Evolutionary Pressures on the Yeast Transcriptome

Dominique Chu and Anton Salykin

Abstract—Codon usage bias (CUB) is the well known phenomenon that the frequency of synonymous codons is unequal. This is presumably the result of adaptive pressures favouring some codons over others. The underlying reason for this pressure is unknown, although a large number of possible driver mechanisms have been proposed; one of them is the decoding time. The standard model to calculate decoding time is the Gromadski-Rodnina model. Yet, recently, there have been a number of studies arguing to the effect that this conventional speed-model is not relevant to understand the dynamics of translation. However, results remain inconclusive so far. This contribution takes a novel approach to address this issue based on comparing mRNA with random synonymous variants to estimate the evolutionary pressures that have acted on the transcriptome. It emerges that over 70 percent of ORFs have been subject to a strong selection pressure for translation speed and that there is also a strong selection pressure for the avoidance of traffic jams. Finally, it is also shown that both homogeneous and very heterogeneous transcripts are over-represented. These results corroborate the validity of the Gromadski-Rodnina model.

Index Terms—Translation, evolution, saccharomyces cerevisiae

1 Introduction

The genetic code is highly degenerate—there are 20 amino-acids but 64 codons. Consequently, it is inevitable that each amino acid sequence can be encoded by a very large number of different mRNAs, the so-called *synonymous* mRNAs. Large scale analyses of codons have shown that individual species prefer some codons over others. This is commonly referred to as the *codon usage bias* (CUB). While the bare fact of CUB is well established, its underlying biological reasons are not. A number of drivers of the CUB have been proposed, including the abundance of isoacceptor tRNA, pre-mRNA level selection, mRNA concentration [9], mRNA secondary structure [32], the efficiency of translation initiation [25], GC content [16], gene length [19], translation error [26], [30], protein structure [20], [34] and others [12], [22].

Perhaps one of the more important drivers of the CUB is the decoding time [27]. The current best understanding of the factors determining the decoding time go back to a model by Gromadski and Rodnina [14]. The central idea of the model is that cognate aa-tRNA species compete with near matches (the so-called *near-cognate* aa-tRNA) for access to the ribosome. The latter are thought to occupy the ribosomal A-site for significant amounts of time before eventually unbinding; while bound they prevent access for the cognate aa-tRNA [11] and crucially prevent the translation of the message from proceeding.

- *D. Chu is with the School of Computer Science, University of Kent, CT2 7NF, Canterbury, United Kingdom. E-mail: d.f.chu@kent.ac.uk.*
- *A. Salykin is with the Department of Biology, Faculty of Medicine, Masaryk University, Kamenice 5/A3, Brno, Czech Republic. E-mail: asalykin@med.muni.cz.*

Manuscript received 7 Dec. 2014; revised 22 Feb. 2015; accepted 27 Mar. 2015. Date of publication 5 Apr. 2015; date of current version 5 Oct. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2420554

For many codons, near-cognates are much more abundant than cognates. Even though each near-cognate occupies the ribosome for a short time only, collectively they cause a major bottleneck for translation as a whole [3]. Consequently, the elongation time depends primarily on the ratio of cognate to near-cognates rather than on the absolute number of cognates. This ratio varies strongly between codons. The model of cognate/near-cognate interaction has recently been corroborated experimentally [3].

Following from that is the key prediction of the Gromadski-Rodnina model that the decoding time may vary strongly even between synonymous codons. For example, in *Saccharomyces cerevisiae* the fastest codon (AGA) is read nearly 44 times faster than the slowest one (CUC), according to the model. Similarly, among the synonymous codon sequences for a given protein the predicted translation times (i.e., average time to read one codon) of the fastest sequence may be as much as five times lower than that of the slowest. Despite these large differences, the importance of speed for the evolution of CUB is currently unclear. The *prima facie* argument why translation speed should be selected for is as follows [21], [27]: Faster translation speeds lead to higher achievable translation rates given a fixed ribosome pool; hence by decreasing the time required for a ribosome to read a transcript, the cell can reduce the number of ribosomes while keeping the translation rate fixed. Given that ribosomes are metabolically costly [3, SI], it would seem natural to assume that there is a strong adaptive pressure towards faster mRNAs.

While this resource argument seems clear, it is unclear at present whether there really are speed differences between sequences and what precisely causes them. So far, empirical methods have not been able to settle this question. Most studies in the field do not explicitly consider the decoding speed, but instead rely on various measures of codon adaptedness. Two of the best known ones are the *codon adaptation*

index (CAI) [28] measuring how coding of an mRNA deviates from some highly expressed reference genes, and the *tRNA adaptation index* (tAI) [10] measuring gene adaptation in relation to the tRNA composition of the cell. These are often used as proxies for decoding speed and are able to predict various transcriptomic and proteomic key measures, including expression levels of both mRNA and protein [12]. It can also be shown that both of these adaptedness measures are good predictors for decoding speed (see Table 2 below).

