CodonZ

Installation and Operation Guide

CodonZ is software for replacing the codons in a given gene with a set of codons designed to improve protein expression levels. This manual covers setting up and operating CodonZ.

CodonZ is written by Harold Burgess (NICHD, haroldburgess@mail.nih.gov).

9/1/2014 version 2.1

Table of Contents

1. Installation

1. CodonZ is written in IDL and packaged as runtime code that can operate on any computer with the free 'virtual machine' platform from IDL. CodonZ will need to be installed on a PC, running Windows XP or 7. Download CodonZ runtime code (contained in the file 'codonz2.sav') from:
https://science.nichd.nih.gov/confluence/display/burgess/Software

2. Create a directory C:\Program Files\CodonZ. Then place the codonz2.sav file in the CodonZ folder

3. Download the IDL virtual machine: go to the Exelis Visual Information Solutions website (http://www.exelisvis.com/) and if you do not already have an account, register as a new user (click on "My Account" and follow the instructions for setting up a new account). You will receive an email with a link to verify your account, which will allow you to login to the website. Approval and the ability to download can take an additional 24 hours, however. It seems that although an initial verification email is sent, there is no email notification for approval being granted, so periodically check whether downloads are now accessible once you have logged in to the Exelis website. Then download IDL (this will be more than 1 Gb). You may have to perform a full IDL installation, but you do not need to purchase a license to run IDL runtime code.

4. Install software that calculates the RNA secondary structure. Our initial analyses were performed using the UNAfold package. However you may now need to pay a fee to use this software. We have therefore modified CodonZ to use the (currently) free software package ViennaRNA. Either package is fine and you do not need both. CodonZ will detect which package you have installed. If you have both installed, it will use UNAfold by default.

4. 1. Installation of UNAfold. Download at:
http://mfold.rna.albany.edu/?q=DINAMelt/software
or
http://dinamelt.rit.albany.edu/download.php

Install to into a directory called C:\Program Files\UNAFold
Note that this may not be the default directory! For some computers the default will be C:\Program Files (x86) \UNAFold, and you need to change the installation parameters. Specifically, CodonZ will look for this file:
C:\Program Files\UNAFold\bin\hybrid-ss-min.exe

4. 2. Installation of ViennaRNA. Download at:
http://www.tbi.univie.ac.at/RNA/index.html#download

Install to the default directory. CodonZ will look for this file:
C:\Program Files\ViennaRNA Package\RNAfold.exe

2. Codon modification

2. 1. Quick start

1. Paste the amino acid sequence you want to encode into the box on the left. If the sequence contains spaces or line breaks, press <u>Cleanup</u> to remove these.

If you are starting with a DNA sequence, paste it into the box on the right. Press 'Cleanup' to remove spaces, numbers and line breaks. Then press '>>> Prot' to translate to the cognate amino acid sequence.

2. Under the <u>Organism</u> drop-down menu on top, select Zebrafish/Mouse as appropriate.

3. Press the <u>Optimize</u> to produce the DNA sequence with codons optimized according to default settings. CodonZ will repeatedly run UNAfold or ViennaRNA software for secondary structure prediction, so that the taskbar and/or windows will appear to flicker on screen. This is normal and will take a couple of minutes. A progress indicator will replace the information about the current DNA sequence.

4. Copy the new DNA sequence from the box on the right and send it for synthesis.

2. 2. Overview of user interface

Elements of the main interface are highlighted in Figure 1. The two major text boxes on the left and right are for protein and nucleotide sequence respectively. The small text box below the protein box is for nucleotide sequences that are to be avoided when generating nucleotide sequence. The text box below the nucleotide sequence box will be used in future implementations for specifying regions of the nucleotide sequence that should be filled with minor codons.

