January 2010

The length distributions of non-coding and coding sequences in relation to gene expression: A study on Arabidopsis thaliana

R Caldwell  
*University of Wollongong, rac34@uowmail.edu.au*

J Kongcharoen  
*University of Wollongong, jk117@uowmail.edu.au*

Y Lin  
*University of Wollongong, yanxia@uow.edu.au*

R Zhang  
*University of Wollongong, rzhang@uow.edu.au*

Follow this and additional works at: [https://ro.uow.edu.au/scipapers](https://ro.uow.edu.au/scipapers)

Part of the Life Sciences Commons, Physical Sciences and Mathematics Commons, and the Social and Behavioral Sciences Commons

**Recommended Citation**

Caldwell, R; Kongcharoen, J; Lin, Y; and Zhang, R: The length distributions of non-coding and coding sequences in relation to gene expression: A study on Arabidopsis thaliana 2010, 127-133.  
https://ro.uow.edu.au/scipapers/505

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
The length distributions of non-coding and coding sequences in relation to gene expression: A study on Arabidopsis thaliana

Keywords
distributions, coding, length, study, non, gene, expression, relation, arabidopsis, sequences, thaliana

Disciplines
Life Sciences | Physical Sciences and Mathematics | Social and Behavioral Sciences

Publication Details

This conference paper is available at Research Online: https://ro.uow.edu.au/scipapers/505
The Length Distributions of Non-Coding and Coding Sequences in Relation to Gene Expression: A Study on Arabidopsis thaliana

R. Caldwell1, J. Kongcharoen2, Y. Lin2, and R. Zhang1
1School of Biological Sciences, University of Wollongong, Wollongong, NSW, Australia
2School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia

Abstract - The availability of complete genomic sequences as well as other gene related databases, such as gene expression datasets, allows for extended exploration of the production of gene products. Using the length distribution data of the coding and non-coding regions of a gene and it corresponding gene expression values the aim was to determine if there was a relationship between these values. Pearson’s correlation showed that the coding and 5’ UTR lengths are negatively correlated in relation to gene expression intensity, however the 3’ UTR length showed positive correlation. This was confirmed by using quantile regression analysis on the data. The results support the role and importance of these length regions in the regulation of gene expression.

Keywords: non-coding sequence, coding sequence, untranslated regions, gene expression, quantile regression

1 Introduction

Gene expression is a process pivotal to all living organisms. The process involves the fabrication of functional gene products such as proteins, which is template by DNA via messenger RNA. The process involves several phases, and has strict levels of control at the transcriptional and translational initiation and elongation stages, proving more complex in eukaryotes than prokaryotes. Importance in the stability of the mRNA, localization, and translational efficiency has been allied with the un-translated regions of the gene, the 5’ and 3’ UTR [1] [2]. The production and regulation of gene products is a very important and complicated process that implores attention.

The availability of numerous complete genomic sequences, as well as the gene expression data offered from microarray expression experiments permits advanced statistical and biological analysis in relation to gene expression. Much research has been published on the analysis of microarray data, focusing on classification and clustering of data [3] [4]; influence of introns and un-translated regions (UTRs) [5] [6] [7]; codon usage, composition and evolution [8] [9] [10] [11]. Variations in protein size, coding and non-coding sequence length and intron number and size differ significantly among living organisms. Research has found within complex organisms such as Homo sapiens and Caenorhabditis elegans that highly expressed genes typically boast fewer and shorter introns, shorter coding sequences, shorter intergenic regions, shorter mRNA and shorter 3’ UTRs [1]. In other animal studies, it has been reported that there has been negative correlations between gene expression and coding sequence and protein length, indicating that highly expressed genes are smaller in length [12] [8] [9]. Ren [10] established that within plant genomes highly expressed genes comprise additional and longer introns, and primary transcripts, suggesting highly expressed genes are larger, in contrast to animal genes.

The intention of this research was to take the non-coding and coding sequence length data obtained from the internet of the Arabidopsis thaliana and use it to examine if there is a relationship between the length distributions of these gene regions and gene expression intensity.

