Mapping the translation initiation landscape of an \textit{S. cerevisiae} gene using fluorescent proteins

Tuval Ben-Yehezkel\textsuperscript{a,b,1}, Hadas Zur\textsuperscript{c,1}, Tzippy Marx\textsuperscript{b}, Ehud Shapiro\textsuperscript{a,b}, Tamir Tuller\textsuperscript{d,e,*}

\textsuperscript{a} Computer Science & Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
\textsuperscript{b} Biological Chemistry, Weizmann Institute of Science, Rehovot, Israel
\textsuperscript{c} Computer Science, Tel Aviv University, Tel Aviv, Israel
\textsuperscript{d} Biomedical Engineering, Tel Aviv University, Tel Aviv, Israel
\textsuperscript{e} The Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel

\textbf{A R T I C L E   I N F O}

Article history:
Received 13 March 2013
Accepted 17 May 2013
Available online xxxx

Keywords:
\textit{S. cerevisiae}
Translation
Out-of-frame codons
Genomics
Scanning model
5'UTR

\textbf{A B S T R A C T}

Accurate and efficient gene expression requires that protein translation initiates from mRNA transcripts with high fidelity. At the same time, indiscriminate initiation of translation from multiple ATG start-sites per transcript has been demonstrated, raising fundamental questions regarding the rate and rationale governing alternative translation initiation. We devised a sensitive fluorescent reporter assay for monitoring alternative translation initiation. To demonstrate it, we map the translation initiation landscape of a \textit{Saccharomyces cerevisiae} gene (RMD1) with a typical ATG sequence context profile. We found that up to 3%-5% of translation initiation events occur from alternative out-of-frame start codons downstream of the main ATG. Initiation from these codons follows the ribosome scanning model: initiation rates from different start sites are determined by ATG order, rather than their context strength. Genomic analysis of \textit{S. cerevisiae} further supports the scanning model: ATG codons downstream rather than upstream of the main ATG tend to have higher context scores.

1. Introduction

Gene translation is a central metabolic and regulatory process in all living organisms. A central open question in this field relates to the fidelity and regulation of translation initiation [1–3].

In eukaryotes, canonical initiation of translation is believed to involve scanning of an mRNA transcript with the pre-initiation complex until a start codon, which is typically an AUG codon with the ‘correct’ surrounding sequence context, is approached [4–6]. At this point the large ribosomal subunit associates and elongation commences. According to the first AUG-rule, AUG codons closer to the 5'UTR end of the transcript have a higher probability for initiating translation [7]. However, several deviations from this canonical model have been reported: for example, initiations from non-AUG codons (AUG/CUG/UG) [7–10]: specifically, it was also shown for a specific yeast gene, ALA1, that initiation from all single codons resulting from a single mutation of AUG is possible [8]. Recent studies have demonstrated that mammalian genes with non-AUG N-terminal extensions evolve under strong purifying selection [11] and suggest that frequently non-AUG codons induce translation initiation [9,10]. There are also reported cases of leaky scanning where AUG codons with sub-optimal contexts are skipped and translation initiates at a downstream AUG [7]. Finally, there are known mechanisms of translation initiation from internal ribosome entry sites (abbreviated as IRES). An IRES is a nucleotide sequence that enables translation initiation from the middle of an mRNA transcript [12–14], potentially not involving mRNA scanning.

Recently, studies based on the ribosomal profiling approach have suggested that, in mammals, translation initiation events and programmed frame shifts may occur downstream from the main START codon [9,15,16]. In yeast however, it has been suggested that leaky scanning rarely occurs [7], and cases of translation initiation at any measurable rate from alternative codons in native yeast genes that are not in the frame of the main ORF [17,18] have not, to our knowledge, been reported.

In this study we devised a modular fluorescent reporter assay that experimentally monitors the rate at which translation initiates at alternative ATG codons, thereby accurately mapping the translation initiation landscape from genes with START codons not in the reading frame of the main ORF. We applied it to the yeast RMD1 gene and show that translation initiates from alternative ATG codons downstream of the main ATG and not in the frame of the main ORF, following a specific mechanistic model of translation initiation. Finally, we discuss the advantages and disadvantages of our approach and possible ramifications related to the reported results.
2. Results

2.1. The context score distribution of AUG codons in the yeast genome

We devised a statistical score for ATG start codons (named context score [19]) based on their similarity to the sequence context of ATG START codons of highly expressed genes (Material and methods) in the genome of Saccharomyces cerevisiae, in which higher context scores denote similarity to the ATG context of highly expressed genes (Material and methods); the approach is similar to comparing the AUG context to the Kozak consensus, but is more sensitive as it includes specific weights for each nucleotide distribution (details in the Material and methods section). Surprisingly, our analysis of the S. cerevisiae genome reveals that the transcripts of this organism are enriched with alternative ATG codons with relatively high context scores (Material and methods), both before and after the conventional START ATG (Figs. 1A–B; see also Supplementary Figs. 2–9). Specifically, when considering the nucleotides at a distance of less than 50 codons upstream (in the 5’UTR) or downstream from the ORF’s conventional START codon, which is the region along the transcript likely to participate in translation initiation, S. cerevisiae’s 5861 protein coding genes include a total of 16,635 alternative ATG codons in the three reading frames, of which 12,484 are out of frame.

