Escuela Profesional de
Ciencia de la Computación

Bioinformatics

Genome by numbers

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín de Arequipa

April 28, 2020

Table of Contents

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Objectives

- Understand the size of our genome.

vmachacaa@unsa.edu.pe

Objectives

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

vmachacaa@unsa.edu.pe

Introduction

The biology of cells: Summary

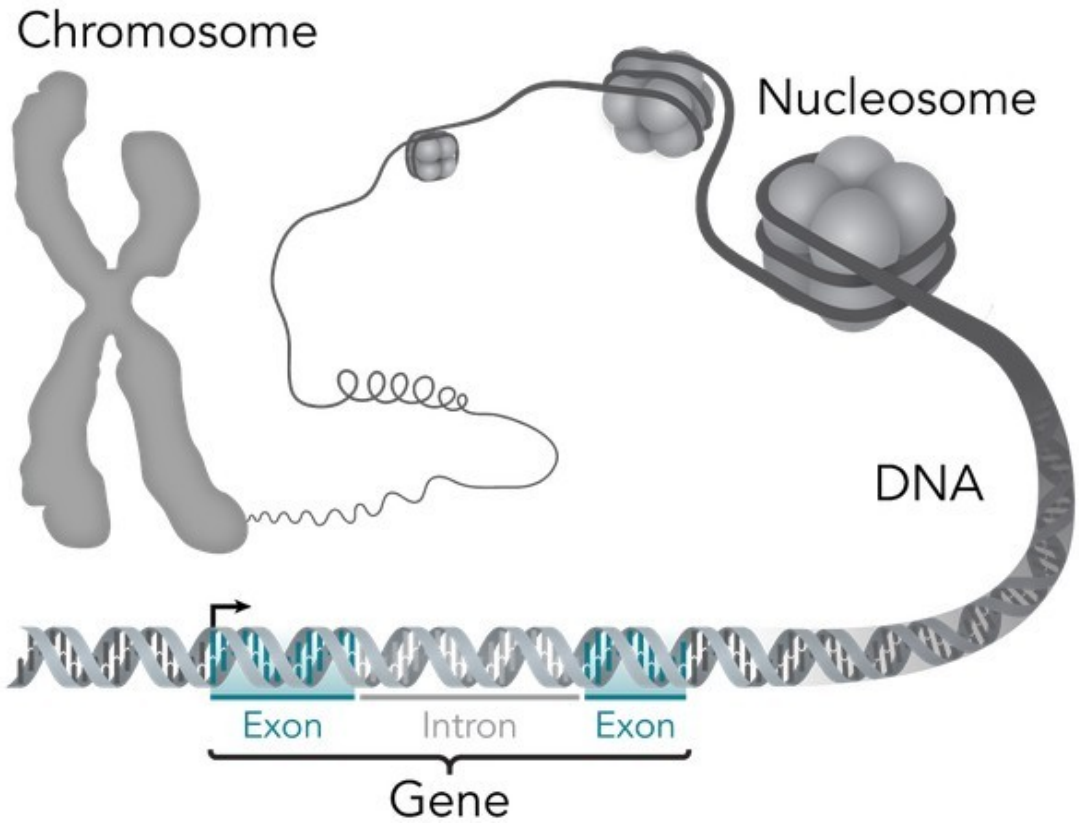

Figure: Chromosome-DNA-gene [1].

