Exploring Bioinformatics: A Project-Based Approach

Key for Web Exploration Questions, Chapter 10

1. List the genes that GENSCAN found within the sequenced region, along with their lengths and the approximate length of the processed mRNAs. Why do the gene arrows point in different directions?

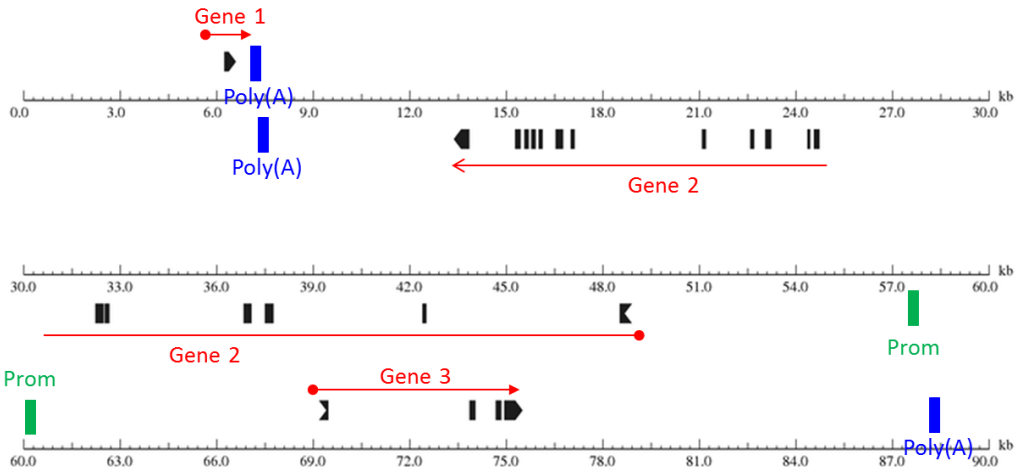
   GENSCAN finds three possible genes in this region (see table at right and graphical representation below).

Gene	Direction	Exons	Length (nt)	Processed Length
1	to right	(1)	130	130
2	to left	18	34,670	3021
3	to right	4	5,859	738

   (1) At the beginning of the sequence is a lone predicted exon; this could be a chance occurrence, but GENSCAN reports that it is followed by a polyadenylation site, so it could well be an exon from a gene that begins in sequence farther upstream that we don't have.

   (2) A multi-exon gene is predicted going from right to left (so, its template strand is on the opposite strand from a left-to-right gene) and taking up much of the sequence. Notice that it begins with a predicted first exon and ends with a predicted last exon and that GENSCAN also finds a predicted promoter before the first exon and a predicted polyadenylation site after the last exon.

   (3) A much shorter predicted gene is found near the end of the sequence, going left to right. This gene has only four exons. It is also preceded by a predicted promoter region and followed by a predicted polyadenylation site.



2. What is the difference between an exon marked Init and an exon marked Intr (in the text output)? Why is this difference significant in predicting genes?

   An exon marked Init is a predicted first exon, starting with a start codon. An exon marked Intr is an internal exon, one between the first and last exons. This is important because there might be an exon-like sequence at random in some intergenic region; if it's not itself an ORF and it's not between a first exon and a last exon, then it can pretty much be ignored.

3. Look at how the predicted proteins begin. Does this information strengthen or weaken the case for any of the genes?

   The proteins from Gene 2 and Gene 3 start with methionine (M or Met), which should be the case if the gene starts with an ATG start codon. Gene 1 does not, but that also makes sense with it being a final exon from some gene that starts in sequence we don't have.

4. What other features did GENSCAN identify (look in the text output)? Do these provide additional support for any of the predicted genes?

   GENSCAN also predicts promoters (Prom) and polyadenylation sites (PlyA; this is a sequence where a eukaryotic mRNA would be cleaved and the poly(A) tail added, essentially like the terminator site). As noted above, these do strengthen the case that Gene 2 and Gene 3 are real, complete genes. Note, though, that the poly(A) site for Gene 3 is pretty far from the last predicted exon (a little odd especially given that the gene seems so compact). The site for Gene 2 is also a little far and given its location,the Gene 1 and Gene 2 mRNAs would overlap.

5. How does the number of genes predicted by AUGUSTUS compare to the results from GENSCAN?

   AUGUSTUS finds only a single gene, with two alternative transcript versions, whereas GENSCAN found three distinct genes. (See why it's desirable to compare different algorithms? The output from GENSCAN seemed so clear and certain before we did this…)

6. How does the structure (e.g., length, number of introns and exons, position in the DNA) of the genes predicted by AUGUSTUS compare to GENSCAN?

