Configuration and instructions

Configuration files

Three configuration files need to be added/edited when adding a new dataset: the data file, the group_config file, the dataset_config file. One data file to be generated by python script: similar IDR data file.

All configuration/data files use human readable, easy to manipulate format. Currently csv format is used, all delimiters for column should be tab (“,” could be included in data file by accident when used as delimiter). Common problems in the data files that might break the code: unexpected symbols, missing values, unexpected data types.

All configuration/data files should be put in “templates” directory.

The dataset_config.csv file

This file list all the datasets, add one row when new dataset added.
First column: dataset name, any name for the dataset, will appear in the drop menu and other place in webpage.
Second column: specify the data file name of new dataset to be read by program;
Third column: specify the group_config file of new dataset to be read;
Fourth column: specify the similar IDR data file of new dataset to be read.

The data file

The main data file for the new dataset. One file for every dataset.
First row is header.

First column: unique name for every IDR, can’t be missing;
Second column: systematic protein name/Uniprot ID, missing value will be filled with “N/A”
Third column: Protein common name/gene name, missing value will be filled with “N/A”
All other columns: the Z-Scores for every feature, missing value will be filled with “0” (and 0 will be to calculate similar IDRs).
Z-Score columns order will be the order to be displayed on the webpage, features in same group should be in adjacent columns in order to properly define groups, groups ordered similarly for mean and log_variance z-scores.
Z-Score columns names should not decorate heavily otherwise the figures and webpage will be messy and hard to understand. But if the mean and log_var z-scores columns named exactly, they will be automatically decorated (“.1” added).

The group_config file
The file to define groups, one file for every dataset. It define groups by specify columns in data file.
First row is header (ignored by program).
First column: specify the columns in the data is z-scores for mean or log variance.
Second column: short name for the group, 1-7 characters. Longer names will cause figure labels overlap. No blank space in short name.
Third column: long name for webpage, more detailed description for features. Blank space allowed.
Don’t need specify mean or variance.
Fourth column: the start column in the data file for this group.
Fifth column: the end column in the data file for this group.
Six column: group id, naturally ordered for all groups.

The similar IDR file(sim file)
generated by findsim.py.

Procedures for adding new dataset to the website

1. Prepare configuration/data file, put in “templates” directory.
2. Run findsim.py with data file in “templates” directory, name the result file and add to dataset.config.csv. Running could take hours.
3. Because can’t pass any parameter to default page, the dataset list needs manually edited in the default index page:
   - Open index.html in templates directory, add <option> new dataset</option> just under the <option>Human</option> line.
   - Open app.py file and add:
     elif dataset=="new dataset":
       dataset=x
     where x is the new dataset number
     just under the following lines:
       elif dataset=="Human":
       dataset=1
   - The alternative for manually editing code is not set search page as default page, then dataset list can pass as parameters to it, that will cost users one more click to begin searching.
4. Run web site by “python app.py” command. In develop mode the website can be seen at localhost:5000

Maintenance

Images will accumulate in “static/image” folder, clear them to free space.