Although the abundance of high context scores, alternative ATG codons near the START codon (Figs. 1A–C) and not in the frame of the main ORF emphasize the discriminatory power of the translation machinery in initiating translation, it also raises the hypothesis that translation may frequently initiate from such alternative START ATGs. The fact that the context score of the ATG codons upstream of the beginning of the main ORF tends to be significantly lower than context scores of ATG codons downstream of the beginning of the main ORF (Figs. 2A–B) supports the scanning model and our hypothesis above.

2.2. A novel fluorescent reporter assay determines the rate at which translation initiates from alternative ATG codons not in the frame of the main ORF in vivo

We provide initial experimental evidence for the hypothesis that translation may frequently initiate from alternative ATG codons not in the frame of the main ORF in S. cerevisiae by measuring the protein output from these codons in an S. cerevisiae gene. The mechanism of translation initiation by transcript scanning suggests that if alternative initiation occurs, it is most likely to commence from those relatively close to the main START ATG, potentially resulting in previously unknown short peptides [7]. These peptides are experimentally challenging to detect and quantify in the native endogenous context because of their infrequency and variability, rendering analyses using ad hoc methods complicated.

To overcome these problems we devised an experimental setup that enables a sensitive and well controlled analysis of alternative translation initiation events from a different frame relatively to the reading frame of the main ORF (Fig. 3) from any gene (that fulfills some additional conditions that will be further discussed) using a unified methodology. Specifically, a set of variants of a gene is generated, each containing its first ~150 nucleotides (which is related to the length of polypeptide needed to fill the ribosomal exit tunnel)

Fig. 1. The number and distribution of the context score of alternative and main START ATG codons (see also Supplementary Figs. 2–9). A. The number and distribution of alternative ATGs 150 nt upstream (in the 5’UTR) and downstream from the start ATGs according to their log–absolute context score. The context scores of the four alternative ATGs at the beginning of the gene RMD1 are marked by red dashed bars; as can be seen ALT1 has a higher context score than the other three, and the context scores of ALT1–3 appear in the middle of the distribution. B. The number and distribution of the 16,635 alternative ATGs according to their log–relative context score, which is the log of their context score divided by the empirical p-value < 0.001 and absolute (KS test, p-val = 2.3 ± 10−7, empirical p-value < 0.001) ATG context scores are higher (more efficient) downstream of the main ORF (relatively to the ones upstream of the main ORF), supporting the scanning model. The values of the context scores include various ‘outliers’ due to the nature of this score (Material and methods); this is the reason that the mean absolute context score of the 5’UTR does not appear in A. [20] fused to a downstream Yellow fluorescent Protein (YFP) reporter gene, and their fluorescent output is recorded.

We assume that the alternative initiation codons that we want to study have a +1 frame shift compared to the frame of the main ORF; cases with alternative initiation events from codons with a +2 frame shift compared to the frame of the main ORF can be analyzed in a similar manner.

The first variant, named ‘Main’, has the first 150 nt of the wild-type (WT) gene fused to the YFP reporter, and thus estimates the protein levels due to translation initiation solely from the main ATG codon or other ATG codons in-frame (Fig. 3(1)). The second variant, named ‘Alt’ also contains the WT gene, but fused to the YFP reporter with a +1 nt frame shift, in order to measure the protein levels only due to translation initiation from all the alternative ATGs of that frame (Fig. 3(2)). The third variant, named ‘TTG’, was constructed from the ‘Alt’ variant by replacing all its alternative ATGs of the +1 frame with TTG (encoding Leucine). Its purpose is to estimate an upper bound on

Please cite this article as: T. Ben-Yehezkel, et al., Genomics (2013), http://dx.doi.org/10.1016/j.ygeno.2013.05.003
the noise and to control for the possibility that the measured protein levels of the ‘Alt’ variant are due to a frame shift in translation following initiation from the traditional main START ATG of the gene (Fig. 3(3) and Material and methods). The fourth variant named ‘None’, contains a promoter-less YFP reporter gene in order to estimate the lower bound on the noise, e.g. due to auto-fluorescence of S. cerevisiae (Fig. 3(4)). Note that the ‘None’ and ‘TTG’ variants are still expected to be very similar and specifically, the lower and upper bounds can be very close or identical, as they both model noise.

Finally, in order to estimate the translation initiation rate from each of the four alternative ATGs we collectively measured using the ‘Alt’ strain, we generated four additional strains, named ‘Alt1’, ‘Alt2’, ‘Alt3’, and ‘Alt4’ (Figs. 3(5)–(8)); Each of these four strains was identical to the ‘TTG’ strain except that in each of them one of the out-of-frame alternative ATGs was mutated back to ATG and the latter three remained TTG. Taken together, a dedicated version of the aforementioned set of reporter variants can accurately map the translation initiation landscape of any gene.