vmachacaa@unsa.edu.pe

Introduction

The biology of cells: Summary

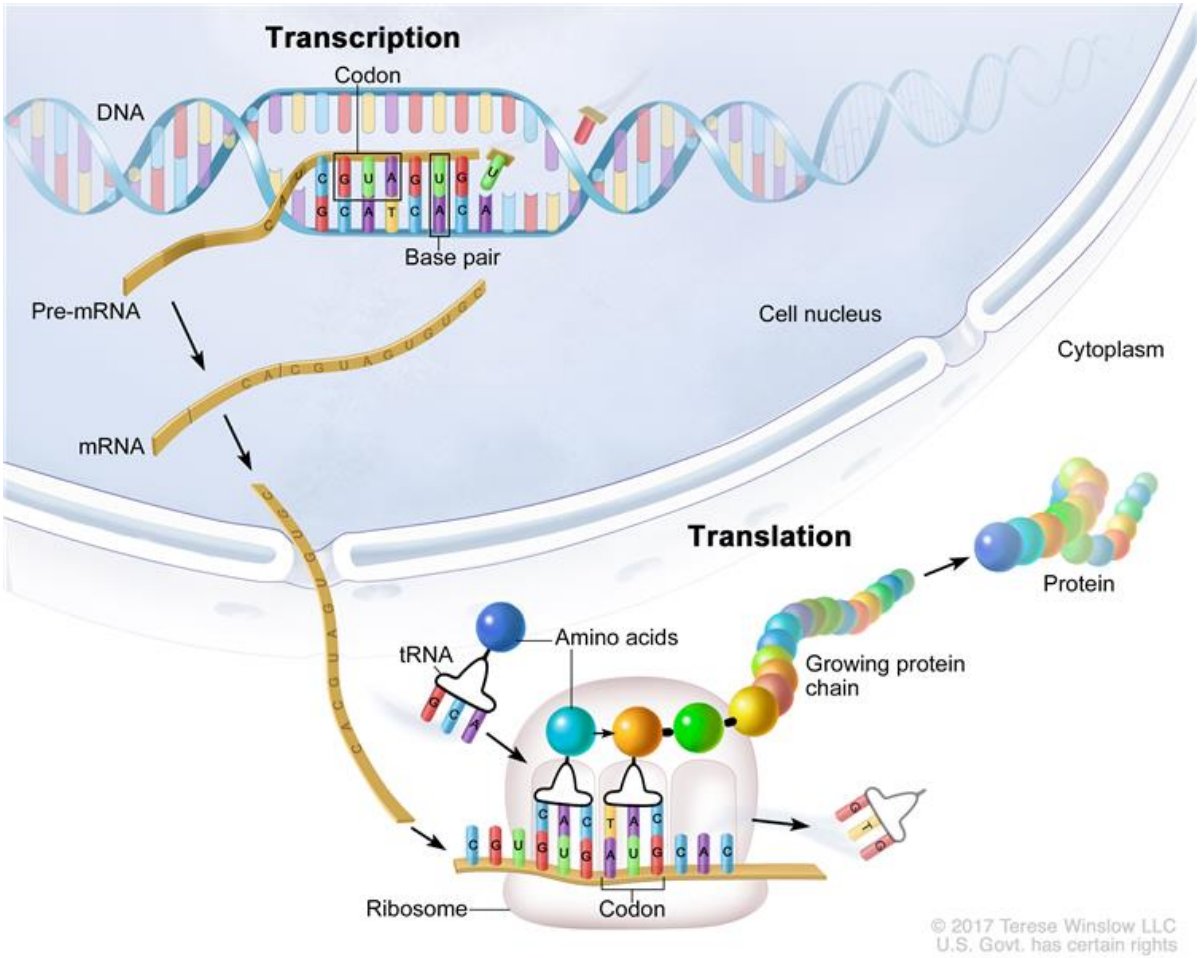

Figure: Transcription and translation [2].

