Review of the paper 'Rate-Limiting Steps in Yeast Protein Translation' by Shah & Plotkin *et al*.

In their paper Rate-Limiting Steps in Yeast Protein Translation, Shah *et al.* propose a computational model for whole-cell translation. This model is used to test a number of hypotheses regarding the contribution of various factors to global protein production in the cell.

As we describe below, the study (at least in its current form), is clearly not suitable to be published in Cell. Specifically, the reported results are either wrong or not new. In addition, the described model has many disadvantages that are not discussed, and it is clearly not evaluated accurately. Furthermore, while the authors claim the model is more comprehensive than previous models, many fundamental aspects mentioned in previous studies are not included in it. Below are some initial points (a partial list).

Diament & Tuller 5.3.2016

Major

---

1) The reported correlation between estimated initiation probabilities and predicted 5'-end mRNA folding energies is significant, but very low ($R=0.125$, $p<1E-13$). Furthermore, the data presented in Figure 5A is in fact binned. This fact is specifically interesting, as the authors discourage the use of bins, but at the same time note that it is "sometimes appropriate to bin the data for visualization purposes". ~60 bins are used for 3,795 values, and while error bars denote the variance in each bin, the "visual representation" gives the impression of a strong relation between the two variables using *very few points*, which is not the case.

In addition, the authors should include a control, e.g. partial correlation, for other important variables, such as the Kozak score (Kozak, *Cell*, 1986), when computing this correlation (which eventually may be significantly lower) as well as CAI, amino acid charge, and GC content, at the beginning of the coding region, etc. These variables are known (and were known at the time of writing this paper) to have typical signals in the 5'-end of the ORF that are correlated with expression level (and initiation rate), for example see (Plotkin and Kudla, *Nat Rev Genet*, 2011).

We also recommend that statistical significance for this correlation be tested using an empirical null model. For example, amino acid preserving sequences in the relevant window can be generated at random to estimate the probability of observing a correlation of 0.125 between inferred initiation rates and mRNA folding energy. This is important since the distribution of amino acids near the N terminal of the protein may induce the folding strength via their specific codon coding them.

When testing 5'/3' UTR length vs. initiation rates, p-values are provided but the statistical test is not defined, and it is not clear what is considered a "shorter" or "longer" UTR (*e.g.*, if a t-test was used).

2) This is not the first study that includes/suggests the analyses of many intracellular components (mRNA and ribosomes). However, previous studies on the topic such as (Brackley et al., *PLoS Comput Biol*, 2011; Chu et al., *Bioinformatics*, 2011; Cook and Zia, *Journal of Statistical Mechanics: Theory and Experiment*, 2009; Cook et al., *Phys. Rev. E*, 2009; Greulich et al., *Phys. Rev. E*, 2012; Heldt and Thiel, 2009; Jonathan Cook and Zia, *Journal of Statistical Mechanics: Theory and Experiment*, 2012; Karr et al., *Cell*, 2012; Mather et al., *Biophysical Journal*, 2013) are not cited.

3) The major disadvantage when developing a model with many parameters is overfitting (Babyak, *Psychosom Med*, 2004), or accumulation of error due to individual errors in the different parameters. Overfitting is a fundamental concept in modeling and it is very surprising that this issue is not mentioned at all; the authors do not bother to discuss this issue, and to demonstrate that this is not the case with the model (as is usually done and should be done in papers in the field).

Specifically, the authors base their simulation on 13 "physical parameters that have been experimentally determined in yeast", however they do not discuss or provide confidence intervals (or error estimates) for these parameters in Table 1, main text, or experimental procedures. The accumulated error for these 13 parameters (and several additional ones appearing in Table S1, some of them determined per-gene for 3,795 genes) could be quite high. Thus, all the results in the paper, such as protein yield increase of transgenes, should be accompanied with confidence intervals, or at the very least, it should be demonstrated that results are robust to changes in parameters.

In addition, it is not clear how stochastic the simulation results are? Even when averaging over the last 500s of the simulation, it should be demonstrated that repeated simulations converge to a similar/same state. "… for each simulation involving transgenes, we used ten sequences of similar CAI values and equal mRNA abundances to represent the transgene, in order to alleviate noisy, sequence-specific effects." It would be good to show the variance between the 10 simulated transcripts (*e.g.*, an error bar). This is clearly a fundamental aspect without which it is impossible to evaluate the results.

