INCA 2.0 User Manual

Welcome

What is INCA?

INCA can be used to analyze synonymous codon usage in a (usually bacterial) genome. It can calculate a number of codon usage indices, and has many different options for graphical display of these values. Finally, it also features a kind of neural network, called the self-organizing map (SOM), which can cluster genes by codon usage.

What are the requirements to run it?

The software runs on all 32 bit versions of Windows. While disk space and memory requirements are relatively low, a powerful processor (Pentium 4 or Athlon) is recommended.

How to obtain INCA?

Through the website at http://www.bioinfo-hr.org/inca or contact the author directly at fsupek@public.srce.hr. The software is free for academic (non-commercial) use; otherwise contact the author to obtain a license.

Ideas and bug reports

Any suggestions and bug reports are certainly welcome. If some minor changes in the software are needed to better suit your work, this can be arranged.

Thanks

I would like to thank my colleagues from the Bioinformatics Group: Maja, Morana and Neno for their unending support, suggestions and constructive criticism; the *el jefe* Kristian Vlahoviček, for being such a kind boss and bearing with me while INCA was being developed. Many thanks go to Gordana Maravić for supplying me with all the literature I needed – or even a bit more - and finally to John Novembre for devoting his time to help me test INCA and (hopefully) make it a more useful piece of software.

Disclaimer

Please cite

Supek F, Vlahovicek K; INCA: synonymous codon usage analysis and clustering by means of self-organizing map; Bioinformatics, 2004

Version history

1.00 **– 21. September 2003**

- first publicly available version

1.02 **– 16. October 2003**

- fixed bug that could cause CAI to be calculated incorrectly
- added the Find feature in the Table tab
- slight speed up when loading genomes

1.10 **– 29. February 2004**

- added support for several other codon usage measures: ENC prime (N_c' or ENC') and MCB
- reorganized the options in the *xy* Plot and Table tabs (axes, columns…) – now there is a separate "relative to" combo box, which means any measure can be computed relative to any of the available gene groups
- added a random nucleotide sequence generator
- added in indicator of quality of SOM clustering ("map resolution")
- changed definition of "G+C content at silent sites" to mean "G and C frequency at 3rd sites where all point mutations are silent" instead of "where transitions are silent" (in this version a more general definition is used, allowing the background frequency of each nucleotide to be determined separately, instead of only being able to calculate G+C vs. A+T)
- fixed bug that could cause errors when using the Table view to view a short genome if a longer one has been opened previously
- fixed bug that caused the Table view to be redrawn twice unnecessarily
- fixed bug: now "txt" is the default extension for exported tables

1.12 **– 12. April 2004**

- added separate "Find first" and "Find next" buttons to the Table view
- fixed bug that could cause INCA to stop responding when generating a small number of sequences
- slight changes in the way ENC' is calculated – the program now mimics John Novembre's script; for details refer to the "Measures of codon usage" section of the Manual
- loading of User codon frequency tables is now more flexible; tables can be taken directly from the Codon Usage Database at http://www.kazusa.or.jp/codon/
- all measures of codon usage can also be calculated against the User freq table
- now all tables: table view export, gene set import/export and codon frequencies import/export use "txt" as default extension
- correction of reference: "B (Karlin et al 1996)" changed to "1998"; this does not in any way affect computation of the "B" statistic

1.12a **– 28. June 2004**

- fixed gene length treshold bug. In prior versions, on loading files the gene length information contained in the *ptt* file was consulted, instead of the actual length of the sequence (this is the Length displayed in the Table and Plot views). The *ptt* value is normally one codon shorter (it doesn't include the final stop codon); this is why in INCA 1.12 setting the treshold to „Disregard genes with less than 100 codons" and opening a file would actually only allow genes 101 codons or longer (100+stop). As of version 1.12a the stop codons count towards the gene length *both* when loading files (ie genes exactly 100 codons long are accepted) *and* in the Table/Plot views.

***1.20* – 13. August 2004**

- new feature: the Cluster/COG tab visually presents connections between clusters and COG functional categories
- new feature: the Optimizer tab improves codon usage of a sequence so as to improve expression of heterologous genes
- fixed bug: INCA got confused when loading FASTA files with lowercase letters. This is now fixed and the program doesn't care whether upper- or lowercase letters are used. However, codons containing letters other than u, g, c, a or t are still ignored.
- the table view now always displays a column with the gene description
- fixed bug: when exporting frequency tables of clusters, if Cluster *n* was chosen, the frequencies of *n+1* were exported; also, exporting codon frequencies from background nucleotide composition was problematic. This now works correctly.

***1.20a* – 19. September 2004**

- fixed issues with the Optimizer:
  - when pasting a sequence with a length in codons not divisible by 3, INCA now truncates one or two nucleotides from the end of the sequence to remove the partial codon; this prevents an "Access violation" from happening when the Optimize button is clicked.
  - if the sequence contained IUB ambiguity codes for nucleotides, ie. AUGC**N**G, INCA would simply skip such characters, which could cause 'frameshifting' downstream of the ambiguity. These are now handled correctly; codons containing ambiguity codes are colored bright green in the Optimizer.

***2.0 beta* – 30. November 2004**

- in this major release of INCA, a large part of the of the code has been rewritten to support numerous new features, including, but not limited to
  - ability to load/unload multiple files (ncbi, kegg, cutg, fasta files)
  - save and load 'projects', import numerical data and codon frequencies
  - create user-defined gene groups, descriptive stats & correlation for groups
  - 3D scatterplots, coloring by any criterion, graphical select & filtering
  - improved SOM, based on the MILC statistic, more visualization criteria
  - principal component analysis (PCA) in plots, tables and SOM
  - a more comprehensive nucleotide sequence generator
  - numerous user interface improvements; currently for Win32, soon for Linux

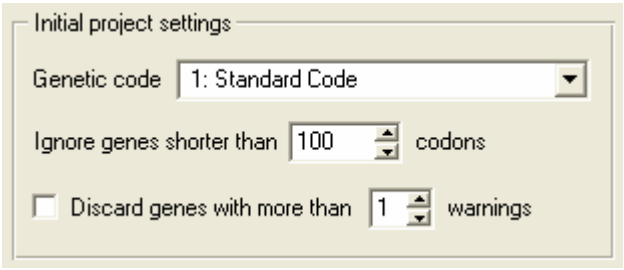
***2.0 final* – 20. August 2005**

- a number of bugs from the beta have been fixed
  - when using PCA, the contribution of each component to the total variance of the data was misestimated; now the correct numbers are shown
  - "Genomes" are now called "Genefiles" to avoid confusion
  - even if only a single gene group was selected for analysis using the SOM, the net was still trained using all loaded genes; this is now fixed
  - fixed issues with positioning of secondary windows: Scatterplot 3d options, SOM init options etc.)
  - fixed issue with occasional wrong scaling of the *x* axis in the Groups/Bins
  - fixed appearance of the Frequency selection window
  - fixed issue with ENC being computed as NaN when Met or Trp were rare (>=1); now ENC can be computed even for genes without those amino acids
  - fixed issue with INCA rejecting lines in FASTA files beginning with an ambiguity code (last two fixes - thanks to Anna Palmé)
  - fixed ordering of Frequency Charts on the corresponding Tab
  - added an option to export codon and amino acid frequencies alongside data from the Gene Browser
  - in the SOM, "Similarity to neighbors" is now more correctly named "Difference from neighbors"

- a Linux version has been made available
    - it might be less stable and behave in unpredictable ways, especially when using the Scatterplot
    - the Web lookup and the link to bioinfo-hr.org will probably not work
    - possible issues with font sizes, depending on the Linux distribution
    - INCA 2 for Linux is a 'permanent beta' version, meaning I will probably not be able to fix most of the stability issues; the Win32 version should be much more reliable

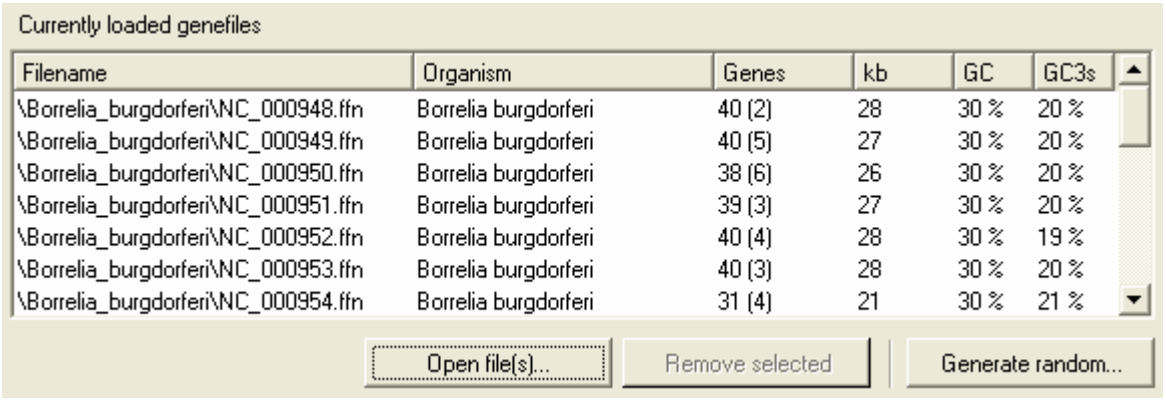
The "Project data" tab

Opening sequence files



Before you start a new project – by opening files containing nucleotide sequences – you should adjust the Initial Project Settings. These include: the genetic code for the organism you're examining (the translation tables were adopted from the NCBI website), the gene length threshold, and the option to discard genes that generated a certain warning count. Warnings are issued if: (i) a gene does not end in a valid stop codon and (ii) the warning count is increased by 1 for each internal stop codon. These options cannot be altered once you've loaded files, or at least until you unload the files, either manually, or by starting a new project.

