Exploring Bioinformatics: A Project-Based Approach

Key for Test Your Understanding Questions, Chapter 10

1. Suppose you use the sliding-window algorithm described to analyze codon bias. At several points in a DNA sequence, you notice that you see a high score in your first window and a low score in your second window. But, when you slide the window by one or two nucleotides, you get low scores in both windows. How would you explain this pattern? How might you want to account for it in deciding where your exon-intron boundaries are?

   This pattern actually makes sense, because the codon usage depends on the reading frame. So, where you see the high score in the first window and a low score in the second, the first window is likely in-frame with the coding sequence and the boundary between the windows is near an intron-exon boundary. But when you shift one or two nucleotides, the first window is then out-of-frame and it’s no surprise that the codon usage pattern no longer matches expectation.

   A way to use this to your advantage in an exon-prediction program is to deliberately look for cases where you get the high-low pattern at multiple points three nucleotides apart, good evidence your upstream window is in a coding sequence.

2. Explain why the codon-usage method is likely to be imprecise in defining exon-intron boundaries.

   Suppose we find a high score in the upstream window and a low score in the downstream window. Is the boundary between them exactly where the exon-intron boundary occurs? Chances are that sliding three or six nucleotides upstream or downstream will not drastically change the result, so where is the actual boundary? Not to mention that the boundary might be in the middle of a codon, not between them.

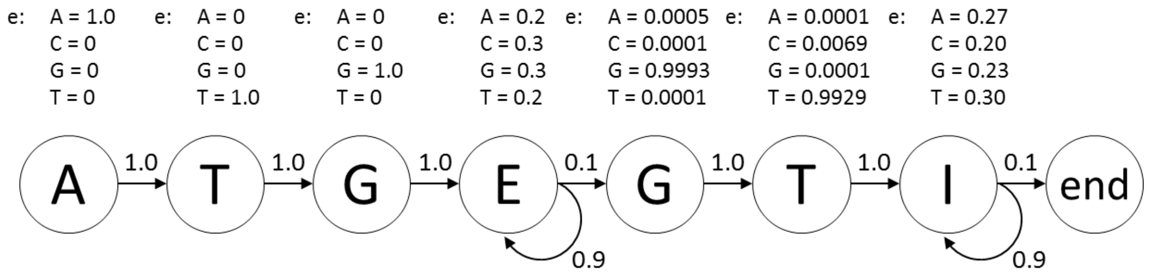
3. CpG island prediction algorithms generally require not only a higher-than-expected frequency of CG pairs but also that the region under examination have an overall higher percentage of G+C than the average in the genome. What is the value of this constraint?

   If a region of high A+T content is being examined, even a few CG pairs will give a CpG frequency much higher than expected, because the A+T-rich region will have a low expected value. We expect the CpG island to be a C+G-rich region, so that should be included in our calculation.

4. CpG islands are associated with promoter regions. How can this help with exon prediction?

   Regardless of what method is used for exon prediction, identifying the first exon can be challenging. However, in eukaryotes, the start codon is (almost) always the *first* AUG codon from the 5′ end of the mRNA, so finding a promoter region can help us figure out where the mRNA might start and thus help identify a genuine start codon.

5. Draw an HMM model which requires an ATG followed by some exon nucleotides, a splice-donor site and then some intron nucleotides.


e: A = 1.0 C = 0 G = 0 T = 0
e: A = 0 C = 0 G = 0 T = 1.0
e: A = 0 C = 0 G = 1.0 T = 0
e: A = 0.2 C = 0.3 G = 0.3 T = 0.2
e: A = 0.0005 C = 0.0001 G = 0.9993 T = 0.0001
e: A = 0.0001 C = 0.0069 G = 0.0001 T = 0.9929
e: A = 0.27 C = 0.20 G = 0.23 T = 0.30
A
1.0
T
1.0
G
1.0
E
0.1
0.9
G
1.0
T
1.0
I
0.1
0.9
end

6. How might the first exon be distinguished from internal exons in an HMM?

Beginning with an ATG and with no preceding splice site; perhaps preceded by a TATA or Inr sequence.

7. Suggest some qualities of a DNA sequence that you would weight positively and some that you would weight negatively in developing a neural network model to identify an exon.

Codon bias could be weighted in here, with close match to the expected coding sequence bias for the organism being weighted positively. Slight G+C bias would also be weighted positively. Containing a good match to the sequences found at the 3′ or 5′ ends of an intron would be weighted negatively, but matches to the exonic portion of that boundary sequence would be weighted positively. Promoter sequences or CpG island upstream would be a positive factor, as would polyadenylation site downstream. Heavily repeated sequences might weight somewhat negatively, as would lack of a match to any protein in a database.