Connectivity Map 2 data package

1 Introduction

The Connectivity Map (version 2) data package provides a reference drug data set for the DrugVsDisease package which is a pipeline for the comparison of drug and disease gene expression profiles where negatively correlated (enriched) profiles can be used to generate hypotheses of drug-repurposing and positively correlated (enriched) profiles may be used to infer side-effects of drugs.

1.1 Drug Signatures

The Connectivity Map version 2.0 rank matrix was used to generate the reference set of drug profiles for DvD. All profiles for a given compound treatment were merged as described in [**?**], giving 1309 ranked expression profiles based on the HGU-133A platform. These profiles were then converted to gene symbols by taking the average rank where multiple probes map to the same gene and removing probes which mapped to more than one gene. These mappings were obtained from BiomaRt. Using these signatures, pairwise similarity scores were calculated using the KS running sum statistic based on the top 100 and bottom 100 genes for each profile. Affinity propagation clustering (provided by the R package apcluster) was used to create the network of drug connections. The 1309 compounds were clustered into 103 clusters, the assignments of each compound are stored in the object drugClusters. cMap2data does not contain any R code and all data objects can be accessed using the data command in R.

```
> data(drugClusters,package="cMap2data")
```

1.2 Cytoscape Information

An associated cytoscape plug-in is available for DvD which also uses the cMap2data package. The cMap2data package contains cytodrug and cytodisease data objects which have the edges in the network along with the distance and Running sum Peak Statistic (RPS). The latter two are used as edge attributes by Cytoscape. The Running sum Peak Statistic takes values 1 or -1 where 1 indicates a positive correlation and -1 a negative correlation. The distance measure gives the strength of this correlation. This data frame is used by the DrugVsDisease package to generate cytoscape sif and edge attribute files. For links out to external web browsers DrugBank, cMap2data also contains search compatible terms for all nodes in the reference data set. (Note that some compounds in the connectivity map are known not to be in DrugBank).

```
> data(cytodrug,package="cMap2data")
> #to get the compound (node) names and corresponding search terms
> data(druglabels,package="cMap2data")
```

References

[1] Hu G, Agarwal P (2009) Human Disease-Drug Network Based on Genomic Expression Profiles, *PLoS ONE*, **4(8)**: e6536.

[2] Shigemizu D, Hu Z, Hung J-H, Huang C-L, Wang Y, et al. (2012) Using Functional Signatures to Identify Repositioned Drugs for Breast, Myelogenous Leukemia and Prostate Cancer. *PLoS Comput Biol* **8(2)**: e1002347.

[3] Sirota M *et al.* (2011) Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. *Sci Trans Med*, **3:96ra77**.

[4] Subramanian A *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**(43), 15545-15550.

[5] Lamb J *et al.* (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, **313**(5795), 1929-1935.

[6] Gentleman R *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**(10), R80.

[7] Parkinson et al. (2010) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucl. Acids Res.*,doi: 10.1093/nar/gkq1040.

[R 2008] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

[9] Cline *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, **2**, 2366-2382.