Translation initiation: a regulatory role for poly(A) tracts in front of the AUG codon in *Saccharomyces cerevisiae*

Xuhua Xia¹, Vivian MacKay², Xiaoquan Yao¹, Jianhua Wu³, Fumihito Miura⁴, Takashi Ito⁴, David R. Morris²

1. Department of Biology, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, Canada, K1N 6N5. Fax: (613) 562-5486, E-mail: xxia@uottawa.ca

2. Department of Biochemistry, UW Box 357350, 1705 NE Pacific St. Seattle WA 98195-7350, USA

3. Institute of Biomechanics and Department of Biomedical Engineering, School of Bioscience and Bioengineering, South China University of Technology, Guangzhou, China

4. Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-Ku, Tokyo 113-0032, Japan

Running title: Pre-AUG poly(A) length and translation initiation

Keywords: yeast, translation initiation, poly(A), ribosomal loading, protein production, 5’ UTR, vaccinia virus
ABSTRACT

5’ UTR serves as the loading dock for ribosomes during translation initiation and is the key site for translation regulation. Many genes in the yeast (Saccharomyces cerevisiae) contain poly(A) tracts in their 5’ UTRs. We studied these pre-AUG poly(A) tracts in a set of 3274 recently identified 5’ UTRs in the yeast to characterize their effect on in vivo protein abundance, ribosomal density and protein synthesis rate in the yeast. The protein abundance and the protein synthesis rate increase with the length of the poly(A), but exhibit a dramatic decrease when the poly(A) length is 12 or greater. The ribosomal density also reaches the lowest level when the poly(A) length is 12 or greater. This supports the hypothesis that a pre-AUG poly(A) tract can bind to translation initiation factors to enhance translation initiation, but a long (≥12) pre-AUG poly(A) tract will bind to Pab1p, whose binding size is 12 consecutive A residues in yeast, resulting in repression of translation. The hypothesis explains why a long pre-AUG poly(A) leads to more efficient translation initiation than a short one when PABP is absent, and why pre-AUG poly(A) is short in the early genes but long in the late genes of vaccinia virus.
INTRODUCTION

Poly(A) tracts in 5’ UTR have been recognized recently as important sites for translation regulation. These poly(A) tracts, referred to hereafter as pre-AUG A_N where N stands for the number of consecutive A nucleotides, can interact with translation initiation factors or poly(A) binding proteins (PABP) to either increase or decrease translation efficiency. Pre-AUG A_N can enhance internal ribosomal entry both in the presence of PABP and eIF-4G in the yeast, *Saccharomyces cerevisiae* (GILBERT et al. 2007) and in the complete absence of PABP and eIF-4G (SHIROIKH and SPIRIN 2008). Translation initiation factors eIF-4B and eIF-4F can bind to poly(A) tracts (GALLIE and TANGUAY 1994), and exogenous poly(A) added to an *in vitro* translation system can inhibit translation initiation (GROSSI DE SA et al. 1988; JACOBSON and FAVREAU 1983; LODISH and NATHAN 1972), most probably by sequestering the translation initiation factors (GALLIE and TANGUAY 1994) and PABP (GILBERT et al. 2007). The inhibiting effect of exogenous poly(A) on translation can be removed by addition of either translation initiation factors eIF-4B and eIF-4F (with eIF-4A) in combination (GALLIE and TANGUAY 1994) or PABP (GILBERT et al. 2007; GROSSI DE SA et al. 1988) which presumably would bind to the exogenous poly(A) and free translation initiation factors sequestered by the exogenous poly(A).

While pre-AUG A_N may improve translation efficiency, a few studies (BAG 2001; BAG and BHATTACHARJEE 2010; MA et al. 2006; MELO et al. 2003a; MELO et al. 2003b; PATEL and BAG 2006; PATEL et al. 2005; WU and BAG 1998) suggest an inhibitory effect of PABP when it binds to a long pre-AUG A_N and presumably interferes with the scanning mechanism of translation initiation. The binding site of Pab1p in the yeast is
about 12 consecutive A nucleotides, with the binding affinity decreasing rapidly with shorter poly(A) until 8 below which there is little affinity (SACHS et al. 1987). This suggests that mRNAs with pre-AUG AN of different lengths may interact differently with yeast Pab1p and have different translation efficiencies.

Several recent technological breakthroughs have eliminated two fundamental difficulties in characterizing the relationship between pre-AUG AN and translation efficiency in *S. cerevisiae*. First, the transcription start site (TSS) necessary for the accurate delineation of 5' UTR and pre-AUG AN has been characterized for thousands of yeast genes by direct mapping of the capped yeast mRNA sequences (MIURA et al. 2006). This approach to delineate TSSs is conceptually more direct and technologically more accurate than the SAGE approach (ZHANG and DIETRICH 2006). A merged set of TSSs from these two studies has recently been compiled (LAWLESS et al. 2009).

The second difficulty in quantifying the relationship between pre-AUG AN and translation efficiency is the lack of large-scale characterization of protein production and ribosomal density on mRNA. This difficulty is alleviated by a recent proteomic study in yeast (GHAEMMAGHAMI et al. 2003) and several studies characterizing the ribosomal density and the protein synthesis rate for thousands of yeast genes (ARAVA et al. 2003; INGOLIA et al. 2009; MACKAY et al. 2004).