2.3. RMD1 as a model gene for studying alternative translation initiation

We analyzed the S. cerevisiae genome to find a model gene for studying alternative initiation of translation from codons not in the frame of the main ORF. We searched for genes with main and alternative ATG context scores that are common in S. cerevisiae, as evident by the distribution of ATG context scores in the S. cerevisiae genome (Fig. 1). We excluded genes whose alternative ATG codons that we aimed to study are far, more than 50 codons, upstream or downstream from the beginning of the ORF, since initiation rarely occurs there naturally. Finally, in order to maximally exploit the accuracy of our approach we (1) searched for genes that had all the studied alternative ATGs in the same reading frame (and out-of-frame compared to the main ORF), (2) excluded genes that had additional in-frame ATGs within 150 nt of the main ORF (to isolate the effect of the main START codon on initiation) and (3) excluded genes that had stop codons in the 5′UTR and/or the first 150 nt of the main ORF in all frames to prevent re-initiation [21,22]; our aim was to...
study initiation and not re-initiation; re-initiation occurs when the ribosome finishes translating an ORF (by approaching a STOP codon) and starts translating a new one downstream of it (by approaching the next START codon); by screening against genes with no stop codons we control for this phenomenon (see Fig. 4 for an illustration of the selection process).

The gene RMD1 (YDL001W), encoding a cytoplasmatic protein required for meiosis, which contains 4 alternative out-of-frame ATGs, was selected for experimental analysis. Fig. 1 shows that the context score of RMD1’s four alternative ATGs closest to the main ATG is similar to the mean context score of alternative ATGs in the S. cerevisiae genome, both in terms of their absolute values (Fig. 1A) and in terms of their relative values compared to the context score of their main ATG (Fig. 1B). The context score of RMD1’s main ATG is also similar to the mean genomic value of main ATGs (Fig. 1C). Finally, the alternative ATG load of RMD1 (4 sites downstream of the main ATG) is also not unusual, as 31% of S. cerevisiae genes harbor at least four alternative ATGs downstream of their main ATG.

We generated the complete set of 8 strains of the RMD1 gene (‘Main’, ‘Alt’, ‘TTG’, ‘No’, ‘Alt1’, ‘Alt2’, ‘Alt3’ and ‘Alt4’) in order to (1) experimentally determine whether translation initiates at its out-of-frame alternative ATG codons, (2) provide a quantitative measure of the process and (3) gain mechanistic insight into the process of translation initiation at these sites.

2.4. Alternative translation initiation from ATGs not in the frame of the main ORF in RMD1

We cultured the complete set of strains and recorded their fluorescence in log-phase using a plate-reader (details in the Material and methods section). Our results demonstrate that the combined protein output due to initiation from the four out-of-frame ATGs (i.e. the ATGs not in the frame of the main ORF) of the ‘Alt’ strain is 3.3% (with STD of only 0.59%) of the output due to initiation from the main ATG of the ‘Main’ strain (Material and methods, Fig. 5). Initiation from TTG codons of the ‘TTG’ strain was orders of magnitudes lower than that of the ‘Alt’ strain. Specifically, a statistical analysis demonstrates that the protein levels resulting from the out-of-frame alternative ATGs of the ‘Alt’ strain are significantly higher than from the TTG codons of the ‘TTG’ strain ($p = 1.3 	imes 10^{-32}$, empirical $p$-value $< 0.001$), and that the ‘TTG’ strain is very similar (less than 0.2% difference) but significantly higher than the promoter-less ‘None’ strain ($p = 1.3 	imes 10^{-22}$, empirical $p$-value $= 0.006$). Thus, our results show for the first time, that in S. cerevisiae significant levels of translation initiation can occur from out-of-frame ATG codons downstream of the main START ATG, and to a much lesser extent possibly even from non-ATG codons, or from out-of-frame ATG codons.

Interestingly, individual analyses of the out-of-frame ‘Alt1–4’ variants show that the protein levels due to initiation from the alternative ATG codons closest to the main ATG of the ORF (‘Alt’ strain) were significantly higher than the other three strains (ALT2–4), which did not exhibit protein levels higher than the noise (‘TTG’ strain). Protein levels due to initiation from ALT1 were estimated to be 5% (with STD of only 0.13%) of the protein levels of the ‘Main’ strain (significantly higher than the ALT variant; $p$-value $= 4.8 	imes 10^{-32}$, empirical $p$-value $< 0.001$; Material and methods, Fig. 5), suggesting that the cumulative translation initiation rate from several alternative ATG codons is not additive, and is probably related to the context and positions of the alternative ATGs as previously suggested [7,17,23]. Only the alternative ATG codon closest to the main START exhibited significant amounts of translation initiation, suggesting that either a small fraction of ribosomes scanning the 5′-UTR of RMD1 mRNAs scan through the main START and initiate translation from the next closest ATG they encounter in the scanning process; or, as we discuss later, it is also possible, at least partially, that the alternative initiation is due to cases of 5′ truncated mRNA transcripts.

In fact, the ATG context score (Material and methods) of the first out-of-frame alternative ATG codon was the lowest of the four, and not extremely or unusually high (Figs. 1A–B; the scores of the alternative ATGs 1–4 after normalizing by the score of the Main START are 0.5, 0.76, 0.59, and 10.6 respectively). This suggests that translation initiation from these out-of-frame codons is determined to a larger extent by their location along the transcript than their ATG context score. This result broadens the classical mechanistic assumption of the scanning model, that translation initiates at the first ATG with a sufficiently high context score, regardless of the context score of downstream ATGs, as it demonstrates that a small fraction of the pre-initiation complexes may scan through an ATG with a relatively good context score (the main ATG), and start from a downstream out-of-frame ATG even if its context score is worse.

