GGG 201B: Exam 1

Gitanshu Munjal

Question 1:

When aligning nucleotides we often use a simple scoring scheme like +1 match, -1 mismatch. But what if we wanted to make something more biologically relevant? Suppose you are building a program like RepeatMasker, which as the name suggests, masks repeat sequences in genomic DNA...

Thinking out loud:

The simplest (conceptually elementary) way to find repeated elements (in order to mask them) would be to align a sequence to itself and search for repeated elements (visualized as lines based on similar regions from a dot matrix in figure 1), record the position of these elements, and subsequently mask these positions from the sequence. Inherently, there are multiple problems with this approach but the two most serious ones are: (1) the amount of processing power required and (2) confounding detection of elements that are not truly "repetitive" but appear to be so.

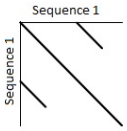

Figure 1: Sequence aligned to itself

The amount of processing power required is directly proportional to the size of the search space and with the present approach the search space, for a genome of size $3x10^9$bp (the human genome for example) would be $9x10^{18}$bp! Running a Smith-Waterman type algorithm on a space of this size might need more computing power than technology can currently provide. However, this problem can be remedied by using a subset of the genome (a library of known repeats for example) as the query. The second problem mentioned earlier can be remedied by defining some thresholds (for sequence length, complexity, and number of occurrences) for what we are willing to accept as a repeat; the single nucleotide "A" in the sequence "AGTCATACCGTATGCATTCGA" should likely not qualify as a "repeat" while a sequence of a thousand consecutive A's should qualify as one but a whole gene duplication should not.

Our approach in finding repeats (and building a repeat masking program) in genomic DNA would thus be to use existing (RepBase for example) or *de novo* (using RepeatModeler for example) repeat libraries as queries against our genomic sequence.

1a. Describe in detail how you would make a scoring matrix for nucleotide sequences. A full credit answer would allow another graduate student to build a scoring matrix from your directions.

If we wanted to make a more biologically relevant scoring scheme (than the mentioned +1/-1), we should calculate pairwise scores for all possible nucleotide (mis)match combinations in a way that emphasizes the biology. As an example, an A-A match should be more valuable in a G-C rich background than a G-G match, transitions should be rare in general, and transversions should be rarer still. One way to calculate such a score is using the formula:

$$S_{ij} = log(\frac{Q_{ij}}{P_i P_j}) = log(\frac{observed frequency}{expected frequency})$$

Where,
S is the score for any two nucleotides i and j
Q_{ij} is the observed substitution frequency (for example from counting in multiple alignments)
P_i is the probability of the nucleotide i (similarly j in case of P_j)

Consider the above formula and a GC rich sequence, the denominator would be smaller for an "A-A" match (smaller frequencies) compared a to "G-G"match and so the score for an "A-A" match would be higher. The rarity of transitions and (higher rarity of transversions) would make the observed frequency small and the overall score negative (more negative for transversions).

Table 1: Nucleotide Scoring Matrix

	A	T	G	C
A				
T				
G				
C				

The grad student could follow the above mentioned game plan and generate the scoring matrix by using the above mentioned formula to fill out the matrix (Table 1) for each row,column(i,j) combination:

1b. Do you make the scoring matrix symmetrical or asymmetrical? Justify your answer.

The above presented idea for scoring matrix is **assymetric** in general. Looking closely at the underlying formula, it follows that $S_{ij} \neq S_{ji}$ for transition and transversion cases under different levels of G-C content. In such cases, the denominator (expected frequencies) would remain the same for S_{ij} & S_{ji} but the numerators would likely be different. Consider the A $\rightarrow$G transition versus the G $\rightarrow$A transition in a G-C rich background for example; we would likely observe more transitions of the first kind than the latter due to simple reason that there are more G's than A's. Subsequently, $S_{AG} \neq S_{GA}$ resulting in an **assymetric matrix** ($ScoringMatrix \neq ScoringMatrix^T$).

1c. How would you determine appropriate gap costs for your scoring matrix?

There does not appear to be a concrete mathematical method to solve for an appropriate gap cost and the choice of gap costs seems somewhat arbitrary even where models (affine, concave, linear) exist. The general idea seems to be to make gaps hard to come by (opening penalty) when close matches are the goal and easier when divergence is the goal. Our goal here is to query a repeat library against a sequence and find "true" repeats to be able to mask them. Under such a circumstance, finding close/exact matches would be the primary goal and so the gap opening penalty should be fairly high and the gap extension penalty once a gap is opened should be comparatively lower. Along these lines, for a given nucleotide composition, we can choose a fixed gap opening penalty of 3 times the worst transition score and a fixed gap extension penalty as 1/3 of the worst transversion score.

1d. Do you expect all positions in a repetitive element to be equally important? In other words, is there any reason to make profile HMMs of repetitive elements? Justify your answer.

