(http://genomes.urv.es/CAIcal)

# TUTORIAL

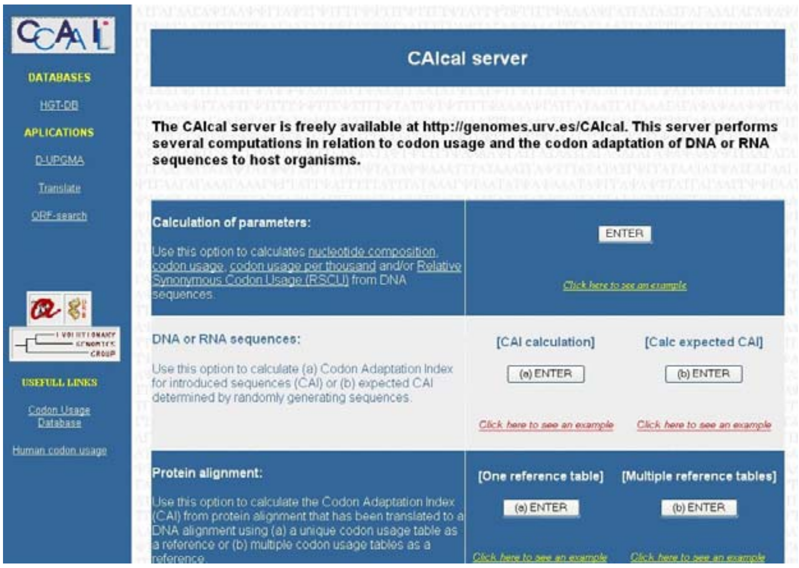
(July 2006)

## Table of contents

## Introduction

### WHAT IS CAICAL?

CAIcal is a web server, freely available at http://genomes.urv.es/CAIcal, that performs several computations in relation to codon usage and the codon adaptation of DNA or RNA sequences to host organisms.



CAIcal server

The CAIcal server is freely available at http://genomes.urv.es/CAIcal. This server performs several computations in relation to codon usage and the codon adaptation of DNA or RNA sequences to host organisms.

Calculation of parameters:

Use this option to calculates nucleotide composition, codon usage, codon usage per thousand and/or Relative Synonymous Codon Usage (RSCU) from DNA sequences.

DNA or RNA sequences:

Use this option to calculate (a) Codon Adaptation Index for introduced sequences (CAI) or (b) expected CAI determined by randomly generating sequences.

Protein alignment:

Use this option to calculate the Codon Adaptation Index (CAI) from protein alignment that has been translated to a DNA alignment using (a) a unique codon usage table as a reference or (b) multiple codon usage tables as a reference.

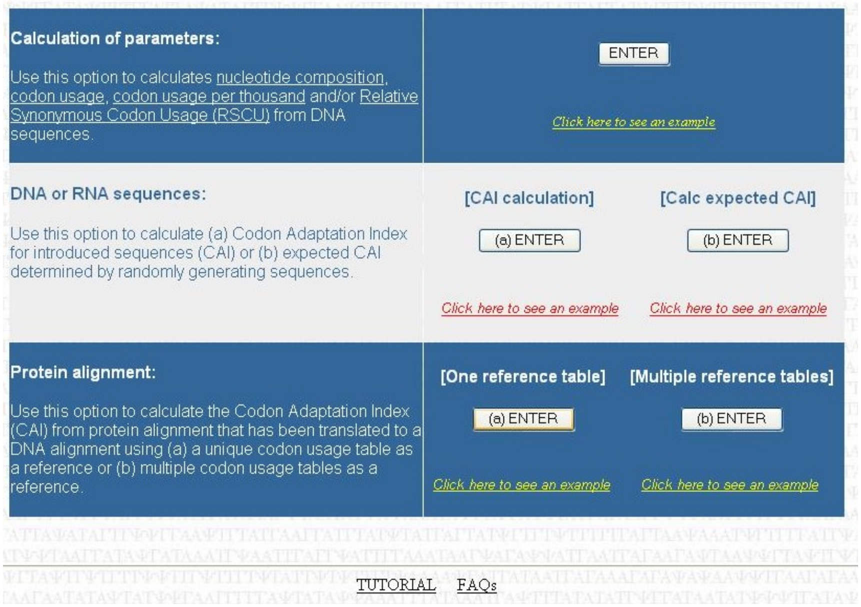
### OPTIONS AVAILABLE

CAIcal has three main options:

a) Calculation of parameters. This initial option provides basic calculations such as nucleotide composition, codon usage, codon usage per thousand and relative synonymous codon usage (RSCU) (see section A of this guide).

b) CAI calculation for FASTA sequences. This section has two options: 1) CAI calculation for DNA or RNA sequences introduced and 2) calculation of an expected CAI value determined by randomly generating sequences (see section B of this guide).

c) CAI calculation for protein alignment translated to DNA alignment. This option provides the use of: 1) one reference codon usage table for all of the sequences or 2) one reference table for each sequence introduced (see section C of this guide).



**Calculation of parameters:**

Use this option to calculates nucleotide composition, codon usage, codon usage per thousand and/or Relative Synonymous Codon Usage (RSCU) from DNA sequences.

**DNA or RNA sequences:**

Use this option to calculate (a) Codon Adaptation Index for introduced sequences (CAI) or (b) expected CAI determined by randomly generating sequences.

**Protein alignment:**

Use this option to calculate the Codon Adaptation Index (CAI) from protein alignment that has been translated to a DNA alignment using (a) a unique codon usage table as a reference or (b) multiple codon usage tables as a reference.

ENTER
Click here to see an example
[CAI calculation]
[Calc expected CAI]
(a) ENTER
(b) ENTER
Click here to see an example
Click here to see an example
[One reference table]
[Multiple reference tables]
(a) ENTER
(b) ENTER
Click here to see an example
Click here to see an example
TUTORIAL
FAQs

We have also developed this Tutorial, a Frequently Asked Questions section and several examples, which are available from the home page of the server. These helpful options will be periodically updated.

## Required inputs

### INPUT REQUIREMENTS

The server first checks whether the query sequences are a DNA or a RNA region. The table below is a summary of the input requirements for each section of CAIcal.

| Input | Gene parameters | CAI calculation | | CAI calculation of alignment | |
|---|---|---|---|---|---|
| | | CAI | Expected CAI | One reference table | Multiple reference tables |
| **DNA or RNA sequences in FASTA format** | 1 | 1 | 1 | 1 | 1 |
| **One codon usage reference table for all sequences** | 0 | 1 | 1 | 1 | 0 |
| **One codon usage reference table for each sequence** | 0 | 0 | 0 | 0 | 1 |
| **Protein alignment** | 0 | 0 | 0 | 1 | 1 |
| **Genetic code selection** | 1 | 1 | 1 | 1 | 1 |
| **Upper confidence limit option** | 0 | 0 | 1 | 0 | 0 |
| **Nucleotide composition and codon usage calculation options** | 1 | 0 | 0 | 0 | 0 |
| | | | | | |

*(**1**: required; **0**: not required)*

### FORMAT OF THE REFERENCE SET

An easy way to introduce the codon usage reference table in CAIcal is to copy and paste the codon usage tables from *Codon Usage Database* (Nakamura et al., 2000). We have therefore added a link to this database in the left frame of the server. The codon usage table from the '*Codon Usage Database*' format allowed in CAIcal is as follows:

**Fields: [triplet] [frequency: per thousand] ([number])...**

Example:

```
UUU 17.4(586747)  UCU 15.0(507382)  UAU 12.1(408578)  UGU 10.5(352664)
UUC 20.4(687969)  UCC 17.7(596425)  UAC 15.3(516505)  UGC 12.6(426761)
UUA  7.5(254407)  UCA 12.1(409879)  UAA  1.1( 35822)  UGA  1.6( 55514)
UUG 12.8(432797)  UCG  4.5(150335)  UAG  0.8( 27554)  UGG 13.3(447152)
```

```
CUU 13.1(440882)  CCU 17.5(589809)  CAU 10.8(363555)  CGU  4.6(155426)
CUC 19.7(664417)  CCC 20.0(675558)  CAC 15.1(509431)  CGC 10.6(357380)
CUA  7.1(240672)  CCA 16.9(569871)  CAA 12.1(408697)  CGA  6.2(208816)
CUG 39.9(1347830) CCG  7.0(237033)  CAG 34.3(1157220) CGG 11.6(390529)

AUU 15.8(532975)  ACU 13.0(438753)  AAU 16.7(563795)  AGU 12.1(408481)
AUC 20.9(705646)  ACC 19.0(641707)  AAC 19.0(642797)  AGC 19.5(656528)
AUA  7.4(249300)  ACA 15.0(504527)  AAA 24.1(812474)  AGA 11.9(402225)
AUG 22.0(744022)  ACG  6.1(205470)  AAG 32.0(1079579) AGG 11.9(402146)

GUU 11.0(370035)  GCU 18.5(624602)  GAU 21.7(732533)  GGU 10.8(364282)
GUC 14.6(491325)  GCC 28.1(947810)  GAC 25.2(850343)  GGC 22.5(758251)
GUA  7.1(238697)  GCA 15.9(537665)  GAA 28.6(964323)  GGA 16.4(553492)
GUG 28.3(956245)  GCG  7.5(253270)  GAG 39.7(1340672) GGG 16.6(558612)
```

We have also introduced another format as follows:

**Fields: [triplet] ([number])...**

Example:

```
TTT (171) TCT (147) TAT (124) TGT (99)
TTC (203) TCC (172) TAC (158) TGC (119)
TTA (73)  TCA (118) TAA (0)   TGA (0)
TTG (125) TCG (45)  TAG (0)   TGG (122)
CTT (127) CCT (175) CAT (104) CGT (47)
CTC (187) CCC (197) CAC (147) CGC (107)
CTA (69)  CCA (170) CAA (121) CGA (63)
CTG (392) CCG (69)  CAG (343) CGG (115)
ATT (165) ACT (131) AAT (174) AGT (121)
ATC (218) ACC (192) AAC (199) AGC (191)
ATA (71)  ACA (150) AAA (248) AGA (113)
ATG (221) ACG (63)  AAG (331) AGG (110)
GTT (111) GCT (185) GAT (230) GGT (112)
GTC (146) GCC (282) GAC (262) GGC (230)
GTA (72)  GCA (160) GAA (301) GGA (168)
GTG (288) GCG (74)  GAG (404) GGG (160)
```

The section that requires more than one codon usage database in the same text box need sequence identification: *[name of sequence]*.

Example:

```
[name_sequence_1]
TTT (171) TCT (147) TAT (124) TGT (99)
TTC (203) TCC (172) TAC (158) TGC (119)
TTA (73)  TCA (118) TAA (0)   TGA (0)
TTG (125) TCG (45)  TAG (0)   TGG (122)
CTT (127) CCT (175) CAT (104) CGT (47)
CTC (187) CCC (197) CAC (147) CGC (107)
CTA (69)  CCA (170) CAA (121) CGA (63)
CTG (392) CCG (69)  CAG (343) CGG (115)
ATT (165) ACT (131) AAT (174) AGT (121)
ATC (218) ACC (192) AAC (199) AGC (191)
ATA (71)  ACA (150) AAA (248) AGA (113)
ATG (221) ACG (63)  AAG (331) AGG (110)
GTT (111) GCT (185) GAT (230) GGT (112)
GTC (146) GCC (282) GAC (262) GGC (230)
GTA (72)  GCA (160) GAA (301) GGA (168)
GTG (288) GCG (74)  GAG (404) GGG (160)


[name_sequence_2]
TTT (171) TCT (147) TAT (124) TGT (99)
TTC (203) TCC (172) TAC (158) TGC (119)
TTA (73)  TCA (118) TAA (0)   TGA (0)
TTG (125) TCG (45)  TAG (0)   TGG (122)
CTT (127) CCT (175) CAT (104) CGT (47)
CTC (187) CCC (197) CAC (147) CGC (107)
CTA (69)  CCA (170) CAA (121) CGA (63)
CTG (392) CCG (69)  CAG (343) CGG (115)
ATT (165) ACT (131) AAT (174) AGT (121)
ATC (218) ACC (192) AAC (199) AGC (191)
ATA (71)  ACA (150) AAA (248) AGA (113)
ATG (221) ACG (63)  AAG (331) AGG (110)
GTT (111) GCT (185) GAT (230) GGT (112)
GTC (146) GCC (282) GAC (262) GGC (230)
```

```
GTA (72)  GCA (160) GAA (301) GGA (168)
GTG (288) GCG (74)  GAG (404) GGG (160)
```

## ERROR AND WARNING MESSAGES

The table below is a brief summary of the main errors and warning of CAIcal.

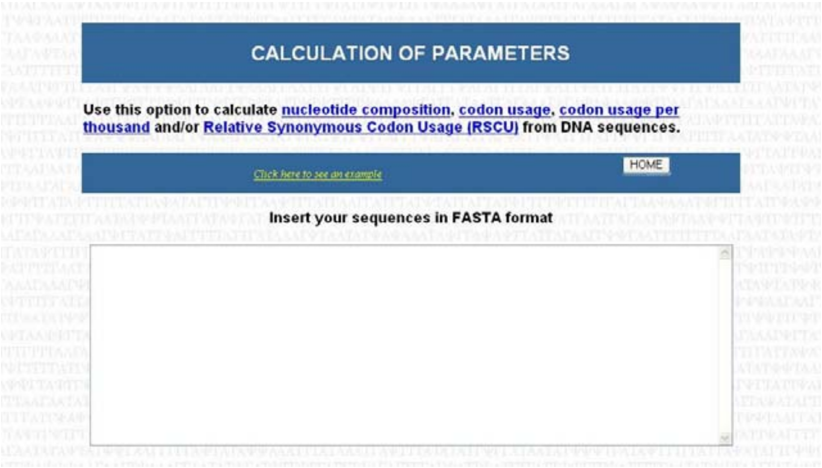
| Option | Genes Parameters | CAI calculation | | CAI calculation of alignment | |
|---|---|---|---|---|---|
| | | CAI | Expected CAI | One reference table | Multiple reference tables |
| **No sequences are introduced** | E | E | E | E | E |
| **No reference table is introduced** | E | E | E | E | E |
| **Sequence is not divisible by three** | E | E | E | E | E |
| **Sequence with more than one stop codon** | W | W | W | W | W |
| **No parameters are checked** | E | - | - | - | - |
| **DNA sequence introduced does not correspond to protein sequence from the alignment.** | - | - | - | E | E |
| **Sequence too long (>10000 nt)** | E | E | E | E | E |
| | | | | | |

*(**E**: error; **W**: warning)*

## SECTION A - Calculation of Parameters

### OVERVIEW

Use this option to calculate nucleotide composition, codon usage, codon usage per thousand and/or relative synonymous codon usage (RSCU) from DNA sequences.