2 Materials and Methods

The gene length was divided into three sections, and data was collected for each region including and excluding introns. The distances were measured in base pairs (bp) of the nucleotide sequence, and the regions were separated into coding and non-coding regions. The first region is situated between the Translation Start Site (TLS) and the Translation Stop Codon (TSC). This region will be referred to as D₁ or coding region length (TLS-TSC distance). The second region encompasses the +1 position after the promoter (the Transcription Start Site (TSS)) to the last nucleotide before the TLS. This region will be referred to as D₂ (TSS-TLS distance). The third region is situated between the translation stop codon (TSC) and the Transcription Termination Site (TTS), and will be referred to as D₃ (TSC-TTS distance). The data collected without introns is denoted as d₁, d₂ and d₃ respectively [13].
2.1 Data Sources

Collection of data for *Arabidopsis thaliana* was extracted from public databases from the world-wide-web. The NCBI reference sequence assembly database provided a comprehensive table on the coding sequence data [http://www.ncbi.nlm.nih.gov/]. To obtain the coding sequence (CDS) length, the start and stop positions referenced from the genome were exploited. The coding sequence lengths that did not contain introns were calculated from the protein length values.

The un-translated regions for *Arabidopsis* (3' and 5' ends) (March 2008), the intron data (June 2009), and the Microarray data [ftp://ftp.arabidopsis.org/home/tair/Microarrays/AFGC/] were collected from the TAIR website [http://www.arabidopsis.org/help/helppages/BLAST_help.jsp #datasets]. The average intensity values from the microarray data were used in this study, and were calculated from a simple average value from all of the AFGC microarray experiments. The average intensity value represents a large range of conditions and tissue types.

2.2 Statistical Analysis

Pearson’s correlation was applied to the datasets for testing the degree of linear relationship between the variables gene expression and the length of each gene region. Calculations were acquired from SPSS at a significance level of 0.1.

Median and mean values were compared due to the nature of data. Skewness and kurtosis formulas were calculated and were used to measure the observations that were clustered around a central point, or to measure the asymmetry of the distribution.

The skewness and kurtosis analysis on the datasets identified left and right skewness in the data, no data was normally distributed. Therefore, to compare means, the Kruskal-Wallis test was performed to compare three or more independent groups of sampled data, which makes no assumptions about the distribution of the data.

In ecology, quantile regression has been proposed and used as a way to discover more useful predictive relationships between variables in cases where there is no relationship or only a weak relationship between the means of such variables [14]. After preliminary analysis of the length distributions and gene expression using standard Pearson’s correlation, quantile regression was used to extend the effect of gene length distribution on the average gene expression intensity. This type of analysis exposes the influence of independent variable(s) on a dependent variable in terms of variation range and conditional distribution status in greater depth [15].

Quantile regression models are used in this study to model average gene expression on the length of non-coding regions (3' UTR and 5' UTR's) and coding regions for *Arabidopsis thaliana* using the dataset without introns. To build up an appropriate quantile regression model for the average gene expression intensity and the length of coding region dataset, we started with the linear quantile regression model. Then we tested the quadratic, the cubic and higher order quantile regressions until an appropriate model was found. The Akaike Information Criterion (AIC) was used as a criterion to assist in the model selection. The following models are used to fit *Arabidopsis thaliana* dataset:

\[
\begin{align*}
Q_{\text{int}}(\tau | d_1) &= \beta_0(\tau) + \beta_1(\tau)d_1 + \beta_2(\tau)d_1^2 + \epsilon(\tau) \\
Q_{\text{int}}(\tau | d_2) &= \beta_0(\tau) + \beta_1(\tau)d_2 + \epsilon(\tau) \\
Q_{\text{int}}(\tau | d_3) &= \beta_0(\tau) + \beta_1(\tau)d_3 + \beta_2(\tau)d_3^2 + \epsilon(\tau)
\end{align*}
\]

where \(Q_{\text{int}}(\tau | d_1), Q_{\text{int}}(\tau | d_2), \text{ and } Q_{\text{int}}(\tau | d_3)\) are the \(\tau^\text{th}\) quantile of the average gene expression intensity on the length of coding region, the length of 5' UTR region and the length of 3' UTR region covariates respectively. \(\beta_0(\tau), \beta_1(\tau), \beta_2(\tau)\) are unknown parameters in the model and need to be estimated. \(\epsilon(\tau)\) is the error term in the model; \(0 < \tau < 1\).

Equations (1) and (3) are quadratic quantile regression models of the average gene expression intensity on the length of coding region and the length of 3' UTR region respectively. Equation (2) is a linear quantile regression model of the average gene expression intensity on the length of 5' UTR region.

