Homework 6

Dieu My Nguyen | MCDB 5520 | April 30, 2018

1. (20 pt: NOTE this question is worth twice normal value) During an analysis of a promoter you identify six sites (shown below) that alter the expression of the promoter when they are deleted.

ACGGAG
ACGTGG
AGGAAG
AGGCAC
ACGCAC
AGGGAC

(a) (3pt) Since the number of sites is small, we will include pseudocounts distributed according to the background nucleotide frequencies in the genome. Assume you are working in a genome where %A=%T=20%, and %G=%C=30%. The sum of all pseudocounts per position should be 1. What are the pseudocount values for each nucleotide?

A: 0.2
C: 0.3
G: 0.3
T: 0.2

(b) (4pt) Fill in the nucleotide count matrix, N(b,i) for this multiple alignment, including pseudocounts.

In [22]:
```
# Import modules
from Bio import motifs
from Bio.Seq import Seq
import numpy as np
```

In [11]:
```
# Set 6 sites
sites = [
        Seq("ACGGAG"),
        Seq("ACGTGG"),
        Seq("AGGAAG"),
        Seq("AGGCAC"),
        Seq("ACGCAC"),
        Seq("AGGGAC")]
```

In [17]:
```
# Set background frequencies
bg_freq = {
    "A" : .2,
    "T" : .2,
    "G" : .3,
    "C" : .3,
}
```

In [18]:
```
# Create motif object
m = motifs.create(sites)
```

Raw count matrix (position frequency matrix, PFM) without pseudocounts:

In [19]:
```
# Note: indexing with 0
print(m.counts)
```

```
        0      1      2      3      4      5
A:   6.00   0.00   0.00   1.00   5.00   0.00
C:   0.00   3.00   0.00   2.00   0.00   3.00
G:   0.00   3.00   6.00   2.00   1.00   3.00
T:   0.00   0.00   0.00   1.00   0.00   0.00
```

Add pseudocounts:

$$N_{b,i} := N_{b,i} + \beta q_b$$

where

$$\beta = 1 \text{ and } q_A = q_T = 0.2, q_G = q_C = 0.3$$

In [28]:
```
# Add pseudocounts to PFM
new_PFM = dict(m.counts)    # copy raw count matrix

for key, val in new_PFM.items():    # add pseudocount from background freque
    pseudo_count = bg_freq[key]
    new_PFM[key] = np.array(val) + pseudo_count
```

Count matrix with pseudocounts:

In [29]:
```
new_PFM
```

Out[29]:
```
{'A': array([6.2, 0.2, 0.2, 1.2, 5.2, 0.2]),
 'C': array([0.3, 3.3, 0.3, 2.3, 0.3, 3.3]),
 'G': array([0.3, 3.3, 6.3, 2.3, 1.3, 3.3]),
 'T': array([0.2, 0.2, 0.2, 1.2, 0.2, 0.2])}
```

(c) (5pt) Now convert the above counts matrix into a probability matrix, recalling that

$$\mathrm{P(b,i)} = \mathrm{N(b,i)} / \sum_{k=1}^{4} N(k,i)$$

Sample calculation for A at 0th position:

$6.2/(6.2 + 0.3 + 0.3 + 0.2) = 6.2/7 = 0.89$

In [30]:
```
pwm = m.counts.normalize(pseudocounts=bg_freq)
```

In [33]:
```
# Python rounded these to 2 decimals
print(pwm)
```

```
        0     1     2     3     4     5
A:   0.89  0.03  0.03  0.17  0.74  0.03
C:   0.04  0.47  0.04  0.33  0.04  0.47
G:   0.04  0.47  0.90  0.33  0.19  0.47
T:   0.03  0.03  0.03  0.17  0.03  0.03
```

(d) (5pt) Now convert the above probabilitiy matrix into a scoring matrix, recalling that S(b,i)=log[P(b,i)/P(b)], where P(b) is defined by the genome's nucleotide frequencies.

Sample calculation for A at 0th position:
$log_2(0.89/0.2) = 2.15$

In [35]:
```
pssm = pwm.log_odds(bg_freq)
```

In [36]:
```
print(pssm)
```

```
        0     1     2     3     4     5
A:   2.15 -2.81 -2.81 -0.22  1.89 -2.81
C:  -2.81  0.65 -2.81  0.13 -2.81  0.65
G:  -2.81  0.65  1.58  0.13 -0.69  0.65
T:  -2.81 -2.81 -2.81 -0.22 -2.81 -2.81
```

(e) (3pt) Consider the following two new sequences:

Sequence 1: TCGGAG
Sequence 2: ACTGAG

Based on the scoring scheme determined in part D, which of these two sequences is a better fit to this motif model?

Sequence 2 is better with a higher score.

In [38]:
```
seq_1 = Seq("TCGGAG", m.alphabet)
seq_2 = Seq("ACTGAG", m.alphabet)
```

In [42]:
```
print("Sequence 1 score: " + str(pssm.calculate(seq_1)))
print("Sequence 2 score: " + str(pssm.calculate(seq_2)))
```

```
Sequence 1 score: 2.1060903
Sequence 2 score: 2.6679692
```

2. (10pt) You notice that only 60% of the peaks detected by ChIP-Seq have the known

transcription factor motif nearby.