CALCULATION OF PARAMETERS

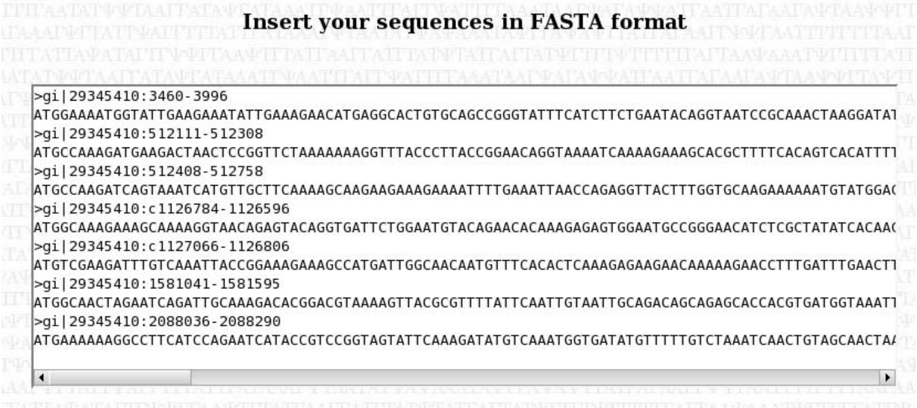
Use this option to calculate nucleotide composition, codon usage, codon usage per thousand and/or Relative Synonymous Codon Usage (RSCU) from DNA sequences.

Insert your sequences in FASTA format

### INPUTS

This section requires three steps:

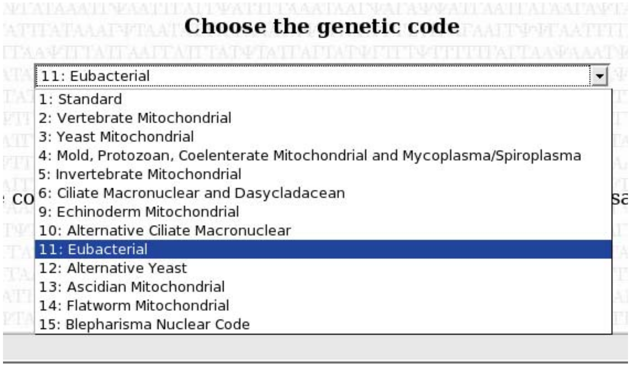
1) Introduction of DNA or RNA sequences in FASTA format in a text box.



Insert your sequences in FASTA format

```
>gi|29345410:3460-3996
ATGGAAAATGGTATTGAAGAAATATTGAAAGAACATGAGGCACTGTGCAGCCGGGTATTTCATCTTCTGAATACAGGTAATCCGCAAACTAAGGATAT
>gi|29345410:512111-512308
ATGCCAAAGATGAAGACTAACTCCGGTTCTAAAAAAAGGTTTACCCTTACCGGAACAGGTAAAATCAAAAGAAAGCACGCTTTTCACAGTCACATTTT
>gi|29345410:512408-512758
ATGCCAAGATCAGTAAATCATGTTGCTTCAAAAGCAAGAAGAAAGAAAATTTTGAAATTAACCAGAGGTTACTTTGGTGCAAGAAAAAATGTATGGAC
>gi|29345410:c1126784-1126596
ATGGCAAAGAAAGCAAAAGGTAACAGAGTACAGGTGATTCTGGAATGTACAGAACACAAAGAGAGTGGAATGCCGGGAACATCTCGCTATATCACAAC
>gi|29345410:c1127066-1126806
ATGTCGAAGATTTGTCAAATTACCGGAAAGAAAGCCATGATTGGCAACAATGTTTCACACTCAAAGAGAAGAACAAAAAGAACCTTTGATTTGAACTT
>gi|29345410:1581041-1581595
ATGGCAACTAGAATCAGATTGCAAAGACACGGACGTAAAAGTTACGCGTTTTATTCAATTGTAATTGCAGACAGCAGAGCACCACGTGATGGTAAATT
>gi|29345410:2088036-2088290
ATGAAAAAAGGCCTTCATCCAGAATCATACCGTCCGGTAGTATTCAAAGATATGTCAAATGGTGATATGTTTTTGTCTAAATCAACTGTAGCAACTAA
```

2) Choose the genetic code corresponding to the sequences introduced. If the genetic code is not choose appropriately, it can generate some error messages.

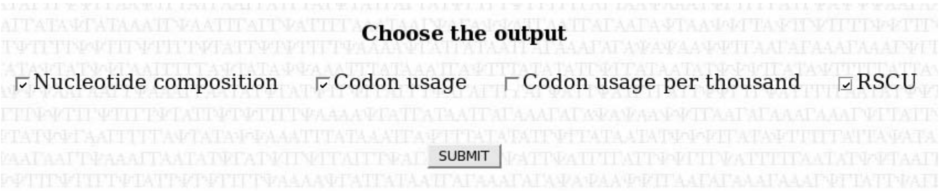


Choose the genetic code

11: Eubacterial

1: Standard
2: Vertebrate Mitochondrial
3: Yeast Mitochondrial
4: Mold, Protozoan, Coelenterate Mitochondrial and Mycoplasma/Spiroplasma
5: Invertebrate Mitochondrial
6: Ciliate Macronuclear and Dasycladacean
9: Echinoderm Mitochondrial
10: Alternative Ciliate Macronuclear
11: Eubacterial
12: Alternative Yeast
13: Ascidian Mitochondrial
14: Flatworm Mitochondrial
15: Blepharisma Nuclear Code

3) Finally, choose at least one of the four outputs available and click on the submit button.



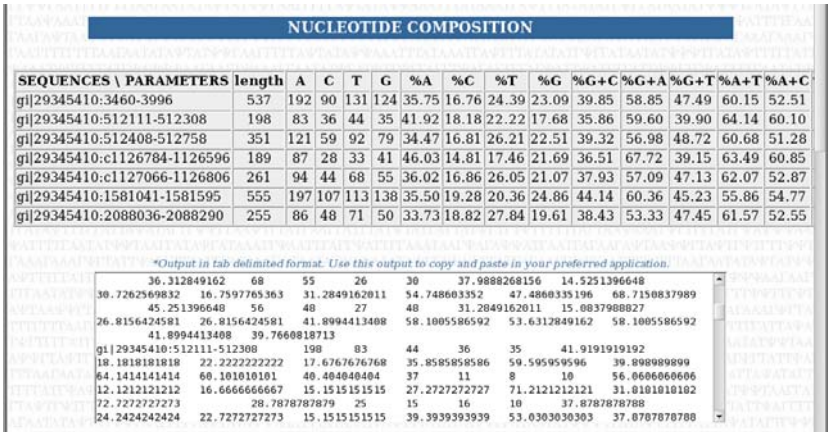
Choose the output

☑ Nucleotide composition ☑ Codon usage □ Codon usage per thousand ☑ RSCU

## OUTPUTS

This section has four outputs:

- Nucleotide composition.



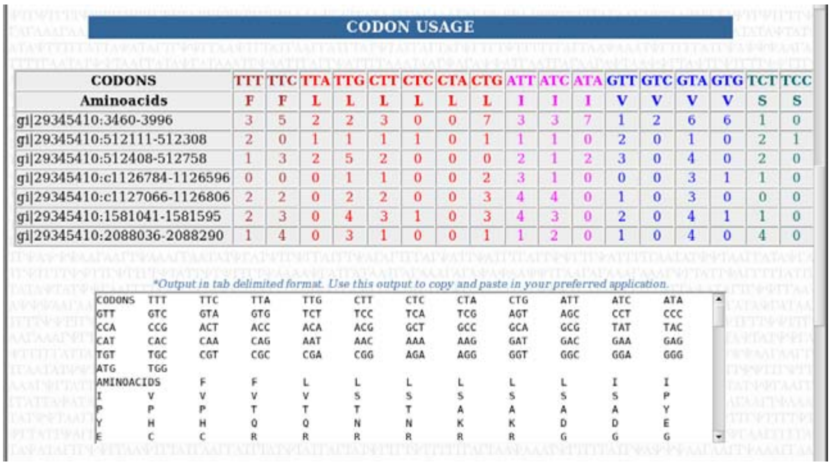
NUCLEOTIDE COMPOSITION

| SEQUENCES \ PARAMETERS | length | A | C | T | G | %A | %C | %T | %G | %G+C | %G+A | %G+T | %A+T | %A+C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| gi\|29345410:3460-3996 | 537 | 192 | 90 | 131 | 124 | 35.75 | 16.76 | 24.39 | 23.09 | 39.85 | 58.85 | 47.49 | 60.15 | 52.51 |
| gi\|29345410:512111-512308 | 198 | 83 | 36 | 44 | 35 | 41.92 | 18.18 | 22.22 | 17.68 | 35.86 | 59.60 | 39.90 | 64.14 | 60.10 |
| gi\|29345410:512408-512758 | 351 | 121 | 59 | 92 | 79 | 34.47 | 16.81 | 26.21 | 22.51 | 39.32 | 56.98 | 48.72 | 60.68 | 51.28 |
| gi\|29345410:c1126784-1126596 | 189 | 87 | 28 | 33 | 41 | 46.03 | 14.81 | 17.46 | 21.69 | 36.51 | 67.72 | 39.15 | 63.49 | 60.85 |
| gi\|29345410:c1127066-1126806 | 261 | 94 | 44 | 68 | 55 | 36.02 | 16.86 | 26.05 | 21.07 | 37.93 | 57.09 | 47.13 | 62.07 | 52.87 |
| gi\|29345410:1581041-1581595 | 555 | 197 | 107 | 113 | 138 | 35.50 | 19.28 | 20.36 | 24.86 | 44.14 | 60.36 | 45.23 | 55.86 | 54.77 |
| gi\|29345410:2088036-2088290 | 255 | 86 | 48 | 71 | 50 | 33.73 | 18.82 | 27.84 | 19.61 | 38.43 | 53.33 | 47.45 | 61.57 | 52.55 |

*Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
        36.312849162    68      55      26      30      37.9888268156   14.5251396648
30.7262569832   16.7597765363   31.2849162011   54.748603352    47.4860335196   68.7150837989
        45.251396648    56      48      27      48      31.2849162011   15.0837988827
26.8156424581   26.8156424581   41.8994413408   58.1005586592   53.6312849162   58.1005586592
        41.8994413408   39.7660818713
gi|29345410:512111-512308       198     83      44      36      35      41.9191919192
18.1818181818   22.2222222222   17.6767676768   35.8585858586   59.595959596    39.898989899
64.1414141414   60.101010101    40.404040404    37      11      8       10      56.0606060606
12.1212121212   16.6666666667   15.1515151515   27.2727272727   71.2121212121   31.8181818182
72.7272727273           28.7878787879   25      15      16      10      37.8787878788
24.2424242424   22.7272727273   15.1515151515   39.3939393939   53.0303030303   37.8787878788
```

- Codon usage.



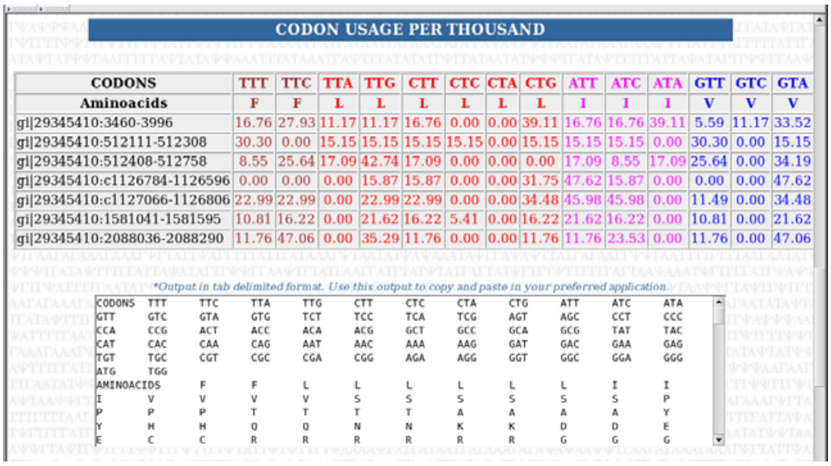
CODON USAGE

| CODONS | TTT | TTC | TTA | TTG | CTT | CTC | CTA | CTG | ATT | ATC | ATA | GTT | GTC | GTA | GTG | TCT | TCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aminoacids | F | F | L | L | L | L | L | L | I | I | I | V | V | V | V | S | S |
| gi\|29345410:3460-3996 | 3 | 5 | 2 | 2 | 3 | 0 | 0 | 7 | 3 | 3 | 7 | 1 | 2 | 6 | 6 | 1 | 0 |
| gi\|29345410:512111-512308 | 2 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 0 | 1 | 0 | 2 | 1 |
| gi\|29345410:512408-512758 | 1 | 3 | 2 | 5 | 2 | 0 | 0 | 0 | 2 | 1 | 2 | 3 | 0 | 4 | 0 | 2 | 0 |
| gi\|29345410:c1126784-1126596 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 3 | 1 | 0 | 0 | 0 | 3 | 1 | 1 | 0 |
| gi\|29345410:c1127066-1126806 | 2 | 2 | 0 | 2 | 2 | 0 | 0 | 3 | 4 | 4 | 0 | 1 | 0 | 3 | 0 | 0 | 0 |
| gi\|29345410:1581041-1581595 | 2 | 3 | 0 | 4 | 3 | 1 | 0 | 3 | 4 | 3 | 0 | 2 | 0 | 4 | 1 | 1 | 0 |
| gi\|29345410:2088036-2088290 | 1 | 4 | 0 | 3 | 1 | 0 | 0 | 1 | 1 | 2 | 0 | 1 | 0 | 4 | 0 | 4 | 0 |

*Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
CODONS  TTT     TTC     TTA     TTG     CTT     CTC     CTA     CTG     ATT     ATC     ATA
GTT     GTC     GTA     GTG     TCT     TCC     TCA     TCG     AGT     AGC     CCT     CCC
CCA     CCG     ACT     ACC     ACA     ACG     GCT     GCC     GCA     GCG     TAT     TAC
CAT     CAC     CAA     CAG     AAT     AAC     AAA     AAG     GAT     GAC     GAA     GAG
TGT     TGC     CGT     CGC     CGA     CGG     AGA     AGG     GGT     GGC     GGA     GGG
ATG     TGG
AMINOACIDS      F       F       L       L       L       L       L       L       I       I
I       V       V       V       V       S       S       S       S       S       S       P
P       P       P       T       T       T       T       A       A       A       A       Y
Y       H       H       Q       Q       N       N       K       K       D       D       E
E       C       C       R       R       R       R       R       R       G       G       G
```

- Codon usage per thousand.



