gecn:
Pairwise integration of gene expression and copy number data

Leo Lahti*and Martin Schäfer

January 2, 2011

1 Introduction

A number of algorithms have been recently suggested for integrating gene expression and DNA copy number observations. This R package provides simulated example data and a comparison pipeline to evaluate the performance of the different approaches.

1.1 Comparison methods

The current version compares the following algorithms: DRI (3), edira (4), intCNGEan (5), SIM (2), pint (1), PMA (6).

2 Examples

This Section shows how to run the simulated data through the comparison pipeline, and to reproduce the results.

2.1 Simulated data

Use of the package is demonstrated with a simulated example data set containing paired observations of gene expression and copy number.

Example data set contains (i) gene expression matrix, (ii) gene copy number matrix, (iii) sample class labels (tumor/normal), and (iv) list of 'known' cancer genes.

2.2 Pollack et al. breast cancer data set

Read Pollack et al. (2002) data set (4719CopyNoGeneDataset.tsv[1]) and a list of known breast cancer genes. Ground truth list of known breast cancer genes[2]

*leo.lahti@iki.fi

[1]http://www.pnas.org/content/suppl/2002/09/23/162471999.DC1/4719CopyNoGeneDatsetLegend.html accessed June 2, 2010.

[2]http://www.nature.com/onc/journal/v18/n56/abs/1203335a.html

('tgdb_by_name.cgi.html' and 'tgdb.txt' [3])

```
> library(gecn)
```

```
pint Copyright (C) 2008-2010 Olli-Pekka Huovilainen and Leo
Lahti.
This program comes with ABSOLUTELY NO WARRANTY.
This is free software, and you are welcome to redistribute it under FreeBSD license, see t
```

```
> data(pollack)
> library("org.Hs.eg.db")
```

3 Methods

Acknowledgements

The package has been supported by EuGESMA COST Action BM0801: European Genetic and Epigenetic Study on AML and MDS.

References

[1] Leo Lahti, Samuel Myllykangas, Sakari Knuutila, and Samuel Kaski. Dependency detection with similarity constraints. In *Proceedings MLSP'09 IEEE International Workshop on Machine Learning for Signal Processing XIX*, pages 89–94, Piscataway, NJ, September 2-4 2009. IEEE Signal Processing Society.

[2] Renée˜X Menezes, Marten Boetzer, Melle Sieswerda, Gert-Jan˜B van Ommen, and Judith˜M Boer. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC bioinformatics*, 10(1):203, January 2009.

[3] Keyan Salari, Robert Tibshirani, and Jonathan˜R Pollack. DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics (Oxford, England)*, 26(3):414–6, February 2010.

[4] Martin Schäfer, Holger Schwender, Sylvia Merk, Claudia Haferlach, Katja Ickstadt, and Martin Dugas. Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics*, 25:3228–3235, 2009.

[5] Wessel˜N. van Wieringen and Mark˜A. van˜de Wiel. Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, 65:19–29, 2009.

[6] Daniela˜M. Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10:515–534, 2009.

[3] http://www.tumor-gene.org/cgi-bin/TGDB/tgdb_by_name.cgi accessed 5.6.2010.