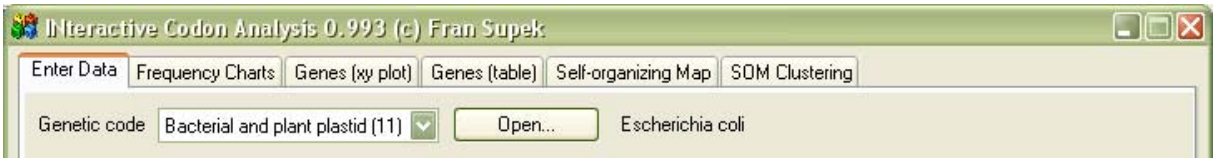
INCA: INteractive Codon Analysis 1.0 short tutorial

© Fran Supek 2003
fsupek@public.srce.hr
http://www.bioinfo-hr.org/inca

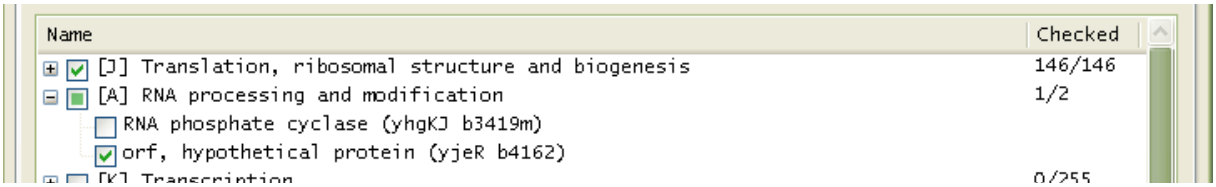
Welcome to the INCA tutorial. This document is meant as a quick, illustrated overview of all the features of the program; it's by no means comprehensive and covers only a fraction of the available options. For a more detailed information source please refer to the User Manual.

All the files necessary for this tutorial can be freely downloaded from the NCBI ftp site. The organism we'll use as an example is the well-known *Escherichia coli* K12. The Internet address of the genome is ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K12/ (please note this may have changed since this document was written; still, you should be able to find these files easily).

The two files we need are the NC_000913.ffn and NC_000913.ptt. The former contains hypothetical cDNA sequences for all *E. coli* genes in FASTA format, and the latter lists information about the genes and their protein products, such as gene name, description and COG functional category. Download the files to a folder on your computer.



Before opening the files, you will need to select the appropriate genetic code. This is not necessary with *E. coli* and most organisms since they use the default universal code. However, the Mycoplasma, mitochondria, some yeasts and some protozoa use alternative codes; keep this in mind.

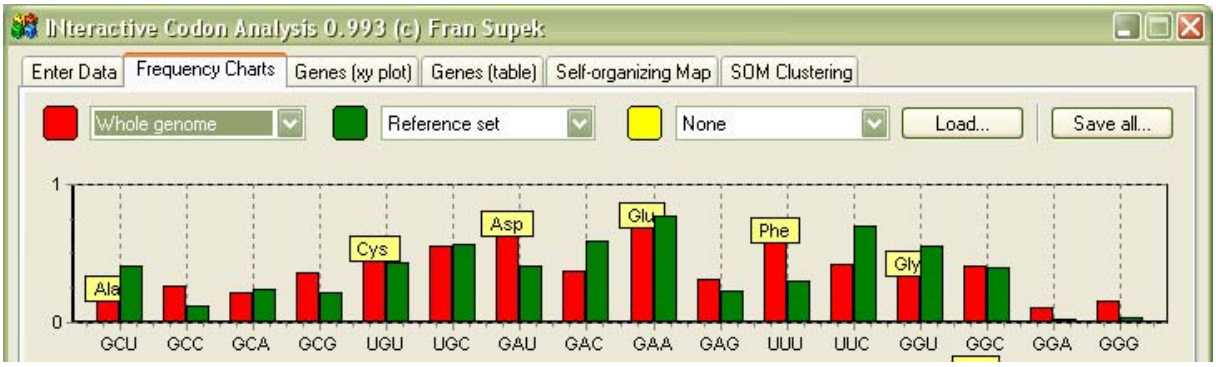


Immediately after parsing the files, INCA computes the codon frequencies for each gene; this is why loading takes a little time – normally a few seconds on a modern computer. Disregard the warnings about detected internal stop codons. All genes are displayed sorted by functional category; the checkboxes next to each gene are used to define the "reference set". Instead of selecting genes one by one, you can also check the whole category, or check all genes whose name or description matches a search criterion.

