INCA 1.20a User Manual

© Fran Supek, 19. September 2004

INCA 1.20a User Manual 1
Welcome 2
What is INCA? 2
What are the requirements to run it? 2
How to obtain INCA? 2
Ideas and bug reports 2
Thanks 2
Disclaimer 2
Please cite 2
Version History 3
The "Enter Data" Tab 5
Opening Files 5
The reference set 6
Gene length threshold 6
Random sequence generator 7
The "Frequency Charts" Tab 7
User-defined codon frequency tables 7
The "Genes (xy plot)" Tab 8
Measures of codon usage 8
Other properties & chart options 9
The “Genes (table)” tab 10
The “Self-organizing Map” tab 11
Initial parameters 11
Visualizing the SOM 11
The “SOM Clustering” tab 12
The Clusters / COGs tab 13
The Optimizer tab 14

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 when using the SOM.

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

This software is provided "as is" and any expressed or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose are disclaimed. In no event shall the regents or contributors be liable for any direct, indirect, incidental, special, exemplary, or consequential damages (including, but not limited to, procurement of substitute goods or services; loss of use, data, or profits; or business interruption) however caused and on any theory of liability, whether in contract, strict liability, or tort (including negligence or otherwise) arising in any way out of the use of this software, even if advised of the possibility of such damage.

Please cite

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

 © Fran Supek 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.

© Fran Supek 2004

The "Enter Data" Tab

Opening Files

The very first thing you need to do, before opening any files, is to select an adequate genetic code for the organism you're examining. INCA supports several commonly-used translation tables. Changing this option after the genome has been loaded has no effect.

Next, you need to select one (or more) files to analyze. After clicking the "Open…" button, the program will ask you to select a file with an "ffn" extension, which contains hypothetical cDNAs for the organism you are examining in FASTA format.

It expects individual genes to be separated with comment lines (beginning with '>'); they look something like this:

```
>K05_orf380, gi|13507739:692-1834
```

or

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

When parsing such 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.

The comment lines should be followed by the nucleotide sequence. During analysis of sequences, INCA ignores codons containing unspecified nucleotides, such as R, Y, or N. It also performs some checks, such as looking for internal stop codons or genes not terminated by stop codons and will report such problems after loading is done.

The software 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 COG 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:

```
17489..18655 + 388 16128013 nhaA b0019 P COG3004
                                         Na+/H antiporter, pH dependent
```

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

More than one file can be loaded; you will need to do this, for example, when examining bacteria with more than one chromosome. Just select all the files using the Shift key in the Open dialog. If the program believes the files come from different organisms it will warn you after loading.

While the files get parsed, codon usage for each gene is also calculated; that's why the process takes a few moments. For example, a 1.7 GHz Pentium 4 loads the *E. coli* genome in 5 seconds.

The reference set

INCA displays all the genes in a list, organized by COG functional category to make it easy to find your way around. Click the plus sign before the category to show all the genes belonging to it. Note that if the genes haven't been assigned to categories, or no ptt file could be found, all the genes will be put under "Not Classified".

Clicking the checkboxes next to individual genes adds or removes them from the reference set; the checkboxes next to categories affect all the genes in the category. Counts are displayed in the rightmost column.

You can use the buttons beneath the list to find a gene of interest. The search looks for the whole phrase (Boolean operators are not supported) in a gene's name, synonym or description properties. It's also possible to check or uncheck all genes matching the search criteria (look at the status bar to see how many genes actually got checked).

The list of the checked genes can be saved to a text file, and its lines look like this:

```
b0473    htpG        chaperone Hsp90, heat shock protein C 62.5
```

That kind of files can be applied to the current reference set, by adding or subtracting it from the set. If you wish to create a gene set file manually, keep in mind the following: all lines beginning with the > sign and empty lines are ignored. The first valid line contains the organism name to warn for possible mismatches; second and all other lines contain information about genes. Only the first word in the line (up to a tab, comma or space) is the search criterion; the rest of the line is ignored. Here's an example of a valid file:

```
>this file contains some E. coli chaperones
Escherichia coli

dnaJ           chaperone with DnaK; heat shock protein
b0436
fimC (b4316)
```

Result: the software will look for the dnaJ, b0436 and fimC genes.

Gene length threshold

At the bottom of the window is an important option which enables you to exclude all genes shorter than the specified length from further examination. This is necessary because short genes have a limited codon sampling and therefore aren't suitable for codon usage calculations. Changing this number will immediately show you how many genes got rejected. Please note that when increasing the threshold, the genes that belonged to the reference set and fall below the new threshold get removed from the reference set; setting the threshold back to the lower value won't restore the old reference set.

