

Documentation for running the Perl scripts to evaluate the fraction of co-regulated from clustering results using SCPD

Steps (outline)

1. Apply clustering algorithms to the microarray data, and massage the output format.
2. Download the perl scripts, input data and gold standard from SCPD.
 - On linux: `tar -xzvf perl.tar.gz`
 - On windows: WinZip can uncompress "perl.tar.gz"
3. Compute the TP rates from clustering results using "evaluate.pl"
4. Compute the mean and stdev from random partitions using "randomPartition3.pl"
5. Compute the z-scores using "normalizeTF.pl"

Example:

"sampled_5E_1_OutCompleteLinkCorr_25.txt" is an example clustering result on the randomly sampled compendium data with 215 genes at $E = 5$. Hierarchical complete-link using correlation was used to produce 25 clusters.

Format of output files from clustering algorithms:

- When clustering genes: a tab-delimited file "Out<algorithm>_<number of clusters>.txt" with the following headings
GeneID<tab>ORF<tab>GeneName<tab>Cluster#(number of clusters<newline>
- GeneID's are unique identifiers for genes. They follow the order of the genes appearing in the input file. For example, the first gene in the file is 0, and the second gene is 1 etc.
- Genes with the same cluster number belong to the same cluster. For example, gene #164 and gene #169 were assigned to the same cluster in "sampled_5E_1_OutCompleteLinkCorr_25.txt".

Running ./evaluateTF.pl on linux/unix platform

- Compute TP rates from clustering results
- USAGE: `evaluateTF.pl [options]`
Options: `-cluster <cluster filename prefix>`
`-NoC <NoC_lo> <NoC_step> <NoC_hi>`
`-n <# randomly sampled datasets, typically 100 in our study>`
`-E <# experiments (E)>`
`-G <# genes in this current data subset>`
`-tf <gold standard from SCPD or YPD>`
- Example using the randomly sampled compendium data with 215 genes at $E = 5$, evaluated using SCPD:
`./evaluateTF.pl -cluster OutCompleteLinkCorr -NoC 5 5 100 -n 100 -E 5 -G 215 -tf merged_list_orf_tf_compendium215_noTATA.txt`
- Output: "percent_shareTF_OutCompleteLinkCorr.txt"
 - TP rate for each NoC and each randomly sampled dataset

Running ./randomPartition3.pl on linux/unix platform

- Compute mean and stdev of TP rates from random partitions
- USAGE: `randomPartition3.pl [options]`
Options: `-cluster <cluster filename prefix>`

-NoC <NoC_lo> <NoC_step> <NoC_hi>
 -n <# randomly sampled datasets, typically 100 in our study>

-N <# random partitions used to compute the mean and SD, typically 1000 in our study>

-E <# experiments (E)>
 -G <# genes in this current data subset>
 -tf <gold standard from SCPD or YPD>

- Example using the randomly sampled compendium data with 215 genes at E = 5, evaluated using SCPD:

```
./randomPartition3.pl -cluster OutCompleteLinkCorr -NoC 5 5 100 -n 100 -N 1000 -E 5 -G 215 -tf merged_list_orf_tf_compendium215_noTATA.txt
```

- Output files:
 - "random_avg_OutCompleteLinkCorr.txt": mean for each NoC and each randomly sampled dataset
 - random_sd_OutCompleteLinkCorr.txt": stdev for each NoC and each randomly sampled dataset

Running ./normalizeTF.pl on linux/unix platform

- Compute the z-scores using results from evaluateTF.pl and randomPartition3.pl
- Assumes output files from evaluateTF.pl and randomPartition3.pl in the current directory
- USAGE: randomPartition3.pl [options]
 Options: -cluster <cluster filename prefix>
 -NoC <NoC_lo> <NoC_step> <NoC_hi>
 -n <# randomly sampled datasets, typically 100 in our study>
- Example using the randomly sampled compendium data with 215 genes at E = 5, evaluated using SCPD:

```
./normalizeTF.pl -cluster OutCompleteLinkCorr -NoC 5 5 100 -n 100
```

- Output: "normalized_shareTF_OutCompleteLinkCorr.txt":
 - Row → z-scores from different # clusters
 - Columns → z-scores from each of the randomly sampled dataset

Running these perl scripts in Windows

1. Install Active Perl: <http://www.activestate.com/>
2. Replace the first line of each perl script with "#!perl.exe"