Through a detailed analysis of sequence features in characterized 5’ UTRs and their relationship to *in vivo* protein production and ribosomal loading, two interesting patterns were revealed. First, the frequency of nucleotide A increased dramatically towards the initiation codon, with a concurrent decrease in nucleotide U. This trend is particularly pronounced in highly expressed genes relative to lowly expressed genes when gene
expression is measured by either the protein abundance, by the predicted rate of protein synthesis based on mRNA abundance, by the ribosomal density, or by the codon adaptation index which is highly correlated with transcript and protein abundance in the yeast (COGLAN and WOLFE 2000; DURET and MOUCHIROUD 1999). Second, the protein abundance and the protein synthesis rate both increase with the length of pre-AUG AN up to N = 11 and decrease dramatically for genes having a pre-AUG AN with N ≥ 12. The ribosomal density also reaches the lowest level for genes having a pre-AUG AN with N ≥ 12.

MATERIALS AND METHODS

The 5’ UTR sequences were extracted by using the specification in Table 4 of Miura et al. (MIURA et al. 2006) and the yeast genomic sequences retrieved from UTGB (http://yeast.utgenome.org/). A total of 3274 genes have their TSS characterized, with many having multiple TSSs and consequently multiple 5’ UTRs. To avoid overrepresentation of the sequence patterns of genes with multiple 5’ UTRs, only the longest 5’ UTR for each gene is used. The 5’ UTR sequences are available as a FASTA file (MiuraTSS.FAS in Supplemental Materials).

Three measures of translation efficiency were used for quantifying the relationship between pre-AUG AN and translation efficiency. The first is codon adaptation index (CAI) (SHARP and LI 1987) with its improved version (XIA 2007) implemented in DAMBE (XIA 2001; XIA and XIE 2001). We retrieved the genomic sequences of the 16 chromosomes of S. cerevisiae from the genome database at NCBI, extracted coding sequences (CDSs) and calculated their corresponding CAI values by using DAMBE with the Eysc_h.cut codon usage table. The benefit of using CAI is that it can be computed for
any gene with a codon frequency distribution, whereas experimentally measured protein expression or ribosomal loading data are often limited to relatively highly expressed proteins or transcripts.

The second measure of translation efficiency is the protein abundance, in units of molecules/cell, measured experimentally in a large-scale quantification of yeast protein abundance (GHAEMMAGHAMI et al. 2003). The dataset contains protein abundance data for 3850 yeast genes after excluding 18 genes that do not have a matched name in the current yeast database. For characterizing the relationship between the protein abundance and the length of pre-AUG $A_N$, the protein abundance was log-transformed to stabilize the variance and to linearize the relationship. The protein abundance data used in this paper is available as GhaemmaghamiProtein.xls in Supplemental Materials.

The third measure of translation efficiency is based on large-scale characterizations of ribosomal loading on gene-specific mRNA sequences (ARAVA et al. 2003; INGOLIA et al. 2009; MACKAY et al. 2004). The experimental data for the ribosomal density and the predicted protein synthesis rate in two experimental conditions (mating pheromone treatment and control) were reliably measured for 3916 genes (Supplemental Table II in MACKAY et al. 2004). Another data set characterizing ribosomal density, with a similar method but in a different laboratory (ARAVA et al. 2003), as well as a data set characterizing ribosomal loading with the quite different deep sequencing method (INGOLIA et al. 2009), were also analyzed and compared against results from MacKay et al. (2004). These three data sets, referred to hereafter as MacKayData, AravaData and IngoliaData are noteworthy in that they are highly concordant (Concordance.pdf in
Supplemental Materials). All other data used this paper are in AllData.xls (Supplemental Materials).

RESULTS

*Overrepresentation of A in 5’ UTRs*

The nucleotide frequency distribution over 150 5’ UTR sites immediately upstream of the translation initiation codon shows a dramatic increase of nucleotide A towards the translation initiation codon starting from site -40, with a concurrent decrease in nucleotide U (Fig. 1). The nucleotide frequency distribution before and after site -40 are highly significantly different ($\chi^2 = 2815.84, \text{DF} = 3, p < 0.0001$). A similar pattern was also observed in a previous study (SHABALINA et al. 2004) which suggested that the 30-nt segment immediately upstream of the initiation codon was functionally important in eukaryotic genes, although that study suffered from the fact that sequences upstream of the initiation codon in the genome are not necessarily part of the 5’ UTR because many yeast genes have 5’ UTRs shorter than 30 nt.

If the pattern of increasing A and decreasing U towards the initiation codon (Fig. 1) is related to translation efficiency, then the pattern should be more pronounced in highly expressed genes than in lowly expressed genes. To test this prediction, we compared the nucleotide distribution between the 500 genes with the highest CAI values (High-Expression Group or HEG) and the 500 genes with the lowest CAI values (Low-Expression Group or LEG). Genes in HEG exhibit the pattern much more dramatically than those in the LEG (Fig. 2). The nucleotide frequency distribution between HEG and LEG are highly significantly different ($\chi^2 = 592.74, \text{DF} = 3, p < 0.0001$). The same
pattern is observed when the predicted protein synthesis rate (Mackay et al. 2004) is used as a proxy of translation efficiency (Fig. 3).

*Genes having a pre-AUG $A_N$ with $N < 12$ produce more proteins than those with $N \geq 12$*

Genes with a pre-AUG $A_{12}$ have protein abundance higher than those with a pre-AUG $A_{<12}$ (Table 1), with the two groups differing from each other significantly ($F = 4.96, \text{DF}_{\text{model}} = 1, \text{DF}_{\text{error}} = 2218, p = 0.0261$). There is no significant difference either among poly(A) length groups 2, 3, …, 11, or among the poly(A) length groups with the length equal to, or greater than, 12.