Yes there is a need for making profile HMMs for repetitive elements, one example supporting this comes from transposable elements. Remnants of transposable elements at various decay levels make up a lot of the repetitive DNA. Such regions would be hard to capture using methods that depend on sequence homology where TE families are represented by a single consensus sequence making it hard to find distantly related sequences (Wheeler 2013). Profile HMMs are probabilistic models that use (statistical) features of all sequences from a multiple alignment to build the profile to be used for scoring potential matches to new sequences (Durbin 1998).

References

- Durbin, Richard, ed. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.
- Wheeler, Travis J., et al. "Dfam: a database of repetitive DNA based on profile hidden Markov models." *Nucleic acids research* 41.D1 (2013): D70-D82.
- http://openwetware.org/wiki/Wikiomics:Repeat_finding
- GGG 201B Class Notes (Winter 2015)
- http://repeatmasker.org/webrepeatmaskerhelp.html

Question 2:

Studies on intron-mediated enhancement have shown that exons near the start of transcription are different from other introns. Surely you must also wonder about introns near the end of a gene...

2a. How would you determine if introns at the ends of genes are different from other introns? Be very specific in your answer with enough detail that another graduate student could implement your plan.

Are introns at ends of genes different from other introns?

If you look (at) it then you should have put a number on it!

We can begin to answer the above question by calculating the relative entropy between distal introns (at ends of genes) and all introns. The first step would be the need to define "distal" (covered in next section). The relative entropy (Kullback-Leibler distance) is then given by:

$$D(P||Q) = \sum P_i log_2 \frac{P_i}{Q_i}$$

Where,
D is the distance
P is the frequency distribution of nucleotides in distal introns
Q is the frequency distribution of nucleotides in all introns
P_i is the probability of a given nucleotide in the distal introns
Q_i is the probability of a given nucleotide in all introns

Developing the term-IMEter

One could begin by looking at the effect of compositional differences using an approach similar to the one used by *Rose et al. (2008)* and build a word based discriminator exactly like their "IMEter". The IMEter considers frequencies of all possible words of a given length k (set parameter) to return a log-odds score. Similar to their approach (with slight modification), one could calculate IMEter scores as the difference between the sum of logarithms of observed occurence/expected occurence in distal introns (frequency of subsequence of a given length k in promoter-distal introns/frequency of same subsequence in sequences with composition similar to all introns) and the sum of logarithms of observed occurence/expected occurence in proximal introns (frequency of subsequence of a given length k in promoter-proximal introns/frequency of same subsequence in sequences with composition similar to all introns). Mathematically, this can be described as:

$$S = \sum_{i=1+D}^{i \leq L-K-A} log(\frac{Q_{wi}}{P_{wi}})$$

Where,
S is IMEter score
i is the position in sequence
L is the length of intron
K is the word size
w_i is a word of length K starting at position i
D is the length of splice donor site consensus
A is the length of splice acceptor site consensus
P is the frequency distribution of words of size K in promoter-proximal introns
Q is the frequency distribution of words of size K in distal introns

(Equation adapted from *Rose et al. (2008)* except numerator and denominator inverted so distal introns are expected to have a positive score while proximal ones are expected to have a negative score).

Similar to *Rose et al. (2008)*, we can use introns with known positions (relative to transcription start sites) and parent genes represented by full length cDNAs as a starting point for our model and use half of these data for training and the other half for testing.

2b. Describe how you could determine the optimal K-mer size for such a term-IMEter.

Similar to the protocol used by *Rose et al. (2008)*, we could try multiple different combinations of parameter values (k-mer size and nucleotide position of delineation between proximal and distal introns) and test the resulting predictions in a linear regression against actual experimental data to find optimal values by maximizing the correlation (R^2).

2c. How would you determine if there are specific motifs in terminal introns?

Motif finding can be done by analyzing motif signals (words and Position Weight Matrices). One could look for the most frequent words in an enumeration of all words but this is likely not computationally efficient for our (and most other) purposes. *Rose et al. 2008* use the probabilistic NestedMICA method *(Down and Hubbar 2005)* that models motifs as PWMs for motif discovery. This method implements simultaneous (as opposed to step-wise) motif discovery using Nested Sampling (Skilling J) which has the consequence of enhancing sensitivity and scalability *(Down et al. 2007)*. We can implement the same strategy to look for motifs in terminal introns.

2d. Describe an experiment you would perform to determine if terminal introns alter gene expression.

I would conduct the experiment in a model system like *Arabidopsis thaliana*. I would use terminal introns from a few different genes and insert single copies of each at different locations (proximal,central,distal) relative to the promoter within a gene that I can easily measure mRNA accumulation for and compare these measurements with the mRNA accumulation for the same gene when intronless. No difference would tell me that terminal introns do not alter gene expression, constant increase/decrease for a given intron when slided one single copy at a time across the gene would tell me that the position has no effect but maybe intron composition does. If different introns show the same effects, I would look for common motifs between them. The wet lab protocols for this kind of study are detailed in *Rose et al. (2004)*.