3 Results

3.1 The average expression level of genes in relation to the length of the coding sequence region

For the coding sequence, the length data was split into 5 length categories (Table 1). In the lowest category ≤ 100 there were no values obtained, as the coding sequence started at length values above 100. In each of the datasets that included (data not shown) and excluded introns it was observed that the shorter the coding sequence length for each gene, there was an increase in the average intensity (gene expression). Pearson correlation confirmed this with negatively significant results (\(r = -0.108\) without introns; \(r = -0.87\) with introns).

3.2 The average expression level of genes in relation to the length of the 5' UTR region

The data for the 5' UTR lengths (Table 2) show a similar trend as seen in the coding sequence data, with the shorter the length of the 5' UTR the higher the average intensity values. When Pearson's correlation was applied to this dataset, significant negative correlations were also
observed ($r = -0.045$ without introns; $r = -0.036$ with introns). The dataset without introns exhibited a larger variation in gene expression from each length category than the dataset with introns (data not shown). It was also observed that the length values from greater than 100 to less than or equal to 1000 did not vary considerably in the average intensity values.

### 3.3 The average expression level of genes in

**Table 1** – Coding sequence of *Arabidopsis thaliana* without introns. The average length (bp) in comparison to average intensity (gene expression) in 5 length regions (sample size $N=17,405$).

<table>
<thead>
<tr>
<th>Length (bp) of Coding Sequence (d1)*</th>
<th>Gene Number</th>
<th>Average Length</th>
<th>Median Length</th>
<th>Average Intensity</th>
<th>Median Intensity</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 100, $\leq$ 250</td>
<td>70</td>
<td>210</td>
<td>210</td>
<td>11746</td>
<td>8803</td>
<td>-0.289</td>
<td>0.873</td>
</tr>
<tr>
<td>&gt; 250, $\leq$ 500</td>
<td>1478</td>
<td>407</td>
<td>417</td>
<td>9406</td>
<td>7786</td>
<td>-0.409</td>
<td>1.155</td>
</tr>
<tr>
<td>&gt; 500, $\leq$ 1000</td>
<td>5057</td>
<td>778</td>
<td>789</td>
<td>8162</td>
<td>5658</td>
<td>-0.260</td>
<td>1.464</td>
</tr>
<tr>
<td>$\geq$ 1001</td>
<td>19800</td>
<td>1619</td>
<td>1416</td>
<td>7720</td>
<td>5051</td>
<td>3.706</td>
<td>1.739</td>
</tr>
</tbody>
</table>

Pearson Correlation: $r = -0.108$* (*Correlation is significant at the 0.01 level (2-tailed))

*No sequences fall into $\leq$ 100 bp subset

**Table 2** – 5’ Un-translated region of *Arabidopsis thaliana* without introns. Average length (bp) in comparison to average intensity (gene expression). Data was split into 5 regions, with a sample size of $N=17,405$.

<table>
<thead>
<tr>
<th>Length (bp) of 5’ UTR (d2)</th>
<th>Gene Number</th>
<th>Average Length</th>
<th>Median Length</th>
<th>Average Intensity</th>
<th>Median Intensity</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq$ 100</td>
<td>6833</td>
<td>65</td>
<td>70</td>
<td>8589</td>
<td>5804</td>
<td>-0.677</td>
<td>1.522</td>
</tr>
<tr>
<td>&gt; 100, $\leq$ 250</td>
<td>7612</td>
<td>157</td>
<td>146</td>
<td>7548</td>
<td>5308</td>
<td>0.562</td>
<td>1.643</td>
</tr>
<tr>
<td>&gt; 250, $\leq$ 500</td>
<td>2445</td>
<td>332</td>
<td>314</td>
<td>7658</td>
<td>4910</td>
<td>0.852</td>
<td>1.368</td>
</tr>
<tr>
<td>&gt; 500, $\leq$ 1000</td>
<td>485</td>
<td>645</td>
<td>589</td>
<td>7637</td>
<td>4808</td>
<td>0.909</td>
<td>1.647</td>
</tr>
<tr>
<td>$\geq$ 1001</td>
<td>30</td>
<td>2312</td>
<td>2424</td>
<td>3881</td>
<td>3119</td>
<td>-0.152</td>
<td>4.509</td>
</tr>
</tbody>
</table>

Pearson Correlation: $r = -0.045$* (*Correlation is significant at the 0.01 level (2-tailed))

**Table 3** – 3’ Un-translated region of *Arabidopsis thaliana* without introns. The average length (bp) in comparison to average intensity (gene expression). Data was split into 5 length regions (Sample size $N=17,405$).

