



Universidad Nacional de San Agustín de Arequipa

Bioinformatics

Genome by numbers

MSc. Vicente Machaca Arceda

July 2, 2020

Table of Contents



Introduction

Objectives

The biology of cells: Summary

Numbers and Databases

Genomes by numbers

Databases

DNA sequence formats

Table of Contents



Introduction

Objectives

The biology of cells: Summary

Numbers and Databases

Genomes by numbers

Databases

DNA sequence formats

Objectives



- ▶ Understand the size of our genome.

Objectives



- ▶ Understand the size of our genome.
- ▶ Understand the lack of research in Genomics.

Introduction

The biology of cells: Summary

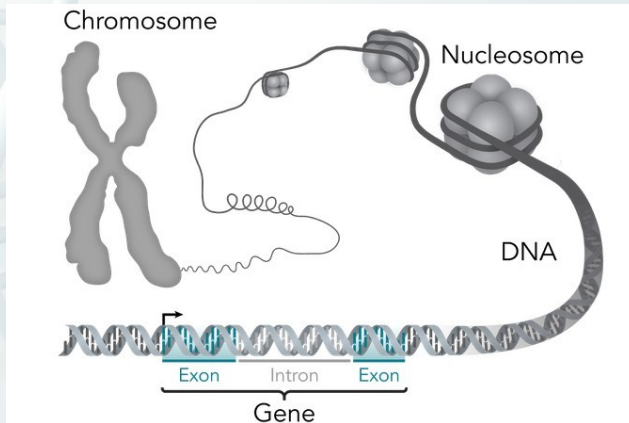


Figure: Chromosome-DNA-gene [1].

Introduction

The biology of cells: Summary

5

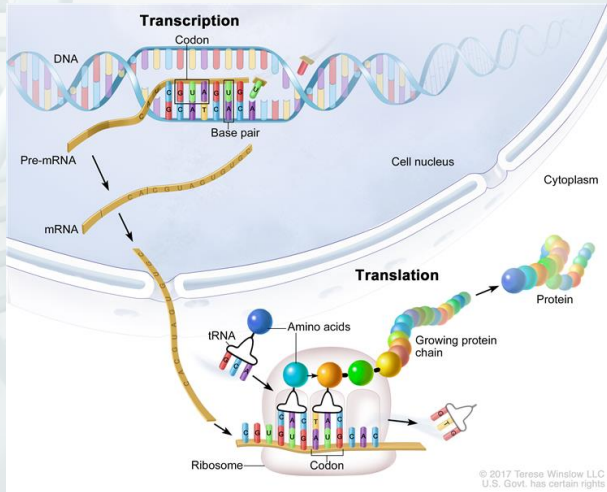


Figure: Transcription and translation [2].

Introduction

The biology of cells: Summary

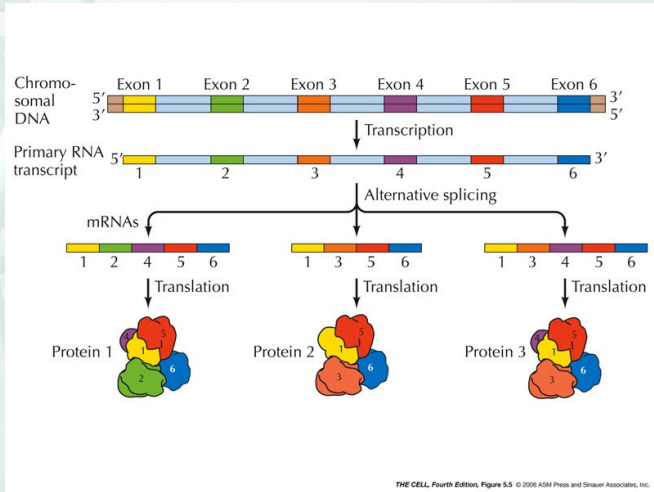


Figure: Alternative splicing [3].

Table of Contents



Introduction

Objectives

The biology of cells: Summary

Numbers and Databases

Genomes by numbers

Databases

DNA sequence formats

Numbers and Databases

Numbers of base pairs



The human genome is made of ~**3.2 billions bp** of DNA.
~6.4 billions of nucleotides [4].

