Bioinformatics

Sequence analysis

Assis. Prof. Dr. Labeed Al - Saad

The objectives

- What is a sequence?

- What is sequence processing?

- What are the main lines of sequence analysis?

- What are the key steps of sequence analysis?

What is a sequence?

A sequence is an ordered collection of items. In simpler terms, it's a list of things where order matters. Here are some key aspects of sequences:

▪ Order matters: The order in which items appear in the sequence is important. For instance, the sequence "1, 2, 3" is different from "3, 2, 1".

▪ Repetition allowed: Elements can repeat within a sequence. The sequence "apple, banana, apple" is perfectly valid.

▪ Finite or infinite: A sequence can have a fixed number of elements (finite) or continue indefinitely (infinite). Examples of finite sequences include shopping lists or steps in a recipe. The sequence of natural numbers (1, 2, 3, ...) is an infinite sequence.

What is a sequence?

Sequences are fundamental concepts in various fields, including:

- **Mathematics:** They are used to study series, calculus, and other areas.

- **Computer Science:** They are used in algorithms, data structures, and text processing.

- **Biology:** DNA, RNA, and protein sequences are crucial for understanding biological processes.

- **Music:** The order of notes forms a musical sequence.

What is a sequence?

There are different ways to represent a sequence depending on the context:

- Listing all elements explicitly (e.g., 1, 2, 3, ...).

- Using a mathematical formula to define how each element is generated (e.g., for the sequence 2, 4, 6, ..., the rule could be "next number is current number + 2").

What is a sequence?

··· GTGCATCTGACTCCTGAGGAGAAG ···   DNA

··· CACGTAGACTGAGGACTCCTCTTC ···

⬇ Transcription

··· GUGCAUCUGACUCCUGAGGAGAAG ··· RNA

⬇ ⬇ ⬇ ⬇ ⬇ ⬇ ⬇ ⬇

Translation

··· V   H   L   T   P   E   E   K ··· Protein

What is sequence processing?

Sequence processing refers to the broad field of techniques used to analyze ordered lists of data.

This data can be anything from DNA or protein sequences in biology to text strings in computer science. The key idea is that the order of the elements in the sequence matters and can hold significant meaning.

What is sequence processing?

Here's a breakdown of the core aspects of sequence processing:

- Data types: Sequence processing can handle various data types, including:

  - ❖ Biological sequences: DNA, RNA, proteins.

  - ❖ Text data: Sentences, paragraphs, code.

  - ❖ Time series data: Stock prices, sensor readings.

  - ❖ Any data where the order is important.

What is sequence processing?

- Goals: The goals of sequence processing can vary depending on the data type and field. Here are some common examples:

  ❖ Finding patterns: Identify recurring motifs or structures within a sequence. (e.g., finding specific genes in DNA).

  ❖ Classification: Categorize a sequence based on its content. (e.g., classifying spam emails).

  ❖ Prediction: Forecast the next element in a sequence. (e.g., predicting the next word in a sentence).

  ❖ Alignment: Compare multiple sequences to identify similarities and differences. (e.g., comparing DNA sequences across species).

  ❖ Assembly: Reconstruct a larger sequence from smaller fragments. (e.g., assembling a genome from short DNA reads).

What is sequence processing?

- **Techniques:** A wide range of techniques are used in sequence processing, including:

  - ❖ **Dynamic programming:** Efficiently solving problems by breaking them down into smaller subproblems. (e.g., sequence alignment).

  - ❖ **Hidden Markov Models (HMMs):** Statistical models for representing sequences with hidden states. (e.g., gene prediction).

  - ❖ **Machine learning:** Algorithms that learn from data to perform tasks like classification or prediction. (e.g., sentiment analysis of text).

  - ❖ **Deep learning:** A subfield of machine learning using complex neural networks for tasks like motif discovery or protein structure prediction.
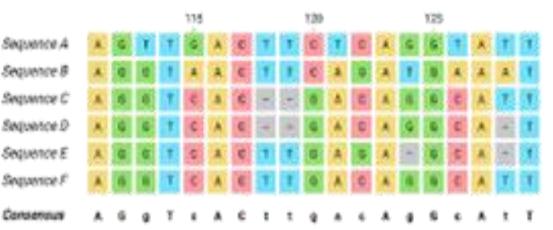
What is sequence processing?

Sequence processing is a fundamental concept in various fields, and the specific techniques used will depend on the specific application.

What are the main lines of sequence analysis?

The main lines of sequence analysis:

■ Sequence alignment: This is the process of arranging two or more sequences of DNA, RNA, or protein to identify regions of similarity. These similarities may be due to functional, structural, or evolutionary relationships between the sequences

Multiple Sequence Alignment

https://assets-global.website-files.com/621e95f9ac30687a56e4297e/64a8d74bcf720ca09f0f023c_V2_1676212488941_acecc7dd-4fb7-40ba-b8b9-968ab7adaeb7.png

What are the main lines of sequence analysis?

- **Assembly:** This is the process of reconstructing a long sequence from shorter overlapping fragments. This is commonly used in genome sequencing, where the genome is sequenced in small pieces and then assembled into a complete sequence.

- **Annotation:** This is the process of identifying and labeling the features of a sequence, such as genes, regulatory elements, and repetitive elements.

What are the main lines of sequence analysis?

- **Variant calling:** This is the process of identifying differences between two or more sequences. This is commonly used in cancer research to identify mutations in cancer cells.

- **Motif finding:** This is the process of identifying short, recurring sequences that may have a functional or regulatory role.

- **Phylogenetic analysis:** This is the process of reconstructing the evolutionary relationships between different species. This is often done by comparing the sequences of genes that are common to all of the species being studied.

What are the key steps of sequence analysis?

The key steps of sequence analysis can be broken down into three main stages:

- Data preparation:

  - ❖ Data Acquisition: Obtain raw sequence data from databases, experiments, or sequencing technologies.

  - ❖ Data Preprocessing: Clean and preprocess the data by removing noise, errors, or irrelevant information.

  - ❖ Assembly for genomes would be another crucial step here, where short sequence reads are stitched together into a larger, contiguous sequence

What are the key steps of sequence analysis?

- Analysis:

  ❖ This is the core stage where you extract information from your sequences.

  ❖ Depending on your goal, different techniques are employed:

    ➢ Alignment: Comparing your sequence to existing databases or other sequences of interest to identify similarities and differences.

    ➢ Motif finding: Searching for short, recurring patterns within the sequence that might have a specific function.

What are the key steps of sequence analysis?

- Interpretation:

  ❖ This final stage involves making sense of the results from the analysis stage.

  ❖ You might use your findings to:

    ➤ Identify genes or proteins within your sequence.

    ➤ Understand the function or structure of a molecule.

    ➤ Explore evolutionary relationships between different organisms.

  ❖ This stage often involves referencing biological databases and prior research to put your results in context.