02-510/710: Computational Genomics, Spring 2023

HW3: Epigenetics

Version: 1
Due: 23:59 EST, Mar 27, 2023 on Gradescope

Topics in this assignment:

1. Hidden Markov Models
2. Motif-Finding
3. ChIP-Seq Data Analysis

What to hand in.

- One write-up (in pdf format) addressing each of following questions.
- All source code. If the skeleton is provided, you just need to complete the script and send it back. Your code is tested by autograder, please be careful with your main script name and output format.

Submit the following file which contain the completed code and the pdf file to gradescope separately.

./S2023HW3.pdf
./HMM.py

Please note that all the solutions and code must be your own. We will check for plagiarism after the final submission.

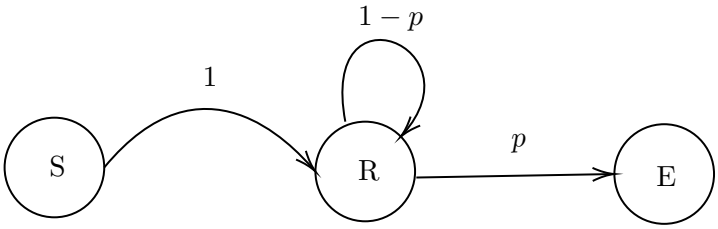


Figure 1: Random Genome HMM

1. **[38 points] Hidden Markov Models**

In this question, we will explore Hidden Markov Models (HMM) and their use for genome annotation.

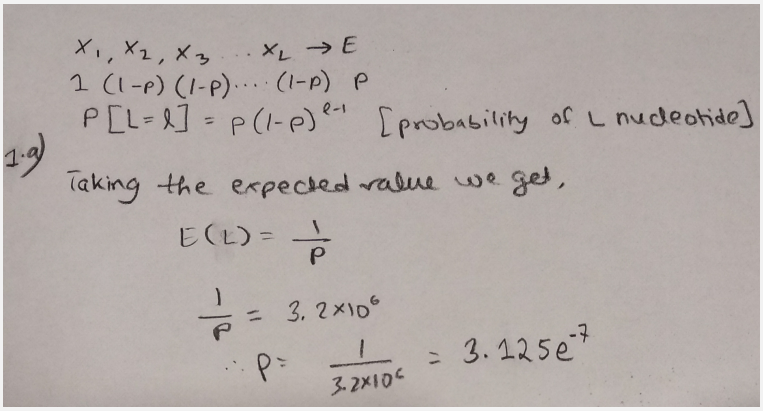
Warm-Up

(a) (5 points) Consider the state transition diagram for the very simple HMM shown in Figure 1. The state S is a silent start state, and the state E is a silent end state. The state R, which stands for *random*, emits nucleotides with the following probabilities:

nucleotide	emission probability
A	0.20
C	0.30
G	0.30
T	0.20

The human genome is approximately 3.2 billion base pairs long. Suppose we generate a genome using the HMM in Figure 1. Find the value of p such that the expected length of this random genome is equal to the size of the human genome.

Solution



(b) (1 point) Find the expected GC content of the genome generated in the previous part (write the answer as a percentage).

Solution

0.3 +0.3 = 0.6, which is 60 percent

Supervised Learning with HMMs

(c) (7 points) Consider the problem of supervised learning using HMMs. Specifically, we are given a set of n observation sequences $O^{(i)} = o_1^{(i)}, \ldots, o_{T_i}^{(i)}$ drawn from an alphabet set O, along with state annotation data $Q^{(i)} = q_1^{(i)}, \ldots, q_{T_i}^{(i)}, (i = 1, \ldots, n)$, from a set of possible states Q. Here T_i is the length of the i-th observation. Show that the maximum likelihood estimates for the HMM are

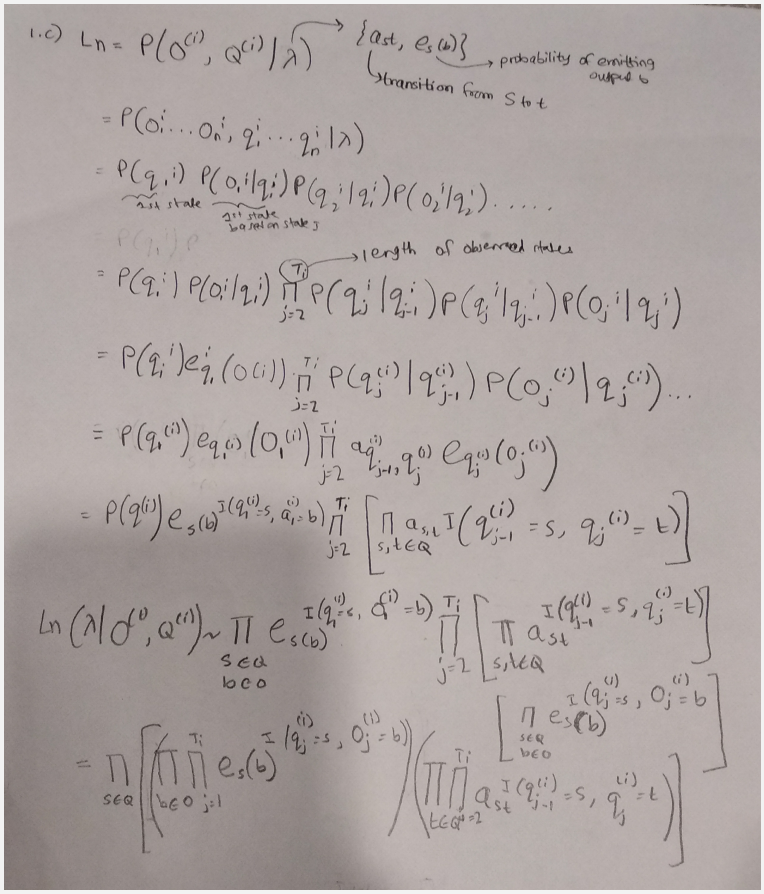
$$\hat{a}_{st} = \frac{\sum_{i=1}^{n} \sum_{j=2}^{T_i} \mathbb{I}\left\{q_{j-1}^{(i)} = s, q_j^{(i)} = t\right\}}{\sum_{i=1}^{n} \sum_{j=2}^{T_i} \sum_{t' \in Q} \mathbb{I}\left\{q_{j-1}^{(i)} = s, q_j^{(i)} = t'\right\}},$$