One peculiar data point in Table 1 is the protein abundance for genes with a pre-AUG $A_{10}$. These genes have a singularly low protein abundance, and include a gene with the lowest protein abundance ($YJL084C/ALY2$, alpha arrestin) of only 3.8950. The $A_{10}$ in this gene proceeds “GA”, i.e., in the configuration of AAAAAAAAAAGA. There are two other yeast genes ($YPL271W$ and $YBL101C$) sharing this AAAAAAAAAABA configuration (where $B$ means “not A”, i.e., C, G, or U) with protein abundance values equal to 8.3610 and 6.6839 (which is also low), respectively. Also, one of the $A_{10}$ genes ($YMR181C$) is starvation-induced (Gilbert et al. 2007) and has a relatively low protein abundance as well (= 6.5800). Without these four genes, the mean protein abundance for $A_{10}$ genes is 8.1528.

The observation above led us to ask whether the pre-AUG $A_N$ needs to be contiguous for its effect on protein abundance. We compiled genes with a pre-AUG 12mer that has 11 A residues but with a non-A breaking the contiguity, i.e., in the configuration of

$ABBBBBBBBBBB, AABBBBBBBBB, AAAABBBBBBB, \ldots$,
AAAAAAAAAAAABA and compared their protein abundance with those of genes containing a pre-AUG A_{12}. Protein abundance is significantly higher in the former (mean = 8.5970, n = 43) than in the latter (mean = 7.2013, n = 12), with t = 2.3453, DF = 53, and p = 0.0228 (two-tailed test). Thus, contiguity in poly(A) appears important for decreasing protein production. However, protein abundance is higher for genes with a non-A close to the 5’ end than with a non-A at the 3’ end, i.e., ABAAAAAAAAAA and AABAAAAAAAAAAA genes produce more proteins (mean = 8.6399, n = 9) than AAAAAAAAAAAAAB and AAAAAAAAAAAABAA genes (mean = 6.9207, n = 6), with t = 2.2314, DF = 13, p = 0.0439 (two-tailed test). This suggests that where contiguity is broken may be important.

*Genes with a pre-AUG A_{11} produce more proteins than those with a pre-AUG A_{12}*

The binding affinity of Pab1p approaches maximum when the poly(A) length reaches 12 but decreases quickly when the poly(A) length is 11 or shorter, with hardly any affinity when the poly(A) length is 8 (SACHS *et al.* 1987). So we examined whether genes with a pre-AUG A_{11}, on average, produce more proteins than genes with a pre-AUG A_{12}.

Among the yeast genes with characterized protein abundance (GHAEMMAGHAMI *et al.* 2003), 12 have a pre-AUG A_{12} and 18 have a pre-AUG A_{11} (including the *Pab1/YER165W* mRNA, Table 2). Genes with a pre-AUG A_{11} have higher protein abundance than genes with a pre-AUG A_{12} (t = 3.5827, DF = 28, p = 0.0013, two-tailed test, Table 2). This is consistent with the interpretation that mRNAs with a pre-AUG A_{<12} may enhance translation, but mRNAs with a pre-AUG A_{\geq12} may be subject to translation repression mediated by PABP binding to the pre-AUG poly(A) and interfering with the
ribosomal scanning. The mean protein abundance values are 7.2013 for genes with a pre-AUG A12 and 9.2160 for genes with a pre-AUG A11 (Table 2). Converted back to the original scale, the former is 1341.12 molecules/cell and the latter is 10056.42 molecules/cell.

The analysis above based on protein abundance has the problem that protein abundance is affected by a number of other factors such as translation elongation and mRNA abundance. For example, the protein abundance is highly correlated with CAI which is a measure of translation elongation efficiency (Fig. 4). It is therefore possible that the difference in protein abundance between the A11 genes and the A12 genes may simply be due to the mRNAs from the A11 genes having higher translation elongation efficiency than those from the A12 genes. It is known that different genes can differ much in translation elongation efficiency because their differential codon adaptation to the tRNA pool (van Weringh et al. 2011; Xia 1998; Xia 2005).

To control for the potential difference in translation elongation efficiency between the A11 and A12 genes, we have performed an analysis of covariance (ANCOVA) to test the difference in protein abundance between the A11 and A12 genes with CAI as a covariate. The relationship between protein production and CAI (Fig. 4) is highly significant (slope = 9.8479, p < 0.0001), and the intercept for A11 genes is highly significantly greater than that for A12 genes by 1.4108 (p = 0.0015). Thus, the A11 genes still have significantly higher protein abundance than the A12 genes when translation elongation efficiency has been controlled for by using CAI as a covariate.

We have also performed a similar ANCOVA by using characterized mRNA abundance (Holstege et al. 1998) as a covariate. Protein abundance is significantly
correlated with the mRNA abundance (Fig. 5) for yeast genes with either an A₁₁ or an A₁₂ in their 5’ UTR, with a slope of 0.9455 (p < 0.0001). The intercept for A₁₁ genes is significantly greater than that for A₁₂ genes by 1.1096 (p = 0.0274). Thus, the A₁₁ genes produce more proteins than the A₁₂ genes, when mRNA abundance has been controlled for. A similar pattern is observed when the mRNA abundance characterized by Miura et al. (2008) is used.

Protein synthesis rate and ribosomal density depend on the length of pre-AUG A₅

One problem with the analysis above is that protein abundance depends on both protein production and protein degradation. For example, cyclins typically have a half-life in minutes (Aviram et al. 2008) and are consequently expected to have relatively low abundance even if translation initiation and elongation of their mRNAs are efficient. Thus, the low protein abundance in the A₁₂ genes relative to the A₁₁ genes could be due to more rapid degradation of proteins from the A₁₂ genes than those from the A₁₁ genes.

