Application of Bioinformatics in Genetics Research

Instructors:

Dr. Matt Gitzendannert

Dr. Raad Gharaibeh

Dr. Lei Zhou (Course director)

Course web page: http://zhoulab.net/GMS6014/home.html for classroom practices, homework, etc.

Application of Bioinformatics in Genetic Research

Time and location:

MWF : 12:00-1:00

CGRC-291

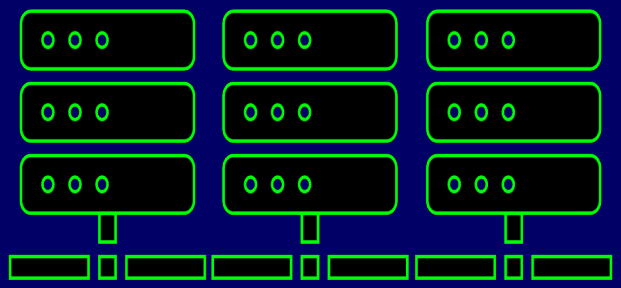
Evaluation

- 50% classroom participation
  - Discussion.
  - Be ready to share your screen.
- 50% homework

Required facility

- **Your own laptop**
  - **Browser(s)**
  - **text editor**
  - **Some programs**

FTP programs such as **FileZilla**



- **HiPerGator**
  - **All Linux/Unix programs**
  - **Large data set process.**

Practice

Practice: Download and install a text editor

Rule of thumb for doing your own bioinformatics:

- Make a folder for each program / project.
- Do NOT have space in folder and file name, consider using "_" to separate words.

History of bioinformatics – sequence analysis

- Sequence comparison
    - Similarity search
    - Phylogenetic analysis
- Structure predication
- Gene prediction
- Genomics, omics, and systems biology

Bioinformatics in the post genome era

The opportunity provided by genome sequence and genomic / proteomic technology is matched by the challenge to bioinformatics / computational biology

- **Information Representation.**

  - many new types of data, such as *Function*, *Location*, *Interaction*, *Regulatory pathway*, *Expression profile, etc. needs to be recorded*

- **Data Management**

  - Infrastructure for inputting, managing, access and retrieval of relevant information in a "sea of databases". Cloud computing.

- **Systematics**

Bioinformatics in the post genome era

- Whole genome sequencing - SNP and whole genome wide association studies.
- Genomic/proteomic expression profiling (RNA and protein levels).
- Epigenomics, Comparative genomics, …
- Regulatory pathway simulation – systems biology.

$1,000 genome and … $500,000 analysis ?

Overwhelmed by data?

Objectives of GMS6014

- Basic skills for retrieving and storing data, using web-based bioinformatics tools.
- Ability to install and run standalone local applications.
- Understanding the basis of bioinformatics applications using sequence similarity search as the example.
- An introduction to HTS analysis & HiPerGator

Sequence Representation - nucleotide

```
N G R C W T G Y C Y

A G A C A T G C C C
C   G   T     T   T
G
T
```

For complete list of IUB/IUPAC nucleic acid codes, see
http://www.ncbi.nlm.nih.gov/blast/fasta.shtml

Sequence Representation - amino acids

Q:

What's the common property of these amino acids ?

1. D, E
2. I, L, V, M, F

Sequence Representation - amino acids

Example:

```
W D L L A Q I L C Y A L R I Y
W R F L A T V V L E T L R Q Y
W K F L A I T M C K V L K Q F
R C L L C N K L Y Y L
L N R L L A E L Y E V
L R L L Q Q Q Q M V L
```

```
W D L L A Q I L C Y A L R I Y
W R F L A T V V L E T L R Q Y
W K F L A I T M C K V L K Q F
R C L L C N K L Y Y L L R K V
L N R L L A E L Y E V L C H I
L R L L Q Q Q Q M V L Q R Q Y
```

Coloring based on aa property.

Representation of sequence – sequence file format

1.) FASTA – simple and clean

> gene_name, (other info)
MASASASKJHKLJLKJLDSDFSF
SSDSASFSFD…

Practice / DIY: retrieve sequence in Fasta format and save the file in the local computer.

How to store sequence files

- Pure text format is clean and allows down stream sequence analysis.
- .doc or .rtf allows formatting during annotation – however, extra information are inserted thus NOT suitable for computational analysis.
- !! No space in file or folder name!!- Or trouble will find you.

Practice – file types

- Using file Finder (Mac) or Explorer (PC) to view downloaded files.
- Change the "Tools→Folder Options" so that the file **extensions (.xxx) are revealed**.
- Edit the downloaded sequence file in MS Word, highlight a section of the sequence with Bold font or color and save as .doc
- Open the .doc file in NotePad (PC) or TextEditor (Mac) – observe the inserted characters.

Practice – file types (Cont.)

- Load the "Mysequence.doc" file to Webcutter using "Choose file" and then "Upload sequence file".
  -Notice that the "sequence" in the sequence box are nonsense characters.
- Clear input; Browse and then load the .txt file. Run an analysis.

Keep you sequences as .seq or .fasta file for downstream analysis.

Representation of sequence

The need to represent associated info with sequence

- Structured data entry
- Specialized databases
  - 3-d Structure
  - Mutation / Diseases
  - Protein family / Protein domain
  - Interaction
  - Pathway
  - ….

Representation of sequence

The need to represent associated info with sequence

- Structured data entry
- Specialized databases
- Complex / customized data structure

  - Object-oriented data representation

(Mount, p44-45)