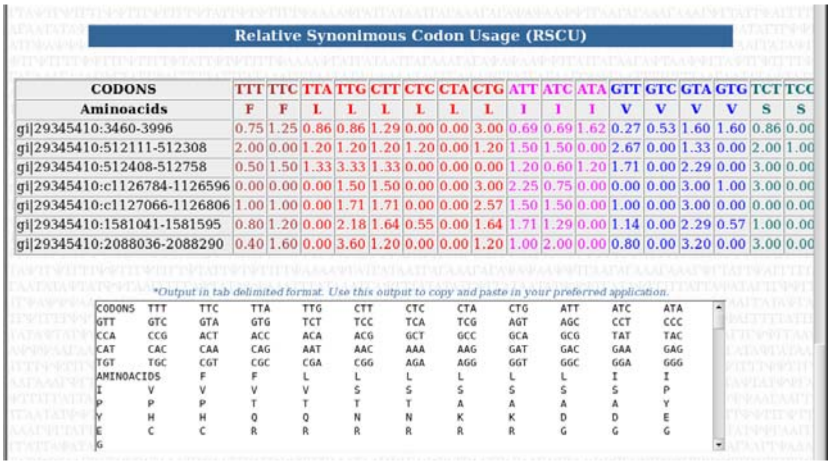
CODON USAGE PER THOUSAND

| CODONS | TTT | TTC | TTA | TTG | CTT | CTC | CTA | CTG | ATT | ATC | ATA | GTT | GTC | GTA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aminoacids | F | F | L | L | L | L | L | L | I | I | I | V | V | V |
| gi\|29345410:3460-3996 | 16.76 | 27.93 | 11.17 | 11.17 | 16.76 | 0.00 | 0.00 | 39.11 | 16.76 | 16.76 | 39.11 | 5.59 | 11.17 | 33.52 |
| gi\|29345410:512111-512308 | 30.30 | 0.00 | 15.15 | 15.15 | 15.15 | 15.15 | 0.00 | 15.15 | 15.15 | 15.15 | 0.00 | 30.30 | 0.00 | 15.15 |
| gi\|29345410:512408-512758 | 8.55 | 25.64 | 17.09 | 42.74 | 17.09 | 0.00 | 0.00 | 0.00 | 17.09 | 8.55 | 17.09 | 25.64 | 0.00 | 34.19 |
| gi\|29345410:c1126784-1126596 | 0.00 | 0.00 | 0.00 | 15.87 | 15.87 | 0.00 | 0.00 | 31.75 | 47.62 | 15.87 | 0.00 | 0.00 | 0.00 | 47.62 |
| gi\|29345410:c1127066-1126806 | 22.99 | 22.99 | 0.00 | 22.99 | 22.99 | 0.00 | 0.00 | 34.48 | 45.98 | 45.98 | 0.00 | 11.49 | 0.00 | 34.48 |
| gi\|29345410:1581041-1581595 | 10.81 | 16.22 | 0.00 | 21.62 | 16.22 | 5.41 | 0.00 | 16.22 | 21.62 | 16.22 | 0.00 | 10.81 | 0.00 | 21.62 |
| gi\|29345410:2088036-2088290 | 11.76 | 47.06 | 0.00 | 35.29 | 11.76 | 0.00 | 0.00 | 11.76 | 11.76 | 23.53 | 0.00 | 11.76 | 0.00 | 47.06 |

*Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
CODONS  TTT     TTC     TTA     TTG     CTT     CTC     CTA     CTG     ATT     ATC     ATA
GTT     GTC     GTA     GTG     TCT     TCC     TCA     TCG     AGT     AGC     CCT     CCC
CCA     CCG     ACT     ACC     ACA     ACG     GCT     GCC     GCA     GCG     TAT     TAC
CAT     CAC     CAA     CAG     AAT     AAC     AAA     AAG     GAT     GAC     GAA     GAG
TGT     TGC     CGT     CGC     CGA     CGG     AGA     AGG     GGT     GGC     GGA     GGG
ATG     TGG
AMINOACIDS      F       F       L       L       L       L       L       L       I       I
I       V       V       V       V       S       S       S       S       S       S       P
P       P       P       T       T       T       T       A       A       A       A       Y
Y       H       H       Q       Q       N       N       K       K       D       D       E
E       C       C       R       R       R       R       R       R       G       G       G
```

- Relative Synonymous Codon Usage.



Relative Synonimous Codon Usage (RSCU)

| CODONS | TTT | TTC | TTA | TTG | CTT | CTC | CTA | CTG | ATT | ATC | ATA | GTT | GTC | GTA | GTG | TCT | TCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aminoacids | F | F | L | L | L | L | L | L | I | I | I | V | V | V | V | S | S |
| gi\|29345410:3460-3996 | 0.75 | 1.25 | 0.86 | 0.86 | 1.29 | 0.00 | 0.00 | 3.00 | 0.69 | 0.69 | 1.62 | 0.27 | 0.53 | 1.60 | 1.60 | 0.86 | 0.00 |
| gi\|29345410:512111-512308 | 2.00 | 0.00 | 1.20 | 1.20 | 1.20 | 1.20 | 0.00 | 1.20 | 1.50 | 1.50 | 0.00 | 2.67 | 0.00 | 1.33 | 0.00 | 2.00 | 1.00 |
| gi\|29345410:512408-512758 | 0.50 | 1.50 | 1.33 | 3.33 | 1.33 | 0.00 | 0.00 | 0.00 | 1.20 | 0.60 | 1.20 | 1.71 | 0.00 | 2.29 | 0.00 | 3.00 | 0.00 |
| gi\|29345410:c1126784-1126596 | 0.00 | 0.00 | 0.00 | 1.50 | 1.50 | 0.00 | 0.00 | 3.00 | 2.25 | 0.75 | 0.00 | 0.00 | 0.00 | 3.00 | 1.00 | 3.00 | 0.00 |
| gi\|29345410:c1127066-1126806 | 1.00 | 1.00 | 0.00 | 1.71 | 1.71 | 0.00 | 0.00 | 2.57 | 1.50 | 1.50 | 0.00 | 1.00 | 0.00 | 3.00 | 0.00 | 0.00 | 0.00 |
| gi\|29345410:1581041-1581595 | 0.80 | 1.20 | 0.00 | 2.18 | 1.64 | 0.55 | 0.00 | 1.64 | 1.71 | 1.29 | 0.00 | 1.14 | 0.00 | 2.29 | 0.57 | 1.00 | 0.00 |
| gi\|29345410:2088036-2088290 | 0.40 | 1.60 | 0.00 | 3.60 | 1.20 | 0.00 | 0.00 | 1.20 | 1.00 | 2.00 | 0.00 | 0.80 | 0.00 | 3.20 | 0.00 | 3.00 | 0.00 |

*Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
CODONS  TTT     TTC     TTA     TTG     CTT     CTC     CTA     CTG     ATT     ATC     ATA
GTT     GTC     GTA     GTG     TCT     TCC     TCA     TCG     AGT     AGC     CCT     CCC
CCA     CCG     ACT     ACC     ACA     ACG     GCT     GCC     GCA     GCG     TAT     TAC
CAT     CAC     CAA     CAG     AAT     AAC     AAA     AAG     GAT     GAC     GAA     GAG
TGT     TGC     CGT     CGC     CGA     CGG     AGA     AGG     GGT     GGC     GGA     GGG
AMINOACIDS      F       F       L       L       L       L       L       L       I       I
I       V       V       V       V       S       S       S       S       S       S       P
P       P       P       T       T       T       T       A       A       A       A       Y
Y       H       H       Q       Q       N       N       K       K       D       D       E
E       C       C       R       R       R       R       R       R       G       G       G
G
```

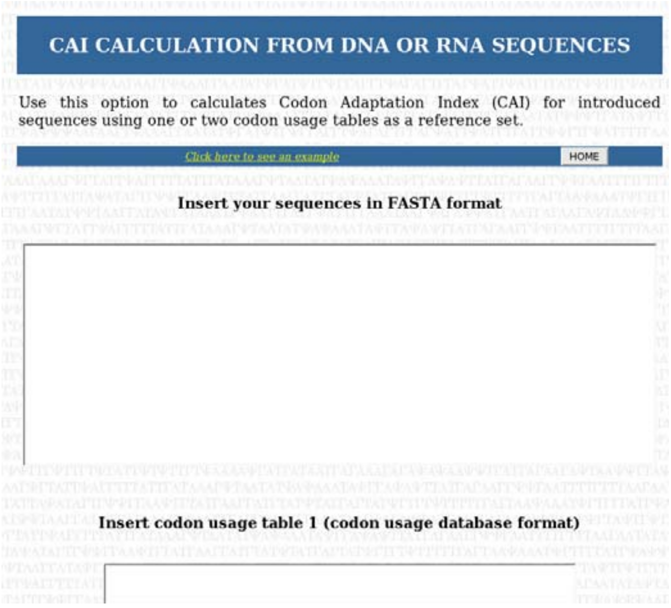
This section includes an output in tab-delimited format for each calculation. This output can be used to copy and paste into other applications.

# SECTION B - CAI calculation for FASTA sequences

## B1. CAI calculation

### OVERVIEW

Use this option to calculate the Codon Adaptation Index (CAI) for introduced sequences using one or two codon usage reference tables as a reference set.



CAI CALCULATION FROM DNA OR RNA SEQUENCES

Use this option to calculates Codon Adaptation Index (CAI) for introduced sequences using one or two codon usage tables as a reference set.

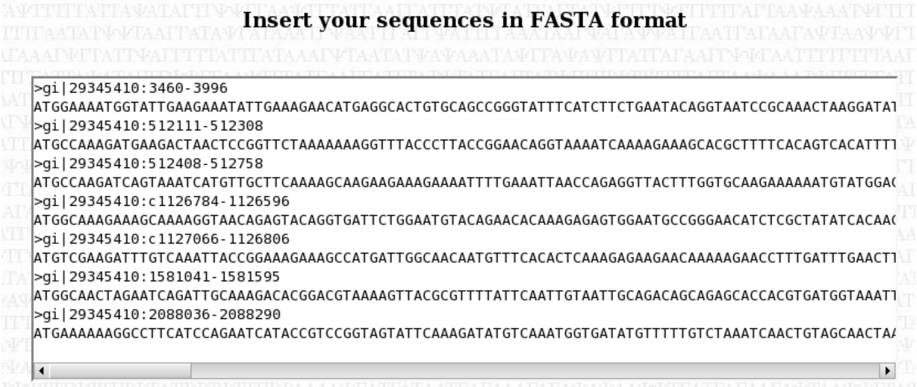
Insert your sequences in FASTA format

Insert codon usage table 1 (codon usage database format)

### INPUTS

This section requires four steps:

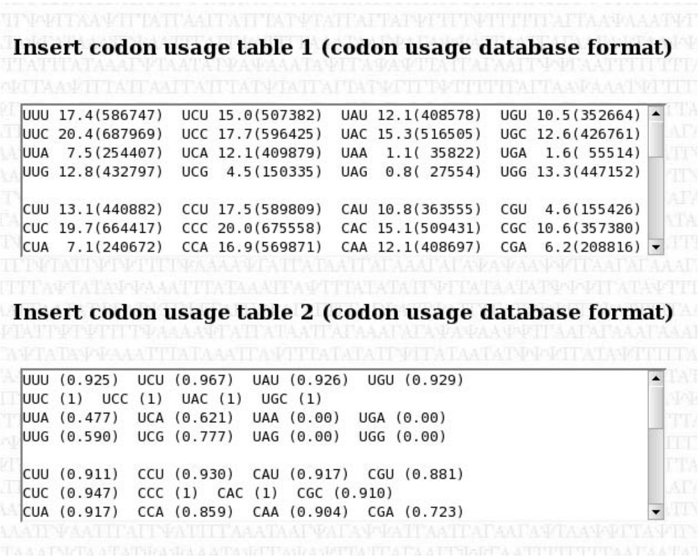
1) Introduction DNA or RNA sequences in FASTA format in a text box.



Insert your sequences in FASTA format

```
>gi|29345410:3460-3996
ATGGAAAATGGTATTGAAGAAATATTGAAAGAACATGAGGCACTGTGCAGCCGGGTATTTCATCTTCTGAATACAGGTAATCCGCAAACTAAGGATAT
>gi|29345410:512111-512308
ATGCCAAAGATGAAGACTAACTCCGGTTCTAAAAAAAGGTTTACCCTTACCGGAACAGGTAAAATCAAAAGAAAGCACGCTTTTCACAGTCACATTTT
>gi|29345410:512408-512758
ATGCCAAGATCAGTAAATCATGTTGCTTCAAAAGCAAGAAGAAAGAAAATTTTGAAATTAACCAGAGGTTACTTTGGTGCAAGAAAAAATGTATGGAC
>gi|29345410:c1126784-1126596
ATGGCAAAGAAAGCAAAAGGTAACAGAGTACAGGTGATTCTGGAATGTACAGAACACAAAGAGAGTGGAATGCCGGGAACATCTCGCTATATCACAAC
>gi|29345410:c1127066-1126806
ATGTCGAAGATTTGTCAAATTACCGGAAAGAAAGCCATGATTGGCAACAATGTTTCACACTCAAAGAGAAGAACAAAAAGAACCTTTGATTTGAACTT
>gi|29345410:1581041-1581595
ATGGCAACTAGAATCAGATTGCAAAGACACGGACGTAAAAGTTACGCGTTTTATTCAATTGTAATTGCAGACAGCAGAGCACCACGTGATGGTAAATT
>gi|29345410:2088036-2088290
ATGAAAAAAGGCCTTCATCCAGAATCATACCGTCCGGTAGTATTCAAAGATATGTCAAATGGTGATATGTTTTTGTCTAAATCAACTGTAGCAACTAA
```

2) Insert one or two codon usage reference tables. These reference tables can be obtained from codon usage databases or created by the user.