Use the "Open file(s)…" button to select files containing the sequences to analyze. The dialog that pops up supports selection of multiple files (click and drag, or use shift+arrow keys). You can also load more sequence files later, or remove the ones you don't need.



The "Genes" column first lists the number of sequences ('genes') that conform to the specified criteria regarding gene length and warning count, while the number in parenthesis is the number of genes that were discarded. "kb" is the total length of all sequences in kb, "GC" is the total G+C %, while "GC3s" takes only 3rd sites of fourfold degenerate amino acids into account when computing the G+C %.

The files you load are expected to contain only sequences of coding regions of genes, i.e. hypothetical cDNAs. INCA supports several file formats, for instance the commonly used FASTA format. Such files may contain many nucleotide sequences separated with comment lines (beginning with '>'); they may look like this:

```
>BBP07, gi|11497060:4936-5397
GTGAAAATGAGTGAACAAGAAAGCTTACAAGCACAAGTTGCAGGAGAAGAAGAACTTTTAGTAACAAAAC
TCCAGAATTTT...
```

or

```
>(gi|16127994:1080570-1080686, 1080677-1081408)
ATGGACAAACTAGAATTTAAAATGGAATTGGAAATTGGGTGGTTTGGTGGTCGTGCAGGTATTGCTAAAA
TGCATGAAAAAGGGAGT...
```

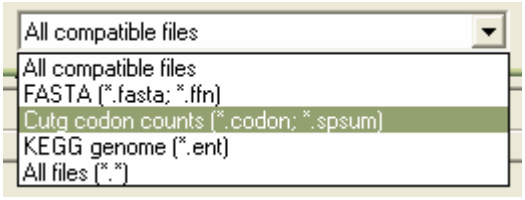
When parsing the comment lines, numbers after the colon (if present) are examined and the gene's starting and ending positions determined. If the line contains a comma, the part before it is the gene's name, and the part after its description; if there is no comma, all of the text is assigned to both properties. During analysis of the sequences, INCA ignores codons containing unspecified nucleotides, such as R, Y, or N.

The FASTA file should have in the ".ffn" or ".fasta" extensions; after loading it, INCA also automatically looks for a file with the same filename, but with a ".ptt" extension in the same directory. It contains additional information about the genes and their protein products, e.g. which KOG category they belong to, and their exact name and short description. While not necessary, it's highly recommended to have this file loaded. The file's lines look something like this:

```
15215..15418  +  67  11497073  blyA  BBP23  -  -  pore-forming hemolysin
15422..15769  +  115 11497074  blyB  BBP24  -  -  hemolysin accessory protein
```

The genomes found on the NCBI website (or ftp server) contain such "ffn" and "ptt" file pairs which can be used in INCA without modification.

Another supported format is the CUTG database (http://www.kazusa.or.jp/codon/ or ftp://ftp.kazusa.or.jp/pub/codon/current/) codon count file. The format is similar to FASTA, except the sequences are replaced by codon counts. INCA doesn't care if only counts are available instead of whole sequences, since codon usage is examined at gene level only.

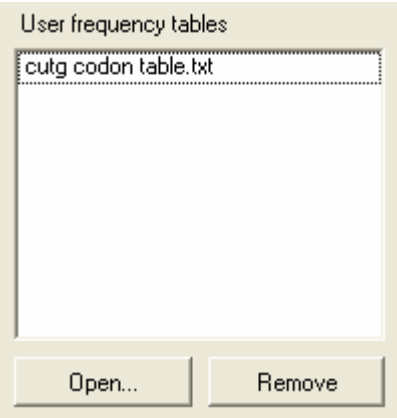


Finally, INCA supports the KEGG (http://www.genome.jp/kegg/kegg2.html or ftp://ftp.genome.jp/pub/kegg/genomes/genes) database format for genomes, having an ".ent" extension. A field in such a file looks like this:

```
ENTRY       BG11037          CDS       B.subtilis
NAME        aadK
DEFINITION  aminoglycoside 6-adenylyltransferase [EC:2.7.7.-]
POSITION    complement(2734911..2735762)
DBLINKS     BSORF: BG11037
            SubtiList: BG11037
            NCBI-GI: 2635124
            UniProt: P17585
CODON_USAGE       T               C               A               G
          T  15   3   6   7   5   0   7   3  14   5   0   0   1   3   0   7
          C   7   2   2   0   5   1   3   1   2   0   3   4   3   2   2   1
          A  10   7   1  13   6   2   2   1  11   7  10  10   2   0   5   2
          G   2   1   8   2   3   2   6   0  16   5  18   6   5   1   1   5
AASEQ       284
            MRSEQEMMDIFLDFALNDERIRLVTLEGSRTNRNIPPDNFQDYDISYFVTDVESFKENDQ
            WLEIFG...
NTSEQ        852
            atgcgaagtgagcaggaaatgatggacatttttttggactttgctttgaacgatgagaga
            atccgattggtcactttggaagggtcacgtacaaacagaaatatccctcctgacaacttc
            caagattatgacatctcgtattttgtaactgatgtagaatcttttaaagaaaatgatcag
            cctcccgaa...
```

While the sequence files get parsed, codon usage for each gene is also calculated; that's why the process normally takes a few moments. For example, a 1.7 GHz Pentium 4 would load the *E. coli* genome in 6 seconds.

User-defined codon frequency tables



The "User frequency tables" option might be useful, for example, if you want to compare the codon usage of a gene to codon frequencies derived from intergenic sequences, or to a frequency table derived from nucleotide proportions. Before that you need to load the frequencies from a text file.

When parsing such files, INCA skips empty lines and lines beginning with >. Each line should contain a codon and a number – its frequency, and they should be separated by at least a space, tab or comma. It is allowed to have other text on the same line, as well as more than one codon/frequency combination. For example:

```
GCU A Ala  0.2457 0.5342 0.2500
```

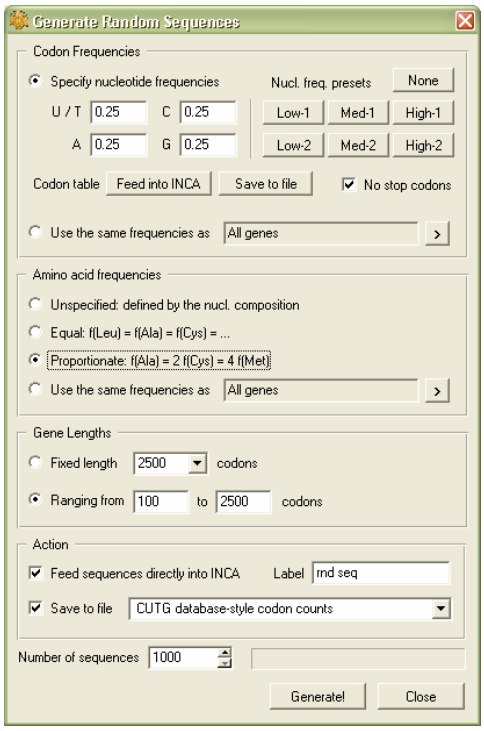
This is a valid line which assigns the frequency 0.2457 to the codon GCU (it is also an example of the result of INCA's "Export frequencies…" feature, found on the Frequencies tab). When importing frequency tables, the frequencies need not add up to 1 – INCA will renormalize them as necessary. However, the program doesn't perform any other checks on the file; for example, it is possible to load a file that uses a different genetic code than the current one, or a file with some missing items. However, doing so will result in unpredictable behavior later.

The Codon Usage Database at http://www.kazusa.or.jp/codon/ is a good source of codon usage tables for any organism with a decent number of GenBank entries (sequenced genes). Such tables may be copied directly from the webpage, and pasted into Notepad or any other plain text editor, the file saved as .txt and loaded into INCA. An example follows:

```
UUU 26.1(165603)  UCU 23.4(148415)  UAU 18.8(119063)  UGU  8.0( 50774)
UUC 18.4(116350)  UCC 14.2( 90112)  UAC 14.8( 93512)  UGC  4.8( 30129)
UUA 26.2(165767)  UCA 18.7(118252)  UAA  1.0(  6577)  UGA  0.7(  4228)
UUG 27.1(171689)  UCG  8.6( 54410)  UAG  0.5(  3225)  UGG 10.4( 65773)

CUU 12.2( 77618)  CCU 13.5( 85873)  CAU 13.7( 86808)  CGU  6.5( 40983)
CUC  5.5( 34623)  CCC  6.8( 43168)  CAC  7.8( 49226)  CGC  2.6( 16476)
CUA 13.4( 84725)  CCA 18.2(115180)  CAA 27.3(173198)  CGA  3.0( 19096)
CUG 10.5( 66507)  CCG  5.3( 33455)  CAG 12.2( 77248)  CGG  1.8( 11129)

AUU 30.1(190778)  ACU 20.2(128142)  AAU 35.8(227047)  AGU 14.2( 89977)
AUC 17.1(108376)  ACC 12.6( 79919)  AAC 24.9(157543)  AGC  9.7( 61571)
AUA 17.8(112939)  ACA 17.7(112497)  AAA 42.0(266207)  AGA 21.3(134769)
AUG 20.9(132697)  ACG  8.0( 50540)  AAG 30.8(195515)  AGG  9.3( 58657)

GUU 22.0(139144)  GCU 21.1(133757)  GAU 37.7(238723)  GGU 23.9(151342)
GUC 11.7( 73874)  GCC 12.6( 79840)  GAC 20.3(128583)  GGC  9.8( 62203)
GUA 11.8( 74933)  GCA 16.2(102692)  GAA 45.7(289705)  GGA 10.9( 69211)
GUG 10.8( 68221)  GCG  6.2( 39200)  GAG 19.2(121817)  GGG  6.0( 38291)
```

The Random sequence generator can also be used to generate frequency tables; refer to the next section for more information.