The ultimate solution to avoid the shortcomings above is to use a more direct measure of translation efficiency than protein abundance to avoid the confounding effect of mRNA abundance or protein degradation. Such measures, expressed as translation efficiency based on ribosomal density, have become available recently for yeast genes (Arava et al. 2003; Ingolia et al. 2009; MacKay et al. 2004; Serikawa et al. 2003).

Among the 3916 genes in MacKayData, 2139 genes have their transcription start sites experimentally determined by Miura et al. (2006). The predicted protein synthesis rate was measured under two experimental conditions (control yeast cells and those treated with a mating pheromone). The two measured rates, however, are similar, with the Pearson correlation being 0.9317 for log-transformed data. We simply took the
average of the two rates, referred to hereafter as MeanRate (of protein synthesis), and studied its relationship to the length of pre-AUG \( A_N \).

MeanRate increases with the length of pre-AUG \( A_N \), but decreases dramatically when poly(A) length reaches 12 (Fig. 6). The regression line in Fig. 6 is highly significant (\( p < 0.0001 \)) for genes whose pre-AUG \( A_N \) is no longer than 11. Thus, a pre-AUG \( A_N \), up to 11, tends to increase the protein synthesis rate, but a long (\( \geq 12 \)) pre-AUG \( A_N \) tends to decrease the protein synthesis rate.

We have also examined the relationship between the experimentally measured ribosomal density and the length of the pre-AUG \( A_N \). Ingolia et al. (2009) characterized ribosomal density for yeast genes in both rich and starvation conditions. Here we use their ribosomal density for 5311 yeast genes characterized under two rich conditions, log-transformed to stabilize variance. While ribosomal density fluctuate for genes with the pre-AUG \( A_N \) length (Fig. 7), the mean ribosomal density for genes with a pre-AUG \( A_{<12} \) (=3.6879) is significantly greater than that (=3.0136) for genes with a pre-AUG \( A_{\geq12} \) (\( t = 2.1932, \text{DF} = 2911, p = 0.0284 \), two-tailed test).

**Is low protein abundance in cyclin due to rapid degradation?**

The cyclin gene \( PCL5/YHR071W \) has a pre-AUG \( A_{29} \), which is the longest among all yeast genes. The Pcl5 protein has a half-life of only 2-3 minutes, and its low abundance is usually attributed to its rapid degradation (Aviram et al. 2008). Our results suggest that the low abundance may be partially attributable to inefficient translation initiation presumably mediated by the binding of Pab1p to the long pre-AUG \( A_{29} \). The ribosomal occupancy is extremely low for \( PCL5 \) mRNA relative to other yeast genes (Fig. 8).
DISCUSSION

Interaction between Pre-AUG poly(A) and Pab1p

Our finding that yeast genes with a pre-AUG A_{212} have a much reduced predicted protein synthesis rate (Figs. 6) and ribosomal density (Fig. 7) is consistent with the hypothesis that, while a pre-AUG A_N may enhance translation by binding to translation initiation factors (GALLIE and TANGUAY 1994; SHIROKIKH and SPIRIN 2008), a long pre-AUG A_N, by binding tightly to PABP, may repress translation by interfering with ribosomal scanning (BAK 2001; MA et al. 2006; MELO et al. 2003a; MELO et al. 2003b; PATEL and BAK 2006; PATEL et al. 2005; SACHS et al. 1987; WU and BAG 1998). This hypothesis would predict that removing PABP would remove its inhibitory effect on mRNA with long pre-AUG A_N. This is exactly what has been observed in a previous in vitro experiment (SHIROKIKH and SPIRIN 2008) without PABP, where the translation-enhancing effect is greater for longer poly(A) than for shorter poly(A), i.e., the ranking order of translation initiation efficiency is A_{25} > A_{12} > A_5.

Our results suggest an alternative to the dominant hypothesis concerning the function of PABP on translation initiation. Several studies have shown that exogenous poly(A) can inhibit translation initiation (GROSSI DE SA et al. 1988; JACOBSON and FAVREAU 1983; LODISH and NATHAN 1972), and the inhibitive effect can be eliminated by the addition of PABP (GILBERT et al. 2007; GROSSI DE SA et al. 1988). Similarly, non-coding poly(A) sequences, such as BC1 and BC200 RNA expressed in neurons, are known to bind PABP (MUDDASHETTY et al. 2002) and to inhibit translation initiation when highly expressed (KONDRASHOV et al. 2005; WANG et al. 2005; WANG et al. 2002). The dominant hypothesis is that PABP, in addition to its function in mRNA stabilization and
circularization, also serves as a translation initiation factor (Kahvejian et al. 2005; Khanam et al. 2006) that functions by binding to pre-AUG An. Thus, either exogenous poly(A) RNA or intrinsically produced poly(A) RNA such as BC1 and BC200 RNA that sequesters PABP would reduce translation initiation (Gilbert et al. 2007; Khanam et al. 2006). Consistent with this hypothesis, adding PABP eliminated the inhibitive effect of the exogenous poly(A) RNA (Gilbert et al. 2007; Grossi de Sa et al. 1988). The hypothesis also explains why poly(A) is over-represented in 5’ UTR in yeast genes, especially those highly expressed ones because such pre-AUG An would gain enhanced translation initiation by interacting with PABP. However, this hypothesis has three difficulties. First, it cannot explain why genes with a long pre-AUG An have a reduced protein synthesis rate as well as a reduced ribosomal density shown in this paper. Second, it cannot explain why, in the complete absence of PABP, pre-AUG An can still enhance translation initiation for both capped and uncapped mRNA (Shirokikh and Spirin 2008). Third, it cannot explain why adding translation initiation factors eIF-4B and eIF-4F (including eIF-4A) in combination also eliminated the inhibitive effect of exogenous poly(A) RNA on translation initiation (Gallie and Tanguay 1994). Our new hypothesis is that pre-AUG An binds to translation initiation factors such as eIF-4B and eIF-4F to facilitate translation initiation. Exogenous or intrinsic poly(A) RNAs can inhibit translation initiation not only because they compete for PABP but also because they would sequester the translation initiation factors eIF-4B and eIF-4F. Adding PABP can eliminate the inhibitive effect of exogenous poly(A) RNA because PABP would bind to the poly(A) and free the translation initiation factors sequestered by these poly(A) RNAs. This new hypothesis, which was implicitly proposed in a previous study (Gallie and
Tanguay 1994) demonstrating the binding of poly(A) RNA to eIF-4B and eIF-4F, eliminates all three difficulties plaguing the other hypothesis.