Using *E.coli* as a host Kudla et al. [17] measured the translation rates of an extensive library of synonymous sequences with widely varying predicted speeds. The authors reported no correlation between codon adaptedness and translation rate. Similarly, Qian et al. [23] demonstrated experimentally that the codon adaptedness of an open reading frame (ORF) is not a good predictor for the translation rate. Based on a thorough bioinformatics analysis of footprinting data Charneski and Hurst [1] found that codon adaptedness is not a good predictor for codon speed at all, but that amino-acid charge is. Indeed, these authors pointed out that besides charge, all other previously considered predictors of codon-speed are insignificant. Hence, based on their findings one should *not* expect transcripts to be optimised for (Gromadski-Rodnina) speed.

This partial evidence contrasts with received wisdom in biotechnology where codons of recombinant proteins are engineered routinely to maximise expression [15] indicating a link between codon reading time and expression rate. Also, recently Chu et al. [3] showed for a Firefly Luciferase transcript in a yeast host system that the translation speed as shown by the Gromadski-Rodnina model is a very good predictor for the expression rate. Even stronger evidence is given by Tuller et al. [33] who found a correlation between codon adaptedness and expression level in a genome wide study involving both *Saccharomyces cerevisiae* and *E.coli*.

Closely related to speed is another dynamical effect of translation. So-called "traffic jams" are collisions between two or more ribosomes occupying the same mRNA. Such traffic jams tend to reduce the translation rate per ribosome and therefore are a source of inefficiency in the translational machinery. At present it is not entirely clear whether or not traffic jams actually have a large dynamical impact on translation. There are three main cellular parameters that influence the propensity for traffic jams: ribosome numbers, the translation initiation rate and the speed distribution of codons on the transcript. One can hypothesize that the cell has evolutionarily optimised all three of these aspects to minimise traffic jams.

Altogether, despite significant research efforts in this field, empirical methods have so far not been able to decide the selection pressures driving the evolution of CUB. In this contribution we will therefore take a novel approach. We will consider the transcripts themselves and investigate to what extent they carry the hallmarks of past adaptive pressures. Rather than considering footprinting data that reports ribosome locations or study the expression levels of various genes, we compare transcript sequences of *Saccharomyces cerevisiae* to random synonymous variants. This comparison then allows some conclusions about the evolutionary origin of the sequences in questions. For example, if there is a an evolutionary pressure to increase the speed of transcripts, as predicted by the Gromadski-Rodnina model, then one would also expect that actual yeast sequences tend to be faster than random variants. If on the other hand Gromadski-Rodnina speed is irrelevant for the dynamics of translation, then the *Saccharomyces cerevisiae* transcriptome would show no particularly strong bias towards faster speeds.

To do this comparison we considered for each yeast transcript 1,000 random synonymous variants. These variants were obtained by replacing WT codons by random synonyms. For each of the variant sequences we calculated the quantity under investigation and ranked the wild-type sequence against all the variants. We then took the rank of the WT sequence as an indicator for the evolutionary pressure that this particular sequence experiences.

Here, we consider three key measures: Average codon speed, as measured by the Gromadski-Rodnina model, the propensity of sequences for traffic jams (as derived from a model based on the Gromadski-Rodnina model) and the evenness of the decoding speed distribution across the transcript (in short: *sequence homogeneity*). We found a very strong selection pressure for reading speed. We also found that sequences optimised for traffic jams are over-represented. Surprisingly it also emerged that both heterogeneous and homogeneous sequences are over-represented.

2 Materials and Methods

2.1 Sequences

The tAI values for yeast have been obtained from a link provided in the supplementary material in [18].[1] Yeast mRNA sequences are downloaded from the SGD database [2]. The CAI values have been calculated using EMBOSS and their standard comparison sequence for *Saccharomyces cerevisiae*. The codon reading speeds are obtained from the computational model described in [7].

2.2 Rank Analysis

For the rank analysis of the decoding speed we generated for each ORF 1,000 random synonymous sequences. We calculated for each the reading time per codon by adding the average reading times of each codon and dividing the sum by the length of the transcript. For the rank analysis of the standard deviation, we calculated the standard deviation of the decoding times of codons across each transcript and divided this by the mean reading time per codon. Then we ranked sequences with respect to this measure. In order to obtain the ranks for ribosome sequestration, we simulated every ORF and each of its 1,000 synonymous variants using a dynamic model of yeast translation based on the Gromadski-Rodnina model to obtain the average number of ribosomes on each sequence. The simulation model is described in [7]. For these particular simulations we assumed a saturating ribosome affinity of 1 for each sequence and 200,000 ribosomes. This high number of ribosomes effectively decoupled the individual transcripts. We simulated the system for a period of 1,500 s and calculated the average number of

1. http://longitude.weizmann.ac.il/pub/papers/Man2007_tai/suppl/

TABLE 1
The Abbreviations lSD and hSD Indicate the Set of Low and High Standard Deviation Corresponding to the ORFs Ranking Top 2 Percent and Bottom 10 Percent with Respect to Standard Deviation

	number	fraction
number ORFS	5862	1
Speed	4109	0.70
lSD	664	0.11
hSD	1054	0.18
Jam	433	0.07