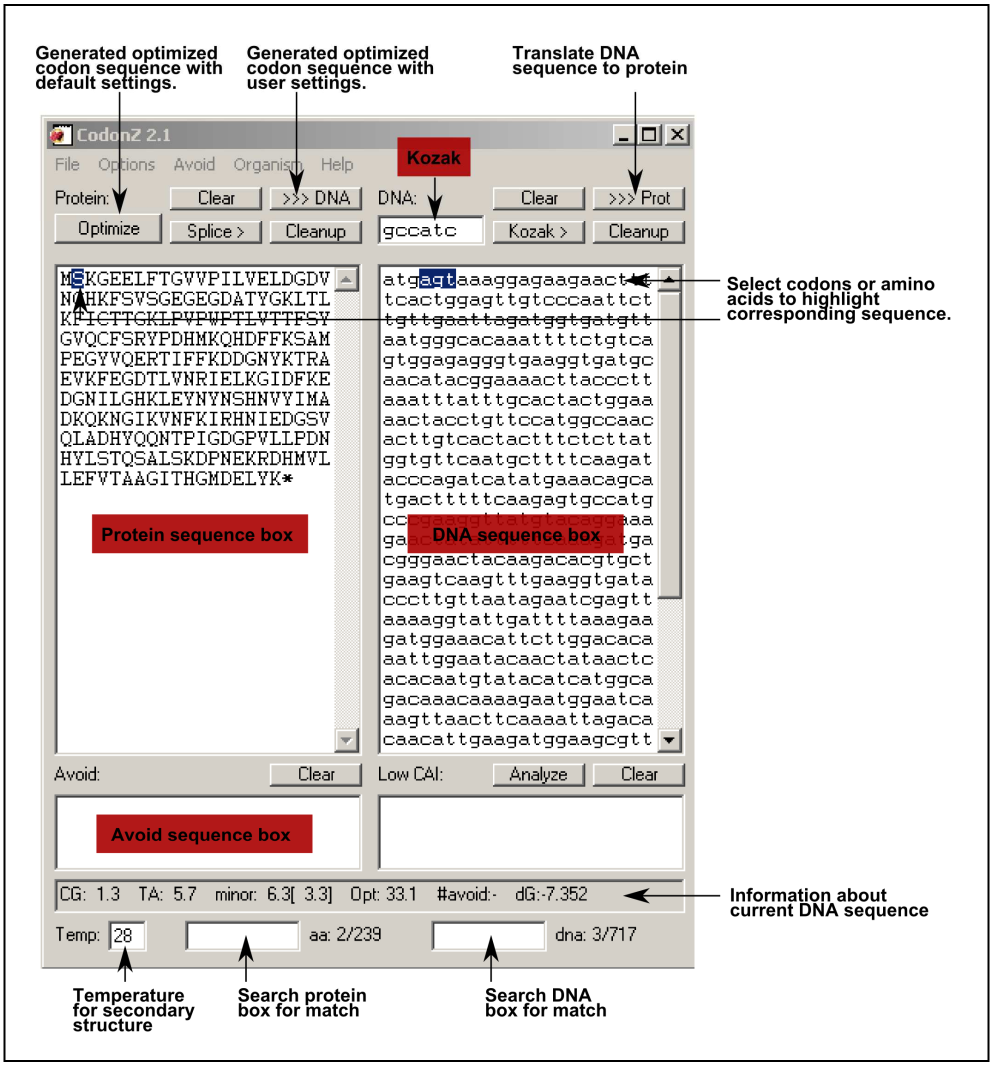
The Clear buttons above each box delete text in the corresponding box.

The Cleanup buttons remove any non-amino acid text from the left box, or any non-nucleotide text from the right box. This is useful if you are pasting sequences with white-space or numbers.

DNA sequence can be translated to protein with the >>> Prot button. The >>> DNA button generates a nucleotide sequence for the protein, using the user specified parameters.

The status bar below the Avoid and Low CAI boxes contains information about the current nucleotide sequence, in order:

- The percentage of CG dinucleotides
- The percentage of TA dinucleotides
- The percentage of minor [and rare] codons
- The percentage of optimal codons
- The number of sequences matching an entry in the Avoid box.
- The free energy of the minimum energy secondary structure for the specified Kozak and nucleotide sequence.



Generated optimized codon sequence with default settings.

Generated optimized codon sequence with user settings.

Translate DNA sequence to protein

Select codons or amino acids to highlight corresponding sequence.

