Noah Cape pest/lab 01
compBio 315
Feb 13, 2023

Problem Set Questions

0) Unix commands. Identify the function of these Unix/Terminal/operators:
`ls` list the contents of your current dir

`cd` change directory

`cp` copy contents of a source file into a destination file

`mv` move files

`rm` removes non directory files

`mkdir` creates a directory in current dir

`pwd` prints current working directory

`scp` secure file copy between hosts on network

`lpr` print files, sent to default destination or named dest

`cat` concatenate and print files, read files and writing them to standard output

`more` forward movement reading through a file in cmdline

`grep` search given input files with specified pattern

`head` display the first n lines of a file (default to 10)

`tail` display the last n lines of a file (defualt to 10)

`man` read online manual pages

`chmod` change file modes and access control lists

`|` Pipe connecting stdout of command to stdin of next command

`&` Run command in the background, shell does not wait for command to finish

`>` Output redirection operator, overwriting files contents

`<` Input redirection

`>>` Output redirection, concatenate to files current contents

`*` Match zero or more characters

`?` Match one character

`;` Run multiple commands and execute the second no matter the status of the first

1) Get to know DNA

(a) Directionality of the DNA backbone: The two DNA strands run opposite to eachother, 5'->3', major strand, then the minor strand which is upside down compared to major stand where the major stand is 5' at top minor is 3'.

(b) The four bases of DNA are A, T, G, and C, RNA has A, U, G, and C. The two groups that bases fall into are purines (A, G) and prymidines (C, T, U). Purines consist of two carbon nitrogen ring bases and Pyrimidines consists of one.

(c) Chromosome: Strucutre of DNA molecules and proteins, Gene: One contiguous stretch of DNA to corresponds to a different kind of protien., Operon: Unit made of linked genes, Codon: nucleotide triplet, Nucleotide: Basic DNA molecule consisting of sugar, phosphate and its base.

(d) Typical nucleotides in chromosome: 50 mbp to 250 mbp, gene: 10-15 kpb, operon: 5 * avg gene length ~> 50-75 kbp, codon: 3.

(e) What makes DNA an acid: the phosphate groups on in the DNA backbone have a net negative charge, since the basic component (nitrogen groups) form the inside of the double helix therefore with the phosphate groups having an negative charge and being exposed to the environment compared to the nitrogen bases makes DNA an acid.

2) Define the terms

- Oligo: polynucleotide which contains a relatively small number of nucleotides
- primer: short single stranded DNA fragment used for lab techniques, usually target certain places in genome to bind to (18-25 bp)
- dNTP: Deoxynucleotide triphosphates, bulding block of DNA, lose two phosphate groups when incorperated into DNA during replication.
- ddNTP: dideoxynucleotides lack the 3' hydroxyl group that inhibit further polymerization of DNA backbone, used in Sanger sequencing method.
- electrophoresis: Lab technique to seperate DNA, RNA, or protein molecules based on size and electrical charge. Smaller molecules move through gel faster than larger molecules.

- restriction enzyme fragment: Fragment of DNA from cutting of DNA by restriction enzyme. Matches to certain patterns in DNA and cuts DNA at those sites.
- pyrosequencing: Sequencing method of DNA that detects light emitted during the sequencial addition of nucleotides.
- primer walking: Sequencing technique that uses a series of Sanger sequencing reactions to clone a gene.
- ligation: Joining of two nucleic acid fragments through the action of an enzyme.
- mate-paired ends / paired-end sequencing: Methodologies that give information about two read belonging to a pair. The DNA is sheared into random fragments and then both ends of each fragment are sequenced. Mate pair tags that are sequenced belong to a larger molecule (btw 2 and 10 kbps) In Paired end sequences both reverse and forward templates, each end is seperately sequences and the two sequences are known as paired end reads. Distance between paired end reads is about 300bp.

3) Key innovations in DNa sequencing technologies:
- Paper suggests "the first implementation of this approach produces approximately 10 billion reads per run, with a turnaround time of under 20 hrs per run for 300 bp reads, and with base quality similar to existing platforms (Q30 >85%), at a price of $1/Gb."
- Longer reads (300bp) opening up analysis that cannot be achieved in short reads (100bp)

4) DNA with p_a = p_c = p_g = p_t = 0.25
- Liklihood of start codon (0.25 * 0.25 * 0.25) = 0.015625
- 64/3 bp (~21) If a stop codon happens occurs 3/64

Lab Questions:

Nucleotide Frequencies:

In bacteria there is a higher GC content, and in human there is a higher AT content
Human Chromosome 1 the highest (p(N_1N_2)/p(N_1)p(N_2)) is CG

ORF:

In bacteria if an ORF is longer than 190 (Human ~205), I have some confidence that it is not just random (using a chi square test)
The distribution of ORF lengths is clearly exponential, tending quicly to zero. Huge amount of ORFs in the 1-(~60) range, then a fast drop off. In the human chromosome 1 there is an ORF that is 300,000 bp long. See photo of MG1655

Count:

If I ran the command ("ATGCCCCTAT).count("CC") it would only count 2. Therefore count does not use a sliding window search, and no overlapping values. For nucleotide it is the same, and the same for dinucleotides that are not double of same base. But misses a bunch with same base.

Human Chromosome 19

```
{'A': 24.3377, 'T': 24.4025, 'G': 22.8503, 'C': 22.7947}
```

```
AT 0.8512720988254474
AG 1.1928805739656705
AC 0.8157860298998123
AA 1.1405659639840535
TG 1.213987264920778
TC 0.980170019539124
TA 0.6747116711251331
TT 1.1425715194290085
GC 0.9750052100688936
GA 0.9831262059818626
GT 0.8151171531373451
GG 1.2403476626188779
CA 1.215065883541543
CT 1.1915017914465016
CG 0.3240467974618719
CC 1.2429683590624645
Most Exceptional: CG
```

Human Chromosome 13:

```
{'A': 26.2736, 'T': 26.3742, 'G': 16.5555, 'C': 16.473}
```

Way lower GC content

```
AT 0.8923265607328001
AG 1.1394642385281524
AC 0.8524700465504546
AA 1.1127032075494108
TG 1.200773019188618
TC 0.9856701226739538
TA 0.7665252859943031
TT 1.115505307749047
GC 1.0326232313411714
GA 0.9906829298966879
GT 0.848570154351487
GG 1.2235644207009015
CA 1.2034131528139644
CT 1.1389910377136299
```

```
CG 0.23142465818722607
CC 1.225458974484744
Most Exceptional: CG
```

E Coli.

```
{'A': 24.6193, 'T': 24.59, 'G': 25.3668, 'C': 25.4239}
```

Much higher GC content

```
AT 1.1030241182542915
AG 0.8210836166140543
AC 0.883808642778375
AA 1.201435906353755
TG 1.1134497505616
TC 0.921466055442626
TA 0.7545332655226582
TT 1.2099230707202004
GC 1.2831161406396345
GA 0.9224060103638554
GT 0.8831468171588427
GG 0.9048282166424415
CA 1.1197704366470012
CT 0.8137896652021194
CG 1.1584831522898624
CC 0.9059924661357719
Most Exceptional: GC
```

In all cases GC/CG is the most exceptional, in human because it is so much lower than expected and in E Coli, because it is higher