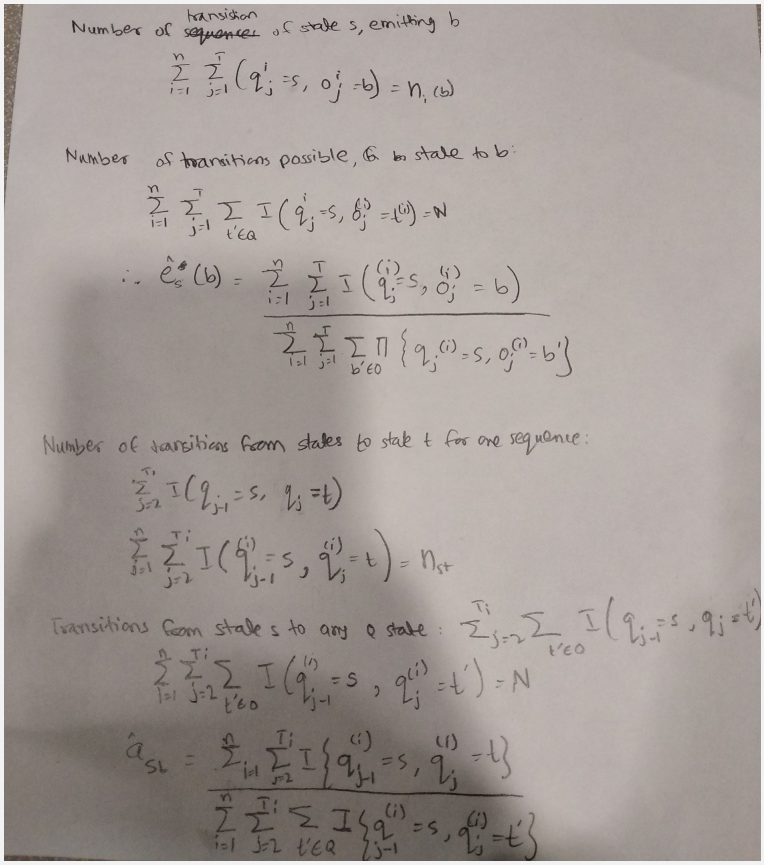
and

$$\hat{e}_s(b) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{T_i} \mathbb{I}\left\{q_j^{(i)} = s, o_j^{(i)} = b\right\}}{\sum_{i=1}^{n} \sum_{j=1}^{T_i} \sum_{b' \in O} \mathbb{I}\left\{q_j^{(i)} = s, o_j^{(i)} = b'\right\}}.$$

Here, $\mathbb{I}$ is the indicator function, which takes value 1 when its argument is true and 0 otherwise. (The notation is explained in slide 32 of the Feb 27 lecture).

Solution





HMMs for Genome Annotation

In lecture, we saw that we can use HMMs for annotating epigenetic states in the genome, where our observed output is a combination of genetic and epigenetic markers in a given set of genome annotation tracks. In this problem, we will consider a simplified scenario in which we observe only one feature of the genome.

Consider the problem of crossing a vampire with a werewolf, which produces a monstrosity that we shall name the "werepyre" (1 point extra-credit: suggest an alternative name for this hybrid beast). To study this creature, we find werepyre tissue samples from multiple werepyres, and assemble their genomes. We shall denote the i-th genome as $S^{(i)} = (s_1^{(i)}, s_2^{(i)}, \ldots, s_n^{(i)})$, $s_j^{(i)} \in \{A, C, G, T\}$. Our goal is to determine which bases of each genome come from the vampire genome, and which come from the werewolf genome.

(d) (5 points) Suppose for the moment that we also have access to the sequences of vampire and werewolf genomes. Outline a strategy we might use to determine the origins of each base in the werepyre genome. (*Hint*: global alignment of the werepyre genomes to each of the ancestral genomes is unlikely to work).