<table>
<thead>
<tr>
<th>Length (bp) of 3’ UTR (d3)</th>
<th>Gene Number</th>
<th>Average Length</th>
<th>Median Length</th>
<th>Average Intensity</th>
<th>Median Intensity</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq$ 100</td>
<td>541</td>
<td>65</td>
<td>74</td>
<td>6878</td>
<td>3663</td>
<td>-0.625</td>
<td>1.426</td>
</tr>
<tr>
<td>&gt; 100, $\leq$ 250</td>
<td>7949</td>
<td>192</td>
<td>197</td>
<td>7044</td>
<td>4738</td>
<td>-0.464</td>
<td>1.715</td>
</tr>
<tr>
<td>&gt; 250, $\leq$ 500</td>
<td>7614</td>
<td>335</td>
<td>318</td>
<td>8852</td>
<td>6145</td>
<td>0.747</td>
<td>1.503</td>
</tr>
<tr>
<td>&gt; 500, $\leq$ 1000</td>
<td>1161</td>
<td>631</td>
<td>599</td>
<td>9382</td>
<td>6239</td>
<td>1.010</td>
<td>1.508</td>
</tr>
<tr>
<td>$\geq$ 1001</td>
<td>140</td>
<td>1327</td>
<td>1212</td>
<td>9782</td>
<td>7302</td>
<td>0.481</td>
<td>0.829</td>
</tr>
</tbody>
</table>

Pearson Correlation: $r = 0.105$* (*Correlation is significant at the 0.01 level (2-tailed))

**relation to the length of the 3’ UTR region**

The analysis for the 3’ UTR data (Table 3) showed a very distinctive result. The shorter the 3’ UTR length the lower the average gene expression. Therefore indicating a positive correlation. The Pearson correlation $r$ values were also positively significant ($r = 0.105$ without introns; $r = 0.065$ with introns).
3.4 Quantile regression analysis – lengths of coding and non-coding regions

Quantile regression was conducted on the length data for the coding region, 5'UTR region and 3'UTR region in relation to gene expression (Figure 1 (a), (b), (c)). The coefficients for \( d_1 \) and \( d_2 \) in models (1) and (2) are negative for all quantile cases, however the coefficients for \( d_3 \) (in model (3)) are positive. This indicates that the length of the coding region and the length of 5'UTR region (without introns) are negatively related to the quantiles of the average gene expression intensity while the length of 3'UTR region (without introns) are positively related. The patterns observed (Figure 1) shows the values of the quantile of the average gene expression intensity decreases as the value of \( d_1 \) (a) or \( d_2 \) (b) increases. However, as the value of \( d_3 \) (c) increases so does the value of the quantiles of the average gene expression intensity increase only in the length range of 0 to 1000 bp. As \( d_3 \) increases after 1000 bp the quantile of the average gene expression intensity decreases. Therefore, the larger the quantile, the faster the quantile line proceeds down, \( d_3 \) increases, while the quantile lines are steadier for \( d_2 \). After initial increases, the average gene expression intensity decreases as \( d_3 \) increases.

![Figure 1](image1.png)

**Figure 1**: The quantile curves of the average gene expression intensity on the length of coding region (a); the length of 5'UTR region (b) and the length of 3'UTR region (c). The conditional quantiles include the range of 0.3 to 0.7 in quantile increments of 0.1.

4 Discussion

Our study using the average gene expression intensity data of *Arabidopsis thaliana* has verified previous research [9] [12] [16] that there is negative correlation between the length of the coding sequence (\( d_1 \)) as well as the 5' un-
translated region \((d_1)\) and gene expression levels. Further analysis has also found that the 3′ UTR showed a positive correlation. The use of a non-linear model to identify correlations between these regions has been demonstrated here. Previous research conducted by us found that there is a non-linear function relationship between the coding sequence length and the 5′ UTR region [13], and supports the fact that there is a non-linear relationship in the Arabidopsis data in relation to gene expression. Using quantile regression modeling, to further test this correlation, it has confirmed the results, and is capable of aiding in the investigation of coding and non-coding length distributions on gene expression.