3. Discussion

We describe a novel approach for monitoring translation initiation from alternative out-of-frame ATG codons. We demonstrate the method by implementing it on the model S. cerevisiae gene RMD1. Previous papers have suggested based on the ribosomal profiling
approach that in mammals there are alternative initiation events near the beginning of the START codon of the main ORF [9,16]. Here we used an approach which is based on estimation of protein levels and is, therefore, complementary to the ribosomal profiling approach, both providing an additional novel layer of experimental validation for this phenomenon, and testing whether it occurs in relatively simple eukaryotic translation systems.

Our computational analyses of alternative START sites in the S. cerevisiae genome (Figs. 1A, B and 2) both emphasize the discriminatory power exhibited by the translation machinery in initiating translation at specific START sites, and raise questions regarding the actual fidelity of the process. Specifically, there is a global signal of lower median scores of upstream alternative sites than the score of downstream sites; this is in accordance with what the scanning model assumes — binding of the ribosomal pre-initiation complex to the first ATG downstream of the cap site with a strong enough context sequence score. However, this signal is not very strong and coherent: first, the correlation coefficient between the context score and ribosomal density is significant (0.19; see Material and methods section); second, as can be seen in Figs. 1–2, the rather wide range of context scores for the main START sites overlaps that of alternative sites. Thus, we expect, specifically in genes with lower context scores for the main ATG, higher usage of potential alternative start codons. There are several additional factors other than the context score that may favor the main ATG site over other Main ATGs. For example, it has been recently discovered that small peptides from the out-of-frame and in-frame alternative ATG codons we identified, are within the range of naturally occurring peptides expressed from the S. cerevisiae genome (see Material and methods).

In addition, the reported signal of translation from ‘out-of-frame’ ATG codons is rather low, less than 5%, and is based on a long list of normalizations and statistical tests (Material and methods), and is significant based on accepted statistical standards. It is possible that the reported signal is an upper bound and the actual signal of alternative translation is even lower than reported; this possibility emphasizes the fidelity of translation initiation. On the other hand, it is possible that future implementations of the method reported here on other genes will yield a much stronger signal of translation from out-of-frame ATG codons. Thus, further studies that will include implementation of our methods on a large set of genes in S. cerevisiae and/or other organisms will shed additional light on the content and abundance of such out-of-frame translated proteins.

It is important to remember that all biological methods, even the very common ones, usually include biases. For example, it is known that DNA chips that have been used in thousands of studies in recent years include biases [33], and various aspects of the recent ribosomal profiling approach apparently include biases [9,16,34]. Our approach can estimate protein levels and translation rates due to alternative START codons that are downstream (as was demonstrated in the current study) but also upstream of the beginning of the ORF, which are not in the same frame of the main ORF. However, in some cases it may be difficult to determine the exact mechanism that causes the alternative initiation. In the case of alternative initiation for a codon that is downstream of the beginning of the ORF, if the analyzed protein levels are low it is possible that this signal is due to skipping of the main start codon (i.e. leaky scanning). However, it is also possible, at least partially, that the alternative initiation is due to cases of 5’ truncated mRNA transcripts; this could occur by cap-independent translation of degradation intermediates, cryptic promoters or aberrant splicing/processing; such truncated transcripts may not include initiation from out-of-frame ATGs have yet unknown regulatory roles in S. cerevisiae.

While results from the model RMD1 gene cannot be readily generalized to the entire transcriptome of S. cerevisiae, its fairly common (1) absolute values of alternative ATG context scores (see Fig. 1A); (2) relative values of alternative ATG context scores compared to their respective ‘Main’ ATG (see Fig. 1B), as well as (3) its relatively average main ATG context score compared to other Main ATGs in the genome (see Fig. 1C), bolsters the notion that other S. cerevisiae transcripts may also exhibit similar behavior.

Our analyses show that the lengths of peptides potentially generated from the out-of-frame and in-frame alternative ATG codons we identified, are within the range of naturally occurring peptides expressed from the S. cerevisiae genome (see Material and methods).

Please cite this article as: T. Ben-Yehezkel, et al., Genomics (2013), http://dx.doi.org/10.1016/j.ygeno.2013.05.003
the main START and thus translation from these transcripts does not involve leaky scanning of their main START. It is important to mention that currently, the typical abundance of such 5’ truncated mRNA transcripts is unknown; if their relative abundance is less than 1% (as in the cases of transcription elongation errors which are estimated to be 1:10,000 [12]), this mechanism will become negligible relatively to the leaky scanning explanation. We believe, however, that the ability to detect the alternative initiation rate is important regardless of the biophysical mechanisms causing it.

One possible source of bias in our method is the fact that the proteins corresponding to the different constructs usually have different N-terminals. An interesting direction for improving our approach is by including a ‘stop–go self-cleaving peptide’ [35]. Specifically, an augmented version of our method can include a self-cleaving peptide that will cleave the N-terminal of the generated peptide after translation such that it will be identical for all variants. However, data from such constructs may be problematic and/or biased in different ways. For example, first, 2A peptides used in stop–go constructs are a significant insertional mutation (approximately 60 bp) into our constructs, rendering them largely different from the native gene we attempt to provide data for. Second, the 2A insertion may affect the activity and stability of our constructs in unpredictable ways (see [36]).

