HW3: Motif Finding, BLAST, ChIP-seq, Clustering, and GO Term Enrichment

Due: Friday, October 29, 2021 at 5pm

Submission Guidelines

Please read the instructions carefully! Name all files using the following convention:

- lastName_firstName_shortDescriptionOfFile.extension.

You must submit **one** pdf file (with all solutions to questions) and **one** file containing your code. If using R this can either mean including:

- one pdf file and one .R file **or**
- one pdf file and one .Rmd (R markdown) file.

1 Motif Finding

1.1 Theoretical Model

Suppose you have a transcription factor that binds the following set of sequences:
ATGT
ACGT
ACCT
ATCT

1. Build a position frequency matrix to model this transcription factor, using a pseudocount of 0.25.
2. Suppose that the genome this transcription factor is found in has a nucleotide composition of A 40%, C 10%, G 10% and T 40%. Give log-odds scores for searching for binding sites for this transcription factor within the sequence TATGT.

1.2 Uncovering Regulators

A set of genes in *E. coli* were found to be co-regulated across a variety of conditions. The regulatory regions of these genes are found in the file: **DNA-seqs.txt**. Use the MEME and Tomtom tools from the MEME Suite to uncover the putative regulator. Require that the motif be found at least once in each sequence, and that one motif be searched for. Please read carefully the tools description (to decide on the steps required to solve this question) and document the input and output at each step.

2 BLAST

2.1 Multiple testing in BLAST

1. The non-redundant BLAST database has $6 * 10^7$ nucleotide sequences. Suppose you BLAST your sequence and get a sequence hit with a p-value of 10^{-10}. You have performed $6 * 10^7$ sequence comparisons, what is the probability that at least one of the observations will be called significant by chance? You can give just a derivation of this calculation. Hint: the p-value is the probability of making a mistake in a sequence comparison, all sequence comparisons are independent, and you would like to derive the probability of making at least one mistake.

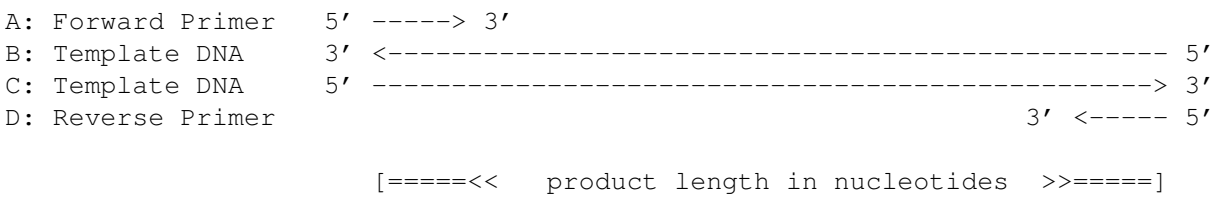
2. Instead of using p-values, BLAST reports E-values. What is it, and what is your estimate of the E-value based on part 1. (Note that BLAST actually uses a different calculation to estimate an E-value).

2.2 Primers for COVID screening by RT-PCR

RT-PCR is widely used to test people for COVID-19. The specificity of the test depends on the ability to discriminate between viral and human RNA.

2.2.1 Background on PCR

Sequence specificity of PCR To run a PCR reaction, researchers design short sequences of DNA called primers. The sequence of a pair of primer determines which DNA templates they can be used to detect. The needed primer-template relationship is shown in the cartoon below.

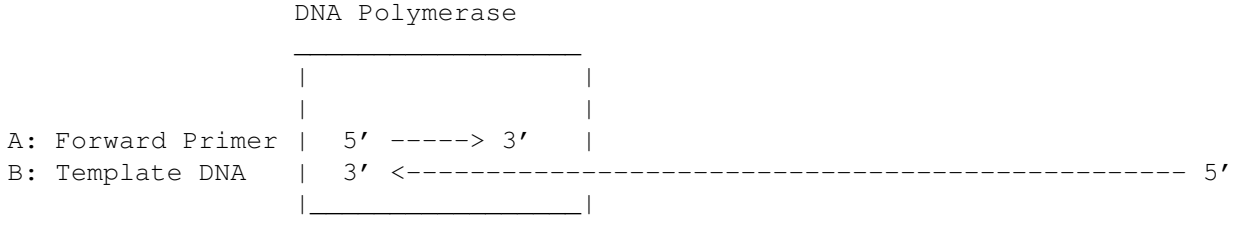


Here, the template DNA can be detected because it has complementary primers with valid position and orientation. Arrows represent the orientation of DNA molecules, with the arrowhead at the 3' end. Adjacent arrows with opposite directions represent reverse complement sequences. For example, sequence (A) is the reverse complement of a portion of sequence (B).