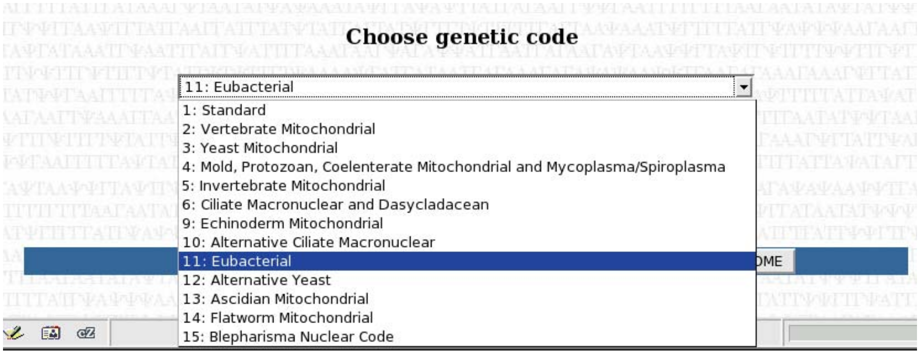
Insert codon usage table 1 (codon usage database format)

```
UUU 17.4(586747)  UCU 15.0(507382)  UAU 12.1(408578)  UGU 10.5(352664)
UUC 20.4(687969)  UCC 17.7(596425)  UAC 15.3(516505)  UGC 12.6(426761)
UUA  7.5(254407)  UCA 12.1(409879)  UAA  1.1( 35822)  UGA  1.6( 55514)
UUG 12.8(432797)  UCG  4.5(150335)  UAG  0.8( 27554)  UGG 13.3(447152)

CUU 13.1(440882)  CCU 17.5(589809)  CAU 10.8(363555)  CGU  4.6(155426)
CUC 19.7(664417)  CCC 20.0(675558)  CAC 15.1(509431)  CGC 10.6(357380)
CUA  7.1(240672)  CCA 16.9(569871)  CAA 12.1(408697)  CGA  6.2(208816)
```

Insert codon usage table 2 (codon usage database format)

```
UUU (0.925)  UCU (0.967)  UAU (0.926)  UGU (0.929)
UUC (1)  UCC (1)  UAC (1)  UGC (1)
UUA (0.477)  UCA (0.621)  UAA (0.00)  UGA (0.00)
UUG (0.590)  UCG (0.777)  UAG (0.00)  UGG (0.00)

CUU (0.911)  CCU (0.930)  CAU (0.917)  CGU (0.881)
CUC (0.947)  CCC (1)  CAC (1)  CGC (0.910)
CUA (0.917)  CCA (0.859)  CAA (0.904)  CGA (0.723)
```

3) Choose the genetic code corresponding to the sequences introduced. Choosing an unsuitable genetic code may generate errors messages.



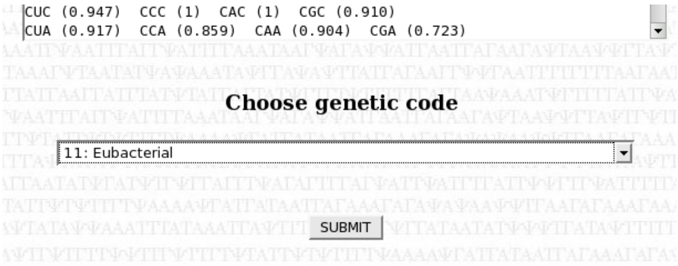
Choose genetic code

11: Eubacterial

1: Standard
2: Vertebrate Mitochondrial
3: Yeast Mitochondrial
4: Mold, Protozoan, Coelenterate Mitochondrial and Mycoplasma/Spiroplasma
5: Invertebrate Mitochondrial
6: Ciliate Macronuclear and Dasycladacean
9: Echinoderm Mitochondrial
10: Alternative Ciliate Macronuclear
11: Eubacterial
12: Alternative Yeast
13: Ascidian Mitochondrial
14: Flatworm Mitochondrial
15: Blepharisma Nuclear Code
OME

4)

5) Click on the submit button.



```
CUC (0.947)  CCC (1)  CAC (1)  CGC (0.910)
CUA (0.917)  CCA (0.859)  CAA (0.904)  CGA (0.723)
```

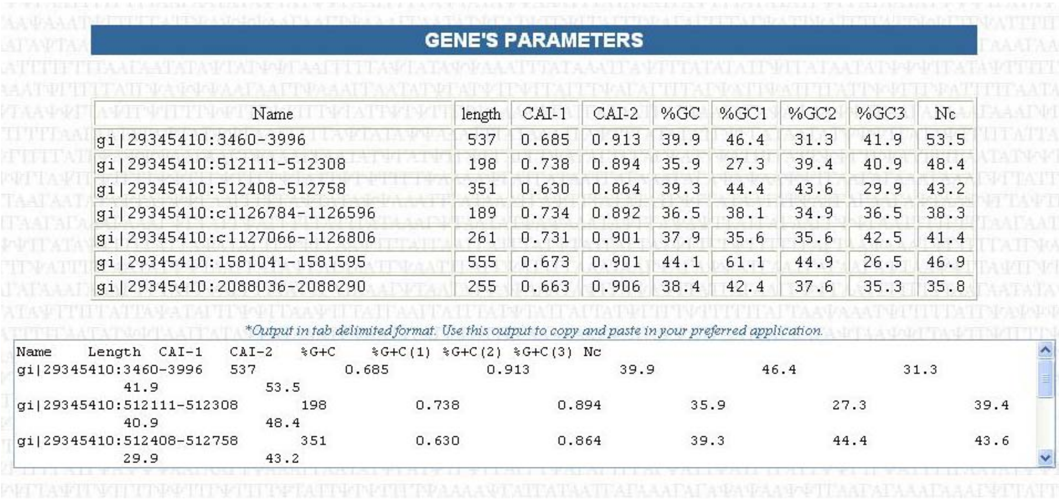
Choose genetic code

11: Eubacterial

**OUTPUTS**

The first output in this section are gene parameters such as CAI (Sharp and Li 1987) (CAI-1 and CAI-2 correspond to adaptation to codon usage reference tables 1 and 2 respectively), the effective number of codons (Nc) (Wright

1990) or G+C percentage. Additionally there is an output in tab delimited format. This output can be used to copy and paste into other applications.



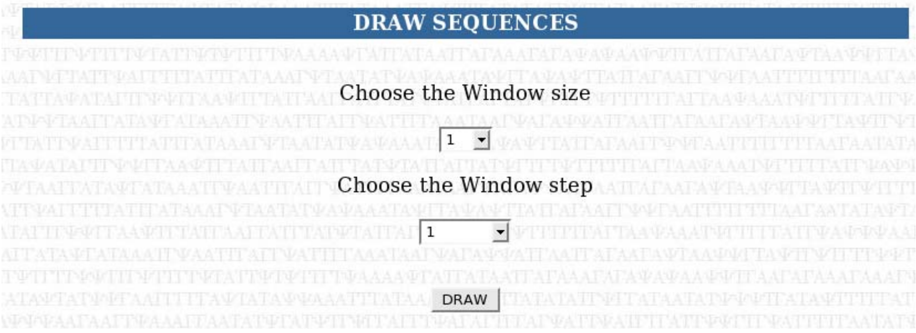
GENE'S PARAMETERS

| Name | length | CAI-1 | CAI-2 | %GC | %GC1 | %GC2 | %GC3 | Nc |
|---|---|---|---|---|---|---|---|---|
| gi\|29345410:3460-3996 | 537 | 0.685 | 0.913 | 39.9 | 46.4 | 31.3 | 41.9 | 53.5 |
| gi\|29345410:512111-512308 | 198 | 0.738 | 0.894 | 35.9 | 27.3 | 39.4 | 40.9 | 48.4 |
| gi\|29345410:512408-512758 | 351 | 0.630 | 0.864 | 39.3 | 44.4 | 43.6 | 29.9 | 43.2 |
| gi\|29345410:c1126784-1126596 | 189 | 0.734 | 0.892 | 36.5 | 38.1 | 34.9 | 36.5 | 38.3 |
| gi\|29345410:c1127066-1126806 | 261 | 0.731 | 0.901 | 37.9 | 35.6 | 35.6 | 42.5 | 41.4 |
| gi\|29345410:1581041-1581595 | 555 | 0.673 | 0.901 | 44.1 | 61.1 | 44.9 | 26.5 | 46.9 |
| gi\|29345410:2088036-2088290 | 255 | 0.663 | 0.906 | 38.4 | 42.4 | 37.6 | 35.3 | 35.8 |

*Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
Name     Length  CAI-1    CAI-2    %G+C      %G+C(1) %G+C(2) %G+C(3) Nc
gi|29345410:3460-3996   537          0.685          0.913          39.9          46.4          31.3
          41.9          53.5
gi|29345410:512111-512308       198          0.738          0.894          35.9          27.3          39.4
          40.9          48.4
gi|29345410:512408-512758       351          0.630          0.864          39.3          44.4          43.6
          29.9          43.2
```

This output provides a graphically visualization of the weight of each codon along a DNA sequence. The window size and length can be defined by the user.



DRAW SEQUENCES

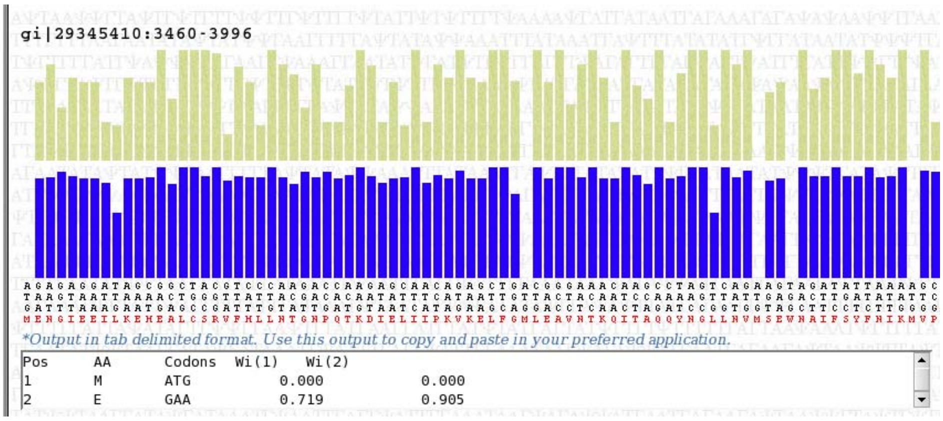
Choose the Window size

1

Choose the Window step

1

If the user has pasted two reference tables, the values for reference tables 1 and 2 are represented in yellow and blue bars, respectively. The values used to represent each figure are also included in a text box in tab-delimited format.

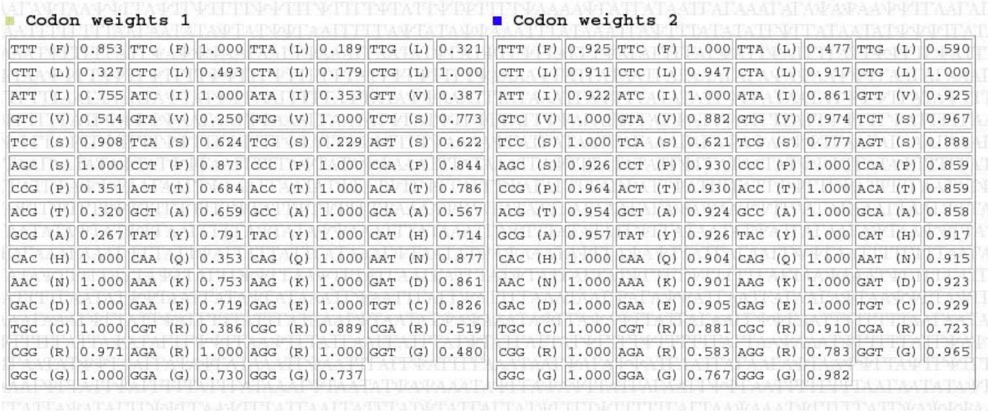


gi|29345410:3460-3996
AGAGAGGATAGCGGCTACGTCCCAAGACCAAGAGCAACAGAGCTGACGGGAAACAAGCCTAGTCAGAAGTAGATATTAAAAGC
TAAGTAATTAAAACTGGGTTATTACGACACAATATTTCATAATTGTTACTACAATCCAAAAGTTATTGAGACTTGATATATTC
GATTTAAAGAATGAGCCGATTTGTATTGATGTAATCATAGAAGCAGGACCTCAACTAGATCCGGTGGTAGCTTCCTCTTGGGG
MENGIEEILKEHEALCSRVFHLLNTGNPQTKDIELIIPKVKELFGMLEAVNTKQITAQQYNGLLNVMSEWNAIFSYFNIKMVP

*Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
Pos     AA      Codons  Wi(1)   Wi(2)
1       M       ATG          0.000          0.000
2       E       GAA          0.719          0.905
```

The graphical representation includes a table with the weight of each codon.