(a)

	intersection	all 1	all 2	fraction 1	fraction 2	p-values
Speed, hSD	889	4109	1054	0.22	0.84	$<$ 2.2E-16
Speed, lSD	538	4109	664	0.13	0.81	$<$ 8.763E-11
Jam, hSD	151	433	1054	0.35	0.14	$<$ 2.2E-16
Jam, lSD	48	433	664	0.11	0.07	$<$ 0.9313
Speed, Jam	376	4109	433	0.09	0.87	$<$ 4.109E-15

(b)

"Jam" and "Speed" indicate the top 2 percent ranked ORFs with respect to ribosome sequestration and average decoding speed. ***(a)*** *The number of genes ranked within the top 2 percent within the random ensemble. The total number of ORFs considered was 5862.* ***(b)*** *The intersection set. Roughly 80 percent of the SD and Jam optimised ORFs are also Speed optimised. Column "intersection" shows the number of ORFs that are contained in both sets; all 1 and all 2 is the number of ORFs that belong to the first and second set only (referring to the order in the first column). The column "fraction 1" shows the number in "intersection" as a fraction of "all 1." If this number is much higher than the relevant entry under "fraction" from table (a) then this indicates that the intersection is larger than one would expect at random. The last column reports the significance level of the overlap using the χ^2 test (see Materials and Methods for details); according to this all overlaps are significant except for (Jam, lSD).*

ribosomes on a transcript from a record of the number of ribosome on a sequence over the entire simulation time.

2.3 Overlap between Sequences

To determine the significance between two sets of rankings, we performed a χ^2-test (standard `R` function `chisq.test()`) using the contingency table of the form:

	P_2	$\neg P_2$
P_1	I	$A1 - I$
$\neg P_1$	$A2 - I$	$5862 - (A1 + A2 - I)$.

Here P_1 would be the property first listed in Table 1b (i.e., Speed, Speed, Jam, Jam, Speed). The meaning of $\neg P_1$ is "not having property P_1." Correspondingly, P_2 stands for the property listed second in Table 1b (i.e., hSD, lSD, hSD, lSD, Jam). The symbols A_1 represents the column all1; correspondingly A_2; I is the intersection column. The cell ($\neg P_2$, $\neg P_1$) computes the number of ORFs that are neither in P_1 nor in P_2.

3 Results

3.1 Selection Pressures

We compared each codon sequence of *Saccharomyces cerevisiae* with an ensemble of 1,000 random synonymous variants (see Materials and Methods for details). We ranked each yeast sequence and an ensemble of synonymous variants with respect to the average translation speed. In this way we obtained altogether about 5,600 rankings. This analysis suggests a strong selection for high speed: Over 70 percent of sequences are ranked within the top 2 percent within their respective random ensembles (Fig. 1). If there were no adaptive pressure for speed only 2 percent of sequences would be expected to be ranked within the top 2 percent of their ensembles.

Using comparisons with random sequences it is also possible to check whether yeast ORFs are selected to reduce traffic jams. To this end we determined by simulation the average number of ribosomes on each yeast ORF and on all of the random variants. Given this it was possible to determine the ribosomal occupation ratio (ROR) which is the ratio of the mean number of ribosomes (N_{focal}) on the focal sequence divided by the average of the numbers of ribosomes on random sequences (N_i) across the ensemble of random control sequences.

$$\text{ROR} = \frac{N_{\text{focal}}}{\frac{1}{1{,}000}\sum_{i=1}^{1{,}000} N_i}. \tag{1}$$

To understand how typical the actual yeast transcripts are among their synonymous codons, we considered the distribution of RORs across the transcriptome. In particular we compared two distributions: First, the distribution when the focal sequence for each transcript was a sequence chosen randomly from the set of random variants. Second, the distribution when the focal sequence was the actual yeast transcript. In the first case one would expect the ROR values to follow a unimodal distribution around a mean of 1 and a characteristic standard deviation. The second case may or may not yield a different distribution. If it does, then this indicates that the actual yeast sequence is different from the random sequences with respect to the ROR.

Fig. 2 compares the two distributions graphically. Both are approximated very well by a normal distribution around 1, but the real RORs distribute with a much wider standard deviation than the RORs derived from the random variants. This means that a large proportion of the real sequences are atypical among their synonymous mutants, but there does not seem to be a systematic bias to a higher or lower propensity to traffic jams.

To obtain additional insight, we ranked each of the ORFs within the 1,000 random synonymous variants with respect

