RSCU_COMPARER: A NEW STATISTICAL TOOL FOR PRACTICAL ANALYSIS OF CODON USAGE

Vladimirov N.V.*, Kochetov A.V., Grigorovich D.A., Matushkin Yu.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
*Corresponding author: e-mail: nikita@bionet.nsc.ru

Key words: RSCU, synonymous codons, t-test, gene expression, transgene

SUMMARY

Motivation: Many existing methods for statistical analysis of codon usage are efficient for deep investigation of codon usage patterns, but not for practical use in bioengineering. Practitioners are interested in simple and easy-to-use methods for determining the most unfavorable codons in a heterologous gene for its expression in a particular organism.

Results: A new statistical method for practical codon usage analysis is proposed. It allows a user to reveal statistically significant differences of codon usage patterns between two samples of highly and lowly expressed genes (provided by the user), and to determine the set of optimal and suboptimal codons that should be preserved or avoided in heterologous gene engineering. The novelty of the method is based on observation that distribution of log_{10}(RSCU) values is close to normal (RSCU, relative synonymous codon usage). This transformation allows parametric statistics, such as Student’s t-test, to be applied to comparison of codon usage patterns.

The program RSCU_comparer may be applied in practical bioengineering studies for selection of optimal and non-optimal codons in particular species using training sets of highly and lowly expressed genes.


INTRODUCTION

It is well known that the usage of synonymous codons in eukaryotic genes varies greatly in a species-specific manner. In many prokaryotic and in some eukaryotic organisms (e.g. *Saccharomyces cerevisiae, Drosophila melanogaster*) highly expressed genes are characterized by a preferential usage of more frequent synonymous codons. Currently, the relationship between translation elongation efficiency and preferences in the usage of synonymous codons in genes of most eukaryotic organisms is known poorly. However, an adjustment of the synonymous codon content is frequently used in gene engineering experiments to enhance a transgene expression rate. An increased content of species-specific synonymous codons preferentially used by highly expressed genes of a host organism may improve transgene expression characteristics.

METHODS AND ALGORITHMS

Training sets of highly (H) and lowly (L) expressed genes are provided by the user. For both samples the RSCU indices of codon usage bias (relative synonymous codon usage) (Sharp, Li, 1987) are calculated for 59 codons that have synonymous alternatives:
Here $X_i$ is the number of $i$-th codon occurrences in the (H) or (L) gene set, and $n_i$ – the number of alternative synonymous codons for $i$-th codon (from 2 to 6).

In such a way, codon usage of (H) and (L) samples is represented by two 59-dimensional vectors: $RSCU^{(H)} = (x_1, \ldots, x_{59})$ and $RSCU^{(L)} = (y_1, \ldots, y_{59})$, respectively.

The idea of the method is based on observation that distributions of $\log_{10}(RSCU)$ values are usually close to the normal distribution (see Examples). The log-normality of $RSCU^{(H)}$ and $RSCU^{(L)}$ allows to apply parametric statistics for comparison of codon usage patterns between (H) and (L) sets. For this purpose we used paired Student's t-test implemented in R statistical package (www.r-project.org). Since paired t-test requires normality of paired differences, at first we perform the Kolmogorov-Smirnov (K-S) test for normality of paired differences $D = \log_{10}(RSCU^{(H)})-\log_{10}(RSCU^{(L)}) = (d_1, \ldots, d_{59})$. For this purpose we use K-S test also implemented in R.

The vector $(d_1, \ldots, d_{59})$ reflects the difference in codon usage frequencies between (H) and (L) samples. The smallest $d_i$ values correspond to the suboptimal codons. Selection of suboptimal codons is performed by taking a lower $Q_{sub}$-quantile of $d_i$ distribution. Codons whose $d_i$ fall into the quantile limits are considered suboptimal. Avoidance of these codons in a heterologous gene may increase its translation elongation rate in the selected organism. In the same manner, the optimal codons (with high $d_i$ values) are selected using the upper $Q_{opt}$-quantile (10% by default).

EXAMPLES

The closeness of $\log_{10}(RSCU)$ to normal distribution may be seen on the following example (Fig. 1). We selected three samples of genes from *E. coli* representing highly (High), lowly (Low) expressed and randomly selected (Rand) genes. Each set contains 20 genes. The High set consists of ribosomal genes, the Low set includes genes with the lowest elongation efficiency index (EEI) (Likhoshvai, Matushkin, 2002) that estimates codon bias, and Rand includes genes randomly selected using the RSA-tools (http://rsat.ulb.ac.be/rsat/).

The closeness of $\log_{10}(RSCU)$ distributions to normal is confirmed by the Kolmogorov-Smirnov test for normality:

- High: $D_{KS} = 0.148$, $p = 0.164$; Low: $D_{KS} = 0.125$, $p = 0.309$; Rand: $D_{KS} = 0.0979$, $p = 0.623$.

The K-S test for normality of paired differences $(d_1, \ldots, d_{59})$ between High, Low and Rand samples:

- (H-L) $D_{KS} = 0.150$, $p = 0.200$; (H-R) $D_{KS} = 0.164$, $p = 0.128$; (L-R) $D_{KS} = 0.0907$, $p = 0.716$.

