Electronic Supplementary Material (ESI) for Molecular BioSystems.
This journal is © The Royal Society of Chemistry 2014

Supplementary of Most Associations between Transcript Features and Gene Expression are Monotonic

Gilad Shaham[1] and Tamir Tuller[1,2]

[1]Department of Biomedical Engineering, the Engineering Faculty, Tel Aviv University, Israel. [2]The Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv.

TT: tamirtul@post.tau.ac.il

Additional transcript features

The following transcript features were analyzed, but we did not discuss them in the main text. The full results of all the transcript features with enough data points are available in tables S1-4

GC Content: the percentage of nitrogenous bases on a DNA molecule that are either guanine or cytosine. GC pairs are bound by three hydrogen bonds, while AT pairs are bound by two hydrogen bonds, thus DNA with high GC-content is more stable than DNA with low GC-content. Four features are the entire 5'UTR the last 40nt of the 5'UTR, the entire 3'UTR and the first 40nt of the 3'UTR. For the ORF there are 101 features, the entire UTR and a window of 40nt for the first one hundred nt. Additional combined feature are the GC content of the entire transcript, the last 40nt of 5'UTR and first 40nt of ORF, entire 5'UTR and first 40nt of ORF, entire 5'UTR and entire ORF, last 40nt of ORF and first 40nt of 3'UTR, last 40nt of ORF and entire 3'UTR, and entire ORF and entire 3'UTR

Predicted folding energy: an approximation of the mRNA secondary structure, the Matlab rnafold function (Matlab Bioinformatics Toolbox), which predicts the folding energy of the secondary structure associated with the minimum free energy for an RNA sequence (or subsequence).The folding energy was estimated for all the sliding windows of length 40nt and averaged the resultant folding energy prediction of all the windows induced by the sequence. Specifically, mean predicted local folding energies used in this study were performed based on the entire transcript of each gene (*i.e.*, including 5'UTR, ORF, and 3'UTR) and for each region (5'UTR, ORF, and 3'UTR) separately.

ATG context scores: According to the canonical model of translation initiation step the Ribosome scans the nucleotides from the 5' end towards the 3' until it recognizes the ATG start codon which represents the beginning of the open reading frame (ORF). The recognition of the start codon triggers the elongation step. However, ATG codons are expected to be present in all possible reading frames, upstream and downstream of the main START ATG. It is possible that the Ribosome would miss the start ATG of the ORF and start translation at a different location, possibly in frame shifts which would result in a completely different amino-acid sequence. Thus, to prevent the translation of undesired proteins, there is selection for nucleotide sequences with low affinity to the pre-initiation complex near the beginning of the ORF. an ATG context score was devised[1] based on their similarity to the sequence context of main ATG START codons of highly expressed genes, in which higher context scores denote similarity to the ATG context of highly expressed genes.

For each of the following segments both the mean and maximum ATG context score and the mean and maximum context score relative to the start ATG were calculated. The calculation was performed for all 3 reading frames separately and combination of the three. The segments selected were: entire 5'UTR, the last 30 codons of the 5'UTR, the entire ORF, first 200 codons of the ORF, first 30 codons of the ORF, entire 3'UTR, last 30 codons of the 5'UTR and first 30 codons of the 3'UTR, last 30 codons of the ORF and first 30 codons of the 3'UTR.

An additional feature per gene was to simply calculate the ATG context of the start ATG.

Number of ATGs: defined for the same regions as those of the context score features.

Distance of first ATG: Calculated for the 5'UTR, 3'UTR and ORF from the main start ATG, for all 3 reading frames separately and combination of the three.

Number of base pairs: A base pair (bp) is the linking between two nitrogenous bases on RNA strands that are connected via hydrogen bonds. In the canonical Watson-Crick DNA base pairing, adenine (A) forms a base pair with uracil (U) and guanine (G) forms a base pair with cytosine (C). The calculation was based on the bracket notation of the measured folding energy (PARS) or predicted measurements and was calculated for the 5'UTR, ORF, 3'UTR and combined for the entire transcript.