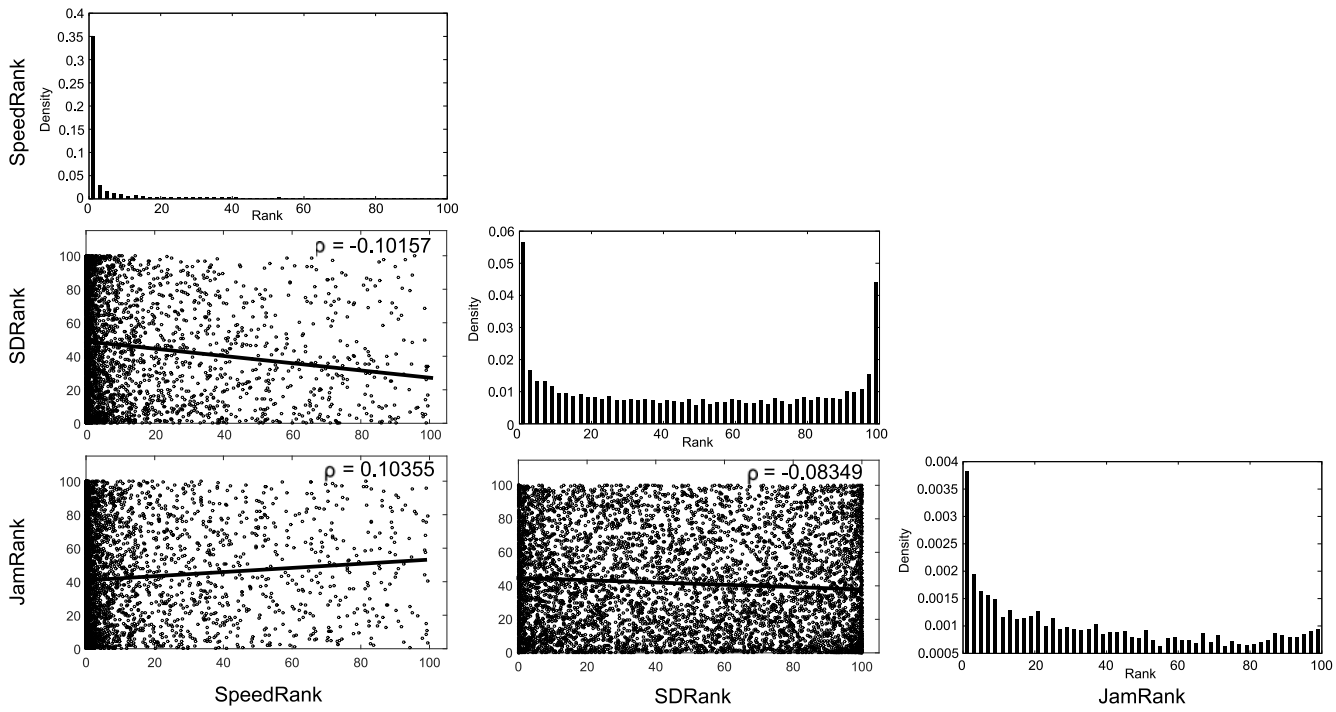

Fig. 1. For each yeast ORF 1,000 random synonymous variants were produced. Within the sample we determined the rank of the actual sequence for speed, standard deviation, and ribosome sequestration/propensity for traffic jams. A low rank (i.e., 1,2,3,...) indicates a sequence with relatively high speed, low standard deviation and a low ribosome sequestration respectively. The diagonal shows the distribution of ranks for all yeast ORFs. The off-diagonal graphs are correlation plots for the entire set of ORFs. The bottom right graph shows how the ranks with respect to speed correlate with the rank with respect to traffic jams. The figures also indicates the correlation coefficient. All ranks are weakly, but statistically significantly correlated ($P < 0.05$).

to the average number of ribosomes on the sequences over the simulation period (i.e., the propensity for traffic jams). A high rank means a relatively low propensity for traffic jams when compared to a random control. We found that high ranks are statistically over-represented in the sample (Fig. 1), although the bias is not as strong as in the case of the speed ranks. Slightly more than 7 percent of all ORFs are ranked within the top 2 percent and 20 percent within the top 10 percent.

A common mathematical model for traffic jams of this sort are TASEPs. These are idealised systems of entities performing a unidirectional one dimensional discrete random walk in continuous time. Many variants of this model can be solved analytically, mainly because they assume a totally homogeneous transition rate, i.e., the average time required to hop from one site to the next is the same for all hops. As such TASEPs are a simplified model of the dynamics of ribosomes on the transcript. Using numerical simulations it can be shown that fully homogeneous transition rates maximise the throughput compared to non-homogeneous transition rates (with the same overall average transition time) [6].

Based on this we hypothesise that codon sequences may be tuned by evolution for greater sequence homogeneity