This is what we'll do right now. Enter the word "ribosomal" into the search box, and click "Check all". The status bar should read "43 genes checked". The reference set is normally

defined to contain highly expressed genes, ribosomal genes being prime examples; these genes usually show very specific codon preferences. Note that this kind of text search is a crude way of selecting a certain class of genes, and prone to errors - as we will demonstrate later.

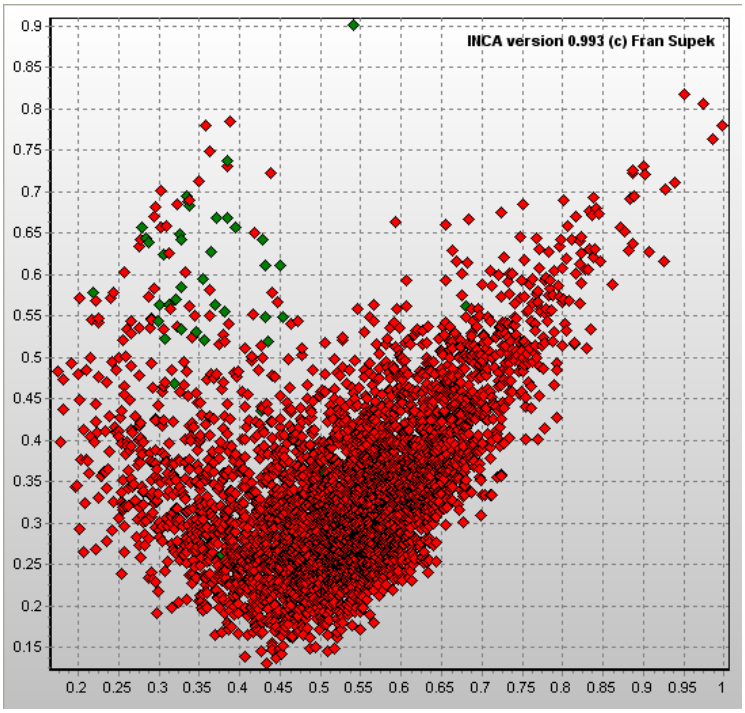
Now select the “Frequency charts” tab. There are 3 drop-down lists at the top of the screen. Set the first one to “Whole genome”, and the second to “Reference set”. You should see something like this.

