intcomp:
Benchmarking tool for gene expression - copy number data integration models

Leo Lahti*and Martin Schäfer

May 28, 2011

1 Introduction

Several algorithms have been suggested to integrate gene expression and DNA copy number measurement to discover novel cancer-associated genes and chromosomal regions. However, quantitative comparison of these models has been missing. This R package provides a benchmarking pipeline for integrative cancer gene detection methods. Cancer gene detection performance of each algorithm is evaluated by comparing the prioritized candidate gene list from each method to a golden standard list of known cancer genes in simulated and real data sets.

The package is experimental beta-release and provided as is. The main purpose of this vignette is to provide reference to algorithmic details of the benchmarking framework used in (1) and sufficient documentation of the functionality such that similar experiments can be replicated by experienced R users.

1.1 Comparison methods

The current version implements the comparison of the following ge/cn integration algorithms: CNAmet (2; 3), DRI (4), edira (5), intCNGEan (6), Ortiz-Estevez (7), pint (8), PMA (9), PREDA/SODEGIR (10), and SIM (11). Multiple variants are evaluated for some methods (1).

2 Examples

2.1 Installing dependencies

The benchmarking pipeline depends on various external R packages. Install the dependencies from within R using the following commands:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite(c("biomaRt", "DNAcopy", "DRI", "edira", "hgu133ahsentrezg.db",
+     "intCNGEan", "org.Hs.eg.db", "PMA", "SIM", "PREDA", "CGHcall",
+     "CNAmet"))
```

*leo.lahti@iki.fi

You may need to install the following packages manually: curl library[1] and the R packages XML[2], RCurl[3], edira[4], intCNGEan[5], org.Hs.eg.db[6], PREDA/SODEGIR[7], and CNAmet[8].

3 Running the comparison tests

3.1 Loading example data

The package contains the data sets from (12) and (13). These two data sets are available from public sources and have been used to validate multiple alternative integration algorithms.

3.1.1 Pollack et al. (2002)

Loading Pollack et al. (2002) (13) data set[9].

```
> library(intcomp)
> data(pollack)
> dat <- read.pollack(chrs = 1:22, CopyNoGeneDataset4719, clone2geneid)
> ge <- dat$ge$data
> cn.raw <- dat$cn$data
> gene.info <- dat$ge$info
```

3.1.2 Hyman et al. (2002)

Loading Hyman et al. (2002) (12) data set[10].

```
> data(hyman)
> library("org.Hs.eg.db")
> dat <- read.hyman(cdna, cgh, genenames, chrs = 1:22, as.list(org.Hs.egALIAS2EG))
> ge <- dat$ge
> cn.raw <- dat$cn
> gene.info <- dat$cn.raw$info
```

3.2 Preprocessing copy number data

Some methods require segmented and/or called copy number data, obtained with

[1] http://curl.haxx.se/download.html

[2] http://cran.r-project.org/web/packages/XML/index.html

[3] http://www.omegahat.org/RCurl/

[4] http://www.statistik.tu-dortmund.de/~schaefer/

[5] http://www.few.vu.nl/~wvanwie/software/intCNGEan/intCNGEan.html

[6] http://www.bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html

[7] http://www.xlab.unimo.it/PREDA/PREDAinstall.R

[8] http://csbi.ltdk.helsinki.fi/CNAmet/

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

[10] HymancdnaDataA.tab, HymancghDataA.tab and HymanAcc.mat obtained from http://www.ece.ucsb.edu/pubs/ieee/index.shtml accessed June 2, 2010.

```
> cgh <- process.copynumber(cn.raw)
> cn.seg <- list(data = assayDataElement(cgh, "segmented"), info = gene.info)
> cn.call <- list(data = assayDataElement(cgh, "calls"), info = gene.info)
> rownames(cn.call$data) <- rownames(cn.seg$data) <- rownames(ge$data)
> cn.call$info <- cn.seg$info <- cn.raw$info
```

3.3 Golden standard list of known cancer genes

The golden standard list of known breast cancer genes from The Breast Cancer Gene Database[11] (14) was downloaded and stored to the tgdb object. This script details how to load the list, convert to Entrez GeneIDs and select the genes that are included in the experimental data:

```
> library("org.Hs.eg.db")
> data(tgdb)
> cancerGenes <- get.brca.genes(rownames(dat$ge$data), as.list(org.Hs.egALIAS2EG),
+     tgdb)
```

4 Running the benchmarking pipeline

The example data set now contains (i) gene expression data (ge), (ii) gene copy number data (cn), (iii) optional sample class labels (tumor/normal), and (iv) golden standard list of known cancer genes. The ge and cn data sets are lists containing *data* and *info* fields. The probes in gene expression and gene copy number are paired; *data* is a data matrix with gene expression (ge$data) or gene copy number (cn$data) data; *info* field is a data frame containing additional information about genes: *loc* indicates the genomic location of the probes in base pairs (numeric); *chr* and *arm* are factors indicating the chromosome and chromosomal arm of the probe, respectively. To run the benchmarking tests for all methods, use:

```
> methods <- c("CNAmet", "edira")
> res <- test.geneorder.pipeline(ge = ge, cn.raw = cn.raw, cn.seg = cn.seg,
+     cn.call = cn.call, cghCall = cgh, Labels = NULL, cancerGenes = cancerGenes,
+     nperm = 20, input = "real", version = "normal", methods = methods)
> auc.ordered <- sort(unlist(res$auc))
```

The following *methods* are available: *OrtizEstevez*, *intcngean*, *edira*, *pint*, *SIM.window}, SIM.full}, DRI.cp, DRI.cs, PMA.raw, CNAmet, PREDA and for two-group comparisons (requiring Labels) DRI.ct, DRI.ss, DRI.srank, DRI.sraw. The cancer gene prioritization of each method is compared to the golden standard list of known cancer genes; the result contains running times of the algorithms and the AUC values from ROC analysis (1). The AUC values provide quantitative estimates of model performance in cancer gene detection and provide the basis for comparisons.*

[11] http://www.tumor-gene.org/cgi-bin/TGDB/tgdb_by_name.cgi accessed 5.6.2010; ’tgdb_by_name.cgi.html’ and ’tgdb.txt’

Acknowledgements

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

References

[1] *L.~Lahti, M.~Schäfer, H.-U. Klein, S.~Bicciato, and M.~Dugas, "Cancer gene prioritization by multi-platform data integration: a comparative review. upcoming,"*

[2] *S.~Hautaniemi, M.~Ringnér, P.~Kauraniemi, R.~Autio, H.~Edgren, O.~Yli-Harja, J.~Astola, A.~Kallioniemi, and O.~Kallioniemi, "A strategy for identifying putative causes of gene expression variation in human cancers,"* Journal of the Franklin Institute*, vol.~341, no.~1-2, pp.~77–88, 2004.*

[3] *R.~Louhimo and S.~Hautaniemi, "CNAmet: an R package for integrating copy number, methylation and expression data,"* Bioinformatics*, vol.~27, no.~6, pp.~887–888, 2011.*

[4] *K.~Salari, R.~Tibshirani, and J.~R. Pollack, "DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data.,"* Bioinformatics*, vol.~26, no.~3, pp.~414–6, 2010.*

[5] *M.~Schäfer, H.~Schwender, S.~Merk, C.~Haferlach, K.~Ickstadt, and M.~Dugas, "Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities,"* Bioinformatics*, vol.~25, no.~24, pp.~3228–3235, 2009.*

[6] *W.~N. van Wieringen and M.~A. van~de Wiel, "Nonparametric testing for DNA copy number induced differential mRNA gene expression,"* Biometrics*, vol.~65, pp.~19–29, 2009.*

[7] *M.~Ortiz-Estevez, J.~De~Las~Rivas, C.~Fontanillo, and A.~Rubio, "Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression,"* Genomics*, vol.~97, pp.~86–93, 2011.*

[8] *L.~Lahti, S.~Myllykangas, S.~Knuutila, and S.~Kaski, "Dependency detection with similarity constraints," in* Proceedings MLSP'09 IEEE International Workshop on Machine Learning for Signal Processing XIX*, (Piscataway, NJ), pp.~89–94, IEEE Signal Processing Society, September 2-4 2009. Implementation available in pint package of R/BioConductor http://www.bioconductor.org/packages/release/bioc/html/pint.html.*

[9] *D.~M. Witten, R.~Tibshirani, and T.~Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,"* Biostatistics*, vol.~10, no.~3, pp.~515–534, 2009.*

[10] *S.~Bicciato, R.~Spinelli, M.~Zampieri, E.~Mangano, F.~Ferrari, L.~Beltrame, I.~Cifola, C.~Peano, A.~Solari, and C.~Battaglia, "A computational procedure to identify significant overlap of differentially*

expressed and genomic imbalanced regions in cancer datasets," Nucleic Acids Research*, vol.~37, pp.~5057–5070, 2009.*

[11] R.~X. Menezes, M.~Boetzer, M.~Sieswerda, G.-J.~B. van Ommen, and J.~M. Boer, "Integrated analysis of DNA copy number and gene expression microarray data using gene sets.," BMC bioinformatics*, vol.~10, no.~1, p.~203, 2009.*

[12] E.~Hyman, P.~Kauraniemi, S.~Hautaniemi, M.~Wolf, S.~Mousses, E.~Rozenblum, M.~Ringner, G.~Sauter, O.~Monni, A.~Elkahloun, O.-P. Kallioniemi, and A.~Kallioniemi, "Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer," Cancer Res.*, vol.~62, pp.~6240–6245, Nov. 2002.*

[13] J.~R. Pollack, T.~Sø~rlie, C.~M. Perou, C.~A. Rees, S.~S. Jeffrey, P.~E. Lonning, R.~Tibshirani, D.~Botstein, A.-L. Bø~rresen Dale, and P.~O. Brown, "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.," Proceedings of the National Academy of Sciences of the United States of America*, vol.~99, pp.~12963–8, Oct. 2002.*

[14] R.~Baasiri, S.~Glasser, D.~Steffen, and D.~Wheeler, "The Breast Cancer Gene Database: a collaborative information resource," Oncogene*, vol.~18, no.~56, pp.~7958–7965, 1999.*