Iterative network guided cMapping and validation

Supplementary Material and Methods - Supplementary Code: R objects documentation

This document describes functions, scripts and data objects used in the software enclosed to the paper entitled *A semi-supervised approach for refining transcriptional signatures of drug response and repositioning predictions*, by Francesco Iorio et al, submitted as research paper to PLoS ONE.

Author: Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

April 28, 2014

`DRUG_COMMUNITIES` *cMap drug communities*

Description

Data frame containing the community identifiers for 1,233 (out of 1,309) drugs from the Connectivity Map (cMap) dataset [1,2].
These communities have been obtained as described in [3,4].
Row names correspond to drug names. This data frame has been assembled using R and the data in the supplementary materials of [3] publicly available at [5] and [6].

Format

A data frame with 1233 observations on the following 2 variables, specifing for each drug:

`cID` The numerical identifiers of the community

`DRUGS` The drug name

References

[1] Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer, 7, 54-60.

[2] Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929.

[3] Iorio,F. et al. (2009) Identifying network of drug mode of action by gene expression profiling. Journal of Computational Biology, 16, 241-251.

[4] Iorio,F. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences, 107, 14621.

[5] http://www.pnas.org/content/107/33/14621.long?tab=ds

[6] http://mantra.tigem.it/About/AboutPnas.aspx.

DRUG_DISTANCES *cMap drug distances*

Description

1,233 x 1,233 double matrix containing the pair-wise distance scores for the 1,233 drugs in the cMap dataset [1,2].
These distances have been computed among drug prototype ranked lists (PRLs) (contained in the DRUG_PRLs object) assembled as described in [3,4] and the supplementary material and methods of our paper.
Row and column names correspond to drug names.
The entry in the i,j position of the matrix contains the distance between the i-th and the j-th drug.

References

[1] Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer, 7, 54-60.

[2] Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929.

[3] Iorio,F. et al. (2009) Identifying network of drug mode of action by gene expression profiling. Journal of Computational Biology, 16, 241-251.

[4] Iorio,F. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences, 107, 14621.

DRUG_PRLs *cMap drug prototype ranked lists*

Description

22,283 x 1,309 data frame containing the prototype ranked lists (PRLs) for all the drugs in the cMap dataset [1,2]. For each drug, the PRL consists of a genome-wide list of affyMetrix HG-U133A probe-sets identifiers sorted according to their consensual differential expression upon treatment with the drug under consideration, across a set of human cancer cell lines.

These PRLs have been assembled by post-processing the cMap gene expression profiles through the Kru-Bor method described in [4, 5] and the supplementary material and methods of our paper. Column names of the data frame correspond to drug names.

Format

A data frame with 22283 observations on 1309 variables.

[drug name]**a character vector containing 22,283 microrray probe-sets names**

References

[1] Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer, 7, 54-60.

[2] Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929.

[3] Iorio,F. et al. (2009) Identifying network of drug mode of action by gene expression profiling. Journal of Computational Biology, 16, 241-251.

[4] Iorio,F. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences, 107, 14621.

`GDSC_CELL_LINE_ANNOTATIONS`

GDSC cell line annotations

Description

1,471 x 5 data frame, containing the tissue of origin annotations, sample names and COSMIC [1] identifiers for all the cell lines in the GDSC [2] panel.
Row names correspond to COSMIC identifiers.

Format

A data frame with 1471 observations on the following 5 variables, specifying for each cell line:

`Cell.line.name` a character vector containing the cell line name

`Analysis.Set.Name` a character vector containing the cell line names the cell line name

`COSMIC.ID` a character vector containing the COSMIC identifier of the cell lines

`GDSC.description_1` a character vector containing the description of the tissue of origin of the cell line

`GDSC.description_2` a character vector containing the description of the tissue of origin of the cell line (at a different level of specificity)

References

[1] Forbes,S.A. et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res, 39, D945-50.

[2] Garnett,M.J. et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483, 570-575.

GDSC_DRUG_ANNOTATIONS *GDSC drug annotations*

Description

3 x 4 data frame, containing the annotations, and target information for docetaxel, vinorelbine, and paclitaxel.
Row names correspond to internal drug identifiers.

Format

A data frame with 3 observations on the following 4 variables, specifying of each drug:

`DRUG_NAME` The name of the drug

`SYNONYMS` Drug name synonyms

`BRAND_NAME` Brand name of the drug

`PUTATIVE_TARGET` Putative targets of the drug

GDSC_DRUG_SCREENING_DATA *GDSC drug screening data*

Description

Data structure containing the GDSC [1] screening data for docetaxel, vinorelbine, and paclitaxel, across the 1,074 human cancer cell lines in the panel.
It contains five 1,074 x 3 double matrix (`IC50s`, `IC90s`, `AUC`, `SLOPE`, and `maxConc`) with cell line COSMIC [2] identifiers as row names and drug internal identifiers as column names.

Format

The entry i,j of each of these contained double matrix, indicates for the treatment of the i-th cell line with the j-th drug:

`IC50s` the half-maximal inhibitory concentration

`IC90s` the 90% inhibitory concentration

`AUC` the normalised area under the dose/response curve

`SLOPE` the maximal concentration tested

References

[1] Garnett,M.J. et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483, 570-575.

GDSC_basalRanked_lists

Ranked lists of genes based on expression level statistics of the GDSC cell lines

Description

17641 x 715 string matrix, containing genome-wide ranked lists of genes (one for each cell line in the GDSC [1] panel) sorted according to their expression level statistics computed as described in the supplementary material and methods of our paper.
Column names correspond to COSMIC [2] cell line identifiers.

References

[1] Garnett,M.J. et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483, 570-575.

[2] Forbes,S.A. et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res, 39, D945-50.

GDSC_basalEXP

GDSC cell lines basal expression

Description

Basal expression profiles of the cell lines in the GDSC [1] panel.

Format

17,641 x 715 double matrix, containing the pre-processed basal expression profiles of the cell lines in the GDSC [1] panel.
Row names correspond to genes and column names correspond to cell line COSMIC [2] identifiers. This matrix has been assembled by downloading the raw gene expression data publicly available at the ArrayExpress repository [3] (accession number: E-MTAB-783) and by pre-processing it as described in the supplementary material and methods of our paper.

References

[1] Garnett,M.J. et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483, 570-575.

[2] Forbes,S.A. et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res, 39, D945-50.

[3] Parkinson,H. et al. (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res, 39, D1002-4.

`affy_ps_annotation` *affyMetrix probe-sets annotation*

Description

Data frame containing the annotation for the Affymetrix probe-sets in the HG-U133A platform, used to in the connectivity map project [1,2]

Format

A data frame with 22277 observations on the following 2 variables.

`V1` HUGO symbol(s) for the gene(s) mapped by the probe-sets

`V1` annotation(s) of the gene(s) mapped by the probe-sets

Row names correspond to the probe-sets identifiers

References

[1] Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer, 7, 54<e2><80><93>60.

[2] Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929.

`enrichedMOAs` *Modes of action enriched in the drug communities*

Description

String vector containing the modes-of-action or the features enriched in the drug communities described in [1].
Names of the vector correspond to community identifiers. This vector has been assembled by using R and the supplementary material of the publication [1], publicly available at [2].

References

[1] Iorio,F. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences, 107, 14621.

[2] `http://www.pnas.org/content/107/33/14621.long?tab=ds`

Index