Information about current DNA sequence

Temperature for secondary structure

Search protein box for match

Search DNA box for match

Kozak: gccatc

Temp: 28

```
MSKGEELFTGVVPILVELDGDV
NGHKFSVSGEGEGDATYGKLTL
KFICTTGKLPVPWPTLVTTFSY
GVQCFSRYPDHMKQHDFFKSAM
PEGYVQERTIFFKDDGNYKTRA
EVKFEGDTLVNRIELKGIDFKE
DGNILGHKLEYNYNSHNVYIMA
DKQKNGIKVNFKIRHNIEDGSV
QLADHYQQNTPIGDGPVLLPDN
HYLSTQSALSKDPNEKRDHMVL
LEFVTAAGITHGMDELYK*
```

```
atgagtaaaggagaagaactt
tcactggagttgtcccaattct
tgttgaattagatggtgatgtt
aatgggcacaaattttctgtca
gtggagagggtgaaggtgatgc
aacatacggaaaacttaccctt
aaatttatttgcactactggaa
aactacctgttccatggccaac
acttgtcactactttctcttat
ggtgttcaatgcttttcaagat
acccagatcatatgaaacagca
tgactttttcaagagtgccatg
ccgaaggttatgtacaggaaa
ga[illegible]tga
cgggaactacaagacacgtgct
gaagtcaagtttgaaggtgata
cccttgttaatagaatcgagtt
aaaaggtattgattttaaagaa
gatggaaacattcttggacaca
aattggaatacaactataactc
acacaatgtatacatcatggca
gacaaacaaaagaatggaatca
aagttaacttcaaaattagaca
caacattgaagatggaagcgtt
```

```
CG: 1.3   TA: 5.7   minor: 6.3[ 3.3]   Opt: 33.1   #avoid:-   dG:-7.352
```

Figure 1. Main user interface of CodonZ.

3. Options

Several options allow user control of codon choice for nucleotide sequence. These options are accessed via the dropdown menus at the top of the screen.

3. 1. Avoid CG/TA

When checked, CodonZ will attempt to avoid CG or TA dinucleotide sequences.

3. 2. Avoid swaps

When checked, CodonZ will maximize repeated use of same codon for a given amino acid. The first time an amino acid is specified, the optimal codon is chosen. If, in avoiding CG dinucleotides, a different codon (codon B) must be chosen, then the next time that amino acid appears, codon B will be selected rather than the optimal codon. This is a simple way to maximize tRNA cycling.

By default this option is not used for codon optimization.

3. 3. Avoid minor codons

Checking this options will force the algorithm to avoid codons which are used at less than 1% of total codon usage. Minor codons include the smaller set of 'rare' codons which are those which we operationally define as those which our analyses show are significantly avoided by highly expressed genes.

3. 5. Max folding

This option will adjust the secondary structure of the mRNA around the ATG to prevent tight folding. For this to work, you must have installed either the ViennaRNA or UNAfold package. If both are installed, select one using the Options menu. Results will not be identical but should be similar.

CodonZ performs a semi-random search to find the nucleotide sequence with the least secondary structure. Because the search is semi-random, slightly different solutions are likely to be found with repeated runs. The process takes a couple of minutes and during this time, a progress indicator replaces the DNA sequence information bar and should advance.

When you press Optimize, CodonZ goes through the following procedure.

1. Generate nucleotide sequence for the entire protein using options to avoid CG/TA dinucleotides, minor codons, restriction enzyme sites, splice sites and degradation sites. Save an internal copy of the resulting DNA sequence.

2. Keep the first 13 amino acids and delete all amino acids after those.

3. Clear all RE sites and turn off the avoid CG/TA and minor codons options.

4. Turn on the option to maximize folding energy (ie to find the minimum energy structure with the greatest free energy) and search for the best solution.

5. Replace the first 13 codons in the saved sequence with those maximizing folding energy.

3. 4. Kozak

The Kozak button simply cycles through different frequently used Kozak sequences in the Kozak box to allow you to quickly evaluate the effect of altering the Kozak sequence on the mRNA folding energy. A slightly different set of Kozak-like sequences are used by fish and mouse and the Kozak button will therefore cycle through different entries depending on the species selected.

For zebrafish, although gccatc is the most frequently used Kozak-like sequence for genes that are highly expressed, gccacc is sometimes a good choice. This sequence is still similar to the zebrafish consensus. If the start codon is followed by a codon starting with a 'G', then the sequence will be gccaccATGG, containing an NcoI site which can facilitate cloning.

3. 5. Sequences to avoid

After optimizing the sequence, CodonZ will try to remove any entry that appears in the Avoid box.

