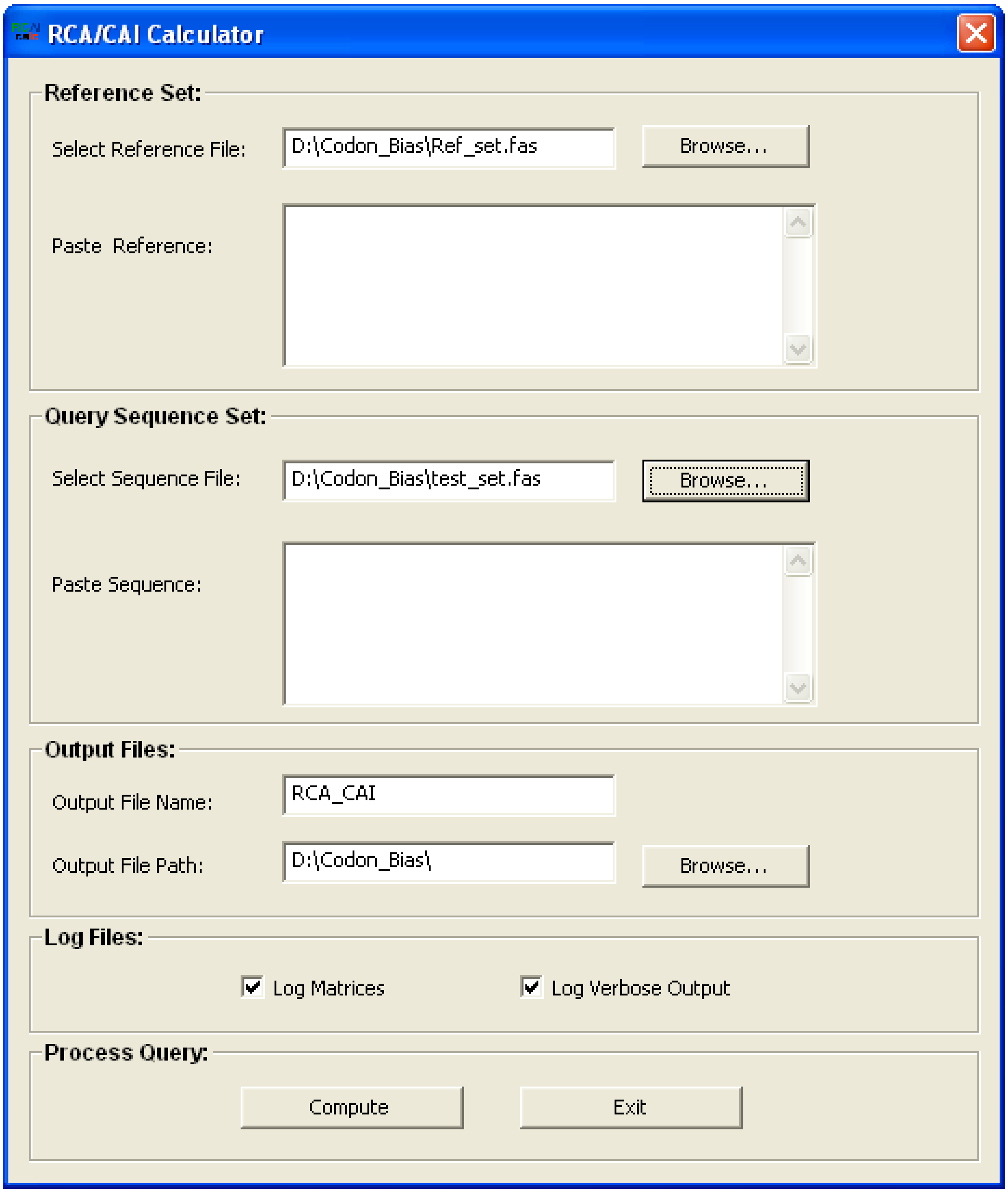
RCAI-CALC User Manual

Introduction:

RCAI-CALC is a Microsoft Windows based application. The application outputs the Codon Adaptation Index (CAI) (Sharp and Li 1987) and Relative Codon Adaptation (RCA) (Fox and Erill 2010) values for a given set of Query Sequences. The indices are calculated using a reference set of highly expressed genes. The application allows the user to select a file or paste the reference and query sequence sets and calculate the CAI and RCA values.



