Comparative Gene Ontology Website User Guide

http://www.comparativego.com/

Author: Mario Fruzangohar

This website has been designed for researchers who want to study the association between gene ontology and lists of genes or samples in order to understand the biological pathways involved or to select important genes for further analysis.

Gene Ontology (GO) is divided into three name spaces: Molecular Function, Biological Process and Cellular Component which we abbreviate as MF, BP and CC respectively.

Note 1: All the genes used as illustrations in this manual were extracted from the *Streptococcus* bacterial genome. Currently, the database associated with this web server contains information from bacteria, viruses, zebrafish, yeast and humans. However, it is eventually envisaged to contain information on all available species.

Note 2: In this document we use the terms “gene list” and “sample” interchangeably. Also, the term “GO enrichment” means GO enrichment at the protein level. The enrichment method is explained in the publication describing the Comparative GO database.

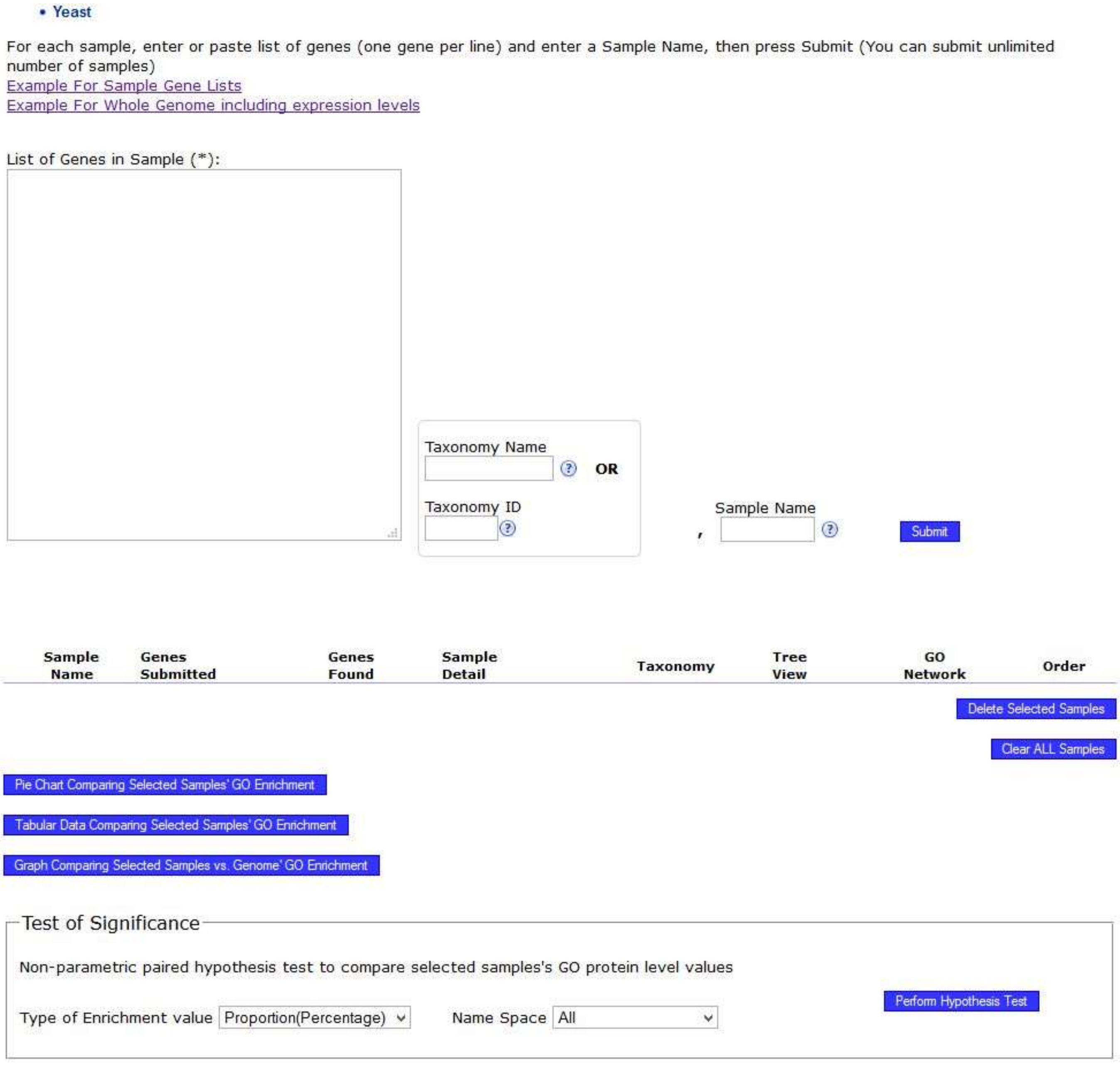
Note 3: Figures in this manual may not contain real and accurate data.

For the sample data we have used Streptococcus pneumoniae TIGR4 ordered locus gene names where genes have been over-expressed in three samples. Actual data is presented at the bottom of this manual so that researchers can copy and paste this into the website to reproduce the analyses in the manual. Additional examples of whole genome gene lists are available and accessible from the web server home page.

Home Page

Submitting a gene list or sample

When visiting the home page of Comparative GO initially, it appears as in Figure 1:



- Yeast

For each sample, enter or paste list of genes (one gene per line) and enter a Sample Name, then press Submit (You can submit unlimited number of samples)

Example For Sample Gene Lists

Example For Whole Genome including expression levels

List of Genes in Sample (*):

Sample Name	Genes Submitted	Genes Found	Sample Detail	Taxonomy	Tree View	GO Network	Order

Test of Significance

Non-parametric paired hypothesis test to compare selected samples's GO protein level values

Type of Enrichment value: Proportion(Percentage)

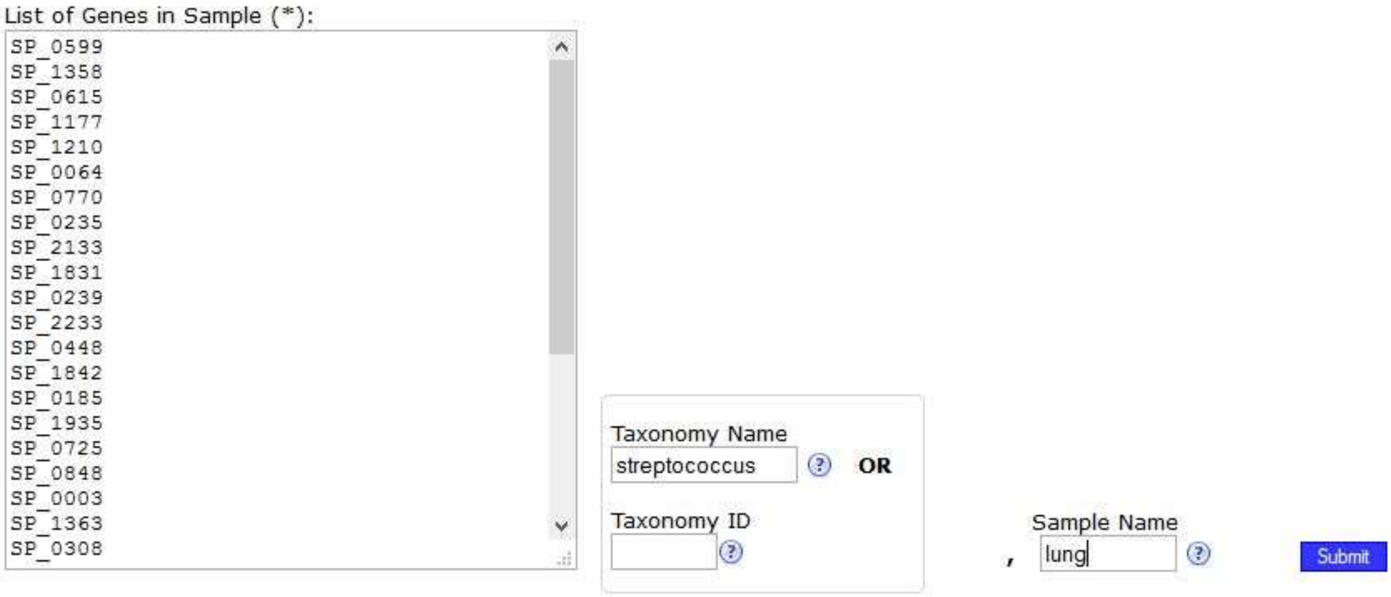
Name Space: All

Taxonomy Name
?
OR
Taxonomy ID
?
,
Sample Name
?
Submit
Delete Selected Samples
Clear ALL Samples
Pie Chart Comparing Selected Samples' GO Enrichment
Tabular Data Comparing Selected Samples' GO Enrichment
Graph Comparing Selected Samples vs. Genome' GO Enrichment
Perform Hypothesis Test

Figure 1

This website can accept lists of genes in samples in two formats. For the first format, genes are entered one per line in the “List of Genes in Sample” box provided as shown in Figure 2A. In the second format, genes and their expression level coefficients are entered in one line separated by a Space or Tab character (i.e. blank space character), as shown in Figure 2B. If expression levels are available for genes, it is better to include these expression levels since this produces more accurate reports. Optimally, expression levels should be normalised. Expression levels can be taken from microarray results or RNA-Seq RPKM/FPKM counts. If a gene does not have an expression level, the system will assume its coefficient equals 1. Thus, the system can accept a list that includes genes with and without expression levels.

To produce Figure 2A we selected a list of genes from a text editor in multiline format where there was one gene per line. The selection was then copied and pasted into the “List of Genes in Sample” box. A meaningful name was entered into the “Sample Name” box, (i.e. in this case “lung” was entered) before clicking on “Submit”. In our example we submitted an additional two samples that we named “brain” and “blood”.



List of Genes in Sample (*):

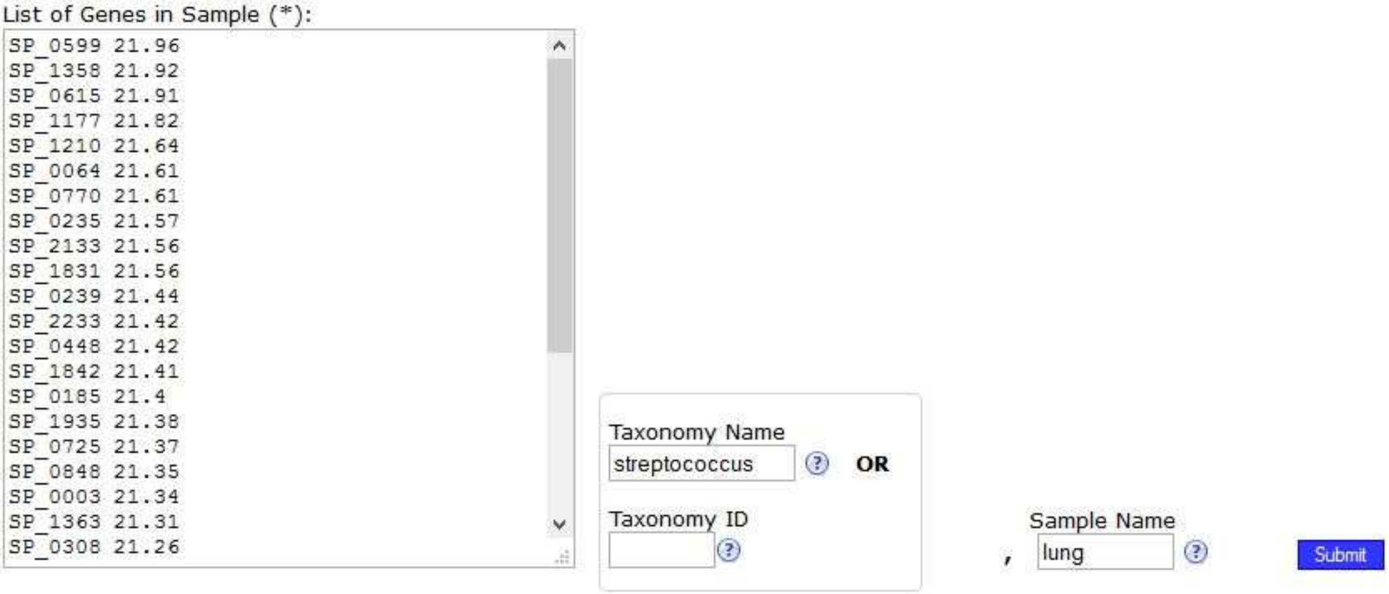
```
SP_0599
SP_1358
SP_0615
SP_1177
SP_1210
SP_0064
SP_0770
SP_0235
SP_2133
SP_1831
SP_0239
SP_2233
SP_0448
SP_1842
SP_0185
SP_1935
SP_0725
SP_0848
SP_0003
SP_1363
SP_0308
```