Random sequence generator



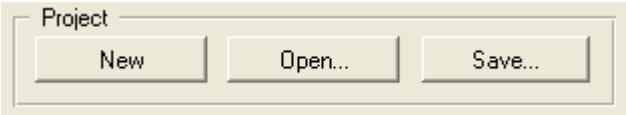
This feature of INCA is used to generate sequences of chosen lengths and nucleotide compositions. You can specify the frequency of each nucleotide separately (the values need not add up to 1) or use one of the presets. These will produce sequences with varying G+C content and skew; refer to Comeron & Aguade (1998) in *J Mol Evol* for more information about the presets. An option is also offered that causes the generated sequences to mimic the codon preferences of a specified gene group.

The codon table derived from the specified nucleotide frequencies can, if desired, be saved to a tab-delimited text file, or fed directly into INCA.

When the "Unspecified" option is chosen for the amino acid frequencies, they will vary to accommodate the nucleotide frequencies requested by the user. For example, if a high G+C content value is specified, the amino acids encoded by the G and C rich codon are more frequently used. The "Equal" and "Proportionate" options force the amino acid frequencies to stay fixed regardless of the nucleotide frequencies. In that case, the overall nucleotide composition of the genome will *not* match the specified ratios, because the use of specific nucleotides is in part dictated by the use of amino acids. However, nucleotide use at silent sites *will* reflect the specified values. The last option in this group constrains amino acid frequencies to be the same as in the specified gene group.

The generated sequences can be loaded directly into INCA - the effect is the same as if you opened a file containing the new sequences. They can also be saved to a FASTA file, either as DNA or RNA, or to a CUTG database codon count file.

Working with projects



A "Project" is a file, saved with the ".i2p" extension that contains the snapshot of your current work in INCA. A project file, once saved, may be opened later to continue exactly where you left off, or may be transferred to another computer. The i2p files are text files, and therefore quite large, but compress well using zip or gz.

The following things are saved to the project file:

- codon usage of all of the loaded genes
- arrangement of genes into groups (the reference set, clusters, bins…)
- loaded user frequency tables
- current state of the self-organizing map
- imported user data

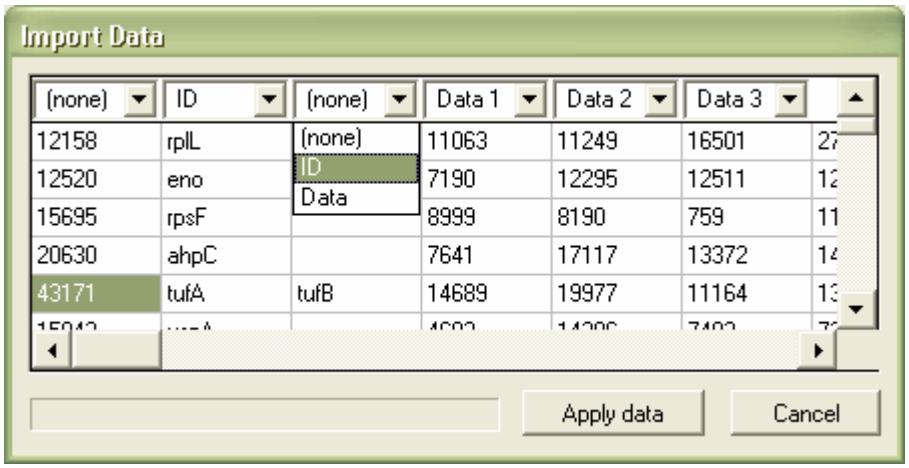
Importing user data

The "User data" feature on the "Project data" tab may be used to load and kind of numerical data connected to the genes you are currently examining. You might, for instance, want to import a file containing mRNA abundance data from a microarray experiment to see if it correlates with codon usage.

First, load all the files containing gene sequences you intend to analyze, let's say the *E. coli K12* genome. Then click the "Import data" button and choose a tab delimited text file, where each row describes a gene and columns contain one or more values. A sample fragment of such a file follows:

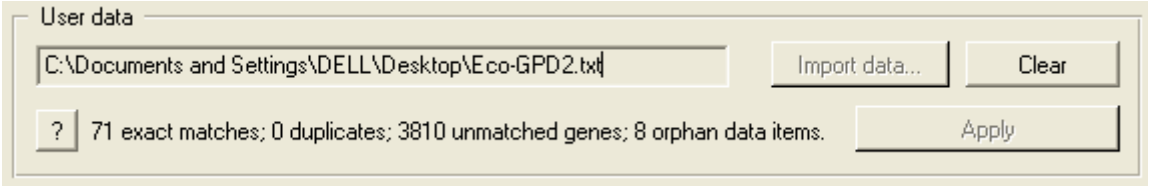
```
12158     rplL          11063  11249  16501  27910  10088  909.91 925.23 1357.18
12520     eno           7190   12295  12511  12152  7190   574.31 982.07 999.30
15695     rpsF          8999   8190   759    1155   8039   573.39 521.83 48.33
20630     ahpC          7641   17117  13372  14671  1605   370.40 829.69 648.20
43171     tufA   tufB   14689  19977  11164  13808  5582   340.25 462.74 258.59
15842     uspA          4682   14296  7402   7221   3235   295.56 902.38 467.23
```

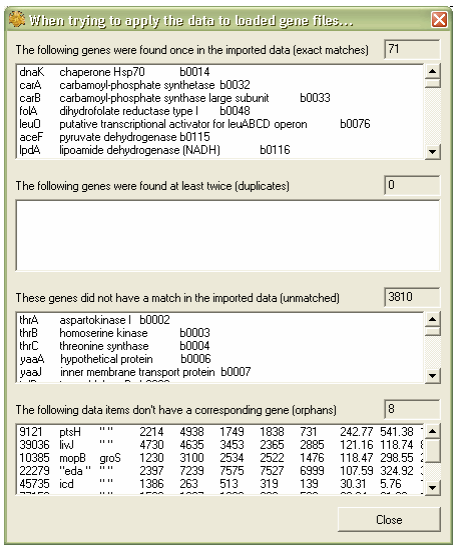
This file contains a series of protein abundance data for *E. coli* grown under different condition. Next, a dialog pops up where INCA asks you to tell it what the columns in the file contain. You can resize the dialog to see more of your data at once.



Choosing "(none)" will ignore the column altogether; "Data" denotes the column contains numerical data; "ID" means this column contains the gene identifier, such as the name of the gene. When matching ID values in this column to loaded genes, the genes' *Name* and *Synonym* fields are examined case-insensitively for an exact match; e.g. "eno" will match "ENO", but won't match "enolase". If you mark more than one column as "ID", all of them will be used in the search and a match in either "ID" column means a gene is 'found'. For instance, the fifth row in the above file would match the tufA and tufB genes, and assign to both the values 14689, 19977, 11164 and so on.

Be warned that you may have to wait a while after clicking the "Apply data" button, especially if you have lots of genes loaded, or the table you're importing is large. After the process is done, you'll see a short report in the bottom of the window:





Exact matches are the genes for which exactly one row in the imported data file has been found. If two or more rows matched a single gene, it's counted as a *duplicate*, and it will be associated with data from the last matching row in the table. *Unmatched* are the loaded genes that haven't been found in the imported file, and *orphans* are items in the imported file which haven't been found among the loaded genes. Clicking the "?" button will reveal what genes belong to what category.

The "Clear" button removes from memory the imported file, and all the numerical data that has been associated with genes. The "Apply" button is used when you load a gene file *after* importing user data, to scan the already imported data for possible matches among the newly loaded genes.

The imported data can be easily added as a column to the Gene Browser, or used for visualization in the Scatterplot; refer to the Gene groups and properties section for more information.

Gene groups and properties

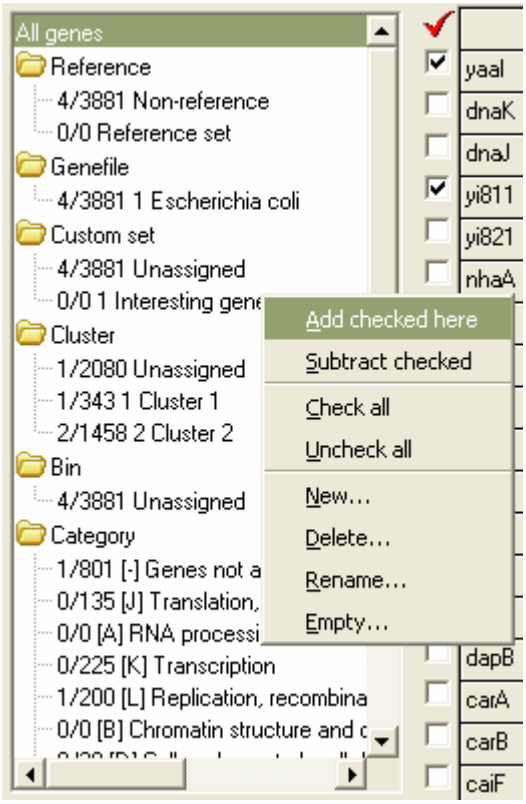
The information in this section is not specific for a single part of INCA; you will need to, for instance, choose a gene group to examine in the Gene Browser, one to visualize in the Scatterplot, or a group to analyze using the Self-organizing map. There are several types of gene groups:

- *All genes* – a single group containing all loaded genes
- *Reference* – the user must manually add genes to the second one to create the reference set, normally consisting of highly expressed genes. Always contains two groups:
  - the *Non-reference* group
  - the *Reference set* group
- *Genefile* – each FASTA (or CUTG or KEGG) file you loaded is displayed as a single gene group.
- *Custom set* – contains an arbitrary number of groups, created and deleted by the user at will; the genes may also freely be added or removed to any such group.
- *Cluster* – generated automatically by the Self-organizing map.
- *Bin* – generated automatically by the Binning feature.
- *Category* – the KOG functional category, determined at the time of loading from the ".ptt" file if it exists.

Each loaded gene must belong to at least one group of a type i.e. it must belong to: either the reference set or the non-reference set; at least one genefile and so on. If a gene hasn't been assigned to a bin, a cluster or to a custom set, it can be found in the Unassigned bin/cluster/custom set.

