithms, future wet lab work could be done to elucidate if posttranslational modifications are impacting the nonconformance rate of parent peptides. Since glycosylation is such a common posttranslational modification this could be a good assessment. Some gel slice trials would be repeated and this time a step including enzymatic removal of glycans would be included. The resulting conformance score would indicate if the number of nonconforming peptides was reduced.

4.4 Methods

Protein Sample Preparation
100ml of YSH474 cells were grown to mid-log phase in YPD and lysed with RIPA buffer (150mM NaCl, 1% Igepal, 0.1% SDS, 50mM Tris pH 8.0) and acid washed glass beads. To prevent degradation, protease inhibitors (Roche) and 500 ul of 17 mg/ml PMSF were added to the lysis buffer. The sample was lysed with a vortex by alternating 15 seconds of vortexing with 30-60 seconds of cooling in ice. The lysate was spun at 5,000 RPM after confirmation of lysis. Samples of 500 ug or 1000 ug were run (at 4°C) on 4-20% SDS-PAGE gels (Bio-Rad) with a protein standard marker run next to each sample (Bio-Rad) to serve as a guide for excision of sample bands (Shevchenko 2007). Molecular weight size ranges (25-37, 37-50, and 50-75 kDa) were selected for analysis. After excision, bands were reduced with 10mM DTT for 30 min at 56°C and alkylated with 55mM iodoacetamide for 30 minutes at room temperature. A trypsin solution of 13ng/ul was made in 10mM ammonium bicarbonate and was incubated with the acetonitrile dehydrated gel pieces for 1.5 hours until full absorption. The gel pieces were incubated overnight in a 37°C water bath for optimum trypsin digestion. The peptides were extracted with a ratio of 1:2 formic acid and acetonitrile and vacufuged to complete dryness. After resuspension in 0.1% TFA the peptides were loaded onto a c18 (Michrom) packed nanospray column (Polymicro). Samples were run with a 180 minute gradient on a LCQ Deca XP (Thermo). Gels with visible degradation had limited conformance to parent protein masses and therefore it was important to reduce degradation by keeping samples cold and using protease inhibitors. Gels were discarded and not analyzed if their appearance suggested visible degradation.
Algorithm analysis

Sequest was run using the Proteome Discoverer 1.2 Platform and OMSSA was run using the Wesleyan maintained cluster. For both algorithms we included the following optional mass increases to peptides: (i) N-terminal amino acid +42 Da for acetylation modification (endogenous); (ii) N-terminal methionine oxidation +16 Da (endogenous). A static modification of cysteine (carbamidomethyl +57 (from reduction/alkylation)) was also included. The other parameters are organized below by algorithm and in the case of OMSAA by run.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sequest</th>
<th>OMSSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum missed cleavage sites</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Precursor mass tolerance</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Fragment mass tolerance</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Precursor Ion Search Type</td>
<td>mono</td>
<td>avg</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>avg</td>
</tr>
<tr>
<td>Fragment Ion Search Type</td>
<td>mono</td>
<td>mono</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>mono</td>
</tr>
<tr>
<td>Mass tolerance charge scaling</td>
<td>N/A</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>none</td>
<td>none</td>
</tr>
</tbody>
</table>

Stored Procedure: Miin Lin

The stored procedure was written in SQL and began by setting the probability threshold. The Sequest b/y ion screen had a probability value threshold of 0.081 and the a/b/y ion screen had a value of 0.164. For both the b/y and a/b/y ion screen done in OMSSA the probability threshold value was 1 except for implementation 2 in the a/b/y screen. All of the peptides identified within the threshold were pulled into a table where any peptides that matched to more than one parent protein were removed. The molecular weights of the parent proteins of the remaining peptides were calculated based from the protein sequence. Each protein was queried to see if its molecular weight fit into the specified size range inputted by the user. The peptides...
were then uploaded to the database and the conformance score calculated where each peptide was counted once per MS/MS trial.
4.5 Figures
Figure 4.1: Gel Slice Size Ranges

Lysate  Ladder

- 50-75 kDa
- 37-50 kDa
- 25-37 kDa
Figure 4.1

An example of a SDS-PAGE gel, trial 2. The three size ranges are outline. The gel pieces were cut based on the marker lanes and then cut into smaller pieces for trypsin digestion (Shevchenko 2007)
Figure 4.2: Peptide Match Types and Scoring Uncertainties
Figure 4.2