Solution

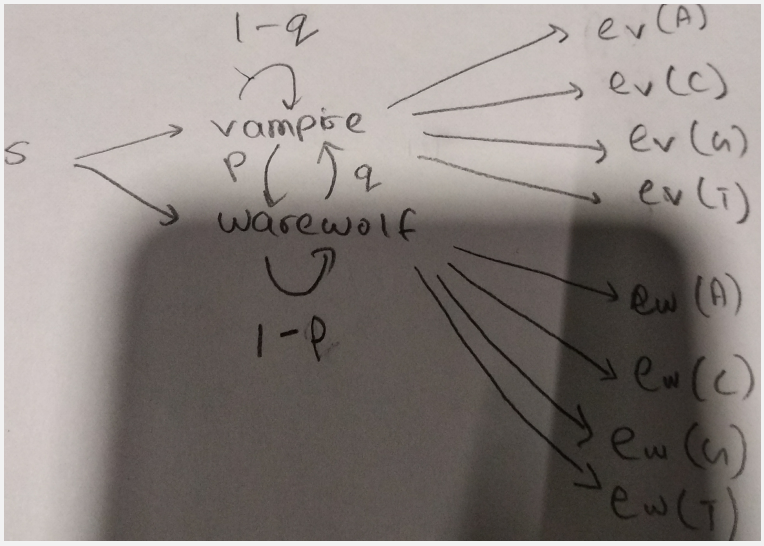
We can divide the werepyre genome into smaller segments of length k, where k is a small enough number to capture the local variations between the ancestral genomes and the werepyre genome. Then we perform a local alignment of each segment against the vampire and werewolf genomes separately using a suitable alignment algorithm such as Smith-Waterman. And for each segment, we can compute the alignment score for both

the vampire and werewolf genomes. If the alignment score for one of the ancestral genomes is significantly higher than the other, then we can infer that the segment likely originated from the ancestral genome with the higher alignment score. We would do this for all the segments of the werepyre genome and merge the segments together, assigning each segment to either the vampire or werewolf genome based on the results. If there are any gaps or unassigned segments in the werepyre genome, we can perform the allignment algorithm with larger segment length or a different alignment algorithm until all segments are assigned to either the vampire or werewolf genome. Here we in the procedure we are assuming that we have access to the sequences of the ancestral genomes. (Otherwise, we could first perform a phylogenetic analysis or implement a clustering method to infer the ancestral genomes and then perform the alignment.
The fusion between vampire and warewolf is known as Pricolici according to Romaninan folklore. It can also be termed as vaewolves.

(e) (5 points) Understandably, we have not yet found a vampire or werewolf whose genome we can sequence. Thus, we cannot use the approach we just outlined above. Fortunately, we happen to know that the GC content of these two genomes are drastically different from each other. Based on this information, design a Hidden Markov Model we can use to annotate each base of the werepyre genome as originating from either the vampire or werewolf genome. Show its state transition diagram below, and specify all of the necessary parameters.

Solution

There are two hidden states: S1 and S2. Here, S1 represents the state where the current base in the werepyre genome is more likely to have originated from the vampire genome, and S2 represents the state where the current base is more likely to have originated from the werewolf genome. The observations would be nucleotides A, C, G, and T at each position in the werepyre genome. And the transition probabilities between the two states can be defined as: $P(S1 \rightarrow S1) = 0.9$.This means if the current base is more likely to have originated from the vampire genome, the next base is also more likely to have originated from the vampire genome. $P(S1 \rightarrow S2) = 0.1$.Where there is a small chance that the current base is actually from the werewolf genome. $P(S2 \rightarrow S2) = 0.8$.Implying if the current base is more likely to have originated from the werewolf genome, the next base is also more likely to have originated from the werewolf genome. $P(S2 \rightarrow S1) = 0.2$Represents there is a small chance that the current base is actually from the vampire genome. The emission probabilities are the conditional probabilities of observing a nucleotide given the hidden state. The GC content of the vampire and werewolf genomes are significantly different, so we can use this information to define the emission probabilities as follows: $P(A|S1) = P(T|S1) = (1 - q)/2$, where q is the GC content of the vampire genome. This means if the current base is more likely to have originated from the vampire genome, the probability of observing A or T is $(1 - q)/2$. $P(C|S1) = P(G|S1) = q/2$. If the current base is more likely to have originated from the vampire genome, the probability of observing C or G is $q/2$. $P(A|S2) = P(T|S2) = p/2$, where p is the GC content of the werewolf genome and if current base is likely to have originated from the werewolf genome, then the probability of observing A or T is $p/2$. $P(C|S2) = P(G|S2) = (1 - p)/2$. So, if the current base originated from the werewolf genome, the probability of observing C or G is $(1 - p)/2$.



(f) (1 point) Notice that we do not have annotated werepyre genomes from which we can estimate our parameters. Which of the following algorithms should we use to learn our parameters?