Normally, a value of 100 codons is recommended.

© Fran Supek 2004

Random sequence generator

This feature of INCA is used to generate nucleotide 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) for more information about the presets. An option is also offered that causes the generated sequences to match the codon preferences of the currently loaded genome.

The amino acid frequencies normally 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 "Force uniform amino acid composition" check box enables 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 generated sequences can be loaded directly into INCA or saved into a FASTA formatted file, either as DNA or RNA.

The "Frequency Charts" Tab

At the top of the window there are three drop-down lists, and to each assigned a color. Selecting an option from the list shows the chart of codon frequencies pertaining to the option.

Selecting “Whole genome”, for example, shows the average codon frequencies across the genome, (more exactly, the genes above the set length threshold). “Normal usage” displays the theoretical frequencies for each amino acid in the absence of any codon bias, where frequency=1/number of codons for the amino acid. Before selecting “Reference set”, you will need to define it (if you haven't done so already). “Selected neuron” enables you to have an idea of what's happening at a certain spot in the self-organizing map; more on this later.

Finally, the “Amino freq, norm” option shows amino acid frequencies in the genome, adjusted so that the most common amino acid has a frequency of 1. This is done simply to improve the clarity of the chart.

Moving the mouse pointer over individual chart values will show the exact heights of the columns on the status bar. The only exceptions are amino acid frequencies – the chart shows the normalized values, but the status bar displays the exact numbers.

User-defined codon frequency tables

The “User freq table” option might be useful, for example, if you want to compare the codon usage of a certain genome, or a part of it, with codon usage of another organism. Before that you'll have to load the frequencies from a 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 what a part of the exported file, obtained by the use of "Save all..." button looks like. Codons are sorted by the one-letter abbreviation of the corresponding amino acid, just like in the chart. (INCA uses X for stop codons). Only the currently visible columns will be exported.

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 genome, or a file with some missing items. However, doing so will surely yield erroneous results 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 GeneBank entries (sequenced genes). Such tables may be copied and pasted into Notepad, or any other plain text editor, the file saved as .txt and loaded into INCA.

The "Genes (xy plot)" Tab

This tab contains many options for visualizing genes in a Cartesian coordinate system by assigning certain properties of each gene to its X or Y coordinate. These properties can be selected from the "X axis" and "Y axis" drop-down boxes.

Measures of codon usage

The "B" value is a measure of codon bias of one gene group relative to another group (single genes can also be used), taking into account both the codon frequencies and the amino acid composition. It ranges from 0-2, but only in extreme cases exceeds 1. Its advantages are that it's easy (fast) to compute, intuitive to use (has a linear scale) and has proven useful in applications such as predicting gene expression. It was introduced by Karlin, Mrázek and Campbell in 1998; refer to any of Karlin's recent papers for the exact methodology behind it, for example "Codon usages in different gene classes of the *Escherichia coli* genome" (Mol Microbiol, 1998).

A word of caution: 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'.

INCA uses this newer B measure extensively – you can compute and plot B for any gene against the genome average, the reference set or any of the clusters generated by the self-organizing map (more on that later). The "E" value is computed as the ratio between B(genome) and B(reference) and in certain organisms correlates quite well with the expression levels of a gene, provided the reference set contains highly expressed genes. The "MCB" is conceptually similar to "B" – it also measures the difference in codon usage between two genes or gene groups, but relies on a modified chi-square formula; for more details refer to Urrutia & Hurst (2001).

 © Fran Supek 2004

The "Gene length" and "Gene position" options are mostly self-explanatory; note that 'position' refers to the beginning of a gene.

"$(G+C)_3$ 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 a measure of background nucleotide composition. When calculated by INCA it is similar to, but more accurate than the more general total $(G+C)_3$ used by some authors. $(G+C)_3$ 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.

The "Codon Adaptation Index" refers to the CAI, a measure introduced by Sharp and Li in 1987, which predicts gene expression based on the codon frequencies of a small set of highly expressed genes. When computing CAI, the reference set is expected to contain such genes. Another commonly used statistic is the "ENC" (effective number of codons, also abbreviated N_c), as defined by Wright in 1990, which describes the deviation of codon usage of a gene from uniform usage. It ranges from 20 when only one codon is used for each amino acid, to 61 when there is no codon bias.

A modification of the original ENC, called ENC' (ENC prime) allows to specify – in the "relative to" drop-down list – a standard different than uniform usage (i.e. no bias). This is commonly used to calculate the effective number of codons correctly in cases of uneven nucleotide composition, such as in GC biased genomes. The method to calculate ENC' was first described in a paper by J. Novembre in Mol Biol Evol (2002), who also developed a script that computes ENC' from nucleotide sequences, but using a slightly altered 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".