1. Parent proteins are classified as conforming or nonconforming based on their size (kDa). There are different classifications of matches in each set contributing to uncertainty. The conforming set will contain protein matches to the correct size range and it will also contain false matches that are the result of inaccurate matching by the peptide search algorithm. The nonconforming set will also contain false matches as well as real matches that are found outside of their size range because of aberrant running or posttranslational modification.
Figure 4.3: Glycosylation Analysis

1

![Percent Serine and Threonine of Conforming and Nonconforming Peptides]

2

![Percent Serine and Threonine in Nonconforming Peptides]
**Figure 4.3**

1. Nonconforming peptides have a slight increase in the percent of serines and threonines out of the total number of amino acids when compared to conforming peptides. 2. When the nonconforming peptides are broken down into those that are above the molecular mass range (above) and below the range (below) the above set has the highest percent of serines and threonines.
**Table 4.1: Individual Gel Slice Trials: BY Screen**

<table>
<thead>
<tr>
<th>MS Trial</th>
<th>Conforming peptides</th>
<th>Nonconforming peptides</th>
<th>FIR (%)</th>
<th>Conformance score</th>
<th>SizeRangeMin</th>
<th>SizeRangeMax</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>148</td>
<td>23</td>
<td>4.1</td>
<td>86.5</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>82</td>
<td>12</td>
<td>8.5</td>
<td>87.2</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>3</td>
<td>146</td>
<td>31</td>
<td>4.5</td>
<td>82.5</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>385</td>
<td>111</td>
<td>1.8</td>
<td>77.6</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>129</td>
<td>18</td>
<td>6.1</td>
<td>87.8</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>182</td>
<td>21</td>
<td>6.4</td>
<td>89.7</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>7</td>
<td>160</td>
<td>21</td>
<td>4.4</td>
<td>88.4</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>131</td>
<td>21</td>
<td>3.3</td>
<td>86.2</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>214</td>
<td>20</td>
<td>1.8</td>
<td>91.5</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>10</td>
<td>116</td>
<td>21</td>
<td>10.2</td>
<td>84.7</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>11</td>
<td>294</td>
<td>67</td>
<td>1.4</td>
<td>81.4</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>12</td>
<td>324</td>
<td>56</td>
<td>1.8</td>
<td>85.3</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>13</td>
<td>137</td>
<td>12</td>
<td>2.0</td>
<td>91.9</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>14</td>
<td>90</td>
<td>14</td>
<td>2.9</td>
<td>86.5</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>15</td>
<td>92</td>
<td>13</td>
<td>10.5</td>
<td>87.6</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>16</td>
<td>88</td>
<td>12</td>
<td>8.0</td>
<td>88.0</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>17</td>
<td>134</td>
<td>15</td>
<td>6.0</td>
<td>89.9</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>18</td>
<td>115</td>
<td>22</td>
<td>5.1</td>
<td>83.9</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>19</td>
<td>324</td>
<td>52</td>
<td>2.7</td>
<td>86.2</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>20</td>
<td>158</td>
<td>9</td>
<td>0.6</td>
<td>94.6</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>21</td>
<td>131</td>
<td>9</td>
<td>6.4</td>
<td>93.6</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>22</td>
<td>224</td>
<td>40</td>
<td>3.0</td>
<td>84.8</td>
<td>37</td>
<td>50</td>
</tr>
</tbody>
</table>
Table 4.1

The list of 22 gel slice trials and their individual conformance scores. Nine trials were taken from the 25-37 kDa size range, eight were from 37-50 kDa, and five were from 50-75 kDa.
### Table 4.2: Overall Conformance Scores for Sequest and OMSSA

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Distinct peptides</th>
<th>Conforming peptides</th>
<th>Nonconforming peptides</th>
<th>Conformance score</th>
<th>FIR(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequest</td>
<td>4480</td>
<td>3781</td>
<td>699</td>
<td>84.4</td>
<td>6.0</td>
</tr>
<tr>
<td>OMSSA-0</td>
<td>3060</td>
<td>2717</td>
<td>343</td>
<td>88.8</td>
<td>2.4</td>
</tr>
<tr>
<td>OMSSA-1</td>
<td>3644</td>
<td>3196</td>
<td>448</td>
<td>87.7</td>
<td>3.6</td>
</tr>
<tr>
<td>OMSSA-2</td>
<td>2134</td>
<td>1893</td>
<td>241</td>
<td>88.7</td>
<td>5.4</td>
</tr>
<tr>
<td>OMSSA-3</td>
<td>3295</td>
<td>2887</td>
<td>408</td>
<td>87.6</td>
<td>3.9</td>
</tr>
<tr>
<td>OMSSA-4</td>
<td>3035</td>
<td>2696</td>
<td>339</td>
<td>88.8</td>
<td>2.2</td>
</tr>
</tbody>
</table>
Table 4.2