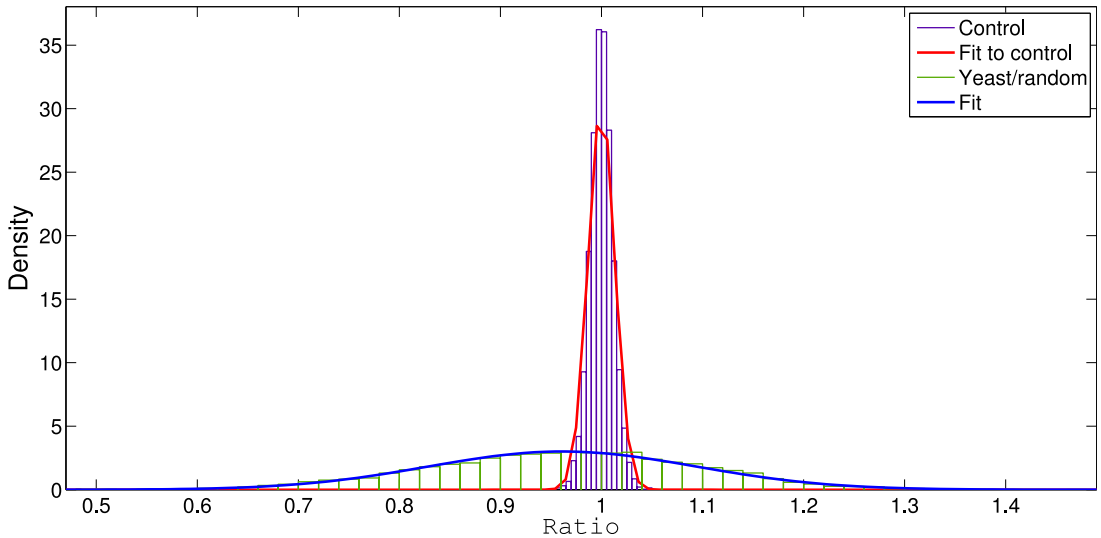

Fig. 2. This graph shows the distribution of the ribosomes occupation ratios across the entire yeast genome (see main text for an explanation). The graph indicates that real ORFs are untypical compared to random synonyms with respect to their propensity to form traffic jams.

(and hence maximal throughput). Here we measure homogeneity simply as the standard deviation of the codon speeds divided by the average codon speed over the ORF sequence. Surprisingly, from our analysis it emerged that both very heterogeneous and very homogeneous sequences are statistically over-represented compared to a random sample (Fig. 1) with 11 percent and 18 percent of sequences ranked within the top 2 percent and bottom 10 percent of their random control ensemble respectively.

The ranks of ORFs within the individual sub-sets (i.e., high speed, high/low standard deviation, low occupation) are only weakly correlated (Pearson's coefficient between -0.1 and 0.1). A different picture emerged when we analysed in more detail the ORFs shared by different sets (i.e., top 2 percent for speed, traffic jams, standard deviation (lSD) and the bottom 10 percent of standard deviation (hSD)). The highest ranking (i.e., top 2 percent) ORFs for standard deviation and traffic jams do not overlap significantly (Table 1b). The two sets share 48 ORFs. This is 11 percent and 7 percent of all elements within the Jam and lSD group and corresponds exactly to the frequencies of Jam and lSD within all ORFs as tabulated in table 1a. The two properties are statistically independent. The overlap between all other sets is larger than what one would expect at random (p-values < 0.05; see table 1b). Surprisingly, the sequences that minimise traffic jams are substantially (factor of 2) over-represented within the heterogeneous sequences and vice versa about twice as many of the jam optimised sequences are unusually heterogeneous. Similarly unexpected is that the high-speed codons are over-represented both in the very homogeneous and the most heterogeneous sequences. Furthermore, among the heterogeneous sequences 84 percent are speed optimised. This value climbs to over 90 percent if we consider the top 2 percent of the heterogenous sequences, rather than the top 10 percent (data not shown).

4 Discussion

In this article we analyse mRNA sequences in *Saccharomyces cerevisiae* to understand some of the evolutionary pressures that shaped the genome. In particular we analysed the average reading speed, the propensity for traffic jams and the sequence homogeneity with respect to speed. The analysis shows that a very large proportion of sequences are speed optimised relative to random synonyms. This result was surprising. While it can be easily argued theoretically that high speed is beneficial for the cell (see above), it was not at all clear that sequences actually are speed optimised to the extent observed. Traditionally, the speed is often measured by tAI or CAI values. These measures show a mono-modal distribution across the transcriptome. This means that a few sequences are highly enriched in terms of the fastest codons, a few are highly enriched in terms of the slowest codons and the vast majority is somewhere in-between.

Yet, this representation does not tell the whole story. It hides the fact that even among this vast majority of sequences with seemingly unremarkable speed many are much faster than one would expect by random. Over 70 percent of yeast transcripts are ranked within the top 2 percent of synonymous sequences. This suggests strongly that yeast sequences are under a significant selective pressure to optimise speed.

This pressure towards increased speed is not surprising *per se*. After all, high reading speed means that costs due to ribosome sequestration can be reduced. In this sense, it is only to be expected that the reading speed is maximised. However, what makes our results more significant is that they are based on the assumption of the Rodnina-Gromadski model. This model has for a long time been assumed to be relevant for the description of decoding speeds, but its relevance has recently become increasingly unclear, c.f. the above cited article by Charneski and Hurst [1]. If the speed model was indeed irrelevant then one would not have expected to see in our study, which was based on the Gromadski Rodnina model, a selection pressure for high speed. However, such a selection pressure is present in the data, suggesting that the model is of relevance.

The next question is then whether or not the avoidance of traffic jams has materially shaped the evolution of CUB. This is a somewhat more subtle question because beside the CUB there are a number of other factors that influence the propensity for traffic jams. There has been some recent interest in theoretical [8], [13], [24] and empirical [29] studies investigating ribosome-ribosome interactions. Yet, it still remains unclear whether traffic jams are a dominant effect impacting on the dynamics of translation in yeast and other organisms. At least in yeast, circumstantial evidence points to jams as a sub-ordinate effect [4], [5], [29]. Ribosomes are an expensive cellular resource and it appears that there are not enough of them available in a cell to cause substantial traffic jams. Our results corroborate this conjecture. The rank analysis of yeast ORFs suggests that there is only a moderate selection pressure to reduce polysome. This would be consistent with the cell avoiding traffic jams by tuning ribosome numbers and affinities. The CUB is only optimised in some genes to further reduce traffic jams.