Taxonomy Name: streptococcus OR

Taxonomy ID:

Sample Name: lung

Figure 2A



List of Genes in Sample (*):

```
SP_0599 21.96
SP_1358 21.92
SP_0615 21.91
SP_1177 21.82
SP_1210 21.64
SP_0064 21.61
SP_0770 21.61
SP_0235 21.57
SP_2133 21.56
SP_1831 21.56
SP_0239 21.44
SP_2233 21.42
SP_0448 21.42
SP_1842 21.41
SP_0185 21.4
SP_1935 21.38
SP_0725 21.37
SP_0848 21.35
SP_0003 21.34
SP_1363 21.31
SP_0308 21.26
```

Taxonomy Name: streptococcus OR

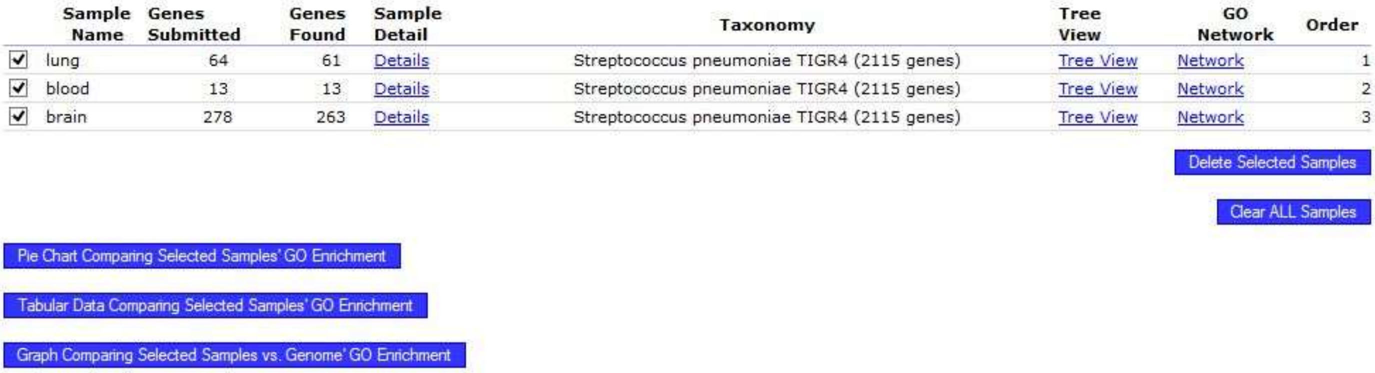
Taxonomy ID:

Sample Name: lung

Figure 2B

There is no limit to the number of samples that can be submitted for comparison to each other. If a unique Sample Name is not provided for a submitted list of genes then the system will choose a sequential name (Sample 1,2,3...) for it automatically. The order of list submission is not important. As we explain later, samples can be selected in any order that is suitable for biological study and comparisons.

After each sample submission, information is provided on the samples as shown in Figure 3. For each list, the first column is a check box for selecting that sample for further processing, the second column is the sample name itself, the third column shows the number of genes have been submitted in the sample, the fourth column shows the number of corresponding genes found in the database, the fifth column contains a link to another page providing details about the sample, the sixth column shows the taxonomy name of the organism to which the genes belong as well as the total number genes in the whole genome of that organism, the seventh column gives a link to the gene ontology tree for that sample, the eighth column contains a link to build and draw regulatory networks between genes and GO groups of the sample and, finally, the last column shows the order of sample selection or the order in which the check boxes in the first column were clicked to select the samples.



	Sample Name	Genes Submitted	Genes Found	Sample Detail	Taxonomy	Tree View	GO Network	Order
☑	lung	64	61	Details	Streptococcus pneumoniae TIGR4 (2115 genes)	Tree View	Network	1
☑	blood	13	13	Details	Streptococcus pneumoniae TIGR4 (2115 genes)	Tree View	Network	2
☑	brain	278	263	Details	Streptococcus pneumoniae TIGR4 (2115 genes)	Tree View	Network	3

Delete Selected Samples
Clear ALL Samples
Pie Chart Comparing Selected Samples' GO Enrichment
Tabular Data Comparing Selected Samples' GO Enrichment
Graph Comparing Selected Samples vs. Genome' GO Enrichment

Figure 3

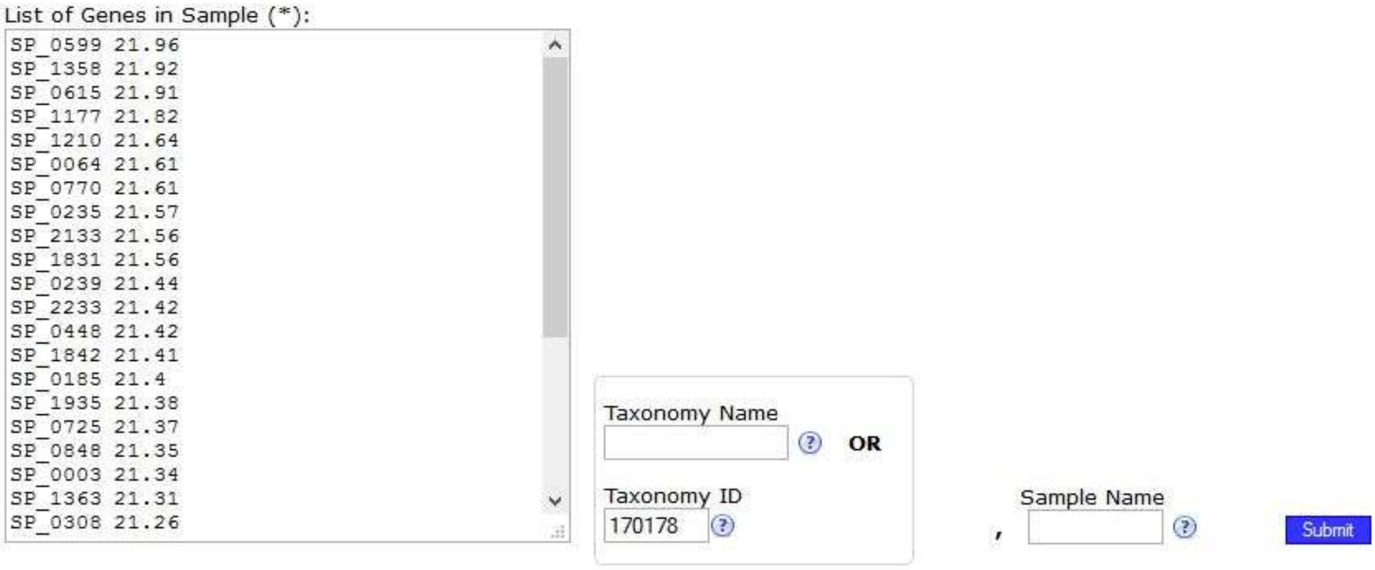
Taxonomy selection mechanism

Each taxonomy ID has its own Gene Ontology records in the database. In species that have many strains, each strain has its own GO records. I that case, selecting a strain with the largest number of GO records may give a more informative analysis. To improve taxonomy selection, the system follows two different strategies.

The first strategy is employed when a user is uncertain about the taxonomy ID but knows part of the taxonomy name. For example, if a user wants to submit a list of genes related to the Streptococcus bacterial genus, they can leave the Taxonomy ID blank and type “streptococcus” under Taxonomy Name field as shown in Figures 2A and 2B. If the specified taxonomy name has multiple strains in the database, then the system will use the taxonomy that has the greatest number of matching genes in the database. In our example, the system finds the Streptococcus strain “Streptococcus pneumoniae TIGR4” (taxonomy id 170187).

The system will accept non-scientific names for model organisms such as; “human”, “zebrafish” and “yeast”. If the user enters a name that belongs to the top of a taxonomy tree (i.e. too general a name), then the system must search many records in the database and, potentially, will take longer to find best taxonomy.

The second strategy is employed when a user is certain about the taxonomy ID. The taxonomy ID is entered at the time of submission as shown in Figure 4.



List of Genes in Sample (*):

```
SP_0599 21.96
SP_1358 21.92
SP_0615 21.91
SP_1177 21.82
SP_1210 21.64
SP_0064 21.61
SP_0770 21.61
SP_0235 21.57
SP_2133 21.56
SP_1831 21.56
SP_0239 21.44
SP_2233 21.42
SP_0448 21.42
SP_1842 21.41
SP_0185 21.4
SP_1935 21.38
SP_0725 21.37
SP_0848 21.35
SP_0003 21.34
SP_1363 21.31
SP_0308 21.26
```

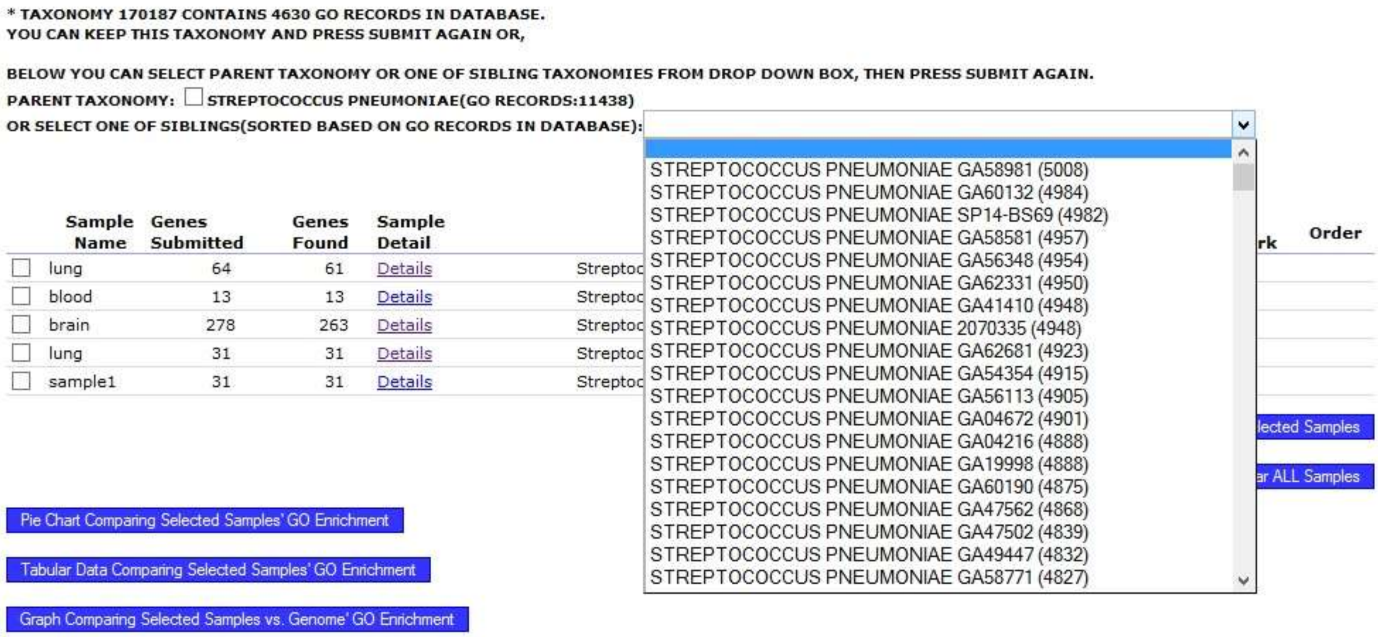
Taxonomy Name: OR

Taxonomy ID: 170178

Sample Name:

Figure 4

The user is then shown a number of GO records from the parent taxonomy (in our example, the parent taxonomy is 1313), and is also shown GO records for all the strains in the family (i.e. sibling strains) as in Figure 5.