In removing these sequencing, the algorithm will avoid using minor and rare codons but introduce CG or TA dinucleotides if needed.

The most common sequences to avoid are:
- restriction sites that complicate cloning
- splice sites that may give rise to spurious transcripts

- AUUUA motifs that may destabilize RNA.

Note that the algorithm used for this process may not find an sequence that avoids all entries in the Avoid box, as in some cases, altering one codon to avoid an entry introduces a match to another entry.

3. 6. Splice

Most introns occur inside exons, rather than in untranslated regions. The Splice button assists in designing coding sequence that can accommodate the insertion of an intron that will be seamlessly excised without changing the coding sequence. Because a sequence matching the splice donor consensus is CAG.gtaagt and an good splice acceptor is cag.G, introns can be cloned into PstI (CTGCAG) sites which are in a CTGCAGG sequence

Example, target sequence is `##CTGCAGG##`

Cut with PstI gives `##CTGCA` and `GG##`
Add fragment with PstI overhands: `Ggtaagtnnnnn...nnnnnctgca`

Ligation: `##CTGCAGgtaagtnnnnn...nnnnnctgcaGG##`
Exons are underlined, so after splicing you get `##CTGCAGG##`

Pressing the Splice will highlight amino acid sequences where you can alter the codon usage to create a CTGCAGG sequence. Edit the DNA sequence manually to create these sequences.

3. 8. Low CAI

This feature is under development. Clusters of rare codons can be needed to allow correct protein folding. You can identify regions of low codon adaptation using the 'Analyze' button and copy regions of low adaptation into the box below the DNA sequence area. During codon optimization, these sequence areas will be populated by minor codons.

4. Selecting a species specific codon use

4. 1. For zebrafish or mouse

To specify the codon usage database for analyzing DNA in the DNA window, or that will be used when translating from protein sequence, select zebrafish or mouse from the Organism menu. For these species, the codon usage is determined by our analysis of highly expressed genes. This analysis also includes which codons are designated as 'minor and rare' codons.

Because the stop codon is recognized by release factors that in many organisms have a specific 4 nucleotide preference, the DNA sequence will contain an extra base (in lowercase). Thus when optimizing using zebrafish and mouse codon frequencies, where the preference is TAAA, the DNA sequence will appear as 'TAAa'.