It is possible that our approach under-estimates the selection against traffic jams. From previous numerical analyses [5] it is apparent that there is a system-wide shortage of ribosomes. In our traffic jam simulations we chose the parameters such that ribosome initiation is saturating; physiological rates are nearly certainly much lower [5], [29]. Consequently, our approach may not detect some of the jam-reducing motifs in the yeast genome that are only effective when the initiation pressure from ribosomes is lower. Further research is required to settle this question.

Intriguingly, we also found that very homogeneous and very heterogeneous sequences are over-represented in the yeast genome. It is not clear why this is the case. Based on theoretical considerations one could conjecture that homogeneity is selected for in order to reduce traffic jams. However, if this were the case, then one would also expect a substantial overlap between the least jam-prone sequences and the most homogeneous ones. No significant overlap exists. Another possibility is that homogeneity is but a side effect of high speed. When more of the codons are faster, then this will tend to remove speed variations along the sequence. Indeed, fast sequences are over-represented among the homogeneous sequences, but only slightly so. However, note that within the heterogeneous sequences (bottom 10 percent) fast sequences are even more over-represented (table 1).

The apparent selection for heterogeneity in ORFs could be a side effect of Tuller's "slow ramps" [31], i.e., stretches of

TABLE 2
Comparing How Various Measures of Codon Adaptedness Correlate with Protein and mRNA Abundance

	CAI	tAI	time	mrna
protein	0.47	0.43	-0.39	0.39
mrna	0.77	0.66	-0.61	1.00
CAI	1.00	0.91	-0.85	0.77
tAI	0.91	1.00	-0.91	0.66

The table matrix lists the Pearson correlations between various key-characteristics of ORFs.

poorly adapted codons at the beginning of the sequence. If this was the case, then one would expect that within the tail ends of ORFs heterogeneous sequences are no longer over-represented. Indeed, we find that the over-representation is reduced, but some still remains (see Fig. 3). Ramps would naturally tend to increase the heterogeneity of the sequence. Our data does not indicate a significant correlation between selection for reduced traffic jams and mRNA, protein expression respectively (data not shown). It is interesting to note that sequences with high standard deviation are over-represented among those with a low propensity for traffic jams (table 1). Only 7 percent of all sequences are within the top 2 percent with respect to jams, but among the highly heterogeneous ones, 14 percent are. This suggests that heterogeneity could have a function in avoiding traffic jams.

Traditionally CUB is investigated using measures of codon adaptedness (e.g., CAI, tAI and others) rather than the speed directly. A tacit assumption that seems to be made frequently is that adaptedness reflects speed. Indeed, we found a very high correlation (Pearson > 0.9) between tAI/CAI and average codon speed (Table 2). Similarly, it is now well established that for many organisms there is a strong correlation between various measures of codon adaptedness and mRNA/protein expression ([18] and Table 2). Consequently, decoding speed and mRNA/protein abundances also correlate although in a somewhat weakened form.

This correlation between codon adaptedness and expression rate has often been interpreted in that highly expressed proteins are under a higher selection pressure than lower expressed ones and consequently more able to counteract random drift [12]. While this seems plausible at first, it is not at all clear that the assumption that rarely expressed genes are under a lower selective pressure is true. One could just as easily make the opposite argument that rare conditions are often associated with stress and require even more efficient translation. Another possibility is to argue that the CUB has evolved as a regulator for gene expression and low expression is a consequence rather than a cause of the lower codon adaptedness.

Our ranking data can give some insight here because the speed-rank of an ORF within a random ensemble is indicative of the selection pressure for faster translation. We find that the correlation between speed-rank and protein/mRNA expression is weak. For proteins expressed below 14,000 the speed-rank and the protein expression rate correlate with a Pearson coefficient of ≈ -0.08; the same value for mRNA is ≈ 0.06 . However, within the dataset there are 473 proteins expressed at 14,000 or higher. Within this subset all but four ORFs rank within the top 5 percent

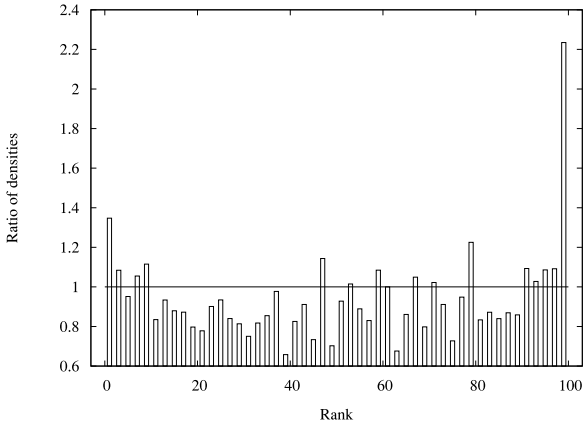

Fig. 3. The ratio of the histograms for the standard deviation for the full sequence (as in Fig. 1) and the same histogram but only using the last 40 codons (data not shown). A value of 1 would mean that for this particular bin the two samples have the same density of entries; a value greater than one means that there are more of those bin values in the full sequence than in the last 40 codons only. The graph indicates that highly heterogeneous sequences are more common when the full sequence is considered.

according to speed. This means that very highly expressed genes are nearly always adapted for speed. Similarly for the mRNA. In our dataset there are 249 ORFs with more than six copies (on average); of those only four rank outside the top 4 percent according to speed. Unlike the direct correlation between codon adaptedness and expression levels, this suggests that only the highest expressed genes are subject to a particularly strong selection pressure for speed.