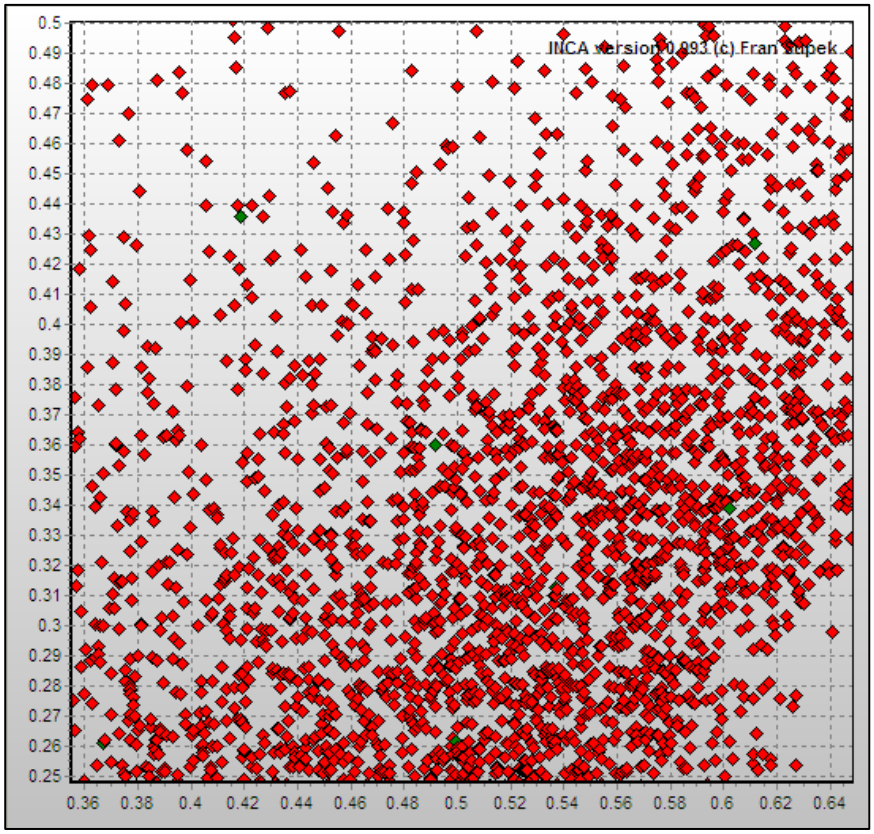


The codons are sorted by the one-letter abbreviation of the amino acid they encode; move the mouse over the columns and INCA displays the exact value of each frequency in the status bar. In this case, the chart shows that the reference set indeed has a preference for a certain set of codons: UGC for cysteine, GAC for aspartate and UUC for phenylalanine.

Switch to the “Genes (*xy* plot)” tab; the individual genes are now shown as dots. Their *x* coordinates defaults to codon usage difference from the reference set, and *y* to difference from the average of the genome. The plot displays a pattern reminiscent of a crescent moon, whose left “horn” consists mostly of ribosomal set genes (green dots).

Moving the mouse over dots shows additional information about the gene in question. Notice a green dot at *x*=0.49, *y*=0.36 (if you have trouble spotting this gene among the thousands of *E. coli* genes, use the Sonar and search for “rimL”). Now zoom in to be able to see the individual dots in the vicinity better; click a point to the upper left of the gene, hold the mouse button, drag to form a box around the gene and release the mouse button. Doing the same thing in the opposite direction un-zooms. Clicking and holding the right mouse button scrolls in the direction of the mouse movement.