* TAXONOMY 170187 CONTAINS 4630 GO RECORDS IN DATABASE.
YOU CAN KEEP THIS TAXONOMY AND PRESS SUBMIT AGAIN OR,

BELOW YOU CAN SELECT PARENT TAXONOMY OR ONE OF SIBLING TAXONOMIES FROM DROP DOWN BOX, THEN PRESS SUBMIT AGAIN.
PARENT TAXONOMY: □ STREPTOCOCCUS PNEUMONIAE(GO RECORDS:11438)
OR SELECT ONE OF SIBLINGS(SORTED BASED ON GO RECORDS IN DATABASE):

STREPTOCOCCUS PNEUMONIAE GA58981 (5008)
STREPTOCOCCUS PNEUMONIAE GA60132 (4984)
STREPTOCOCCUS PNEUMONIAE SP14-BS69 (4982)
STREPTOCOCCUS PNEUMONIAE GA58581 (4957)
STREPTOCOCCUS PNEUMONIAE GA56348 (4954)
STREPTOCOCCUS PNEUMONIAE GA62331 (4950)
STREPTOCOCCUS PNEUMONIAE GA41410 (4948)
STREPTOCOCCUS PNEUMONIAE 2070335 (4948)
STREPTOCOCCUS PNEUMONIAE GA62681 (4923)
STREPTOCOCCUS PNEUMONIAE GA54354 (4915)
STREPTOCOCCUS PNEUMONIAE GA56113 (4905)
STREPTOCOCCUS PNEUMONIAE GA04672 (4901)
STREPTOCOCCUS PNEUMONIAE GA04216 (4888)
STREPTOCOCCUS PNEUMONIAE GA19998 (4888)
STREPTOCOCCUS PNEUMONIAE GA60190 (4875)
STREPTOCOCCUS PNEUMONIAE GA47562 (4868)
STREPTOCOCCUS PNEUMONIAE GA47502 (4839)
STREPTOCOCCUS PNEUMONIAE GA49447 (4832)
STREPTOCOCCUS PNEUMONIAE GA58771 (4827)

	Sample Name	Genes Submitted	Genes Found	Sample Detail		rk	Order
□	lung	64	61	Details	Streptoc		
□	blood	13	13	Details	Streptoc		
□	brain	278	263	Details	Streptoc		
□	lung	31	31	Details	Streptoc		
□	sample1	31	31	Details	Streptoc		

lected Samples
ar ALL Samples
Pie Chart Comparing Selected Samples' GO Enrichment
Tabular Data Comparing Selected Samples' GO Enrichment
Graph Comparing Selected Samples vs. Genome' GO Enrichment

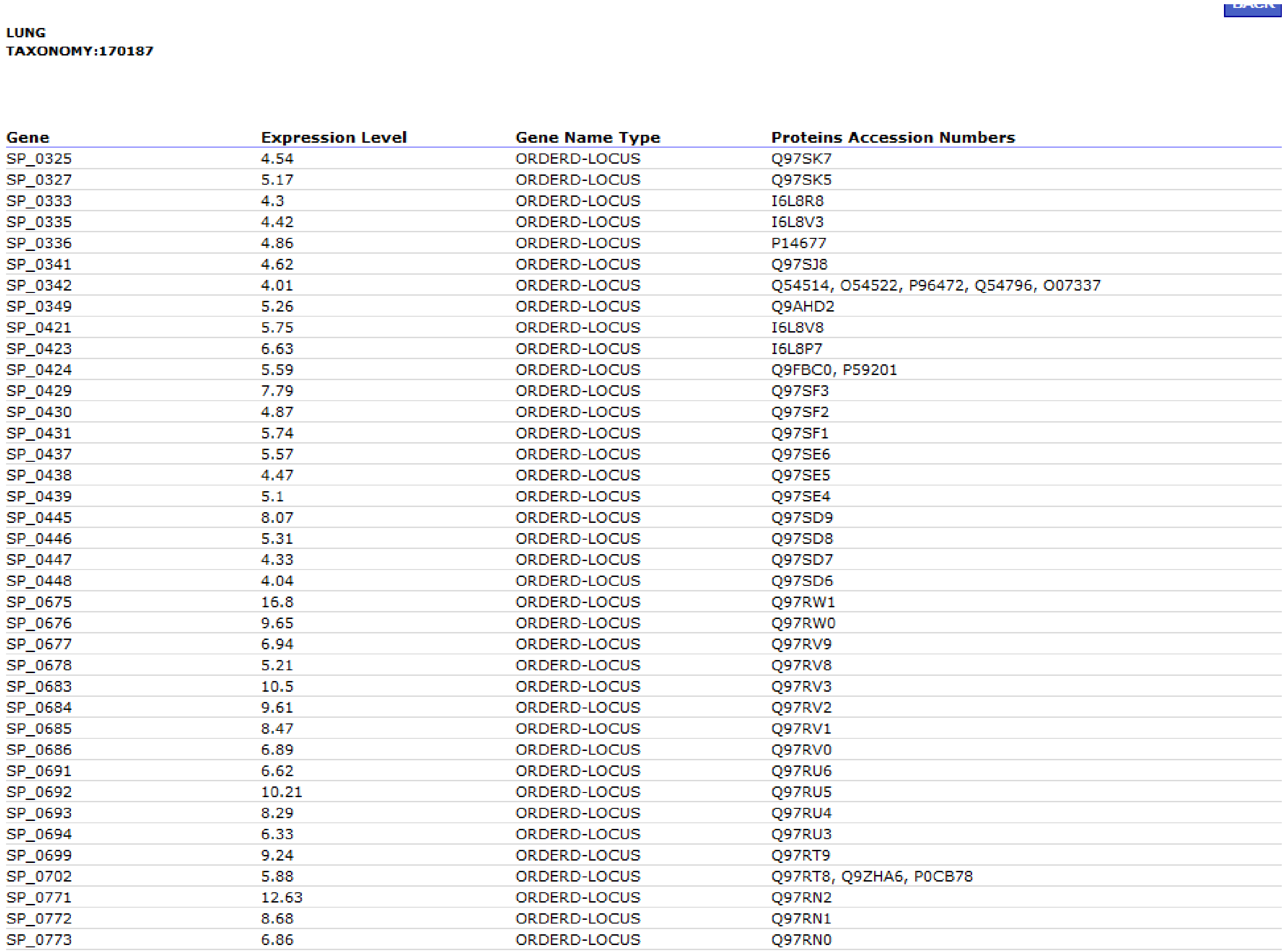
Figure 5

At this stage the user can select the parent taxonomy checkbox or any of the siblings in the drop down box. If the user does not select any of them (i.e. indicating that they are happy with our own taxonomy id), then the user can just click on the submit button again and the system will proceed using the original taxonomy id.

Details of a Sample

In the example shown, the user wished to see details of the “lung” sample. Clicking on the Details link took them to a page such as that shown in Figure 6. As one can see in the image, the page shows the list of found genes along with their expression level, type of gene name and finally the accession numbers of the proteins that are produced by that gene. In this system the “Gene Name Type” can be one of five possibilities: (1) Primary (2) Synonym (3) Ordered Locus (4) ORF or (5) Uniprot Protein Accession ID

If some genes were not found in the database, they will be listed at the end of the ‘Detail’ page.



LUNG
TAXONOMY:170187

Gene	Expression Level	Gene Name Type	Proteins Accession Numbers
SP_0325	4.54	ORDERD-LOCUS	Q97SK7
SP_0327	5.17	ORDERD-LOCUS	Q97SK5
SP_0333	4.3	ORDERD-LOCUS	I6L8R8
SP_0335	4.42	ORDERD-LOCUS	I6L8V3
SP_0336	4.86	ORDERD-LOCUS	P14677
SP_0341	4.62	ORDERD-LOCUS	Q97SJ8
SP_0342	4.01	ORDERD-LOCUS	Q54514, O54522, P96472, Q54796, O07337
SP_0349	5.26	ORDERD-LOCUS	Q9AHD2
SP_0421	5.75	ORDERD-LOCUS	I6L8V8
SP_0423	6.63	ORDERD-LOCUS	I6L8P7
SP_0424	5.59	ORDERD-LOCUS	Q9FBC0, P59201
SP_0429	7.79	ORDERD-LOCUS	Q97SF3
SP_0430	4.87	ORDERD-LOCUS	Q97SF2
SP_0431	5.74	ORDERD-LOCUS	Q97SF1
SP_0437	5.57	ORDERD-LOCUS	Q97SE6
SP_0438	4.47	ORDERD-LOCUS	Q97SE5
SP_0439	5.1	ORDERD-LOCUS	Q97SE4
SP_0445	8.07	ORDERD-LOCUS	Q97SD9
SP_0446	5.31	ORDERD-LOCUS	Q97SD8
SP_0447	4.33	ORDERD-LOCUS	Q97SD7
SP_0448	4.04	ORDERD-LOCUS	Q97SD6
SP_0675	16.8	ORDERD-LOCUS	Q97RW1
SP_0676	9.65	ORDERD-LOCUS	Q97RW0
SP_0677	6.94	ORDERD-LOCUS	Q97RV9
SP_0678	5.21	ORDERD-LOCUS	Q97RV8
SP_0683	10.5	ORDERD-LOCUS	Q97RV3
SP_0684	9.61	ORDERD-LOCUS	Q97RV2
SP_0685	8.47	ORDERD-LOCUS	Q97RV1
SP_0686	6.89	ORDERD-LOCUS	Q97RV0
SP_0691	6.62	ORDERD-LOCUS	Q97RU6
SP_0692	10.21	ORDERD-LOCUS	Q97RU5
SP_0693	8.29	ORDERD-LOCUS	Q97RU4
SP_0694	6.33	ORDERD-LOCUS	Q97RU3
SP_0699	9.24	ORDERD-LOCUS	Q97RT9
SP_0702	5.88	ORDERD-LOCUS	Q97RT8, Q9ZHA6, P0CB78
SP_0771	12.63	ORDERD-LOCUS	Q97RN2
SP_0772	8.68	ORDERD-LOCUS	Q97RN1
SP_0773	6.86	ORDERD-LOCUS	Q97RN0

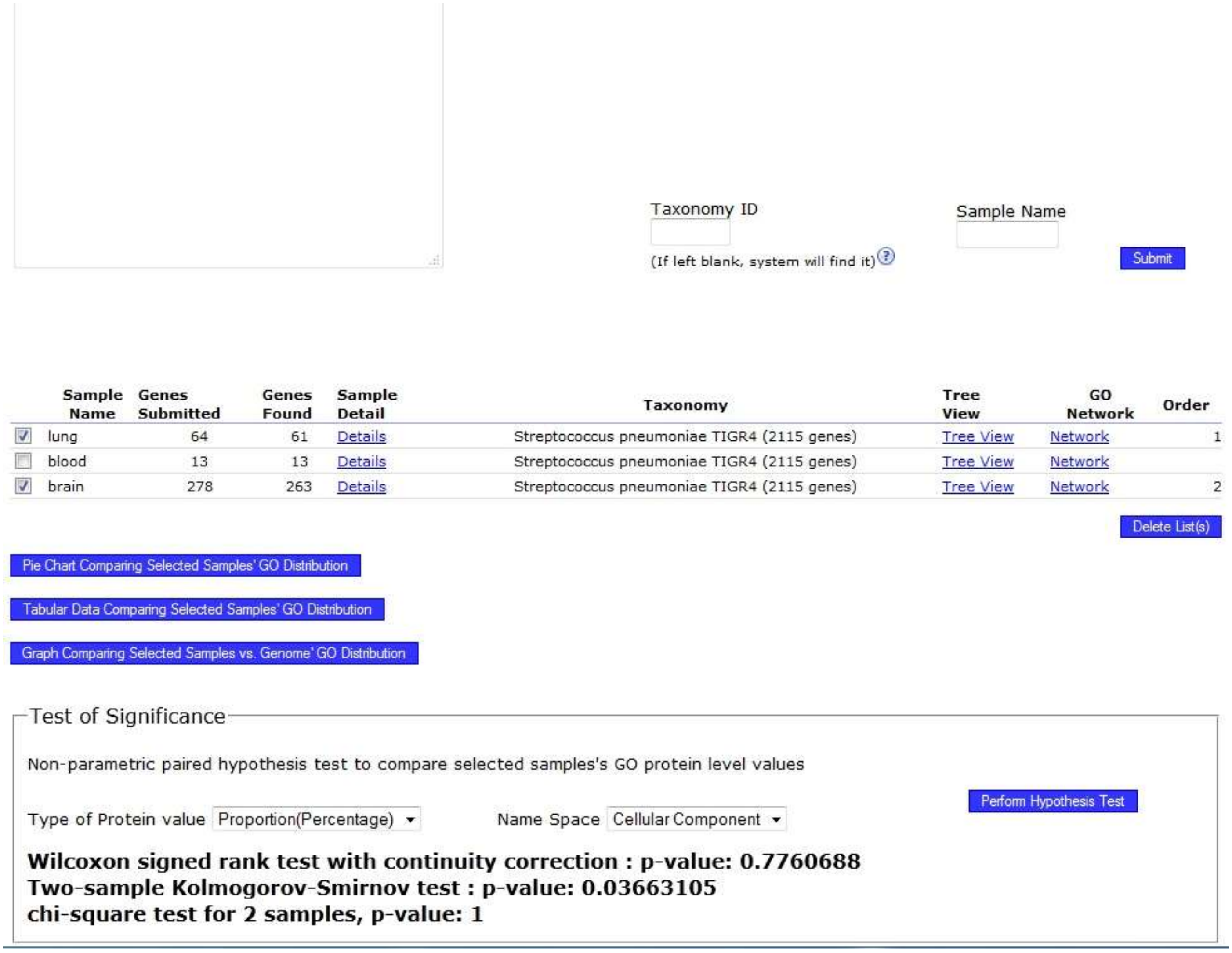
Figure 6

Hypothesis test to compare overall Samples’ GO distribution

In order to compare the overall GO distribution between 2 or more samples, the system provides a non-parametric hypothesis testing tool on the home page. In general, we cannot assume that the distribution of GO groups in a sample is normal. As shown in Figure 7, we have selected 2 samples (lung and brain). At the bottom of the home page there is a group box titled ‘Test of Significance ’. Inside that, there are two drop-down boxes. First one can choose the type of GO enrichment values (original protein level or percentage). It is important to note that the original protein levels of GO groups between two samples can change significantly but that their percentages do not. The second drop-down box can limit the GO name space. It is a good practice to perform hypothesis testing in a separate name space, because the GO enrichments of each name space can change independently.

