Eukaryotic Gene Finding

Overview

In this exercise you will:

- Download a sequence file from blackboard
- Predict possible exon structures in two eukaryotic genomic sequences
- Use gene finding programs Genscan, HMMgene and NetGene2
- Evaluate the exon prediction scores
- Evaluate splice site scores
- Evaluate coding region potential

Genscan

- Go to the Genscan server at: http://genes.mit.edu/GENSCAN.html
- Copy-and-paste sequence (including the FASTA header line starting with >gi....) into the DNA field
- Press Run Genscan (use default options)
- Read the explanation at the bottom of the screen and answer the following questions:

  1 How many exons are predicted in Sequence 1 ?
  2 What are the begin and end positions ?
  3 For the possible exons, what is the probability of each
  4 On which strand (+ or -) is the gene located?
  5 Write down the first 6 amino acids and the total length of the predicted protein sequence

HMMgene

- Go to the HMMgene server at: http://www.cbs.dtu.dk/services/HMMgene/
- Copy-and-paste sequence (including the header line starting with >gi....) into the field named Sequence(s) in FASTA format Press Submit sequence (use default options)
  Important! - Wait for prediction to finish
- The link named Explanation of output format will take you to a HELP/DOCUMENTATION page that will explain the output format (This is NOT the prediction on your sequence)
- Go back to the prediction page and answer the following questions:

  1 How many exons are predicted in Sequence ?
  2 What are the begin and end positions ?
  3 For the possible exons, what is the probability of each
  4 On which strand (+ or -) is the gene located?
  5 Compare the exon-intron boundaries with those obtained by Genscan. Do they agree for all exons?

NetGene2

- NetGene predicts potential donor and acceptor splice sites as well as protein coding potential. It does not predict a complete exon-intron gene structure
- Go to the NetGene2 server at: http://www.cbs.dtu.dk/services/NetGene2/
- Cut-and-paste sequence (starting with the header line >gi....) into the field named Sequence
- Press Send file (use default selection of human) and wait for prediction to finish.
- Scroll down to Donor splice sites, direct strand (Direct = + strand; do not look at the predictions for complement(-)strand in this exercise)
- NetGene2 presents you with scores for many potential donor and acceptor splice sites. Consult your results obtained using Genscan and HMMgene, answer the following questions

  1. Based on the predictions from Genscan/HMMgene, at which position do you expect to find a donor splice site?
  2. If NetGene predicts a donor splice site at this position, what is then the confidence score?
  3. Scroll down to Acceptor splice sites, direct strand
  4. Consult your results obtained using Genscan and HMMgene
  5. Based on the predictions from Genscan/HMMgene, at which position do you expect to find an acceptor splice site?
  6. If NetGene predicts an acceptor splice site at this position, what is then the confidence score?

Output of Genescan

- Gn.Ex : gene number, exon number (for reference)
- Type : Init = Initial exon (ATG to 5' splice site)
- Intr = Internal exon (3' splice site to 5' splice site)
- Term = Terminal exon (3' splice site to stop codon)
- Sngl = Single-exon gene (ATG to stop)
- Prom = Promoter (TATA box / initation site)
- PlyA = poly-A signal (consensus: AATAAA)
- S : DNA strand (+ = input strand; - = opposite strand)
- Begin : beginning of exon or signal (numbered on input strand)
- End : end point of exon or signal (numbered on input strand)
- Len : length of exon or signal (bp)
- Fr : reading frame: äbsolute reading frameërelative to start of sequence.

  For example, if nucleotides 1,2,3 of the sequence are read as a codon, that's called reading frame 0. If 2,3,4 are read as a codon, that's reading frame 1. If 3,4,5 are read as a codon, that's reading frame 2, and so on. This information, together with the starting and ending positions of the exon, is sufficient to give the amino acid sequence encoded by the exon.

- Ph : net phase of exon (exon length modulo 3)
  For example, an exon of length 15 bp has net phase 0 since 15 is divisible by 3, an exon of length 16 bp has net phase 1 because 16 divided by 3 leaves a remainder of 1, an exon of length 17 bp has net phase 2, and an exon of length 18 bp has net phase 0 again. The point of this is that exons whose net phase is 0 can be omitted from the gene without disrupting the reading frame: such exons are candidates for being either 1) incorrect, or 2) alternatively spliced.
- I/Ac : initiation signal or 3' splice site score (If below zero, probably not a real acceptor site.)
- Do/T : 5' splice site or termination signal score (If below zero, probably not a real donor site.)
- CodRg : coding region score (tenth bit units)
  Low coding region scores may indicate potentially incorrect predictions or genes with unusual amino acid and/or codon usage patterns.
- P : probability of exon (sum over all parses containing exon)
  This quantity is close to the actual probability that the predicted exon is correct.
- Tscr : exon score (depends on length, I/Ac, Do/T and CodRg scores)