(a) (5pt) Give at least two explanations for the remaining 40%.

• False positive ChIP signals. The 40% of peaks may not indicate true motifs with proteins binding. The false discovery rate (FDR) chosen to perform peak calling might have been too high and thus substantial enrichments were classified as peaks.

• Novel genes are that epigenetically inactive and/or haven't been discovered/annotated.

(b) (5pt) How would you test (experimentally or computationally) for the explanations you suggested in part A?

• Use stricter FDR to get the best peaks that may hold biological truth. There might also be errors in the reference genome. For example, multicopy sequences may have been incorrectly assembled into a single copy. If this is the case, we could try the masking procedure documented in this paper: False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions by Pickrell et al (2011)

• Match the unknown motifs to known motifs by HOMER using similarity. If the unknown motif is conserved, the motif finding is relatively easier than if mutations have occurred at this location.

• Overall, repeating the study experimentally and computationally is usually a good idea.

3. (10pts)Your colleague professor Stu Dent generated a genome-wide DNA methylation map for normal colon cells using MRE-seq (a restriction enzyme approach) and MeDIPseq (an immunoprecipitation approach). In an intergenic region, he found an interesting locus. This locus is about 20kb. On one end of the locus, there is a 2kb CpG rich stretch that has both relatively low MRE-seq and MeDIP-seq signals. The rest 18kb has high level of MeDIP-seq signals.

(a) (3pt) Why might you suspect that this region encodes for a novel gene?

High MeDIP-seq signals indicate high methylation levels in the 18kb region of the locus. CpG methylation is a central mechanism of epigenetic gene regulation. This region with high CpG methylation may indicate low expression level of a possibly novel gene that is inactivated epigenetically.

(b) (4pt) You decide to look at histone modification patterns across this region for more evidence. There are several genome-wide datasets available for this cell type: H3K4me1, H3K4me3, H3K27me3, H3K9me3, H3K36me3, and H3K9Ac. Which histone mark would you investigate for this locus and why?

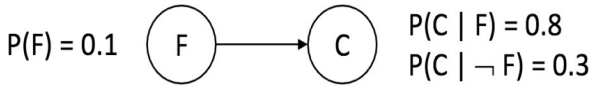
I'd use H3K27me3, which is known to inhibit transcription. When it's trimethylated, it's associated with inactive gene promoters. We can use this histone mark to look for this inactive gene that may be novel. (source: https://www.ncbi.nlm.nih.gov/pubmed/21652639 (https://www.ncbi.nlm.nih.gov/pubmed/21652639))

(c) (3pt) Suggest at least one other source of data (NOT annotation or bisulfide sequencing)

that may help you confirm or refute your suspicion, and why you think it may help?

Since our candidate gene is unknown and possibly novel, we may have 2 approaches for DNA methylation analysis: 1) whole genome methylation profiling and 2) searching for differentially methylated regions. For the latter, we can do bisulfide sequencing, but if not, we can use enzyme cleavage or m-CIP. Cleavage of DNA by a restriction enzyme may be blocked or impaired when the recognition sequence is methylated. From the effect of the cleavage activity of the commonly used pair of restriction enzymes MspI and HpaII, we could infer the sequence's DNA methylation status, allowing locus-specific discrimination between the methylated and unmethylated sequence. Otherwise, we could use another bisulfite-free method, Methylated CpG immunoprecipitation or m-CIP, to obtain enrichments of methylated CpGs by using a protein that binds methylated CpGs in DNA (MBD2). To confirm our suspicion, we would hope to see high levels of methylated CpGs in the region of interest.

4. (10pt) Consider the following simple Bayesian network diagram:



Where F is "having the flu" and C is "coughing" and both are binary (yes/no) variables.

(a) (3 pt)Given the diagram, what is the probability of both having the flu and a cough e.g. P(F,C)?

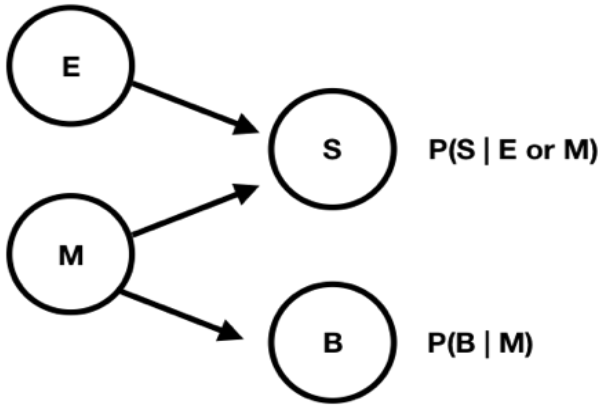
$P(C|F) = P(C,F) / P(F)$
$P(C|F) \times P(F) = P(C,F)$
$(0.8)(0.1) = 0.08$

(b) (3 pt) What is the probability of having a cough e.g. P(C) from the diagram?