Note that the only two types that allow the genes to be assigned to a group manually are the *reference* and the *custom set* group types. The latter also requires the user to manually create groups, as they are not created automatically, like the *clusters*, *bins* or *categories*.

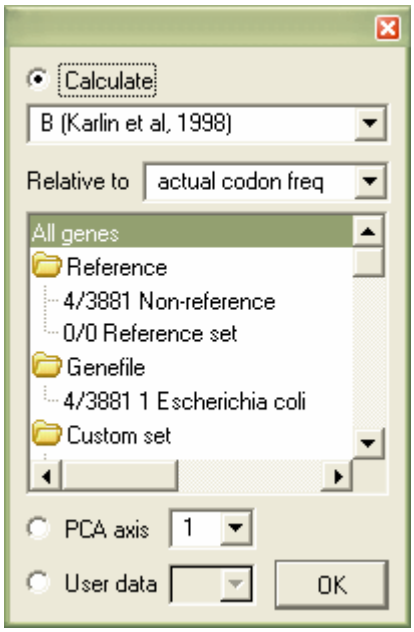
To the right you see a view of an example arrangement of genes into groups. Right-clicking on a group pops up a menu with options that should be self-explanatory. Before the name of each group, two numbers are displayed. The first one is the number of checked (see the Gene Browser section) genes, and the second one after the backslash is the total number of genes in the group.



A *gene property* is any numerical value associated with the gene. This might assume:

- A directly observable characteristic of a gene, such as its location on the chromosome (the start position), its length in codons, or the predicted hydrophobicity of its protein product (after Kyte and Doolitle)
- A measure of codon usage. Such measures are methods of quantifying codon usage patterns of a gene by comparing them to an expected codon distribution, the result being a single scalar value. INCA calculates many measures of codon usage (also called: indices); for an overview and comparison of the methods, refer to one of the following papers:
    - Comeron JM and Aguade M: An Evaluation of Measures of Synonymous Codon Usage Bias; *J Mol Evol* (1998) 47:268–274
    - Novembre JA: Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias; *Mol Evol Biol* (2001) 19(8):1390–1394
    - Supek F and Vlahovicek K: Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity; *BMC Bioinformatics* (2005) 6:182
- User data imported from a text file (see Importing user data)
- A PCA (*principal component analysis*) axis

Gene properties are the values assigned to axes of the Scatterplot, they can be used to visualize the Self-organizing map, or to divide the genes into Bins (among other things). A typical dialog to select a gene property looks like this:



The uppermost drop-down list is used to select the property. Some of them are affected by the choice of "Relative to", such as the measure called B selected in the picture which compares the codon usage of a gene to an expected pattern of codon usage. Others, such as gene length, are obviously not affected by the choice.

Implementation of some of the options should be described in more detail:

••• Karlin and Mrazek in 1996 published a J Mol Biol paper titled "What drives codon choices in human genes?" which describes a different method to calculate a statistic also named **"B"**. It has, to my knowledge, later been completely abandoned by the authors in favor of the 1998 version of "B", also called 'codon bias between gene groups'. This is the version computed by INCA.

••• "**$(G+C)_3$ at silent sites**" shows the proportion of G or C nucleotides at the 3rd position in codons where 3rd position mutations are silent. This value is an estimate of background nucleotide composition. When calculated by INCA it is somewhat different than the more general total $(G+C)_3$ used by some authors, which is to a degree affected by choice of amino acids. $(G+C)_3$ at silent sites is also sometimes defined and calculated as total $(G+C)_3$ minus the frequencies of Trp and Met, which is an adequate approximation when the universal genetic code is used, but differs slightly from INCA's more precise definition.

••• a method called **ENC'** (ENC prime) allows calculation of 'effective number of codons' in comparison to a standard different than uniform usage. This is commonly used

to more correctly quantify overall amount of codon bias in cases of uneven nucleotide composition, such as in GC-biased genomes. The method to calculate ENC' was described in a paper by J. Novembre in Mol Biol Evol (2002), who also developed a script that computes ENC' from nucleotide sequences using a slightly different method (available from http://ib.berkeley.edu/labs/slatkin/novembre/) Up to and including version 1.10, INCA followed the instructions to compute ENC' from the paper; since version 1.12, the program behaves exactly like Novembre's script. The differences between the two versions are, however, very small, and appear only in one of the two following cases:

a) when ENC' of a gene is close to 61 (i.e. very little difference in codon usage) the discrepancies between the versions are minor and amount to several tenths of one ENC' unit per gene

b) if a gene has exactly 5 amino acids of a certain type (e.g. 5 times Ala) the discrepancies are usually larger, but rarely exceed two ENC' units

Qualitatively, the results should remain similar. The method implemented in INCA 1.12 and newer is considered "more correct".

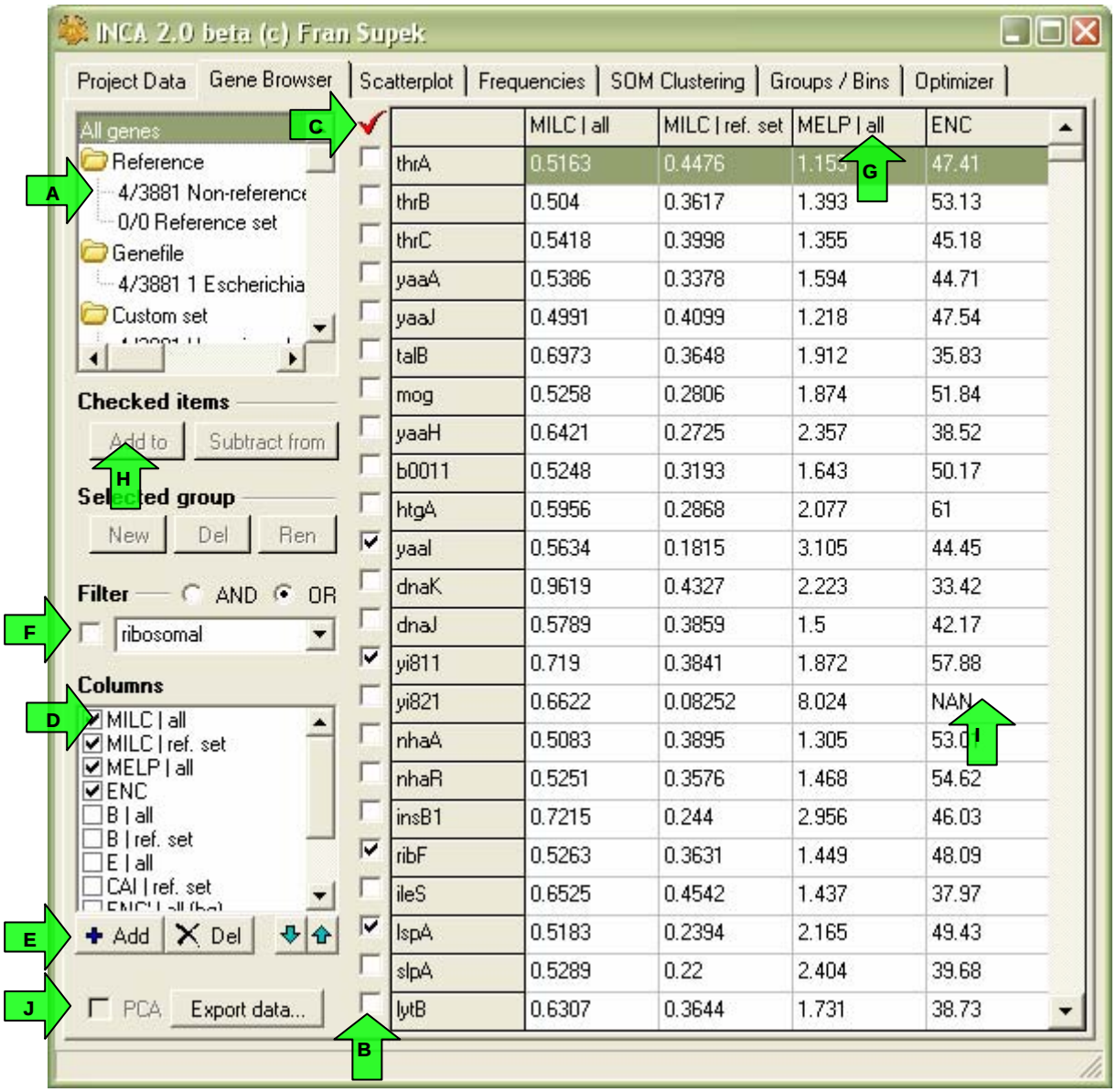
••• some measures of codon usage supported by INCA attempt to quantitatively predict the levels of expression of a gene; these are: **Fop, CAI, E, MELP** and **GCB**. All of them assume a reference set was defined, that includes genes known to be highly expressed (normally - the ribosomal protein genes).

When computing any of these expression predictors, INCA looks at "the reference set" group; if it's empty, they won't be computed correctly. Fop and CAI are not affected by the choice of "relative to" as the only parameters they require are the codon frequencies of the reference set. The other predictors are affected by the choice. For instance, E is calculated as:

$$E(gene) = \frac{B(gene \mid group)}{B(gene \mid ref.set)}$$

Frequencies of the reference set for calculating the denominator are determined by examining "the reference set" group; the "*group*" in the numerator is what we've chosen as "Relative to…" Normally, this should be the genome of the gene's host organism, or at least a codon table (see User-defined codon frequency tables) describing its codon usage pattern. The same rule applies to MELP and GCB, as they are calculated similarly.

The “Gene Browser” tab



A. Use the tree to select the gene group you wish to view. More than one group may be selected, e.g. Clusters 1 and 3, but not 0, 2 and 4, by ctrl+clicking or shift+arrow keys on the keyboard.