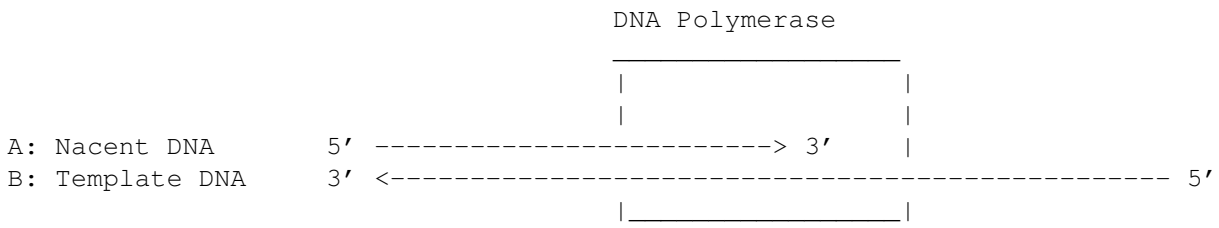
Molecular biology of PCR Note: this background is not essential to do this problem, but you may find it helpful to understand why the problem is interesting.

The position and orientation requirements for PCR are dictated by the underlying molecular biology. PCR works by synthesizing new DNA from existing templates. This is done by the enzyme DNA polymerase, represented by a rectangle in the cartoons below. There are two critical steps:

1. *annealing:* the primer, template, and polymerase come into contact with one another. Sequence complementarity between the primer and template stabilizes the needed interactions. Here (A) and (B) are shown, but the process is the same for primer (D) and template strand (C).



2. *elongation*: DNA polymerase adds nucleotides, one-by-one, to the 3' end of the nascent DNA strand. The primer and template are both required to begin extension.



If the nascent DNA strand (A) grows long enough, then it can serve as template for the other primer (D). In other words, the sequence of the nascent strand (A) is the same as (C). Each template can yield one additional molecule of template DNA. When multiple rounds of PCR are done, the number of template molecules can grow exponentially.

An incorrect orientation of the primers prevents this exponential amplification of the template. Similarly, if the nascent DNA molecules don't span the sequence between the two primers, amplification will fail.

A method for primer design If a sequence similar to the primer is found in sequences other than the one you wish to amplify, then the PCR process may lead to off-target amplifications. NCBI provides primer-BLAST to help researchers design primers that avoid common pitfalls. The tool proceeds in two steps. First, a researcher inputs a sequence they would like to amplify. This sequence is passed to `Primer3` to obtain a set of candidate primer pairs. These pairs are selected, in part, based on a thermodynamic model of annealing. Next, `BLAST` is used to screen for off-target sequences that could be amplified. In the example above, (A)-(A), (D)-(D), or (A)-(D) primer pairs could all potentially prime off-target amplification. Once primers are designed, they need to be verified empirically. Having good candidate primers makes this process easier.

2.2.2 Question on primer design

Use the web interface for *primer-BLAST* to find sequences of the SARS-CoV-2 genome that are amplified by the set of primers below. For your query, use the following inputs:

- PCR template (specified as NCBI reference sequence id for the SARS-Cov-2 genome): `NC_045512.2`
- forward primer: `GAC CCC AAA ATC AGC GAA AT`
- reverse primer: `TCT GGT TAC TGC CAG TTG AAT CTG`
- primer pair specificity checking parameters: Organism = `humans (taxid:9606)`

The remaining inputs can be set to their default values. Primer-BLAST will output a substantial amount of information. To answer this question, you will need to identify a few pieces of information. Answer the following questions in the context of the BLAST results:

1. Which SARS-CoV-2 gene is amplified by this primer pair? To answer this question, you can use the *graphical view of primer pairs* output. Genes are represented by green lines. Primer pairs are represented by blue arrowheads near the bottom of the output. The output also has a button that can help you zoom-in on the relevant region of the SARS-CoV-2 genome.
2. Are these primers specific to SARS-CoV-2, or do they amplify human mRNA sequences?
3. Consider shortening each of these primers to include only the first 18 nucleotides. Find the human transcripts that could be amplified by this PCR reaction. What are their lengths? The PCR reaction offers some control over the maximum product length. Is there a product length cutoff that allows you to exclude human transcripts?

references

- The primer pair was obtained from the following CDC article: *Research Use Only 2019-Novel Coronavirus (2019-nCoV) Real-time RT-PCR Primers and Probes.*

  `https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html`

- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G., 2012. *Primer3—new capabilities and interfaces.* Nucleic acids research, 40(15), pp.e115-e115.

3 P-values in a Simulated ChIP-seq Gene

Suppose there is a gene that has a mean number of ChIP-seq reads of $\lambda_1 = 10$ in one condition and $\lambda_2 = 50$ in another condition. We will simulate random data from such an experiment with $n = 4$ replicates for each condition. The goal is to compute a p-value that helps you distinguish between the null hypothesis that these counts came from conditions where the actual mean read value is the same ($\lambda_1 = \lambda_2$), versus the alternative hypothesis that the mean read values are different between the two conditions ($\lambda_1 \neq \lambda_2$). Write an R script that does the following:

1. Simulate random Poisson counts. Draw random counts assuming they follow the Poisson distribution with parameter λ as above (hint: use the R function `rpois`). Draw $n = 4$ replicates for each condition. Report a table with the counts you obtained on one random trial.