$P(C) = P(F,C) + P(\text{not } F,C)$
$P(\text{not } F,C) = P(\text{not } F) \times P(C \mid \text{not } F) = 0.9 \times 0.3 = 0.27$
$P(C) = 0.08 + 0.27 = 0.35$

(c) (4 pt)Draw a Bayesian network to encode the following statement. Be sure to state all assumptions included in your diagram.

"A smell of sulphur (S) can be caused either by rotten eggs (E) or as a sign of the doom brought by the Mayan Apocalypse (M). The Mayan Apocalypse also causes the oceans to boil (B)."



5. Read "The what, where, how and why of gene ontology -- a primer for bioinformaticians" (PDF is available on Canvas as GOprimer.pdf) and answer the following questions:

a) (1pt) What are the three ontologies in GO?

Molecular function, biological process, and cellular component.

b) (1pt) The relationships within GO form what kind of graph?

Directed acrylic graph (DAG)

c) (1pt) What is the main difference between the full and filtered GO?

The filtered GO doesn't contain any "has_part" or inter-ontology relationships. Many analysis tools can only use the filtered GO.

d) (1pt) Who does most of the current GO annotation?

Professional curators examining the literature.

e) (1pt) According to Figure 5, what is the most common type of evidence for information within GO?

The IEA evidence code.

f) (2pt) Give one pro and one con of the information content measure of GO terms.

Pro: IC-based measures are less influenced by the idiosyncrasies of the ontology structure than the graph-based measures

Con: IC-based measures are still biased, because some terms are used more often and some research areas receive more attention than others.

g) (3pt) List 3 major criticism of GO.

Annotations only describe the normal, healthy functioning of genes. Data on functional coordination between multi-function genes are not explicitly stored. Until recently no relationships between the 3 ontologies were recorded. Now, inter-ontology relationships are only recorded in the full GO, which is not used by all analysis tools.

6. (Advanced: 10pt) There are two competing methods for chromatin state annotation, Segway

(Hoffman et. al. Nat. Methods 2012) and ChromHMM (Ernst & Kellis Nat. Biotech. 2010). Interestingly, the two groups came together to publish a paper (on Canvas) that compares the two approaches (Hoffman et. al. NAR 2012). Read the NAR paper and answer the following questions:

A. (3pt) Describe the difference between supervised and unsupervised methods.

Supervised methods find instances of one or more pre-determined classes of elements. These methods have been widely used for automatic gene finding that can recognize protein-coding transcripts using sequence features, cDNA sequence and evolutionary conservation of known examples. But because these methods require a training set of known examples, they cannot discover novel types of functional elements.

Unsupervised methods seek to simultaneously discover functional classes and annotate their instances de novo, without needing previously defined classes or known examples. In this way, they can avoid bias toward well-understood phenomena.

B. (2pt) According to Figure 3, in what state do most phenotype-associated SNPs reside?

Mostly in Enh state.

C. (2pt) The authors avoid (quite emphatically) declaring either method as superior. Based on the results presented, which method is better and why?

The authors advise the reader to examine both chromatin state annotations in regions of interest since both might capture a different aspect of the biology. Segway produces smaller segments than ChromHMM. Segway has high resolution, while ChromHMM has high continuity.

Some considerations of the results: For genes, both methods devote to identifying and characterizing protein-coding genes. For promoters, Segway can find high-resolution patterns at TSSs localized to the level of stable nucleosome-free promoter regions, while ChromHMM finds larger TSS-associated segments. Both discovered candidate transcriptional terminators, but here, ChromHMM also discovered the accumulation of a Pol2 label. Taken together, since Segway and uses a variety of data types as input, it might be trained to decipher more biolgical nuances at higher resolution than ChromHMM which is trained with only histone modification data.

D. (3pt) What are the inherent tradeoffs in the number of states? [Here they use 25 states, but the original ChromHMM paper used 51 and in other papers they use 12, 16, 19, and 21.]

In the paper, the number of chromatin states was a compromise between these inherent tradeoffs: Capturing the potential complexity of chromatin mark combinations (need large number of states) and generating models that are make genomic features overall interpretable (need small number of states). The authors chose 25, which they think is suitable for the number of chromatin marks and other input data tracks, allowing annotation of the likely functional roles of each state. Additional chromatin states might be discovered as the number of chromatin properties that can be elucidated increases.

7. (Extra Credit) Consider the following (insanely) simple grammar for RNA secondary structure prediction:

S → aSa'S | aS | ε

The first rule (S → aSa'S) captures both bifurcation (splitting for multiple stem structures) and base pairing! Draw a parse tree indicating how this structure:

Screen Shot 2018-04-29 at 20.06.39

would be generated from this simple (but admittedly a bit odd) grammar. Recall: the parse tree must begin with S, each non-terminal (in this case "S") is replaced at each step of tree with a single rule from the grammar, and all branches must end with ε. Hint: We will discuss secondary structure of RNAs starting on April 23rd.