4) All the conclusions reported in this paper based on the model are either wrong or not new. Thus, the ability of this model to provide novel conclusions is not convincing at all. For example (points mentioned in the abstract), the correlation between initiation and folding energy has been suggested many times before (Jia and Li, *FEBS Lett.*, 2005; Kudla et al., *Science*, 2009; Sagliocco et al., *J. Biol. Chem.*, 1993; Schauder and McCarthy, *Gene*, 1989; Wang and Wessler, *Plant Physiol.*, 2001) and is very low ($r = 0.125$); the idea that the ramp is caused by rapid initiation is wrong, among others, due to the fact that the authors did not normalize each profile by dividing it by the mean as was performed in previous studies they cite. The conclusions

regarding the fact that “protein production in healthy yeast cells is typically limited by the availability of free ribosomes, whereas protein production under periods of stress can sometimes be rescued by reducing initiation or elongation rates” is not new (Bergmann and Lodish, *J. Biol. Chem.*, 1979) and can’t be accurately evaluated via the model due to overfitting issues mentioned above, and missing fundamental aspects mentioned below and above.

5) The authors ignore previous studies that suggested that elongation is not only due to adaptation to the tRNA pool. For example, the model does not consider important aspects related to the elongation rate that were previously reported and suggested to have stronger effect on elongation than tRNA levels. Expressly, the effect of mRNA folding (Nackley et al., *Science*, 2006; Plotkin and Kudla, *Nat Rev Genet*, 2011; Pop et al., *Molecular Systems Biology*, 2014; Tuller et al., *Genome Biology*, 2011; Yang et al., *PLoS Biol*, 2014) and amino acid content (Charneski and Hurst, *PLoS Biol*, 2013; Lu and Deutsch, *J. Mol. Biol.*, 2008; Lu et al., *Journal of Molecular Biology*, 2007; Muto and Ito, *Biochemical and Biophysical Research Communications*, 2008; Pavlov et al., *Proc. Natl. Acad. Sci. U.S.A.*, 2009), as well as wobble basepairing (Stadler and Fire, *RNA*, 2011). In addition, the analysis is partially based on Ribo-Seq data, but measurements are not used to infer reliable codon decoding rates. Previous studies (e.g. Charneski and Hurst, 2013) suggested that these aspects are significantly more important than the tRNA pool adaptation.

Thus, without considering these fundamental aspects it is not clear what we can learn from this model (!).

6) The analysis of the ramp is clearly wrong and was not performed as in previous studies (see, explanations in Figure 5A in (Tuller and Zur, *Nucl. Acids Res.*, 2015)). Ribosome occupancies must be normalized per gene by its average ribosome occupancy before averaging each codon position across the transcriptome (this is how this was performed in previous studies e.g. (Ingolia et al., *Science*, 2009) (!)). This simple normalization ensures, for instance, that shorter genes with elevated ribosome densities cannot bias the analysis (and indeed did not bias previous reports). Furthermore, the analysis should include only highly expressed genes, because of the very limited coverage (only <7% of the nucleotides are covered with reads (!) for the bottom 80% of genes) of the ribosome profiling approach, and specifically the data in (Ingolia et al., *Science*, 2009). Thus, the entire section dedicated to this issue is not relevant/correct. We must say we are surprised that the authors are not familiar with the details and methodologies used in previous studies.

7) The reported results contradict previous papers of the authors’ themselves (Kudla et al., *Science*, 2009), but this is not discussed. In this paper (Kudla et al., *Science*, 2009), the authors claimed that slower codons (with lower adaptation to the tRNA pool/codon bias) are related to

higher ribosomal densities and vice versa. The 5' end of the ORF has been shown to be enriched with slower codons (Plotkin and Kudla, *Nat Rev Genet*, 2011). Thus, by concluding that the 5' ramp is not related at all to codon usage bias in this Cell paper, the authors contradict themselves (as it is related to the claim that codon bias and ribosome densities are unrelated).

8) The reported results are also contradicted by papers the authors published after this study (Weinberg et al., *Cell Reports*, 2016).

9) The performance related to all the aspects of the model should be comprehensively compared with that of other models, including simpler ones such as TASEP, deterministic TASEP, RFM, tAI, etc., and include a *balanced* discussion about the relative advantages of the different models. Otherwise, the advantages of the proposed model are not clear and misleading. The following are some points to be considered: running time, number of parameters needed and the availability of the parameters, the possibility of analytical analysis of the model, accurate comparison of different predictions using real and simulated data, etc.

In addition, the inferred initiation rates should be compared with previous reported initiation rates based on other models, such as (Ciandrini et al., *PLoS Comput Biol*, 2013; von der Haar, *BMC Syst Biol*, 2008).

10) Codon Bias and Transgene Expression: In order to test the effect of codon optimization it would have been better (and more interesting) to use the equilibrium codon decoding rates of the simulation instead of CAI. It is only assumed, but not shown, that CAI is correlated with the simulated codon decoding rates.

11) Very few verifications for the model are provided in the study: “Simulated initiation rates are successfully inferred using the model equations, and the average codon translation rate and the mean distance between consecutive bound ribosomes are in agreement with previous reports”. The authors could use their GFP library to compute correlations between, *e.g.*, protein production and their model predictions (as well as compare that with the performance of simpler models).