The overall conformance scores for the b/y screen of Sequest and OMSSA. The conformance score was calculated by dividing the total number of peptides by the number of conforming peptides. The FIR is within 2.2-6.0 % for each trial since one probability threshold was used for all trials.
Chapter 5

Conclusions and Future Directions
The work in this thesis began as a means to identify genes that were potentially under-annotated. As this hypothesis was confirmed with successful mutagenesis experiments, our approach changed to accommodate more identifications of under-annotation through peptide mass spectrometry. Identification of a substantial number of downPeptides demonstrated that translation at downstream sites is more common in budding yeast than expected. With this conclusion we realized there was a need to better understand our methods of assessment and verification, including the generation of a new means for assessing peptide identification algorithms.

Bioinformatic evidence described in Chapter 2 suggests that some budding yeast genes are under-annotated. An analysis of mRNA sequences aligned at the annAUG found that mRNAs with 5’ UTR upAUGs tended to have lower information at positions -3 and +5 (Michael Weir), suggesting that some of the annAUGs are mis-annotated or under-annotated. Initiation at upstream sites is further supported by the presence of detected peptides mapping to sequence upstream of the annotated protein start (unpublished data) and ribosome profiles (Ingolia 2009) that detect 80 S translating ribosomes in the 5’ UTRs of some yeast mRNA transcripts. Mutagenesis experiments produced striking results when protein was detected from upAUG candidate genes with mutant annAUGs. The detection of protein strongly suggests that these candidates use an upAUG or upstream noncanonical site to initiate translation.
Investigation of upstream initiation is of great importance as it will provide a more complete understanding of translation initiation mechanisms, but more importantly it can provide information about the function and translational mechanism of how some genes are translated and regulated. Kozak, analyzing a set of eukaryotic mRNA sequences, found that only a small fraction contained upAUGs. However, a dramatic increase in the presence of at least one upAUG was seen in the mRNAs of proto-onco genes, transcription factors, and receptor proteins (Kozak 1987, 1991, Morris 2000). Initiation at an upAUG within these transcripts can result in the translation of a short upORF that reduces initiation at the annAUG (Kozak 1991) through competition for competent ribosomes. The resulting reduction in translation of the annORF is a mechanism of translational regulation. The use of upAUGs may play an important regulatory role in genes where transcriptional and translational control is crucial, such as genes that are involved in the cell cycle. A refined annotation that identifies functional upAUGs would greatly enhance our understanding of how certain genes are regulated.

The MS/MS identification of N-termini has proved to be a powerful method for identifying cases of dnAUG initiation. This screen could be expanded to identify translation initiation sites resulting from other non-annAUG sites of initiation. It would be especially interesting to expand on the analysis of upAUG and non-AUG initiation. An analysis could be done with the existing glutaraldehyde-treated lysate and PeptideAtlas spectra used in Chapter 3. Genes with putative upAUG initiation sites were identified in Chapter 2 and a screen for upPeptides would only increase the
number of identified upstream initiation sites. The high frequency of dnAUG initiation was unexpected, and could not have been seen if candidates were analyzed initially on an individual basis.

Although a screen of MS/MS spectra for non-standard protein translation is a powerful approach, the results are even more convincing when paired with molecular genetic approaches. Partial purification of upAUG candidates with a C-terminal TAP tag provides confirmation that translation does initiate upstream with detection of the identified upPeptides. In other experiments, the mutagenesis of a candidate’s annAUG can confirm the use of an upAUG if protein is detected and found at the correct size. If the upAUG is used only some of time, a quantitative Western blot can be used to measure the abundance of protein from upAUG initiation.

