Assignment 5 BioInformatics

Filippo Casari

November 2022

1 Point 1

Question

Terminology review: differentiate DNA, gene, and genome. Your explanation must demonstrate that you clearly understand what each term means distinctively of the others. (max. 150 words).

Answer

- **DNA**: Nucleotides are a group of complex chemicals that make up DNA. It serves as a set of instructions for how to create and sustain you and contains the genetic information of life.

  Nearly every human cell contains DNA in the nucleus, which is the center of the cell.

  Except for identical twins, everyone's DNA is distinct.

- **Gene**: Although the DNA contained in our cells functions as a molecular instruction manual, it is not just a collection of unorganized letters. It is divided into small sections, or paragraphs, each of which contains a series of instructions on how to create a particular element of you.

  This tiny paragraph is a gene, which is a small piece of DNA.

  According to scientists, our genetic code consists of about 23,000 genes.

  Our cells employ genes as instructions to create molecules known as proteins. Proteins play a wide variety of roles. They support the structure and function of cells as well as their communication.

  Some genes instruct cells on how to produce proteins necessary for cell division and development.

- **Genome**: A genome is an organism's whole DNA composition.

  The genome would be the complete instruction manual if the DNA code were a collection of rules that were meticulously organized into paragraphs (genes) and chapters (chromosomes).

> The arrangement of the human genome, chromosomes, and genes is essentially universal. The only thing that differs somewhat is the DNA code, which is the text on the page. That is what distinguishes us.
>
> In 2003, the human genome was first sequenced. Scientists have read every letter that makes up our genome as part of this endeavor, but this is pointless if we don't know what it means.
>
> On this, scientists are now working. They are gradually learning more about every component of our genome.

2 Point 2

Question

Briefly describe gene prediction and why we do it. (max. 150 words)

Answer

Finding the genomic DNA regions that encode genes is known as gene prediction or gene finding in computational biology. In addition to RNA genes and protein-coding genes, this may also include predictions of additional functional components like regulatory regions. Once a species' genome has been sequenced, one of the earliest and most crucial steps in comprehending it is gene discovery. Gene discovery was first focused on laborious research on living cells and creatures.

A genetic map indicating the general placement of known genes in relation to one another might be created by combining data from numerous such studies and statistical analysis of the rates of homologous recombination of several different genes. Today, gene discovery has been reframed as a mostly computational task given the availability of a complete genome sequence and sophisticated computing resources to the scientific community. It's important to distinguish between figuring out whether a sequence is functional and figuring out what a gene or its product does. One of the crucial phases in genome annotation, after sequence assembly, non-coding region filtering, and repeat masking, is gene prediction.

3 Point 3

Question

Describe and differentiate the three main classifications of gene prediction programs: ab initio based, homology based, and consensus based. (max. 150 words)

Answer

- Gene structure is used as a template to find genes in one class of computational approaches for gene identification, also known as ab **initio prediction**. Signal sensors and content sensors are two different types of sequence information that are used in ab initio gene predictions. Short sequence motifs like splice sites, branch points, polypyrimidine tracts, start codons, and stop codons are referred to as signal sensors. Exon detection must rely on the content sensors, which are the patterns of codon usage particular to a species, and enable statistical detection methods to identify coding sequences from the surrounding non-coding regions.

- **Homology-based**: The alignment of a protein (or RNA sequence in the form of full-length mRNA, cDNA, or EST) with the genome sequence that we want to annotate is used in these methods to predict a gene. The prediction is guided by the known sequence, often known as evidence.

- **Consensus-based** methods incorporate both ab initio and homology-based

4 Point 4

Question

In gene prediction, discuss the characteristics in a DNA sequence that have been statistically shown to point to or act as a kind of marker or clue signifying a coding region, i.e., genes/exons (max. 150 words).

Answer

Starting by defining the ORF. It is the part of DNA that has to be translated. In non-coding DNA, stop codons are found every 20 codons or so. Therefore, an ORF threshold of 50–60 codons indicates the presence of a gene coding region. The third codon location in coding sequences is G or C rather than A or T. That is why we have to look for G and C in the sequence.
We have to underline 2 statistics to consider that part of DNA relevant:

- The third codon position is where the presence of G or C is evaluated by **GC bias**.

- **TESTCODE**: determines whether the same nucleotide is repeated at the third codon position.

However these metrics could miss atypical genes.

5 Point 5

Question

Why do you think that gene prediction is a difficult task? (max. 300 words)

Answer

Gene prediction is usually a difficult task because of the poor presence of conservative motifs. However, it is easier for prokaryotes than eukaryotes in which we have introns, exons and higher variability of the genetic code. Indeed, a splicing mechanism removes non-coding sequences (introns) from pre-mRNA after the transcription of protein coding regions is initiated at certain promoter sequences, leaving the protein-encoding exons. The mature mRNA that results from the removal of the introns and some other alterations to the mature RNA can then be translated in the 5 to 3 direction, typically from the first start codon to the first stop codon. The ORF corresponding to an encoded gene will be broken up by the presence of introns that typically produce stop codons since intron sequences are present in the genomic DNA sequences of eukaryotes. That is the reason why gene prediction for eukaryotes is really difficult.

6 Bibliography

- Professor's slides
- Notes
- *https : //en.wikipedia.org/wiki/Gene_prediction*
- *https : //www.sciencedirect.com/topics/medicine − and − dentistry/gene − prediction*