Int. J. Bioinformatics Research and Applications, Vol.

Detecting slow-translating regions in *E.coli*

Lalit Ponnala

Computational Biology Service Unit,
Cornell University,
Ithaca, NY 14853, USA
E-mail: lp257@cornell.edu

Abstract: We present an application of the spatial scan statistic to identifying clusters of slow-translating codons in *E.coli*. We use an estimate of the time taken to process each codon based on the availability of transfer RNA. We analyse our findings using experimental measurements, and provide support for a well-known hypothesis related to the process of protein production.

Keywords: codon; translation; protein; spatial scan; likelihood.

Reference to this paper should be made as follows: Ponnala, L. (2010) 'Detecting slow-translating regions in *E.coli*', *Int. J. Bioinformatics Research and Applications*, Vol.

Biographical notes: Lalit Ponnala is a Research Associate of the Biotechnology Institute at Cornell University. He received his PhD Degree in 2007 from North Carolina State University and his doctoral thesis was on modelling ribosome-mRNA interaction using signal processing methods. His current research interests include post-transcriptional regulation and the applications of high-throughput sequencing.

1 Introduction

The conversion of DNA to protein involves two intermediate stages: *transcription*, where DNA is converted to RNA, and *translation*, where RNA is converted to protein. During the latter stage (translation of RNA to protein), the choice of codons plays an important role (Lithwick and Margalit, 2003; Karlin et al., 2001), and the concentration of transfer RNAs (tRNAs) becomes critical to the efficiency of protein production (Ikemura, 1985; Varenne et al., 1984). It has been observed that codons having low tRNA availability are used less frequently (Ikemura, 1981) and are translated at a slower rate than other codons (Marin, 2008). Clusters of such 'rare codons' can pause the ribosome for extended periods of time, leading to frameshift, ribosomal drop-off (abrupt termination) or incorrect amino acid incorporation leading to incorrect folding of protein (Wen et al., 2008). Methods to enhance protein production frequently rely on recoding the amino acid sequence using more abundant codons that have higher tRNA availability (Jana and Deb, 2005; Hatfield and Roth, 2007). It is interesting to note that slow-translating codons have been put to constructive use such as in attenuating the spread of viruses (Coleman et al., 2008). More recently, such codons have been

found to affect the solubility of recombinant proteins (Rosano and Ceccarelli, 2009) and give protein domains time to fold (Zhang et al., 2009).

Our aim in this paper is to demonstrate the use of a spatial scan statistic to identify clusters of slow-translating codons. We also examine the validity of the detected clusters by relating their features to measured levels of mRNA and protein in *E.coli*.

2 Methods

It has been shown that the time taken for the arrival and binding of cognate tRNA is inversely proportional to the concentration of that tRNA (Sorensen et al., 1989; Kurland, 1991). Varenne et al. (1984, p.549), state that "the degree of slackening in ribosome movement is almost proportional to the inverse of tRNA concentrations". In order to use these observations quantitatively, we first need to assess the amount of tRNA available to translate each codon.

2.1 Estimation of waiting time

We use published measurements of tRNA concentration in *E.coli* (Dong et al., 1996, Table 5). Based on the mapping between codons and tRNA isoacceptors (Dong et al., 1996, Table 2), we calculate the amount of tRNA available for each codon (r_i) as per the following rules, as described in Zhang et al. (2009):

- if a codon is recognised by more than one isoacceptor, the amount of tRNA available to it is the sum of their concentrations
- if multiple codons are recognised by the same isoacceptor, its concentration is distributed among them in the ratio of their codon usage frequencies

In accordance with the observation of inverse proportionality discussed above, we estimate the 'waiting time' for each codon to be $t_i = 1/r_i$.

We find that using tRNA concentrations from any of the five growth-rates (Dong et al., 1996, Table 5) yields substantially similar results. In this paper, we use t_i values calculated from tRNA concentrations averaged across growth rates.

The level of waiting time varies significantly across codons (see Table 1), and, as explained earlier, the codons having high waiting time are found to be rarely used in the genome.

