Homework 5

Instructions. As usual, you may work together on homework, but you must write up the solutions on your own. If you turn in an electronic copy, please provide it in pdf format. This homework is due by the beginning of class on Tuesday, March 28th.

1. The TAL effectors are DNA binding proteins found in *Xanthomonas* plant pathogens that interact with host genes in order to induce disease (or defense mechanisms) in the host. Each TAL consists of an N terminal domain, a variable number of 34 amino acid repeats, and a C terminal domain. Residues 12 and 13, we will call them *diresidues*, of each 34 amino acid repeat are highly variable. Your academic predecessor, Moscou and Bogdanove [1], hypothesized that these diresidues may directly interact with a nucleotide in the binding site. If so, a TAL effector with 17 repeats should bind a site of 17 contiguous nucleotides. Moscou's goal was to confirm this hypothesis and identify the diresidue/nucleotide mapping preferences.

   Suppose you have done much biological legwork and identified n TAL effectors, *and* you know which host gene is targeted by each of these TALs. Suppose the ith TAL effector has L_i repeats, with diresidue sequence $w_i = (w_{i1}, \ldots, w_{iL_i})$, where w_{ij} represents one diresidue. Let $x_i = (x_{i1}, \ldots, x_{iN_i})$ be the $N_i \approx 1000$ nucleotides immediately upstream of the translation start site of the gene known to be targeted by the ith TAL effector. We will assume that each targeted gene has one binding site for the corresponding TAL effector fully contained within this sequence.

   (a) Assume an iid model for the background sequence. Assume diresidues bind independently, but allow each diresidue to have distinct preferences for the four nucleotides. Perform a detailed derivation of an EM algorithm to identify the diresidue nucleotide preferences. In other words, obtain formulae for the observed data likelihood, the complete data likelihood, the expected complete log likelihood, and the analytic formula for the M step.

   (b) [*optional*] Write a program to implement the EM algorithm for the data that was available to Moscou in 2009. The data on $n = 10$ TAL effectors are provided in two files. File `tal_idata.txt` contains data on the diresidues in each TAL effector. The first line is the number of TAL effectors (10) and the number of unique diresidues (8). The next line reports the number of repeats in each of the 10 TALs. The next 10 lines report the sequence of diresidues in each TAL. The last line reports the identity of the 8 diresidues in terms of amino acids. The '*' stands for deletion of the corresponding residue. File `promoter_data.txt` contains the promoter data. The first line is the number of promoters (10 matching, in order, the 10 TAL effectors of the first dataset). The second line is the length of each promoter. The next 10 lines give the nucleotides of the promoter sequences, encoded as $A = 0, C = 1, G = 2$, and $T = 3$.

   This question is optional because there is *just barely* enough data to get this EM algorithm to work, and there are implementation details that are excellent lessons

in the reality of research, but potential time sinks. Rather than bury you in implementation details, we are going to implement an EM to solve a more topical problem.

2. One of your current colleagues (about to graduate) is working on the problem of detecting differentially expressed isoforms. An isoform is a processed mRNA product of a gene. Many genes have multiple isoforms that use different transcription start sites, different exon/introns, or untranslated regions. It is of interest to detect conditions under which isoform expression or isoform relative abundances change.

   Suppose you have mapped Illumina paired-end sequencing data to a reference genome, where all possible isoforms are annotated. Assume there is no ambiguity in mapping and ignore overlapping genes so you can identify all *read-pairs* that map to a particular gene. Because isoforms share exons, you may not know the isoform source of the *read-pair*.

   An isoform can be thought of as a sequence of alternating exons and introns. A set of isoforms at a locus can be reduced to a splice graph consisting of exons (or subexons) as vertices and edges connecting exons across introns. A subexon is a maximal portion of an exon that appears intact in *all* isoforms. An exon is split into two subexons, for example, if it contains an alternative splice site that is used in some isoforms. Edges have length zero when traversing subexons within an exon. A read-pair r_i can be represented as a unique set of ordered nodes, called an subexon path. Notice, if the underlying sequence of a read-pair contains an unsequenced portion, such as the insert, the subexon path of the read-pair r_i is an incomplete observation of the unobserved set of subexons from which r_i is generated. However, when given the isoform source, a subexon path become a complete observation.

   Focus on the read-pairs $\{r_i\}$ from a single gene g and suppose there are L distinct subexon paths assigned to read-pairs aligned to this gene. The observed data $\{\boldsymbol{y}_i\}$ identify the subexon path of the read-pair r_i, *i.e.*, $\boldsymbol{y}_i$ is an L-dimensional vector, one of the standard basis vectors of L-dimensional Euclidean space. Specifically, if the observed subexon path is l, then $y_{il} = 1$ and $y_{il'} = 0$ for all $l' \neq l$. Each read-pair is considered as generating from one, possibly unobserved, isoform. Given the isoform source, observation $\boldsymbol{y}_i$ is the realization of Multinoulli($\boldsymbol{\theta}_{k\cdot}$), where θ_{kl} is the probability of generating a fragment from subexon path l given that it arises from isoform k. Under the assumption that fragments are sampled uniformly for sequencing, θ_{kl} is a constant computable from the annotation.

   Our goal is to estimate $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, where K is the total number of isoforms and π_k is the probability of sampling the kth isoform, which depends on the relative abundance η_k of isoform k and its length. Since we know the isoform lengths and have functional invariance of MLEs, MLE $\hat{\boldsymbol{\pi}}$ can be transformed into MLE $\hat{\boldsymbol{\eta}}$, which is valuable biological knowledge.

   The data file `isoform.csv` on Blackboard contains information parsed from an RNA-Seq simulator. Each row is the summary information for one subexon path. The columns are: (1) gene, (2) number of read-pairs mapped to the gene, (3) isoform ids, (4) θ_{kl}, (5) the true $\boldsymbol{\pi}$, (6) the subexon path, (7) and the observed count, where entries (3)–(5) each

represent K columns. The sum of the last column should equal column (2) for each gene.

Formulate a model for the data of one gene, derive and implement an EM algorithm to estimate $\hat{\boldsymbol{\pi}}$. Select a highly sampled gene with multiple isoforms and estimate $\hat{\boldsymbol{\pi}}$ for that gene. How close to the truth are you?

References

[1] Matthew J Moscou and Adam J Bogdanove. "A simple cipher governs DNA recognition by TAL effectors." In: *Science* 326.5959 (2009), p. 1501. DOI: 10.1126/science.1178817. URL: http://dx.doi.org/10.1126/science.1178817.