The "relative to" drop-down list setting has an effect only in calculations of B, E, ENC', CAI and MCB.

Other properties & chart options

"Hydrophobicity" is calculated as the arithmetic mean of the sum of the hydropathic indices of each amino acid of the hypothetical protein product, as defined by Kyte and Doolittle in 1982.

The last four options "Reference set", "Cluster no.", "COG category" and "COG super-category" assign an integer index to each gene depending on the group it belongs to. Note that genes in the reference set are assigned a value of -1 and other genes a value of 0 when that criterion is applied. These options are far more useful if grouping is applied (see below).

Moving the mouse over the points (genes) in the plot displays additional information in the space to the left of the plot, and the gene description is displayed in the status

bar. A region of the chart can be zoomed by clicking and dragging the mouse pointer from the upper left corner of the desired viewport to the bottom right corner. Dragging the mouse in the opposite direction un-zooms the chart. Clicking and holding the right mouse button while moving the mouse scrolls the chart in different directions.

To the left of the chart, an r-squared statistic, a measure of linear correlation of the two properties, is calculated and displayed

Values distributed along the X axis can be grouped into an arbitrary number of groups. Each group covers an equal range of the property, and they are displayed as columns which replace the standard xy plot. The color of each column varies from green to black, depending on how many genes contribute to it. The height of the column is the average value of Y property, and the 'roof' shows the standard deviation of the group.

The number of groups is fixed if one of the 'ordinal' properties (e.g. COG category) is chosen for the X axis.

Options offered by the "Color dots by…" box allow you to more easily distinguish between certain groups of genes, such as the reference group, different clusters, or functional categories. The effect of clicking a gene in the plot depends on the option selected in this box:

- Membership in reference: add or remove a gene from the reference set
- E: change the cutoff value for determining highly expressed genes
- Strand: clicking does nothing
- COG categories, all / single: select the category of the clicked gene as criterion
- Clusters all / single: select the cluster of the clicked gene as criteria

The "Sonar" feature at the bottom of the screen is used to locate individual genes in the chart. The location of the first gene matching the search criterion is revealed by the moving circle. "Export to EMF…" saves the chart as an Enhanced Metafile, a graphics format that is widely supported and enables pictures to be resized without a loss of quality.

The "Genes (table)" tab

Here INCA displays very similar information to the last tab, but this time in tabular format. The rows of the table normally contain single genes, and the columns showing individual properties of the genes can be showed or hidden at will. The list of the available columns contains some commonly used properties, but can be expanded with additional ones.

Items can be sorted and grouped by any of the properties. There is an additional option available when grouping is enabled. It determines the display of the standard deviation of values within a group – it can be hidden, shown in parenthesis beside the average value, or shown instead of it.

The "Export table…" button saves the table, with the current display options applied, to a tab delimited text file, which can easily be loaded into a spreadsheet or a statistical analysis program.

 © Fran Supek 2004

The “Self-organizing Map” tab

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). 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 (lowest B value) 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.

Initial parameters

These settings influence the performance of the SOM profoundly; sometimes, a certain amount of experimentation is necessary to get the desired results. "Edge length" sets the size of the network, while "Fill with" determines initial contents of the network.

“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.

Visualizing the SOM

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

- “Similarity to neighbors” colors the neurons according to how similar they are with their surroundings, where white denotes the biggest likeness and black the smallest. Similarity is determined by calculating the “B” value. This method of visualization is sometimes called the “U-matrix”.
- “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.

- “Difference from genomic bias” means cells more similar to the average codon usage of the genome are darker, and those that differ are drawn lighter.
- “Difference from user table” is analogous to the above, but it uses the user-defined codon frequency table instead of the genome for comparison.
- “Gene density, genes/neuron” shows as brighter 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 that.
- “Gene density (reference set)” is similar to the last option, but takes into consideration only the genes selected as the reference set. This may be useful if you, for example, need to see if the selected genes have clustered together and where exactly.

Checking the “Threshold” 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 visualization - become white, while the others are colored dark grey (black). The cutoff value can be changed using the scrollbar below. The legend accompanying it shows to which colors the extreme values of the scrollbar correspond.

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.

Clicking a single cell in SOM selects it, which means you can view its codon frequencies in the “Frequency charts” tab. The selected neuron is marked by a red border around it.

The button marked “Reset net” will return the SOM to its initial state. Resetting the network will also void any clustering data (see next tab for details).