   Both programs find a long gene going right-to-left composed of multiple exons. GENSCAN predicts 18 exons, whereas AUGUSTUS finds only 13 that meet its criteria. AUGUSTUS also finds a potential alternative first exon and so constructs a potential alternative transcript that has only 10 exons; this possible alternative first exon appears to be one that GENSCAN considered to be an ordinary, internal exon. GENSCAN's two additional genes go left-to-right, and AUGUSTUS does not recognize these as strong enough to report based on its algorithms.

7. How do the predicted proteins compare? Clearly, they're not identical, but do they appear related? For example, are they basically the same protein with perhaps some different splicing choices, or do they come from entirely different reading frames or even regions of the DNA? (You can of course use EMBOSS or BLAST to directly compare the proteins or their exons if you wish.)

   For the long gene, the two predicted proteins are clearly closely related, but different splicing choices have been made.

8. Describe the gene that you conclude may be important in influenza resistance: total length, number of exons, processed length, number of amino acids, etc.

There is room for interpretation here, so there is no one right answer. For my analysis, I felt skeptical about how much longer the first two introns were for AUGUSTUS' predicted longer transcript as compared to all the other introns. Given that AUGUSTUS gave lower likelihood to its first two exons and there is a high G+C content within the longest intron that could suggest a promoter region, I decided to use AUGUSTUS' shorter transcript for this analysis and ignore the several exons that GENSCAN finds within this region. However, in the absence of any other supporting data, one could equally well make the case that the additional exons found by GENSCAN in this region suggest that the gene is indeed long and that AUGUSTUS failed to find this useful evidence.

Based on my assumption, the gene (actually, the part encoding the pre-mRNA; we have no promoter, etc. data yet) is approximately 11,500 bp long and is composed of 10 exons. The processed mRNA would be 1,470 nt long and encode a protein of 489 amino acids.

9. Do the CpG islands within the sequenced region support your hypothesis about the genes that are found here? Do they provide any information that might help distinguish between the GENSCAN and AUGUSTUS results?

   The major question dividing the two analyses is how far the gene extends: only as far as AUGUSTUS' shorter transcript, or farther out where AUGUSTUS finds two more exons and GENSCAN finds several more. The largest CPG island is between the start of AUGUSTUS' shorter transcript and the second exon of its longer transcript, and there are two shorter islands within the longer transcript as well. This tends to favor the idea that AUGUSTUS' shorter transcript is the correct one, but then another set of potential CpG islands occurs farther upstream, supporting the longer transcript or perhaps suggesting that both could be transcribed actively from different promoters.

10. Higher scores in the NNPP results mean putative promoters that better match the criteria. Note on your map where the strongest predicted promoters are. The large letters the predicted transcriptional start sites. Can you see good matches to the consensus TATA box sequence (tATAWAW) upstream of potential translational starts?

    There are a surprising number of predicted potential promoter sites—plenty to support any of the genes found by the exon-prediction programs. Unfortunately this evidence is not as helpful as we would like.

11. How does the number of promoters returned by TSSG compare with the NNPP results? What else is different about the TSSG results, and how might this difference be useful?

    TSSG found 18 promoters, with the strongest ones all between nucleotides 31,800 and 34,300. This would tend to support the hypothesis that AUGUSTUS' shorter transcript is the best model. TSSG also returns locations of potential TATA boxes and other transcription-factor binding sites, and these also suggest a good possibility that the promoters in this region are valid. There is some support for a promoter upstream of AUGUSTUS' longer transcript as well, but the prediction is less strong.

12. Higher scores from TSSG again represent better promoter predictions. Do any of the high-scoring promoters match up (at least approximately) with high-scoring promoters from NNPP?

    TSSG's predicted promoters actually match up fairly poorly with NNPP's: the strongest TSSG promoters occur in a region where CpG islands are found but where NNPP fails to predict any promoter sites. NNPP does find several promoters a little farther upstream, though.

13. Does your expression analysis help to reconcile the differences between the GENSCAN and AUGUSTUS predictions?

    As noted above, there is some evidence that AUGUSTUS' shorter transcript is the best.

14. Choose the gene you believe is founded on the most solid evidence, obtain its coding sequence, and use BLAST and OMIM to find out what is known about the gene. Hae you actually identified a gene that makes sense in the context of influenza resistance?

    Using AUGUSTUS' shorter transcript, BLAST finds a good match to a protein referred to in GenBank as "Influenza virus NS1A-binding protein isoform X1." As the name suggests, this protein is known to bind a key influenza virus protein and thus could potentially be involved in resistance to the virus. OMIM suggests that the protein plays a role in the replication of influenza virus, so mutating the gene could potentially interfere with virus replication.

    However, the alignment is not perfect, and there are long gaps in the middle of the aligned sequence. This suggests that AUGUSTUS' prediction does not entirely agree with the consensus gene prediction used in the human genome annotation for a couple of the exons. You may want to play with GENSCAN's predicted exons and see if they give a "better" result by the standards of the existing genome annotation.