vmachacaa@unsa.edu.pe

Introduction

The biology of cells: Summary

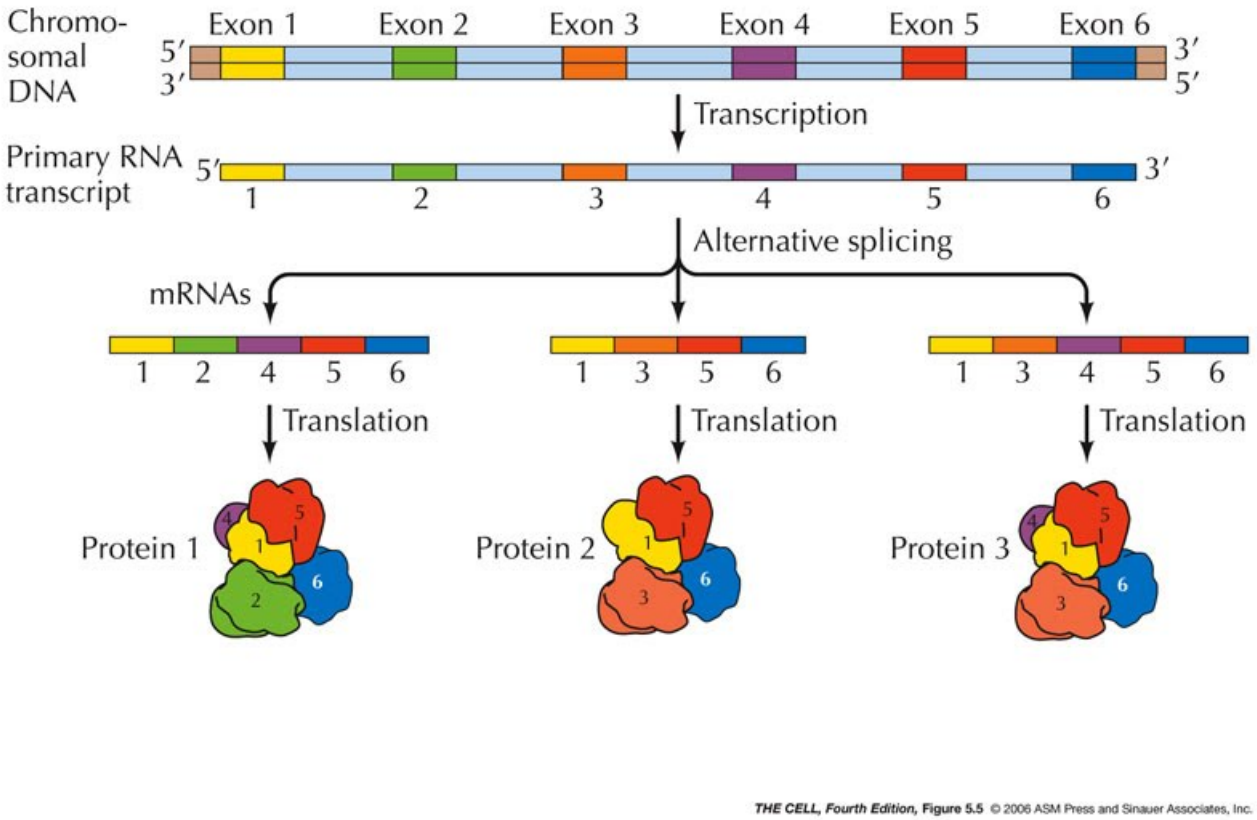

Figure: Alternative splicing [3].

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

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].

vmachacaa@unsa.edu.pe

Numbers and Databases

Numbers of genes

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

vmachacaa@unsa.edu.pe

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].