After validation of normality of $(d_1, \ldots, d_{59})$, the user can apply t-test for comparison of codon usage patterns:

1) High vs Low: $T = -2.68$, $p = 0.009$, samples are significantly different (threshold is 0.05).
2) High vs Random: $T = -3.58$, $p = 0.0007$, samples are significantly different.
3) Random vs Low: $T = 1.31$, $p = 0.19$, difference is non-significant.

The paired $t$ test allows to discriminate between gene sets characterized by a considerable difference in the usage of synonymous codons (High versus Rand, High versus Low for the *E. coli* example). To confirm the applicability of our method to different organisms with known correlation between codon usage bias and gene expression, we analyzed samples “High” and “Rand” from *E. coli*, *B. subtilis*, and *H. influenzae* for positive examples, and from *H. pylori* for a negative one. In each organism
“High” gene set comprised 40 ribosomal genes, and “Rand” gene set consisted of 40 randomly selected genes. The results are shown in Table 1. The K-S test shows normality of all paired differences, and the t-test shows highly significant differences in codon usage of the bacteria except *H. pylori*, which is known to lack translational codon bias.

![Figure 1. Distributions of RSCU and log10(RSCU) values (x-axis) for 3 sets from *E. coli*. The y-axis corresponds to the number of occurrences.](image)

Table 1. Kolmogorov-Smirnov test for \((d_1, \ldots, d_{59})\) normality, and the t-test for differences between “High” and “Rand” samples in bacteria

<table>
<thead>
<tr>
<th>Organism</th>
<th>K-S test, p-values</th>
<th>t-test, p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>0.59</td>
<td>0.0006</td>
</tr>
<tr>
<td><em>B. subtilis</em></td>
<td>0.64</td>
<td>0.01</td>
</tr>
<tr>
<td><em>H. influenzae</em></td>
<td>0.56</td>
<td>0.009</td>
</tr>
<tr>
<td><em>H. pylori</em></td>
<td>0.31</td>
<td>0.9</td>
</tr>
</tbody>
</table>

**IMPLEMENTATION**

RSCU values are calculated with C-program, and transferred into R script that filters the data and computes statistics with t-test() and ks.test() functions. The program takes 2 sets of coding sequences (CDS) assumed to be highly (H) and lowly (L) expressed in the organism of interest. CDS should be submitted in FASTA format.

The program computes:
1. K-S test for normality of (H), (L) and (H-L) log10(RSCU) distributions,
2. paired t-test for mean of RSCU differences between (H) and (L) samples,
3. codons completely absent in (H) set (if any),
4. the most unfavorable codons (defined by \(Q_{sub}\)),
5. the most favorable codons (defined by \(Q_{opt}\)).

**User-defined parameters**

1. \(p_{max}\), a threshold of significance for \(p\)-values (default 0.05);
2. \(Q_{sub}\), a quantile threshold for suboptimal codons selection (default 10 %);
3. \(Q_{opt}\), a quantile threshold for optimal codons selection (default 10 %).
Limitations

In general, avoidance of the most unfavorable codons may be useful even if the difference between the (H) and (L) samples is not statistically significant, because the level of significance depends on the sample sizes and the selection procedure. The selection of appropriate "High" and "Low" expression gene sets may be an uneasy task because of a difference in tissue-specific expression rate (for multicellular organisms), possible difference between expression of members of multigene families, etc.

It should also be noted that in some organisms codon content may not correlate with translation level significantly (possibly because the rate-limiting stage is translation initiation, or for some other reasons).

DISCUSSION

The novelty of the method is based on observation that distribution of $\log_{10}(\text{RSCU})$ values is close to normal. This transformation allows application of parametric statistics, such as Student’s t-test, to comparison of codon usage patterns. A similar approach is widely used in statistical analysis of microarray data (Stekel, 2003), where experimental data on gene expression are normalized (including log-transformation) and compared with reference samples. The (H) set may include genes for ribosomal proteins, heat-shock proteins, translation elongation factors, actins, tubulins or other abundant proteins. The (L) set may include genes encoding transcriptional factors, growth factors, receptors, protein kinases (Kochetov et al., 1998) or randomly selected genes. Also, the (H) and (L) gene sets may be composed on the base of a microarray, SAGE or proteomic data. We recommend to use sets containing at least 20 genes for representative statistics.

RSCU_comparer may be applied in practical bioengineering studies for selection of optimal and non-optimal codons in particular species using training sets of highly and lowly expressed genes. It provides an opportunity to select unfavorable synonymous codons to be replaced for enhancing expression of a transgene CDS.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (grants Nos 05-04-48207, 06-04-49556), the Program of Russian Academy of Sciences (Dynamics of Plant, Animal, and Human Gene Pools), and the Program of RAS Presidium “Origin and evolution of biosphere” (Evolution of molecular-genetic systems: computer analysis and modeling) No. 10104-34/P-18/155-270/1105-06-001/28/2006-1. We thank SB RAS Complex Integration Program (No. 5.3) and Ministry of Industry, Science and Technologies of the Russian Federation (grant No. 2275.2003.4) for partial support.

REFERENCES