After selecting samples and parameters, the user clicks the ‘Perform Hypothesis Test’ button. The results of 3 different tests are shown, the Wilcoxon signed rank test, the KS paired test and the chi-square test for 2 samples. It is important to note that the null hypothesis assumes that the GO enrichment distributions are the same or have not changed significantly.

Unlike the overall GO distribution hypothesis test, we can perform hypothesis tests for each individual GO group using another part of the system explained in the section ‘Comparing Sample versus Genome’.



Taxonomy ID
(If left blank, system will find it)
Sample Name
Submit

	Sample Name	Genes Submitted	Genes Found	Sample Detail	Taxonomy	Tree View	GO Network	Order
☑	lung	64	61	Details	Streptococcus pneumoniae TIGR4 (2115 genes)	Tree View	Network	1
□	blood	13	13	Details	Streptococcus pneumoniae TIGR4 (2115 genes)	Tree View	Network	
☑	brain	278	263	Details	Streptococcus pneumoniae TIGR4 (2115 genes)	Tree View	Network	2

Delete List(s)
Pie Chart Comparing Selected Samples' GO Distribution
Tabular Data Comparing Selected Samples' GO Distribution
Graph Comparing Selected Samples vs. Genome' GO Distribution
Test of Significance

Non-parametric paired hypothesis test to compare selected samples's GO protein level values

Type of Protein value: Proportion(Percentage)

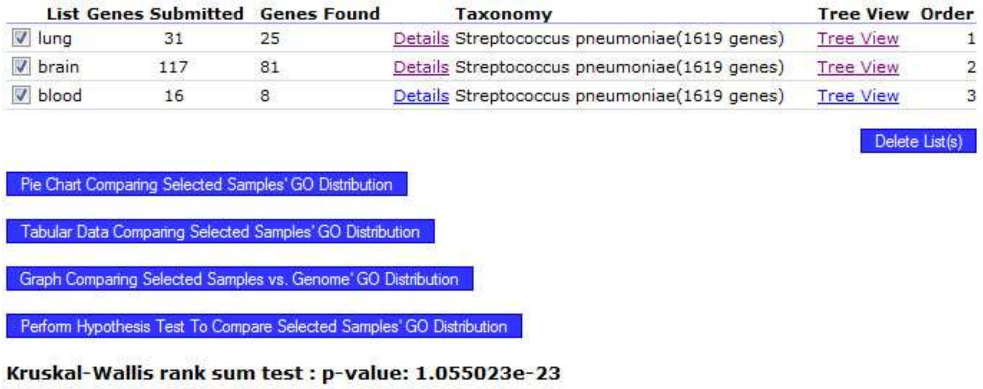
Name Space: Cellular Component

Wilcoxon signed rank test with continuity correction : p-value: 0.7760688
Two-sample Kolmogorov-Smirnov test : p-value: 0.03663105
chi-square test for 2 samples, p-value: 1

Figure 7

Interpretation of p-values depends on levels of significance.

In the next example the user selects multiple (three) samples. A non-parametric test is then performed with results similar to what is seen in Figure 6.



	List	Genes Submitted	Genes Found		Taxonomy	Tree View	Order
☑	lung	31	25	Details	Streptococcus pneumoniae(1619 genes)	Tree View	1
☑	brain	117	81	Details	Streptococcus pneumoniae(1619 genes)	Tree View	2
☑	blood	16	8	Details	Streptococcus pneumoniae(1619 genes)	Tree View	3

Delete List(s)
Pie Chart Comparing Selected Samples' GO Distribution
Tabular Data Comparing Selected Samples' GO Distribution
Graph Comparing Selected Samples vs. Genome' GO Distribution
Perform Hypothesis Test To Compare Selected Samples' GO Distribution

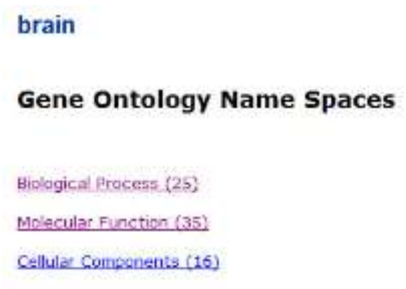
Kruskal-Wallis rank sum test : p-value: 1.055023e-23

Figure 8

Tree view Presentation of Gene Ontology

Gene ontology groups build classic tree structures. The Comparative GO system can present such structures and GO enrichments levels.

For example clicking on the ‘Tree View’ link of the sample ‘brain’ leads us an intermediate page with three links, MF, BP and CC, as shown in Figure 9. The numbers in parentheses next to each gene ontology represent the GO enrichment of that gene ontology name space.



brain

Gene Ontology Name Spaces

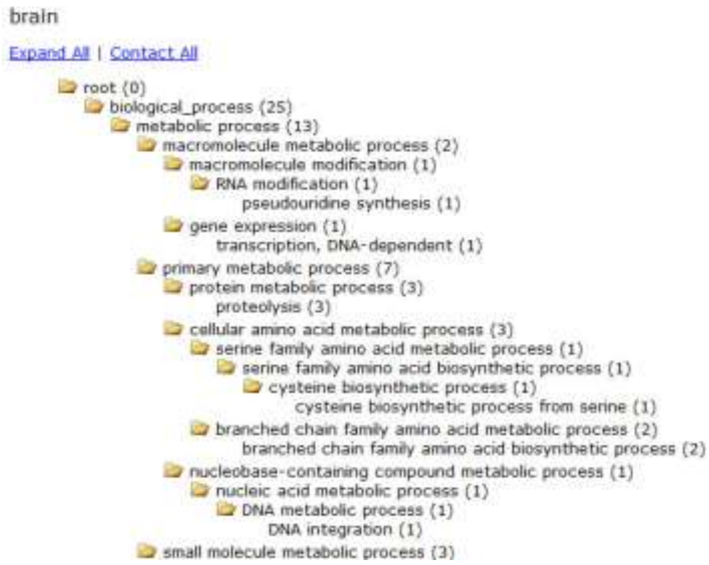
Biological Process (25)

Molecular Function (35)

Cellular Components (16)

Figure 9

Next, clicking on BP gives a display of the BP tree structure as shown in Figure 10. On that page one can selectively collapse any gene ontology group by clicking on its folder icon in order to see its sub groups. In each gene ontology group, there is number that shows the protein enrichment of that group. To collapse all nodes one can click Expand All and to close all nodes one can click Contact All.



brain

Expand All | Contact All

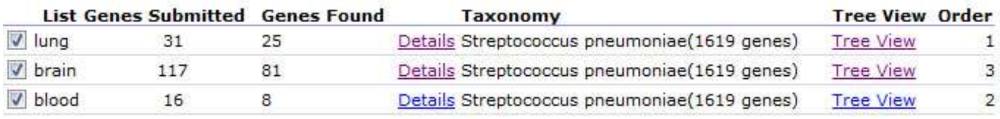
```
root (0)
  biological_process (25)
    metabolic process (13)
      macromolecule metabolic process (2)
        macromolecule modification (1)
          RNA modification (1)
            pseudouridine synthesis (1)
        gene expression (1)
          transcription, DNA-dependent (1)
      primary metabolic process (7)
        protein metabolic process (3)
          proteolysis (3)
        cellular amino acid metabolic process (3)
          serine family amino acid metabolic process (1)
            serine family amino acid biosynthetic process (1)
              cysteine biosynthetic process (1)
                cysteine biosynthetic process from serine (1)
          branched chain family amino acid metabolic process (2)
            branched chain family amino acid biosynthetic process (2)
        nucleobase-containing compound metabolic process (1)
          nucleic acid metabolic process (1)
            DNA metabolic process (1)
              DNA integration (1)
      small molecule metabolic process (3)
```

Figure 10

Note: At any stage, one can return to the home page by clicking the back button at the top right hand corner of the page.

Comparing multiple samples

To select multiple samples on the home page, just click on the left most column checkbox. The order in which check boxes are clicked is shown in the right hand column. One may wish to select samples in a specific order for reasons such as the samples are related to different time stamps, or the samples are from different tissues where metabolic pathways are in a specific order. Figure 11 shows an example of selection order, where a user first clicked on ‘lung’ then ‘blood’ and finally ‘brain’.



	List	Genes Submitted	Genes Found	Taxonomy	Tree View	Order
☑	lung	31	25	Details Streptococcus pneumoniae(1619 genes)	Tree View	1
☑	brain	117	81	Details Streptococcus pneumoniae(1619 genes)	Tree View	3
☑	blood	16	8	Details Streptococcus pneumoniae(1619 genes)	Tree View	2

Figure 11

Gene Ontology Pie Chart

An informative method for comparing multiple samples is to compare their GO enrichment proportions side by side. Comparative GO provides comprehensive and interactive tools to perform these comparisons. At the home page, samples are selected in the desired order, and then the blue button ‘Pie Chart Comparing Selected Samples’ GO Distribution’ is clicked. This opens a page similar to that shown in Figure 12.

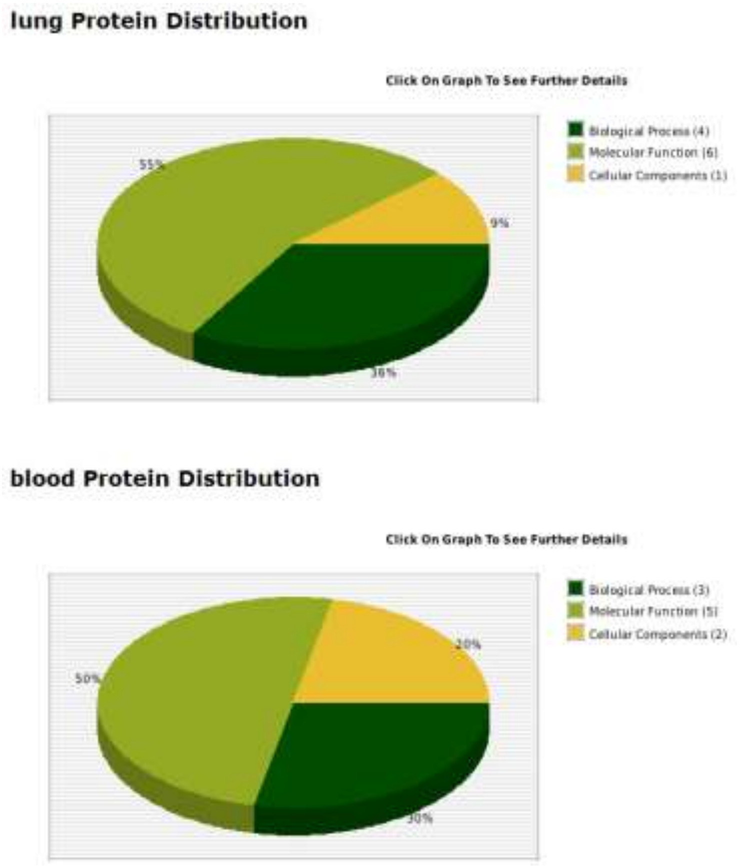
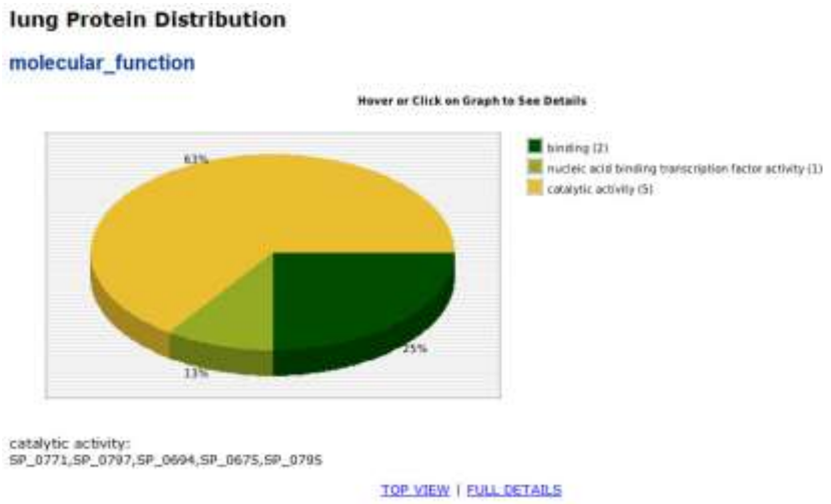


Figure 12

Figure 12 shows a top view comparison of GO name spaces among samples with enrichment levels stated in the legend and percentages shown on the pie chart itself. Each slice of the pie chart is mouse sensitive. When the mouse is hovered over a slice the genes involved are revealed. If one clicks on a slice further details are given of the related gene ontology (one level deeper in the GO tree as shown in Figure 15a). For example, if MF of the ‘lung’ pie chart is clicked that takes the user to something similar to Figure 13.