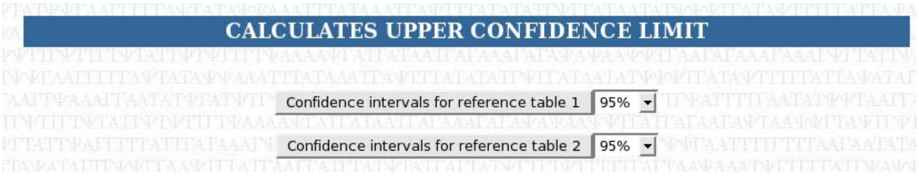
🟩 Codon weights 1

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TTT (F) | 0.853 | TTC (F) | 1.000 | TTA (L) | 0.189 | TTG (L) | 0.321 |
| CTT (L) | 0.327 | CTC (L) | 0.493 | CTA (L) | 0.179 | CTG (L) | 1.000 |
| ATT (I) | 0.755 | ATC (I) | 1.000 | ATA (I) | 0.353 | GTT (V) | 0.387 |
| GTC (V) | 0.514 | GTA (V) | 0.250 | GTG (V) | 1.000 | TCT (S) | 0.773 |
| TCC (S) | 0.908 | TCA (S) | 0.624 | TCG (S) | 0.229 | AGT (S) | 0.622 |
| AGC (S) | 1.000 | CCT (P) | 0.873 | CCC (P) | 1.000 | CCA (P) | 0.844 |
| CCG (P) | 0.351 | ACT (T) | 0.684 | ACC (T) | 1.000 | ACA (T) | 0.786 |
| ACG (T) | 0.320 | GCT (A) | 0.659 | GCC (A) | 1.000 | GCA (A) | 0.567 |
| GCG (A) | 0.267 | TAT (Y) | 0.791 | TAC (Y) | 1.000 | CAT (H) | 0.714 |
| CAC (H) | 1.000 | CAA (Q) | 0.353 | CAG (Q) | 1.000 | AAT (N) | 0.877 |
| AAC (N) | 1.000 | AAA (K) | 0.753 | AAG (K) | 1.000 | GAT (D) | 0.861 |
| GAC (D) | 1.000 | GAA (E) | 0.719 | GAG (E) | 1.000 | TGT (C) | 0.826 |
| TGC (C) | 1.000 | CGT (R) | 0.386 | CGC (R) | 0.889 | CGA (R) | 0.519 |
| CGG (R) | 0.971 | AGA (R) | 1.000 | AGG (R) | 1.000 | GGT (G) | 0.480 |
| GGC (G) | 1.000 | GGA (G) | 0.730 | GGG (G) | 0.737 | | |

🟦 Codon weights 2

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TTT (F) | 0.925 | TTC (F) | 1.000 | TTA (L) | 0.477 | TTG (L) | 0.590 |
| CTT (L) | 0.911 | CTC (L) | 0.947 | CTA (L) | 0.917 | CTG (L) | 1.000 |
| ATT (I) | 0.922 | ATC (I) | 1.000 | ATA (I) | 0.861 | GTT (V) | 0.925 |
| GTC (V) | 1.000 | GTA (V) | 0.882 | GTG (V) | 0.974 | TCT (S) | 0.967 |
| TCC (S) | 1.000 | TCA (S) | 0.621 | TCG (S) | 0.777 | AGT (S) | 0.888 |
| AGC (S) | 0.926 | CCT (P) | 0.930 | CCC (P) | 1.000 | CCA (P) | 0.859 |
| CCG (P) | 0.964 | ACT (T) | 0.930 | ACC (T) | 1.000 | ACA (T) | 0.859 |
| ACG (T) | 0.954 | GCT (A) | 0.924 | GCC (A) | 1.000 | GCA (A) | 0.858 |
| GCG (A) | 0.957 | TAT (Y) | 0.926 | TAC (Y) | 1.000 | CAT (H) | 0.917 |
| CAC (H) | 1.000 | CAA (Q) | 0.904 | CAG (Q) | 1.000 | AAT (N) | 0.915 |
| AAC (N) | 1.000 | AAA (K) | 0.901 | AAG (K) | 1.000 | GAT (D) | 0.923 |
| GAC (D) | 1.000 | GAA (E) | 0.905 | GAG (E) | 1.000 | TGT (C) | 0.929 |
| TGC (C) | 1.000 | CGT (R) | 0.881 | CGC (R) | 0.910 | CGA (R) | 0.723 |
| CGG (R) | 1.000 | AGA (R) | 0.583 | AGG (R) | 0.783 | GGT (G) | 0.965 |
| GGC (G) | 1.000 | GGA (G) | 0.767 | GGG (G) | 0.982 | | |

Another option is the calculation of the upper tolerance limit for the sequences introduced at 90%, 95% or 99% levels of confidence (see section B2 – calculation of expected CAI).



CALCULATES UPPER CONFIDENCE LIMIT

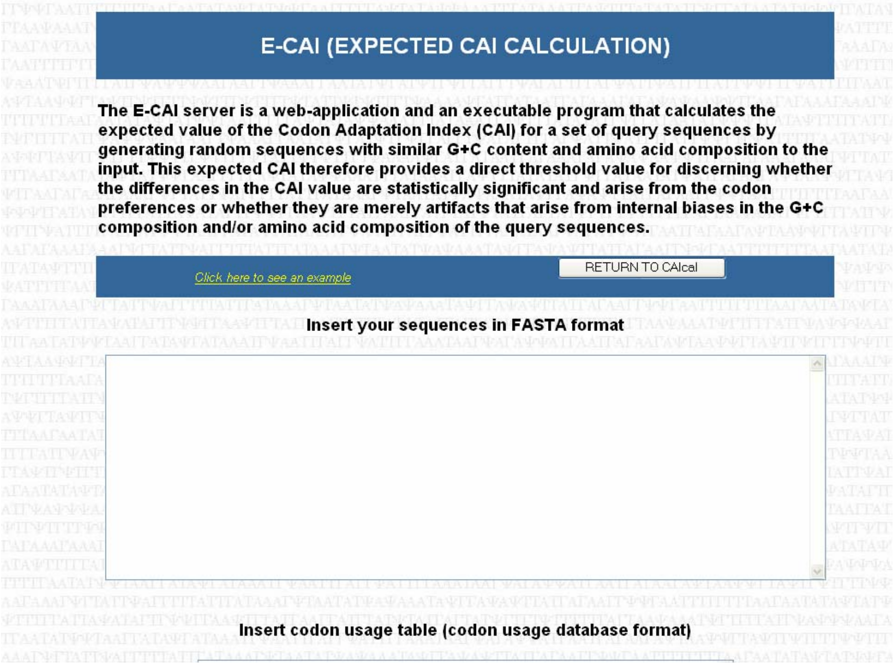
Confidence intervals for reference table 1: 95%

Confidence intervals for reference table 2: 95%

## B2. Expected CAI

### OVERVIEW

Use this option to calculate an expected CAI value determined by randomly generating 500 sequences with the same G+C content and amino acid composition as the query sequence.



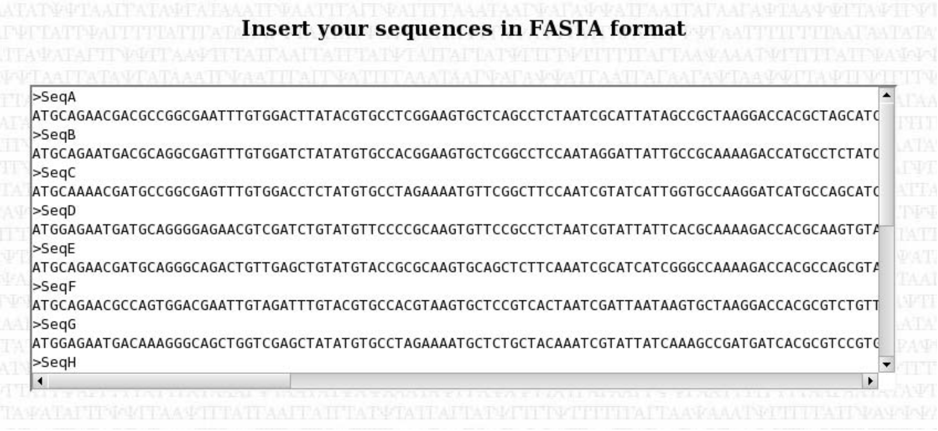
The E-CAI server is a web-application and an executable program that calculates the expected value of the Codon Adaptation Index (CAI) for a set of query sequences by generating random sequences with similar G+C content and amino acid composition to the input. This expected CAI therefore provides a direct threshold value for discerning whether the differences in the CAI value are statistically significant and arise from the codon preferences or whether they are merely artifacts that arise from internal biases in the G+C composition and/or amino acid composition of the query sequences.

Insert your sequences in FASTA format

Insert codon usage table (codon usage database format)

## INPUTS

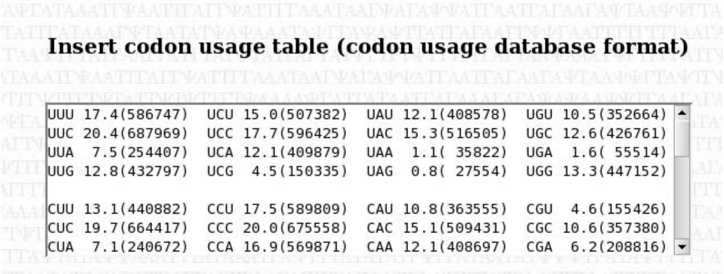
1) Introduce DNA or RNA sequences in FASTA format in the text box. These sequences will be used in this section as a reference to create 500 random sequences.



Insert your sequences in FASTA format

```
>SeqA
ATGCAGAACGACGCCGGCGAATTTGTGGACTTATACGTGCCTCGGAAGTGCTCAGCCTCTAATCGCATTATAGCCGCTAAGGACCACGCTAGCATC
>SeqB
ATGCAGAATGACGCAGGCGAGTTTGTGGATCTATATGTGCCACGGAAGTGCTCGGCCTCCAATAGGATTATTGCCGCAAAAGACCATGCCTCTATC
>SeqC
ATGCAAAACGATGCCGGCGAGTTTGTGGACCTCTATGTGCCTAGAAAATGTTCGGCTTCCAATCGTATCATTGGTGCCAAGGATCATGCCAGCATC
>SeqD
ATGGAGAATGATGCAGGGGAGAACGTCGATCTGTATGTTCCCCGCAAGTGTTCCGCCTCTAATCGTATTATTCACGCAAAAGACCACGCAAGTGTA
>SeqE
ATGCAGAACGATGCAGGGCAGACTGTTGAGCTGTATGTACCGCGCAAGTGCAGCTCTTCAAATCGCATCATCGGGCCAAAAGACCACGCCAGCGTA
>SeqF
ATGCAGAACGCCAGTGGACGAATTGTAGATTTGTACGTGCCACGTAAGTGCTCCGTCACTAATCGATTAATAAGTGCTAAGGACCACGCGTCTGTT
>SeqG
ATGGAGAATGACAAAGGGCAGCTGGTCGAGCTATATGTGCCTAGAAAATGCTCTGCTACAAATCGTATTATCAAAGCCGATGATCACGCGTCCGTC
>SeqH
```

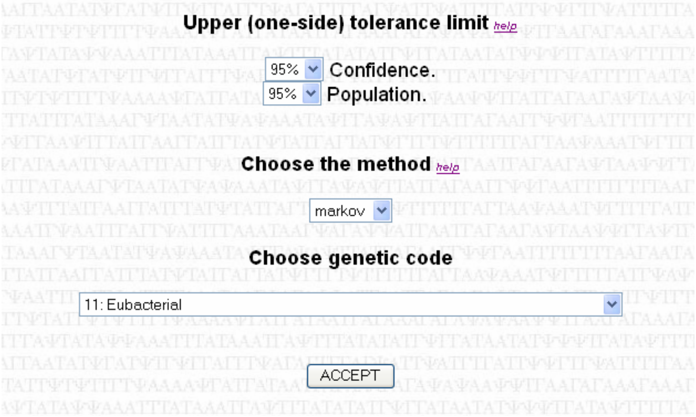
2) Insert the codon usage reference tables. These reference tables can be obtained from codon usage databases or created by the user.



**Insert codon usage table (codon usage database format)**

```
UUU 17.4(586747)  UCU 15.0(507382)  UAU 12.1(408578)  UGU 10.5(352664)
UUC 20.4(687969)  UCC 17.7(596425)  UAC 15.3(516505)  UGC 12.6(426761)
UUA  7.5(254407)  UCA 12.1(409879)  UAA  1.1( 35822)  UGA  1.6( 55514)
UUG 12.8(432797)  UCG  4.5(150335)  UAG  0.8( 27554)  UGG 13.3(447152)

CUU 13.1(440882)  CCU 17.5(589809)  CAU 10.8(363555)  CGU  4.6(155426)
CUC 19.7(664417)  CCC 20.0(675558)  CAC 15.1(509431)  CGC 10.6(357380)
CUA  7.1(240672)  CCA 16.9(569871)  CAA 12.1(408697)  CGA  6.2(208816)
```

3) Choose the upper confidence limit at 90%, 95% or 99%, the Markov or Poisson Method and the appropriate genetic code and click on the accept button.



**Upper (one-side) tolerance limit** help

95% Confidence.

95% Population.

**Choose the method** help

markov

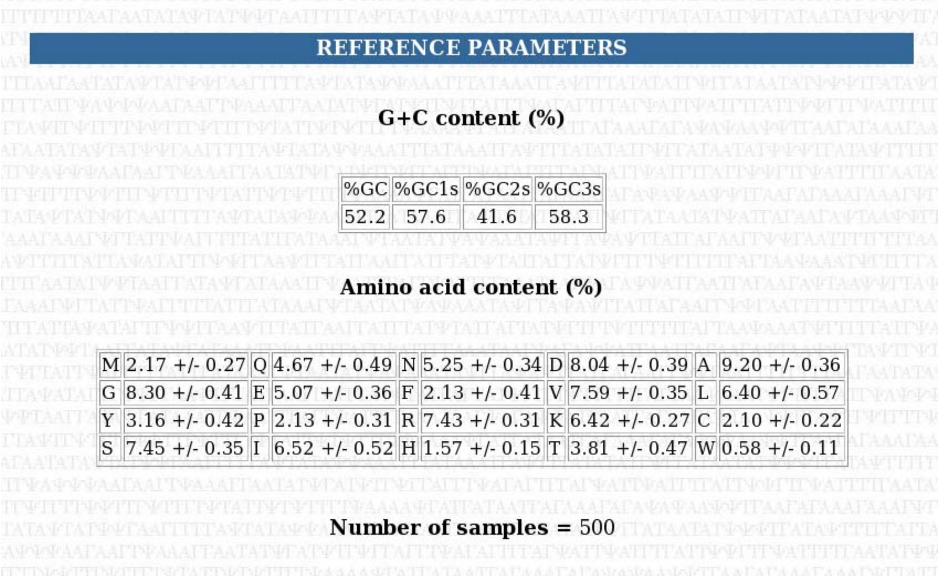
**Choose genetic code**

11: Eubacterial

**OUTPUTS**

There are two outputs in this section. The first of these is related to the parameters used to create random sequences and the second output is the expected CAI calculated.