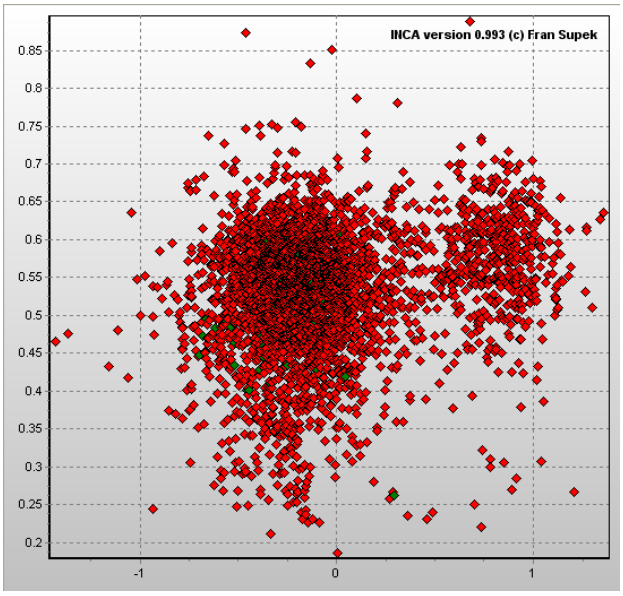


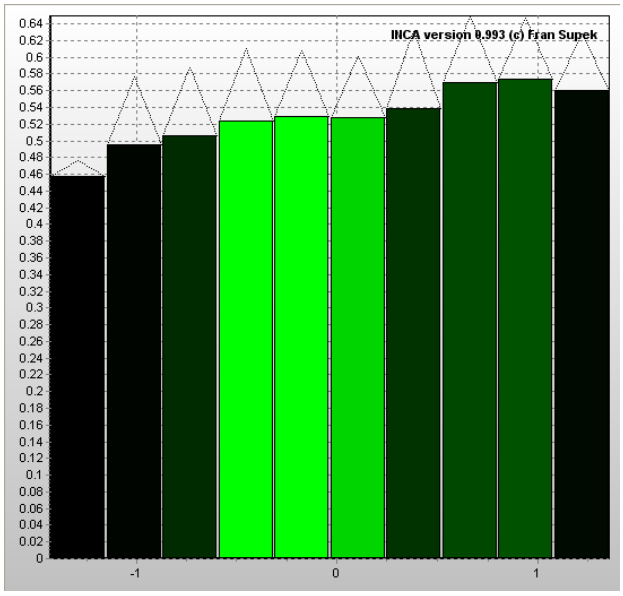


If you position your mouse pointer over the green dot, INCA should state this is the rimL gene; taking a look at the description in the status bar reveals that this is actually not a ribosomal protein, but an acetyl transferase. Now click the dot to remove it from the reference set; notice how all the other genes change their *x* coordinate slightly to reflect the change in the reference set. The smaller the ref. set is, the larger the change caused by adding or removing a gene.

If you examine the other green dots outside the "left horn", you'll mostly find more modification enzymes and some putative ribosomal proteins, not actual ones. This is a striking example how codon usage can reveal the exact nature of genes and their products.

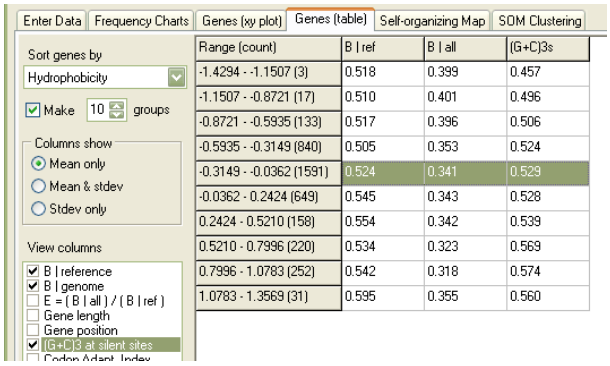
Now try this: assign the "Hydrophobicity" setting to the *x* axis, and "(G+C)3 at silent sites" to the *y* axis. You should see a pattern resembling the left picture.





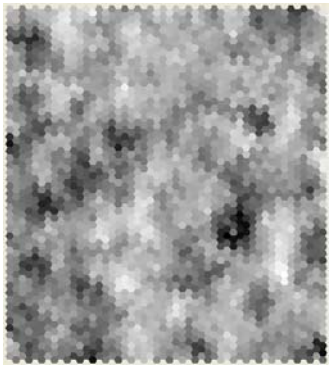
Even from this it could be concluded that the hydrophobic group of proteins have higher GC content. Note this is not simply a consequence of their biased amino acid composition, since bases are counted only at silent (or synonymous) sites. Click the “Make *n* groups” checkbox to divide all the genes into groups based on their *x* axis value and show them as columns. Now it’s even more obvious there is a trend toward more GC use in hydrophobic proteins in columns 8 and 9 compared to columns 4, 5 and 6 which contain the majority of other genes.

The “Genes (table)” tab displays the same information in numeric format. Switch to the tab, select “Hydrophobicity” as the sort criterion, click the “Make *n* groups” checkbox and also be sure to instruct INCA to show the GC3s column. From the table we can conclude that the hydrophobic proteins generally have GC3s content around 0.57, while the others have this value at 0.52 to 0.53.

