Tools and Algorithms in Bioinformatics

GCBA815/MCGB815/BMI815, Fall 2017

Week-2: Data Formats and UNIX Tools

Babu Guda, Ph.D.
Professor, Genetics, Cell Biology & Anatomy
Director, Bioinformatics and Systems Biology Core
University of Nebraska Medical Center

Topics:

- Standard data formats to represent protein/nucleic acid data
- Introduction to UNIX
- Hands-on practice with UNIX commands

Useful Links:

Tutorials Point:
http://www.tutorialspoint.com/unix/unix-getting-started.htm

UNIX tutorial for beginners
http://www.open-of-course.org/courses/course/view.php?id=45

Examples of Data Formats

- Sequence or annotation data
  - NCBI (GenBank)
  - UniProt/SwissProt
  - PIR
  - FASTA
- Protein structure data
  - CIF- Crystallographic Information File
- Genomics data
  - Fastq
  - GTF (Gene Transfer Format)

Verbose Formats

GenBank

```
LOCUS       AAB94881      842 aa                    BCT       05-JAN-1998
DEFINITION  DNA polymerase [Cenarchaeum symbiosum].
ACCESSION   AAB94881
PID         g2599106
VERSION     AAB94881.1  GI:2599106
DBSOURCE    locus AF028831 accession AF028831.1
KEYWORDS    .
SOURCE      Cenarchaeum symbiosum.
```

UniProt/SwissProt

```
ID   DHE2_CLOSY     STANDARD;      PRT;   449 AA.
AC   P24295;
DT   01-MAR-1992 (REL. 21, CREATED)
DT   01-APR-1993 (REL. 25, LAST SEQUENCE UPDATE)
DT   01-NOV-1997 (REL. 35, LAST ANNOTATION UPDATE)
DE   NAD-SPECIFIC GLUTAMATE DEHYDROGENASE (EC 1.4.1.2) (NAD-GDH).
GN   GDH.
OS   CLOSTRIDIUM SYMBIOSUM (BACTEROIDES SYMBIOSUS).
OC   PROKARYOTA; FIRMICUTES; ENDOSPORE-FORMING RODS AND COCCI; BACILLACEAE.
```

Condensed Formats

PIR

```
>P1;CATPAA Chloramphenicol acetyltransferase (EC 2.3.1.28) - E. coli
plasmids  M E K K I T G Y T T V D I S Q W H R K E H F E A F Q S V A Q  C T Y N
Q T V Q L D I T A F L K T V K K N K H K F Y P A F I  H I L A R L M N A H P E F
R M A M K D G E L V I W D S V H P  C Y T V F H E Q T E T F S S L W S E Y H D D
F R Q F L H I Y  S Q D V A C Y G E N L A Y F P K G F I E N M F F V S A N P
W  V S F T S F D L N V A N M D N F F A P V F T M G K Y Y T Q G  D K V L M P L
A I Q V H H A V C D G F H V G R M L N E L Q Q  Y C D E W Q G G A *
```

FASTA

```
>CATPAA Chloramphenicol acetyltransferase (EC 2.3.1.28) - E. coli plasmids
MEKKITGYTTVDISQWHRKEHFEAFQSVAQCTYNQTVQLDITAFLKTVKKNKHKFYPAFI
HILARLMNAHPEFRMAMKDGELVIWDSVHPCYTVFHEQTETFSSLWSEYHDDFRQFLHIY
SQDVACYGENLAYFPKGFIENMFFVSANPWVSFTSFDLNVANMDNFFAPVFTMGKYYTQG
```


