Introduction to Bio-Informatics
Lecture 4

Lecture by: Ahmad R. Naghsh-Nilchi, PhD
Department of Artificial Intelligence
Faculty of Computer Engineering
University of Isfahan

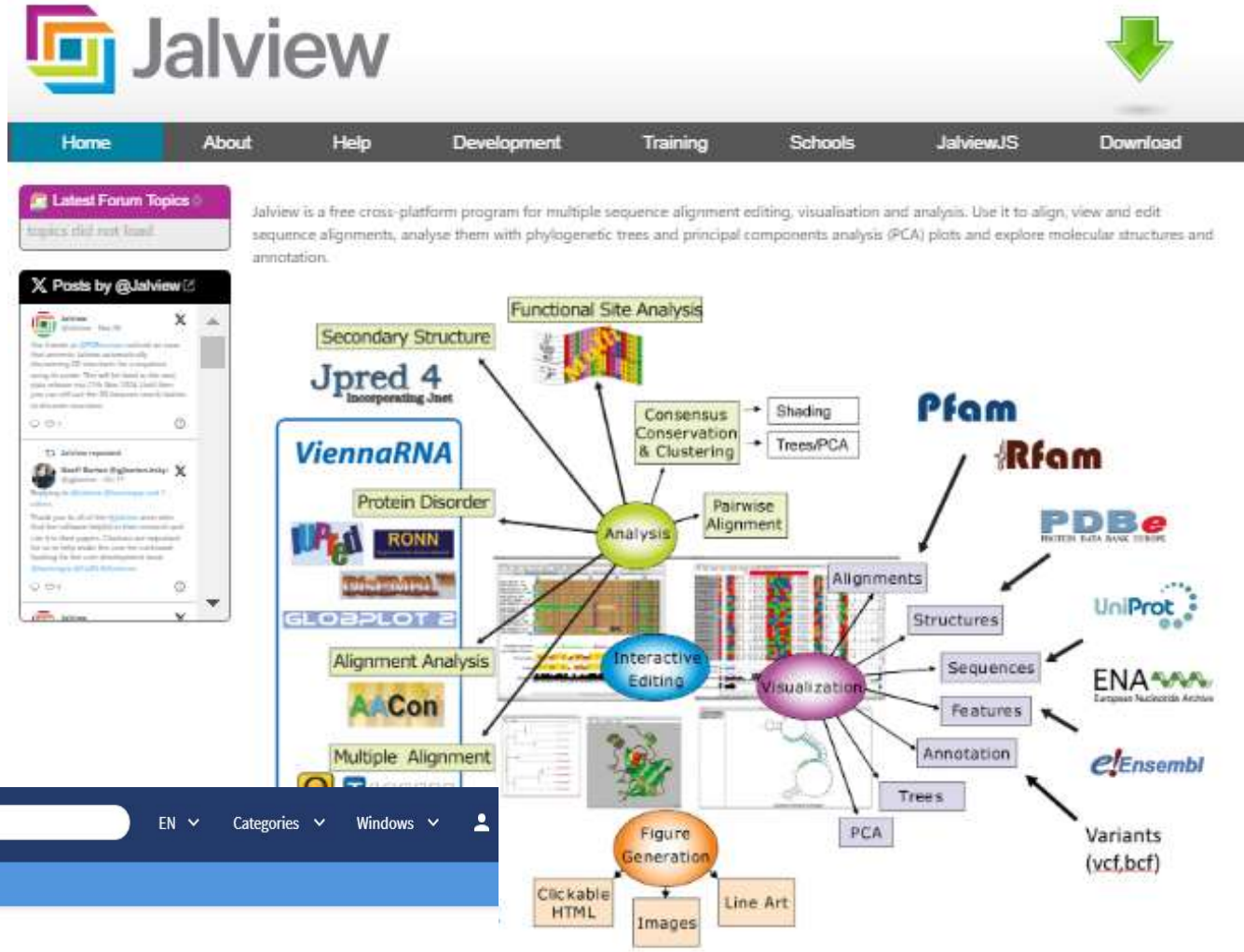
In the last session, we discussed:

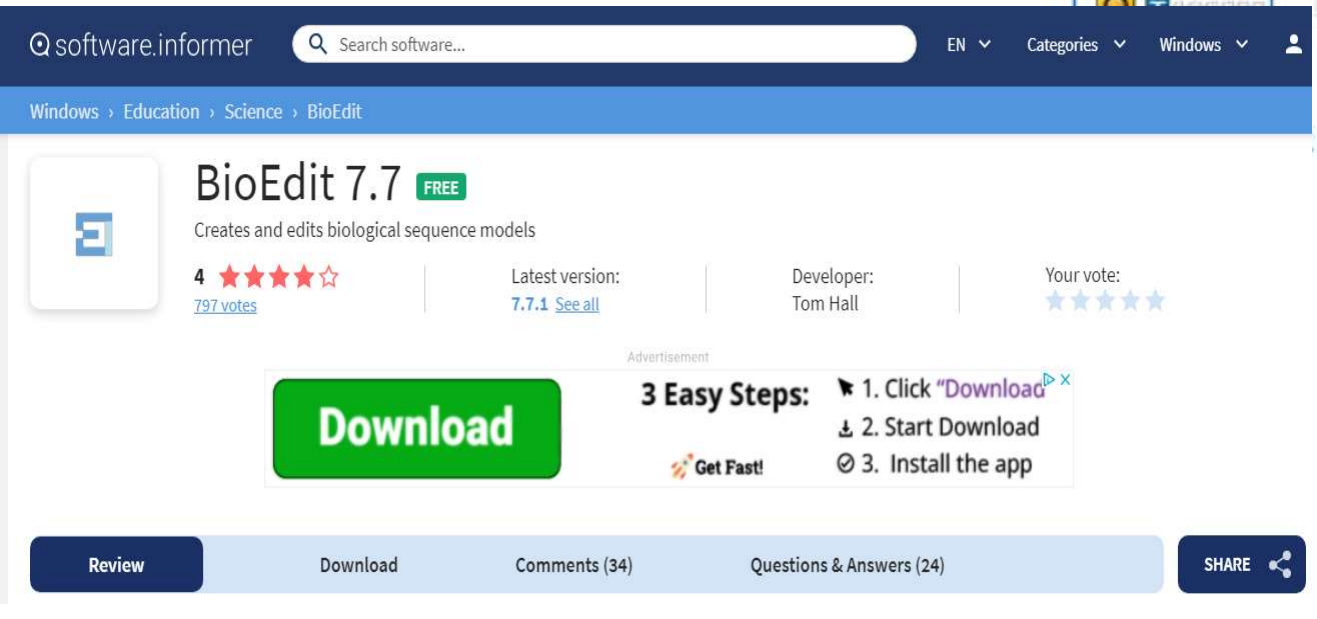
Proteins Multiple Alignments

- Discussed how to perform a multiple alignment.
- Multiple alignments are used to
  - Identify sequence positions where specific amino acids matter for the structural integrity or the function of a given protein
  - Define specific sequence signatures for protein families
  - Classify sequences and build evolutionary trees.

Also Introduced Software:

- ❑ Jalview
- ❑ BioEdit





Nucleotide Sequence Databases

- Nucleotide sequence databases are structured repositories that store and organize sequences of nucleotides, the building blocks of DNA and RNA.
- They facilitate the retrieval, analysis, and comparison of genetic information, serving as resources for genomics, molecular biology, and related fields.
- Their primary function is to store and provide access to nucleotide sequences, but their value is far beyond mere storage.
- Researchers use them for analyses, including sequence alignment and functional annotation.
- By providing access to vast collections of sequence data from various organisms, they support applications, including gene identification, evolutionary studies, and the development of biotechnological tools.

Importance in biological research and genomics

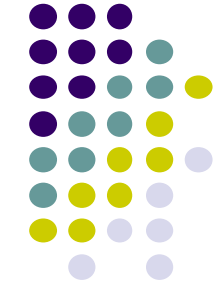
- These fields are pivotal in advancing our understanding of life at a molecular level.
- The details of inheritance, evolution, and disease mechanisms are uncovered through genomics.
- This knowledge is essential for developing targeted therapies, enhancing agricultural practices, and improving environmental conservation efforts.
- Genomic data has the potential for discoveries that can improve human health and biodiversity.

