Charlee Cobb - Transcriptomics, Exercise 11

Charlee Cobb

2023-04-10

M11_v2_description) Different levels of genome annotation. M11_v2_1) What are the two different annotations we will try to predict?

The goal is to predict structrual annotation and functional annotation

M11_v2_2) Why does it take consortiums so long to release genome annotations?

They have to be accurate because scientists use their annotations as valididated functions. So consortiums have to be caustious of how they release the data so they take more time to be as accurate as possible.

M11_v2_3) What features can we use to identify where a gene maybe located in the genome?

We can look for the promoter region or polyA signal of a gene, alternative spliced RNA, and mRNA to identify the location of a gene in the genome.

M11_v3_description) Detecting non-coding elements in the genome M11_v3_1) What is a pseudogene?

Pseudogenes are like genes but are missing features from a functional gene. These genes still get processed and transcribed but they don't produce a function.

M11_v3_2) What features can you look for to identify pseudogenes?

Two key features of pseudogenes include absence of 5' promoters and introns. There will often be a segment of the polyA tail inserted into the genome because of reverse transcrpition of the psuedogene that gets integrated into the genome.

M11_v3_3) What proportion of RNA in your cell is from non-coding RNAs?

98% of transcripts consist of non-coding RNAs

M11_v4_description) Different gene structure finding strategies M11_v4_1) Which gene features can help us correctly predict the structure of the gene?

Gene features that help us predict the structure of a gene include open reading frames and identifying the start and stop codons and identifying well studied splice sites.

M11_v4_2) What kind of rules can you apply

You would use rules that build a neural network like decision trees. Rules include wheter or not the sequence is an intron or exon, labeling promoters, and giving known patterns of start and stop codons and splice regions.

M11_v5_description) Using computational methods to predict genes. M11_v5_1) What are the probabilities that are needed for HMM predictions?

Probabiliity of state is the central theme for probability characteristics. This includes probability of transistion from ones state to another, probability of being an intron or exon, probability of going from intergenic to a promoter region and after the promoter going into the five prime UTR, and the probability that the exon will move to an intron.

M11_v5_2) what are the different parts of a neural network?

In a neural network, there is typically an input layer, n number of hidden layers, and an output layer.

M11_v5_3) What is the difference between neural network and HMM approaches in predicting gene structures.

Neural networks will use provided weights on input and will balance them by taking the best score and output the decision it's most confident in. The HMM takes a limited number of inputs and focuses on the probability of state and transition.

M11_v5_4) Give two features of a gene that would make it difficult to predict its structure correctly.

Introns and Exons containing UTRs are difficult to predict, and smaller genes are not statistically significant, so they are often missed.

M11_v6_description) Using experimental methods to predict genes. M11_v6_1) List all the ways an RNA sequence can help correctly predict the gene structure.

Identify boundaries of exons and introns, find alternative splice sites

M11_v6_2) What is an EST and how is it different from RNA-seq?

EST involves extracting TNA from different developmental stages and tissues, making a cDNA library, selecting clones at random, and pass one or both ends to get the result sequence that is called expressed sequence tag. The main difference between RNA-seq and EST is that RNA-seq can capture quantification of the RNA whereas EST was solely used for detection of RNA.

M11_v6_3) What are the advantages and disadvantages of EST sequencing?

Some advantages of EST include an inexpensive method of sequencing, certainty that a sequence comes from a transcribed gene, information directly related to the tissue and developmental stage, and long continuous sequence that includes introns. Disadvnatages include not being able to quantify the amount of RNA in a sample, no regulatory information is available, not capturing all the genes in a sample, and the EST sequences were not always full length.

M11_v7_description) Using comparative genomics to predict genes. M11_v7_1) What is the basic assumption behind using comparative genomics for predicting gene structure or important genome features?

The assumption for using comparative genomics for predicting or importing gene structure and features is that a species genome is highly conserved, so the regions in a genome are useful. Otherwise they would have diverged or become an area for deleterious mutations.

M11_v7_2) What can you capture using comparative genomics but not using experimental methods such as ESTs?

You can capture evolutionary information, and you can see similarities and differences between different species.

M11_v8_description) Gene definition M11_v8_1) What is the difference between a "putative" and an "unknown" gene?

A putative gene is a segment of DNA that is thought to be a gene, but the function of the gene remains unknown. An unknown gene is simply a gene sequence that hasn't been discovered.

M11_v8_2) What evidence do we have that proves hypothetical genes exist?

Hypothetical genes are based on comuptational prediction. Evidence for these genes can be found in experiements that look for these computational predictions.

M11_v9_description) Protein domain identification M11_v9_1) How long can a protein domain be?

Protein domains range from 25 to 500 amino acid residues.

M11_v9_2) What is Interpro and how can we use it to predict gene function?

Interpro is a database where you can enter your protein sequence and then search for the domains that match an entry in the Interpro database. These entries are tied to GO terms that help identify the function of the protein created by your sequence.

M11_v10_description) RNA-seq de novo transcript workflow. M11_v10_1) Describe the workflow of the RNA-seq analysis that requires de novo assembly. How is this different from the workflow where you do have the reference genome?

The steps of de novo assembly in RNA-seq are as follows: Taking fastq files of RNA-seq and run the files through an assembly software like Trinity. Then you take the output contigs and map the transcripts. You then can put your contig sequences into InterproScan to get functional prediction of all the discovered genes, or you can quantify the gene counts with packages like DESeq2. Finally, can use the quantification output for Gene set enrichment analysis. When you have a reference genome, the main difference is that you use an aligner and align the reads in a fastq file to the reference instead of matching the pieces together.