Note: Remember that each pie chart updates independently, and it does not update the whole page. (This makes navigation more rapid.)

In Figure 13, the top pie chart presents GO groups of Molecular Function for the ‘lung’ sample beside their GO enrichment level. If one hovers the mouse over a slice (e.g. Catalytic Activity in Figure 13), then a list of genes participating in ‘Catalytic Activity’ is revealed underneath the pie chart.



lung Protein Distribution

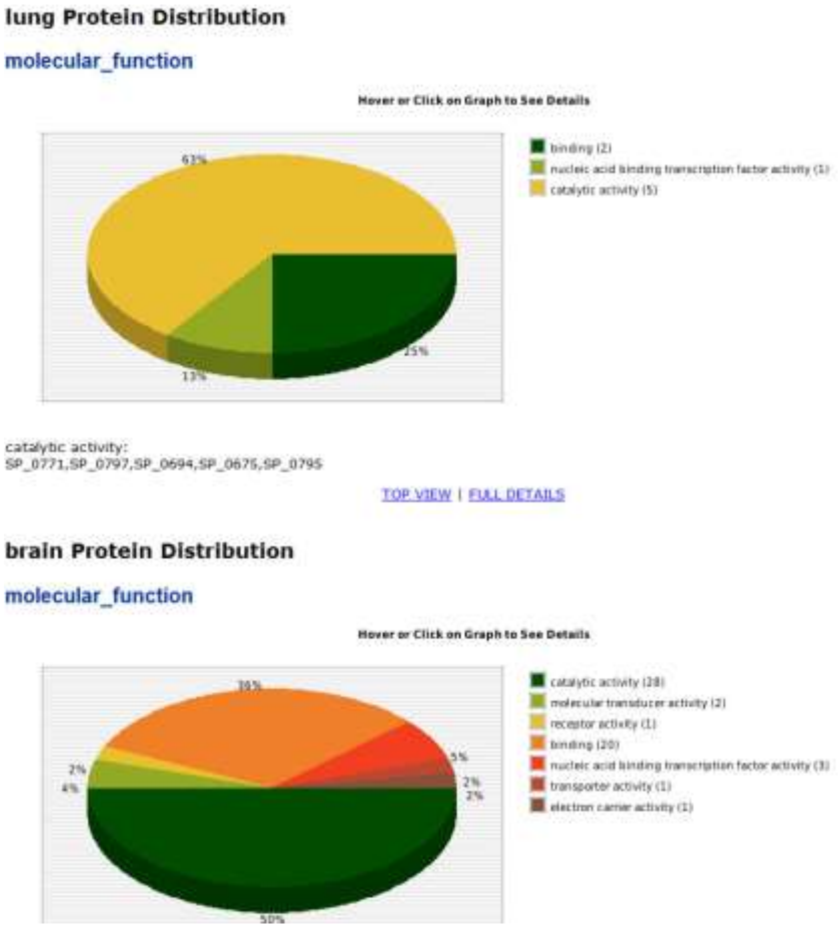
molecular_function

Hover or Click on Graph to See Details

catalytic activity:
SP_0771,SP_0797,SP_0694,SP_0675,SP_0795

TOP VIEW | FULL DETAILS

Figure 13



lung Protein Distribution

molecular_function

Hover or Click on Graph to See Details

63%
25%
13%
🟩 binding (2)
🟩 nucleic acid binding transcription factor activity (1)
🟨 catalytic activity (5)

catalytic activity:
SP_0771,SP_0797,SP_0694,SP_0675,SP_0795

TOP VIEW | FULL DETAILS

brain Protein Distribution

molecular_function

Hover or Click on Graph to See Details

36%
5%
2%
2%
2%
4%
50%
🟩 catalytic activity (28)
🟩 molecular transducer activity (2)
🟨 receptor activity (1)
🟧 binding (20)
🟥 nucleic acid binding transcription factor activity (3)
🟥 transporter activity (1)
🟫 electron carrier activity (1)

Figure 14

As a second example we shown the Molecular function of ‘lung’ and ‘blood’ samples together and by clicking the MF slice in both pie charts, something similar to Figure 14 is shown. One can now compare the two samples to find any biological significance and also to see the genes that are related to them.

It is a good practice to compare multiple samples' GO enrichment at the same level (depth) of the GO tree.

In the next section we explain the mechanics behind Pie Chart navigation.

Navigation across Gene Ontology tree

When navigating along Pie Charts to see different levels of GO groups one is actually traversing along the GO tree from top to bottom and vice versa. A Gene Ontology tree is organized hierarchically as in Figure 15(a). An example of navigation along this tree is depicted by the grey colour nodes in figure 15(a). At any given node of the tree, one can see child nodes and their GO enrichments. From a Pie Chart point of view, at each level one clicks on a slice to see details of that slice all the way down until one reach the outermost level (the 'leaves') of the tree. Comparative GO provides up and down buttons underneath each pie chart for navigation. At each level, there is also a 'Top View' link which transfers view to the top of tree, similar to the view in Figure 12. As one can see in Figures 13 and 14, there is a link 'Full Details' underneath each chart. This link presents the most detailed GO groups of each subspace MF, BP or CC. In other words, considering the Gene Ontology tree in Figure 15(b), greyed nodes or leaves of the tree will be displayed. One example of this presentation is shown in Figure 16, where a user desires to see full details of the MF name space in the 'lung' sample.

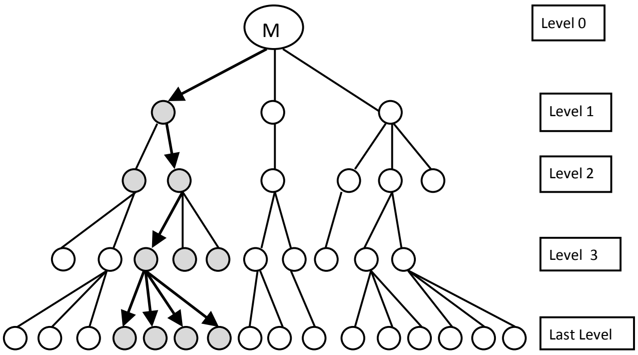


Figure 15(a)

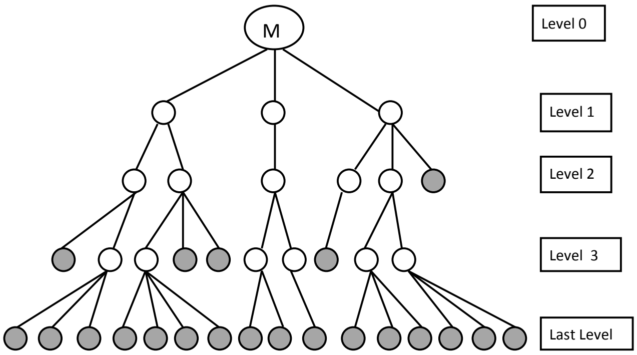
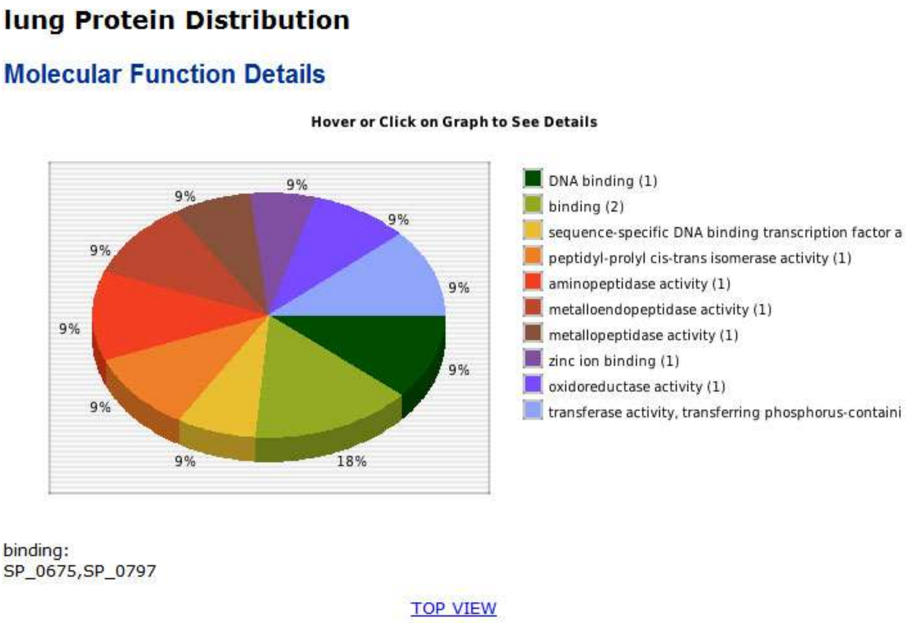


Figure 15(b) leaves are grey nodes



lung Protein Distribution

Molecular Function Details

Hover or Click on Graph to See Details

- 🟩 DNA binding (1)
- 🟩 binding (2)
- 🟨 sequence-specific DNA binding transcription factor a
- 🟧 peptidyl-prolyl cis-trans isomerase activity (1)
- 🟥 aminopeptidase activity (1)
- 🟥 metalloendopeptidase activity (1)
- 🟫 metallopeptidase activity (1)
- 🟪 zinc ion binding (1)
- 🟪 oxidoreductase activity (1)
- 🟦 transferase activity, transferring phosphorus-containi

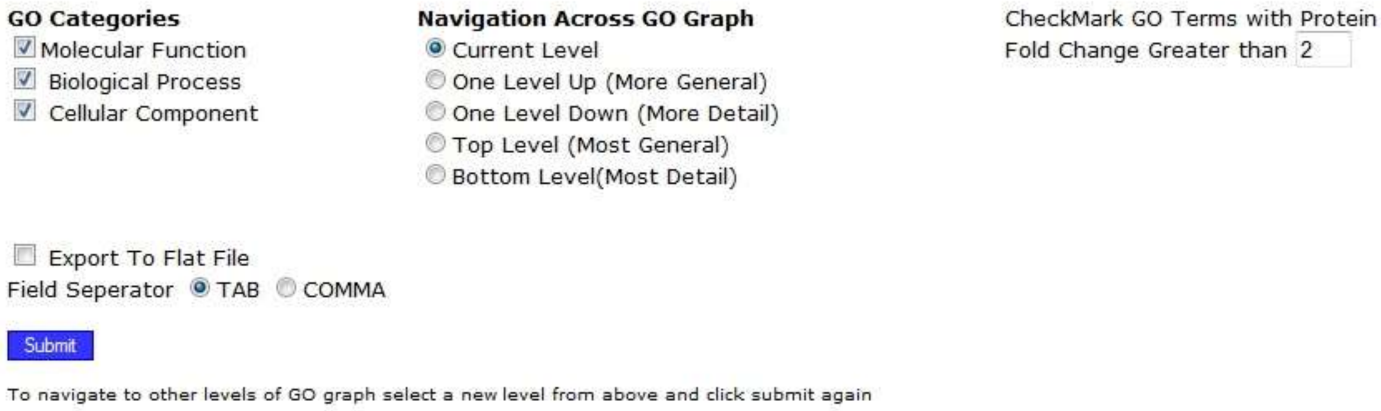
binding:
SP_0675,SP_0797

TOP VIEW

Figure 16

Tabular Gene Ontology Enrichment Analysis and Gene Selection

In this analysis, Comparative GO provides information in table format for comparing the function of genes e.g. to allow better gene selection and/or more insight into biological mechanisms. To perform this analysis, one first selects samples in the desired order at the home page, then clicks the blue button 'Tabular Data Comparing Selected Samples' GO Distribution'. This action takes the user to a page similar that shown in Figure 17.