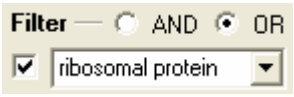
B. The checkboxes are used to set the ‘checked’ status of genes. Use this to mark a number of genes, and then move them to another gene group, e.g. the reference set, by selecting the it and clicking **[H]**.

C. Checks or unchecks the currently visible genes, i.e. those in the chosen groups **[A]** and satisfying the filter **[F]** criteria. The change in checked item count is immediately shown in the gene group tree **[A]**.

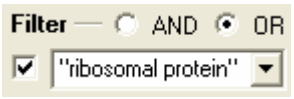
D. Select items which should be shown as columns in the browser. Be aware that enabling lots of columns will slow down INCA.

E. Add columns to display any of the available gene properties, selected using the standard dialog (see Gene groups and properties section).

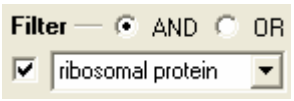
F. This option, when used, shows only the genes conforming to the filter criteria, i.e. containing the specified string in the gene name, synonym or description. Examples:



matches “hypothetical protein” and “Ribosomal subunit protein”



wouldn’t match any of the above, but would match “putative ribosomal protein”



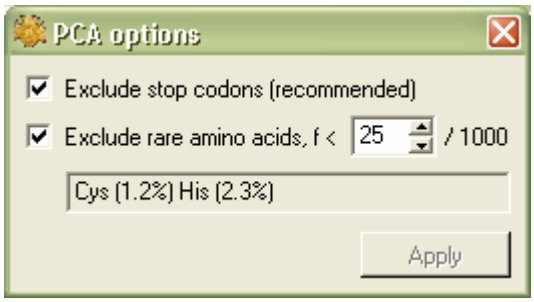
this would match “Protein, ribosomal” but wouldn’t match “hypothetical protein”

G. Click a column header to sort the table by that property; sorting is always done in ascending order (A to Z). An asterisk after the column header indicates it’s currently the sort criterion. Click and drag a column border in the header row to resize that column.

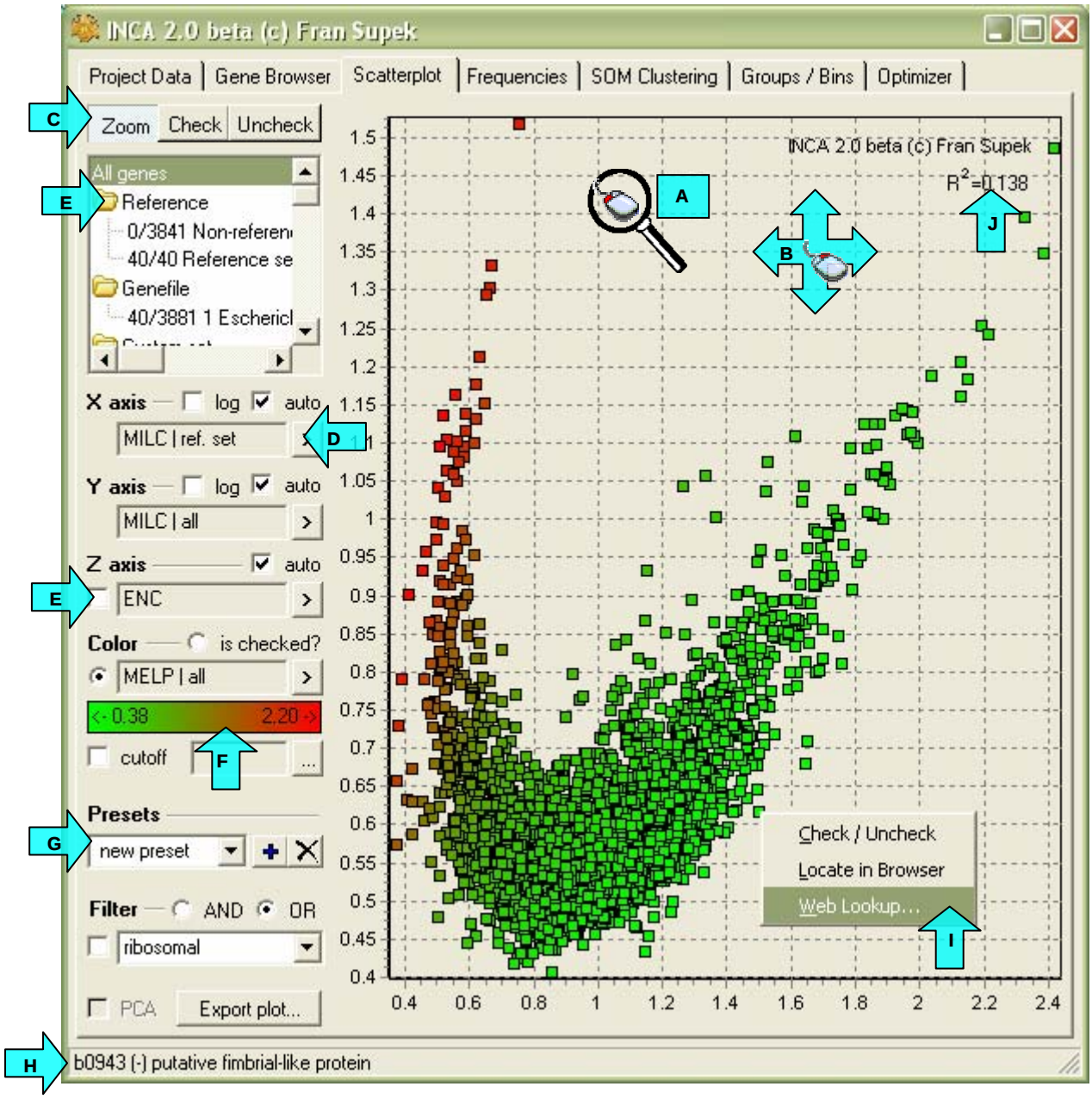
H. Adds checked genes to, or removes them from, the currently selected group. Same options are available by right-clicking gene groups in the tree.

I. “NAN” indicates a value couldn’t be computed (Not A Number).

J. This checkbox is checked when a visible column is set to display a PCA (*principal component analysis*) axis, which forces PCA recalculation anytime the active gene group changes and slows down Gene Browser refresh. You cannot manually activate this checkbox to use PCA; instead, add or show a column containing a PCA axis. While PCA is active, a small dialog box with additional options will be shown.



The “Scatterplot” tab



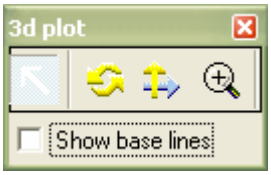
A. Clicking the left mouse button on an empty spot and dragging to the lower right zooms in; clicking and dragging to the upper right zooms out.

B. Clicking the right mouse button on an empty spot and dragging – scrolls the plot.

C. If “Check” or “Uncheck” is selected instead of “Zoom”, clicking and dragging with the left mouse button will check or uncheck genes in the target rectangle.

D. The “>” buttons allows a gene property assigned to an axis to be changed.

E. Checking this checkbox activates the 3D-plot, and shows an additional toolbar that enables spatial rotation and panning.



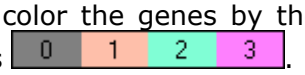
F. Clicking the gradient sets a cutoff value for the coloring; below is pure green, above is pure red. The “…” button enables for the cutoff to be set by entering a value manually.

G. Presets are useful for quickly changing between different plots. A preset contains information about what's assigned to each axis, whether the Z axis is visible, and the coloring criteria (property, cutoff). For example, to quickly activate PCA, select the "PCA" preset from the drop-down box. The presets you create using the "+" button remain saved in the i2settings.ini file, so they can be reused next time you start INCA.

H. The status bar displays: "synonym (name) description" for a gene that's currently under the mouse pointer.

I. Right-clicking a gene in the plot displays the pop-up menu. "Check/Uncheck" will affect the clicked gene only. "Locate in Browser" switches to the Gene Browser tab and highlights the clicked gene in the table. "Web Lookup" works only if INCA knows the gene's protein product's GI number, this will be the case if you loaded a KEGG file (.ent), or a FASTA file (.ffn) accompanied by a .ptt file.

J. This is the squared value of the linear (Pearson) correlation coefficient of the *x* and *y* axes; higher values indicate better correlation.

K. Clicking a gene group type will display all the genes, but color them by membership in groups of that type. For instance, clicking "Clusters" will color the genes by the cluster they belong to. The gradient **[F]** will then look like this 

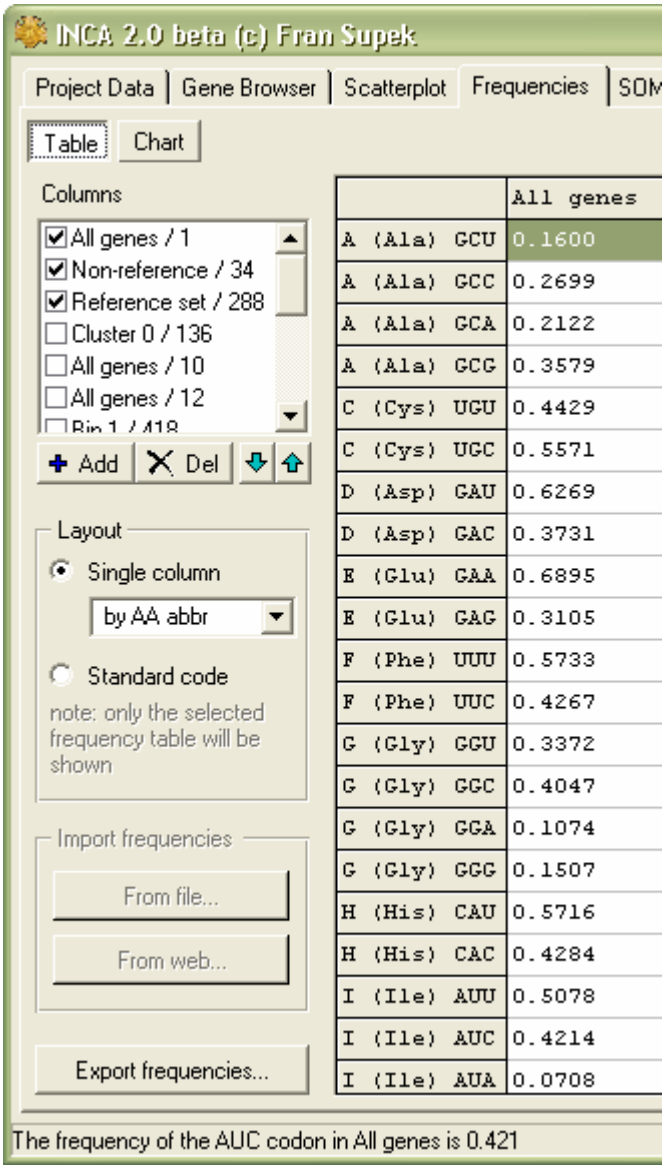
.

