Khan Inan
Exercise 11
BI-GY 7633 Transcriptomics

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

The two different annotations are structural annotations and functional annotations

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

It is mainly because the automation process is still not completely refined due to it being difficult to annotate large genomes that are fragmented. Additionally since there can be errors due to contamination in genomic assemblies, this also delays annotation releases.

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

Genes can be inspected and also compared to a map to determine the location of genes on the chromosome. These maps can offer things like cytogenetic location in order to cross check and validate the locations of genes.

M11_v3_description) Detecting non-coding elements in the genome

M11_v3_1) What is a pseudogene?

A pseudogene is best describe an an imperfect copy of a part of a gene that is ultimately functional

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

Pseudogenes usually give themselves away via certain mutations like frameshifts and also stop codon that are premature.

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

The vast majority of the RNA is non-coding, more specifically around 98% of the total

M11_v4_description) Different gene structure finding strategies

M11_v4_1) Which gene features can help us correctly predict the structure of the gene?

There are certain features of a gene that are always constant and can be predicted such as the presence of a start and stop codon

M11_v4_2) What kind of rules can you apply

As I said early the start and stop codons are always present consistently. Additionally there are also other rules like nucleotide ratios according Chargaff's rules

M11_v5_description) Using computational methods to predict genes.

M11_v5_1) What are the probabilities that are needed for HMM predictions?

The three probabilities are (A, B, π). A is the matrix of state probability, B is the vector of state probability and π is the vector of initial state probability

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

There are input layers for the input fields, hidden layers and lastly output layers for the fields that are the targets

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

I would say the main difference is where HMM and neural networks take their inputs from. An HMM can take user inputs and sets of strings of inputs for its model, meanwhile a neural network can take from a database or anything that is high dimensional.

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

The two biggest features are repetitions as well as the difficulty is locating mutations on the sequence. Both of these require thorough vetting and cross-checking to make sure they are taken into consideration

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.

Using free energy minimization is the most common way and it essentially involves approximating, mapping as well as matching and translating for determining gene structure

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

Essentially EST results in a sequence that can be used during the microarray process, but meanwhile RNA-seq just combines both of these processes together into one.

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

The main advantage of EST is how quick, efficient and cheap it is to do. The main disadvantage is that EST is always a single pass process, and this can lead to various sequencing errors, some which can even be difficult to correct without repeating the sequencing process.

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 biggest assumption behind comparative genomics is that there is always some relation between different species genetically, and these relations can be used to gain a better understanding of which sections of genomes are coding and non-coding

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

The main thing that comparative genomics can capture compared to other methods is the actually function of the genes being analyzed, and the role that these genes play is biological systems and behaviors

M11_v8_description) Gene definition

M11_v8_1) What is the difference between a “putative” and an “unknown” gene?

A putative gene is a section of DNA that is believed to be a gene while an unknown gene is a gene that is unknown as to its purpose or function

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

The evidence lies in thousands of sequences used in algorithms that show that there are genes where we cannot determine their function solely through comparisons and similarities

M11_v9_description) Protein domain identification

M11_v9_1) How long can a protein domain be?

They are usually 100 nucleotides long but the absolute maximum is around 200. It can range from 50-200 amino acids

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

Interpro essentially does a lot of computational work and it groups genes and proteins into families and groups. It uses these groups to determine gene functions

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 workflow that involves de novo assembly is different from the workflow where you have a reference genome because you are essentially developing your own reference genome, so to speak. It is because of this that De novo assembly requires immense amount of repetition, because we are developing a standard to compare our results to, where on the other hand, if we had a reference genome, that would be our point of reference