Types of Nucleotide Sequence Databases

- Primary Type
  - GenBank
  - EMBL Nucleotide Sequence Database
  - DDBJ (DNA Data Bank of Japan)

- Secondary Type
  - UniProt
  - SRA (Sequence Read Archive)
  - NCBI RefSeq

Primary vs. Secondary Nucleotide Sequences Databases



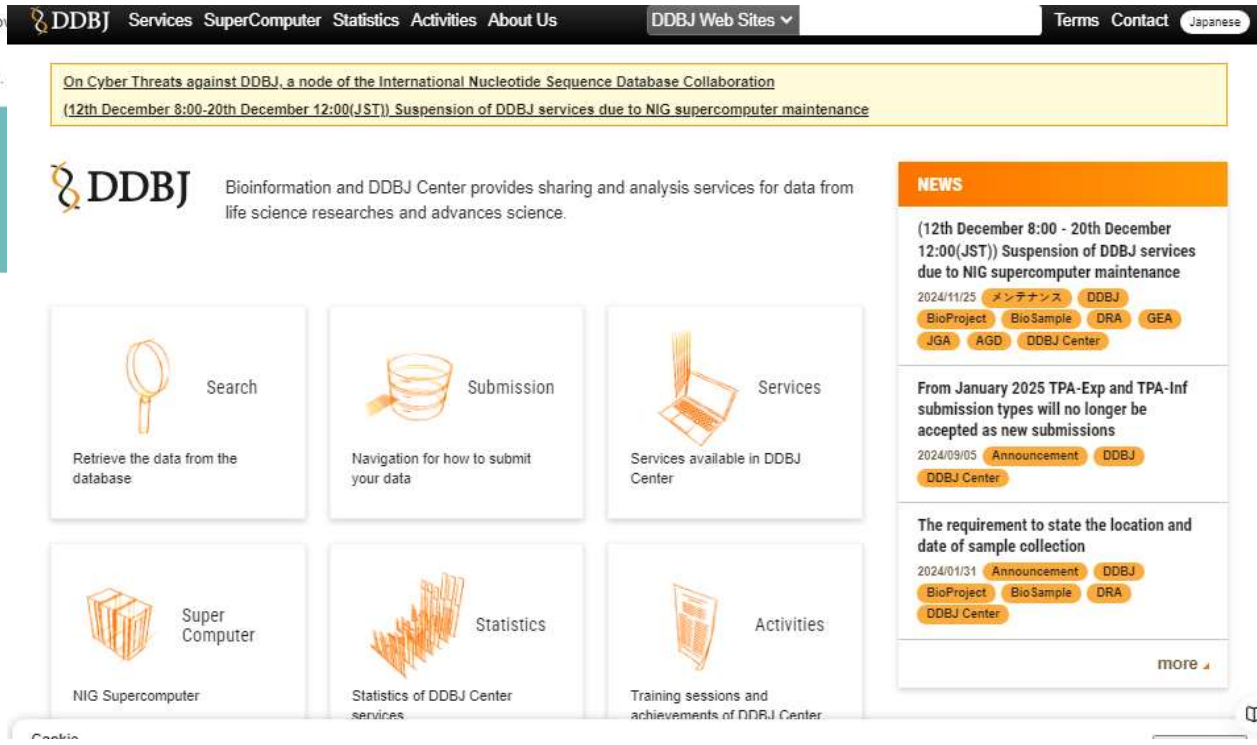
- Primary databases focus on raw sequence data
  - Comprehensive and publicly accessible
  - Allows sharing of sequence data worldwide

- Secondary ones organize information on gene functions, interactions, and evolutionary relationships.
  - Allow access to annotated sequences – let perform comparative analyses & gain insights into genetic variability, functional genomics & molecular biology.
  - Serve as vital tools for data integration and exploration,

European and Japanese Primary Databases:

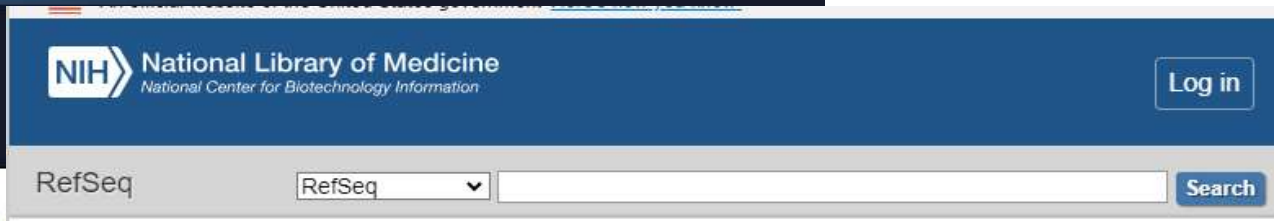
https://www.ebi.ac.uk/ena/browser/home

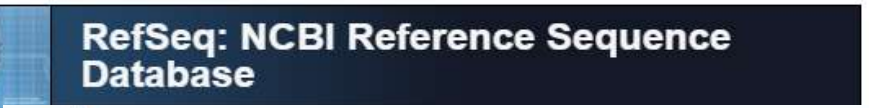
https://www.ddbj.nig.ac.jp/index-e.html



Getting Started

- Documentation
- How to submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA