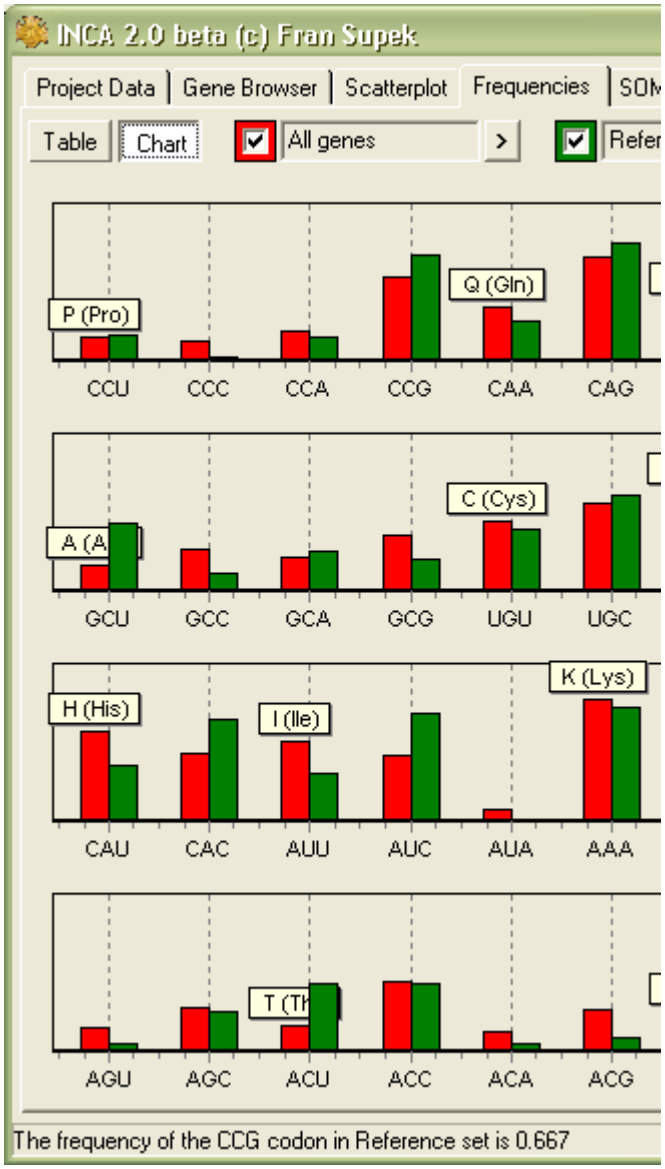
Other options, such as selecting a gene group to view in the group tree, the filter feature, or PCA, are very similar to the ones in the Gene Browser; refer to that section for more information.

The "Frequencies" tab

This part of INCA is used to easily side-by-side compare, either graphically or numerically, codon frequencies of gene groups of any type. Hint: to examine codon usage for a single gene, create a Custom set in the Gene Browser and add only that one gene to it; switch to Frequencies and choose the Custom set you just created.

There are two display modes, changed by clicking the "Table" or "Chart" buttons.





The "Table" mode is similar to the Gene Browser, only this time when adding columns you are asked to specify a gene group, for example the reference set, or Cluster 4. The "Standard code" layout shows only a single selected (highlighted) column, regardless of it having a checkmark next to it or not. The "Export frequencies…" saves only the currently shown columns to a tab-delimited text file.

The "Chart" mode can be used to compare up to three gene groups; codons are sorted by amino acid. Move the mouse pointer over a column to show the exact amount in the status bar (bottom of window).

The “SOM Clustering” tab

Quick start

1. Make sure you’ve loaded the gene files you intend to analyze and switch to the SOM tab.
2. Click the “Run w/o Vis” button (“Run w/Vis” is slower) and wait until SOM finishes training and recognition. Time required depends on number of genes analyzed and processor speed.
3. Adjust the cutoff value by clicking the cutoff bar, or manually specify a value by clicking the “…” button.

4. When satisfied with clustering proposal regarding cluster number, map and gene coverage, click the “Accept clusters” button.
5. Examine clusters in the Groups/Bins tab, Gene Browser (add a column with the *Membership: Cluster x* property) or Scatterplot (by e.g. coloring genes using the same property).

SOM introduction

Self-organizing maps (SOMs) are also named Kohonen maps, after their inventor, Teuvo Kohonen. A SOM is a variety of neural network often used for converting high-dimensional data (such as codon usage frequencies) into one- or two-dimensional maps than can be easily visualized. Additionally, a SOM can be used for grouping data into clusters.

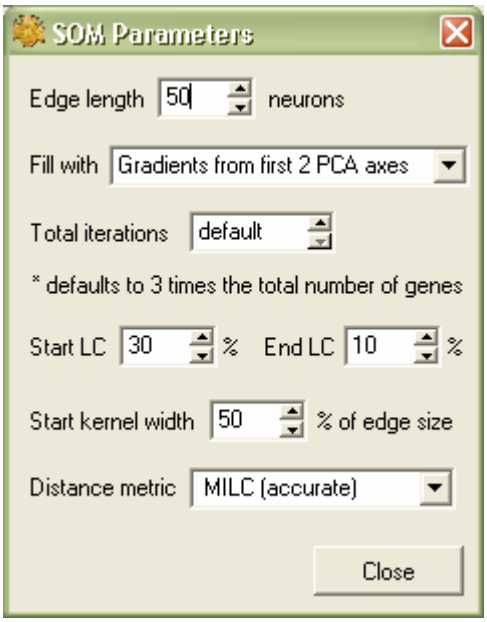
In short, INCA’s SOM consists of a layer of ‘neurons’, and each one has a table of codon frequencies associated with it. Before training has commenced, each neuron is usually assigned a random codon frequency table (other initial settings are possible, such as PCA initialization). The training lasts for a predefined number of cycles; each cycle, a random sample is picked out from the genome. The network is searched for the neuron closest in codon usage to the chosen gene, and then that neuron and its neighboring neurons are updated to become even more similar to the chosen gene (learning). This process is repeated many times while the strength of the learning process and the neighborhood radius decrease. In the end, the map should display a pattern of areas of similar codon usage, to which genes are then ‘mapped’ – each gene is assigned to the neuron most like it (recognition process).

Initial parameters

These settings influence the performance of the SOM profoundly; sometimes, a certain amount of experimentation is necessary to get the desired results. Click the “Parameters” button *before* starting training to access a dialog with the following settings:

- “Edge length” sets the size of the network, default value is 50; large values should in theory produce better results.
- “Fill with” determines initial contents of the network. The default is to initialize the map using PCA, which is recommended. Other settings include random or uniform values.

- "Random seed" is used to initialize the pseudorandom number generator so that the results of the SOM could be reproducible, i.e. the neural network, when run on the same genome with the same initial parameters (including the random seed) will produce identical results. On the other hand, changing just the seed while leaving all other parameters constant is a way to influence the outcome of the training.
- "Total iterations" is self-explanatory; default value is approximately 3 times the number of genes.
- "Start LC" and "End LC" refer to the learning coefficient; this is the number that determines how strongly the input (a randomly selected gene sample) affects the target neuron and its neighbors.

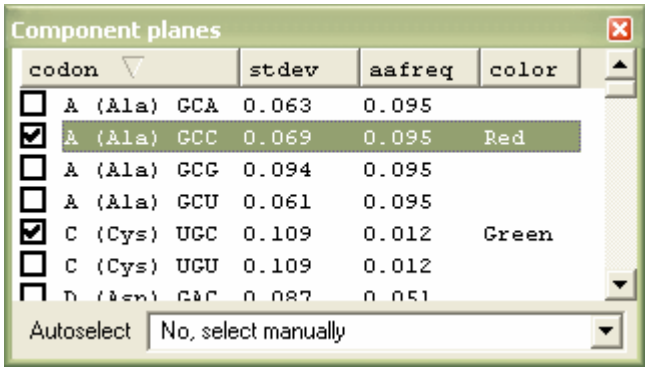


- "Start kernel width" determines the starting size of the neighborhood (or kernel); ending width is always limited to a single neuron.
- "Distance metric" has two choices. To quantify the distance in codon usage of a gene and a neuron, or between neurons (neighbor similarity), "B" uses a simple sum of differences of codon frequencies, while "MILC" uses a log-ratio chi-squared statistic. The latter may be more accurate, but training will be approximately 5 times slower.

Visualizing the SOM

The four buttons under the initial parameters are used to start, pause or resume training. During (and after) training, the network can be visualized in a variety of ways.