Table 1 Estimated waiting time for each codon in *E. coli*

Codon	*Estimated waiting time*	*Codon*	*Estimated waiting time*
GCU	0.306576	AUU	0.123096
GCA	0.231183	AUA	0.924197
GCG	0.136638	CUG	0.049041
GCC	0.379939	CUC	0.468743
CGU	0.117252	CUU	0.477867
CGC	0.111511	CUA	5.466705
CGA	0.720644	UUG	0.105162

Table 1 Estimated waiting time for each codon in *E. coli* (continued)

Codon	*Estimated waiting time*	*Codon*	*Estimated waiting time*
CGG	0.470367	UUA	0.537975
AGA	0.344590	AAA	0.165436
AGG	0.551268	AAG	0.545443
AAC	0.355854	AUG	0.081288
AAU	0.441723	UUC	0.559073
GAC	0.258931	UUU	0.416090
GAU	0.154800	CCG	0.239277
UGC	0.306448	CCC	0.769168
UGU	0.387176	CCU	0.486752
CAA	0.306560	CCA	2.091173
CAG	0.241896	UCA	0.591308
GAA	0.071980	UCU	0.280434
GAG	0.160381	UCG	0.294545
GGA	0.276070	AGC	0.314375
GGG	0.195399	AGU	0.586482
GGC	0.101103	UCC	0.632165
GGU	0.121756	ACC	0.286792
CAC	0.796637	ACU	0.392728
CAU	0.602873	ACG	0.229557
AUC	0.148451	ACA	1.038801
UGG	0.273523	GUA	0.342367
UAC	0.302471	GUG	0.140865
UAU	0.230845	GUU	0.127673
GUC	0.406562		

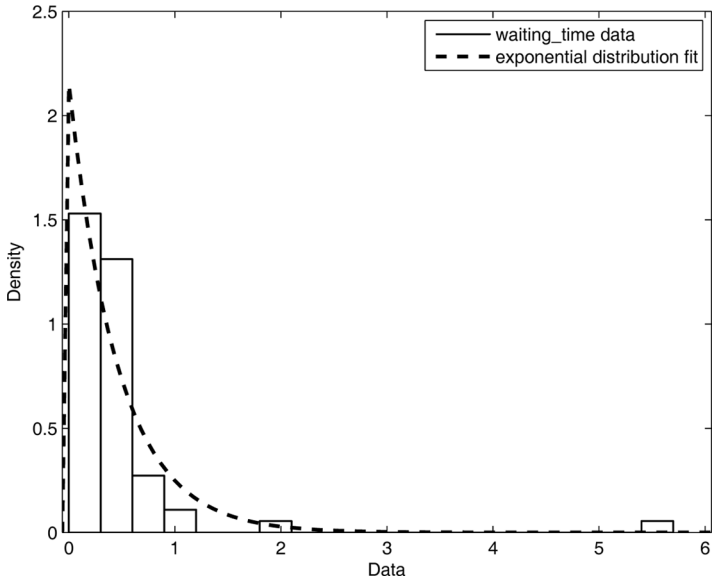
2.2 Cluster detection

In order to find clusters of slow-translating codons, we use the spatial scan method proposed in Huang et al. (2007). We are interested in finding clusters of consecutive codons along a gene that collectively have a higher average waiting time than the rest of the codons in the same gene.

It seems reasonable to assume that the estimated waiting time of each codon, t_i, follows an exponential distribution (see Figure 1). Instead of assigning binary values $(1/0)$ to denote the presence of slow-translating codons, using the estimated waiting time has some advantages. It circumvents the problem of having to choose a cut-off point and prevents the loss of information inherent in converting continuous measurements to discrete values (Huang et al., 2007).

Let us consider a gene G that has N codons. The spatial scan method begins by choosing concentric (one-dimensional) 'zones' around each codon. Let n_{in} denote the number of codons in a zone Z and let θ_{in} denote the average waiting time of all these codons. Correspondingly, we use n_{out} to denote the number of codons outside Z, with θ_{out} as their average waiting time. In our statistical test, the null hypothesis states that $\theta_{\text{in}} = \theta_{\text{out}}$, i.e., the average waiting time is the same inside and outside the selected zone. The alternative hypothesis is that codons within zone Z have a higher waiting time leading to a slow-translating cluster, i.e., $\theta_{\text{in}} > \theta_{\text{out}}$.