A prediction of the argument that higher expressed genes are under a stronger selection pressure is that random drift must be an important factor shaping codon usage. If this was the case then one would expect longer ORFs to be altogether less well adapted than shorter ORFs, because the former contain many more micro-states corresponding to average sequences. Hence, one would predict a negative correlation between gene length and adaptedness. We could not find such a correlation for the tAI (Pearson correlation coefficient between tAI and length of ORF is 0.08) or the length of the ORF and the average decoding time (Pearson ≈ 0.04), whereas predicted speed rank and gene length are mildly correlated (Pearson ≈ -0.29). So, genetic drift has shaped the genome, but the rank analysis does not indicate a strong role.

5 Conclusion

When the speed-predictions of the Gromadski-Rodnina model are used, then a strong adaptive pressure towards increased speed is detectable in the transcriptome of *Saccharomyces cerevisiae*. Similarly, we could find a mild adaptive pressure towards reduced traffic jams and high and low sequence homogeneity.

While this indicates that the Gromadski-Rodnina model is relevant for the dynamics of translation in yeast, a number of open questions remain. It is conceivable (though unlikely) that additional drivers of CUB evolution somewhat bias this investigation. In our random variants we have not taken into account GC contents or mRNA secondary structure, nor have we allowed non-synonymous but inconsequential substitutions. Future research needs to clarify whether or not taking into account these effects makes a material difference.

ACKNOWLEDGMENTS

Dominique Chu is the corresponding author.

REFERENCES

[1] C. Charneski and L. Hurst, "Positively charged residues are the major determinants of ribosomal velocity," *PLoS Biol.*, vol. 11, no. 3, p. e1001508, Mar. 2013.

[2] J. Cherry, E. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. Chan, K. Christie, M. Costanzo, S. Dwight, S. Engel, D. Fisk, J. Hirschman, B. Hitz, K. Karra, C. Krieger, S. Miyasato, R. Nash, J. Park, M. Skrzypek, M. Simison, S. Weng, and E. Wong, "Saccharomyces genome database: The genomics resource of budding yeast," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D700–D705, Jan. 2012.

[3] D. Chu, D. Barnes, and T. von der Haar. (2011). The role of tRNA and ribosome competition in coupling the expression of different mRNAs in Saccharomyces cerevisiae. *Nucleic Acids Res.* [Online]. *39(15)*, pp. 6705–6714. Available: http://nar.oxfordjournals.org/content/39/15/6705.abstract

[4] D. Chu, E. Kazana, N. Bellanger, T. Singh, M. Tuite, and T. von der Haar, "Translation elongation can control translation initiation on eukaryotic MRNAs," *EMBO J.*, vol. 33, no. 1, pp. 21–34, Jan. 2014.

[5] D. Chu and T. von der Haar, "The architecture of eukaryotic translation," *Nucleic Acids Res.*, vol. 40, no. 20, pp. 10098–10106, Nov. 2012.

[6] D. Chu and T. von der Haar, "Charting the dynamics of translation," *Biosystems*, vol. 119, pp. 1–9, 2014.

[7] D. Chu, N. Zabet, and T. von der Haar. (2012). A novel and versatile computational tool to model translation. *Bioinformatics* [Online]. *28(2)*, pp. 292–293. Available: http://bioinformatics.oxfordjournals.org/content/28/2/292.abstract

[8] L. Ciandrini, I. Stansfield, and M. Romano, "Role of the particle's stepping cycle in an asymmetric exclusion process: A model of mRNA translation," *Phys. Rev. E*, vol. 81, no. 5 Pt 1, p. 051904, May. 2010.

[9] A. Coghlan and K. H. Wolfe, "Relationship of codon bias to mRNA concentration and protein length in saccharomyces cerevisiae," *Yeast*, vol. 16, no. 12, pp. 1131–1145, Sep. 2000.

[10] M. dos Reis, L. Wernisch, and R. Savva, "Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome," *Nucleic Acids Res.*, vol. 31, no. 23, pp. 6976–6985, Dec. 2003.

[11] A. Fluitt, E. Pienaar, and H. Viljoen, "Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis," *Comput. Biol. Chem.*, vol. 31, nos. 5/6, pp. 335–346, Oct. 2007.

[12] H. Gingold and Y. Pilpel, "Determinants of translation efficiency and accuracy," *Mol. Syst. Biol.*, vol. 7, p. 481, Apr. 2011.

[13] P. Greulich, L. L. Ciandrini, R. Allen, and M. Romano, "Mixed population of competing totally asymmetric simple exclusion processes with a shared reservoir of particles," *Phys. Rev. E*, vol. 85, p. 011142, Jan. 2012.