Using RefSeq

- About RefSeq
- Human Reference Genome
- Prokaryotic RefSeq Genomes
- FAQ
- NCBI Handbook
- Factsheet

Databases Applications:

- Genomics and Genome Assembly:
  - These databases store and analyze genomic data, which is essential for understanding their structure, function, and evolution.
- Gene Expression Analysis:
  - One can identify patterns of gene expression to help understand how genes are regulated and respond to different conditions.
- Mutation Analysis:
  - Nucleotide sequence databases can be used to identify genetic mutations associated with diseases, which can help diagnose and treat genetic disorders.
- Phylogenetic Analysis:
  - By comparing sequences from different organisms, one can infer evolutionary relationships and reconstruct phylogenetic trees.

- Gene Function Prediction:
  - They can be used to predict the function of genes based on their sequence similarity to known genes.
- Drug Discovery:
  - By analyzing the sequences of disease genes, one can identify potential drug targets and design new therapies.
- Forensic Analysis:
  - They can be used in forensic science to analyze DNA evidence and identify individuals.
- Synthetic Biology:
  - They can be used to design and construct new biological pathways, circuits, and organisms.

- Cancer Research:
  - By analyzing the sequences of cancer genomes, it is possible to identify genetic mutations and develop new treatments.
- Personalized Medicine:
  - They can be used to analyze an individual's genetic data and provide personalized treatment recommendations.
- Infectious Disease Research:
  - By analyzing the sequences of pathogens, researchers can understand the evolution and transmission of infectious diseases.
- Agricultural Research:
  - Nucleotide sequence databases can be used to improve crop yields, disease resistance, and nutritional content.

- Environmental Monitoring:
  - By analyzing the sequences of environmental samples, researchers can monitor the presence and abundance of different species.
- Education and Training:
  - Nucleotide sequence databases can be used to teach students about genomics, bioinformatics, and molecular biology.

Challenges of Nucleotide Sequence Databases

- Data Quality and Accuracy:
  - Ensuring the accuracy and quality of the sequence data is a significant challenge, as errors can occur during sequencing, assembly, and annotation.
- Data Volume and Complexity:
  - The rapid growth of sequence data poses significant challenges for data storage, processing, and analysis.
- Data Integration and Interoperability:
  - Integrating data from different sources and formats can be difficult, making it challenging to compare and analyze sequences across different databases.

- Annotation and Curation:
  - Accurate annotation and curation of sequence data are time-consuming and require significant expertise.
- Database Maintenance and Updates:
  - Regular updates and maintenance of databases are needed to ensure that the data remains accurate and relevant.
- Data Security and Access Control:
  - Ensuring the security and integrity of sensitive sequence data, such as human genomic data, is essential.
- Intellectual Property and Ownership:
  - Resolving issues related to intellectual property and ownership of sequence data can be complex.

Nucleotide Sequence Databases Considerations:

- Ethical and Social Implications:
  - The use of nucleotide sequence databases raises ethical and social concerns, such as the potential for genetic discrimination or the misuse of genomic data.
- Privacy and Confidentiality:
  - Ensuring the privacy and confidentiality of individuals whose genomic data is stored in databases is crucial.
- Data Sharing and Collaboration:
  - Encouraging data sharing and collaboration among researchers while ensuring data security and integrity is essential.

- Standardization and Harmonization:
  - Standardizing and harmonizing data formats, annotation, and analysis methods is necessary to facilitate data comparison and integration.
- Education and Training:
  - Providing education & training for researchers, clinicians, and students on the use and interpretation is vital.
- Funding and Sustainability:
  - Ensuring long-term funding and sustainability is essential to maintain their availability and utility.
- Regulatory Frameworks:
  - Establishing regulatory frameworks to govern their use and ensure compliance with laws and guidelines.

