ESTIMATION GRAPHICAL MODEL

26 OKTOBER, 2021

Mona Azadkia, Ahmad Ehyaei

ETH zürich

Contents

Chapter1

Introduction

Chapter2

Review and background

Chapter3

Methods

Chapter4

Theoretical properties

Chapter5

Simulation

Chapter6

Application to TCGA data

For download the RNA-seq breast cancer data from Cancer Genome Atlas (TCGA) database we use `TCGAbiolinks` package (Silva et al. 2016).

6.1 Download Data

6.1.1 TCGA Portal

TCGAbiolinks' purpose is to make it easier to access GDC data, build preprocessing methods, and provide multiple methods for analysis and visualization. For install package run below commands:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("TCGAbiolinks")
```

Fist we download clinical data related to `TCGA-BRCA` study

```
library(TCGAbiolinks)

clinical_data <- TCGAbiolinks::getLinkedOmicsData(
  project = "TCGA-BRCA",
  dataset = "Clinical"
)
readr::write_rds(clinical_data,"data/clinical_brca.rds")
```

Table 6.1: TCGA BRCA Clinical Data Header

attrib_name	TCGA.5L.AAT0	TCGA.5L.AAT1
years_to_birth	42	63
Tumor_purity	0.6501	0.5553
pathologic_stage	stageii	stageiv
pathology_T_stage	t2	t2
pathology_N_stage	n0	n0
pathology_M_stage	m0	m1
histological_type	infiltratinglobularcarcinoma	infiltratinglobularcarcinoma
number_of_lymph_nodes	0	0
PAM50		
ER.Status		
PR.Status		
HER2.Status		
gender	female	female
radiation_therapy	yes	no
race	white	white
ethnicity	hispanicorlatino	hispanicorlatino
Median_overall_survival	0	0
overall_survival	1477	1471
status	0	0
overallsurvival	1477,0	1471,0

The above table contains clinical data of patients. The name of each column is related to the patient's id.

The GDC mRNA quantification analysis pipeline measures gene level expression in HT-Seq raw read count, Fragments per Kilobase of transcript per Million mapped reads (FPKM), and FPKM-UQ (upper quartile normalization). For see more information see mRNA Analysis Pipeline GDC document.

Definition 6.1.1: HT Seq Normalization

RNA-Seq expression level read counts produced by HT-Seq are normalized using two similar methods: FPKM and FPKM-UQ. Normalized values should be used only within the context of the entire gene set. Users are encouraged to normalize raw read count values if a subset of genes is investigated.

FPKM:

The Fragments per Kilobase of transcript per Million mapped reads (FPKM) calculation normalizes read count by dividing it by the gene length and the total number of reads mapped to protein-coding genes.

$$FPKM = \frac{RC_g * 10^9}{RC_{pc} * L} \tag{0.3}$$

Upper Quartile FPKM

The upper quartile FPKM (FPKM-UQ) is a modified FPKM calculation in which the total protein-coding read count is replaced by the 75th percentile read count value for the sample.

$$FPKM - UQ = \frac{RC_g * 10^9}{RC_{g75} * L}$$

RC_g: Number of reads mapped to the gene
RC_{pc}: Number of reads mapped to all protein-coding genes
RC_{g75}: The 75th percentile read count value for genes in the sample
L: Length of the gene in base pairs; Calculated as the sum of all exons in a gene
The read count is multiplied by a scalar 10^9 during normalization to account for the kilobase and 'million mapped reads' units.
mRNA Expression HT-Seq Normalization

```
measurements = c( "HTSeq - Counts", "HTSeq - FPKM", "HTSeq - FPKM-UQ")
for( m in measurements){
  query <- GDCquery(
    project = "TCGA-BRCA",
    data.category = "Transcriptome Profiling",
    data.type = "Gene Expression Quantification",
    workflow.type = "HTSeq - FPKM-UQ",
    barcode =  colnames(clinical_data)[-1] # List of all patients id.
  )