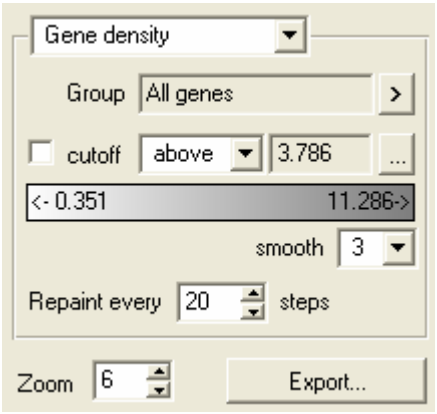
- "Difference from neighbors" colors the neurons according to how similar they are to their surroundings, where smaller numbers denote a stronger likeness and larger numbers a larger difference. This method of visualization is sometimes called the "U-matrix", and allows areas consistent in codon usage to be detected.
- "Component planes" can simultaneously show up to three codon frequencies and their distribution across the network. When you select this option, a dialog containing all the codons will pop up, allowing you to mark the ones you want shown. Each one is then assigned a color: red, green or blue. There is also an option for automatic selection of codons, which takes into account the variation of each codon frequency throughout the map and the frequency of the amino acid it codes for (these are the two values in the second and third column). Additionally, the list of the codons can be sorted by values in each column.



- "Standard properties" allows you to select any of the available gene properties to visualize the map. For instance, you may want to select "MILC | all genes" to detect areas that differ from the average codon usage of the genome you're analyzing. Or, if

you select "MILC | User table 1" you can highlight areas of the map similar to codon usage patterns of another genome, whose frequency table you loaded into INCA previously.

- "Gene density, genes/neuron" shows as brighter (higher values) the neurons that have a higher number of genes mapped to them. Since mapping is a process that occurs after the training is over, it doesn't make much sense to use this option before then. By choosing a specific gene group instead of "All genes", you may reveal the areas rich in e.g. reference set genes.



Checking the "cutoff" option will have as a result, instead of a spectrum of values being shown, that a number be chosen as the cutoff value. Neurons below it - or above, depending on the current setting - become white (in cluster), while the others are colored dark grey (not in cluster). The exact value can be changed by clicking the cutoff bar. The extreme values of the spectrum are indicated.

The "Zoom" setting is purely visual and has no effect on the functionality of the SOM: Finally, "Smooth" specifies how much (and if at all) should the map be blurred, which has an effect of making the edges of different areas smoother. This is particularly useful when using one of the "Gene density" visualization options to define clusters.

Keep in mind that visualizing the network while it is being trained will slow the training process considerably; to speed it up, increase the "Redraw net every *n* steps" value, or turn off visualization altogether.

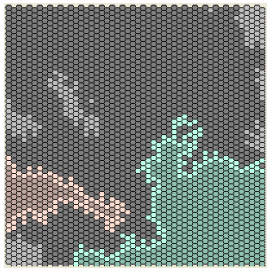
Clustering using the SOM

After the self-organizing map has finished training and all of the genes have been mapped to the network, those genes can be divided into clusters based on their position in the map. The current visualization settings are used to separate areas of the map, together with genes mapped to them, into clusters.

For this, it is necessary that the "cutoff" option be checked and a cutoff value specified. Two common strategies used in clustering are:

- Take the areas which seem consistent in codon usage patterns i.e. the neighboring neurons are quite similar, and use them to create clusters. Use the "Similarity to neighbors" criterion, turn on the cutoff, and make sure to select "below" – this is because *lower* values indicate higher similarity (less distance). This is the default strategy.
- Determine the areas more densely populated with genes, and use them to create clusters. Use the "Gene density" criterion, turn on the cutoff, and select "above". You will also probably need to increase "Smooth" to 2 or 3.

There is an option for rejecting areas smaller than a certain percentage of the map as cluster candidates; the default value is 20/1000 of the map. If you change it, you need to click the cutoff bar again to apply the new setting.



If the map is finished with the training and recognition phases, and "cutoff" is checked, INCA should automatically display the proposed clusters in different colors and inform you about number of clusters, map coverage and percentage of genes in clusters. The dark grey neurons are areas between clusters, and light gray ones are clusters that have been rejected based on the size threshold. All genes found in these grey areas will be put into Cluster 0 (Unassigned).

When you're satisfied with the clustering results, click the Accept clusters button.

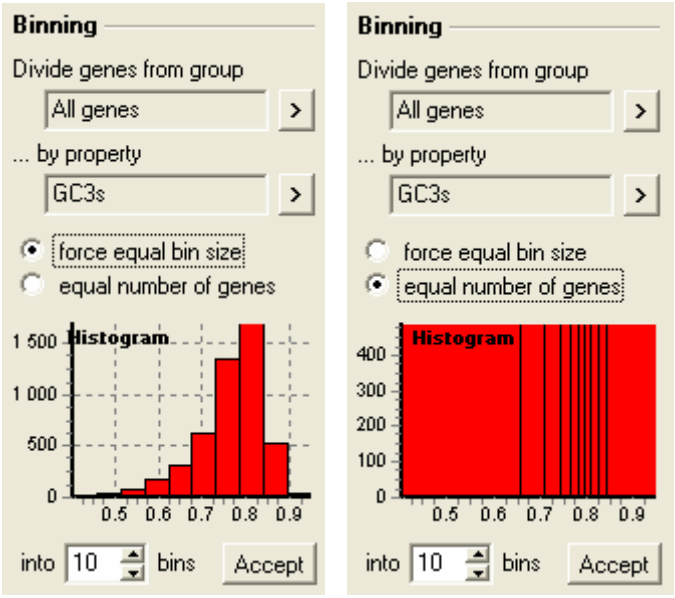
The Reset SOM command will return the self-organizing map to its initial state. Resetting the SOM will not void current cluster assignments (that you created by clicking "Accept clusters").

The “Groups/Bins” tab

This part of INCA is used to:

- examine a specified property of gene groups (*descriptive statistics*)
- discover correlation between groups (*contingency tables*)
- divide genes into groups based on a property (*binning*)

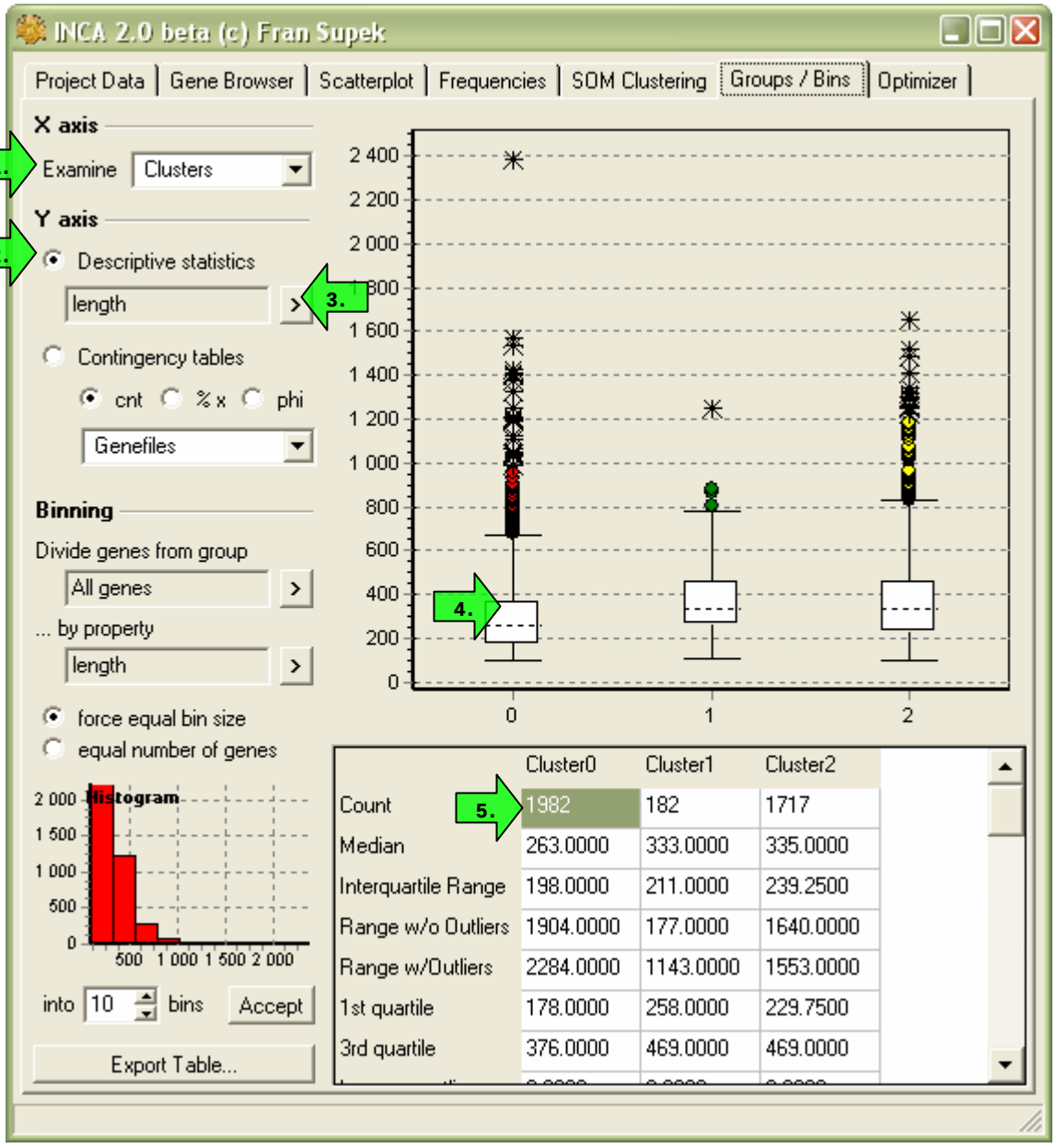
Binning



The binning feature is useful when examining the behavior of groups of genes that differ in a certain property, for example to determine whether longer genes have stronger codon usage patterns than shorter ones.

