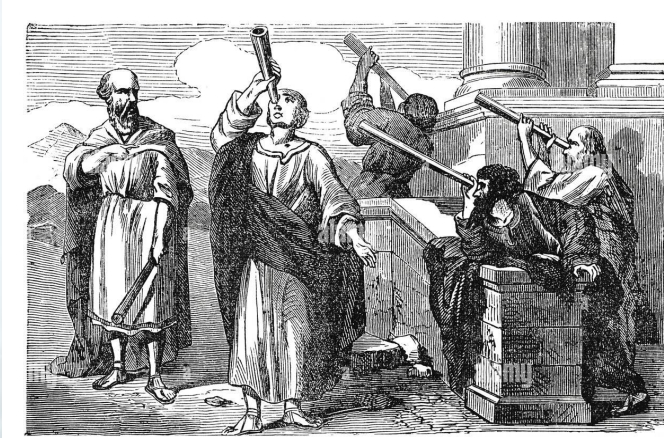# Biological Big Data Mining

Data collection and data analysis were actually taking place from ancient time, even if they were in a primitive form.


Many ancient human civilizations had gained important knowledge by observing the planets and the stars. By analyzing these observations they were able to accurately predict the time of the seasonal changes over a year.
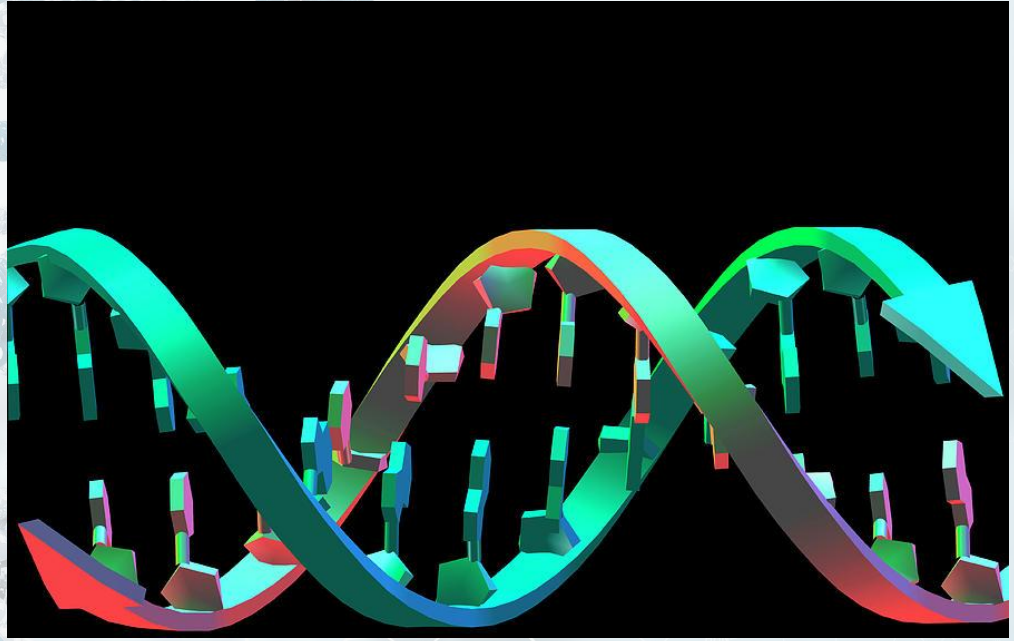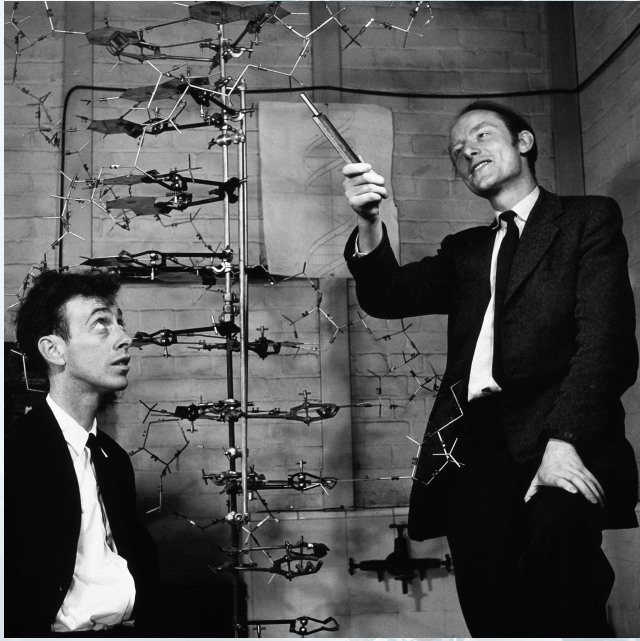
In the 20th century the important discoveries concerning DNA, such as the clarification of the correct double-helix model of DNA structure (Watson & Crick, 1953) established molecular biology as one of the most important research fields of biology. These discoveries attracted much attention and changed the direction of research in biology, as well as in medicine.