The HIV-1 genome is made of ~**20k bp** of DNA.
Meanwhile, the COVID-19 is made of ~**32k bp** [5].

Numbers and Databases

Numbers of genes



There are approximately **19000** to **25000** genes.
No one knows for sure [4].

Numbers and Databases

Percentage of protein-coding genes



Only ~**1 per cent** of the human genome correspond to protein-coding genes. [4].

Human genes have dozens of introns, each of which can be tens of thousands of nucleotides. Distinguishing exons from introns and other forms of non-coding DNA is challenging [4].

Numbers and Databases

Databases



Database	Description
GenBank	Genetic sequence database
BLAST	Finds regions of similarity between sequences
ViPR	Viral genomes database
TCGA	The Cancer Genome Atlas
ICGC	International Cancer Genome Consortium

Numbers and Databases

Databases types



Primary	Secondary
GenBank	RefSeq
UniProt	Genes
PubMed	Taxon
PMC	OMIM
Intact	ICGC



Downloading Sequencing Data: Unsustainable Model



Figure: Downloading sequencing data

Numbers and Databases

Cloud computing and new software paradigm



- ▶ Data sets are in Petabyte and soon Exabyte scale.
- ▶ Data (and the security rules that come with it) will be somewhere (not in our own data centre), and you will move your software to it.

Numbers and Databases

Cloud computing and new software paradigm



- ▶ FASTA, FASTAQ.
- ▶ EMBL.
- ▶ GCG.
- ▶ GenBank.

Numbers and Databases

FASTA format



```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin  
ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC  
CCTGGAGGGGTGGCCCCACCGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC  
CTCCTGACTTTCTCGCTTGTTGTTGAGTGGACCTCCCAGGCCAGTGCCGGGGCCCTCATAGGAGAGG  
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC  
CTGCAGGAACCTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAACCTCACCCATGAATGCTCACGCAAG  
TTTAATTACAGACCTGAA
```

Figure: FASTA format example

Numbers and Databases

EMBL format



```
ID      AB000263 standard; RNA; PRI; 368 BP.
XX
AC      AB000263;
XX
DE      Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ      Sequence 368 BP;
acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg      60
ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg      120
caggaataag gaaaagcagc ctctgactt tctcgccttg gtggtttgag tggacctccc      180
aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag      240
gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga      300
agaccttctc ctcttgcaaa taaaacctca cccatgaatg ctcacgcaag ttttaattaca      360
gacctgaa
```

Figure: EMBL format example

Numbers and Databases

A2M format



The **A2M format** is used as the primary format for multiple alignments of protein or nucleic-acid sequences. For proteins, the legal alphabet is:

- ▶ ACDEFGHIKLMNPQRSTVWY for amino acids
- ▶ X for any amino acid
- ▶ B for N or D
- ▶ Z for Q or E
- ▶ O for creating a free-insertion module (FIM)

For nucleic acids, the legal alphabet in SAM is:

- ▶ ACGTU for nucleotides (with T and U considered equivalent)
- ▶ Y for C or T
- ▶ R for A or G
- ▶ N for any nucleotide
- ▶ O for creating a free-insertion module (FIM)



Figure: GOBLET



Register to the following courses and bring yours certificated of accomplish:

- ▶ Introduction to Genomics (4 hours)



- [1] Wikicommons, “Chromosome-dna-gene,” <https://commons.wikimedia.org/wiki/File:Chromosome-DNA-gene.png>, 2020, accessed: 2020-03-20.
- [2] NCI, “Nci dictionary of cancer terms,” <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription>, 2020, accessed: 2020-03-20.
- [3] G. BIO, “Gen. bio,” <https://sites.google.com/site/bio1040genbio2/home>, 2020, accessed: 2020-03-20.
- [4] J. M. Archibald, *Genomics: A Very Short Introduction*. Oxford University Press, 2018, vol. 559.
- [5] G. S. Randhawa, M. P. Soltysiak, H. El Roz, C. P. de Souza, K. A. Hill, and L. Kari, “Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study,” *bioRxiv*, 2020.

References II



A white 3D DNA double helix structure is positioned on the left side of the image, extending from the top to the bottom. The background is a light blue gradient. The text "Thank you" is centered in the middle of the image.

Thank you