

## README file for CodonLogo

### Availability:

CodonLogo is available for download at <http://recode.ucc.ie/CodonLogo> and can be used for a local installation. CodonLogo is also available from the Galaxy toolshed at <http://toolshed.g2.bx.psu.edu/>.

This document shows how to make use of different features of CodonLogo (for the web based version). Arguments that need to be passed to the command line based version of the program can be found in a table near the end of this manual.

**Input file:** By default CodonLogo works with alignments in nbrf, fasta, clustal, phylip, genbank, stockholm, msf, nexus and table format. In case the input file is not in any of the specified formats, CodonLogo attempts to parse any text file passed to it but may not succeed. Users can also convert their alignments into one of the formats acceptable to CodonLogo through alignment format conversion tools. For example:

<http://genome.nci.nih.gov/tools/reformat.html>

<http://www-bimas.cit.nih.gov/molbio/readseq/>

**Alignment format:** Since CodonLogo has been designed to highlight codon conservation patterns, the input alignments should ideally be codon based. This also implies that gaps in the alignments should occur in groups of three. Users can refer to Code2aln[1] or TranslatorX[2] to align their sequences and generate codon based alignments. If you have alignments where gaps do not occur in groups of three, refer to “**Allow gaps which are not in groups of 3/ allow codons with ambiguous nucleotides**” section described later.

**Title for output Sequence Logo:** This is the name that is given to the output file (codon logo image).

**Output format for image:** CodonLogo is currently capable of producing outputs in a variety of formats. These include a) PNG (screen quality) b) PNG (high quality for print) c) PDF d) JPG e) EPS f) Text format: Selection of this option does not produce an image; instead it produces a text file describing frequency of codons and information content of the alignment.

**Bitmap resolution in dots per inch (DPI):** This is set to 96 by default. This value can be increased to generate high resolution images.

**Frame to view the alignment:** Specifies one of three reading frames as coding. This is used as a phase to define sequence of codons. If reverse complement is chosen, sequence of codons can be defined in all six frames.

**Number of stacks per line:** Specifies the number of stacks (codon columns) per line in an output Codonlogo. The default value is set to 20.

**Display codon numbers along x-axis:** ON (select “Yes”) or OFF (select “No”). This allows the program to write the codon numbers/indexes on the x-axis in the output codon logo.

**X-axis label:** Specifies label for X-axis

**Y-axis label:** Specifies label for Y-axis.

**The fine print:** This is a string that appears at the bottom of the output codon logo that is generated. Default is set to “CodonLogo 1.1”.

**Width of a logo stack:** Specifies the width of the stack. Especially useful when the input alignment is long. Default is 20.

**Height of a logo stack:** Specifies the height of the logo stack. Default is set to 100.

**Display entropy scale along y-axis?:** Default is “Yes”.

**Display error bars?:** Specifies whether to display 95% confidence interval . Default is “Yes”.

**Allow gaps which are not in groups of 3/ allow codons with ambiguous nucleotides:** Yes(default): Do not issue any warning if the program encounters a non-standard codon or incomplete codon.  
No:Issue a warning but do not halt the program.  
“Warning:Incomplete or non GATUC codon detected:” With details of the sequence and position of the codon.

**Include entire sequence (default) or specify a subsequence range to use:** Allows to trim alignments and restrict Codon Logo generation to a specific region of the alignment. The subsequence range can be specified in ‘Index\_start’ and ‘Index\_end’ fields.

**Select expected composition model:**

We provide several background models for CodonLogo and allow users to specify their own model. User specified model should be a tab delimited text file with 64 rows (for each codon) and two columns, first corresponding to the codon and second corresponding to its frequency per thousand codons. Refer to the Codon Usage Database for more details on the format and codon usage tables for different organisms (<http://www.kazusa.or.jp/codon/>).

An alignment with few sequences would not show meaningful codon conservation patterns because of the small sample correction that the program would enforce by default. In such a scenario, it is necessary to turn off this feature. In addition to this, due to biases in codon usage of real protein coding sequences, codon frequencies are not equal. To account for this, we also provide background models for coding sequences of three representative organisms i.e. human (*H. sapiens*), yeast (*S. cerevisiae*) and bacteria (*E. coli*). In total, there are 6 options:

- 1) No small sample correction
- 2) Equiprobable
- 3) Human
- 4) Yeast

- 5) *E. coli*
- 6) User specified

Other than Option 1 (No small sample correction), the rest will force the program to apply small sample correction.

**Choose colors:** Specifies colors for displaying codons in CSS2 syntax. The default scheme is based on physicochemical properties of amino acids encoded by codons according to the standard genetic code table.

1. Polar positive H, K, R -Light Blue
2. Polar negative D, E -Red
3. Polar neutral S, T, N, Q -Green
4. Non-polar aliphatic A, V, L, I, M -Blue
5. Non-polar aromatic F, W, Y -Magenta P, G -Brown C -Yellow
6. Stop codons -Black

**Output CodonLogo size:** The size of the output codon logo. Default is Large.

The following table shows the various arguments for the options discussed above for the command line version of CodonLogo

Full help can be accessed from the command line with the command

codonlogo --help

S. No	Feature	Arguments for the command line based version
1	Input file	-f [filename]
2	Title for output sequence logo	-R [filename]
3	Output format for image	-F [eps (default), png, png_print, pdf, jpeg, txt]
4	Bitmap resolution (DPI)	--resolution [integer]
5	Frame to view the alignment	-m [0-5]
6	Number of stacks per line	-n [integer]
7	Display codon numbers along x-axis	-X [True/False]
8	X-axis label	-x [Text]
9	Y-axis label	-y [Text]
10	The fine print	--fineprint [Text]
11	Width of the logo stack	-W [float]
12	Height of the logo stack	-H [float]
13	Display entropy scale along y-axis	-Y [True/False]
14	Display error bars	--errorbars [True/False]
15	Allow gaps which are not in groups of 3/ allow codons with ambiguous nucleotides	-G [True/False]
16	Set lower bound of sequence to display.	-l [integer]

17	Set upper bound of sequence to display.	-u [integer]
18	Select expected composition:	--comp ['equiprobable' or 'none' ]. This will be overridden if a file is specified with -R
19	Specify a file of prior probabilities.	-R [filename]. For <i>E.coli</i> , <i>H. sapiens</i> or <i>S. cerevisiae</i> , use --composition escherichiacoli, --composition homosapiens and --composition saccharomycescerevisiae respectively. This will override --comp if used at the same time.
20	Choose colors	-C [codon] [color] [description] example: -C AAA #FF0000 black
21	Output CodonLogo size	--size [small, medium or large]

### Contents of the **example** directory

- 1) The input sequence alignment that was used to generate the codon logo for the Figure 1B in the manuscript
- 2) The codon logo generated as a result of running CodonLogo on this input alignment.
- 3) A workflow which may be imported into galaxy to create the figures in part B of the figure.

The workflow is also accessible from

<http://beamish.ucc.ie:8080/u/david/w/figure>

### References:

- 1) Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 2010 Jul;38(Web Server issue):W7-13. Epub 2010 Apr 30. PubMed PMID: 20435676.
- 2) Stocsits RR, Hofacker IL, Fried C, Stadler PF. Multiple sequence alignments of partially coding nucleic acid sequences. *BMC Bioinformatics.* 2005 Jun 28;6:160. PubMed PMID: 15985156.