The central dogma of molecular biology, as coined and re-stated by Francis Crick (1958; 1970), describes the flow of the biological information
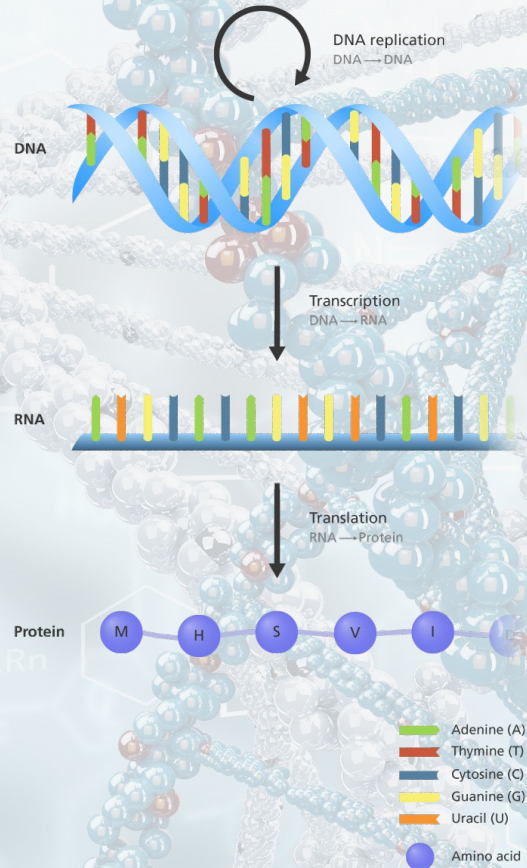

The general transfers that take place in most organisms are described by the filled arrows. In particular, DNA is transcribed into RNA that is finally translated into protein. The circular arrow around DNA denotes its replicability.

Data mining emerged in order to cope with the challenges that traditional data analysis techniques where facing up when dealing with large amounts of often peculiar data.
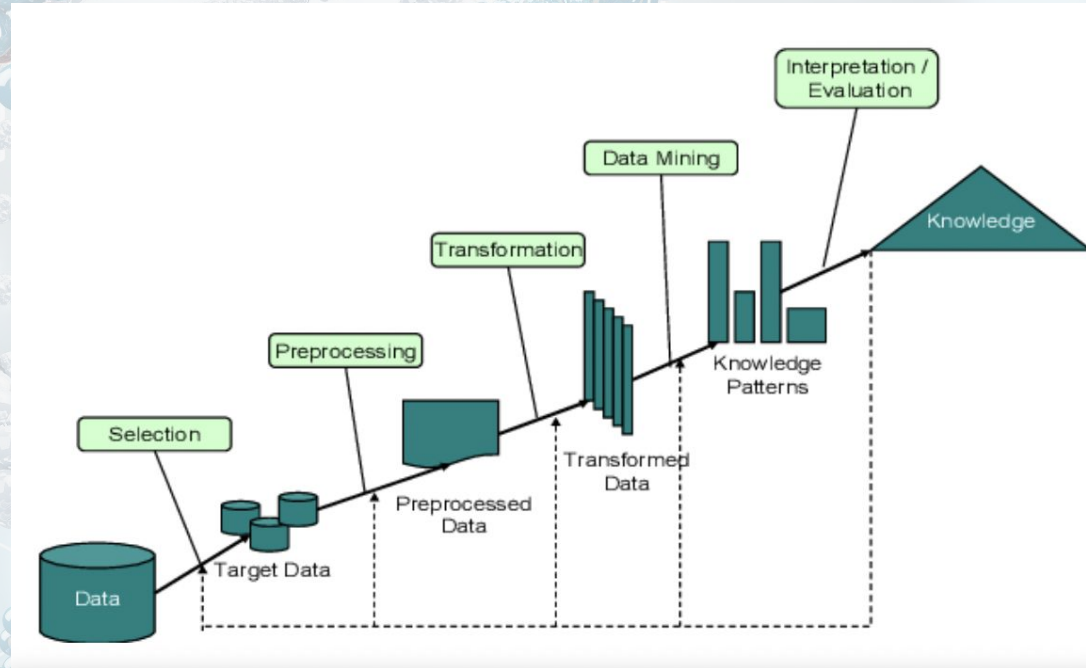

The pre-processing phase usually includes the selection of an appropriate portion of data, the cleaning of the selected data, as well as the transformation of data. The post-processing phase deals with the management of the produced patterns and models and focuses on the evaluation and interpretation of data mining results.

