

# A Browser-based Functional Classification SuperViewer for *Arabidopsis* Genomics

Nicholas Provart<sup>1</sup>, Tong Zhu<sup>2</sup>

**Keywords:** functional classification, data visualization, *Arabidopsis thaliana*

## 1 Introduction.

Functional classification is commonly applied to data generated by microarray gene expression profiling experiments. Indeed, one of the most often used bioinformatics tools after BLAST and clustering is the functional classification pie chart. However, presentation of the absolute numbers of genes in a given cluster falling into each functional classification category may be misleading or mask difference under a given treatment. Another way of examining this sort of information is to normalize to the number of genes in each class present on the chip. In this way, differences are more easily perceived. We present here a web-based tool, the Functional Classification SuperViewer, which performs this normalization, bootstraps the dataset to provide a confidence estimate for the accuracy of the output, generates a dynamic graph summarizing the output for easy incorporation into reports, and provides links to TAIR [1] and other databases for individual IDs entered. Furthermore, expression values may also be submitted along with the IDs, and the values will be displayed with a colour-scale background, along with a functional classification bar-code.

## 2 Software and files.

The program is written in Perl and utilizes GD.pm [2] module for dynamic graph generation. Access is via a web-browser. For GeneChip IDs or AGI numbers keyed to the Munich Information Center for Protein Sequences (MIPS, [3]) dataset, the current MIPS classifications for 25450 genes in the MAtDB as of March 13, 2002, were downloaded as an XML file from [http://biors.gsf.de:8111/searchtool/searchtool.cgi?request=login\\_guest&file=frameset](http://biors.gsf.de:8111/searchtool/searchtool.cgi?request=login_guest&file=frameset). Those IDs falling into classification categories other than 'unclassified' and 'classification not yet clear-cut' were removed from these two categories. Data were reformatted into a flat file format which is then used by the program during classification. *Arabidopsis* annotations were obtained from The *Arabidopsis* Information Resource (TAIR, [1]). [ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Genes/ORF\\_annotations/ATH1.pep.0172002.total](ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Genes/ORF_annotations/ATH1.pep.0172002.total).

## 3 Results and Discussion.

A class score for normalization was calculated based on the following equation.  $N$  is number.

$$\text{Score}_{\text{class}} = [ N_{\text{class}}(\text{inputset}) / N_{\text{classified}}(\text{inputset}) ] / [ N_{\text{class}}(25\text{K}) / N_{\text{classified}}(25\text{K}) ]$$

Furthermore, the input set was bootstrapped one hundred times by sampling the input set (with repeats) and then reclassifying each set so generated. The standard deviation for the scores

---

<sup>1</sup> Department of Botany, University of Toronto, 25 Willcocks St., Toronto, ON. M5S 3B2, CANADA.  
E-mail: [provart@botany.utoronto.ca](mailto:provart@botany.utoronto.ca)

<sup>2</sup> Torrey Mesa Research Institute of Syngenta, AG. 3115 Merryfield Row, San Diego, CA. 92121, USA.  
E-mail: [tong.zhu@syngenta.com](mailto:tong.zhu@syngenta.com)

generated from the bootstrap sets is displayed along with the normalized class score. In this way classes represented by small numbers of genes on the chip may be easily identified (these tend to generate high scores or low scores if over- or under-represented but the SD will show them to be spurious).

Below is an example of the dynamically generated output. Left is shown the non-normalized numbers, right an example of normalized scores. As may be seen, it is readily apparent that protein biosynthesis is being actively up-regulated in this example. The gene list for this example was generated from a microarray experiment looking at chilled wild-type Arabidopsis plants, and such a response is well-known in other organisms.

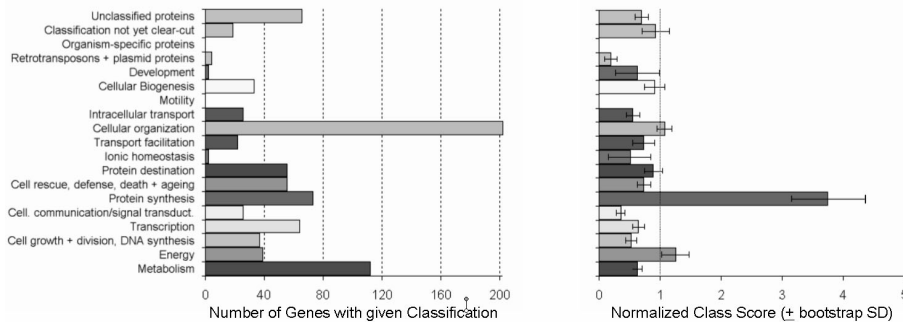


Figure 1: Example Output of the Functional Classification SuperViewer showing non-normalized (left panel) and normalized (right panel) results.