Figure 1 Histogram of estimated waiting time and exponential distribution fit, log-likelihood $= -14.1$



The likelihood that a zone Z contains a cluster of slow-translating codons is given by the ratio (Huang et al., 2007):

$$\lambda(Z) = \frac{\left(\frac{n_{in}}{\sum_{i\in Z} t_i}\right)^{n_{\text{in}}} \left(\frac{n_{\text{out}}}{\sum_{i\notin Z} t_i}\right)^{n_{\text{out}}}}{\left(\frac{N}{\sum_{i\in G} t_i}\right)^{N}} I\left(\sum_{i\in Z} t_i/n_{\text{in}} > \sum_{i\notin Z} t_i/n_{\text{out}}\right)$$

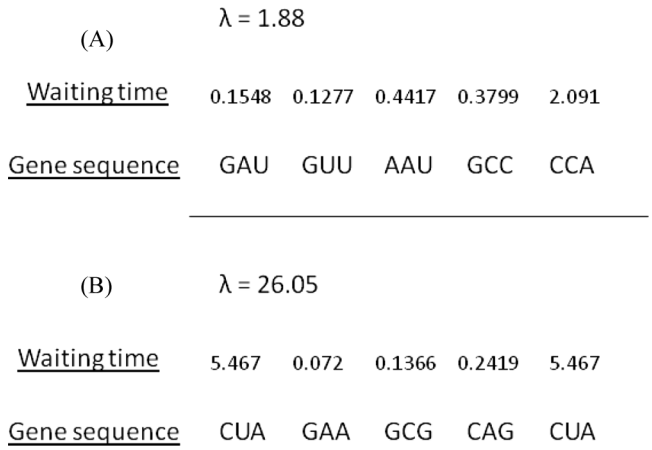
where $I(\cdot)$ is the indicator function. Note that in this formula, the likelihood is invariant to scaling the estimated waiting time, i.e., even if the t_i values were multiplied by a constant K, the calculated $\lambda(Z)$ would remain unchanged. The likelihood $\lambda(Z)$ quantifies the waiting time of a set of consecutive codons Z relative to the rest of the gene. It can be used a metric to compare various zones in terms of their overall waiting time.

By incrementing the zone-width by one codon in each direction, we can calculate the likelihood of all possible zones. Ideally we should consider sets of candidate zones centred on each and every codon, but this has two disdvantages:

- it drastically increases the run-time of the algorithm
- it could lead to detection of *spurious* clusters (i.e., regions that do not contain any genuine slow-translating codons, but contain codons that are translated *relatively* slower than the other codons in the gene), which may not be meaningful biologically (see Figure 2).

In our implementation of the spatial scan algorithm, we evaluate the likelihood of zones centred on codons having $t_i > 0.1$, which includes all but three of the most abundant codons (*GAA*, *CUG* and *AUG*).

Figure 2 (A) A spurious cluster found in the gene *leuE*, containing codons whose waiting times are not very high and (B) a genuine cluster found in the gene *yheS*



According to the method described in Huang et al. (2007), significant clusters are to be found via Monte-Carlo testing. The distribution of the likelihood ratio λ under the null hypothesis must be simulated by creating random permutations of the waiting times. The problem with applying this procedure in our case is that the amino acid sequence may no longer be preserved in the permutations we create. A codon at position i may not be replaceable by another codon that codes for the same amino acid and has the randomly-assigned value of waiting time! So instead of creating random permutations, we take the top 100 most-likely (highest λ) zones, and process them by applying the following rules, in order:

- retain zones of higher likelihood
- if likelihood is the same, retain the smaller zone
- if likelihood is the same and the zones are of the same size, then retain the one having larger total waiting time $\sum_{i \in Z} t_i$.

These rules are based on the reasoning that we are interested in finding small, dense clusters of slow-translating codons, as opposed to large regions containing only a few such codons. We follow the above rules in order to eliminate cases where a zone contains or is contained in another zone or overlaps with it for more than half its size. The candidate zones that pass the above steps are identified as distinct slow-translating codon clusters.

3 Results

We randomly selected genes of varying length from the *E.coli* genome, GenBank accession NC_000913. With each codon we associated a processing time t_i, as shown in Table 1. We excluded stop codons and genes having frameshift (*prfB*) from our analysis. We then performed cluster detection in each gene using software code written in MATLAB (see http://sites.google.com/site/jbrpaper/).