- Baum-Welch Algorithm
- Maximum Likelihood Estimation
- Posterior Decoding
- Expectation Minimization

Solution

Baum-Welch Algorithm

(g) (10 points) The Baum-Welch algorithm is a special case of the Expectation Maximization algorithm that we can use to perform unsupervised learning with HMMs. Instead, we will implement a similar algorithm known as Viterbi Training, which has the advantage of being more intuitive, as well as being computationally faster. Unlike the Baum-Welch algorithm, Viterbi Training makes hard assignments to the latent variables (hidden states) in the E-step. The algorithm runs as follows, starting with an initial guess λ_0 of the parameters:

i. Run Viterbi decoding on each observed sequences using the current parameters λ
ii. Using the annotations generated in step (i), re-estimate λ using maximum likelihood estimation
iii. Repeat steps (i)-(ii) until convergence

Implement the Viterbi Training algorithm and use it on the provided training genome sequences (`werepyre-train.fasta`), using the initial parameters to the provided in the file (`viterbi-init.txt`). To accomplish this, you'll need to implement the Viterbi decoding and Maximum Likelihood Estimation algorithms for HMMs.

Using the estimated parameters, perform Viterbi Decoding on the provided test genome (`werepyre-test.fasta`). Provide an annotation track for the test genome, labeling each base as originating from either a vampire or werewolf genome. The true labels are given in the file `werepyre-test-annotation.fasta`. What percentage of your inferred labels match the true labels?

Solution

(h) (4 points) The true parameter values are provided in `viterbi-true.txt`, If we run the Viterbi Decoding with the true parameter values, what percentage of the inferred labels match the true labels? Are you surprised by the result?

Solution

2. **[51 points] Motif Finding**

In lecture we were introduced to an algorithm used for finding motifs in DNA sequences based on Expectation Maximization (EM). This algorithm forms the basis of the MEME Suite (**M**ultiple **E**M for **M**otif **E**lucidation), one of the most widely used softwares in genomics. Several good papers are available for understanding the algorithm, including the ones here and here.

In this problem, we will examine in some detail the underlying statistical model and assumptions that inform our approach to motif finding. We will also consider an approach to determining the statistical significance of a found motif. We'll start off by considering position-weight matrices in order to understand the statistical model underlying motif finding algorithms.

The Position-Weight Matrix

Consider a biological motif of length W, $M = (M_1, \ldots, M_W)$, where $M_i \in \{A, C, G, T\}$. Our model for biological motifs is that each M_i is a Multinoulli-distributed random with its own probability distribution over the nucleotides A, C, G, and T. We can equivalently represent this motif as a position weight matrix $(M)_{ij}$, for which

$$M_{ij} = \mathbb{P}(M_j = i),$$

where $j = 1, \ldots, W$ and $i \in \{A, C, G, T\}$. Consider the PWM below for a motif of length $W = 6$:

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccccc} M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \end{array} \\ \begin{bmatrix} 0.25 & 0.5 & 0.95 & 0.5 & 0 & 0.03 \\ 0 & 0 & 0.03 & 0 & 0.05 & 0.07 \\ 0 & 0 & 0 & 0.95 & 0.95 & 0.90 \\ 0.75 & 0.5 & 0.02 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{1}$$

One of the key assumptions we make when modeling motifs is that $M_i \perp M_j$ for $i \neq j$; that is, the distributions of the nucleotides in each position of the motif are independent of one another.

(a) (3 points) Find the probability $\mathbb{P}(M = TTAACC)$.

Solution

= 0.75*0.05*0.95*0.5*0.05*0.07 = 0.00006234375

(b) (1 point) Provide the Consensus Sequence corresponding to the PWM above; i.e., the sequence $m = (m_1, \ldots, m_5)$ such that $\mathbb{P}(M = m)$ is maximized.

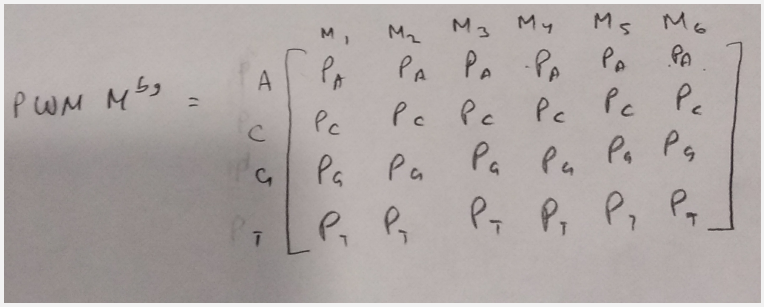
Solution

To obtain the consensus sequence corresponding to a given PWM, we choose the most probable nucleotide at each position of the PWM. P(M = TAAGGG) or P(M = TTAGGG)

(c) (3 points) As we saw in lecture, one problem with relying on PWMs alone is that we do not take into account the background frequency of each nucleotide in the genome. We can correct for this by introducing a new PWM M^{bg}, which gives the probability of generating a sequence $X = (X_1, \ldots, X_W)$ if we randomly put together a sequence of length W by sampling each

nucleotide independently from the background distribution. Suppose our background nucleotide frequencies are given by p_A, p_G, p_C, p_T. Provide the background PWM M^{bg}.