1) Reference parameters from the introduced sequences, i.e. G+C percentage and amino acid composition



**REFERENCE PARAMETERS**

**G+C content (%)**

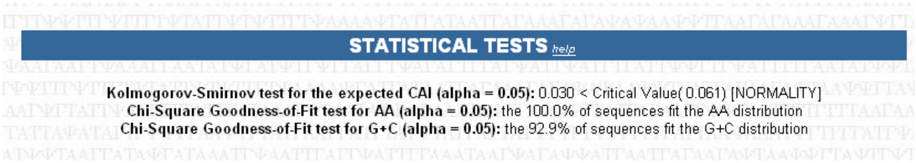
| %GC | %GC1s | %GC2s | %GC3s |
|---|---|---|---|
| 52.2 | 57.6 | 41.6 | 58.3 |

**Amino acid content (%)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| M | 2.17 +/- 0.27 | Q | 4.67 +/- 0.49 | N | 5.25 +/- 0.34 | D | 8.04 +/- 0.39 | A | 9.20 +/- 0.36 |
| G | 8.30 +/- 0.41 | E | 5.07 +/- 0.36 | F | 2.13 +/- 0.41 | V | 7.59 +/- 0.35 | L | 6.40 +/- 0.57 |
| Y | 3.16 +/- 0.42 | P | 2.13 +/- 0.31 | R | 7.43 +/- 0.31 | K | 6.42 +/- 0.27 | C | 2.10 +/- 0.22 |
| S | 7.45 +/- 0.35 | I | 6.52 +/- 0.52 | H | 1.57 +/- 0.15 | T | 3.81 +/- 0.47 | W | 0.58 +/- 0.11 |

**Number of samples =** 500

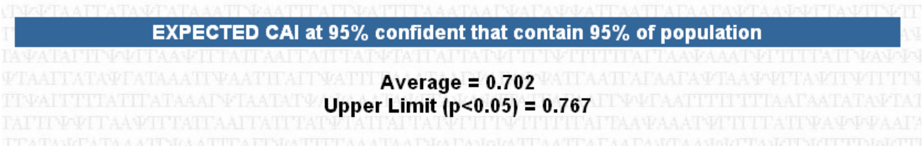
2) Statistical parameters: chi-square goodness-of-fit test and Kolmogorov-Smirnov test. A chi-square test is conducted to compare the goodness-of-fit between the amino acid frequencies or G+C content of each sequence of the query and their mean values. To check whether the CAI of the randomly generated sequences follow a normal distribution, a Kolmogorov-Smirnov test is made.



**STATISTICAL TESTS** help

**Kolmogorov-Smirnov test for the expected CAI (alpha = 0.05):** 0.030 < Critical Value( 0.061) [NORMALITY]
**Chi-Square Goodness-of-Fit test for AA (alpha = 0.05):** the 100.0% of sequences fit the AA distribution
**Chi-Square Goodness-of-Fit test for G+C (alpha = 0.05):** the 92.9% of sequences fit the G+C distribution

3) The figure below is an example of the calculated expected CAI value. The value of the expected CAI at a 95% level of confidence that contain 95% of population is *0.767*. Also included is the CAI average from the random sequences.



**EXPECTED CAI at 95% confident that contain 95% of population**

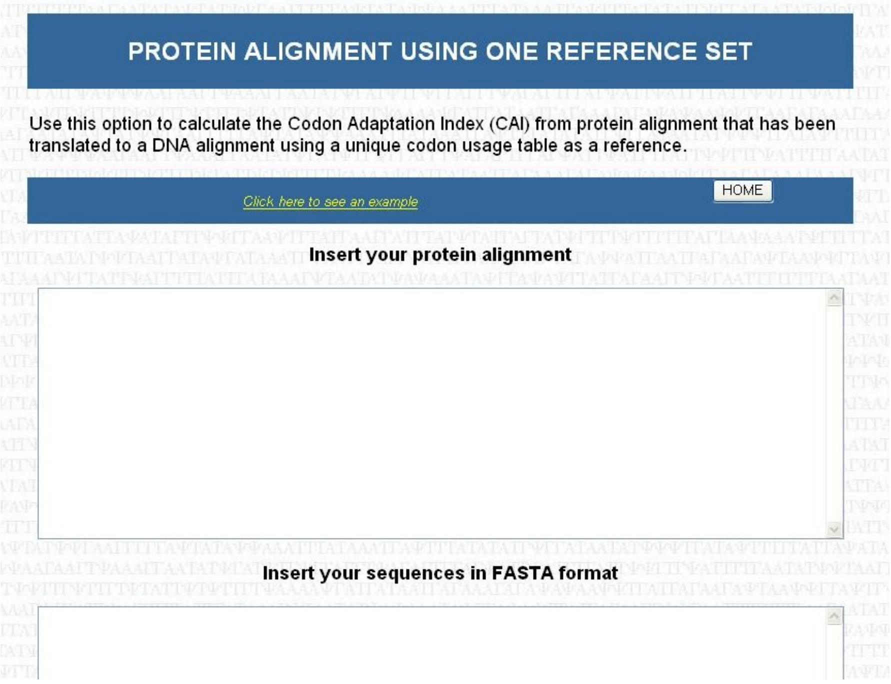
**Average = 0.702**
**Upper Limit (p<0.05) = 0.767**

# SECTION C - CAI calculation for protein alignment

## C1. Using one reference table

### OVERVIEW

Use this option to calculate the Codon Adaptation Index (CAI) from protein alignment translated to DNA using a unique codon usage table as reference.



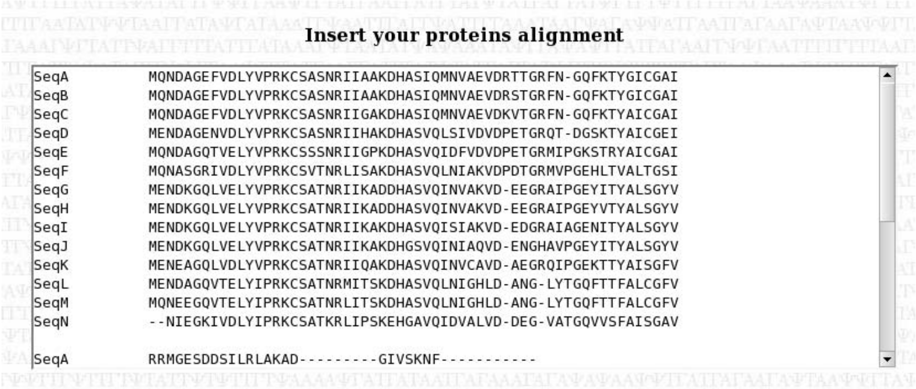
PROTEIN ALIGNMENT USING ONE REFERENCE SET

Use this option to calculate the Codon Adaptation Index (CAI) from protein alignment that has been translated to a DNA alignment using a unique codon usage table as a reference.

Click here to see an example
HOME
Insert your protein alignment
Insert your sequences in FASTA format

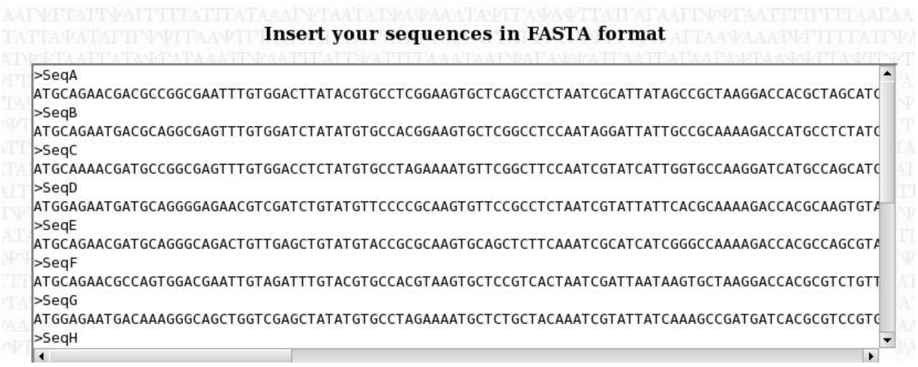
### INPUTS

1) Introduce the protein alignment in the text box.



```
SeqA        MQNDAGEFVDLYVPRKCSASNRIIAAKDHASIQMNVAEVDRTTGRFN-GQFKTYGICGAI
SeqB        MQNDAGEFVDLYVPRKCSASNRIIAAKDHASIQMNVAEVDRSTGRFN-GQFKTYGICGAI
SeqC        MQNDAGEFVDLYVPRKCSASNRIIGAKDHASIQMNVAEVDKVTGRFN-GQFKTYAICGAI
SeqD        MENDAGENVDLYVPRKCSASNRIIHAKDHASVQLSIVDVDPETGRQT-DGSKTYAICGEI
SeqE        MQNDAGQTVELYVPRKCSSSNRIIGPKDHASVQIDFVDVDPETGRMIPGKSTRYAICGAI
SeqF        MQNASGRIVDLYVPRKCSVTNRLISAKDHASVQLNIAKVDPDTGRMVPGEHLTVALTGSI
SeqG        MENDKGQLVELYVPRKCSATNRIIKADDHASVQINVAKVD-EEGRAIPGEYITYALSGYV
SeqH        MENDKGQLVELYVPRKCSATNRIIKADDHASVQINVAKVD-EEGRAIPGEYVTYALSGYV
SeqI        MENDKGQLVELYVPRKCSATNRIIKAKDHASVQISIAKVD-EDGRAIAGENITYALSGYV
SeqJ        MENDKGQLVELYVPRKCSATNRIIKAKDHGSVQINIAQVD-ENGHAVPGEYITYALSGYV
SeqK        MENEAGQLVDLYVPRKCSATNRIIQAKDHASVQINVCAVD-AEGRQIPGEKTTYAISGFV
SeqL        MENDAGQVTELYIPRKCSATNRMITSKDHASVQLNIGHLD-ANG-LYTGQFTTFALCGFV
SeqM        MQNEEGQVTELYIPRKCSATNRLITSKDHASVQLNIGHLD-ANG-LYTGQFTTFALCGFV
SeqN        --NIEGKIVDLYIPRKCSATKRLIPSKEHGAVQIDVALVD-DEG-VATGQVVSFAISGAV

SeqA        RRMGESDDSILRLAKAD---------GIVSKNF-----------
```

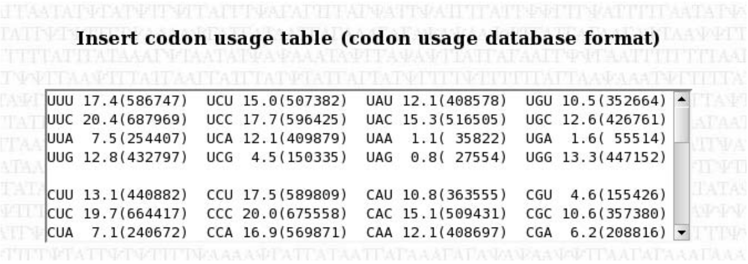
2) Insert the codon usage reference tables. These reference tables can be obtained from codon usage databases or created by the user.



Insert your sequences in FASTA format

```
>SeqA
ATGCAGAACGACGCCGGCGAATTTGTGGACTTATACGTGCCTCGGAAGTGCTCAGCCTCTAATCGCATTATAGCCGCTAAGGACCACGCTAGCATC
>SeqB
ATGCAGAATGACGCAGGCGAGTTTGTGGATCTATATGTGCCACGGAAGTGCTCGGCCTCCAATAGGATTATTGCCGCAAAAGACCATGCCTCTATC
>SeqC
ATGCAAAACGATGCCGGCGAGTTTGTGGACCTCTATGTGCCTAGAAAATGTTCGGCTTCCAATCGTATCATTGGTGCCAAGGATCATGCCAGCATC
>SeqD
ATGGAGAATGATGCAGGGGAGAACGTCGATCTGTATGTTCCCCGCAAGTGTTCCGCCTCTAATCGTATTATTCACGCAAAAGACCACGCAAGTGTA
>SeqE
ATGCAGAACGATGCAGGGCAGACTGTTGAGCTGTATGTACCGCGCAAGTGCAGCTCTTCAAATCGCATCATCGGGCCAAAAGACCACGCCAGCGTA
>SeqF
ATGCAGAACGCCAGTGGACGAATTGTAGATTTGTACGTGCCACGTAAGTGCTCCGTCACTAATCGATTAATAAGTGCTAAGGACCACGCGTCTGTT
>SeqG
ATGGAGAATGACAAAGGGCAGCTGGTCGAGCTATATGTGCCTAGAAAATGCTCTGCTACAAATCGTATTATCAAAGCCGATGATCACGCGTCCGTC
>SeqH
```

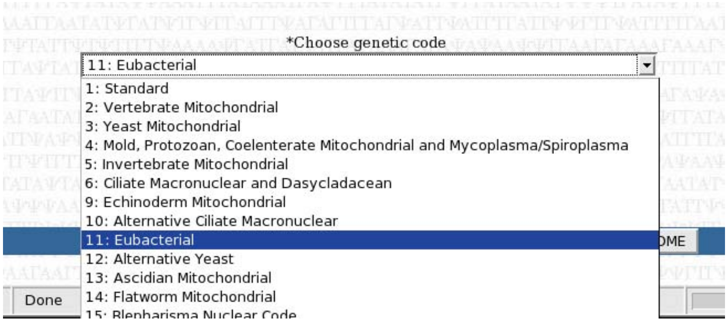
3) Insert the codon usage reference table. These reference tables can be obtained from codon usage database or created by the user.