Finally, cleavage efficiency of the inserted peptide is not guaranteed to be fully efficient in constructs with different N-terminals and may therefore skew our results or create an expression bottleneck that masks our results (see [37,35]).

In summary, the quantification of translation initiation is experimentally challenging, fraught with potential biases, with each experimental approach having its advantages and drawbacks. Specifically in the case of stop–go constructs, we evaluate that these may potentially insert more bias in the system compared to the relatively small differences in the N-terminals of our constructs.

4. Material and methods

4.1. Genomic analysis

4.1.1. Coding sequences and UTRs

The coding sequences and UTRs of S. cerevisiae were downloaded from BioMart [38], and the UTR lengths were based on the study of [39]. The analysis regarding the ATG distribution upstream of the beginning of the ORF was based on the information of the UTR length from [39] and included only the 5’ UTR region.

4.1.2. ATG context score

We investigated the ATG context of the start codon and ATGs upstream and downstream from it, in a genome-wide manner, by devising a measure based on the ATG context of the start codon, called the ATG context score. We define an ATG context score (CS) according to the following algorithm (results are robust to versions of this procedure):

1. Select percentage of highly expressed/translated genes.
2. Calculate a position specific scoring matrix (PSSM) based on the nucleotide context around the start codon of the selected highly expressed genes. This PSSM represents the nucleotides necessary for highly efficient translation (see Fig. 6).
3. Calculate the context score per ATG position according to the PSSM for the rest of the genes.

The training set for the position specific scoring matrix (PSSM) calculation is based on the selection of the top 2% of highly expressed genes, according to the product of ribosomal density (RD) and mRNA levels (ML) [39–41], which represent the total flux of ribosomes over a gene.

However, similar results were achieved when selecting the highly expressed group of genes based solely on RD or ML, and also when based on other measures of expression, such as protein abundance etc.

We consider two large scale ribosomal density (RD) measurements (the number of ribosomes occupying the transcript divided by its length); each generated by a different technology.

The first dataset was generated more recently by Ingolia et al. [41] and the second by Arava et al. [40]. We average across the two RD datasets (after normalizing each dataset by its mean), in order to minimize experimental noise. Similar results were obtained when we analyzed each dataset separately. We average across four datasets of protein abundance measurements [42–44], similarly to RD. Similar results were obtained when we analyzed each dataset separately. We considered two large scale measurements of mRNA levels from [41 and 39].

Similarly to RD, we averaged across the two datasets of mRNA levels. In order to avoid over fitting, we select the top 4% highly expressed genes according to the product of RD and ML, and then randomly select half of that group as the training set. Different percentages of training sets sizes (up to 50%) were also tested in the same manner, and the results are robust to training set size variance.

The PSSM was calculated according to the nucleotide (C, A, T, G) appearance probability (frequency), of the training set (2% selected at (1)), for 6 nts before and 3 nts after the first ATG (start) codon, based on the length of the potential optimal ATG context proposed

Fig. 6. ATG context score PSSM for S. cerevisiae. The figure includes for each position near the beginning of the ORF the distribution of the four nucleotides in highly expressed genes (details in the Material and methods section).

Please cite this article as: T. Ben-Yehezkel et al., Genomics (2013), http://dx.doi.org/10.1016/j.ygeno.2013.05.003
by Hamilton et al. [45]. We achieved similar results when considering various context lengths. Interval values between 3 and 48 nt around the start codon were examined, for symmetric and asymmetric combinations, in jumps of 3 nt, and additionally in jumps of 1 nt from 1 nt up to 10 nts, and the results are robust to interval variance. These intervals were tested for all the aforementioned training set sizes in (1), and the results remain robust. The resultant PSSM appears in Fig. 6.

For each gene, the ATG context score is calculated per ATG position, according to the highly expressed genes PSSM calculated in (2): \[ \text{ATGCSj} = \exp(\sum \log(p_{ij})) \], \( j \) is the gene index, \( i \) the nucleotide position, \( p_i \) the probability that the \( i \)th nucleotide of the \( j \)th gene appears in the \( i \)th position. A relative context score is calculated by normalizing by the main ATG's context score: \[ \text{ATGCSj} = \exp(\sum \log(p_{ij})) / \exp(\text{startATGs}) \). We considered up to the last 6 nt of the UTR and the first 3 nt of the ORF. Note that a PSSM with different background weights for different nucleotides (i.e. ATGcsj = exp(\( \sum \log(p_{ij}) \)) where \( b_{ij} \) is the genomic background frequency over the region of the PSSM (up to the last 6 nt of the 5'UTR and the first 3 nt of the ORF after the ATG) of nucleotide \( i \) in gene \( j \) did not improve the correlation between the ATG context score and the ribosomal density (\( r = 0.19 \) \( p = 6.3 \times 10^{-20} \) for the first model vs. \( r = 0.17 \) \( p = 7.8 \times 10^{-20} \) for the second model); thus, we remained with the simpler model (Occam’s razor).