The “SOM Clustering” tab

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. This tab displays a (visually) smaller version of the same SOM as the last tab, and attempts to separate areas of it, together with genes mapped to them, into clusters.

The “Method” field contains the same items as the “Visualization method” on the previous tab, which means any method of visualization can also be used as a criterion for clustering. The most common method is the “Similarity to neighbors” (U-matrix) is used, where 'white' areas of relatively uniform codon usage get separated with 'black' areas with rapidly changing codon usage.

The “Smooth”, “Zoom” and “Threshold” are analogous to the same options on the previous tab, except that “Threshold” is always on – it needs to be so that continuous areas of 'white' can be declared clusters. The definition of 'white' depends on the method used; sometimes it's the areas above the cutoff value, like in “Gene density”; on other occasions it's the areas below the value, such as in “Similarity to genomic bias” - it's the areas deviant enough from the average that become clusters. Above the scrollbar, INCA shows you which the case is currently.

 © Fran Supek 2004

There is an option for rejecting areas smaller than a certain percentage of the map as cluster candidates.

The data shown in the bottom of the window: number of clusters and percentage of genes in clusters can help you decide when you're satisfied with the clustering results. Clicking "Report" will perform some calculations and show you additional information about each cluster, including Cluster 0 (genes not belonging to any cluster). This information includes:

- "Prominent genes" are determined by computing a modified E value – the B(genome) is divided with B(cluster); genes with the highest E are displayed.
- "Important codon frequencies" shows the frequencies that deviate the most from the average frequencies.
- "COG categories" displays the percentage of the genes in current cluster belonging to a certain functional category compared to the percentage of that category across the genome. The categories (up to 5) that differ significantly from the expected distribution are shown. Significance is determined by means of chi-square test with the condition that $p<0.05$.
- "Reference set" is analogous to the previous option, but shows the percentage of genes that the reference set occupies, instead of COG categories.

The "Map resolution" is a measure of the quality of the clusters generated by the self-organizing map. This is an average modified E value (the same as the one used to determine "Prominent genes") of all genes. In other words, it shows how many times the codon usage of a gene, on average, is more similar to the cluster the gene belongs to than to the whole genome.

Note that changing any of the clustering settings, either on this or on the previous tab, will erase the report (you will be asked for confirmation before this happens).

When viewing information about genes in the XY Plot or Table, before clustering is attempted each gene has a default value of -1; after clustering, genes assigned to a cluster have values of 1 and upward, an all the genes outside clusters have a value of 0.

The Clusters / COGs tab

This allows you to visually determine if the SOM clusters are associated with the COG functional categories. The chart only displays half of the categories at once; simply click anywhere on the chart to show the other 13 categories. The chart can show one of the three things:

1. Relative values – percentages: Shows, for example, what part of Cluster 1 consists of genes in the J category. (this is relative to the *cluster*, not to the *category*)
2. Absolute values – no of genes: Shows how many genes belong to each cluster/category combination.
3. Phi stat (measure of association): This is a normalized chi-square value, and can therefore be compared between samples – it's independent of sample size. A larger phi implies stronger association, whether positive or negative.

If you're unsure of what the chart is showing, move the mouse pointer over a column and an explanation will be shown in the status bar.

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' in the wet-lab 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 (up to two orders of magnitude) improvements in the amount of produced protein.

INCA expects that the sequence you want to optimize is on the Clipboard – you can Copy it from any text editor, and then click the "Paste raw" button in INCA. The text you copied should contain nothing except a sequence of nucleotide and line breaks (*raw* data, no FASTA-style comment lines).

You can optimize the sequence to agree with the codon frequencies of the

- current genome ("Genome freq"),
- the reference set ("Ref. set freq"),
- or the user frequencies you loaded from a file previously ("User freq").

Depending on which you choose, codons will be colored according to how rare (or optimal) they are. Keep in mind that the lines with white background (odd) contain the original sequence, while those with grey background (even) reflect the changes you made to the sequence.

Codons containing IUB nucleotide ambiguity codes, such as N (any nucleotide), R (A or G), Y (C or T) or others are colored bright green.

You can make changes:

- to the current codon, by right clicking it and choosing a more frequent one,
- or to the whole sequence, by clicking the Optimize button

The chart in the bottom of the window shows two sets of columns: the white ones in the background represent the normal (unbiased) frequency for a codon, i.e. 0.5 for each of the two asparagine codons. The other set of columns take the same colors as the codons – i.e. red for very rare codons, and their height is the frequency of the codon in the set you chose in the "Compare with" combobox.

When you're done with the optimization, click one of the "Copy…" buttons to transfer the adapted sequence to the Clipboard.

 © Fran Supek 2004