The large-scale screen of glutaraldehyde-modified lysate and PeptideAtlas spectra, described in Chapter 3, identified many candidates for dnAUG initiation. Of the detected N-termini, a larger than expected number were found to result from dnAUG initiation. Ribosome profile data, TRII score, and CAI analysis (Justin Cherny) further supported the conclusion that these dnAUGs are functional as translation initiation start sites. The increased frequency of frame-1 detected downPeptides close to the annAUG suggests that there is some selection for these dnAUGs. It is possible that the presence of a close, in-frame dnAUG essentially provides two start sites for ribosomes to initiate from (Kochetov 2005, Kozak 2002). In exchange for the slight N-terminal truncation, protein production is increased; a tradeoff that is worthwhile if the N-terminus has no functional role. This is especially
necessary if the annAUG is in sub-optimal context, which may cause some fraction of ribosomes to bypass it.

In cases like the translation of Histone H4 in yeast, the N-terminus has a functional role (Kayne1988) and a rigidly defined annAUG is crucial. A truncated N-terminus, resulting from initiation at a dnAUG, could impart an alternate function to a protein. In addition, initiation at a dnAUG could serve to produce a truncated protein which may remain in the cytosol while the longer version may localize to the ER, mitochondria, or nucleus due to a signal sequence (Bugler 1991, Xiao 2000, Emanuelsson 2000) (Supplemental figures S7 and S8, Appendix).

Moving further downstream from the annAUG, the presence of frame-2 and frame-3 dnAUGs increases. While frame-1 downPeptides were analyzed in Chapter 3 as the most common of the detected downPeptides, frame -2/3 may have some cellular role. Analysis of these peptides (Justin Cherny) showed an increase in frame-2/3 peptide length compared to non-detected theoretical frame-2/3 sequences (Figure 3CD), suggesting that they may have some functionality. Preliminary analysis by Justin Cherny, Miin Lin, and Mary Vallo suggests that sets of frame-2/3 peptides may have signal sequences that facilitate localization to cellular membranes such as to the endoplasmic reticulum (ER). Once within the ER lumen the peptides can be folded and transported to the plasma membrane via a secretory pathway (Ron 2007) or secreted out of the cell. Whether membrane bound or released they may act as a signal to neighboring cells (Turely 2000) similar to what is seen with antigen presentation in the immune system (Watts 1985). These peptides may also have some
role as cell-penetrating peptides (CPPs) which facilitate the entry of protein and other molecules into cells (Richard 2003). Frame-2/3 peptides that remain within the ER lumen may have a regulatory role, for example in the ER’s Unfolded Protein Response (UPR) which causes protein production to slow followed by reduced transport of proteins into the ER (Ron 2007). To enrich for frame-2/3 peptides, cell fractionation could be used to isolate membrane bound proteins for MS/MS. We could expect to see an increase in frame-2/3 peptides when compared to a cytoplasmic fraction, and evidence of different isoforms (down and annotated) of the same protein in the different fractions.

In both the data from the glutaraldehyde modified set and the PeptideAtlas, the number of N-termini detected only represented a fraction of the total number of budding yeast genes. DnAUG initiation appears to be more common than expected and an increase in total N-termini detected would be very valuable in terms of a final assessment of dnAUG initiation frequency. This could involve fractionation of glutaraldehyde-modified lysate for better coverage and experimentation with different algorithms, different parameter settings, and posttranslational modification specification. Any new parameter setting can be assessed first with gel slice spectra to test their ability to contribute to accurate peptide identifications.

An analysis in other organisms would also contribute to an understanding about the frequency and mechanism of non-standard protein translation. A screen for nonstandard translation sites could be very informative in Drosophila melanogaster, especially since the glutaraldehydation of N-termini could be done and there are
publically available spectra available (www.PeptideAtlas.org). Sampling of different developmental tissues could provide a more complete picture of N-terminal peptides with the possibility that upAUG and dnAUG translation may be used at different stages of development or within different tissues.

With the surprising number of dnAUG initiation sites identified and the bioinformatic evidence supporting both dnAUG and upAUG initiation, this suggests that more and more cases will be identified. Ribosome profile data (Ingolia 2009) has described non-standard protein translation on a larger scale, including the presence of nonAUG initiation sites. Translational regulation, controlled at the level of the ribosome and the mRNA, can provide processes to promote or inhibit protein production in a number of ways including protein start site selection, mRNA secondary structure, and mRNA binding proteins (Jackson 2010). The use of alternate initiation sites within a gene provides yet another layer of regulation that instead modulates the sequence of the resulting protein, potentially giving rise to proteins with different functions and localizations. The use of alternate initiation sites increases a cell’s ability to regulate protein product through start site selection and may be as powerful as transcriptional regulation for determining the sequence of a protein.
References