We found at least one cluster of slow-translating codons in each of the 494 genes we examined. The detected clusters had an average size of about 7 codons, and a standard-deviation of 11 codons, indicating a wide range of sizes. The observation that some of the detected clusters tend to be large is in agreement with Clarke IV and Clark (2008). The largest cluster was found in the gene *fliC*, which encodes a structural protein (flagellin). It had a size of 181 codons and an average waiting time of 0.35. In contrast, the cluster that had the highest average waiting time of 4.44 was found to span 5 codons and was detected in the gene *leuL*, which produces an operon leader peptide. The *eno* (enolase) gene was found to contain the maximum number of slow-translating clusters (15).

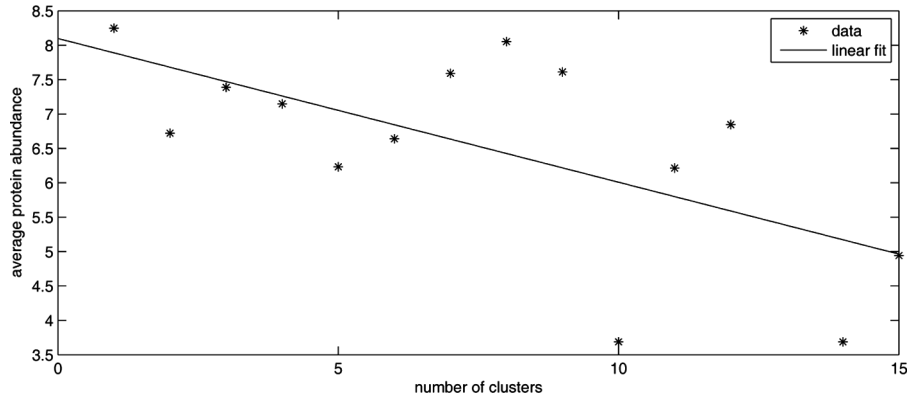
The average distance between non-overlapping clusters was found to be about 80 codons. Longer genes (> 300 aa in length) were found to have a slightly higher number of clusters overall. Roughly 22% of all detected clusters occur near gene-starts, i.e., within the first 50 aa. We observed similar trends when genes of varying physiological role (as per the classification in Riley (1998)) were examined closely.

As explained earlier, the clustering of slow-translating codons tends to suppress the protein levels of expressed genes, i.e., it has a negative impact on mRNA-protein conversion. To test the validity of this observation in light of our detected clusters, we used measurements of mRNA expression level (Selinger et al., 2000) and protein abundance (Link et al., 1997) in *E.coli*. These were the only datasets we could find where mRNA and protein measurements were made under similar experimental conditions, with the *E.coli* cells in stationary phase. Using the formula given in Selinger et al. (2000, p.1267), we estimated the number of RNA copies per cell from the published SP-intensities (http://arep.med.harvard.edu/ExpressDB/EDS37/). The N-abd values provided in Link et al. (1997, Table 4) quantify protein abundance in number of molecules per cell. Since both the mRNA level and protein abundance measurements are expressed in *per cell* units, we can compare them gene-wise without any further data-processing. Based on name-match with the GenBank annotation, we were able to extract all the relevant information for 92 genes in *E.coli*. We detected clusters in all of these genes, varying in number (N_{clust}) from 1 to 15.

We found that genes containing more slow-translating clusters (higher values of N_{clust}) had lesser protein abundance on the average (see Figure 3), but this could also be due to lower levels of mRNA in those genes. In order to clearly quantify the relationship between protein abundance and the number of clusters, we need to measure the correlation between them while controlling for mRNA level. Partial correlation analysis enables us to do this without having to take the ratio of potentially noisy variables (Chen and Zheng, 2009).

For each value of N_{clust}, we calculated the average protein abundance and average mRNA level of genes that had the specified number of clusters. We then evaluated the partial correlation between N_{clust} and protein-abundance while controlling for mRNA-level, and found that it is significantly negative: $\rho = -0.614585$, p-value $= 0.025411$. This finding validates the hypothesis proposed in earlier work (Kane, 1995) that the presence of slow-translating codon clusters tends to hamper protein production. By detecting clusters of slow translating codons using a spatial scan statistic, we are able to support the observations from earlier studies, and validate them with available experimental data.