Finally, when we considered a slightly different variant of the context score the results remain very similar:

Suppose that for a certain gene the nt corresponding to position \( i \) in the PSSM is missing for a certain ATG (due to short 5'UTR). The score of the missing positions will be computed as follows:

\[ S = \log \left( \frac{p \times p_a \times p_c \times p_g}{p_1 \times p_2 \times p_3 \times p_4} \right) \]

Let \( p_1, p_2, p_3, \) and \( p_4 \) be the distribution of the 4 nucleotides (A, C, T, G) in the genome in this position (all genes). Let \( p_1' \), \( p_2' \), \( p_3' \), and \( p_4' \) be the four probabilities of the PSSM in this position (based on highly expressed genes). The expected PSSM score in such a position is:

\[ \exp(\sum \log(p_{ij}'))/\exp(\text{startATGs}) \]

This analysis of the ATG context scores in the 5'UTR and the first 50 codons of the ORF, ending at nucleotides 150, 151 and 152 respectively to the frame shifts, with the GFP protein concatenated at the end.

We examined gene candidates according to the number of alternative ATGs in the 5'UTR and ORF, the best/worst/median alternative ATG context score; the distance of the first stop codon from the alternative ATG, the metabolic cost of the peptide defined between the alternative ATG and the first stop codon; the distance of an alternative ATG from the start of the GFP; the % of the mRNA levels of the gene compared to the gene TEF (YPR080W) (as this gene is known to have a strong promoter); the number/distance of relevant STOP codons (i.e. the first stop codon after each alternative ATG), the rank of the gene's protein levels; and the rank of the gene's RD.

4.1.5. Weighted number of ATGs

This analysis (Supplementary Figs. 3, 4, 6, 7B, 8B, 9) was performed by weighting the mRNA levels of the transcripts assuming a total of 60,000 transcripts and based on the mRNA levels from [41]. The results were similar to the ones reported in the main text.

4.1.6. Analysis of length distribution of peptides potentially generated from alternative ATGs in the S. cerevisiae genome

We considered the distance of out-of-frame ATGs from STOP codons in their frame, and the distance of in-frame alternative ATGs from the traditional START ATGs (Material and methods). Our analysis shows that of the alternative (out-of-frame and in-frame) 16,635 ATG codons, 12,484 are out-of-frame alternative ATG codons with a mean distance of 14.83 codons to the closest STOP codon, and 4151 are alternative in-frame ATG codons with a mean distance of 26.95 codons to the closest main START ATG (see also Supplementary figures). In comparison, the mean length of annotated S. cerevisiae peptides is 496.93 with STD 383.07, and with minimal ORF length of only 17 codons (close to the values above).

4.1.7. Empirical p-value based on bootstrapping

4.1.7.1. Context score empirical p-values. Let \( Nu \) and \( No \) denote sets of ATG (absolute or relative) context scores in the 5'UTRs and ORFs respectively; let \( N \) denote the group of ATGs that appear in \( Nu \) or in \( No \) \( \{|N| = |Nu| + |No|\} \). We describe here the procedure for the medians comparisons, but the same procedure was performed for the means comparisons.

Let \( d = \text{median}(\text{No}) - \text{median}(\text{Nu}) \)

Repeat the following 1000 times:

1) Sample a set \( Nui \) of size \( |Nu| \) from \( N \); the rest of the ATG scores from \( N \) are denoted as a set \( Noi \)
2) Compute \( d_i = \text{median}(\text{Noi}) - \text{median}(\text{Nui}) \)

The empirical p-value is \( \text{no. of cases where } d_i \geq d \) / 1000. The empirical p-values for both comparisons (absolute and relative) and when considering mean or median are \( p < 0.001 \) (in none of the random comparisons were the differences between the ORF and the 5'UTR mean/median larger or equal to the original one).

The same procedure was applied in all cases, absolute or relative, mean or median.

Please cite this article as: T. Ben-Yehezkel, et al., Genomics (2013), http://dx.doi.org/10.1016/j.ygeno.2013.05.003
Let \( d = \text{mean}(Y_1) - \text{mean}(Y_2) \).

Repeat the following 1000 times:

1. Sample a set \( Y_1 \) of size \(|Y_1|\) from \( Y \); the rest of the YFP measurements from \( Y \) are denoted as a set \( Y_2 \).
2. Compute \( d_t = \text{mean}(Y_1) - \text{mean}(Y_2) \).

The empirical p-value is \((\text{no. of cases where } d_t \geq d) / 1000\).

### 4.1.7.3. The expected relations between initiation rates and YFP levels.

In this subsection we briefly discuss the relation between translation initiation rates and protein concentration/abundance. Specifically, we will explain why we expect that protein abundance should stand in high positive correlation with translation initiation rates in our study.

Protein abundance levels are determined by a balance between protein production and degradation rates. Fixing the degradation rate, protein abundance levels will rise when the production rate is increased. Fixing the production rate, protein abundance levels will decrease when the degradation rate is increased.