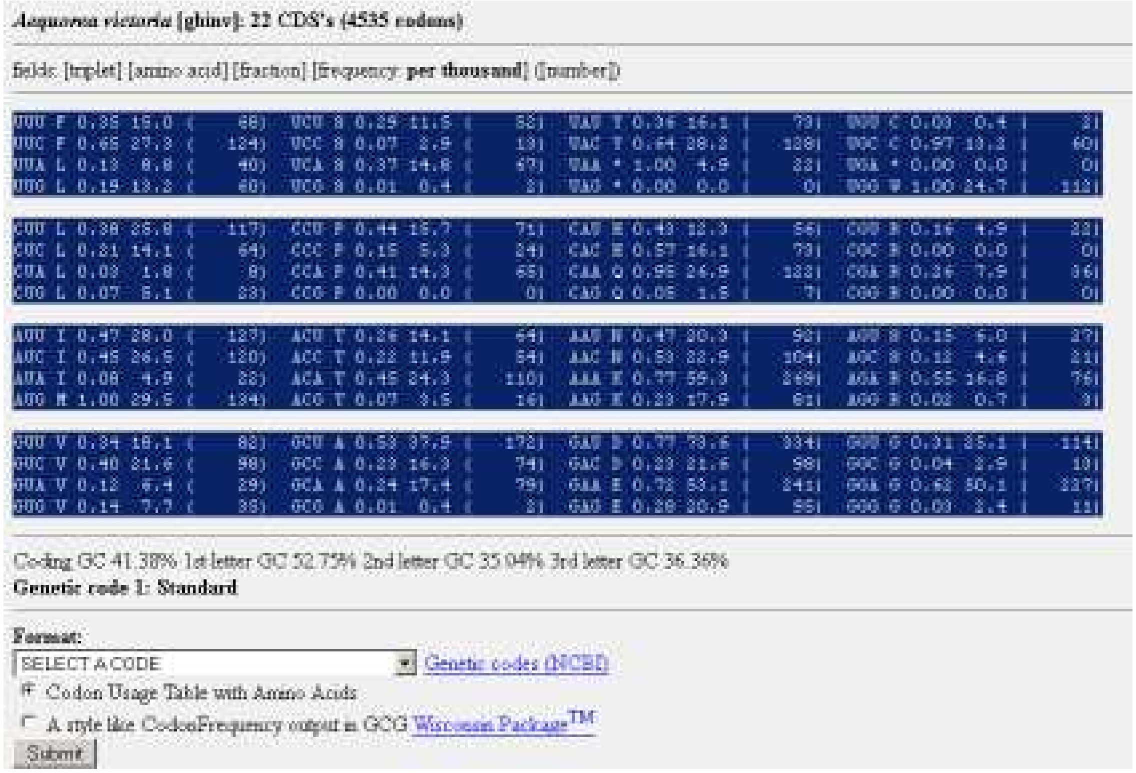
4. 2. For other organisms

For other organisms you can specify the codon usage table using these steps:

1. Find the codon usage for the organism at http://www.kazusa.or.jp/codon/

2. Under 'Format' on the webpage, select 'Standard' and 'Codon Usage Table with Amino Acids' then press Submit

3. Select the entire table from UUU to the final bottom right bracket (see image below) then copy to the clipboard using ctrl-C.



Aequorea victoria [gbinv]: 22 CDS's (4535 codons)

fields: [triplet] [amino acid] [fraction] [frequency: **per thousand**] ([number])

```
UUU F 0.35 15.0 (    68)  UCU S 0.29 11.5 (    52)  UAU Y 0.36 16.1 (    73)  UGU C 0.03  0.4 (     2)
UUC F 0.65 27.3 (   124)  UCC S 0.07  2.9 (    13)  UAC Y 0.64 28.2 (   128)  UGC C 0.97 13.2 (    60)
UUA L 0.13  8.8 (    40)  UCA S 0.37 14.8 (    67)  UAA * 1.00  4.9 (    22)  UGA * 0.00  0.0 (     0)
UUG L 0.19 13.2 (    60)  UCG S 0.01  0.4 (     2)  UAG * 0.00  0.0 (     0)  UGG W 1.00 24.7 (   112)

CUU L 0.38 25.8 (   117)  CCU P 0.44 15.7 (    71)  CAU H 0.43 12.3 (    56)  CGU R 0.16  4.9 (    22)
CUC L 0.21 14.1 (    64)  CCC P 0.15  5.3 (    24)  CAC H 0.57 16.1 (    73)  CGC R 0.00  0.0 (     0)
CUA L 0.03  1.8 (     8)  CCA P 0.41 14.3 (    65)  CAA Q 0.95 26.9 (   122)  CGA R 0.26  7.9 (    36)
CUG L 0.07  5.1 (    23)  CCG P 0.00  0.0 (     0)  CAG Q 0.05  1.5 (     7)  CGG R 0.00  0.0 (     0)

AUU I 0.47 28.0 (   127)  ACU T 0.26 14.1 (    64)  AAU N 0.47 20.3 (    92)  AGU S 0.15  6.0 (    27)
AUC I 0.45 26.5 (   120)  ACC T 0.22 11.9 (    54)  AAC N 0.53 22.9 (   104)  AGC S 0.12  4.6 (    21)
AUA I 0.08  4.9 (    22)  ACA T 0.45 24.3 (   110)  AAA K 0.77 59.3 (   269)  AGA R 0.55 16.8 (    76)
AUG M 1.00 29.5 (   134)  ACG T 0.07  3.5 (    16)  AAG K 0.23 17.9 (    81)  AGG R 0.02  0.7 (     3)

GUU V 0.34 18.1 (    82)  GCU A 0.53 37.9 (   172)  GAU D 0.77 73.6 (   334)  GGU G 0.31 25.1 (   114)
GUC V 0.40 21.6 (    98)  GCC A 0.23 16.3 (    74)  GAC D 0.23 21.6 (    98)  GGC G 0.04  2.9 (    13)
GUA V 0.12  6.4 (    29)  GCA A 0.24 17.4 (    79)  GAA E 0.72 53.1 (   241)  GGA G 0.62 50.1 (   227)
GUG V 0.14  7.7 (    35)  GCG A 0.01  0.4 (     2)  GAG E 0.28 20.9 (    95)  GGG G 0.03  2.4 (    11)
```