In addition, the data appearing in Figure S2 is not cited or provided as an accession number.

12) “Interestingly, we also found a negative correlation between initiation probability and open reading frame (ORF) length (R = -0.56 and p < 10^-15; Figure 5B), even after controlling for mRNA expression level (partial correlation, R = -0.425 and p < 10^-15). This trend suggests that shorter yeast genes have experienced selection for faster initiation, and so it provides a mechanistic explanation for the greater density of ribosomes typically observed on short genes (Arava et al., 2003; Lackner et al., 2007).” This result can be at least partially explained by the fact that in (Ingolia et al., *Science*, 2009) there is an increased ribosome density at the 5' end (partially due to biases as suggested in (Ingolia et al., *Cell*, 2011)). This should have a stronger effect on shorter genes and will be related to higher inferred initiation rates in these genes (higher RC --> higher initiation). The current analysis does not control for this point.

At the time of publishing and working on this study it was already very clear that the higher density of ribosomes at the 5'end of the mRNA was at least partially due to biases in the experiment (see, for example, (Ingolia et al., *Cell*, 2011)); nevertheless, the authors do not mention or consider this fundamental issue in their analysis/conclusions at all (!).

Minor:

13) "Finally, it is important to note that the TASEP-based models of translation (e.g., Reuveni *et al*., 2011) cannot, even in principle, be used to assess whether protein production is limited by available ribosomes because such models assume a fixed, inexhaustible supply of free ribosomes.". There are dozens of TASEP-based models of translation but the authors cited only one. For example, see (Chou, *Biophys J*, 2003; Chou and Lakatos, *Phys. Rev. Lett.*, 2004; Ciandrini et al., *Phys. Rev. E*, 2010, *PLoS Comput Biol*, 2013; Dong et al., *J Stat Phys*, 2007; Garai et al., *Phys. Rev. E*, 2009; Kemp et al., *Mol. Microbiol.*, 2013; MacDonald and Gibbs, *Biopolymers*, 1969; Romano et al., *Phys. Rev. Lett.*, 2009; Shaw et al., *Phys. Rev. E*, 2003; Zia et al., *J Stat Phys*, 2011). There is an entire section (Comparison with Ribosome Flow Model of Translation) in the extended experimental procedures in which the proposed model is compared exclusively with RFM in terms of the size of ribosome pool needed to reconcile the 2 models. Other computational models for translation are not mentioned.

14) Minor ramp issue:

"The ability of our model to recapitulate this striking spatial trend is nontrivial because we did not use any position-specific information from the ribosomal profiling data." It should be noted that previous much simpler models for translation, e.g. tAI values, also managed to do this. See also comment 12.

References

Babyak, M.A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med *66*, 411–421.

Bergmann, J.E., and Lodish, H.F. (1979). A kinetic model of protein synthesis. Application to hemoglobin synthesis and translational control. J. Biol. Chem. *254*, 11927–11937.

Brackley, C.A., Romano, M.C., and Thiel, M. (2011). The Dynamics of Supply and Demand in mRNA Translation. PLoS Comput Biol *7*, e1002203.

Charneski, C.A., and Hurst, L.D. (2013). Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. PLoS Biol *11*, e1001508.

Chou, T. (2003). Ribosome Recycling, Diffusion, and mRNA Loop Formation in Translational Regulation. Biophys J *85*, 755–773.

Chou, T., and Lakatos, G. (2004). Clustered Bottlenecks in mRNA Translation and Protein Synthesis. Phys. Rev. Lett. *93*, 198101.

Chu, D., Zabet, N., and Haar, T. von der (2011). A novel and versatile computational tool to model translation. Bioinformatics btr650.

Ciandrini, L., Stansfield, I., and Romano, M.C. (2010). Role of the particle's stepping cycle in an asymmetric exclusion process: A model of mRNA translation. Phys. Rev. E *81*, 051904.

Ciandrini, L., Stansfield, I., and Romano, M.C. (2013). Ribosome Traffic on mRNAs Maps to Gene Ontology: Genome-wide Quantification of Translation Initiation Rates and Polysome Size Regulation. PLoS Comput Biol *9*, e1002866.

Cook, L.J., and Zia, R.K.P. (2009). Feedback and Fluctuations in a Totally Asymmetric Simple Exclusion Process with Finite Resources. Journal of Statistical Mechanics: Theory and Experiment *2009*, P02012.

Cook, L.J., Zia, R.K.P., and Schmittmann, B. (2009). Competition between multiple totally asymmetric simple exclusion processes for a finite pool of resources. Phys. Rev. E *80*, 031142.

Dong, J.J., Schmittmann, B., and Zia, R.K.P. (2007). Towards a Model for Protein Production Rates. J Stat Phys *128*, 21–34.