Let \( c_i \) denote the concentration of protein \( i \) and let us assume that this protein is translated from a certain mRNA transcript with \( m_i \) copies in the cell. In general, the dynamics of this process may be described by the following differential equation:

\[
\frac{dc_i}{dt} = R_i - D(c_i),
\]

where \( R_i \) and \( D(c_i) \) are the translation rate per mRNA molecule and the degradation rate of protein \( i \) correspondingly. One common choice for \( D(c_i) \) is:

\[
D(c_i) = d_i \cdot c_i(t) \quad \text{where } d_i > 0 \quad \text{is constant}.
\]

The steady state solution of the above differential equation is:

\[
d_i \cdot c_i^{ss} = R_i \cdot m_i \quad \text{where } c_i^{ss} \text{ is the steady state concentration of protein } i \quad \text{(that is measured by the YFP levels)}. \]

All the variants studied in this research have very similar ORFs with up to four point mutations and/or indels at their beginning; the mutations are related to the ATGs and thus expected to be related to initiation rates (and not elongation or termination rates). Thus, in our case \( R_i \) is assumed to be mostly initiation rate, we expect that the few very specific mutations/indels introduced in the experiment will not affect degradation rates or mRNA levels. Thus, from the equation above we get that \( R_i = \frac{d_i \cdot c_i^{ss}}{m_i} \) and indeed \( R_{\infty} \propto c_i^{ss} \).

### 4.2. The different variants and the rationale behind them

The analysis included eight different variants of the gene RMD1 (see also Fig. 3). Each variant containing ~150 nucleotides of the beginning of this gene (Fig. 3) fused with a YFP gene, and measured their fluorescent levels:

1. The first variant (named Main) wasn’t mutated compared to the original RMD1 gene and thus estimated the protein levels due to translation initiation solely from the main ATG codon.
2. The second variant (named ALT) was fused to the YFP with a 1 nt frame shift in order to measure the protein levels only due to the alternative ATGs of that frame (Fig. 3).
3. The third variant (named TTG) was engineered to estimate an upper bound on the noise; it was reconstructed from the second variant by replacing all the alternative ATGs with TTG.
4. The fourth variant (named None) was a YFP gene without a promoter; thus, it is a lower bound on the noise.

To estimate the initiation rate due to each one of the alternative ATG codons separately we also generated four additional variants; each variant was similar to the third variant above, but in this case only three of the alternatives ATGs were mutated to TTG. We reconstructed such a variant for each of the four alternative ATGs and named these variants: ALT1 (the alternative ATG closest to the beginning of the ORF is not mutated), ALT2, ALT3, and ALT4 (the alternative ATG furthest from the beginning of the ORF is not mutated) respectively.

The fact that we compared the protein levels of each of the variants to the TTG variants enabled us to estimate the protein levels due to alternative initiation events and not due to frame shifts during translation elongation after initiation (which is not expected to occur when mutating ATG codons to TTGs).

### 4.3. Molecular biology

#### 4.3.1. Methods for DNA construction

Construction of the DNA of the YDL001W variants was performed according to the methods described in [46].

#### 4.3.2. Phosphorylation

300 pmol of single stranded DNA in a 50 μl reaction containing 70 mM Tris–HCl, 10 mM MgCl₂, 7 mM dithiothreitol, pH 7.6 at 37 °C, 1 mM ATP & 10 units T4 Polynucleotide Kinase (NEB) was used. Reaction is incubated at 37 °C for 30 min, then at 42 °C for 10 min and inactivated at 65 °C for 20 min.

#### 4.3.3. Elongation

1 pmol of single stranded DNA of each progenitor in a 25 μl reaction containing 2.5 μl of Hot Start DNA Polymerase (Novagen, 71086–3) reaction according to manufacturer’s guidelines was used. Three cycles of annealing were executed for each elongation to ensure full yield of elongation.

#### 4.3.4. PCR

All PCR reactions were performed in 96 well PCR microplates, using KOD Hot Start DNA Polymerase (Novagen, 71086–3) according to its protocol.

#### 4.3.5. Digestion by Lambda exonuclease

1–5 pmol of 5’ phosphorylated DNA termini in a 30 μl reaction containing 67 mM Glycine–KOH, 2.5 mM MgCl₂, 0.01% Triton X-100, 5 mM 1,4-dithiothreitol, 5.5 units Lambda exonuclease (Epicentre) and SYBR Green diluted 1:50,000. Thermal Cycler program is 37 °C for 15 min, 42 °C for 10 min, enzyme inactivation at 65 °C for 10 min.

#### 4.3.6. Chemical oligonucleotide synthesis

Oligonucleotides for all experiments were ordered from IDT with standard desalting.

#### 4.3.7. Minigenes

Three sequenced fragments of under 300 bp containing segments of the YDL001W variants were ordered as IDT Minigenes.

#### 4.3.8. DNA purification

DNA construction related DNA purification was performed with Qiagen Minielute purification kit using standard protocols.

#### 4.3.9. The master strain

The master strain was created by integrating into the yeast genome a cassette containing a promoter-less YFP, followed by a NAT (Nourseothricin) resistance marker under its own promoter. The entire sequence was inserted into the his3Δ1 locus.

#### 4.3.10. Transformations

The three YDL001W-YFP fusion variants were transformed into the master strain using the LiAc/SS carrier DNA/PEG method following the procedures described in [47].

Cells were plated on solid agar SD-URA selective media and incubated at 30 °C for 3–4 days. Transformant colonies were handpicked and patched on SD-URA + NAT (Werner BioAgents) agar plates.

