Hyun-Joon Yang

yanghyun@usc.edu

BISC 478

HW1

Q1

Compare the sequencing accuracy, average read length, and sequence throughput of Illumina Sequencing, Oxford Nanopore, and PacBio.

- Illumina sequencing allows short reads that are around ~200-300 bp through synthesis. The reads tend to be highly accurate with ~0.1% error rate. High throughput is achieved using bridge amplification.
- Oxford Nanopore allows long reads that are around ~8 kb. The reads have accuracy at around ~85-90%. High throughput is achieved by having nanopores in parallel which have motor proteins that feed the fragment during which electrical signal is used to identify the sequence.
- PacBio allows long reads that are around ~10-16 kb. The reads have accuracy at around 15%, similar to Nanopore, but higher than Illumina. High throughput is achieved using SMRT cells that have wells in each of which is a single molecule of DNA that get sequenced through synthesis.

Q2

Given a random DNA sequence of length N. Assuming equal nucleotide probability

A) [5 points] What is the expected number of occurrences of a k-mer in this sequence? A k-mer is a nucleotide sequence of a certain length k.

binomial distribution formula

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

in this case, $p = (\frac{1}{4})^k$, $n = N - k + 1$

expectation formula

$$E[X] = \sum_{x=1}^{n} x P(X = x)$$

therefore

$$E[X] = \sum_{x=1}^{N-k+1} x \binom{N-k+1}{x} (\frac{1}{4})^{kx} (1 - (\frac{1}{4})^k)^{N-k+1-x}$$

B) [5 points] What is the probability of observing m occurrences of k-mer in this sequence?

$$P(X = m) = \binom{N-k+1}{m} (\frac{1}{4})^{km} (1 - (\frac{1}{4})^k)^{N-k+1-m}$$

Q3

Finding gene information using UCSC Genome browser

Shown below are some of the top studied human genes:

- TP53: tumor protein p53
- IL6: interleukin 6
- ESR1: estrogen receptor 1
- EGFR: epidermal growth factor receptor
- BRCA1: breast cancer 1
- ApoE: apolipoprotein E
- TNF: tumor necrosis factor
- VEGFA: vascular endothelial growth factor
- AKT1: protein kinase B
- MTHFR: methylenetetrahydrofolate reductase

A) [5 points] Select two genes from this list and read about them. Comment why these genes are important?

- TP53
  - TP53 is a gene that encodes for the protein p53 which is crucial in suppressing tumors and thus in cancer prevention.
- AKT1
  - AKT1 is a gene that encodes for ATK1 kinase (aka protein kinase B) which is important for signaling pathways such as regulating cell proliferation, differentiation, and cell survival.

Now pick one gene out of the two. Go to UCSC Genome Browser and select the Human hg38 genome assembly in the Genomes tab. In the search box, type the gene name you picked and select the gene. Use the GENCODE v32 gene track to answer the following questions:

B) [2 points] On what chromosome is this gene?

chromosome 17

C) [2 points] This gene is present on which strand (forward/reverse)?

reverse strand

D) [1 point] How long is the gene in base pairs?

19,137 bp

Q4

In-silico PCR
Polymerase chain reaction is a method used to make many copies of a specific DNA region. In a PCR reaction, the region of DNA that will be copied, or amplified, is determined by the "primer" sequence. PCR primers are short pieces of single-stranded DNA, usually around 20 base pairs in length. Two primers are used in each PCR reaction, and they are designed so that they are complementary to the sense and anti-sense strands of the target DNA.

In-silico PCR is a way to computationally predict what sequences will be amplified without having to perform PCR and sequence the results. We will use the In-Silico PCR tool from the UCSC genome browser. Navigate to:

UCSC Genome Browser → Tools → In-Silico PCR

Make sure you are viewing the human genome version GRCh38/hg38 (hover over "Genomes", and select Human GRCh38/hg38 if this is not the case).

For the forward (left) primer (5'→ 3'), enter:
AGATATTATTTCAGAAGAGAT

For the reverse (right) primer (5'→ 3'), enter:
TGGCTTGGAACGTACATTTAC

Then click submit. Go through the results carefully and pay attention to the regions where the primers bind (upper case indicates matching regions). On the result page, the blue text is a link. Click on it, and

A) [5 points] Describe any salient features of the region that was amplified (e.g. genes). Note, you can click on any gene and it will show you annotations of known functions of the gene, disease associations.

The region contains TAS2R38 gene which is located on the reverse strand of chromosome 7 and is 1,143 bp long. This gene encodes a seven-transmembrane G protein-coupled receptor that controls the ability to taste glucosinolates.

B) [10 points] We can use the UCSC browser to retrieve the DNA sequence for a region. In the search box/position bar you enter an interval:

chr22:23,253,797-23,253,980

Click on click on View, then DNA → get DNA. This will provide you the DNA sequence for this region.

Now design 20-base pair primer sequences (forward and reverse) that would amplify the interval:

chr22:23,253,797-23,253,980

You can verify your results using In-Silico PCR.

Forward: AGATGGCTCGTTCGGAACAC

Reverse: CGCTTTAGTGGACTCCAGGG

Q5

A) [5 points] A base has a PHRED quality score of 20. What is the probability that this base was incorrectly sequenced?

score formula

Q = 20

$$Q = -10 \log_{10}(p)$$

$$20 = -10 \log_{10}(p)$$
$$p = 10^{-20/10}$$
$$p = 1/100 = 0.01$$

B) [10 points] An Illumina FASTQ file has an entry:

@ERR894729.1874/1

AGTCTGTAAAATGACTCTGA

+

????BBBBB($$?????$??

Using the quality score encoding listed in this URL, calculate the expected number of sequencing errors in this read.

expected error

$$E = \sum_{i=1}^{N} p_i$$

In [1]:
```
seq = "AGTCTGTAAAATGACTCTGA"
qs = "????BBBBB($$?????$??"

E = 0
if len(seq) == len(qs):
    for i in qs:
        Q = ord(i) - ord('!')
        p = 10**(-Q/10)
        E += p
        print(i, Q, p, E)
    print(E)
else:
    print('lengths of sequence and quality score do not match')
```

```
? 30 0.001 0.001
? 30 0.001 0.002
? 30 0.001 0.003
? 30 0.001 0.004
B 33 0.0005011872336272725 0.004501187233627272
B 33 0.0005011872336272725 0.005002374467254544 5
B 33 0.0005011872336272725 0.005503561700881817
B 33 0.0005011872336272725 0.006004748934509089
B 33 0.0005011872336272725 0.006505936168136361
( 7 0.19952623149688797 0.20603216766502433
$ 3 0.5011872336272722 0.7072194012922965
$ 3 0.5011872336272722 1.2084066349195688
? 30 0.001 1.2094066349195687
? 30 0.001 1.2104066349195686
? 30 0.001 1.2114066349195685
? 30 0.001 1.2124066349195683
? 30 0.001 1.2134066349195682
$ 3 0.5011872336272722 1.7145938685468405
? 30 0.001 1.7155938685468404
? 30 0.001 1.7165938685468403
1.7165938685468403
```

$$E = 1.7166$$