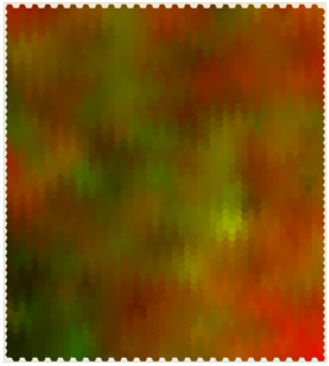


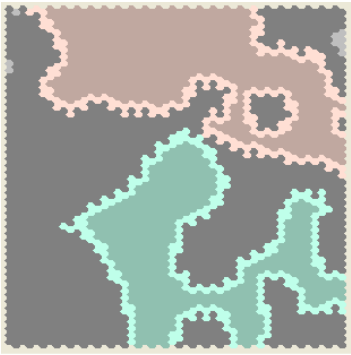
Range (count)	B | ref	B | all	(G+C)3s
-1.4294 - -1.1507 (3)	0.518	0.399	0.457
-1.1507 - -0.8721 (17)	0.510	0.401	0.496
-0.8721 - -0.5935 (133)	0.517	0.396	0.506
-0.5935 - -0.3149 (840)	0.505	0.353	0.524
-0.3149 - -0.0362 (1591)	0.524	0.341	0.529
-0.0362 - 0.2424 (649)	0.545	0.343	0.528
0.2424 - 0.5210 (158)	0.554	0.342	0.539
0.5210 - 0.7996 (220)	0.534	0.323	0.569
0.7996 - 1.0783 (252)	0.542	0.318	0.574
1.0783 - 1.3569 (31)	0.595	0.355	0.560

Finally, let us take a look at the self-organizing map (SOM), a variety of unsupervised neural network used by the software to arrange genes into cluster according to their codon preferences. Click the appropriate tab; leave the initial settings at their defaults. Click the “Run w/Vis” button, and the network begins training itself. If you turn off the visualization, by clicking the “Run w/o Vis” button or selecting the “No Visualization (fast)” method, you will speed up training significantly. After it has completed, you’ll have to wait for a few more seconds for genes to be “mapped” to the network (their *x* and *y* coordinates determined).



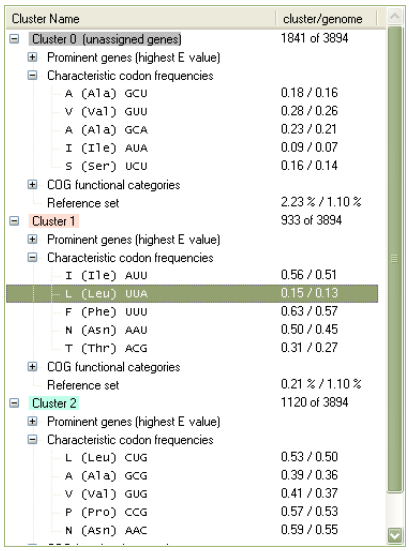
At the default visualization setting, “Similarity to neighbors”, areas with monotonous codon usage are colored white; on the other extreme, areas with rapidly changing codon frequencies are black. At the end of training, the map should look as pictured. By checking the “Threshold” option, only the areas above the chosen cutoff value will be colored white, and others black. Now turn the “Threshold” option back off, and select the “Component planes” method. This shows a list of individual codons, with columns stating how much the frequency of the codon varies around the map (stdev) and how common is the amino acid it encodes. These criteria should help you decide which component planes you want to show, or you can let INCA choose the most interesting planes by using the auto-select option. Set it to “Two codons” and the software will choose CUG (Leu) and GCC (Ala) and assign different colors to them. The most important auto-selected codon is displayed in red, second most important in green and third in blue (although I can’t recommend displaying more than 2 at the same time, because the map gets difficult to interpret). Now set the method back to “Similarity to neighbours”, the “Smooth radius” to 3, threshold to 0.0640 and click the “SOM clustering” tab. Here, each continuous area of “white” will be considered to be a single cluster.





The threshold we have chosen should generate 2 clusters. You can also change the cutoff value on this tab and see the effects it has on number of clusters and their coverage of the genome (see bottom of INCA window). You also have the option to discard clusters smaller than a certain percentage of the map. Currently, 53% of all genes are in clusters. All the genes belonging to neither of the two clusters are assigned to Cluster 0. You may click the “Report >” button to see some additional information about each cluster.

There are several things we can conclude from this report, and the most important one would be to see that the SOM actually generated meaningful results. Arrange the report so it looks like pictured, i.e. expand the “Characteristic codon frequencies” for each cluster. In Cluster 2, a preference is displayed for use of GCG for alanine, while Cluster 0 leans toward using GCU and GCA. Cluster 1 is characterized by using AUU for isoleucine; clusters 1 and 2 differ in their use of codons for asparagine, the former “likes” AAU and the latter AAC. You can also see a sample of genes representative of the entire cluster and check whether there is any association between clusters and gene functions, or with the reference set. Clustering data is now also accessible in the “Genes (table)” tab, from where it can be saved to a text file and analyzed further.



This concludes the tutorial; for more details consult the User Manual.