Practical Bioinformatics (BIOL4220) - Exam 1 Topics

Exam 1 covers topics covered between Lab 01A and Lab 06B. To prepare, you should know:

- How to translate written instructions to Unix command(s)
- How to use the following Unix commands, including the general format of the input they accept, and what output they produce:

  echo, ls, pwd, cat, mv, cp, rm, cd, mkdir, rmdir, cp -R, rm -rf, man, wc, head, tail, diff, grep, sort, cut, uniq, tr, rev

- How to navigate the filesystem and how to locate filesystem resources using relative and absolute file paths, wildcards (`?` and `*`), and special directories (`~`, `.`, and `..`).
- How to use the following Git commands, and what effect they have on the files and branches on local and remote repositories:

  git status, git log, git add, git commit, git diff, git push, git pull, git branch, git merge, git checkout, git revert

- How to read a git status, and determine what commands are needed to synchronize the local repository with a remote repository (e.g. on GitHub).
- How to interpret a Git commit history graph and commit log; in particular, how to identify how commit, branch, and merge commands shape the graph.
- How to use redirects (`>`,`>>`,`<`) and pipes (`|`) to construct complex tasks involving multiple programs and files.
- How the following shell programming features work:

  variables, if-statements, for-loops, user arguments, command substitutions, comments

- How to read a script, how to describe its overall purpose, and how to annotate key commands comments to make the script more human-readable
- How to write and run a script to solve a plainly stated objective
- How to translate nucleotides into amino acids when provided a table for a genetic code.
- How to interpret the basic features of a GenBank record for a nucleotide accession.

- How the relative costs for matches, mismatches, gap-opens, and gap-extensions influence the general structure of an alignment produced by heuristic methods (e.g. MAFFT)

- How to convert a phylogenetic scenario that involves substitution, insertion, and deletion events into an alignment (e.g. such as those scenarios presented in the PRANK paper)

- What the following symbols mean within regular expressions:

  *. \d \w \s [ab] [^a] [m-z] {m} {m,n} * + ? ^ $*

- How to use regex groups in the following contexts:

  (a)bc (a)bc(d) ((a)b)cd (abc|def) (abc){3}

- How to regex to search with `grep`, and how to find-replace with `sed`, including with capture groups and back-references

- How to interpret relationships among taxa from a phylogeny, and how to convert a phylogeny (with branch lengths) into a Newick string

Below are questions that are identical or similar to questions in the lectures/labs that might appear on Exam 1:

- If your filesystem contains

  ```
  /home/mlandis/Biol4220
  /home/mlandis/Biol4220/notes.txt
  /home/mlandis/Biol4220/labs
  /home/mlandis/Biol4220/labs/lab_01A.pdf
  /home/mlandis/Biol4220/labs/lab_01B.pdf
  /home/mlandis/Biol4220/lectures
  /home/mlandis/Biol4220/lectures/lect_01A.pdf
  /home/mlandis/Biol4220/lectures/lect_01B.pdf
  /home/mlandis/Biol4220/lectures/lect_02A_draft.pdf
  ```

  … then name all known directories, all known files, the directory that contains three files, and the directory that contains two other directories.

- Still working with the above filesystem:
  - What single command would delete all pdf files in `/home/mlandis/labs` if you were located in `/home/mlandis/lectures`?

    - If you were located in `/home/mlandis`, how would you move the folder `labs` into `lecture`, and what would the absolute file paths for the lab pdfs become?

- What Git commands would I need to type to ensure data.txt, output.txt, and run.sh have all been saved to the commit history, each in *one of three separate* commits, and that those commits have been replicated to a remote repository (e.g. GitHub)?

- What is the expected output for the following commands?
    - echo "aGcttAcGCaTaC" | tr "t" "u" | tr "T" "U"
    -

- How would you extract a sorted data table that lists the scientific name and adult body mass for all species in order Monotremata in tab-delimited format, then save that output to monotreme_mass.tsv? Example input:

```
$ head data/mammal_data.csv
Order;Scientific_name;AdultBodyMass_g;Max_longevity_d
Rodentia;Eligmodontia typus;17.37;292
Rodentia;Microtus oregoni;20.35;456.25
Rodentia;Peromyscus gossypinus;27.68;471.45833335
Macroscelidea;Elephantulus myurus;59.51;401.5
Rodentia;Peromyscus boylii;23.9;547.5
Rodentia;Phodopus campbelli;27.06;653.95833335
Rodentia;Myodes gapperi;19.83;608.33333335
Eulipotyphla;Sorex palustris;13.07;547.5
Rodentia;Reithrodontomys humulis;8.25;817.90416665
```

- Write a pipeline to compute the number of uniquely named .txt files in a directory that do not contain the text “ignore” in the file name.

- Write a script that swaps the names of two files. For example, if file1.txt contained the text “Hello” and file2.txt contained the text “world!”, then after calling the script part_3/problem_1/run.sh, file1.txt would contain the text “world!” and file2.txt would contain the text “Hello”. Running the script should not result in any other lasting changes to the filesystem (e.g. new permanent folders and/or files, etc.)

- Translate the codons TTA ATT ACC CCA GAA into amino acids using the table for the genetic code, below. (Note, T is encoded as U during transcription of DNA -> RNA).

- Suppose you have the codon AGA, and it was mutated by a single-character nucleotide substitutions. What substitutions ***would not*** change which amino acid that codon encoded?

RNA codon table

1st position	2nd position U	2nd position C	2nd position A	2nd position G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Amino Acids

Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp:Aspartic acid
Cys:Cysteine
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine
Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline
Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosisne
Val: Valine

- You are provided the following file:

  $ cat file.txt
  ```
  gene
  genre
  generic
  energy
  energetic
  ```

  Give a regular expression that only matches “energy and energetic”. Give a regular expression that only matches “generic” and “energy”. (You cannot use grouped OR statements like `(generic|energy)` as solutions.)

- Write a small script using regex that reports whether a GenBank accession is a valid Nucleotide record or not. Valid Nucleotide records follow the format A######, AA#######, or AA#########, with A representing any capital letter and # representing any numeral.

- Write a Unix pipeline that uses grep to print the word count for all occurrences of words that begin with a vowel and end with the letters “ing”. Only match words that are flanked by whitespace characters and are printed in their entirety on one line (i.e. ignore linewrapped words). For example “ Eating “ and “ eating “ would match, but “ eating.” and “ eating,” and “ singing “ would not.

- Report all instances of motifs that repeat xA between 5 and 6 times, and are flanked by the basepair C on both sides; only list the repeating region, not the flanking region. For example, CTATATATATAC would match, and the printed motif would be TATATATATA. The motif CGATAGACATAC would match, and be printed as GATAGACATA. The motif CTATATATATAG would not match.

- You are given the Newick string

  (B:0.12,(((A:0.63,D:0.51)0.99:0.71,F:0.55)0.35:0.32,(C:0.18,E:0.77)0.85:0.46)0.32:0.33);

  Draw the phylogeny as a tree, annotated with branch lengths and clade support. Name the taxa that belong to each of the five clades. Also, report the expected number of substitutions/site that separate taxa A and D. Is that number greater than the expected number of substitutions/site that separate taxa C and E?