vmachacaa@unsa.edu.pe

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

vmachacaa@unsa.edu.pe

Numbers and Databases

Databases types

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

vmachacaa@unsa.edu.pe

Numbers and Databases

Databases

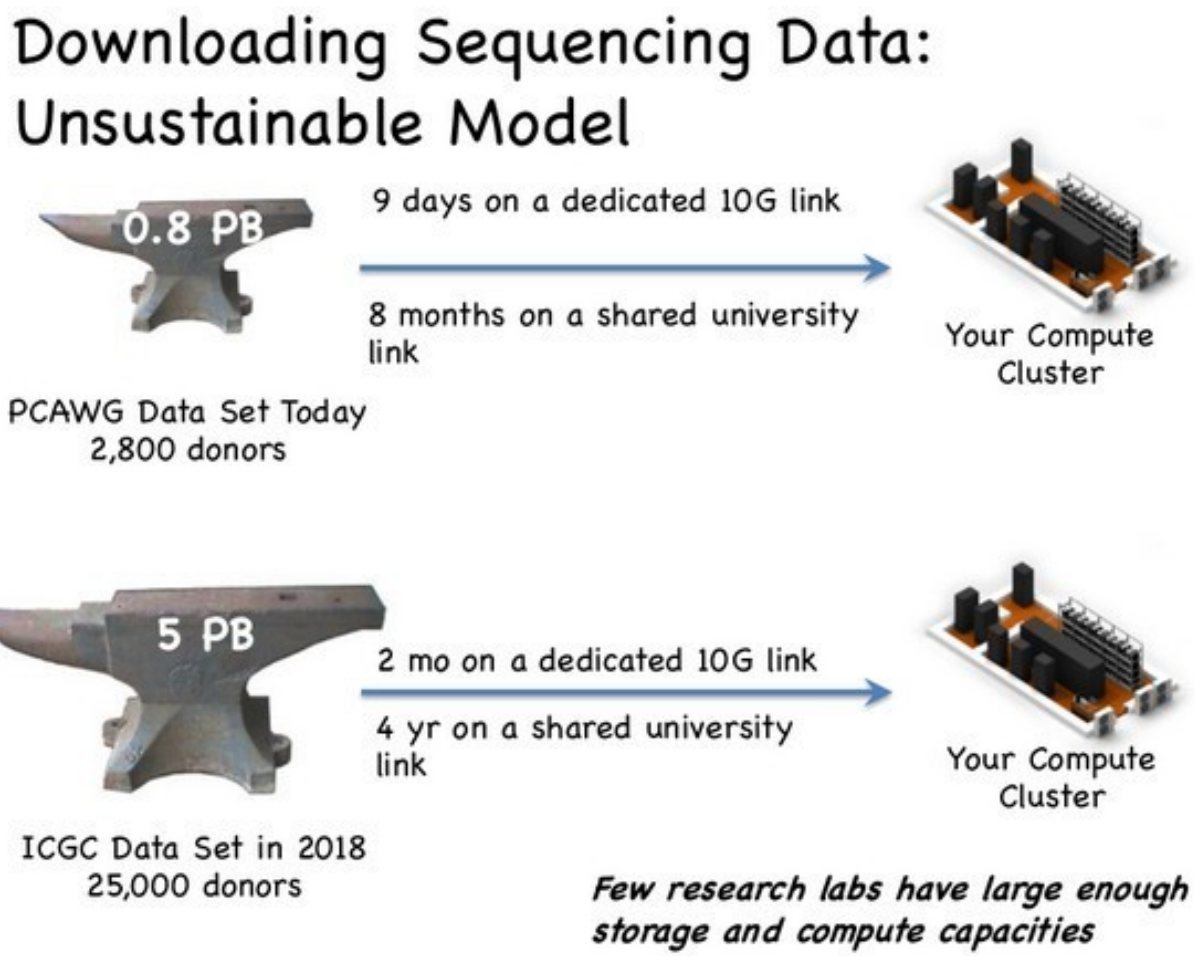

Figure: Downloading sequencing data

vmachacaa@unsa.edu.pe

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.

vmachacaa@unsa.edu.pe

Numbers and Databases

Cloud computing and new software paradigm

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

vmachacaa@unsa.edu.pe

Numbers and Databases

FASTA format

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

Figure: FASTA format example

vmachacaa@unsa.edu.pe

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 ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
```

Figure: EMBL format example

vmachacaa@unsa.edu.pe

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)

vmachacaa@unsa.edu.pe

Numbers and Databases

GOBLET

Figure: GOBLET

vmachacaa@unsa.edu.pe

Bioinformatics

Homework

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

- Introduction to Genomics (4 hours)

vmachacaa@unsa.edu.pe

References I


Wikicommons, “Chromosome-dna-gene,” https://commons.wikimedia.org/wiki/File:Chromosome-DNA-gene.png, 2020, accessed: 2020-03-20.


NCI, “Nci dictionary of cancer terms,” https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription, 2020, accessed: 2020-03-20.


G. BIO, “Gen. bio,” https://sites.google.com/site/bio1040genbio2/home, 2020, accessed: 2020-03-20.


J. M. Archibald, *Genomics: A Very Short Introduction*. Oxford University Press, 2018, vol. 559.

vmachacaa@unsa.edu.pe

References II

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.

vmachacaa@unsa.edu.pe