GO Categories

☑ Molecular Function
☑ Biological Process
☑ Cellular Component

Navigation Across GO Graph

● Current Level
○ One Level Up (More General)
○ One Level Down (More Detail)
○ Top Level (Most General)
○ Bottom Level(Most Detail)

CheckMark GO Terms with Protein Fold Change Greater than 2

□ Export To Flat File
Field Seperator ● TAB ○ COMMA

Submit

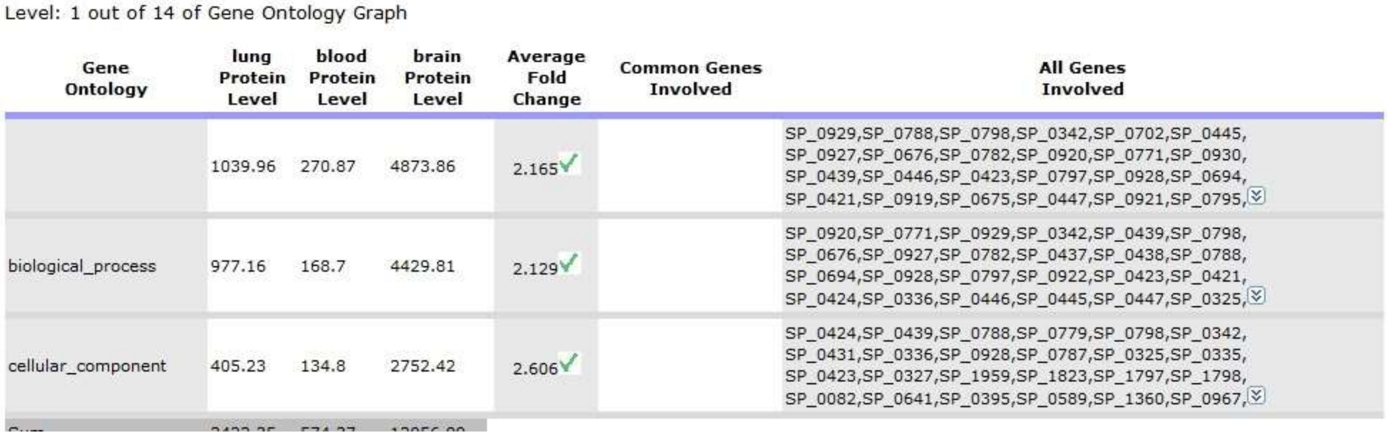
To navigate to other levels of GO graph select a new level from above and click submit again

Figure 17

Before explaining the report criteria in Figure 17, we show the output of the report obtained when the submit button is clicked as shown in Figure 18.

Columns of the Report

The first column of the table gives the gene ontology name. Then for each sample (samples are ordered based on Figure 11) there is a column showing protein the enrichment value related to that ontology. Protein enrichment values are normalized based on the size of sample (the number of genes in the sample). Smaller samples' enrichment values are scaled to higher values to make them comparable with bigger samples. Obviously, the biggest sample's enrichment values remain unchanged. When we compare whole genomes of species under multiple biological conditions, normalization is not applied because sample size is equal to the genome size for all samples. These values in the sample-specific columns are summarized in the 'Average Fold Change' Column that shows the geometric average of enrichment value fold changes among samples in that particular gene ontology. This number can give an overall idea of whether a gene ontology's enrichment value has been increased or decreased, and to what extent this change has been done between samples. This type of enrichment analysis can have significant biological meaning. For example in Figure 18, Biological Process has a 2.129 average fold change that, in fact, is a geometric average of 168.7/977.16 and 4429.81/168.7.



Level: 1 out of 14 of Gene Ontology Graph

Gene Ontology	lung Protein Level	blood Protein Level	brain Protein Level	Average Fold Change	Common Genes Involved	All Genes Involved
	1039.96	270.87	4873.86	2.165✓		SP_0929,SP_0788,SP_0798,SP_0342,SP_0702,SP_0445, SP_0927,SP_0676,SP_0782,SP_0920,SP_0771,SP_0930, SP_0439,SP_0446,SP_0423,SP_0797,SP_0928,SP_0694, SP_0421,SP_0919,SP_0675,SP_0447,SP_0921,SP_0795,
biological_process	977.16	168.7	4429.81	2.129✓		SP_0920,SP_0771,SP_0929,SP_0342,SP_0439,SP_0798, SP_0676,SP_0927,SP_0782,SP_0437,SP_0438,SP_0788, SP_0694,SP_0928,SP_0797,SP_0922,SP_0423,SP_0421, SP_0424,SP_0336,SP_0446,SP_0445,SP_0447,SP_0325,
cellular_component	405.23	134.8	2752.42	2.606✓		SP_0424,SP_0439,SP_0788,SP_0779,SP_0798,SP_0342, SP_0431,SP_0336,SP_0928,SP_0787,SP_0325,SP_0335, SP_0423,SP_0327,SP_1959,SP_1823,SP_1797,SP_1798, SP_0082,SP_0641,SP_0395,SP_0589,SP_1360,SP_0967,
Sum	2422.35	574.37	12056.09			

Figure 18

For each gene ontology, the 'Common Genes Involved' column shows genes that are present in that gene ontology among all samples. In fact, it is the intersection of the gene sets in all samples. In contrast, the 'All Genes' column shows the union of the gene sets. When each sample is equal to the whole genome, these two columns contain the same list of genes.

To save space on the page, each cell shows a maximum of 4 lines. If a user wishes to see the full list of genes they can click on the small arrow icon at the bottom right corner of the cell.

In this report one can navigate to all levels of the GO tree. At each level of the GO tree, all the nodes at that level are shown as in Figure 15(b). In contrast, in the pie chart, we can view a subset of the nodes at any given level as shown in Figure 15(a).

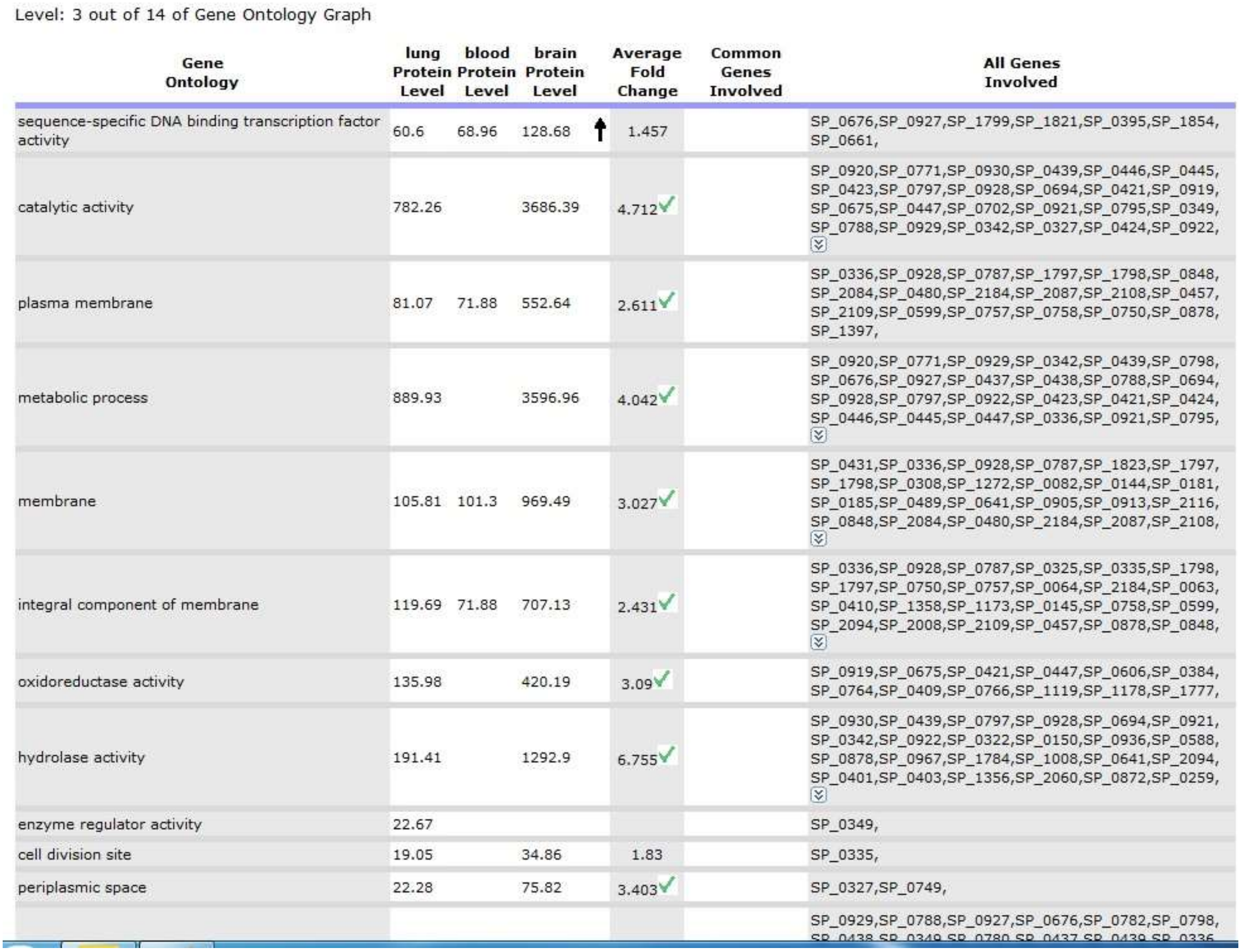
Reporting Criteria

As Figure 17, the first group of check boxes are MF, BP and CC. Deselecting each item filters out that name space from analysis so that one sees fewer items in the result table.

A second group of radio buttons are navigational buttons. 'Current Level' is a pointer that indicates the level of the GO tree being observed. When one first visits the page, Current Level is at the 'top level' (Most General) of the GO tree. 'Bottom Level' (most details) means the last level of the tree (or the leaves of tree as seen as the grey nodes in Figure 15(b)). We can also change the Current Level by selecting 'One Level Up/Down' options.

Next, there is text box with default value 2 that defines a cutoff to highlight GO groups with an average fold change more than 2 or less than (1/2). This can be helpful if we are interested in GO groups with specific fold changes.

For example, Figure 3 shows level 3 of the GO tree. GO groups with green checkmarks have significant average fold changes. GO groups with consistent increases or decreases in enrichment are highlighted by upward/downward arrows (e.g. see sequence-specific DNA binding transaction factor activity in Figure 19).



