

**CS 590 Introduction to Bio-informatics Spring 2021**

2nd Quiz Guide

**Date and time**

Friday Feb. 26th , 4:30 – 5:300 P.M. (ET115)

*Major topics:*

# Finding Genes

* What is a gene?

A gene is a basic physical and functional unit of heredity. Genes are made up of DNA. Some genes are made up of DNA. Some genes act as instructions to make molecules(Phân tử) called proteins. However, many genes do not code for proteins. In human, gene vary in size of a few hundreds’ DNA bases to more than 2 million bases.

* The genetic code: each group of three nucleotides in the DNA sequences codes for different amino acid in the protein sequence. For examples, the first 3 nucleotides, ATG, code for M (methionine).

Find the frequency of DNA “words” of certain length.: **count(seq, wordsize)**

When:

wordsize=1 : bases

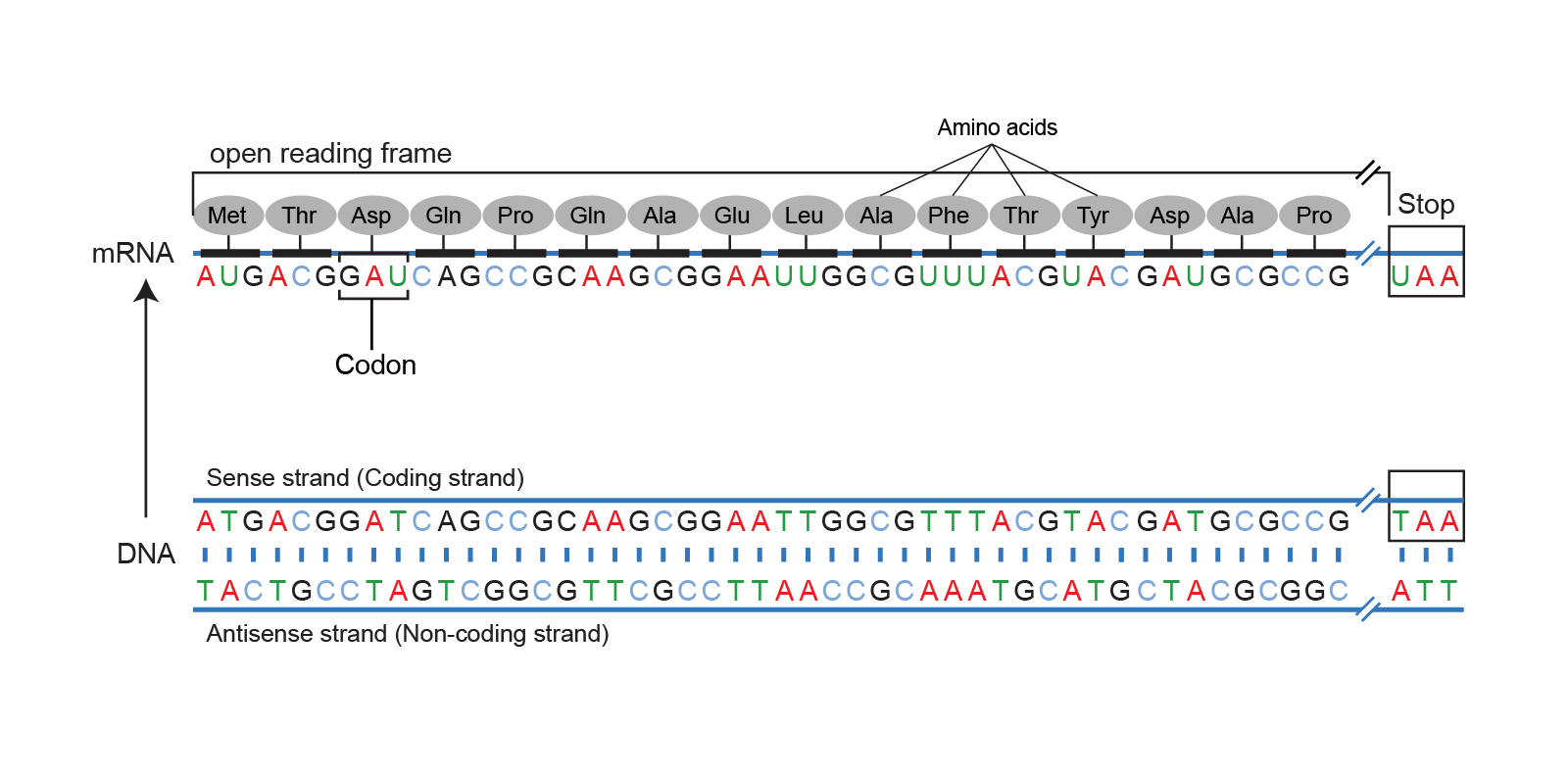
wordsize=2 : bp words

wordsize=3 : codons (a sequence of three nucleotides which together form a unit of genetic code in a DNA or RNA molecule. OR: Each group of three bases (DNA triplets) in mRNA constitutes a codon). Each codon specifies an amino acid.