Coding GC 41.38% 1st letter GC 52.75% 2nd letter GC 35.04% 3rd letter GC 36.36%

Genetic code 1: Standard

Format:

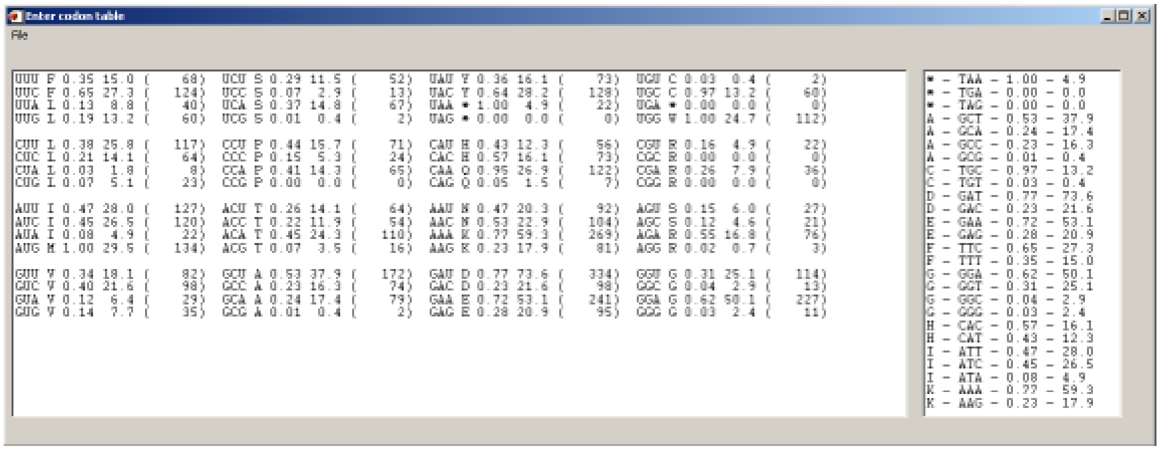
SELECT A CODE

● Codon Usage Table with Amino Acids

○ A style like CodonFrequency output in GCG Wisconsin Package™

4. In CodonZ, select Organism → Enter New

5. Paste the table into the left hand box. If all goes well, you'll see the table reproduced in the right hand box:



```
UUU F 0.35 15.0 (    68)  UCU S 0.29 11.5 (    52)  UAU Y 0.36 16.1 (    73)  UGU C 0.03  0.4 (     2)
UUC F 0.65 27.3 (   124)  UCC S 0.07  2.9 (    13)  UAC Y 0.64 28.2 (   128)  UGC C 0.97 13.2 (    60)
UUA L 0.13  8.8 (    40)  UCA S 0.37 14.8 (    67)  UAA * 1.00  4.9 (    22)  UGA * 0.00  0.0 (     0)
UUG L 0.19 13.2 (    60)  UCG S 0.01  0.4 (     2)  UAG * 0.00  0.0 (     0)  UGG W 1.00 24.7 (   112)

CUU L 0.38 25.8 (   117)  CCU P 0.44 15.7 (    71)  CAU H 0.43 12.3 (    56)  CGU R 0.16  4.9 (    22)
CUC L 0.21 14.1 (    64)  CCC P 0.15  5.3 (    24)  CAC H 0.57 16.1 (    73)  CGC R 0.00  0.0 (     0)
CUA L 0.03  1.8 (     8)  CCA P 0.41 14.3 (    65)  CAA Q 0.95 26.9 (   122)  CGA R 0.26  7.9 (    36)
CUG L 0.07  5.1 (    23)  CCG P 0.00  0.0 (     0)  CAG Q 0.05  1.5 (     7)  CGG R 0.00  0.0 (     0)

AUU I 0.47 28.0 (   127)  ACU T 0.26 14.1 (    64)  AAU N 0.47 20.3 (    92)  AGU S 0.15  6.0 (    27)
AUC I 0.45 26.5 (   120)  ACC T 0.22 11.9 (    54)  AAC N 0.53 22.9 (   104)  AGC S 0.12  4.6 (    21)
AUA I 0.08  4.9 (    22)  ACA T 0.45 24.3 (   110)  AAA K 0.77 59.3 (   269)  AGA R 0.55 16.8 (    76)
AUG M 1.00 29.5 (   134)  ACG T 0.07  3.5 (    16)  AAG K 0.23 17.9 (    81)  AGG R 0.02  0.7 (     3)

GUU V 0.34 18.1 (    82)  GCU A 0.53 37.9 (   172)  GAU D 0.77 73.6 (   334)  GGU G 0.31 25.1 (   114)