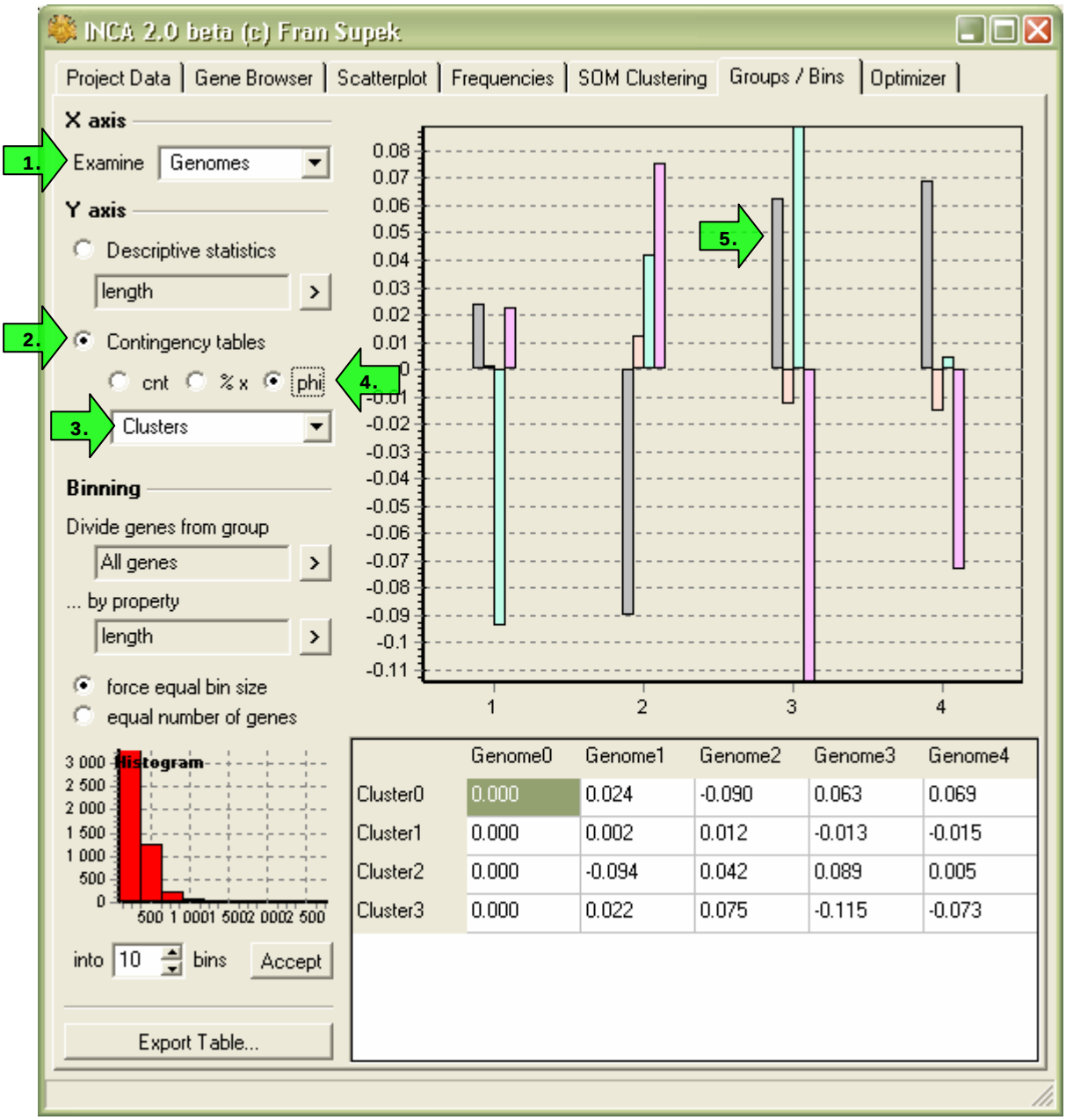
Select a group to subdivide into bins; genes not belonging to that group will remain in Bin 0 (Unassigned). Select a property, and if you’d like the bins to be equally sized, or equally full. When you’re satisfied by the distribution shown in the histogram, click the “Accept” button. Be warned that this will overwrite old binning assignments.

Descriptive statistics



1. Choose a type of gene group to examine.
2. Make sure the "Descriptive statistics" radio button is checked.
3. Select a gene property to examine.
4. The box-and-whisker diagram displays the median value as a dashed line, the 1st and the 3rd quartile as bounds of the white box. The whiskers are the highest non-outlier values (within 1.5 interquartile ranges of the 1st or the 3rd quartile). Colored dots are mild outliers (within 3 interquartile ranges of the 1st or the 3rd quartile) and stars are severe outliers.
5. The same values are also presented in a table.

Contingency tables

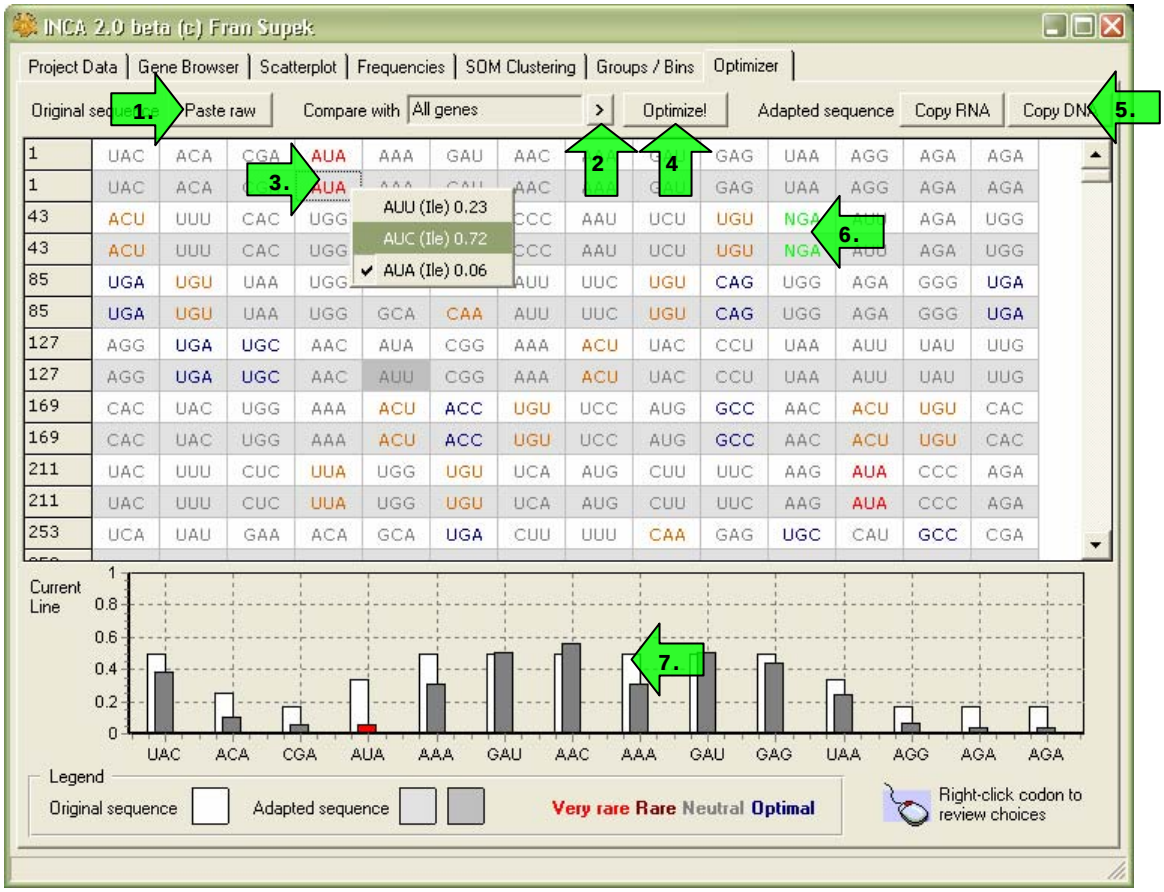


1. Choose a gene group type to examine.
2. Select "Contingency tables".
3. This is the group type **[1.]** is being compared to.
4. The "cnt" option is the true contingency table – it shows how many genes belong to each combination of **[1.]** and **[3.]**. "%x" shows the values as percentages of the size of the group on the *x* axis. "phi" shows a chi-squared value scaled to range from -1 to 1. Larger negative values indicate stronger avoidance, and larger positive values indicate stronger preference. If composition of one group type completely determines the composition of other group type (e.g. only reference set genes are in Cluster 1, and only non-reference set are in Cluster 0), phi would equal 1 or -1.
5. The values are displayed graphically and in the table. If you selected "%x" or "cnt" in **[4]**, the table also displays row and column totals.

The “Optimizer” tab

Attempts to achieve heterologous gene expression are often hindered by the fact that that the gene uses codons which are rare in the new host. This slows translation, and can even result in dysfunctional protein product. Because of this INCA now offers a tool to modify the sequence in question to agree better with codon preferences of the expression system, be it yeast, *E. coli* or human cells.

If the sequence contains only a few rare codons, actual DNA can be ‘corrected’ by oligonucleotide-directed mutagenesis; on the other hand, if the disagreement is more severe, a whole synthetic gene can be made-to-order. These kinds of procedures have often resulted in dramatic improvements in the amount of produced protein.



1. Open the sequence you intend to optimize in a text editor and copy it to the Clipboard. Then click the “Paste raw” button to import it into INCA.
2. Select the target gene group. This should be the reference set (or the whole genome) of the new expression host for the gene.
3. Rare codons, colored red and orange, can be changed by right-clicking each codon. White rows are the original sequence; shaded rows are the adapted sequence. Darker shading indicates a codon has already been changed.
4. The “Optimize” button instantly changes all codons to optimal ones.
5. When you’re done optimizing, copy the new sequence back to the Clipboard.
6. Codons containing IUB ambiguity codes are colored bright green.
7. White bars represent neutral codon frequencies (e.g. 0.5 for Cys, 1 for Met, 0.25 for Val), and the second series of values are the frequencies in the adapted sequence.