PDB File format

```
ATOM      1  N   HIS A   1      49.668  24.248  10.436  1.00 25.00           N
ATOM      2  CA  HIS A   1      50.197  25.578  10.784  1.00 16.00           C
ATOM      3  C   HIS A   1      49.169  26.701  10.917  1.00 16.00           C
ATOM      4  O   HIS A   1      48.241  26.524  11.749  1.00 16.00           O
ATOM      5  CB  HIS A   1      51.312  26.048   9.843  1.00 16.00           C
ATOM      6  CG  HIS A   1      50.958  26.068   8.340  1.00 16.00           C
ATOM      7  ND1 HIS A   1      49.636  26.144   7.860  1.00 16.00           N
ATOM      8  CD2 HIS A   1      51.797  26.043   7.286  1.00 16.00           C
ATOM      9  CE1 HIS A   1      49.691  26.152   6.454  1.00 17.00           C
ATOM     10  NE2 HIS A   1      51.046  26.090   6.098  1.00 17.00           N
ATOM     11  N   SER A   2      49.788  27.850  10.784  1.00 16.00           N
ATOM     12  CA  SER A   2      49.138  29.147  10.620  1.00 15.00           C
ATOM     13  C   SER A   2      47.713  29.006  10.110  1.00 15.00           C
ATOM     14  O   SER A   2      46.740  29.251  10.864  1.00 15.00           O
ATOM     15  CB  SER A   2      49.875  29.930   9.569  1.00 16.00           C
ATOM     16  OG  SER A   2      49.145  31.057   9.176  1.00 19.00           O
ATOM     17  N   GLN A   3      47.620  28.367   8.973  1.00 15.00           N
ATOM     18  CA  GLN A   3      46.287  28.193   8.308  1.00 14.00           C
ATOM     19  C   GLN A   3      45.406  27.172   8.963  1.00 14.00           C
```

Fastq format to represent short-read data

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

~ Highest sequencing quality
! Lowest sequening quality

Order of quality value characters from lowest to highest

```
 !"#$%&'()*+,-./0123456789:;
<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```


GTF- Gene Transfer Format

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

Fields

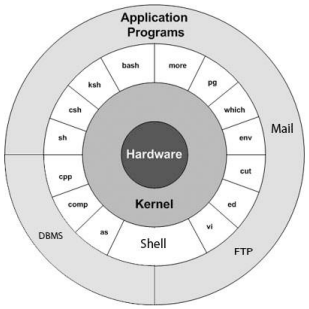
Fields **must** be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '.'

1. **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

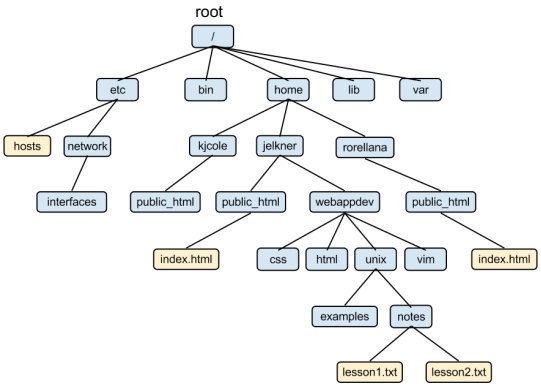
UNIX Architecture

- Developed in 1969 at Bell Labs
- Variants of UNIX- Solaris, AIX, HP-UX, Linux, etc.
- Linux is freely available and popular

- Kernel: interacts with hardware
- Shell: interacts with the user, interprets the commands and calls the programs
- Command: the language to interact with the operating system
- Files: Files are organized into directories in a tree-like structure called the file system



UNIX File Directory Structure



Using the UNIX server

- Download and install SSH Secure Shell client on your desktop
- https://shareware.unc.edu/pub/win/SSHSecureShellClient-3.2.9.exe
- Server name: cbsb.unmc.edu
- Use the supplied credentials to login
- Change your password
- Copy class files from: /storage/share/GCBA815/

```
$ passwd
Changing password for amrood
(current) Unix password:******
New UNIX password:*******
Retype new UNIX password:*******
passwd: all authentication tokens updated  successfully

$
```


Characteristics of UNIX

- Case-sensitive (Amber is different from amber)
- Very short but intuitive (cp for copy, ls for list, etc)
- Multi-functional based on the context ('mv' for move or rename)
- Versatile: perform many tasks depending on the arguments
  - ls command followed by arguments, –a, -l or -al
- Can create a workflow by stitching multiple commands together
  - cat file1 file2 >file3
  - grep ^ID file1 | grep -i 'human' |