Solution



(d) (4 points) Now that we have M^{bg}, we can define a likelihood ratio, LR, for an observed sequence $X = (x_1, \ldots, x_W)$. The LR is calculated as

$$LR(X) = \frac{\mathbb{P}(X|M)}{\mathbb{P}(X|M^{bg})}.$$

We prefer to work with the log-likelihood ratio, LLR, which is simply

$$LLR(X) = \log_2 \frac{\mathbb{P}(X|M)}{\mathbb{P}(X|M^{bg})}.$$

Now define the scoring matrix

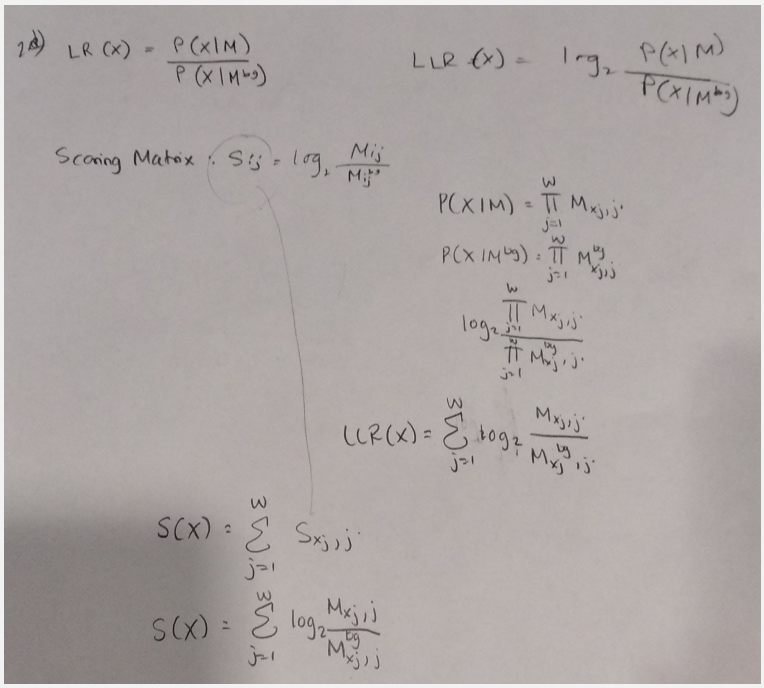
$$S_{ij} = \log_2 \frac{M_{ij}}{M^{bg}_{ij}}.$$

Given S_{ij}, we define the score of the sequence X, $\mathcal{S}(X)$, as

$$\mathcal{S}(X) = \sum_{j=1}^{W} S_{x_j,j}.$$

Show that $\mathcal{S}(X) = LLR(X)$.

Solution



(e) (3 points) Our motivation for defining the LLR is that we want to be able to conclude whether or not X is a true instance of our motif M or not. Suppose that for our observed sequence X we calculate $\mathcal{S}(X) = 2$. How much more likely is it that X is a true instance of M than a background sequence? (In other words, calculate $LR(X)$.

Solution

Putting LLR(X) = 2 equates to 4

(f) (3 points) Now suppose $\mathcal{S}(X) = 0$. What do you conclude?

Solution

Putting LLR(X) = 2 equates to 1

(g) (5 points) The above implies the obvious decision rule:

$$X \text{ is a true instance of } M \text{ if } LLR(X) > 0,$$

or equivalently, if $\mathcal{S}(X) > 0$. Is this a statistically sound approach? Why or why not?

Solution

This is not a statistically sound approach, as no distribution is applied to calculate the probability and the condition that if LRR(X)>0 is very stringent to imply that X is a true instance as there are chances of false positive.

(h) (3 points) To refine our approach above, we cast the problem of deciding whether or not X is a true motif in a statistical hypothesis testing framework. We will test the following hypotheses:

$$H_0\text{: } X \text{ is drawn from the distribution } M^{bg}$$
$$H_1\text{: } X \text{ is drawn from the distribution } M$$

using $\mathcal{S}(X)$ as our test statistic, with significance level α. Our decision rule is

$$\text{reject } H_0 \text{ if } \mathcal{S}(X) > t,$$

where t is the number such that $\mathbb{P}_{H_0}[\mathcal{S}(X) > t] = \alpha$ (this reads as the probability, under the null hypothesis, that our test statistic takes on a value greater than t). Why is this a reasonable decision rule?

Solution

It helps take hold of the randomness and/or variation or circumstance when X is not a true instance. Also, we can take account for the threshold.

(i) (8 points) Since we do not know the sampling distribution of $\mathcal{S}(X)$ under H_0, we need to determine t empirically. Outline an approach to determine t (*Hint*: this is similar in spirit to the permutation testing that we covered in HW 2).

Solution

We randomly sequence from M bg (consisting all the possible combination) and calculate the score and determine the probability, which would satisfy $P_{H_0}[S(x) > t] \leq \alpha$. We generate a graph representing the null hypothesis and determine the region where it satisfies the hypothesis.

Note that in the statistical model above, we made an assumption that any sequence that is not an instance of M is a background sequence (i.e., random). This may not be true in reality, since, for example, X could be an instance of a different motif M'. See this paper if you are curious about how to deal with the problem of finding multiple motifs in the same set of sequences.

Estimating PWMs from Labeled Data

Before considering the more difficult task of unsupervised learning, we'll consider a simpler problem, where we consider the task of learning a position-weight matrix (PWM) from labeled data.