Select Reference File: D:\Codon_Bias\Ref_set.fas

Select Sequence File: D:\Codon_Bias\test_set.fas

Output File Name: RCA_CAI

Output File Path: D:\Codon_Bias\

☑ Log Matrices ☑ Log Verbose Output

Figure 1: Graphical User Interface for RCAI calc

Getting started:

RCAI-CALC provides a Graphical User Interface (GUI) to load the required sequence files (reference and query sequence sets) or to paste these sequences and compute the values of CAI and RCA all the query set sequences.

Main Operation files:

The application starts by loading the *Reference Sequence* and *Query Sequence* sets. The *Reference Sequence set* is composed of genes presenting a very strong bias (major codon bias) due, presumably, to high expression levels. This set is used to calculate codon frequencies and base frequencies for each codon position. The *Query Sequence set* is the set of sequences for which the RCA and CAI indices will be computed.

Reference Sequence set:

This is a set of sequences with a strong bias (major codon bias) that is presumably due to high expression levels. The user can select a reference sequence file or paste the reference sequences in the appropriate text box. Reference sequences can be in two possible formats: FASTA and raw TEXT. In FASTA format each sequence is preceded by an identification line beginning with '>'. Accepted extensions for FASTA files are: FAS/FNA. In raw TEXT sequences must be separated with two consecutive newline characters. The accepted extension for raw text files is TXT. The user can select the reference sequence file using the browse button or paste the reference sequence in the Text box.

Query Sequence set:

This is a given set of sequences on which to calculate the CAI and RCA values. Query sequences can also be in two possible formats: FASTA or raw TEXT. Again, the user can select the query sequence file using the browse button or paste the query sequences in the appropriate text box.

Functionality:

Log options:

The log option determines whether the program should log results during operation. The default Log option is *yes* for both Log Matrices and Log Verbose Output. If the log option is set, the program will generate the following files:

- (Output_FileName)_VerboseOutput.csv

  This file contains, for each sequence, the list of codons used to compute RCA and CAI, together with the natural logarithm (ln) of their w_i and RCA_{xyz} values.

- (Output_FileName)_CodonFreq.csv

  This file contains the codon frequencies derived from the reference set and the natural logarithm (ln) of their w_i and RCA_{xyz} values.
- (Output_FileName)_BaseFreq.csv

  This file contains the base frequencies for each codon position in the reference set.

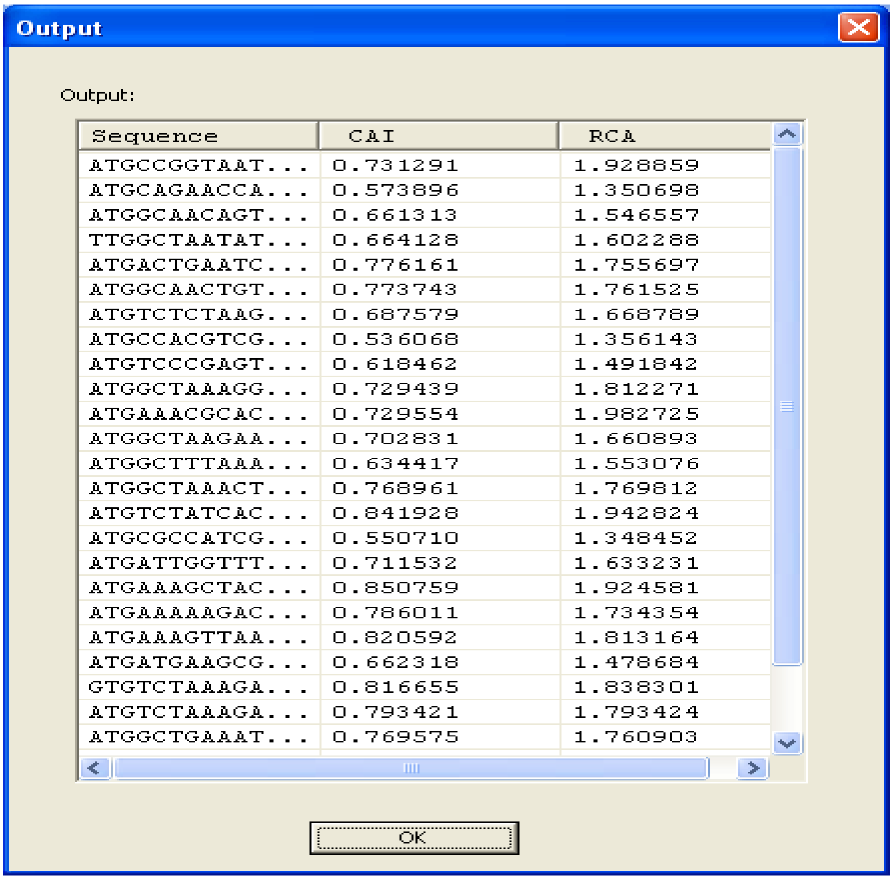
"Compute" Button:

When the user clicks on the "Compute" button, **RCAI-CALC** will scan the reference set and calculate the codon frequencies and base frequencies for each codon position. Sequences may contain IUB degenerate codes for bases, and their frequency contributions will be weighted accordingly (see *Main Operation* below) (Cornish-Bowden 1985). After computing the reference set frequencies, **RCAI-CALC** will compute the natural logarithm of the w_i and RCA_{xyz} values for each codon in the reference set. It then computes the RCA and CAI values on each query sequence with respect to the reference set as an arithmetic mean. The results will be displayed by **RCAI-CALC** in an emergent results window. They will also be saved to a .CSV (comma-separated value) file in the specified output folder. The CSV file can be imported directly into Ms-Excel or any other spreadsheet program for additional processing and analysis.

Display results:

After the computation of the indices is complete, **RCAI-CALC** will display the result in the Output window (see figure below). The results window contains all the sequences in the query set and their RCA and CAI values.

RCAI-CALC will save the results into a comma-separated value (CSV) file *(Output_FileName)_Output.csv* that can be opened directly with spreadsheet software such as Ms-Excel. The output file is saved at the output file path selected by the user.



Sequence	CAI	RCA
ATGCCGGTAAT...	0.731291	1.928859
ATGCAGAACCA...	0.573896	1.350698
ATGGCAACAGT...	0.661313	1.546557
TTGGCTAATAT...	0.664128	1.602288
ATGACTGAATC...	0.776161	1.755697
ATGGCAACTGT...	0.773743	1.761525
ATGTCTCTAAG...	0.687579	1.668789
ATGCCACGTCG...	0.536068	1.356143
ATGTCCCGAGT...	0.618462	1.491842
ATGGCTAAAGG...	0.729439	1.812271
ATGAAACGCAC...	0.729554	1.982725
ATGGCTAAGAA...	0.702831	1.660893
ATGGCTTTAAA...	0.634417	1.553076
ATGGCTAAACT...	0.768961	1.769812
ATGTCTATCAC...	0.841928	1.942824
ATGCGCCATCG...	0.550710	1.348452
ATGATTGGTTT...	0.711532	1.633231
ATGAAAGCTAC...	0.850759	1.924581
ATGAAAAAGAC...	0.786011	1.734354
ATGAAAGTTAA...	0.820592	1.813164
ATGATGAAGCG...	0.662318	1.478684
GTGTCTAAAGA...	0.816655	1.838301
ATGTCTAAAGA...	0.793421	1.793424
ATGGCTGAAAT...	0.769575	1.760903

Figure 2: Sample Output Window

Main operation:

CAI and RCA are effective measures of synonymous codon usage bias (Fox and Erill 2010). Codon usage bias (CUB) is the departure from expected uniformity in the frequency of occurrence of synonymous codons in genomic sequences (Ermolaeva 2001). This bias is due to the redundancy of the genetic code, which allows differential use of synonymous codons, and the difference in availability of the different species of tRNA in a cell, among other factors (Kurland 1991). The CUB has been shown to correlate well with gene expression in prokaryotes. Several independent measures for estimating this bias have been devised over the years. Some of these are estimate bias as the departure from expected codon usage uniformity. Other methods estimate the bias as the divergence in codon usage from a reference group of genes. These reference genes are known to present a strong bias (major codon bias), that is typically attributed to high expression values (Kurland 1991). The Codon Adaptation Index (CAI) and Relative Codon Adaptation (RCA) are two indices based on the use of a reference set.