Evolutionary rates dN, dS and dN/dS: dN/dS is the ratio of the number of non-synonymous substitutions per non-synonymous site (dN) to the number of synonymous substitutions per synonymous site (dS), which can be used as an indicator of selective pressure acting on a protein-coding gene. Comparisons of homologous genes with a high dN/dS ratio are usually said to be evolving under positive selection. The estimations for dN, dS, dN/dS were taken from Wall *et al.*[2]. These three features are relevant only to the ORF and combined predictors

Relative length ratios: The ratio of the length of the 5'UTR to the ORF, the 3'UTR to the ORF and the 5'UTR to the 3'UTR

Effect of zero/default values

When comparing MIC and Spearman to examine whether the association is monotonic or not, sometimes, there is significant impact as to whether values of zero (or default values) are included in the calculation. **Figure S1** illustrates this case.

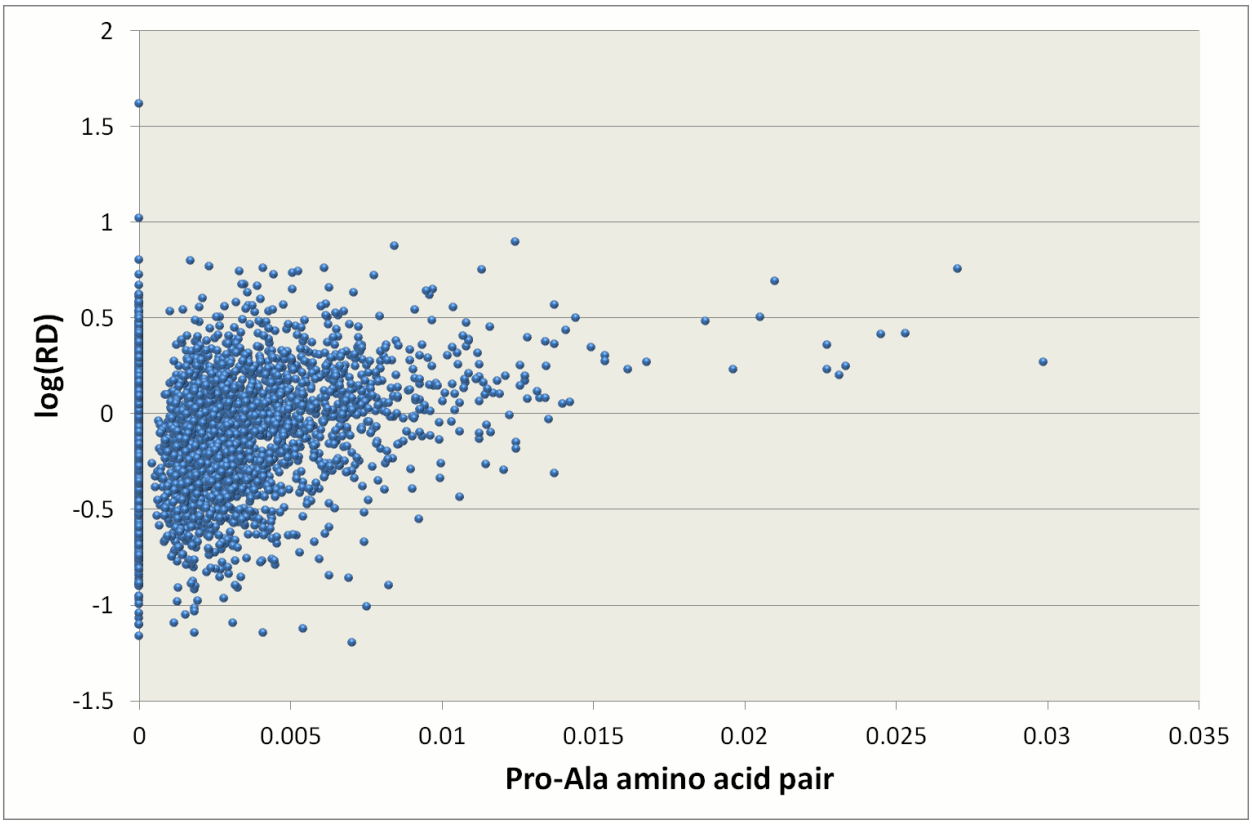

Figure S1 – Example of feature-expression relation between Pro-Ala amino acid pair and RD. Out of a total of 3,682 genes there are 1,566 without this pair. MIC does detect the relationship (MIC value is 0.1061 with empirical p-value $< \mathbf{10^{-5}}$) while the difference to the Spearman score is significant (Spearman is 0.0076, $\boldsymbol{MIC - \rho^2}$ empirical p-value is 0.0004). While considering only non-zero values MIC score remains significant (MIC value is 0.1521 with empirical p-value $< \mathbf{10^{-5}}$), but now the difference is not significant (Spearman is 0.1121, $\boldsymbol{MIC - \rho^2}$ empirical p-value >0.99999).