Presence of a pre-AUG AN appears to be a key feature in a set of internal ribosomal entry sites (IRESs) empirically verified in a recent study on yeast translation (Gilbert et al. 2007). All those poly(A) tracts are shorter than 12 consecutive A’s. These include not only the genes involved in the invasive growth in the yeast, but also transcripts that are routinely transcribed and translated, such as eIF-4G and Pab1 transcripts. The IRES activity mediated by the pre-AUG AN does seem to require Pab1p (Gilbert et al. 2007). A recent study using mRNAs without a poly(A) tail (Kahvejian et al. 2005) suggests that PABP may serve as a translation initiation factor independent of its binding to the poly(A) tail. It is possible that the multiple PABP functions may depend on how strong it binds to pre-AUG AN, with strong binding inhibiting translation and weak binding enhancing translation. It is also possible that the association between the IRES activity and the pre-AUG AN is coincidental. A recent study on IRESs from both the yeast and Drosophila melanogaster shows that IRES activity increases consistently with decreasing stability of secondary structure (Xia and Holcik 2009). A pre-AUG poly(A) would contribute to a weak RNA secondary structure when nucleotide U usage is dramatically reduced (Figs. 1-3).

While there are empirical evidence that mammalian PABP expression may be auto-regulated by PABP binding to the pre-AUG AN of its own mRNA (Bag 2001; Bag and Battacharjee 2010; Ma et al. 2006; Melo et al. 2003a; Melo et al. 2003b; Patel and Bag 2006; Patel et al. 2005; Wu and Bag 1998), there is no evidence that Pab1p abundance in yeast is autoregulated. Pab1p abundance is high, being the top 39th among
the 3841 yeast genes with characterized protein abundance (GHAEMMAGHAMI et al. 2003). Its mRNA ranked the top 114th in ribosomal density among the 5164 yeast genes with ribosomal density characterized by Ingolia et al. (2009). Such a high protein abundance and a high ribosomal density is strong evidence that the high protein abundance in Pab1p does not interfere with the translation of its mRNA. If the auto-regulation requires a pre-AUG A₁₂, then Pab1 mRNA would escape autoregulation because it has only a pre-AUG A₁₁. The mammalian PAPB seems to be less strict on contiguity of poly(A), especially its RRM 3+4 (Khanam et al. 2006).

**Relevance to the translation of early and late genes in vaccinia virus**

The finding that the length of pre-AUG Aₙ is strongly associated with ribosomal loading and protein synthesis sheds light on the evolutionary significance of the difference in the length of pre-AUG Aₙ between early and late genes in vaccinia virus. The early vaccinia viral genes have a pre-AUG Aₙ with 4-14 A residues (Ahn et al. 1990; Ink and Pickup 1990), but the poly(A) tract in late genes are often around 35 A residues (Bertholet et al. 1987; Schwer and Stunnenberg 1988; Schwer et al. 1987). The early viral genes are translated in the presence of abundant PABP which would repress the translation of mRNAs with a long pre-AUG Aₙ. This implies that the transcripts of the viral early genes should have only short poly(A) to avoid repression. In contrast, late viral genes are translated when the cellular protein production has been much reduced, i.e., when PABP is expected to be less abundant. So mRNAs from late viral genes can have long pre-AUG Aₙ without suffering from translation repression mediated by PABP. It has been experimentally demonstrated that, in the absence of
PABP, the translation enhancing effect of pre-AUG AN increases with its length (SHIROKIKH and SPIRIN 2008).

There is some controversy concerning whether the PABP level is reduced during the infection cycle of vaccinia virus. The degradation of host mRNA appears nearly complete 6 hrs after the viral infection as no host poly(A) mRNA is detectable at/after this time (KATSAFANAS and MOSS 2007). Furthermore, a large-scale characterization of mRNA of Hela cells infected with vaccinia virus (YANG et al. 2010) showed that PABP mRNA was reduced to 50% by 4 hrs. Although no mRNA characterization is done after this time, intuition would suggest continued reduction, and such a suggestion is consistent with the finding that no host mRNA is detectable after 6 hrs after the viral infection (KATSAFANAS and MOSS 2007).

The study by Katsafanas et al. (2007) also showed that the viral mRNAs are located in the cavities of viral factories (VFs), where they are transcribed and translated. A number of translation initiation factors such as eIF4E and eIF4G are also localized in these cavities (KATSAFANAS and MOSS 2007; WALSH et al. 2008). In contrast, PABP is localized on the periphery of VF (WALSH et al. 2008) which suggests that PABP does not participate in translation of the viral genes. It is known that vaccinia virus produces poly(A) nontranslated small RNA sequences that selectively inhibit cap-dependent translation of host messages (BABLANIAN and BANERJEE 1986; BABLANIAN et al. 1986; BABLANIAN et al. 1987; BABLANIAN et al. 1993; LU and BABLANIAN 1996), presumably by binding to PABP and preventing it from interacting with other translation initiation factors. Both Rubella virus and Bunyamwera virus inhibit translation of host genes by
producing a capsid protein that binds to PABP and prevents it from binding to other translation initiation factor (BLAKQORI et al. 2009; ILKOW et al. 2008).

