Visualizing RNA-seq data: Pertinence, software, and applications

Exit Seminar
Lindsay Rutter

Program of Study Committee:
Dianne Cook (Major Professor)
Amy Toth (Major Professor)
Heike Hofmann
Daniel Nettleton
James Reecy

May 1, 2018

My Background

- Education
    - B.S. in Bioengineering
      Pennsylvania State University (2003-2007)
    - Major in Bioinformatics and Computational Biology
      Iowa State University (2012-Present)
    - Minor in Statistics
   *Iowa State University* (2012-Present)
- Internships
    - Nagaoka University of Technology (Summer 2012)
    - Okinawa Institute of Science and Technology (Summer 2014)
    - MathWorks (Summer 2016)
    - After Inc. (Fall 2016)
    - Google Summer of Code (Summer 2017)
    - GeneLab (Pending - Summer 2018)

Chapter Outline

- **Chapter 1**
  Visualization methods for genealogical datasets

- **Chapter 2**
  The case for visualization methods in RNA-seq data analysis

- **Chapter 3**
  Software for visualization methods in RNA-seq data analysis

- **Chapter 4**
  Gene expression responses to diet quality and viral infection
  in *Apis mellifera*

Chapter 1 : Visualization methods for genealogical datasets
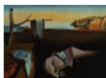
ggenealogy: An R Package for Visualizing Genealogical Data

Lindsay Rutter
Iowa State University

Susan VanderPlas
Iowa State University

Dianne Cook
Monash University

Michelle A. Graham
Iowa State University

Abstract

This paper introduces **ggenealogy** (Rutter *et al.* 2015), a developing R software package that provides tools for searching through genealogical data, generating basic statistics on their graphical structures using parent and child connections, and displaying the results. It is possible to draw the genealogy in relation to variables related to the nodes, and to determine and display the shortest path distances between the nodes. Production of pairwise distance matrices and genealogical diagrams constrained on generation are also available in the visualization toolkit. The tools are being tested on a dataset with milestone cultivars of soybean varieties (Hymowitz *et al.* 1977) as well as on a web-based database of the academic genealogy of mathematicians (North Dakota State University and American Mathematical Society). The latest stable package version is available in source and binary form on the Comprehensive R Archive Network.

Keywords: genealogy, data visualization, statistical graphics, exploratory data analysis, interactive, R.

Chapter 4 : Visualization methods for significance testing of RNA-sequencing data

Completed work

Product	Description	Date
R package	First release of **ggenealogy**, which provides visualization tools for genealogical datasets	March 2015
Presentation	Presented **ggenealogy** at JSM	August 2015
Award	Student paper award at ASA Statistical Computing and Graphics Section	August 2015

Scheduled deliverables

Product	Description	Date
R package	Second release of **ggenealogy** package, which provides visualization tools for genealogical datasets	May 2016
Paper	Submit **ggenealogy** paper to JSS	May 2016
R package	First release of package that provides visualization tools for RNA-sequencing datasets	TBD
Paper	Submit paper about visualization tools for clustering analysis of RNA-sequencing	TBD
Paper	Submit paper about visualization tools for significance testing of RNA-sequencing	TBD

Other work

Product	Description	Date
R package	First release of ePort package that generates electronic reports for instructors to evaluate student performance	July 2016

References

- `ggplot2` (Wickham 2009)
- `GGally` (Schloerke et al. 2016)
- `nullabor` (Wickham et al. 2014)
- `ggbio` (Yin et al. 2012)
- `GGobi` (Swayne et al. 2003)
- `tourr` (Wickham et al. 2011)
- `plotly` (Sievert et al. 2016)
- Parallel coordinate plots (Inselberg 1985, Wegman 1990)
- Visual statistical inference (Chowdhury et al. 2015)

References

- `pedigree` (Coster 2013)
- `kinship2` (Therneau et al. 2015)
- `GraphViz` (Gansner and North 2000)
- `Cytoscape` (Shannon et al. 2003)
- `explorase` (Lawrence et al. 2008)
- `edgeR` (Robinson et al. 2010)
- `DESeq2` (Love at al. 2014)
- `RUVseq` (Risso et al. 2014)
- Gene expression visual inference (Yin et al. 2013)
- Biological clustering : (Newell et al. 2013)

Leaves at 120 minutes - 3 Clusters