Negative correlations were found between the length of the 5′ UTR and coding sequence and gene expression. Raghava [9] reported significant negative correlation in the expression levels and gene length for Saccharomyces cerevisiae, with an r value of -0.18. The observations of the 5′ UTR and coding sequence indicate that for Arabidopsis thaliana may be subject to evolutionary constraints in the management of gene expression. Longer 5′ UTR regions in eukaryotes can produce defective proteins due to a higher instance of mutation to the translation-initiation codons [17]. A theory many have considered is that to reduce the cost of energy in gene expression, natural selection supports shorter proteins and shorter introns [18]. This could undoubtedly be the circumstance with large protein lengths, which could impact on the energy cost of biosynthesis, shorter protein lengths could contribute to higher efficiency in synthesis [19].

Since the 5′ UTR and coding sequences are essential components of the production of proteins in any living organism, it is reasonable to assume that selection act on these sequences of genes to amplify transcription and translation effectiveness. Urrutia & Hurst [20] postulate that due to the small length size of the sequences, in their case, protein size in relation to gene expression that selection is acting on these genes to maximise transcription and translation efficiency. However, a model proposed by Lynch [17], for the evolution of 5′ UTR length suggests that the evolution of the length of this region is influenced by stochastic processes, rendering it selectively neutral [17]. Reuter [21] disputed this model suggesting that UTR length evolution is affected by the gene's function and secondary mRNA structures. The length of the 5′ UTR showed some influence in gene expression, to extend on this research further, gene function may indicate the evolutionary weight to changes in these lengths.

Our results on the 3′ UTR gene regions lengths were reverse to that of the 5′ UTR and coding sequence. They showed a positive correlation between the 3′ UTR length and gene expression intensity levels. 3′ UTRs have been related to the stability of mRNA processing, but it can be difficult to interpret due to the involvement of the mRNA in all processes. The importance of this un-translated region is evident in many studies examining the presence of 3′ UTR in tumor growth [22], 3′-processing end sequences on gene expression in plant cells [23], regulation of mouse \(\kappa\) Opioid receptor gene expression by different 3′ Un-translated regions [24].

Extension of the 3′ UTR has also been allied with a pathway known as nonsense-mediated mRNA decay (NMD), where it was seen in Saccharomyces cerevisiae that 91% of the longer 3′ UTR mRNAs tested were affected by NMD [25]. Mutually, the 5′ and 3′ UTR’s involvement in gene expression is broadened to include further quality control mechanisms to strengthen the dependability of accurate protein formation [26], and length is a contributing factor to these control mechanisms. The lengths of the 3′ UTR’s varies substantially within eukaryote genomes. Humans present longer 3′ UTRs, compared to plants with a difference of 33% in length [27]. The evolution of longer 3′ UTR’s, as seen in humans, may be contributed to the regulation of gene expression which use this increase in length for post-transcriptional control mechanisms [28]. The results for the 3′ UTR for this particular plant species, showing higher levels of gene expression indicates that there are evolutionary forces at work and the increased length plays a role in the regulation of gene expression. Tanguy & Gallie [29] concluded from experiments on carrot protoplasts that there was an increase in stimulated expression by 24.5 fold when the 3′ UTR was increased to 27 bases. The un-translated region influence gene expression by way of RNA stability and translational efficiency [29] [28] (3′ UTR) and facilitating translation (5′ UTR). Our results support the role and importance of these regions in the regulation of gene expression.

5 Conclusion

Using the comprehensive Arabidopsis Information Resource (TAIR) database it has been possible to evaluate the relationship between the coding as well as the non-coding sequences and gene expression in the model plant species Arabidopsis thaliana. The research in this paper has confirmed previous findings on protein lengths, and discovered that the coding and 5′ UTR sequences are comparable, whereas, the 3′ UTR had dissimilar trends. The patterns found have contributed some generalised understanding about the relationship between gene expression and the length distributions of the non-coding regions. To extend on what has been ascertained, a further breakdown in tissue type, environmental conditions, chromosome level and gene function [7] would be beneficial in broadening understanding. An extension on the quantile regression model that uses the interaction of all three regions \((d_1, d_2\) and \(d_3\)) could show which length region has the most influence on the average gene expression intensity.

6 References


April 2003.


[27] Pesole G., Liuni S., Grillo G., & Saccone C. “Structural

