Grinn: a Graph database and R package for omic data integration

Kwanjeera Wanichthanarak, Dmitry Grapov, Johannes F. Fahrmann and Oliver Fiehn

February 1, 2016

Contents

1 INSTALLATION

1.1 INSTALL R package grinn

1) Require R software (3.1.1 or higher)
2) Install R package grinn using the following commands in R terminal:

```
## Install devtools R package, if not exist
> install.packages("devtools")
## Install grinn R package
> library("devtools")
> devtools::install_github("kwanjeeraw/grinn")
```

1.2 INSTALL Neo4j database server *(optional)*

This step is only required when setting up a graph database on local machine. The graph database is a part of the Grinn software tool to support network queries and network integration. Currently the precompiled da-tabases are available for human, arabidopsis, mouse, rat, S. cerevisiae and E. coli k-12 at http://sourceforge.net/projects/grinn/files/grinnDatabases/. The human database is connected by default and can be accessed directly after installing the R package grinn.

1) Require Neo4j-community server (2.2.0 or higher)
2) Download and then unzip the Neo4j server
3) Download the database file, extract and move to Neo4j_directory/data
4) Start the Neo4j server
5) Switch to the local database by using the function `setGrinnDb`:

```
## Set database location
> library(grinn)
> setGrinnDb("http://localhost:7474/db/data/")
## Get current database location
> getGrinnDb()
```

2 USE CASE

Here we provide the use cases and R code corresponding to Results and Discussion in the manuscript. We use transcriptomic data [1] and metabolomic data [2] from the published studies on adenocarcinoma lung cancer comparing normal to tumor tissue to demonstrate the different features of Grinn.

Load required library and the data file:

```
> library(grinn)
## Load data for using in the following use cases
> data("lungCancer")
```

Summary of the objects in the `lungCancer` data:

- `geneDat` = log-transformed transcriptomic data
- `metDat` = covariate adjusted metabolomic data
- `sigGene` = 173 significantly expressed genes comparing between tumor and normal tissues
- `sigMet` = 70 significantly changed metabolites comparing between tumor and normal tissues

2.1 Inferring the comprehensive biochemical network of lung adenocarcinoma from gene- and metabolite-based queries

173 significantly expressed genes and 70 significantly changed metabolites comparing tumor with normal tissues are separately used to query a metabolite-protein-gene network. The resulting biochemical networks are then combined and exported for visualization in Cytoscape [3] (Figure 3 and Supplementary figure 1).

```
## Query a metabolite-protein-gene network by using a set of genes
> gpmNW = fetchGrinnNetwork(txtInput = sigGene$grinnID, from = "gene", to="metabolite")
## Set remark
> gpmNW$edges$from = "gpm"
## Observe output
> head(gpmNW$nodes)
> head(gpmNW$edges)
## Query a metabolite-protein-gene network by using a set of metabolites
> mpgNW = fetchGrinnNetwork(txtInput = sigMet$grinnID, from="metabolite", to="gene")
## Set remark
> mpgNW$edges$from = "mpg"
## Observe output
> head(mpgNW$nodes)
> head(mpgNW$edges)
## Combine the networks
> cmbNW = combineNetwork(gpmNW, mpgNW)
## Observe output
> head(cmbNW$nodes)
> head(cmbNW$edges)
## Export output
> write.table(as.matrix(cmbNW$edges), file="combinedNWEdge.txt", sep="\t", row.names=FALSE, quote=FALSE
> write.table(as.matrix(cmbNW$nodes), file="combinedNWNode.txt", sep="\t", row.names=FALSE, quote=FALSE
```

2.2 Integrative analysis of gene expression and metabolic profiles for biochemical network re-construction

Pairwise correlations among 21,204 genes and 462 metabolites are computed and connected with prede-fined molecular relationships to produce an integrated network of data-driven and knowledge-oriented interactions (Figure 4). The resulting network is exported for visualization in Cytoscape.

```
## Compute an integrated network of correlation-oriented and domain knowledge-based relationships
> intgNW = fetchCorrGrinnNetwork(datX=metDat, datY=geneDat, corrCoef=0.5, pval=0.05, method="spearman",
## Observe output
> head(intgNW$nodes)
> head(intgNW$edges)
## Export output
> write.table(as.matrix(intgNW$edges), file="integratedNWEdge.txt", sep="\t", row.names=FALSE, quote = F
> write.table(as.matrix(intgNW$nodes), file="integratedNWNode.txt", sep="\t", row.names=FALSE, quote = F
```

REFERENCES

1. Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, Siegmund KD, Koss MN, Hagen JA, Lam WL et al: Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. Genome research 2012, 22(7):1197-1211.
2. Wikoff WR, Grapov D, Fahrmann JF, DeFelice B, Rom WN, Pass HI, Kim K, Nguyen U, Taylor SL, Gandara DR et al: Metabolomic markers of altered nucleotide metabolism in early stage adenocarcinoma. Cancer prevention research 2015, 8(5):410-418.
3. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 2003, 13(11):2498-2504.