Insert codon usage table (codon usage database format)

```
UUU 17.4(586747)  UCU 15.0(507382)  UAU 12.1(408578)  UGU 10.5(352664)
UUC 20.4(687969)  UCC 17.7(596425)  UAC 15.3(516505)  UGC 12.6(426761)
UUA  7.5(254407)  UCA 12.1(409879)  UAA  1.1( 35822)  UGA  1.6( 55514)
UUG 12.8(432797)  UCG  4.5(150335)  UAG  0.8( 27554)  UGG 13.3(447152)

CUU 13.1(440882)  CCU 17.5(589809)  CAU 10.8(363555)  CGU  4.6(155426)
CUC 19.7(664417)  CCC 20.0(675558)  CAC 15.1(509431)  CGC 10.6(357380)
CUA  7.1(240672)  CCA 16.9(569871)  CAA 12.1(408697)  CGA  6.2(208816)
```

4) Choose the genetic code corresponding to the sequences introduced. Choosing an unsuitable genetic code may generate error messages.

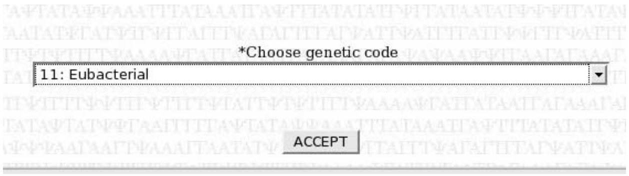


*Choose genetic code

11: Eubacterial

1: Standard
2: Vertebrate Mitochondrial
3: Yeast Mitochondrial
4: Mold, Protozoan, Coelenterate Mitochondrial and Mycoplasma/Spiroplasma
5: Invertebrate Mitochondrial
6: Ciliate Macronuclear and Dasycladacean
9: Echinoderm Mitochondrial
10: Alternative Ciliate Macronuclear
11: Eubacterial
12: Alternative Yeast
13: Ascidian Mitochondrial
14: Flatworm Mitochondrial
15: Blepharisma Nuclear Code
OME
Done

5) Click on the accept button.

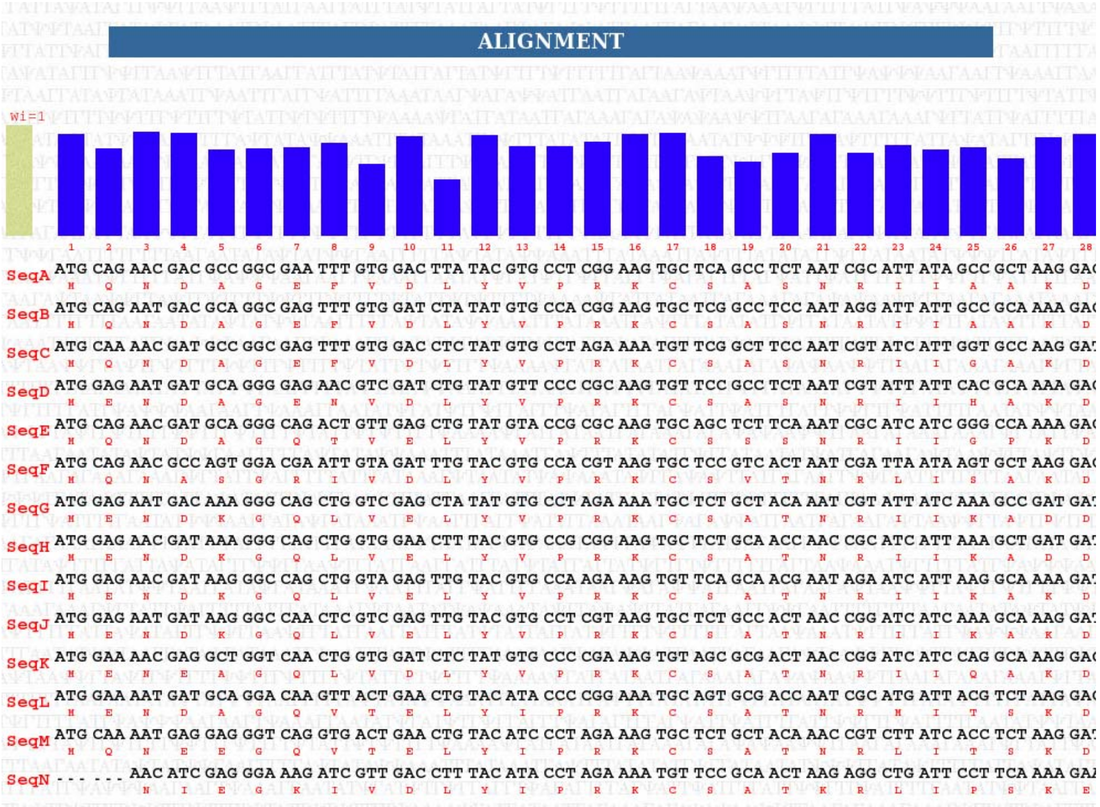


*Choose genetic code

11: Eubacterial

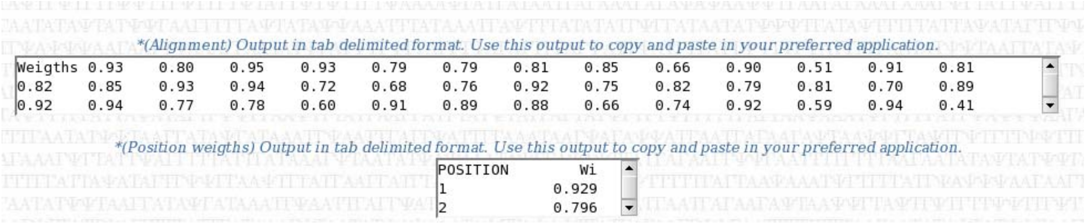
**OUTPUTS**

The main output in this section is the protein alignment translated to DNA with the mean weight of codons along the alignment.



```
SeqA ATG CAG AAC GAC GCC GGC GAA TTT GTG GAC TTA TAC GTG CCT CGG AAG TGC TCA GCC TCT AAT CGC ATT ATA GCC GCT AAG GAC
      M   Q   N   D   A   G   E   F   V   D   L   Y   V   P   R   K   C   S   A   S   N   R   I   I   A   A   K   D
SeqB ATG CAG AAT GAC GCA GGC GAG TTT GTG GAT CTA TAT GTG CCA CGG AAG TGC TCG GCC TCC AAT AGG ATT ATT GCC GCA AAA GAC
      M   Q   N   D   A   G   E   F   V   D   L   Y   V   P   R   K   C   S   A   S   N   R   I   I   A   A   K   D
SeqC ATG CAA AAC GAT GCC GGC GAG TTT GTG GAC CTC TAT GTG CCT AGA AAA TGT TCG GCT TCC AAT CGT ATC ATT GGT GCC AAG GAT
      M   Q   N   D   A   G   E   F   V   D   L   Y   V   P   R   K   C   S   A   S   N   R   I   I   G   A   K   D
SeqD ATG GAG AAT GAT GCA GGG GAG AAC GTC GAT CTG TAT GTT CCC CGC AAG TGT TCC GCC TCT AAT CGT ATT ATT CAC GCA AAA GAC
      M   E   N   D   A   G   E   N   V   D   L   Y   V   P   R   K   C   S   A   S   N   R   I   I   H   A   K   D
SeqE ATG CAG AAC GAT GCA GGG CAG ACT GTT GAG CTG TAT GTA CCG CGC AAG TGC AGC TCT TCA AAT CGC ATC ATC GGG CCA AAA GAC
      M   Q   N   D   A   G   Q   T   V   E   L   Y   V   P   R   K   C   S   S   S   N   R   I   I   G   P   K   D
SeqF ATG CAG AAC GCC AGT GGA CGA ATT GTA GAT TTG TAC GTG CCA CGT AAG TGC TCC GTC ACT AAT CGA TTA ATA AGT GCT AAG GAC
      M   Q   N   A   S   G   R   I   V   D   L   Y   V   P   R   K   C   S   V   T   N   R   L   I   S   A   K   D
SeqG ATG GAG AAT GAC AAA GGG CAG CTG GTC GAG CTA TAT GTG CCT AGA AAA TGC TCT GCT ACA AAT CGT ATT ATC AAA GCC GAT GAT
      M   E   N   D   K   G   Q   L   V   E   L   Y   V   P   R   K   C   S   A   T   N   R   I   I   K   A   D   D
SeqH ATG GAG AAC GAT AAA GGG CAG CTG GTG GAA CTT TAC GTG CCG CGG AAG TGC TCT GCA ACC AAC CGC ATC ATT AAA GCT GAT GAT
      M   E   N   D   K   G   Q   L   V   E   L   Y   V   P   R   K   C   S   A   T   N   R   I   I   K   A   D   D
SeqI ATG GAG AAC GAT AAG GGC CAG CTG GTA GAG TTG TAC GTG CCA AGA AAG TGT TCA GCA ACG AAT AGA ATC ATT AAG GCA AAA GAT
      M   E   N   D   K   G   Q   L   V   E   L   Y   V   P   R   K   C   S   A   T   N   R   I   I   K   A   K   D
SeqJ ATG GAG AAT GAT AAG GGC CAA CTC GTC GAG TTG TAC GTG CCT CGT AAG TGC TCT GCC ACT AAC CGG ATC ATC AAA GCA AAG GAT
      M   E   N   D   K   G   Q   L   V   E   L   Y   V   P   R   K   C   S   A   T   N   R   I   I   K   A   K   D
SeqK ATG GAA AAC GAG GCT GGT CAA CTG GTG GAT CTC TAT GTG CCC CGA AAG TGT AGC GCG ACT AAC CGG ATC ATC CAG GCA AAG GAC
      M   E   N   E   A   G   Q   L   V   D   L   Y   V   P   R   K   C   S   A   T   N   R   I   I   Q   A   K   D
SeqL ATG GAA AAT GAT GCA GGA CAA GTT ACT GAA CTG TAC ATA CCC CGG AAA TGC AGT GCG ACC AAT CGC ATG ATT ACG TCT AAG GAC
      M   E   N   D   A   G   Q   V   T   E   L   Y   I   P   R   K   C   S   A   T   N   R   M   I   T   S   K   D
SeqM ATG CAA AAT GAG GAG GGT CAG GTG ACT GAA CTG TAC ATC CCT AGA AAG TGC TCT GCT ACA AAC CGT CTT ATC ACC TCT AAG GAT
      M   Q   N   E   E   G   Q   V   T   E   L   Y   I   P   R   K   C   S   A   T   N   R   L   I   T   S   K   D
SeqN --- --- AAC ATC GAG GGA AAG ATC GTT GAC CTT TAC ATA CCT AGA AAA TGT TCC GCA ACT AAG AGG CTG ATT CCT TCA AAA GAA
              N   I   E   G   K   I   V   D   L   Y   I   P   R   K   C   S   A   T   K   R   L   I   P   S   K   E
```

We have included two tab-delimited formats, the first one has the complete result and the second one has just the mean weight of the codons and their position.

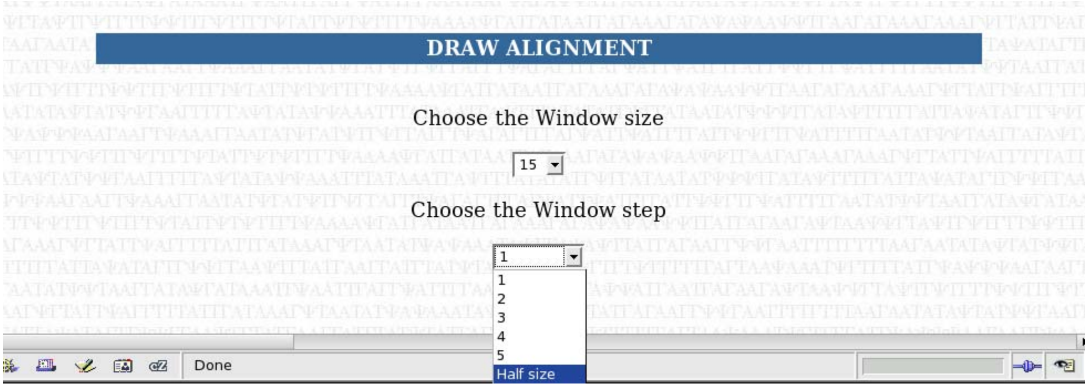

*(Alignment) Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
Weigths 0.93    0.80    0.95    0.93    0.79    0.79    0.81    0.85    0.66    0.90    0.51    0.91    0.81
0.82    0.85    0.93    0.94    0.72    0.68    0.76    0.92    0.75    0.82    0.79    0.81    0.70    0.89
0.92    0.94    0.77    0.78    0.60    0.91    0.89    0.88    0.66    0.74    0.92    0.59    0.94    0.41
```

*(Position weigths) Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
POSITION        Wi
1            0.929
2            0.796
```

The weight of the codons along the alignment can also be visualized by changing the step and the window size.