  GDCdownload(query,files.per.chunk = 100)
   GDCprepare(query = query, summarizedExperiment = FALSE) %>%
  readr::write_rds(sprintf("data/TCGA_BRCA_%s.rds",gsub(" - ","_",m)))
}
```

In the below table, the header of the HT-Seq count data can be found:

mRNA HT-Seq Count Data

ensembl_gene_id	TCGA-3C-AAAU-01A-11R-A41B-07	TCGA-3C-AALI-01A-11R-A41B-07
ENSG00000153786.11	4416	2818
ENSG00000266493.1	0	0
ENSG00000227488.2	0	0

ENSG00000249438.2	0	0
ENSG00000255317.1	0	0
ENSG00000252423.1	0	0
ENSG00000232685.4	0	2
ENSG00000123415.13	1369	1735
ENSG00000200702.1	0	0
ENSG00000181982.16	1812	1510

The `HTSeq-Counts` contains 60488 genes.

6.1.2 Gense Metadata

To use name and id of genes, we download genes metadata tables with `biomaRt` package (Durinck et al. 2005).

```
library(biomaRt)
mart <- useMart(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
listAttributes(mart)
geneTable <- getBM(
  attributes = c("ensembl_gene_id","chromosome_name","start_position",
                 "end_position","strand","band",
                 "hgnc_id","hgnc_symbol"),
                       mart = mart)

geneTable %>%
  readr::write_rds("data/geneTable.rds")
```

Some Fields of Genes Meta Data

ensembl_gene_id	chromosome_name	start_position	end_position	strand	band	hgnc_symbol
ENSG00000210049	MT	577	647	1		MT-TF
ENSG00000211459	MT	648	1601	1		MT-RNR1
ENSG00000210077	MT	1602	1670	1		MT-TV
ENSG00000210082	MT	1671	3229	1		MT-RNR2
ENSG00000209082	MT	3230	3304	1		MT-TL1
ENSG00000198888	MT	3307	4262	1		MT-ND1
ENSG00000210100	MT	4263	4331	1		MT-TI
ENSG00000210107	MT	4329	4400	-1		MT-TQ
ENSG00000210112	MT	4402	4469	1		MT-TM
ENSG00000198763	MT	4470	5511	1		MT-ND2

6.1.3 MSigDB Hallmark 50

> We envision this collection as the starting point for your exploration of the MSigDB resource and GSEA. Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying gene set overlaps and retaining genes that display coordinate expression. The hallmarks reduce noise and redundancy and provide a better delineated biological space for GSEA. We refer to the original overlapping gene sets, from which a hallmark is derived, as its 'founder' sets. Hallmark gene set pages provide links to the corresponding founder sets for deeper follow up. This collection is an initial release of 50 hallmarks which condense information from over 4,000 original overlapping gene sets .(Liberzon et al. 2015)

To download MSigDB we use msigdbr package (Dolgalev 2021).

```
if(!require(msigdbr)) install.packages("msigdbr")
library(msigdbr)

# Retrieve human H (hallmark) gene set
msigdb <- msigdbr(species = "Homo sapiens", category = "H")
readr::write_rds(msigdb,"data/msigdb.rds")
```

Head of Human H (hallmark) Gene Set

gs_name	gene_symbol	entrez_gene	ensembl_gene
HALLMARK_ADIPOGENESIS	ABCA1	19	ENSG00000165029
HALLMARK_ADIPOGENESIS	ABCB8	11194	ENSG00000197150
HALLMARK_ADIPOGENESIS	ACAA2	10449	ENSG00000167315
HALLMARK_ADIPOGENESIS	ACADL	33	ENSG00000115361
HALLMARK_ADIPOGENESIS	ACADM	34	ENSG00000117054
HALLMARK_ADIPOGENESIS	ACADS	35	ENSG00000122971
HALLMARK_ADIPOGENESIS	ACLY	47	ENSG00000131473
HALLMARK_ADIPOGENESIS	ACO2	50	ENSG00000100412
HALLMARK_ADIPOGENESIS	ACOX1	51	ENSG00000161533
HALLMARK_ADIPOGENESIS	ADCY6	112	ENSG00000174233

6.1.4 Pathway Commons

> Pathway Commons is a collection of publicly available pathway data from multiple organisms. Pathway Commons provides a web-based interface that enables biologists to browse and search a comprehensive collection of pathways from multiple sources represented in a common language, a download site that provides integrated bulk sets of pathway information in standard or convenient formats and a web service that software developers can use to conveniently query and access all data. Database providers can share their pathway data via a common repository. Pathways include biochemical reactions, complex assembly, transport and catalysis events and physical interactions involving proteins, DNA, RNA, small molecules and complexes. Pathway Commons aims to collect and integrate all public pathway data available in standard formats. (Cerami et al. 2010)

To access Pathway Commons data we download `PathwayCommons12.All.hgnc.txt.gz` from data page.