CAI at nt 33 *vs.* Expression levels

The following figure presents comparative results of CAI of 40 nt in the ORF at position 33. The non-monotonic p-value and MIC p-value are more significant for the PPR case; non-monotonic p-value is not significant in the other cases. The significant MIC values can be explained by a pattern that is not uniformly random; and thus enables a partial/significant prediction of the value of one variable given the second one.

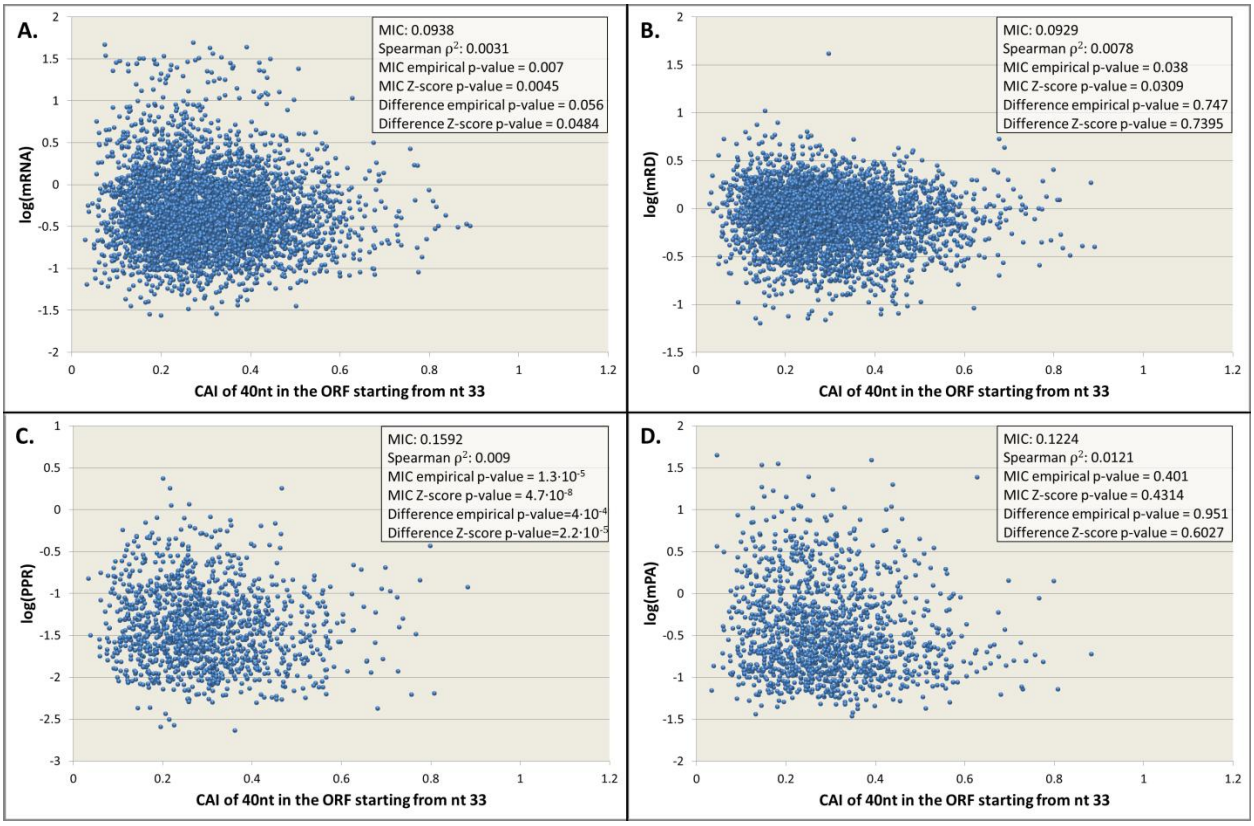

Figure S2 – Comparison of CAI of 40nt in the ORF starting from nt 33 for the different gene expressions. Results are presented for **A.** mRNA, **B.** RD, **C.** PPR, **D.** PA

1. H. Zur and T. Tuller, *PLoS Comput. Biol.*, 2013, **9**, e1003136.
2. D. P. Wall, A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman, *Proc. Natl. Acad. Sci.*, 2005, **102**, 5483–5488.