Figure 3 Relationship between slow-translating clusters and protein yield (without controlling for mRNA level)



4 Discussion

It is easy to identify the presence of slow-translating codons in a gene sequence using some measure of processing time such as codon usage frequency or tRNA concentration. The challenge is to detect *clusters* of such codons, especially when they are embedded among other abundantly-used codons. As discussed in Widmann et al. (2008), clustered slow-translating codons have a greater effect on translational speed than when they are dispersed. There have been studies of the stochastic mechanisms that govern the translation rate, as a function of the spacing between slow-translating codons (Chou and Lakatos, 2004; Dong et al., 2007). While such mechanisms do seem to influence the translation rate, their gross effects are seen in their adverse influence on protein yield. Earlier studies looked specifically for small clusters (Thanaraj and Argos, 1996) or found small clusters (less than 5 codons) by window-averaging which requires assumptions on window size (Zhang and Ignatova, 2009; Clarke IV and Clark, 2008). We have not made any prior assumption on cluster size, and so our algorithm reports all possible clusters, many of which tend to be wider than reported in other studies. Not all yield-impacting clusters contain small, dense collections of slow-translating codons. Clusters that contain other codons can also collectively hamper translation. Such clusters, detected by our spatial-scan method, do have an adverse effect on protein yield as we have shown using experimental measurements.

We have made some assumptions and approximations in the course of our analysis. First of all, we have ignored the time taken for transpeptidation and translocation from our estimates of codon waiting time. As a first approximation, this seems acceptable since (Zhang et al., 2009, p.275) state that "the rate of elongation would be mainly limited by the acquisition of the cognate ternary complex, whereas transpeptidation and tRNA translocation occur much faster". We have also assumed no effect of secondary structure on our waiting time estimates. This too seems reasonable because, according to Sorensen et al. (1989, p.376), the "rates of translating ribosomes are indifferent to the presence of secondary structure in mRNA". It must be noted that the experimental conditions under which the mRNA and protein measurements are made are similar but not identical, and estimating RNA copy number from signal intensity is not without error. The rate of initiation does affect the throughput of

translation, but according to Lithwick and Margalit (2003), this effect (as measured by the binding strength in the Shine-Dalgarno region) is not of much importance in determining translation efficiency. They state that "biased codon usage is the property that is most highly associated with protein expression and that is most conserved".

In spite of novel discoveries supporting other claims, the presence of excessive slow-translating codon clusters negatively impacts protein yield. This seems to be the summary of our results. Further research in this direction could shed light on the processes that are critical for efficient protein production. Looking at the distribution of the identified slow-translating clusters in 3D space might enable a clearer distinction between yield-suppressing clusters and clusters that serve a useful purpose such as enabling proteins to fold. Methods for predicting the abundance of undetected proteins (Nie et al., 2006) do not explicitly consider the occurence of slow-translating codons. This is another area where our demonstrated method could be of use.

References

Chen, L. and Zheng, S. (2009) 'Studying alternative splicing regulatory networks through partial correlation analysis', *Genome. Biol.*, Vol. 10, No. R3, doi:10.1186/gb-2009-10-1-r3.

Chou, T. and Lakatos, G. (2004) 'Clustered bottlenecks in mRNA translation and protein synthesis', *Phys. Rev. Lett.*, Vol. 93, No. 19, p.198101.

Clarke IV, T.F. and Clark, P.L. (2008) 'Rare codons cluster', *PLoS ONE*, Vol. 3, No. 10, p.e3412, doi:10.1371/journal.pone.0003412.

Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E. and Mueller, S. (2008) 'Virus attenuation by genome-scale changes in codon pair bias', *Science*, Vol. 320, No. 5884, pp.1784–1787.

Dong, H., Nilsson, L. and Kurland, C.G. (1996) 'Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates', *J. Mol. Biol.*, Vol. 260, No. 5, pp.649–663.

Dong, J., Schmittmann, B. and Zia, R. (2007) 'Towards a model for protein production rates', *J. Stat. Phys.*, Vol. 128, pp.21–34.