Classes of Living

- Three most basic classes of living organism are
- Prokaryotes
  - Usually bacteria,
- Achaea
  - bacteria-like organisms living in extreme conditions,
- Eukaryotes
  - They go all the way from microscopic yeast to humans, animals, and plants

For Bioinformatics Purposes

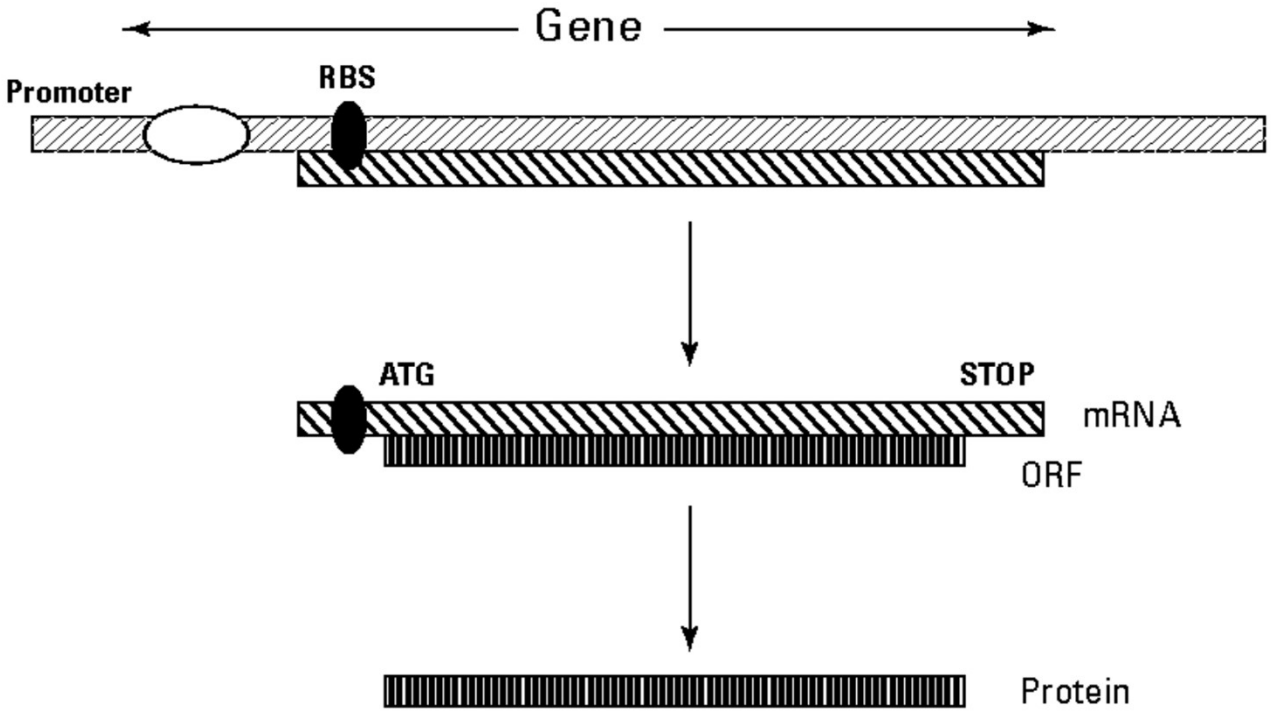
Prokaryotic & Achaea are very similar!

- They are microscopic organisms,
- Their genome is a single, circular DNA molecule,
- Their genome size is in order of a few million base pairs (0.6-8),
- Their gene density – the number of genes per base pair in the genome – is approximately one gene per 1000 base pairs.
- Their genome covers a few useless parts (%70 codes for proteins)
- Their genes do not overlap,
- Their genes are transcribed (copied into messenger RNA (mRNA)) right after a control region called a promoter,
- These mRNA are collinear with the genome sequence, means
  - Genes are in a single piece, not interrupted by noncoding patches (introns)
- Protein sequences derived by translating the longest open reading frame (from ATG to STOP) bridging the gene-transcript sequence

Relationship Between Gene, mRNA, ORF, and Protein:

- Simple Collinear Relationship between bacterial genomic (gene) sequence, the transcript (mRNA), the open reading frame (ORF), and the final protein.

* mRNA sequence gets translated into protein right after a special signal, called Ribosome Binding Site (RBS in the Figure).
* Ribosome is the main piece of machinery of the cell's protein-translation device.