2. Compute and report the sample mean $\bar{c}_i$ and sample variance σ_i^2 in each condition i. Hint: use the R functions `mean` and `var`.

3. Calculate and report the following t-statistic: $t = \frac{\bar{c}_1 - \bar{c}_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\sqrt{n}$

4. Calculate and report the two-tailed p-value of differential protein-DNA interaction using t. Note the desired p-value equals $\Pr(|T| \geq |t|; H_0)$ where t is your t-statistic, and T is a random t-statistic under the null hypothesis H_0. Hint: use the R function `pt`, which is the cumulative of the t-distribution, which your t-statistics follow under the null hypothesis (of equal mean and equal variance between conditions, in this specific case). Use $2(n-1)$ for the degrees of freedom for the t-distribution. Warning: The standard cumulative gives a one-tailed p-value. What does the size of your p-value indicate?

5. Repeat steps 1-4 with $\lambda_1 = \lambda_2 = 10$. Report the t-statistic and p-value you got. What does the size of your p-value indicate?

6. What is the two-tailed p-value of a t-statistic of 3.5 (with the same degrees of freedom as earlier)?

4 GO term enrichment

To work on this problem you will need to download the following datasets and import them into your working environment. Pay attention to the list of files and their import code suggestions:

(a) The list of yeast genes tested in your experiment: assume that all bona-fide yeast genes (verified genes and confirmed yet uncharacterized ORFs) have been tested: **all_yeast_orfs.txt**

(b) The list of genes that are differentially expressed in your experiment:

differentially_expressed_orfs.txt (same as above)

(c) The classification of all yeast genes to GO biological process IDs relevant to yeast S. cerevisiae: **go_bp_matrix.txt**

(d) The mapping between GO IDs and GO terms: **go_bp_to_annotation.txt**

Import suggestion:

```
ORFs = read.table("all_yeast_orfs.txt", stringsAsFactors = FALSE, col.names = "ORF")$
    ORF
DE_ORFs = read.table("differentially_expressed_orfs.txt", stringsAsFactors = FALSE,
    col.names = "ORF")$ORF
GO_Annotation = read.table("go_bp_matrix.txt",stringsAsFactors = FALSE,header=TRUE,
    check.names = FALSE,row.names=1)
GO_Terms = read.table("go_bp_to_annotation.txt",stringsAsFactors = FALSE, sep = "\t",
    check.names=FALSE,quote="",col.names=c("GOterm","Description"))
```

4.1 Data Validation

To verify that your data import was successful, perform the following cross-checks:

(a) Are all genes in your differentially expressed list included in the list of tested genes? If not, which genes are missing from the tested genes? Why do you think they are missing? If possible, fix the differentially expressed list. Otherwise, eliminate these genes from the differentially expressed list.

Hint: remember that you wouldn't find a gene to be differentially expressed unless you tested it. Therefore, the list of differentially expresesd genes should be a subset of the list of tested genes. Also, keep in mind that data input is often imperfect and that this example is designed to simulate raw data.

(b) Calculate the size of each GO term, i.e. the number of genes annotated to it. Plot the distribution of GO term sizes. What is the largest GO term? Report the GO term id, name and size.

(c) Calculate the number of GO terms for gene YDR026C. Is GO:0006725 "cellular aromatic compound metabolic process" among the terms?

4.2 Enrichment analysis

Use the hypergeometric test to calculate the enrichment of each GO term among the list of differentially expressed genes from your experiment. Print out the ranked list of all significant GO terms (sorted by corrected p-value, all p-values < 0.05) in a table with the following columns:

- GO ID
- GO term name
- Number of genes in the background gene list annotated to this GO term
- Number of genes in the differentially expressed gene list annotated to this GO term
- Fold enrichment
- Enrichment p-value
- Enrichment p-value corrected for multiple testing using the Bonferroni method

hints: to use the hypergeometric test, you can define the following:

```
N = number of genes in background
K = number of genes in background with GO term
n = number of differentially-expressed genes
k = number of differentially-expressed genes with GO term
```

Then, you can use the function "phyper" to calculate the p-value:

```
p-value = phyper(k-1, K, N-K, n, lower.tail=FALSE)
```

Also, please use the function "p.adjust" to correct your p-values.

4.3 Theoretical question

Imagine that at some point, you realized that the method used to quantify gene expression was restricted to only non-essential genes in the yeast genome. Explain how this could affect your previous analysis and the obtained p-values.