Hatfield, G.W. and Roth, D.A. (2007) 'Optimizing scaleup yield for protein production: computationally optimized DNA Assembly (CODA) and translation engineering', *Biotechnol. Annu. Rev.*, Vol. 13, pp.27–42.

Huang, L., Kulldorff, M. and Gregorio, D. (2007) 'A spatial scan statistic for survival data', *Biometrics*, Vol. 63, No. 1, pp.109–118.

Ikemura, T. (1981) 'Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system', *J. Mol. Biol.*, Vol. 151, No. 3, pp.389–409.

Ikemura, T. (1985) 'Codon usage and tRNA content in unicellular and multicellular organisms', *Mol. Biol. Evol.*, Vol. 2, No. 1, pp.13–34.

Jana, S. and Deb, J.K. (2005) 'Strategies for efficient production of heterologous proteins in Escherichia coli', *Appl. Microbiol. Biotechnol.*, Vol. 67, No. 3, pp.289–298.

Kane, J.F. (1995) 'Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli', *Curr. Opin. Biotechnol.*, Vol. 6, No. 5, pp.494–500.

Karlin, S., Mrázek, J., Campbell, A. and Kaiser, D. (2001) 'Characterizations of highly expressed genes of four fast-growing bacteria', *J. Bacteriol.*, Vol. 183, No. 17, pp.5025–5040.

Kurland, C.G. (1991) 'Codon bias and gene expression', *FEBS Lett.*, Vol. 285, No. 2, pp.165–169.

Link, A.J., Robison, K. and Church, G. (1997) 'Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli', *Electrophoresis*, Vol. 18, pp.1259–1313.

Lithwick, G. and Margalit, H. (2003) 'Hierarchy of sequence-dependent features associated with prokaryotic translation', *Genome Res.*, Vol. 13, No. 12, pp.2665–2673.

Marin, M. (2008) 'Folding at the rhythm of the rare codon beat', *Biotechnol. J.*, Vol. 3, No. 8, pp.1047–1057.

Nie, L., Wu, G., Brockman, F.J. and Zhang, W. (2006) 'Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins', *Bioinformatics*, Vol. 22, No. 13, pp.1641–1647.

Riley, M. (1998) 'Genes and proteins of Escherichia coli K-12 (GenProtEC)', *Nucleic Acids Res.*, Vol. 6, p.54.

Rosano, G.L. and Ceccarelli, E.A. (2009) 'Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain', *Microb. Cell Fact.*, Vol. 8, No. 41, doi:10.1186/1475-2859-8-41.

Selinger, D.W., Cheung, K.J., Mei, R., Johansson, E.M., Richmond, C.S., Blattner, F.R., Lockhart, D.J. and Church, G.M. (2000) 'RNA expression analysis using a 30 base pair resolution Escherichia coli genome array', *Nat. Biotechnol.*, Vol. 18, No. 12, pp.1262–1268.

Sorensen, M.A., Kurland, C.G. and Pedersen, S. (1989) 'Codon usage determines translation rate in escherichia coli', *J. Mol. Biol.*, Vol. 207, No. 2, pp.365–377.

Thanaraj, T.A. and Argos, P. (1996) 'Ribosome-mediated translational pause and protein domain organization', *Protein Sci.*, Vol. 5, No. 8, pp.1594–612.

Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984) 'Translation is a non-uniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains', *J. Mol. Biol.*, Vol. 180, No. 3, pp.549–576.

Wen, J.D., Lancaster, L., Hodges, C., Zeri, A.C., Yoshimura, S.H., Noller, H.F., Bustamante, C. and Tinoco, I. (2008) 'Following translation by single ribosomes one codon at a time', *Nature*, Vol. 452, No. 7187, pp.598–603.

Widmann, M., Clairo, M., Dippon, J. and Pleiss, J. (2008) 'Analysis of the distribution of functionally relevant rare codons', *BMC Genomics*, Vol. 9, No. 207, doi: 10.1186/1471-2164-9-207.

Zhang, G. and Ignatova, Z. (2009) 'Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis', *PLoS One*, Vol. 4, No. 4, p.e5036.

Zhang, G., Hubalewska, M. and Ignatova, Z. (2009) 'Transient ribosomal attenuation coordinates protein synthesis and co-translational folding', *Nat. Struct. Mol. Biol.*, Vol. 16, No. 3, pp.274–280.