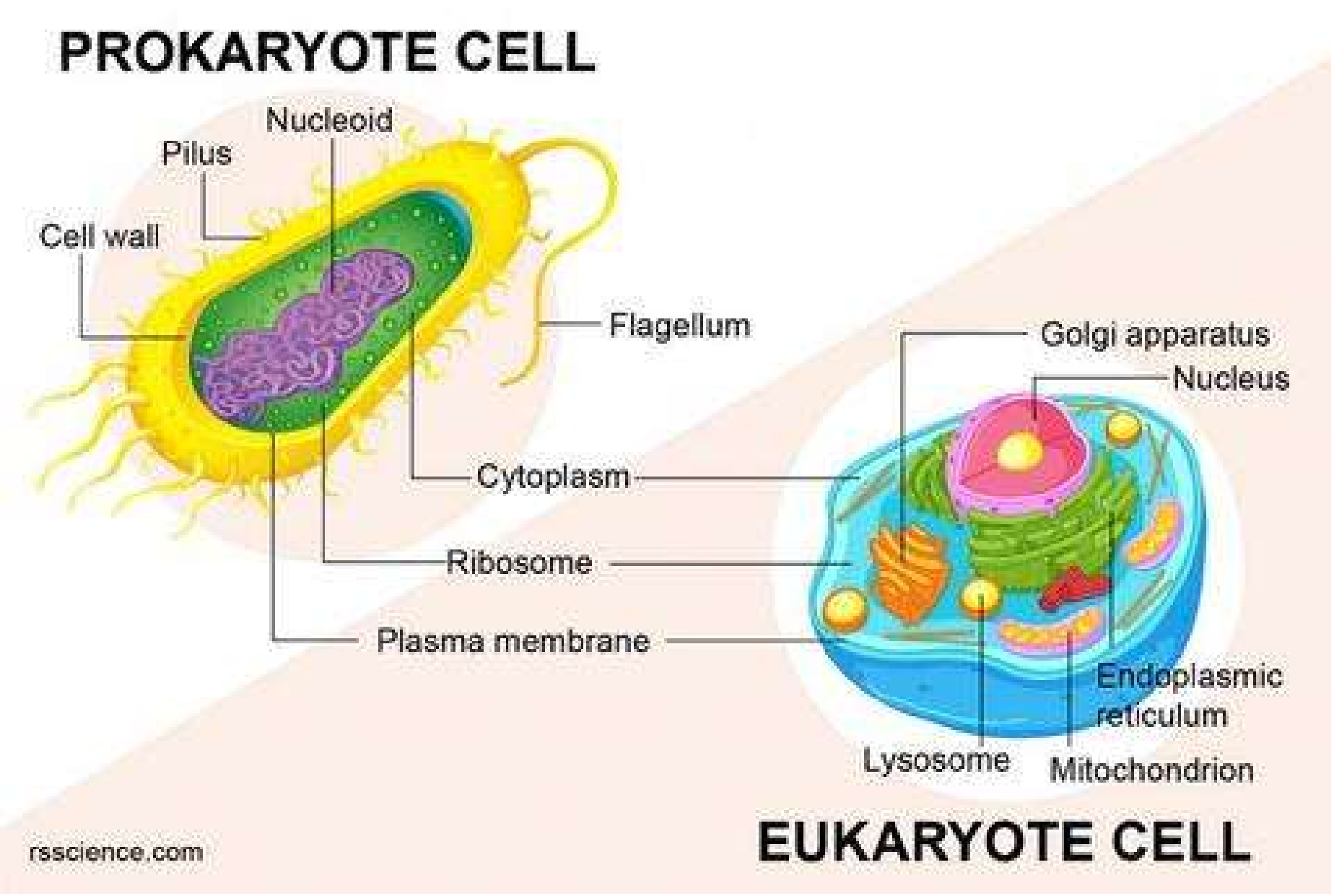
As a result of these relationship:

- Database entries describing a coding Prokaryotic sequence should include three features:
  - The Coordinates of some promoter elements,
  - The coordinates of the RBS,
  - The coordinates of the ORF boundaries,
- That is all we can say regarding the simple architecture of Prokaryotic genes.
- Not all genes encode proteins.
  - For some, the function is directly supported by transcribed RNA molecule, including transfer RNA (tRNA), ribosomal RNA (rRNA), and some other fancy RNAs!