GUC V 0.40 21.6 (    98)  GCC A 0.23 16.3 (    74)  GAC D 0.23 21.6 (    98)  GGC G 0.04  2.9 (    13)
GUA V 0.12  6.4 (    29)  GCA A 0.24 17.4 (    79)  GAA E 0.72 53.1 (   241)  GGA G 0.62 50.1 (   227)
GUG V 0.14  7.7 (    35)  GCG A 0.01  0.4 (     2)  GAG E 0.28 20.9 (    95)  GGG G 0.03  2.4 (    11)
```

```
* - TAA - 1.00 - 4.9
* - TGA - 0.00 - 0.0
* - TAG - 0.00 - 0.0
A - GCT - 0.53 - 37.9
A - GCA - 0.24 - 17.4
A - GCC - 0.23 - 16.3
A - GCG - 0.01 - 0.4
C - TGC - 0.97 - 13.2
C - TGT - 0.03 - 0.4
D - GAT - 0.77 - 73.6
D - GAC - 0.23 - 21.6
E - GAA - 0.72 - 53.1
E - GAG - 0.28 - 20.9
F - TTC - 0.65 - 27.3
F - TTT - 0.35 - 15.0
G - GGA - 0.62 - 50.1
G - GGT - 0.31 - 25.1
G - GGC - 0.04 - 2.9
G - GGG - 0.03 - 2.4
H - CAC - 0.57 - 16.1
H - CAT - 0.43 - 12.3
I - ATT - 0.47 - 28.0
I - ATC - 0.45 - 26.5
I - ATA - 0.08 - 4.9
K - AAA - 0.77 - 59.3
K - AAG - 0.23 - 17.9
```

6. Select File → Quit. Check that the correct codon usage table is loaded under Organism → Display.

5. Editing and proofing sequences

5. 1. Editing sequences

Gene synthesis companies can not generate every sequence. Manual editing of the sequence may be required to re-engineer regions flagged as having too much repetitive sequence or secondary structure.

To assist with this, use the Organism → Display function to open a window showing the codon frequency table currently in use. This table shows the amino acid, codon and relative synonymous codon usage frequency (RSCU). Codons where the RSCU is followed by a ! symbol are minor codons and should be avoided.

Find the nucleotide sequence that needs modifying using the nucleotide search box at the bottom of the window. This will highlight the corresponding amino acids in the protein window.