Suppose we are given N DNA sequences, each of which contains a single, un-gapped promoter sequence of length W, which are marked by annotation. More specifically, our sequence data is denoted as

$$S = (S^{(1)}, \ldots, S^{(N)}) \qquad S^{(n)} = (s_1^{(n)}, \ldots, s_{T_i}^{(n)}) \in \{A, C, G, T\}^{T_n},$$

with corresponding offsets

$$O = (o^{(1)}, \ldots, o^{(N)})$$

As before, T_n is the length of the n-th sequence. The offsets $o^{(n)}$ indicate the position of the first base of the promoter sequence in S_n. Given this data, our task is to estimate the position-weight matrices M corresponding to the promoter sequence, and M^{bg} corresponding to the background nucleotide frequency.

(j) (3 points) First, describe how we would estimate M^{bg} from the data.

Solution

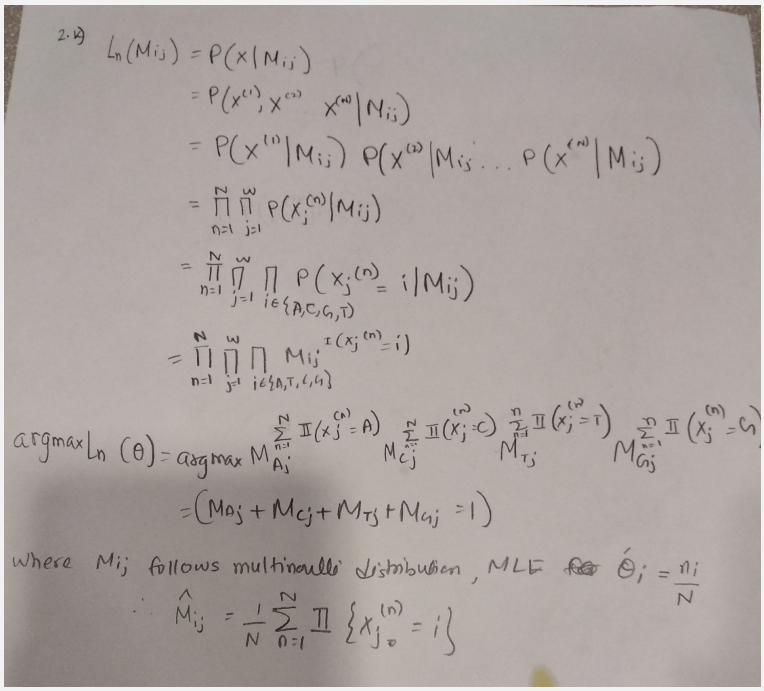
For each sequence, we remove the offset and count the number of A,C,G, T in the remaining sequence and the promoter sequence separately, and get the probability of A in the length of sequence of interest, for both the promoter sequence and the remaining sequence. This would provide the respective M-bg from the data.

(k) (5 points) For simplicity, let us define $X^{(n)} = (s^{(n)}_{o_n}, \ldots, s^{(n)}_{o_n+W-1})$; that is, $X^{(n)}$ is the promoter sequence contained in the n-th larger sequence $S^{(n)}$. We'll denote the j-th base of $X^{(n)}$ as $X^{(n)}_j$. Show that the Maximum Likelihood Estimate for M_{ij}, $i \in \{A, C, G, T\}$, $j = 1, \ldots, W$, is given by

$$\hat{M}_{ij} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left\{X^{(n)}_j = i\right\}. \tag{2}$$

As before, $\mathbb{I}$ is the indicator function.

Solution



One problem with Maximum Likelihood Estimation is that in small datasets, we may set $\hat{M}_{ij} = 0$ if we do not observe any motifs with nucleotide i in position j. To see why this can be problematic, we may ask ourselves an analogous statistical question: if we tossed a coin 3 times and it came up heads each time, would we conclude that the probability of the coin flipping tails is 0, or would we want more data before coming to such a severe conclusion? In practice, we often deal with this issue by using pseudocounts; that is, we assume we have already observed data with α_{ij} motifs containing nucleotide i in position j. The estimate for M_{ij} then becomes

$$\hat{M}_{ij} = \frac{\sum_{n=1}^{N} \mathbb{I}\left\{X^{(n)}_j = i\right\} + \alpha_{ij}}{N + \sum_k \alpha_{kj}}. \tag{3}$$

Though creating fake data in this fashion may seem problematic, we'll show that using pseudocounts actually has a natural probabilistic interpretation coming from Bayesian statistics.

(l) (5 points) Consider a random vector $\mu = (\mu_1, \ldots, \mu_d)$ satisfying the following conditions:

i. $\mu_j \geq 0$ for all $j = 1, \ldots, d$

ii. $\sum_{j=1}^{d} \mu_j = 1$.