[14] K. Gromadski and M. Rodnina, "Kinetic determinants of high-fidelity tRNA discrimination on the ribosome," *Mol. Cell*, vol. 13, no. 2, pp. 191–200, 2004.

[15] C. Gustafsson, S. Govindarajan, and J. Minshull, "Codon bias and heterologous protein expression," *Trends Biotechnol.*, vol. 22, no. 7, pp. 346–353, Jul. 2004.

[16] R. D. Knight, S. J. Freeland, and L. F. Landweber, "A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes," *Genome Biol.*, vol. 2, no. 4, 2001.

[17] G. Kudla, A. Murray, D. Tollervey, and J. Plotkin, "Coding-sequence determinants of gene expression in Escherichia coli," *Science*, vol. 324, no. 5924, pp. 255–258, Apr. 2009.

[18] O. Man and Y. Pilpel, "Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species," *Nat. Genet.*, vol. 39, no. 3, pp. 415–421, Mar. 2007.

[19] E. N. Moriyama and J. R. Powell, "Gene length and codon usage bias in drosophila melanogaster, saccharomyces cerevisiae and escherichia coli," *Nucleic Acids Res.*, vol. 26, no. 13, pp. 3188–3193, Jul. 1998.

[20] P. Mukhopadhyay, S. Basak, and T. Ghosh, "Synonymous codon usage in different protein secondary structural classes of human genes: Implication for increased non-randomness of gc3 rich genes towards protein stability" *J. Biosci.*, vol. 32, no. 5, pp. 947–963, Aug. 2007.

[21] S. Navon and Y. Pilpel, "The role of codon selection in regulation of translation efficiency deduced from synthetic libraries," *Genome Biol.*, vol. 12, no. 2, p. R12, 2011.

[22] E. Novoa and L. Pouplana, "Speeding with control: Codon usage, tRNAs, and ribosomes," *Trends Genet.*, vol. 28, no. 11, pp. 574–581, Nov. 2012.

[23] W. Qian, J. Yang, N. Pearson, C. Maclean, and J. Zhang, "Balanced codon usage optimizes eukaryotic translational efficiency," *PLoS Genet.*, vol. 8, no. 3, p. e1002603, 2012.

[24] M. Romano, M. Thiel, I. Stansfield, and C. Grebogi, "Queueing phase transition: Theory of translation." *Phys. Rev. Lett.*, vol. 102, no. 19, p. 198104, May 2009.

[25] T. Sato, M. Terabe, H. Watanabe, T. Gojobori, C. Hori-Takemoto, and K. Miura, "Codon and base biases after the initiation codon of the open reading frames in the escherichia coli genome and their influence on the translation efficiency," *J. Biochem.*, vol. 129, no. 6, pp. 851–860, Jun. 2001.

[26] P. Shah and M. Gilchrist, "Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias," *PLoS Genet.*, vol. 6, no. 9, p. e1001128, 2010.

[27] P. Shah and M. Gilchrist, "Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 25, pp. 10 231–10 236, Jun. 2011.

[28] P. M. Sharp and W. H. Li, "The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Res.*, vol. 15, no. 3, pp. 1281–1295, Feb. 1987.

[29] M. Siwiak and P. Zielenkiewicz, "A comprehensive, quantitative, and genome-wide model of translation," *PLoS Comput. Biol.*, vol. 6, no. 7, p. e1000865, 2010.

[30] N. Stoletzki and A. Eyre-Walker, "Synonymous codon usage in escherichia coli: Selection for translational accuracy," *Mol. Biol. Evol.*, vol. 24, no. 2, pp. 374–381, Feb. 2007.

[31] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel, "An evolutionarily conserved mechanism for controlling the efficiency of protein translation," *Cell*, vol. 141, no. 2, pp. 344–354, Apr. 2010.

[32] T. Tuller, I. Veksler-Lublinsky, N. Gazit, M. Kupiec, E. Ruppin, and M. Ziv-Ukelson, "Composite effects of gene determinants on the translation speed and density of ribosomes," *Genome Biol.*, vol. 12, no. 11, p. R110, 2011.

[33] T. Tuller, Y. Waldman, M. Kupiec, and E. Ruppin, "Translation efficiency is determined by both codon bias and folding energy," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 8, pp. 3645–3650, Feb. 2010.

[34] T. Xie, D. Ding, X. Tao, and D. Dafu, "The relationship between synonymous codon usage and protein structure," *FEBS Lett.*, vol. 434, nos. 1/2, pp. 93–96, Aug. 1998.

Dominique Chu received the master's degree in theoretical physics from the University of Vienna in 1997 and a doctorate in physics from the University of Bergen in Norway. He is a Lecturer in the School of Computing, University of Kent. He is now working in computational biology.

Anton Salykin received his MSc degree in Medical Biochemisrty from the Northern State Medical University, Russia, in 2009. Then, he started PhD studies in the Department of Biology, Faculty of Medicine, Masaryk University in Czech Republic. Currently, he is interested in metabolic plasticity of pluripotent cells and apply systems biology methods to model their characteristic features.