```
pc = readr::read_delim("raw-data/PathwayCommons12.All.hgnc.txt",delim = "\t")
pc$MEDIATOR_IDS = NULL
readr::write_rds(pc,"data/pathwaycommons.rds")
```

Sample Data of Pathway Commons

PARTICIPANT_A	INTERACTION_TYPE	PARTICIPANT_B	INTERACTION_DATA_SOURCE
A1BG	controls-expression-of	A2M	pid
A1BG	interacts-with	ABCC6	BioGRID
A1BG	interacts-with	ACE2	BIND
A1BG	interacts-with	ADAM10	BIND
A1BG	interacts-with	ADAM17	BIND
A1BG	interacts-with	ADAM9	BIND
A1BG	interacts-with	AGO1	BIND
A1BG	controls-phosphorylation-of	AKT1	pid
A1BG	controls-state-change-of	AKT1	pid
A1BG	interacts-with	ANXA7	IntAct;BioGRID

6.2 Cleansing Data

To replicate (Yang et al. 2021) paper, we consider only genes in RNA-seq whose read number is more than 20 in at least 25% of the samples. We also use `HTSeq-FPKM-UQ`, which is more robust than `HTSeq-Counts` using the total number of reads per sample.

```
# Load gene Table
geneTable = readr::read_rds("data/geneTable.rds") %>%
  select(ensembl_gene_id,hgnc_symbol)

HTSeq_Counts = readr::read_rds("data/TCGA_BRCA_HTSeq_Counts.rds") %>%
  rename( gene_id = X1)

# Find Most Frequent Genes
SeqMat = HTSeq_Counts %>%
  select(starts_with("TCGA")) %>%
  as.matrix()
GeneList = HTSeq_Counts$gene_id[rowSums(SeqMat>20)/ncol(SeqMat)>0.25]

# Filter HTSeq_FPKM_UQ to Most Frequent Genes
readr::read_rds("data/TCGA_BRCA_HTSeq_FPKM_UQ.rds") %>%
  rename( gene_id = X1) %>%
  mutate(ensembl_gene_id = substr(gene_id,1,15)) %>%
  left_join(geneTable,by = "ensembl_gene_id") %>%
  filter(!is.na(hgnc_symbol)) %>%
  filter(gene_id %in% GeneList) %>%
  select(gene_id,hgnc_symbol,starts_with("TCGA")) %>%
  readr::write_rds("data/RNASeq.rds")
```

After frequent restriction, 19879 genes remained.

```
RNASeq = readr::read_rds("data/RNASeq.rds")
```

Reference

Cerami, Ethan G, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. 2010. "Pathway Commons, a Web Resource for Biological Pathway Data." **Nucleic Acids Research** 39 (suppl_1): D685–90.

Dolgalev, Igor. 2021. **Msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format**. https://CRAN.R-project.org/package=msigdbr.

Durinck, Steffen, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. 2005. "BioMart and Bioconductor: A Powerful Link Between Biological Databases and Microarray Data Analysis." **Bioinformatics** 21: 3439–40.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. 2015. "The Molecular Signatures Database Hallmark Gene Set Collection." **Cell Systems** 1 (6): 417–25.

Silva, Tiago C, Colaprico, Antonio, Olsen, Catharina, D'Angelo, et al. 2016. "TCGA Workflow: Analyze Cancer Genomics and Epigenomics Data Using Bioconductor Packages." **F1000Research** 5.

Yang, Jenny, Yang Liu, Yufeng Liu, and Wei Sun. 2021. "Model Free Estimation of Graphical Model Using Gene Expression Data." **The Annals of Applied Statistics** 15 (1): 194–207.