Bioinformatics Tools
High Performance Computing

BMI701 Introduction of Biomedical Informatics
Lab Session 8

Wei-Hung Weng

November 9, 2016

HMS DBMI — MGH LCS

Bioinformatics

- Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines **computer science, statistics, mathematics and engineering** to study and process biological data.

  Wikipedia

Bioinformatics

	Bioinformatics User	Bioinformatics Scientist	Bioinformatics Engineer
(a) An ability to apply knowledge of computing, biology, statistics, and mathematics appropriate to the discipline.		X	X
(b) An ability to analyze a problem and identify and define the computing requirements appropriate to its solution.		X	X
(c) An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs in scientific environments.			X
(d) An ability to use current techniques, skills, and tools necessary for computational biology practice.	X	X	X
(e) An ability to apply mathematical foundations, algorithmic principles, and computer science theory in the modeling and design of computer-based systems in a way that demonstrates comprehension of the tradeoffs involved in design choices.			X
(f) An ability to apply design and development principles in the construction of software systems of varying complexity.			X
(g) An ability to function effectively on teams to accomplish a common goal.	X	X	X
(h) An understanding of professional, ethical, legal, security, and social issues and responsibilities.	X	X	X
(i) An ability to communicate effectively with a range of audiences.	X	X	X
(j) An ability to analyze the local and global impact of bioinformatics and genomics on individuals, organizations, and society.	X	X	X
(k) Recognition of the need for and an ability to engage in continuing professional development.	X	X	X
(l) Detailed understanding of the scientific discovery process and of the role of bioinformatics in it.	X	X	X
(m) An ability to apply statistical research methods in the contexts of molecular biology, genomics, medical, and population genetics research.	X	X	X
(n) Knowledge of general biology, in-depth knowledge of at least one area of biology, and understanding of biological data generation technologies.	X	X	X

Bioinformatics

- Gene expression
  - Microarray (RMA, LIMMA), Ontology and pathway (GSEA, KEGG), Batch removal (COMBAT, SVA), NGS DNA-seq (FastQC, BWA, STAR), RNA-seq (DESeq, Kalisto, Sleuth)
- Transcription factor regulation / Epigenetics
  - ChIP-seq / DNase-seq (MACS, BETA), methylation, HiC
- Genome-wide association
  - GWAS (Plink, GCTA), SNP trait array
- Whole genome sequencing (GATK, Mutect)
- Resources
  - DAVID, GEO, UCSC, CBioPortal

Methods / Algorithms

- Programming
  - R, Bioconductor, Python, Unix, git
- Statistical tests
  - median polish, distribution, qq-plot, t-test, χ^2 test, K-S test, FDR
- Modeling
  - Generalized linear models, logistic regression, LASSO, k-means and hierarchical clustering, PCA, SVM, Burrows-Wheeler alignment, EM, Gibbs sampling, HMM, dynamics programming, survival analysis

Bioinformatics tools

- Google is your good friend
- Bioconductor
- `source("https://bioconductor.org/biocLite.R")`
- `biocLite("PACKAGE_NAME")`
- `library(PACKAGE_NAME)`

High Performance Computing

- Impossible to do all computing on your laptop
- Amazon web service (EC2, RDS, S3, IAM) $$$
    - AWS calculator
- Harvard FAS Odyssey
- HMS Orchestra
- Partners ERISOne
- How is the submission looks like?

Workflow

1. Request for the account
2. Prepare tools
   - FileZilla
   - Putty for Windows, Terminal for Linux/Mac
3. Connect to the server
   - `ssh USERNAME@SERVER_ADDRESS`
   - `ssh -Y jimmy@orchestra.med.harvard.edu` (With X11)
   - `ssh jimmy@erisone.partners.org`
4. Install whatever you need
   - Odyssey and Orchestra have most of the common use tools, e.g. R, python, their packages, ...
   - Create virtual environment to install your tools

Workflow

5. (Optional) Go to interactive node
6. Upload/Edit your script
7. Create a submission file (`.lsf` for ERISOne, `.slurm` for Odyssey, ...)
8. How is the submission looks like?
    - `bsub -q normal -n 4 -R 'rusage[mem=2000]' < job.lsf`
    - `job.lsf` is a normal exec file
9. Use `bjobs` to check the status
10. Use `bkill` to kill the job

- Demonstration

Some Bioinformatics Learning Resources

- edx Bioinformatics
- If you don't have time to take online course
- Coursera Bioinformatics
- Harvard STAT115/BIO512: Computational biology and bioinformatics

Basic Programming / CompSci Courses

- Harvard CS50 (Malan)
- MIT 6.001
- Coursera Python class (Rice)
- Coursera Python class (U Mich)
- Codeacademy, Code school, Datacamp, ...
- Learn Python the Hard Way
- Python Guide

Some Advanced NLP Materials

- NLTK book (very useful!)
- Coursera Jurafusky
- Coursera Radev
- Coursera NLP provided by Michael Collins is also good, but it's gone now (you can try if you can find it)
- Harvard CS287: Natural Language Processing (Rush)
- MIT 6.864: Advanced Natural Language Processing (Barzilay)

Some Visualization Resources

- Harvard CS171: Visualization (Hanspeter)
- Harvard BMI706? (Gehlenborg)

Some Machine Learning Courses

- Coursera Andrew Ng
- Caltech LFD
- Harvard CS109: Data Science (2015 archive)
- Harvard CS181: Machine Learning (Rush, Parkes)
- MIT 6.036: Introduction of Machine Learning (Jaakkola)
- Stanford CS231n: Convolutional Neural Networks for Visual Recognition (Li)
- Bishop - Pattern Recognition and Machine Learning

Take Home Message

- Very brief overview of bioinformatics / computational biology (Courtesy by Dr. XS Liu)
- Tools and algorithms
- How to use cluster
- Some course/book recommendations
- Contact
  - Github repository
  - ckbjimmy@gmail.com
  - Linkedin: Wei-Hung Weng