Documentation for `database.txt`

Description:

The file `database.txt`is a 163MB file which contains a pair-wise list of all the proteins seen in the TAP and HMS-PCI databases as well as the MIPS Complex Catalog as well as the number of trials, t, successes, s, and posterior probability for each pair-wise association. See Gilchrist *et al.* (2003) for more details.

About `database.txt`:

The database is available in gzip compressed form at www.unm.edu/∼compbio/software/Interaction_Assess
The database is comma delimited (i.e., .csv format) and is formatted as follows

		TAP			HMS-PCI			Combined
ORF1	ORF2	t	s	Post. Pr.	t	s	Post. Pr.	Post. Pr
YDR148C	YFL018C	1	1	0.5368207	0	0	0.0018828	0.5368207
YDR148C	YIL125W	2	2	0.9985977	0	0	0.0018828	0.9985977
YDR148C	YCL009C	2	0	0.0002258	0	0	0.0018828	0.0002258
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

The methods used to calculate the posterior probabilites are outlined in Gilchrist *et al.* (2003).

Using `database.txt`

At this point, the database is distributed as a stand alone file without any analysis tools. Given its large size ($> 2,700,000$ lines) we do not recommend using Microsoft Excel for database querying. Instead we recommend that researchers write their own Perl program for database querying. The database should also be importable into standard statistical programs such as SAS, R, or SPlus.

Reference

Gilchrist, M.A., L.A. Salter,and A.Wagner. 2003. A statistical framework for combining and interpreting proteomic datasets, submitted to *Bioinformatics*.