Garai, A., Chowdhury, D., Chowdhury, D., and Ramakrishnan, T.V. (2009). Stochastic kinetics of ribosomes: Single motor properties and collective behavior. Phys. Rev. E *80*, 011908.

Greulich, P., Ciandrini, L., Allen, R.J., and Romano, M.C. (2012). Mixed population of competing totally asymmetric simple exclusion processes with a shared reservoir of particles. Phys. Rev. E *85*, 011142.

von der Haar, T. (2008). A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst Biol *2*, 87.

Heldt, S., and Thiel, M. (2009). A Mathematical Model of Protein Translation and the Competition for Rare Cellular Resources in Response to Amino Acid Starvation. Otto-von-Guericke Universität Magdeburg.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science *324*, 218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell *147*, 789–802.

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

Jonathan Cook, L., and Zia, R.K.P. (2012). Competition for finite resources. Journal of Statistical Mechanics: Theory and Experiment *2012*, P05008.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell *150*, 389–401.

Kemp, A.J., Betney, R., Ciandrini, L., Schwenger, A.C.M., Romano, M.C., and Stansfield, I. (2013). A yeast tRNA mutant that causes pseudohyphal growth exhibits reduced rates of CAG codon translation. Mol. Microbiol. *87*, 284–300.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283–292.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science *324*, 255–258.

Lu, J., and Deutsch, C. (2008). Electrostatics in the ribosomal tunnel modulate chain elongation rates. J. Mol. Biol. *384*, 73–86.

Lu, J., Kobertz, W.R., and Deutsch, C. (2007). Mapping the Electrostatic Potential within the Ribosomal Exit Tunnel. Journal of Molecular Biology *371*, 1378–1391.

MacDonald, C.T., and Gibbs, J.H. (1969). Concerning the kinetics of polypeptide synthesis on polyribosomes. Biopolymers *7*, 707–725.

Mather, W.H., Hasty, J., Tsimring, L.S., and Williams, R.J. (2013). Translational Cross Talk in Gene Networks. Biophysical Journal *104*, 2564–2572.

Muto, H., and Ito, K. (2008). Peptidyl-prolyl-tRNA at the ribosomal P-site reacts poorly with puromycin. Biochemical and Biophysical Research Communications *366*, 1043–1047.

Nackley, A.G., Shabalina, S.A., Tchivileva, I.E., Satterfield, K., Korchynskyi, O., Makarov, S.S., Maixner, W., and Diatchenko, L. (2006). Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science *314*, 1930–1933.

Pavlov, M.Y., Watts, R.E., Tan, Z., Cornish, V.W., Ehrenberg, M., and Forster, A.C. (2009). Slow peptide bond formation by proline and other N-alkylamino acids in translation. Proc. Natl. Acad. Sci. U.S.A. *106*, 50–54.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet *12*, 32–42.

Pop, C., Rouskin, S., Ingolia, N.T., Han, L., Phizicky, E.M., Weissman, J.S., and Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Molecular Systems Biology *10*, 770–770.

Romano, M.C., Thiel, M., Stansfield, I., and Grebogi, C. (2009). Queueing Phase Transition: Theory of Translation. Phys. Rev. Lett. *102*, 198104.

Sagliocco, F.A., Laso, M.R.V., Zhu, D., Tuite, M.F., McCarthy, J.E., and Brown, A.J. (1993). The influence of 5'-secondary structures upon ribosome binding to mRNA during translation in yeast. J. Biol. Chem. *268*, 26522–26530.

Schauder, B., and McCarthy, J.E. (1989). The role of bases upstream of the Shine-Dalgarno region and in the coding sequence in the control of gene expression in Escherichia coli: translation and stability of mRNAs in vivo. Gene *78*, 59–72.

Shaw, L.B., Zia, R.K.P., and Lee, K.H. (2003). Totally asymmetric exclusion process with extended objects: A model for protein synthesis. Phys. Rev. E *68*, 021910.

Stadler, M., and Fire, A. (2011). Wobble base-pairing slows in vivo translation elongation in metazoans. RNA *17*, 2063–2073.

Tuller, T., and Zur, H. (2015). Multiple roles of the coding sequence 5' end in gene expression regulation. Nucl. Acids Res. *43*, 13–28.

Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., and Ziv-Ukelson, M. (2011). Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biology *12*, R110.

Wang, L., and Wessler, S.R. (2001). Role of mRNA secondary structure in translational repression of the maize transcriptional activator Lc(1,2). Plant Physiol. *125*, 1380–1387.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Reports.

Yang, J.-R., Chen, X., and Zhang, J. (2014). Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. PLoS Biol *12*, e1001910.

Zia, R.K.P., Dong, J.J., and Schmittmann, B. (2011). Modeling Translation in Protein Synthesis with TASEP: A Tutorial and Recent Developments. J Stat Phys *144*, 405–428.