* start and stop codons is special codon. **Use the tablecode () in the SeqinR package**

How to find a gene: A protein-coding gene: Start with an ‘ATG’ (Start codon). Followed by an integer (whole) number for codons that code for amino acids. End with “TGA”, “TAA”, or “TAG” (stop codon)

* open reading frames
  + *What is mRNAs*
  + *What is open reading frames ORF?*
  + An open reading frame is a portion of a DNA molecule that when translated into amino acids, contains no stop codons . The genetic code reads DNA sequences in groups of three base pairs, which means that a double-stranded DNA molecule can read in any of six possible reading frames
  + Genes in a genome sequences can occur either on the forward (plus) strand of the DNA, or on the reverse (minus) strand. Use comp() and rev() from the SeqinR package.
  + Many ORFs in a DNA sequence may not correspond to a real genes but happened by change to be found in the sequence (Genes prediction can be true positive gene prediction, or false positive gene prediction).
* Significant ORF:
  + locate an ORF by using the function plotPotentialStartsAndStop2()
  + Computationally, use function findORFsinSeq(): return a list variable: the start positions of ORFs, end positions of those ORF, length of ORFs.
  + Real gene: is the longest ORF found in a random sequences of the same length and nucleotide composition as our original sequence.
  + Methods:
  + Generate a set of random sequences using a multinomial mode▪
  + Use function generateSeqsWithMultinomialModel() to find ORFs in all random sequences
  + Use function findORFsinSeq() to Define a threshold to filter-out all shorter ORFs:
  + function max(): Compare to the longest of the ORF occur in the random sequence.
  + Use quantile(): Compare to the longest 99% of the ORF in the random sequences.



* R examples:
  + use the **matchPatterns()** of BioString packages to find all occurrences (aka “matches” or hits) of a given pattern (typical short) in a (typical long) reference sequences or set of reference sequences (aka the subject)
  + Random DNA sequences: use **sample()** Randomly select bases for a sequence of a specific length.

# Pair-wise Sequence Alignment

* Introduction:

Homologous: Living organisms are the product of an evolutionary history. In classical biology, organs that derive from a recent common ancestor are called homologous. In contrast, organs with similar functionality but anatomically very difference are analogous.

* What is an alignment?

Pairwise Sequence Alignment: Identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (DNA, RNA, protein).

Sequence data is so powerful: Inferring a newly sequenced gene’s function is to find similarities with genes of know function. Estimation protein structure from a related protein with know structure and similar sequence. -> Phylogenetic tree

* Scoring an alignment:
  + Mutation: DNA Sequences and protein sequences encode change in evolutionary time through mutation. The simplest types of mutation are point mutations and insertion/deletions, also know as indels.
  + Simple scoring schema:
    - +1: match premium
    - -u: mismatch penalty
    - -o: indel penalty

Score = #matches -u(#mismatches) -o(#indels)

* Scoring Indels: in nature, number of indels usually **came in as a series.**
* Dot plot:
* A first step in comparing two protein, RNA or DNA sequences
* A graphical method:
* A two-dimensional matrix (like a grid), which has the sequences being compared along the vertical and horizontal axes.
* Allows the comparison of two protein or DNA sequences.
* Identify regions of close similarity:
  + Individual cells are shaded black if residues are identical.
  + Matching sequence segments appear as runs of diagonal lines across the matrix.
* Dynamic programming method:
  + Dynamic Programming Alignment allow Optimal Alignment between two sequences:
    - Needlman and Wunsch Algorithm (1970) for global alignment
    - Smith and Waterman Algorithm for local alignment
* Global alignment: If two sequences are homologous across their entire length.
* Local alignment: If sequences only share homologous regions while other parts of the molecule are quite unrelated.
* Sequence alignment in R

Running time: O(mxn) where m and n are the lengths of the 2 strings

* Statistical significance :

Create random sequences have the same amino acid composition and length as one of the two sequences. Align one sequences with one of the random sequences by the Needlman Wunschal Algorithm. Repeated do it to all the random sequences. Compare the alignment score. If P\_value > 0.05: The alignment score is not statistically significant -> the sequences are not related.

# Pair-wise Protein Sequence Alignment

* Protein sequence data:

Protein is made of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains.

Each group of three bases in mRNA constitutes a codon, each group of three base in mRNA constitutes a codon.

Each codon specifies a particular amino acid.

There are 20 different types of amino acids that can be combined to make a protein.

* Protein sequence databases:

NCBI is database for DNA sequences

RefSeq and Uniprot are manually curated databases.

Additional information added to the entries:

Scientific paper that describe the sequences

The biological function

Expression

Protein to protein interaction.

The amount quality of manually crated protein sequences in Uniprot is much higher than that in RefSeq.

* PAM vs. BLOSUM
* Sequence Alignment
* Statistical significance
* R code

**Ask Professor About the Homework Assignment**

**Protein? DNA or RNA sequences**