That is, μ is drawn from the d-dimensional simplex. The Dirichlet distribution over μ, with parameters $\alpha_1, \ldots, \alpha_d > 0$, has density

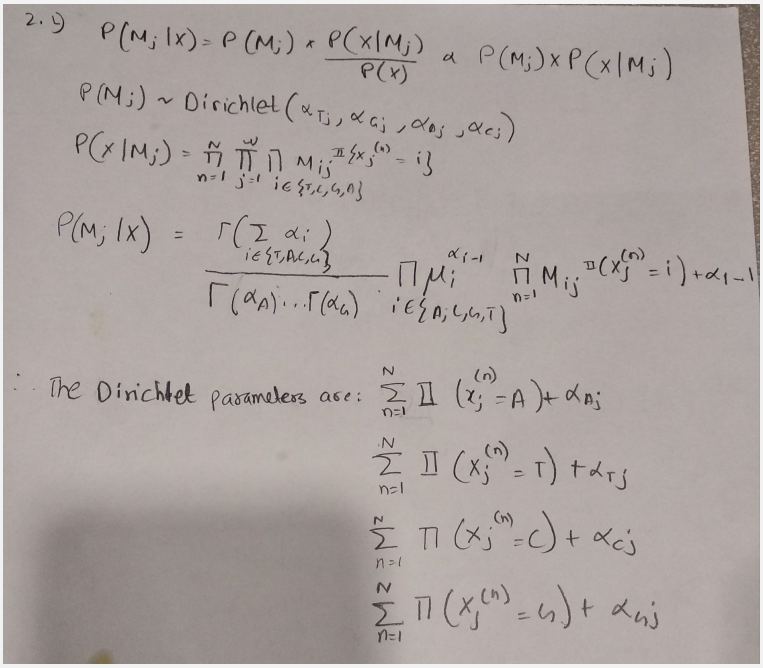
$$g(\mu) = \frac{\Gamma\left(\sum_{j=1}^{d} \alpha_j\right)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_d)} \prod_{j=1}^{d} \mu_j^{\alpha_j - 1}.$$

Consider the distribution $M_j = (M_{Aj}, M_{Cj}, M_{Gj}, M_{Tj})$ over the nucleotides in the j-th position in the motif. Suppose we impose a Dirichlet prior over M_j:

$$M_j \sim \text{Dirichlet}(\alpha_{Aj}, \alpha_{Cj}, \alpha_{Gj}, \alpha_{Tj}).$$

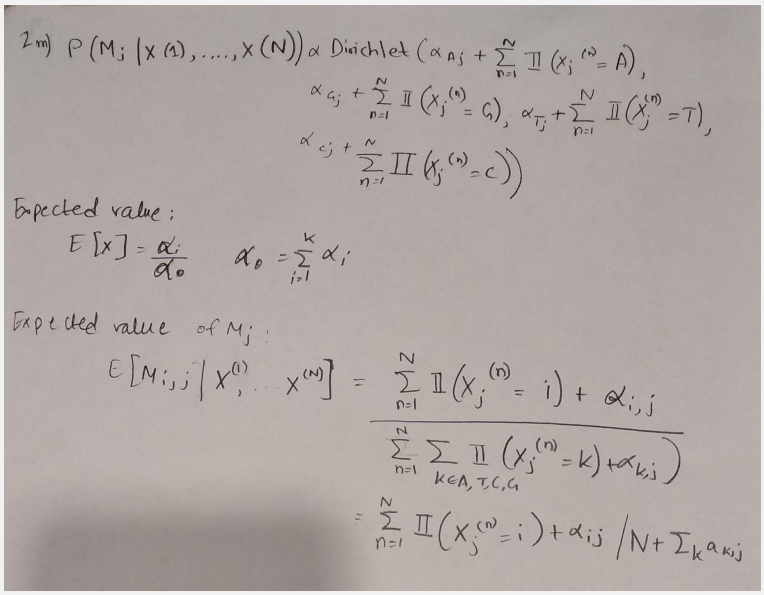
We then observe our motifs $X^{(1)}, \ldots, X^{(N)}$. Show that the posterior distribution of M_j also follows a Dirichlet distribution. What are its parameters?

Solution



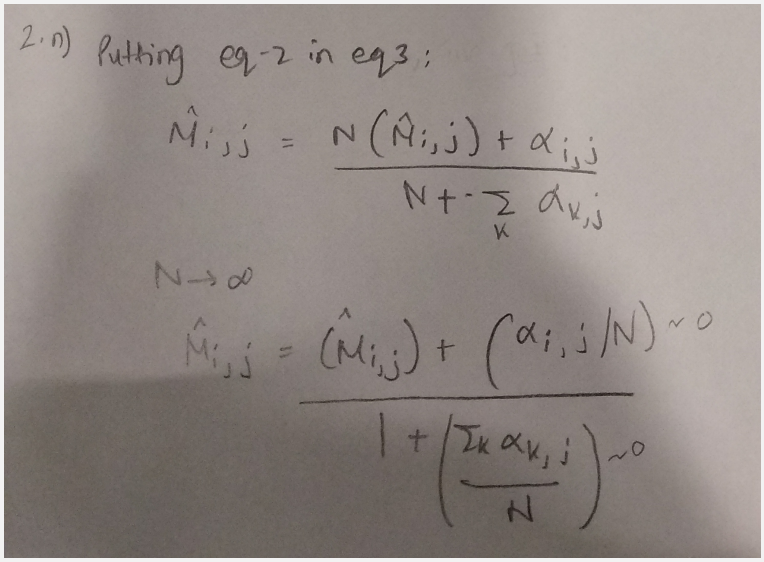
(m) (3 points) Show that the expectation over the posterior distribution of M_j is given by Equation 3.