Please cite this article as: T. Ben-Yehezkel et al., Genomics (2013), http://dx.doi.org/10.1016/j.ygeno.2013.05.003
Correct transformation was verified for the 3 variants by PCR amplification from the yeast's genome and gel electrophoresis. All strains were additionally verified by sequencing.

4.3.11. YFP Library construction

The three synthetic YFP fusion constructs were transformed into a master strain that contained a promoter-less YFP coding sequence at the his3A1 locus. Each synthetic construct contained a URA3 selection marker under its own promoter followed by a TEF promoter, the last 30 bp of YDL001W's 5'UTR, the first 150 bp of YDL001W's ORF and the beginning of the YFP ORF (for recombination purposes).

4.3.12. Sequencing

Several single colonies were picked manually from the plates of each variant, the specific integration locus was PCR amplified from each clone. Correct size amplifications were verified by gel electrophoresis. Amplicons were sequenced in house using Sanger sequencing. Colonies clone. Correct size amplification from the yeast's genome and gel electrophoresis. All strains were additionally verified by sequencing.

4.3.13. Culture and fluorescence measurements

In order to measure growth and fluorescent protein expression, the Singer colony arrayer was used to inoculate all colonies of the library into 100 μl of SD-URA media in a 384 well growth microplate (Greiner bio-one, 781162). Following 24 h of pre-incubation, 5 μl of the yeast cultures was diluted into 80 μl of SD complete media in a 384 well microplate, to reach a starting O.D.600 of ~0.1–0.2.

A microplate reader (Neotec Infinite M200 monochromator) was then set to measure the following parameters in cycles of 10 min: Cell growth (as extracted from absorbance at 600 nm) and YFP expression (Ex. 500 Em. 540). Each cycle contained 4 min of orbital shaking at amplitude of 3 mm. The number of cycles was set to 100 (16h) and the temperature to 30 °C.

The experiment included three plates, each plate included four variants, and 96 wells included yeast cultures for each of the variants. Specifically, the 3 different plates included the following variants (all data in Supplementary Table 1):

- Plate 1: ALT, ALT1, None (denoted as None1), TTG (denoted as TTG1).
- Plate 2: Main, ALT2, None (denoted as None2), TTG (denoted as TTG2).
- Plate 3: ALT3, ALT4, None (denoted as None3), TTG (denoted as TTG4).

4.4. Statistical analysis and normalization

To gain fair comparisons between the protein levels of the different variants we performed several normalization steps. First, we removed from each measurement of each variant outliers with YFP or OD levels that are more than 5 standard deviations from the other measurements of the variant at the same time point. Second, for each measurement, we removed the time points where the OD or YFP is after the peak (see Supplementary Fig. 1). Third, as the YFP/OD levels are generally significantly dependent on the OD levels, we performed the following procedure:

1) We use the TTG variant in plate 1 (TTG1), which has the lowest mean OD levels after the first two stages, as a reference.

2) For each of the other variants we iteratively removed time points with the largest OD until we reached a mean OD similar to the one of the TTG1 variant.

At the end of the third step all the variants have similar mean OD. At the next step, for each of the up to 133 measurements of the variants in the three wells, the YFP levels were normalized by dividing them with the corresponding OD measurements to get an estimation of the protein levels per cell. The variants related to the TTG and the None from the three plates were concatenated respectively after normalizing the values in plates 2 and 3, such that the mean YFP/OD levels of TTG2 and TTG3 are identical to the mean YFP/OD levels of TTG1.

The values of pairs of variants were statistically compared by a Kolmogorov–Smirnov (KS) test.

Similarly, to estimate the ratio between the protein levels due to the first alternative ATG and the main ATG we used the following formula:

\[
\frac{\left[\frac{\text{YFP}_{\text{Alt2}}}{\text{OD}_{\text{Alt2}}} - \frac{\text{YFP}_{\text{Alt3}}}{\text{OD}_{\text{Alt3}}}\right]}{\left[\frac{\text{YFP}_{\text{TTG}}}{\text{OD}_{\text{TTG}}} - \frac{\text{YFP}_{\text{None}}}{\text{OD}_{\text{None}}}\right]} = \frac{\text{Alt2/TTG}}{\text{Alt3/TTG}}
\]

Variants ALT2, ALT3, and ALT4 (related to initiation from the second, third, and fourth ATG, respectively) did not exhibit expression levels higher than the TTG variants in their plate. Thus, we concluded that there is no translation initiation from these ATGs.

Finally, we also utilized the time before t = 40 as we believe that these data are also informative (given the normalizations mentioned above). To prove this point, we computed the correlation between YFP/OD obtained for the strains before time 40 and the ones obtained afterwards. For all the plates the correlations were significantly positive (e.g. a correlation of 0.8 for plate 2; p < 10^{-10}). Thus, since the ranking of strains based on the first and last time points is similar it means that the initial points are informative, and using them should improve the statistical power of the reported analyses. This is the reason we decided to use these points in the analysis.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2013.05.003.

Acknowledgments

We thank Prof. Martin Kupiec, Mr. Ido Yofe, and Mr. Tamir Bieuzner for helpful discussions. This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

References