Codon Adaptation Index (CAI):

CAI uses the reference set in order to compute the codon bias for the given set of query sequences (Sharp and Li 1987). The index is calculated using the frequencies of each codon and the largest frequency among synonymous codons in the reference set. Thus, for any codon *i* and amino acid *j*, we define CAI as:

$$\mathrm{CAI} = \left(\prod_{l=1}^{L} w_i(l) \right)^{1/L} \qquad w_i = \frac{f_{ij}}{\max(f_{xj})},$$

where f_{ij} is the frequency of codon *i* encoding amino acid *j* as observed in the reference set, $\max(f_{xj})$ the largest frequency among the codons encoding amino acid *j* and *L* the length, in codons, of the given gene sequence. The term w_i is called the relative adaptiveness of codon *i* and the CAI is computed as the geometric mean of the w_i values for all codons in the query sequence.

Relative Codon Adaptation (RCA):

RCA is also used to calculate codon usage bias using the reference set (Fox and Erill 2010). The RCA is calculated using the codon frequencies and codon base frequencies for each codon position in the reference set. Again for any given codon *xyz*, we define RCA as:

$$\mathrm{RCA} = \left(\prod_{i=1}^{L} \mathrm{RCA}_{xyz}(l) \right)^{1/L}$$

$$\mathrm{RCA}_{xyz} = \frac{\chi(x,y,z)}{\chi_1(x)\chi_2(y)\chi_3(z)},$$

where X(x,y,z) is the observed frequency of codon xyz in any particular reference gene set, $X_n(m)$ the observed frequency of base m at codon position n in the same reference set, and L the length in codons of the query sequence. Like CAI, RCA is computed as the geometric mean of the RCA_{xyz} term for each codon xyz in the given sequence.

To overcome problems with real number overflow in computing CAI and RCA, the geometric mean of both indices is transformed into an arithmetic mean by first computing the natural logarithm (ln) of the w_i and RCA_{xyz} values for all the codons in the reference set. The query sequences are then scanned and the arithmetic mean of these logarithmic values is computed. Finally, the total RCA or CAI average is restored to linear scale by computing its exponential.

Considerations with regards to absent/stop codons and single codon amino acids

Note that if a certain codon is never used in the reference set then frequency of that codon will be zero (Sharp and Li 1987; Xia 2007). This leads to the counter-intuitive result of a RCA/CAI value of zero, regardless of the deviation. Furthermore, while computing the natural logarithm of w_i and RCA$_{xyz}$ this will result in an infinite value. To overcome this problem **RCAI-CALC** adds a pseudo-count of 1/N (where N is total number of codons in the reference set) to all codon and base frequencies. Stop codons are ignored during computation of CAI/RCA since it cannot be assumed that they are subject to the selective pressures that lead to the observed patterns of codon usage for regular codons. In addition, the frequencies of ATG (Methionine) and TGG (Tryptophan) are not taken into account either when computing CAI/RCA. This is because these are single-codon amino acids and, therefore, their w and RCA_{xyz} terms are always fixed to 1.0 and do not provide information on codon bias (Sharp and Li 1987). While calculating frequencies for each codon, degenerate base codes (IUB-IUPAC) are allowed. Frequencies for these degenerate base codes are computed based on their degeneracy and the possible codons they might encode (Cornish-Bowden 1985). For example:

*AWG*CCGGTAATTAAAGTACGTGAAAACGAGCCGTTCGACGTAGCTCT…

In the above sequence, codon AWG contains the IUB code "W" which has two possible bases "A" and "T". Therefore, 0.5 is added to the frequency of A and T in the second codon position and 0.5 is added to the frequency count for "AAG" and 0.5 for "ATG", the two possible codons encoded by the degenerate codon AWG.

Results:

After computing the values of CAI and RCA, **RCAI-CALC** will save the results into a comma-separated value (CSV) file *(Output_FileName)_Output.csv* that can be opened directly with spreadsheet software such as Ms Excel. The output file contains all the sequences from the query sequences and their respective CAI and RCA values.

References:

- Cornish-Bowden, A. (1985). "Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984." Nucleic Acids Res **13**(9): 3021-30.
- Ermolaeva, M. D. (2001). "Synonymous codon usage in bacteria." Curr Issues Mol Biol **3**(4): 91-7.
- Fox, J. M. and I. Erill (2010). "Relative Codon Adaptation: A Generic Codon Bias Index for Prediction of Gene Expression." DNA Res.
- Kurland, C. G. (1991). "Codon bias and gene expression." FEBS Lett **285**(2): 165-9.
- Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-95.
- Xia, X. (2007). "An improved implementation of codon adaptation index." Evol Bioinform Online **3**: 53-8.