Level: 3 out of 14 of Gene Ontology Graph

Gene Ontology	lung Protein Level	blood Protein Level	brain Protein Level	Average Fold Change	Common Genes Involved	All Genes Involved
sequence-specific DNA binding transcription factor activity	60.6	68.96	128.68 ↑	1.457		SP_0676,SP_0927,SP_1799,SP_1821,SP_0395,SP_1854,SP_0661,
catalytic activity	782.26		3686.39	4.712✓		SP_0920,SP_0771,SP_0930,SP_0439,SP_0446,SP_0445,SP_0423,SP_0797,SP_0928,SP_0694,SP_0421,SP_0919,SP_0675,SP_0447,SP_0702,SP_0921,SP_0795,SP_0349,SP_0788,SP_0929,SP_0342,SP_0327,SP_0424,SP_0922,
plasma membrane	81.07	71.88	552.64	2.611✓		SP_0336,SP_0928,SP_0787,SP_1797,SP_1798,SP_0848,SP_2084,SP_0480,SP_2184,SP_2087,SP_2108,SP_0457,SP_2109,SP_0599,SP_0757,SP_0758,SP_0750,SP_0878,SP_1397,
metabolic process	889.93		3596.96	4.042✓		SP_0920,SP_0771,SP_0929,SP_0342,SP_0439,SP_0798,SP_0676,SP_0927,SP_0437,SP_0438,SP_0788,SP_0694,SP_0928,SP_0797,SP_0922,SP_0423,SP_0421,SP_0424,SP_0446,SP_0445,SP_0447,SP_0336,SP_0921,SP_0795,
membrane	105.81	101.3	969.49	3.027✓		SP_0431,SP_0336,SP_0928,SP_0787,SP_1823,SP_1797,SP_1798,SP_0308,SP_1272,SP_0082,SP_0144,SP_0181,SP_0185,SP_0489,SP_0641,SP_0905,SP_0913,SP_2116,SP_0848,SP_2084,SP_0480,SP_2184,SP_2087,SP_2108,
integral component of membrane	119.69	71.88	707.13	2.431✓		SP_0336,SP_0928,SP_0787,SP_0325,SP_0335,SP_1798,SP_1797,SP_0750,SP_0757,SP_0064,SP_2184,SP_0063,SP_0410,SP_1358,SP_1173,SP_0145,SP_0758,SP_0599,SP_2094,SP_2008,SP_2109,SP_0457,SP_0878,SP_0848,
oxidoreductase activity	135.98		420.19	3.09✓		SP_0919,SP_0675,SP_0421,SP_0447,SP_0606,SP_0384,SP_0764,SP_0409,SP_0766,SP_1119,SP_1178,SP_1777,
hydrolase activity	191.41		1292.9	6.755✓		SP_0930,SP_0439,SP_0797,SP_0928,SP_0694,SP_0921,SP_0342,SP_0922,SP_0322,SP_0150,SP_0936,SP_0588,SP_0878,SP_0967,SP_1784,SP_1008,SP_0641,SP_2094,SP_0401,SP_0403,SP_1356,SP_2060,SP_0872,SP_0259,
enzyme regulator activity	22.67					SP_0349,
cell division site	19.05		34.86	1.83		SP_0335,
periplasmic space	22.28		75.82	3.403✓		SP_0327,SP_0749,
						SP_0929,SP_0788,SP_0927,SP_0676,SP_0782,SP_0798,SP_0438,SP_0349,SP_0780,SP_0437,SP_0439,SP_0336

Figure 19

Comparing Sample versus Genome

Unlike the previous analyses that compared selected samples with each other, below we show a comparison of one sample with its genome. In this way the system can detect GO groups that are over/under represented. Graphical bar charts and a hyper geometric statistical test are employed to detect these groups.

To perform this analysis, samples are selected at home page as explained previously. One then clicks on the ‘Graph comparing Selected Samples vs. Genome GO Distribution’ blue button. In a new page one can then see bar charts that show enrichment levels related to MF, BP and CC from comparisons of the sample versus the relevant genome’s expected value for each sample, as in Figure 20.

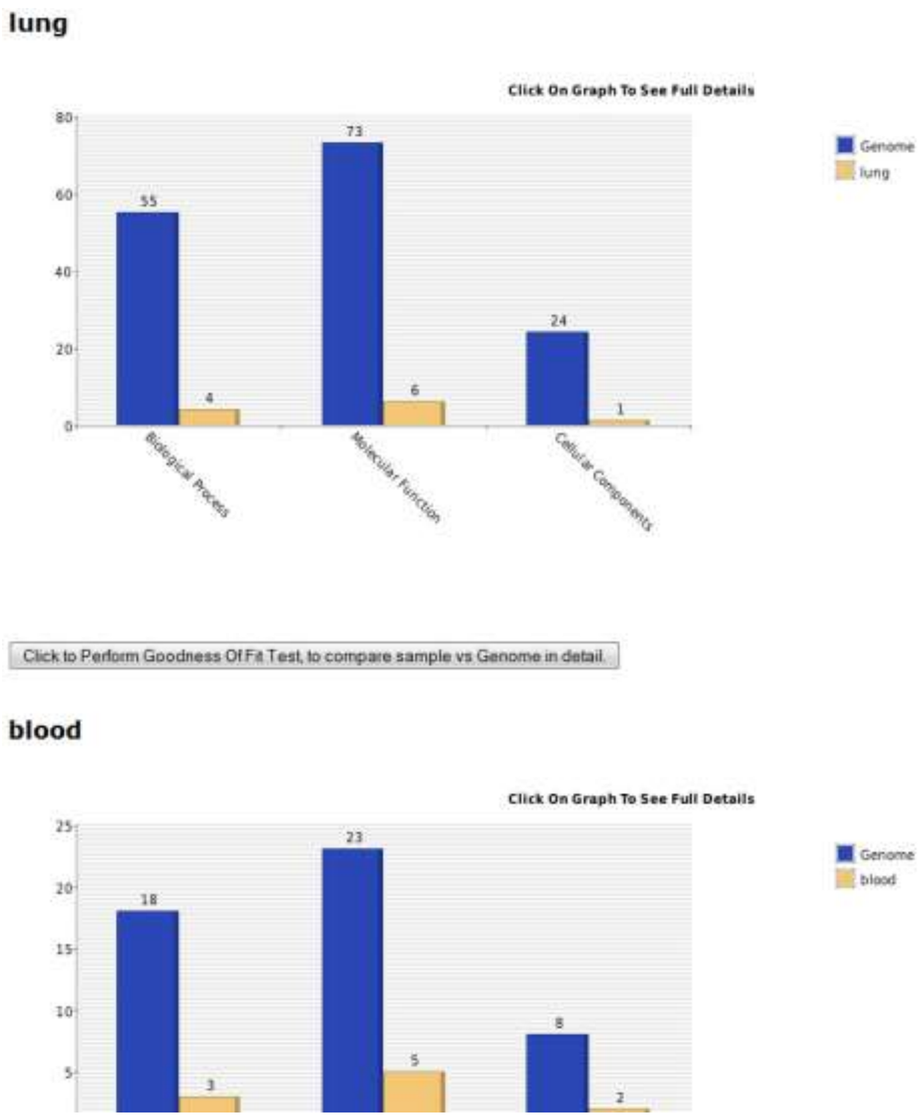

lung
Click On Graph To See Full Details
80
60
40
20
0
55
4
73
6
24
1
Biological Process
Molecular Function
Cellular Components
Genome
lung
Click to Perform Goodness Of Fit Test, to compare sample vs Genome in detail.
blood
Click On Graph To See Full Details
25
20
15
10
5
18
3
23
5
8
2
Genome
blood

Figure 20

Any significant difference in the length of the sample and genome bars is cause for further investigation. For example one may wish to know why MF in the ‘lung’ sample is so different to the genome. Therefore, one clicks on the MF bar in ‘lung’ resulting in a display such as in Figure 21.

Note: If number of GO groups exceeds 20, then Forward/Backward << and >> buttons will be provided underneath the chart to access separate pages . As for the pie charts, each sample bar chart is updated independently and the entire page is not updated when a chart is clicked on.

As seen in Figures 18 and 19, under each chart there is blue button that performs a goodness of fit hypothesis test to compare the sample’s protein distribution versus the genome’s expected protein distribution. Figure 20, shows an example result from performing this test for the ‘lung’ sample.

Navigation in each bar chart is simple and provides a top level view (as Figure 20) or bottom level (full details) view (as Figure 21) of the GO tree. (Middle levels of the GO tree are not accessible.)

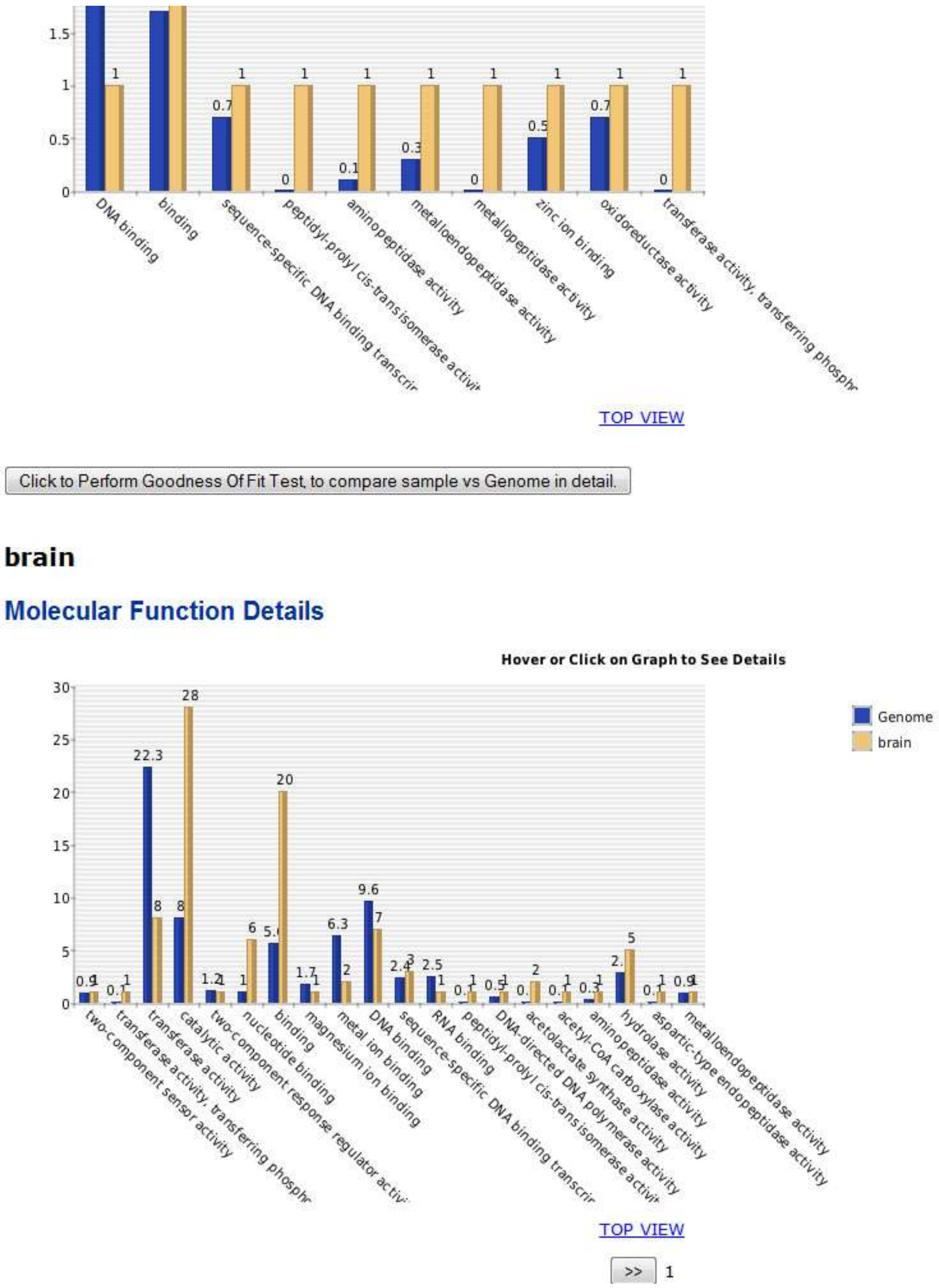
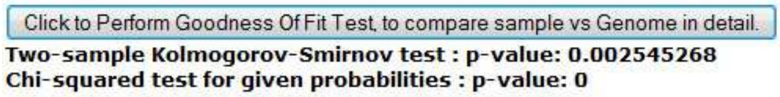

1.5
1
0.5
0
1
0.7
1
0
1
0.1
1
0.3
1
0
1
0.5
1
0.7
1
0
1
DNA binding
binding
sequence-specific DNA binding transcrip
peptidyl-prolyl cis-trans isomerase activit
aminopeptidase activity
metalloendopeptidase activity
metallopeptidase activity
zinc ion binding
oxidoreductase activity
transferase activity, transferring phosph
TOP VIEW
Click to Perform Goodness Of Fit Test, to compare sample vs Genome in detail.
brain
Molecular Function Details
Hover or Click on Graph to See Details
30
25
20
15
10
5
0
0.9 1
0.1 1
22.3 8
8 28
1.2 1
1 6
5.(20
1.7 1
6.3 2
9.6 7
2.4 3
2.5 1
0.1 1
0.5 1
0. 2
0.1 1
0.3 1
2. 5
0.1 1
0.9 1
two-component sensor activity
transferase activity, transferring phosph
transferase activity
catalytic activity
two-component response regulator activ
nucleotide binding
binding
magnesium ion binding
metal ion binding
DNA binding
sequence-specific DNA binding transcrip
RNA binding
peptidyl-prolyl cis-trans isomerase activit
DNA-directed DNA polymerase activity
acetolactate synthase activity
acetyl-CoA carboxylase activity
aminopeptidase activity
hydrolase activity
aspartic-type endopeptidase activity
metalloendopeptidase activity
Genome
brain
TOP VIEW
>> 1

