Data Analysis for Molecular Biology and Biochemistry

MBB 110

Introductory data analysis focusing on molecular biology data sets and examples and including basic programming skills using Python and basic statistics skills using R. Prerequisite: MATH 12 or equivalent is recommended. Students with credit for MBB 243 may not take this course for further credit. CMPT 120 will be accepted in lieu of MBB 110.

Topics

- Flavours of data in molecular biology and biochemistry
- Genomic data
- Fundamentals of R and Python
- Regular expressions and patterns
- Quantitative DNA/RNA sequence analysis
- Exploratory data analysis in the Tidyverse
- Generic visualization methods using ggplot2
- Advanced visualizations for molecular biology

INSTRUCTOR:

Sophie Sneddon ~ Bioinformatician, video game enthusiast, and cat person

sort & uniq

JULIA EVANS
@b0rk

sort sorts its input

$ sort names.txt

the default sort is alphabetical.

---

sort -n
numeric sort

`sort` order	`sort -n` order
12	12
15000	48
48	96
6020	6020
96	15000

---

sort -h: human sort

`sort -n` order	`sort -h` order
15 G	45 K
30 M	30 M
45 K	15 G
200 G	200 G

useful example:
du -sh * | sort -h

---

uniq removes duplicates

a
b
b   =>   a
a   b
c   a
c   c

notice there are still 2 'a's! uniq only uniquifies **adjacent** matching lines

---

sort + uniq = ♥

Pipe something to `sort | uniq` and you'll get a deduplicated list of lines! (sort -u) does the same thing.

b
a
b | sort -u =>   a
a   b

or sort | uniq

---

uniq -c

counts each line it saw.

Recipe: get the top 10 most common lines in a file:

$ sort foo.txt
| uniq -c   I use
| sort -n   ←   this
| tail -n 10   a lot

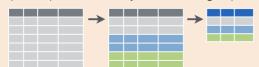
Group Data

dplyr::**group_by(iris, Species)**
  Group data into rows with the same value of Species.
dplyr::**ungroup(iris)**
  Remove grouping information from data frame.

iris %>% group_by(Species) %>% summarise(...)
Compute separate summary row for each group.

iris %>% group_by(Species) %>% mutate(...)
Compute new variables by group.