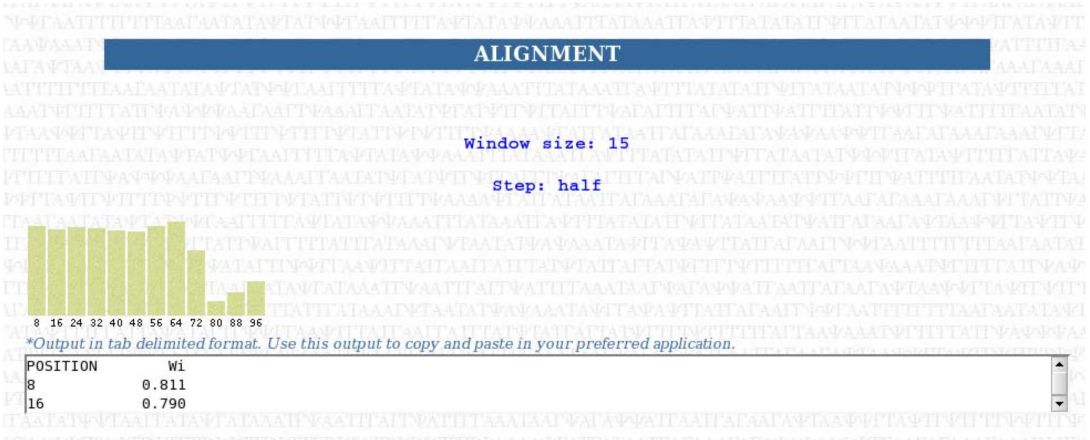


Choose the Window size

15

Choose the Window step

1



ALIGNMENT

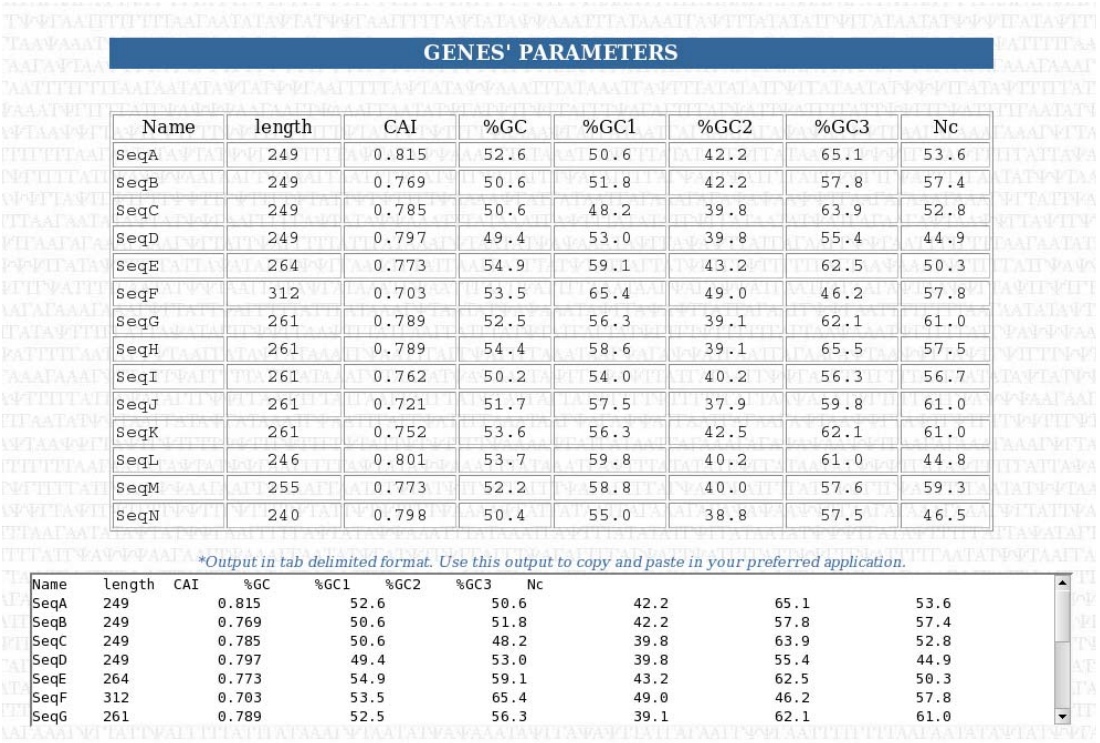
Window size: 15

Step: half

*Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
POSITION        Wi
8               0.811
16              0.790
```

Another output are the gene parameters in a table or in tab-delimited format for copying and pasting them into a spreadsheet program.



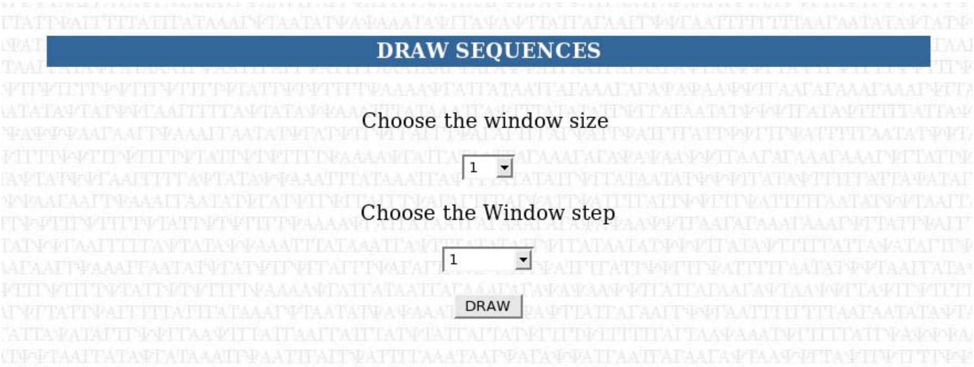
GENES' PARAMETERS

| Name | length | CAI | %GC | %GC1 | %GC2 | %GC3 | Nc |
|---|---|---|---|---|---|---|---|
| SeqA | 249 | 0.815 | 52.6 | 50.6 | 42.2 | 65.1 | 53.6 |
| SeqB | 249 | 0.769 | 50.6 | 51.8 | 42.2 | 57.8 | 57.4 |
| SeqC | 249 | 0.785 | 50.6 | 48.2 | 39.8 | 63.9 | 52.8 |
| SeqD | 249 | 0.797 | 49.4 | 53.0 | 39.8 | 55.4 | 44.9 |
| SeqE | 264 | 0.773 | 54.9 | 59.1 | 43.2 | 62.5 | 50.3 |
| SeqF | 312 | 0.703 | 53.5 | 65.4 | 49.0 | 46.2 | 57.8 |
| SeqG | 261 | 0.789 | 52.5 | 56.3 | 39.1 | 62.1 | 61.0 |
| SeqH | 261 | 0.789 | 54.4 | 58.6 | 39.1 | 65.5 | 57.5 |
| SeqI | 261 | 0.762 | 50.2 | 54.0 | 40.2 | 56.3 | 56.7 |
| SeqJ | 261 | 0.721 | 51.7 | 57.5 | 37.9 | 59.8 | 61.0 |
| SeqK | 261 | 0.752 | 53.6 | 56.3 | 42.5 | 62.1 | 61.0 |
| SeqL | 246 | 0.801 | 53.7 | 59.8 | 40.2 | 61.0 | 44.8 |
| SeqM | 255 | 0.773 | 52.2 | 58.8 | 40.0 | 57.6 | 59.3 |
| SeqN | 240 | 0.798 | 50.4 | 55.0 | 38.8 | 57.5 | 46.5 |

*Output in tab delimited format. Use this output to copy and paste in your preferred application.

```
Name    length  CAI     %GC     %GC1    %GC2    %GC3    Nc
SeqA    249         0.815       52.6        50.6        42.2        65.1        53.6
SeqB    249         0.769       50.6        51.8        42.2        57.8        57.4
SeqC    249         0.785       50.6        48.2        39.8        63.9        52.8
SeqD    249         0.797       49.4        53.0        39.8        55.4        44.9
SeqE    264         0.773       54.9        59.1        43.2        62.5        50.3
SeqF    312         0.703       53.5        65.4        49.0        46.2        57.8
SeqG    261         0.789       52.5        56.3        39.1        62.1        61.0
```

The weight of codons along the sequences can be visualized simply by selecting the window size and the step.



DRAW SEQUENCES

Choose the window size

1

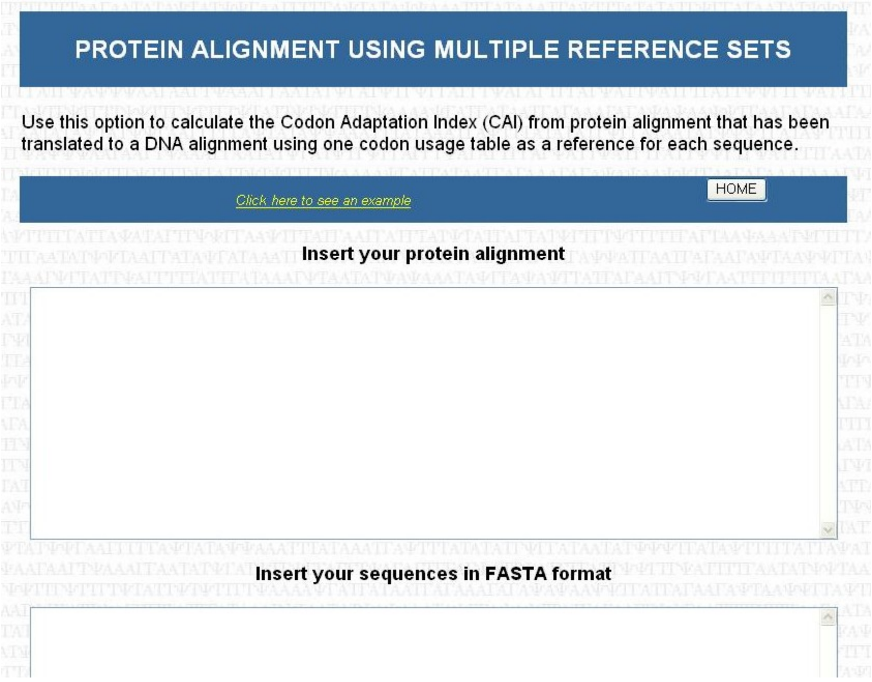
Choose the Window step

1

## C2. Using one reference for each sequence

### OVERVIEW

Use this option to calculate the Codon Adaptation Index (CAI) from protein alignment translated to DNA using a codon usage table as a reference for each sequence.



PROTEIN ALIGNMENT USING MULTIPLE REFERENCE SETS

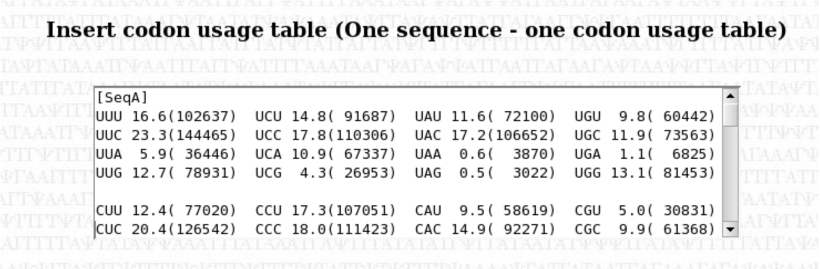
Use this option to calculate the Codon Adaptation Index (CAI) from protein alignment that has been translated to a DNA alignment using one codon usage table as a reference for each sequence.

Insert your protein alignment

Insert your sequences in FASTA format

### INPUTS

This section requires the same inputs as in section C1 but requires just one codon usage table for each sequence introduced. See “Input Requirements” from this tutorial.



Insert codon usage table (One sequence - one codon usage table)

```
[SeqA]
UUU 16.6(102637)  UCU 14.8( 91687)  UAU 11.6( 72100)  UGU  9.8( 60442)
UUC 23.3(144465)  UCC 17.8(110306)  UAC 17.2(106652)  UGC 11.9( 73563)
UUA  5.9( 36446)  UCA 10.9( 67337)  UAA  0.6(  3870)  UGA  1.1(  6825)
UUG 12.7( 78931)  UCG  4.3( 26953)  UAG  0.5(  3022)  UGG 13.1( 81453)

CUU 12.4( 77020)  CCU 17.3(107051)  CAU  9.5( 58619)  CGU  5.0( 30831)
CUC 20.4(126542)  CCC 18.0(111423)  CAC 14.9( 92271)  CGC  9.9( 61368)
```

## Automation of some of the calculations

To allow the calculation of CAI values for hundreds or thousands of sequences on a whole-genome scale and generate an expected value, users of the CAIcal server can download a Perl script that automatically performs these calculations. In addition, several parameters fixed in the CAIcal server (like the number and length of randomly generated sequences) can be specified in this script version.

### How to run it

From the main page of the CAIcal/E-CAI server download the Perl script after introducing your name, institution and e-mail address.

Uncompress the file. In a Linux operative system you can do it by typing: *(Replace * for the appropriate version)*

```
$ tar -xvf CAIcal_ECAI_v*.tar.gz
```

```
$ gunzip CAIcal_ECAI_v*.gz
```

This will generate a directory called CAIcal_v*. Enter to this directory:

```
$ cd CAIcal_ECAI_v*
```

To see how to run CAIcal/E-CAI execute the following command:

```
$ perl CAIcal_ECAI _v*.pl -help
```

To run an example, execute the following command:

```
$ perl CAIcal_ECAI _v*.pl
```

You will need to install a Perl interpret to execute the script in a Windows operative system.

### Parameters to run CAIcal/E-CAI

The script needs several parameters to be executed. These parameters are specified when executing the script:

```
$ perl CAIcal_ECAI_v -e [cai|expected|cai_and_expected] -f
[file_name] -h [file_name] -g [1|4|11] -c [90|95|99] -p
[90|95|99] -o1 [file_name] -o2 [file_name] -o3 [file_name] -n
[number] -l [number] –m [markov/poisson]
```

- PROGRAM TO EXECUTE: -e cai|expected|cai_and_expected <default: cai_and_expected>. The option 'cai' calculates only the CAIs of the query sequences. The ‘expected’ option calculates only an expected CAI and the 'cai_and_expected' option calculates both.

- INPUT1 DATA FILE: -f file_name <default: example.ffn>. The DNA sequences in the input file have to be in fasta format.

- INPUT2 HOST FILE: -h file_name <default: human>. This file has to contain a Codon usage reference table in the Codon Usage Database format.

- GENETIC CODE: -g 1|2|3|4|5|6|9|10|11|12|13|14|15 <default: 1>. This option allows choosing the Standard (1), Eubacteria (11), Mycoplasma (4) and other genetic codes.

- CONFIDENCE LEVEL: -c 90|95|99 <default: 95>.

- PERCENTAGE OF POPULATION OR COVERAGE: -p 90|95|99 <default: 95>.

- OUTPUT1 (CAI: -o1 file_name <default: cai>

- OUTPUT2 (CAI random sequences): -o2 file_name <default:random_sequences_and_cai>

- OUTPUT3(EXPECTED CAI: -o3 file_name <default: expected>

- METHOD: -m markov/poisson <default: markov>

- NUMBER OF RANDOMLY GENERATED SEQUENCES: -n number <default: 1000>

- LENGTH OF RANDOMLY GENERATED SEQUENCES (IN CODONS): -l number <default: 300 (100 in Poisson Method)>

- HELP –help.

## References

- Nakamura, Y., Gojobori, T. and Ikemura, T. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. *Nucl. Acids Res.* 28, 292.
- Sharp, P.M. and Li, W. (1987) The Codon adaptation index -a measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Res.*, 15:1281-1295.
- Wright, F. (1990) The 'effective number of codons' used in a gene. *Gene*, 87:23-29.