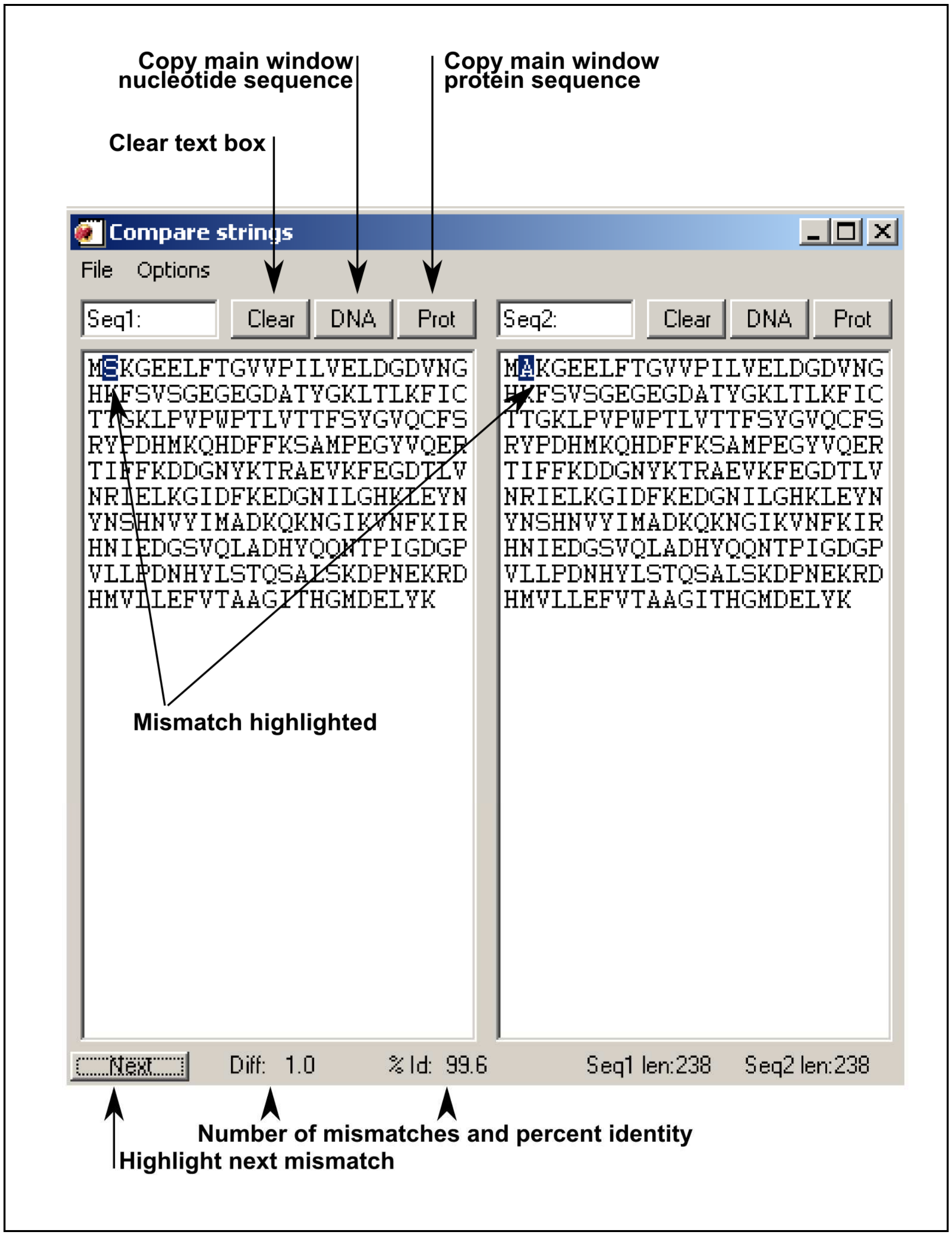
To be sure that you are editing in the correct frame, select a single amino acid in the protein window. This now highlights the corresponding codon in the right window. Choose a new codon from the window showing the codon frequency table.

Populate the Avoid box with entries from the Avoid drop down menu to ensure that you did not inadvertently create a splice sites or degradation sequence.

Finally, double-check that the new nucleotide sequence encodes the correct protein. The tool in the next section can assists with this.

5. 2. Checking sequences

To check the resulting sequences, open the File->Compare utility window. This allows users to copy nucleotide or protein sequences from the main window, and compare two sequences for mismatches. Press the 'DNA' or 'Prot' button above each of the text boxes to copy the sequence from the main window to the corresponding box.



```
MSKGEELFTGVVPILVELDGDVNG
HKFSVSGEGEGDATYGKLTLKFIC
TTGKLPVPWPTLVTTFSYGVQCFS
RYPDHMKQHDFFKSAMPEGYVQER
TIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYN
YNSHNVYIMADKQKNGIKVNFKIR
HNIEDGSVQLADHYQQNTPIGDGP
VLLPDNHYLSTQSALSKDPNEKRD
HMVLLEFVTAAGITHGMDELYK
```

```
MAKGEELFTGVVPILVELDGDVNG
HKFSVSGEGEGDATYGKLTLKFIC
TTGKLPVPWPTLVTTFSYGVQCFS
RYPDHMKQHDFFKSAMPEGYVQER
TIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYN
YNSHNVYIMADKQKNGIKVNFKIR
HNIEDGSVQLADHYQQNTPIGDGP
VLLPDNHYLSTQSALSKDPNEKRD
HMVLLEFVTAAGITHGMDELYK
```

```
Diff:   1.0      % Id:  99.6      Seq1 len:238    Seq2 len:238
```