Walsh et al. (2008) found a persistent level of PABP during the infection cycle of vaccinia virus, but did not provide any evidence that PABP is in fact produced during the viral infection cycle. However, a subsequent paper (PEREZ et al. 2011) from the same lab found that PABP is continuously synthesized whereas the overall production of host proteins is strongly reduced. They suggested that this selective translation of PABP is through the mTOR+4E-BP pathway. However, this suggestion does not seem coherent. Preventing 4E-BP from binding to eIF-4E would lead to a general increase of cap-dependent translation, not selective translation of the PABP mRNA.

In summary, multiple lines of empirical evidence suggest that a pre-AUG $A_N$ shorter than 12 may enhance translation in the yeast. However, yeast genes with a pre-AUG $A_{212}$ tend to be translated inefficiently with a low ribosomal density and output a reduced amount of protein, consistent with the interpretation that such long poly(A) tracts may bind to Pab1p resulting in repression of translation.

ACKNOWLEDGEMENTS

This study is supported by NSERC’s Discovery, RTI and Strategic Research Grants to XX. We thank Y. Fang, M. Ragonnet, A. van Weringh and S. Zhao for discussion, comments and suggestions. B. Moss provided references on PABP and vaccinia virus.
REFERENCES


GROSSI DE SA, M. F., N. STANDART, C. MARTINS DE SA, O. AKHAYAT, M. HUESCA et al., 1988


INK, B. S., and D. J. PICKUP, 1990 Vaccinia virus directs the synthesis of early mRNAs containing 5' poly(A) sequences. Proceedings of the National Academy of Sciences of the United States of America 87: 1536-1540.


KHANAM, T., R. S. MUDDASHETTY, A. KAHVEJIAN, N. SONENBERG and J. BROSIUS, 2006

KONDRAKHIN, A. V., M. KIEFMANN, K. EBNET, T. KHANAM, R. S. MUDDASHETTY et al., 2005


Identification of a C-terminal poly(A)-binding protein (PABP)-PABP interaction domain: role in cooperative binding to poly (A) and efficient cap distal translational repression. J Biol Chem 278: 46357-46368.

MIURA, F., N. KAWAGUCHI, J. SESE, A. TOYODA, M. HATTORI et al., 2006 A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. Proceedings of the National Academy of Sciences 103: 17846-17851.

MIURA, F., N. KAWAGUCHI, M. YOSHIDA, C. UEMATSU, K. KITO et al., 2008 Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. BMC Genomics 9: 574.


PATEL, G. P., and J. BAG, 2006 IMP1 interacts with poly(A)-binding protein (PABP) and the autoregulatory translational control element of PABP-mRNA through the KH III-IV domain. Febs J 273: 5678-5690.


binding protein is necessary and sufficient for RNA binding and cell viability. Mol Cell
Biol 7: 3268-3276.

SCHWER, B., and H. G. STUNNENBERG, 1988 Vaccinia virus late transcripts generated in vitro
have a poly(A) head. Embo J 7: 1183-1190.

SCHWER, B., P. VISCA, J. C. VOS and H. G. STUNNENBERG, 1987 Discontinuous transcription or
RNA processing of vaccinia virus late messengers results in a 5' poly(A) leader. Cell 50:
163-169.

SERIKAWA, K. A., X. L. XU, V. L. MACKAY, G. L. LAW, Q. ZONG et al., 2003 The transcriptome
and its translation during recovery from cell cycle arrest in Saccharomyces cerevisiae.
Mol Cell Proteomics 2: 191-204.

SHABALINA, S. A., A. Y. OGURTSOV, I. B. ROGOZIN, E. V. KOONIN and D. J. LIPMAN, 2004
Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional

SHARP, P. M., and W. H. Li, 1987 The codon Adaptation Index--a measure of directional
synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:
1281-1295.

SHIROKIKH, N. E., and A. S. SPIRIN, 2008 Poly(A) leader of eukaryotic mRNA bypasses the
dependence of translation on initiation factors. Proc Natl Acad Sci U S A 105:
10738-10743.

VAN WERINGH, A., M. RAGONNET-CRONIN, E. PRANCKEVIENIE, M. PAVON-ETERNOD, L.
KLEIMAN et al., 2011 HIV-1 Modulates the tRNA Pool to Improve Translation


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


FIGURES LEGENDS

Fig. 1. The site-specific nucleotide frequencies of 150 nucleotide sites (X-axis), immediately upstream from the start codon AUG in yeast 5’ UTR sequences. Site -1 is represented by 3274 5’ UTR sequences and site -150 is represented by 747 5’ UTR sequences.

Fig. 2. Contrasting site-specific nucleotide frequencies between highly expressed genes (500 genes with the highest CAI values) and lowly expressed genes (500 genes with the lowest CAI values). The increase in A usage towards the initiation AUG is visually more dramatic in highly expressed genes than lowly expressed genes. The nucleotide frequency distribution between HEG and LEG are highly significantly different ($\chi^2 = 592.74$, $DF = 3$, $p < 0.0001$).

Fig. 3. Contrasting site-specific nucleotide frequencies between highly expressed genes (500 genes with the highest predicted protein synthesis rates) and lowly expressed genes (500 genes with the lowest predicted protein synthesis rates). Protein synthesis rates are from Supplemental Table II in MacKay et al. (2004).