References

- Down, Thomas A., and Tim JP Hubbard. "NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence." *Nucleic acids research* 33.5 (2005): 1445-1453.
- Down, Thomas A., et al. "Large-scale discovery of promoter motifs in Drosophila melanogaster." *PLoS computational biology* 3.1 (2007): e7
- Rose, Alan B. "The effect of intron location on intron-mediated enhancement of gene expression in Arabidopsis." *The Plant Journal* 40.5 (2004): 744-751.
- Rose, Alan B., et al. "Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression." *The Plant Cell Online* 20.3 (2008): 543-551.
- Skilling J: Nested Sampling. [http://www.inference.phy.cam.ac.uk/bayesys/] *In American Institute of Physics Conference Series Edited by Fischer R, Preuss R, Toussaint UV.* 2004, 395-405

Question 3:

A flying saucer has been captured and alien life has been found for the first time. As part of the investigative team, you are analyzing alien DNA. You find that there are 5 nucleotides: A, C, G, T, and X. X pairs with itself. Promoters are AT-rich, genes are GC-rich and do not contain X nucleotides, codons are 2 nucleotides long (there are only 16 amino acids), the start codon is GG, the stop codon is XX, introns average 500 nucleotides, exons average 100 nucleotides, splice sites are surprisingly similar to life on Earth, and repetitive elements tend to have a lot of Xs. Design a standard HMM with all these features. You can draw by hand and take a photograph if you prefer.

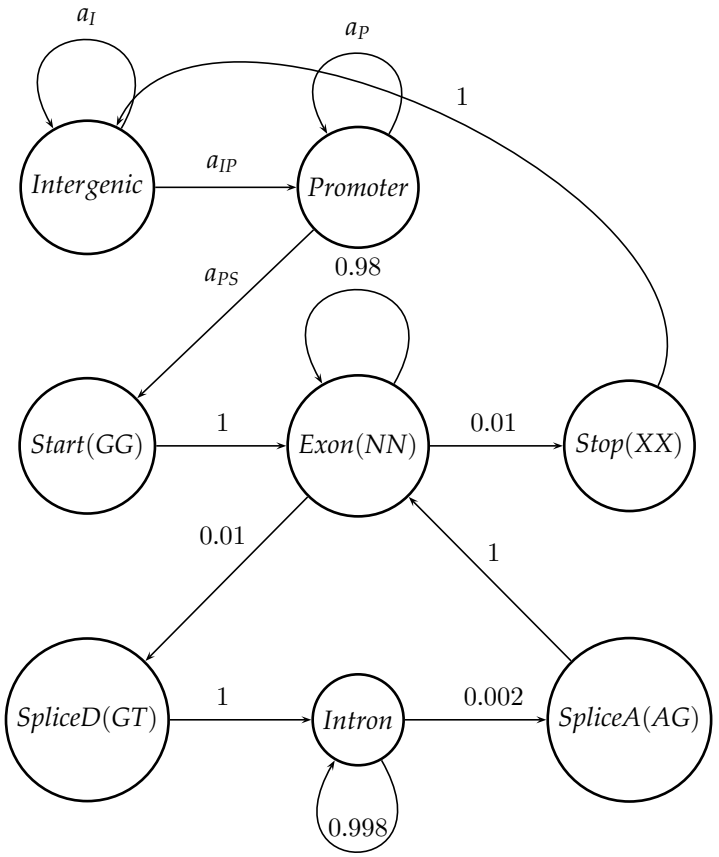

Figure 2: A humble attempt at a hidden Markov model for alien DNA.

Table 2: Emission Freuencies

State	X	T	A	G	C	Total
Intergenic	0.3	0.175	0.175	0.175	0.175	1
Promoter	0.17	0.25	0.25	0.17	0.16	1
Intron	0.2	0.2	0.2	0.2	0.2	1
Start	0	0	0	1	0	1
Stop	1	0	0	0	0	1
Exon	0	0.2	0.2	0.3	0.3	1

States (in the above model; figure 2) with nucleotides in parenthesis sequentially emit two nucleotides (positions as mentioned in parenthesis; where N is any nucleotide) and the transition probability from position 1 to position 2 is 1, arrows coming into these state come to Position 1 and arrows going out of the state leave from Position 2. The exon state for example can be expanded as:

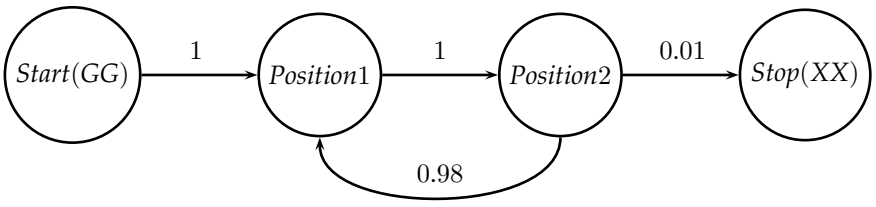

Figure 3: Expanded Exon State (remainder model not shown)

References

- GGG 201B Class Notes (Winter 2015)