Reading into Genes and Genomes: Prokaryotes and Eukaryotes

- Prokaryotes and Eukaryotes are two distinct types of cells that differ in their cellular structure, metabolism, and genetic organization.

- These differences are reflected in their genes and genomes, which have evolved to adapt to their respective environments and lifestyles.


PROKARYOTE CELL
Nucleoid
Pilus
Cell wall
Flagellum
Cytoplasm
Ribosome
Plasma membrane
Golgi apparatus
Nucleus
Endoplasmic reticulum
Lysosome
Mitochondrion
EUKARYOTE CELL
rsscience.com

Prokaryotes

- Lack of a true nucleus:
  - Prokaryotes do not have a membrane-bound nucleus, and their genetic material is found in a single circular chromosome.
- Small genome size:
  - Prokaryotic genomes are small, ranging from 0.5 to 5 million base pairs.
- High gene density:
  - Prokaryotic genomes have a high gene density, with many genes encoded in a small amount of DNA.
- Operons:
  - Prokaryotes often have operons, which are clusters of genes that are co-regulated and transcribed together.
- Horizontal gene transfer:
  - They can exchange genes with other prokaryotes via horizontal gene transfer, allows them to get new traits and adapt to changing environments.

Common Properties of Eukaryotes

- These properties makes them a bit more challenging for genomic and bioinformatics analysis,
- True nucleus:
  - Eukaryotes have a membrane-bound nucleus that contains their genetic material.
- Their genome consists of multiple linear pieces of DNA called chromosomes (up to a hindered million based pairs long)
- Large genome size:
  - Eukaryotic genomes are typically large, ranging from 100 million to 3 billion base pairs - is much bigger than for prokaryotes.
- Low gene density:
  - Eukaryotic genomes have a low gene density, with many non-coding regions and repetitive elements.

Eukaryotes properties continued …

- Their genome is no model of efficiency, containing many useless parts. * Les than %5 of human genome code for protein)
- Gene on opposite DNA strands might overlap
- Genes often exhibit more than one mRNA form (so as protein)
- Genes are transcribed right after a control region (promoter), but far away sequence parts have a strong influence on this process.
- Gene sequences are not collinear with the final mRNA and protein sequences. * Only small bits (the exons) are retained in the mature mRNA that encodes the final product, * Genes are often interrupted by introns, which are non-coding regions that are removed during RNA splicing.
- Gene regulation:
  - Eukaryotes have complex systems of gene regulation, including transcriptional and post-transcriptional control.

Prokaryotic & Eukaryotic: Key differences in gene structure

- Gene length:
  - Eukaryotic genes are generally longer than prokaryotic genes due to the presence of introns.
- Gene organization:
  - Prokaryotic genes are often organized into operons, while eukaryotic genes are typically transcribed individually.
- Promoters and enhancers:
  - Eukaryotic genes have complex promoters and enhancers that regulate gene expression, while prokaryotic genes have simpler promoters.

Prokaryotic & Eukaryotic: Key differences in genome structure

- Chromosome structure:
  - Prokaryotes have a single circular chromosome, while eukaryotes have multiple linear chromosomes.
- Genome size and complexity:
  - Eukaryotic genomes are generally larger and more complex than Prokaryotic genomes.
- Repetitive elements:
  - Eukaryotic genomes contain many repetitive elements, such as transposons and retrotransposons, which are rare in prokaryotes.

Evolutionary implications

- Genome evolution:
  - The differences in genome organization and gene structure between prokaryotes and eukaryotes reflect their distinct evolutionary histories.
- Gene transfer:
  - The ability of prokaryotes to exchange genes through horizontal gene transfer has played a key role in their evolution and adaptation to changing environments.
- Cellular complexity:
  - The development of complex cellular structures and systems in eukaryotes is accompanied by the evolution of more complex genomes and gene regulatory systems.

As a result:

- Genes of higher Eukaryotes (animals) may span up of millions of base pairs
  - For example the human dystrophin gene (its mutation causes a dreadful disease), is 2.2 million base pairs long.
- The relationship among a gene DNA sequence, its primary transcript, the various forms of mature mRNA, and the final protein sequence can be very complex.
- A lot of database entries corresponding to partial gene sequences, or different gene-related objects:
  - such as promoter regions, mRNA, or genome fragments,
- These make their studies challenging!

Thank you

for your attention