Fig. 4. Visualization of the difference in protein abundance between A11 and A12 genes, when codon adaptation index (CAI) is use as a covariate to control for differences in translation elongation efficiency. A11Pred and A12Pred are fitted regression line from an analysis of covariance, and lnProt is the log-transformed protein abundance from Ghaemmaghami et al. (2003). The difference in intercept is 1.4108 ($p = 0.0015$).
Fig. 5. Visualization of the difference in protein abundance between the A_{11} and the A_{12} genes, when mRNA abundance (lnRNA) is used as a covariate to control for the potential difference in mRNA abundance between the A_{11} and the A_{12} genes. A_{11}Pred and A_{12}Pred are fitted regression lines from an analysis of covariance. lnProt is the log-transformed protein abundance from Ghaemmaghami et al. (2003). The difference in intercept is 1.1096 (p = 0.0274). lnRNA is the log-transformed mRNA abundance from Holstege et al. (1998). A similar pattern is observed when mRNA abundance from Miura et al. (2008) is used.

Fig. 6. Protein synthesis rate (lnMeanRate) increases with the length of pre-AUG poly(A), but decreases dramatically when the poly(A) length reaches 12. The straight line indicates the regression line for genes with poly(A) length shorter than 12. The vertical bars show one standard error above and below the mean for each poly(A) length category.

Fig. 7. Relationship between ribosomal density (log-transformed normalized read density) and poly(A) length. The vertical bars show one standard error (SE) above and below the mean for each poly(A) length category. Some neighboring poly(A) length groups were lumped to reduce SE. Ribosomal density for genes in the “12+” group is significantly lower (t = 2.1932, DF = 2911, p = 0.0284, two-tailed test) than that of all other categories treated as one group.
Fig. 8. Frequency distribution of Occupancy (proportion of mRNA associated with polysomes) of 2175 yeast genes (MACKAY et al. 2004) with 5’ UTR characterized by Miura et al. (2006). The mean occupancy of PCL5 mRNA from two measurements is 0.2360.
**Tables**

**Table 1.** Comparison of protein abundance among genes with different lengths (L) of pre-AUG poly(A). The differences are statistically significant ($F = 2.32$, $DF_{model} = 14$, $DF_{error} = 2205$, $p = 0.0036$). The $p$ value would be reduced to 0.0004 if the genes with poly(A) length equal to greater than 12 were lumped into one group. A gene (YJR113C) with an A$_{21}$ in its 5’ UTR was included into the group of genes with a A$_{16}$.

<table>
<thead>
<tr>
<th>L</th>
<th>N</th>
<th>Mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>275</td>
<td>7.9081</td>
<td>1.4264</td>
</tr>
<tr>
<td>3</td>
<td>523</td>
<td>8.0553</td>
<td>1.6128</td>
</tr>
<tr>
<td>4</td>
<td>599</td>
<td>8.1663</td>
<td>1.5917</td>
</tr>
<tr>
<td>5</td>
<td>355</td>
<td>8.2830</td>
<td>1.7301</td>
</tr>
<tr>
<td>6</td>
<td>187</td>
<td>8.1552</td>
<td>1.7907</td>
</tr>
<tr>
<td>7</td>
<td>114</td>
<td>8.1861</td>
<td>1.6532</td>
</tr>
<tr>
<td>8</td>
<td>66</td>
<td>8.6858</td>
<td>1.7825</td>
</tr>
<tr>
<td>9</td>
<td>34</td>
<td>8.3668</td>
<td>1.5573</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>7.7982</td>
<td>1.6347</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>9.2160</td>
<td>1.5686</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>7.2013</td>
<td>1.4117</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>7.6616</td>
<td>0.4500</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>7.4732</td>
<td>1.8465</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>7.6702</td>
<td>1.6804</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>7.8449</td>
<td>1.1042</td>
</tr>
</tbody>
</table>
Table 2. Protein abundance for genes with a poly(A) tract of length 11 or 12 (A_{11} or A_{12}) in their 5' UTR. The protein abundance is significantly higher in A_{11} genes (mean = 8.9694) than in A_{12} genes (mean = 7.4036), based on a two-sample t-test (t = 3.0879, DF = 39, p = 0.0037, two-tailed test). Codon adaptation index (CAI) is included as a covariate for analysis of covariance (see text for details).