Figure 21

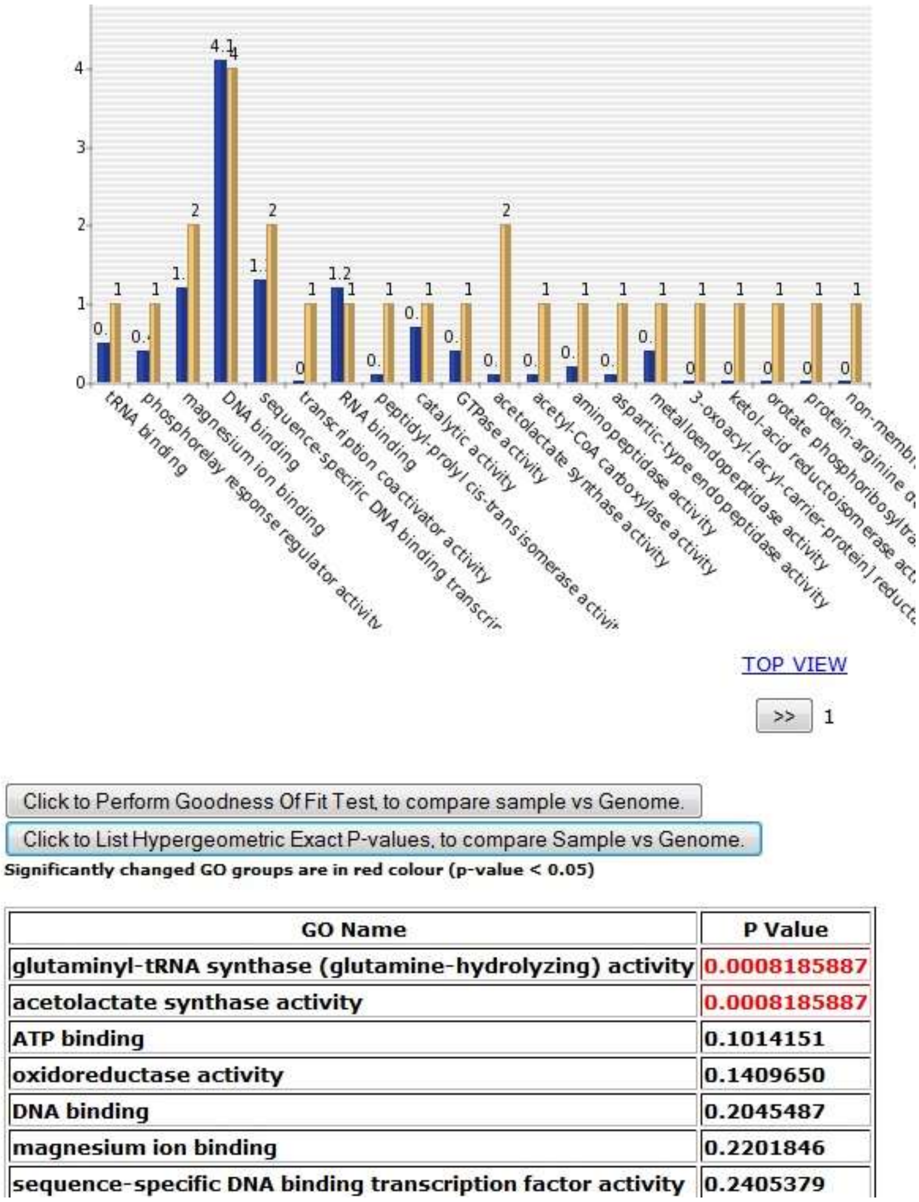
To compare the overall GO enrichment levels of a sample with its expected genome values, a Goodness of Fit test button is provided as Figure 22. The null hypothesis assumes that the sample and genome have similar values. Therefore a p-value of less than 0.05 allows rejection of the null hypothesis. This p-value does not help us to detect over/under represented GO groups in a sample, but can evaluate the overall similarity of all GO groups between sample and genome.


Click to Perform Goodness Of Fit Test, to compare sample vs Genome in detail.

Two-sample Kolmogorov-Smirnov test : p-value: 0.002545268
Chi-squared test for given probabilities : p-value: 0

Figure 22

If the number of GO groups in samples is high, the task of visually comparing bar charts (to detect over/under represented GO groups) can be tedious. Therefore, we have provided a button ‘Hyper Geometric Exact p-values’ as shown in Figure 23. To perform this test one needs to be inside a specific name space (MF BP or CC) before clicking on the button. The test cannot be performed when in the top view. As shown in Figure 23, over/under represented GO groups compared to the genome are highlighted in red and have p-value less than 0.05.


4
3
2
1
0
4.1
4
2
2
2
1.
1.
1.2
1
0.
0.4
0
tRNA binding
phosphorelay response regulator activity
magnesium ion binding
DNA binding
sequence-specific DNA binding transcrip
transcription coactivator activity
RNA binding
peptidyl-prolyl cis-trans isomerase activit
catalytic activity
GTPase activity
acetolactate synthase activity
acetyl-CoA carboxylase activity
aminopeptidase activity
aspartic-type endopeptidase activity
metalloendopeptidase activity
3-oxoacyl-[acyl-carrier-protein] reducta
ketol-acid reductoisomerase act
orotate phosphoribosyltra
protein-arginine d
non-membr
TOP VIEW
>>
1
Click to Perform Goodness Of Fit Test, to compare sample vs Genome.
Click to List Hypergeometric Exact P-values, to compare Sample vs Genome.

Significantly changed GO groups are in red colour (p-value < 0.05)

GO Name	P Value
glutaminyl-tRNA synthase (glutamine-hydrolyzing) activity	0.0008185887
acetolactate synthase activity	0.0008185887
ATP binding	0.1014151
oxidoreductase activity	0.1409650
DNA binding	0.2045487
magnesium ion binding	0.2201846
sequence-specific DNA binding transcription factor activity	0.2405379

Figure 23

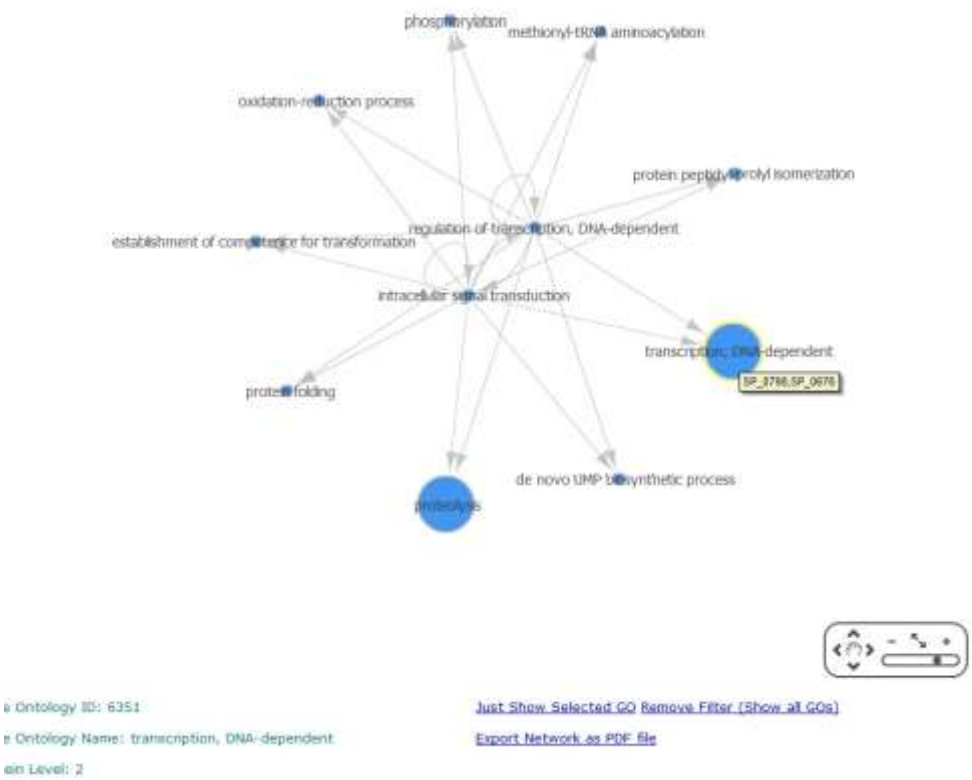
Regulatory Networks of Biological Process Groups

To generate Biological Process networks for each sample, one simply clicks on the Network link for that sample in the home page. The networks generated appear as shown in Figure 24. At first glance, the network in Figure 24 presents three types of information including: 1) regulation of one GO by another GO (grey arrows indicate any regulation, green arrows indicate up-regulation, red arrows indicate down-regulation), 2) the association of GO and genes and 3) the level of GO enrichment represented by the sizes of nodes. The topology of the network can help to reveal novel relationships between groups of genes in a biological pathway. In addition, central GO groups that are associated with genes with the highest connectivity in the network can be used for gene selection.

Note: If ones sample is too large, one should not try to produce a network for it since this may take a long time or consume a lot of memory on ones computer. Best practice is to select a subset of a genome's genes that have a particular biological significance. For example, the first 200 most highly differentially expressed genes can be selected to be analysed in a network.

When the rendered network is very large and complex one can apply a filter to the network by clicking on a specific node. This reveals a new sub network that shows only the relationships of that node with other nodes.

By hovering the mouse over a GO node, genes associated to that GO are displayed as shown in Figure 24.


phosphorylation
methionyl-tRNA aminoacylation
oxidation-reduction process
protein peptidyl-prolyl isomerization
establishment of competence for transformation
regulation of transcription, DNA-dependent
intracellular signal transduction
transcription, DNA-dependent
SP_0798,SP_0676
protein folding
de novo UMP biosynthetic process
proteolysis
Just Show Selected GO
Remove Filter (Show all GOs)
Export Network as PDF file

e Ontology ID: 6351

e Ontology Name: transcription, DNA-dependent

ein Level: 2

Figure 24

It is apparent from Figure 24 that GO terms with the high level of protein enrichment are not essentially located in the centre of a network. This is consistent with our knowledge of the levels of expression of transcription factors.

For additional examples of sample data please refer to the Comparative GO home page.

Appendix: SAMPLE DATA

Lung	Blood	Brain
SP_0432	SP_0211	SP_0133
SP_0440	SP_0263	SP_0225
SP_0675	SP_0538	SP_0325
SP_0676	SP_1044	SP_0326
SP_0677	SP_1045	SP_0327
SP_0678	SP_1109	SP_0328
SP_0683	SP_1329	SP_0333
SP_0684	SP_1430	SP_0334
SP_0685	SP_1517	SP_0335
SP_0686	SP_1545	SP_0336
SP_0692	SP_1673	SP_0341
SP_0693	SP_1752	SP_0342
SP_0694	SP_1860	SP_0343
SP_0699	SP_2074	SP_0344
SP_0702	SP_2182	SP_0349
SP_0771	SP_2237	SP_0350
SP_0772		SP_0351
SP_0773		SP_0352

SP_0774 SP_0779 SP_0780 SP_0781 SP_0782 SP_0787 SP_0788 SP_0789 SP_0790 SP_0795 SP_0796 SP_0797 SP_0798		SP_0421 SP_0422 SP_0423 SP_0424 SP_0429 SP_0430 SP_0431 SP_0432 SP_0437 SP_0438 SP_0439 SP_0440 SP_0445 SP_0446 SP_0447 SP_0448 SP_0579 SP_0580 SP_0581 SP_0582 SP_0587 SP_0589 SP_0590 SP_0595 SP_0596 SP_0597 SP_0603 SP_0604 SP_0605 SP_0606 SP_0675 SP_0676 SP_0677 SP_0678 SP_0683 SP_0684 SP_0685 SP_0686 SP_0691 SP_0692 SP_0693 SP_0694 SP_0699 SP_0700 SP_0701 SP_0702 SP_0739 SP_0740 SP_0741 SP_0742 SP_0747 SP_0748 SP_0749 SP_0750 SP_0755 SP_0756 SP_0757 SP_0758 SP_0763 SP_0764 SP_0765 SP_0766 SP_0771 SP_0772 SP_0773 SP_0774 SP_0779 SP_0780 SP_0781 SP_0782 SP_0787 SP_0788 SP_0789 SP_0790 SP_0795 SP_0796 SP_0797 SP_0798

		SP_0885 SP_0903 SP_0904 SP_0905 SP_0906 SP_0911 SP_0912 SP_0913 SP_0914 SP_0919 SP_0920 SP_0921 SP_0922 SP_0927 SP_0928 SP_0929 SP_0930 SP_1159 SP_1324 SP_1605 SP_2111