Solution



(n) (2 points) Compare the estimators given in Equation 2 and 3. What happens as $N \to \infty$?

Solution



The two point estimates would converge.

3. **[24 points] ChIP-seq analysis.**

We provide you some ChIP-seq peak data in `peak.bed` that contains the chromosome, start and end positions of the ChiP-seq peaks from an experiment. This is real transcription factor ChIP-Seq data from a cancer cell line, and your goal is to figure out if you can distinguish what transcription factor the experiment was performed with. In doing so, you will gain familiarity with some common genomics tools. Most of the parts in this question are pretty open-ended; as long as your answers are well-supported, you will get credit.

Data Processing Steps

Using `peak.bed`, please complete the following tasks.

- Using your tool of choice, extract the first 100 lines of this file to a new file `"trimmed.bed"`.
- Load `"trimmed.bed"` in the UCSC Genome Browser (https://genome.ucsc.edu) as a custom track (use hg19 as the assembly). Get the genomic sequence for each of the intervals through UCSC Genome Browser. You can do that by going to Table Browser in UCSC Genome Browser, select custom track and load `"trimmed.bed"`, select output format as **sequence** and get output. Figure 2 shows how to do this.

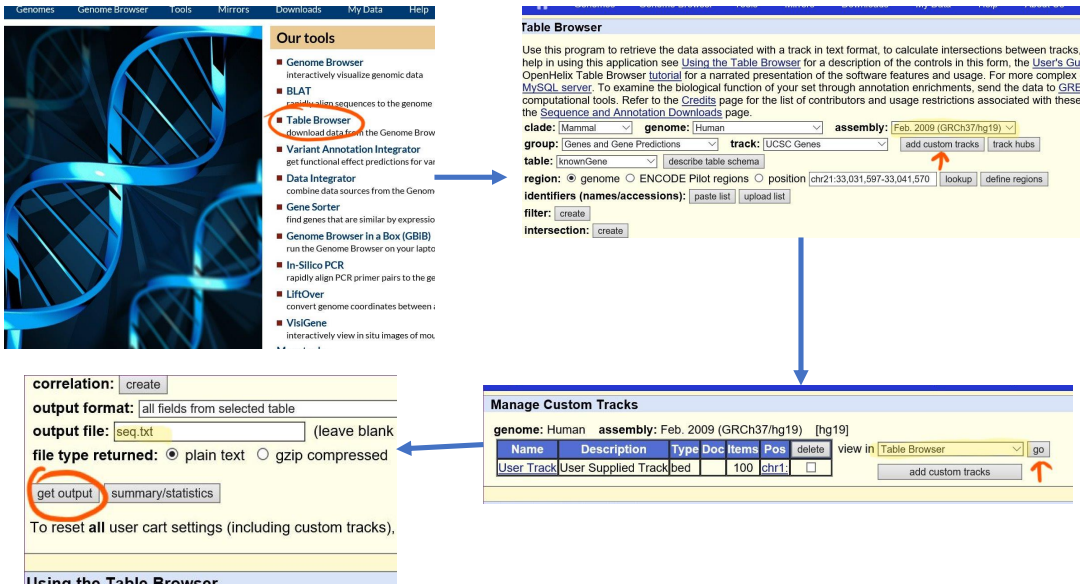


Figure 2: Use UCSC Table Browser to get the sequence.

- Use MEME (http://meme-suite.org/tools/meme) to call enriched motifs based on the output sequence. Use the default setting in MEME but only consider output motif with width from 5 to 9 (set under "Advanced options"). Note that you may still need to adjust the size of the input file to keep the number of characters within 60,000 in the file. You may want to provide your email if you want a notification for when your results are ready from MEME. This may take some time, depending on the server load.
- Using each of your hits from MEME, feed them to TOMTOM by clicking the 'submit/download' arrow to the right of your MEME results. Use all default options for TOMTOM.

Questions

(a) (2 points) Question – Screenshot the output returned by MEME. Based on the motif logos, which one is the most promising hit?

Solution

(b) (6 points) Question – What is the top hit obtained from TomTom for each of the motifs from MEME? Based on this do you agree with your previous assessment? Note that the E-value reported by TOMTOM is the Bonferroni-corrected p-value.

Solution

(c) (8 points) Question – Use the original peak.bed file and use GREAT (http://great.stanford.edu/public/html/) with default settings to identify enriched biological processes and pathways. **Be sure to use GREAT version 3.0.0, with species hg19.** Screenshot the top five Go Biological Process hits. How do these pathways compare to your expectations based on your MEME results? Would you consider these pathways statistically significant? Hint: explain what the important enrichment statistics mean.

Solution

(d) (2 points) Question – Now take a quick look at the hits in the category MSigDB Preturbation category. Explain what type of data is shown here. How much might these hits depend on what type of cancer cells the experiment was done with?

Solution

(e) (6 points) Question – Suppose that you were interested in determining what type of cancer the provided data was collected from. Based on your GREAT analyses, how useful is the provided TF ChIP-Seq data for this purpose? Justify this with some explanation: this will require that you understand what the GO terms are, and what processes your TF(s) are involved with. Then outline a follow up high throughput sequencing experiment from those mentioned in class that might help you with this new objective.

Solution