<table>
<thead>
<tr>
<th>Gene</th>
<th>L_{5' UTR} (1)</th>
<th>MaxL_{PolyA} (2)</th>
<th>lnProt (3)</th>
<th>CAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>YHR082C</td>
<td>867</td>
<td>12</td>
<td>7.1448</td>
<td>0.4409</td>
</tr>
<tr>
<td>YBR077C</td>
<td>68</td>
<td>12</td>
<td>8.1566</td>
<td>0.3919</td>
</tr>
<tr>
<td>YDR295C</td>
<td>266</td>
<td>12</td>
<td>7.3121</td>
<td>0.3714</td>
</tr>
<tr>
<td>YFR047C</td>
<td>43</td>
<td>12</td>
<td>7.8878</td>
<td>0.4590</td>
</tr>
<tr>
<td>YGL006W</td>
<td>282</td>
<td>12</td>
<td>4.5945</td>
<td>0.4082</td>
</tr>
<tr>
<td>YGR238C</td>
<td>133</td>
<td>12</td>
<td>6.1493</td>
<td>0.3607</td>
</tr>
<tr>
<td>YHR043C</td>
<td>112</td>
<td>12</td>
<td>7.7817</td>
<td>0.3977</td>
</tr>
<tr>
<td>YKR097W</td>
<td>175</td>
<td>12</td>
<td>10.2470</td>
<td>0.5721</td>
</tr>
<tr>
<td>YLR190W</td>
<td>461</td>
<td>12</td>
<td>6.9747</td>
<td>0.3952</td>
</tr>
<tr>
<td>YLR192C</td>
<td>109</td>
<td>12</td>
<td>9.7912</td>
<td>0.5625</td>
</tr>
<tr>
<td>YMR016C</td>
<td>684</td>
<td>12</td>
<td>5.7506</td>
<td>0.3615</td>
</tr>
<tr>
<td>YMR251W-A</td>
<td>142</td>
<td>12</td>
<td>8.7225</td>
<td>0.6739</td>
</tr>
<tr>
<td>YNL051W</td>
<td>203</td>
<td>12</td>
<td>6.1493</td>
<td>0.4228</td>
</tr>
<tr>
<td>YNL128W</td>
<td>265</td>
<td>12</td>
<td>6.9881</td>
<td>0.3633</td>
</tr>
<tr>
<td>YBR056W-A</td>
<td>47</td>
<td>11</td>
<td>7.5809</td>
<td>0.4077</td>
</tr>
<tr>
<td>YDL140C</td>
<td>519</td>
<td>11</td>
<td>10.1271</td>
<td>0.4864</td>
</tr>
<tr>
<td>YDL246C</td>
<td>321</td>
<td>11</td>
<td>7.1664</td>
<td>0.4695</td>
</tr>
<tr>
<td>YDR033W</td>
<td>555</td>
<td>11</td>
<td>12.1143</td>
<td>0.7725</td>
</tr>
<tr>
<td>YDR055W</td>
<td>82</td>
<td>11</td>
<td>9.3654</td>
<td>0.5406</td>
</tr>
<tr>
<td>Gene</td>
<td>Length</td>
<td>UTR</td>
<td>Protein Abundance</td>
<td>Fold</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>-----</td>
<td>-------------------</td>
<td>------</td>
</tr>
<tr>
<td>YER115C</td>
<td>294</td>
<td>11</td>
<td>8.1442</td>
<td>0.3465</td>
</tr>
<tr>
<td>YER159C</td>
<td>180</td>
<td>11</td>
<td>8.5424</td>
<td>0.2948</td>
</tr>
<tr>
<td>YER165W</td>
<td>145</td>
<td>11</td>
<td>12.1955</td>
<td>0.7079</td>
</tr>
<tr>
<td>YGL037C</td>
<td>373</td>
<td>11</td>
<td>8.9517</td>
<td>0.5606</td>
</tr>
<tr>
<td>YHL012W</td>
<td>77</td>
<td>11</td>
<td>7.4276</td>
<td>0.3696</td>
</tr>
<tr>
<td>YJR070C</td>
<td>72</td>
<td>11</td>
<td>10.5931</td>
<td>0.6736</td>
</tr>
<tr>
<td>YKL080W</td>
<td>130</td>
<td>11</td>
<td>9.9577</td>
<td>0.4954</td>
</tr>
<tr>
<td>YKL084W</td>
<td>52</td>
<td>11</td>
<td>11.5340</td>
<td>0.4050</td>
</tr>
<tr>
<td>YKL186C</td>
<td>377</td>
<td>11</td>
<td>7.4939</td>
<td>0.3297</td>
</tr>
<tr>
<td>YKR092C</td>
<td>318</td>
<td>11</td>
<td>9.4679</td>
<td>0.5404</td>
</tr>
<tr>
<td>YLL013C</td>
<td>298</td>
<td>11</td>
<td>6.7403</td>
<td>0.4263</td>
</tr>
<tr>
<td>YLL029W</td>
<td>155</td>
<td>11</td>
<td>8.0823</td>
<td>0.3828</td>
</tr>
<tr>
<td>YLR328W</td>
<td>90</td>
<td>11</td>
<td>8.5424</td>
<td>0.4664</td>
</tr>
<tr>
<td>YLR341W</td>
<td>98</td>
<td>11</td>
<td>7.9929</td>
<td>0.4145</td>
</tr>
<tr>
<td>YML091C</td>
<td>316</td>
<td>11</td>
<td>7.0521</td>
<td>0.4622</td>
</tr>
<tr>
<td>YMR070W</td>
<td>213</td>
<td>11</td>
<td>7.4353</td>
<td>0.3715</td>
</tr>
<tr>
<td>YMR291W</td>
<td>735</td>
<td>11</td>
<td>9.2261</td>
<td>0.4517</td>
</tr>
<tr>
<td>YMR296C</td>
<td>211</td>
<td>11</td>
<td>10.0159</td>
<td>0.4094</td>
</tr>
<tr>
<td>YOL155W-A</td>
<td>195</td>
<td>11</td>
<td>8.6049</td>
<td>0.3095</td>
</tr>
<tr>
<td>YPL154C</td>
<td>279</td>
<td>11</td>
<td>7.7964</td>
<td>0.6100</td>
</tr>
<tr>
<td>YPL204W</td>
<td>189</td>
<td>11</td>
<td>9.2396</td>
<td>0.4254</td>
</tr>
<tr>
<td>YPR041W</td>
<td>121</td>
<td>11</td>
<td>10.7845</td>
<td>0.5831</td>
</tr>
</tbody>
</table>

(1) 5’ UTR length
(2) maximum length of poly(A) tracts in 5’ UTR
(3) log-transformed protein abundance
Fig. 1
Fig. 2.
Fig. 3
Fig. 4
Fig. 5
Fig. 6.
Fig. 7
Fig. 8