The HTML output also contains an overview table which indicates which classes a given AGI numbered protein falls into. Furthermore, links are also provided to TAIR and GenBank IDs, and it is also possible to upload expression values associated with each AGI number in the list. The resultant table may be easily copied into Word or other applications.

ID	Value	Classification Key	Subcategory	Annotation
At3g44990	6		other carbohydrate metabolism activities [01.05.99]	xyloglucan endo-transglycosylase ; supported by cDNA: gi_15810248_gb_AY056163.1
At5g3640	5.5		biosynthesis of secondary products derived from L-phenylalanine and L-tyrosine [01.20.35]	MAH20.20 flavonol synthase (FLS) (sp G96330) ; supported by full-length cDNA: Ceres:23924.
At5g42800	5.5		other vitamin, cofactor, and prosthetic group activities [01.07.99]	dihydroflavonol 4-reductase
At3g51240	6		biosynthesis of secondary products derived from L-phenylalanine and L-tyrosine [01.20.35]	flavanone 3-hydroxylase (F3H) ; supported by full-length cDNA: Ceres:36563
At5g13930	2		biosynthesis of secondary products derived from L-phenylalanine and L-tyrosine [01.20.35]	chalcone synthase (naringenin-chalcone synthase) (testa 4 protein) (sp P13114) ; supported by full-length cDNA: Ceres:38370.
At4g22880	4.5		biosynthesis of secondary products derived from L-phenylalanine and L-tyrosine [01.20.35]	putative leucoanthocyanidin dioxygenase (LDOX) ; supported by full-length cDNA: Ceres:42959.
At1g65960	4.5		biosynthesis of secondary products derived from L-phenylalanine and L-tyrosine [01.20.35]	4-coumarate:CoA ligase 3 identical to 4-coumarate:CoA ligase 3 GI:5702190 from [Arabidopsis thaliana] ; supported by cDNA: gi_5702191
At5g13930	2		biosynthesis of secondary products derived from L-phenylalanine and L-tyrosine [01.20.35]	chalcone synthase (naringenin-chalcone synthase) (testa 4 protein) (sp P13114) ; supported by full-length cDNA: Ceres:38370.
At1g64780	2		nitrogen and sulphur transport [01.02.07]	ammonium transporter, putative similar to ammonium transporter (GI:5880357 from [Arabidopsis thaliana] ; supported by cDNA: gi_4324713.
At1g15650	2.5		biosynthesis of secondary products derived from L-phenylalanine and L-tyrosine [01.20.35]	gibberellin 3 beta-hydroxylase, putative similar to gibberellin 3 beta-hydroxylase GI:5932753 from [Arabidopsis thaliana] ; supported by cDNA: gi_1945343_gb_L37126.1_ATHG44A
At2g29300	2.5		other secondary metabolism activities [01.20.99]	putative cytochrome P450
At4g29930	2		C-compound and carbohydrate utilization [01.05.01]	cellulose synthase catalytic subunit - like protein cellulose synthase catalytic subunit (Ath-A), Arabidopsis thaliana, gb:AF027173
At4g34750	2.5		C-compound, carbohydrate transport [01.05.07]	putative sugar transporter
At4g34740	2.5		C-compound and carbohydrate utilization [01.05.01]	TAL20.320 amidophosphoryltransferase 2 precursor
At2g29910	2.5		biosynthesis of secondary products derived from L-phenylalanine and L-tyrosine [01.20.35]	putative cinnamoyl CoA reductase ; supported by full-length cDNA: Ceres:14133.

Table 1: Example HTML output from SuperViewer.

## 4 References.

- [1] The Arabidopsis Information Resource. <http://www.arabidopsis.org/>
- [2] <http://stein.cshl.org/WWW/software/GD/>
- [3] Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. and Frishman, D. 1999. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research* 27: 44-48.