The most common data mining tasks are classification, regression, clustering, and association rules mining

a. Data objects

b. Clustered data objects
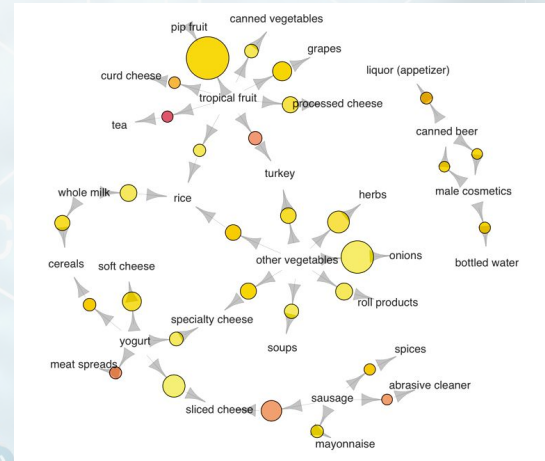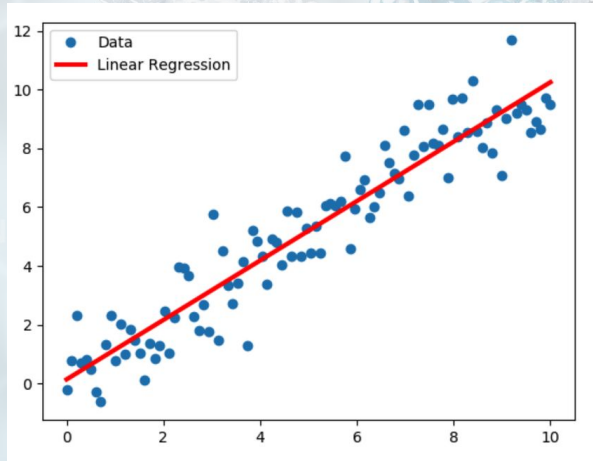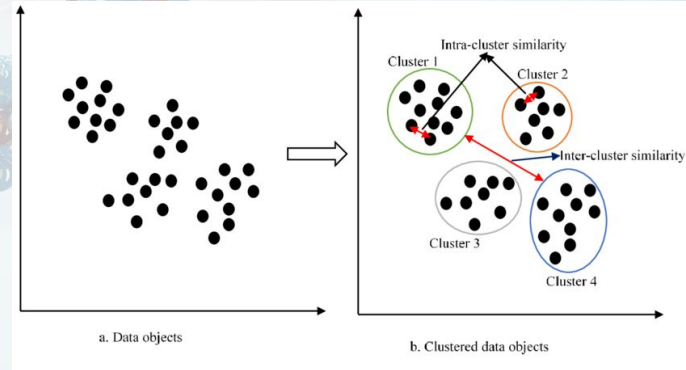
In recent years, rapid developments in genomics and proteomics have generated a large amount of biological data. Drawing conclusions from these data requires sophisticated computational analyses.
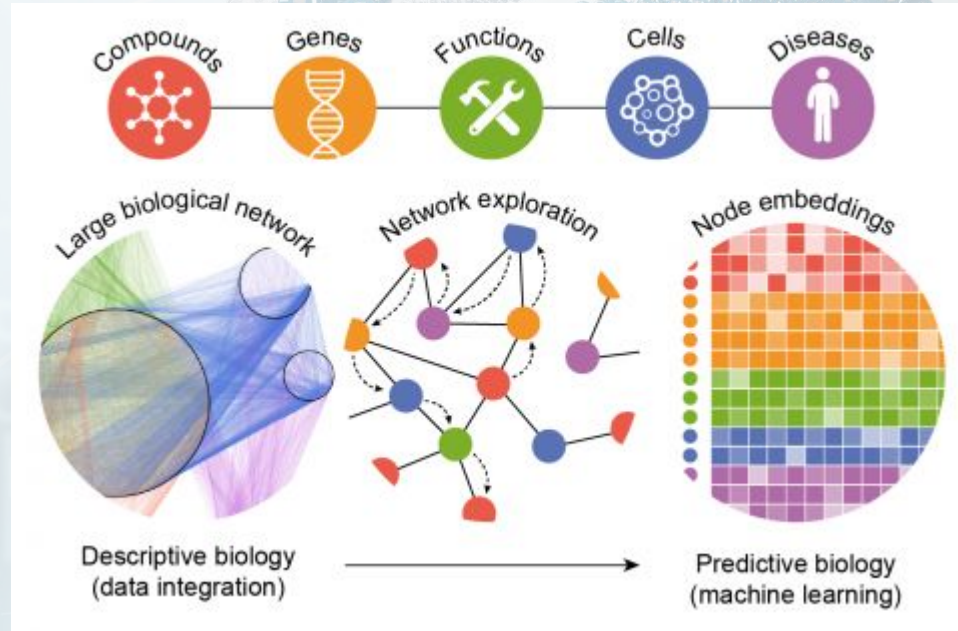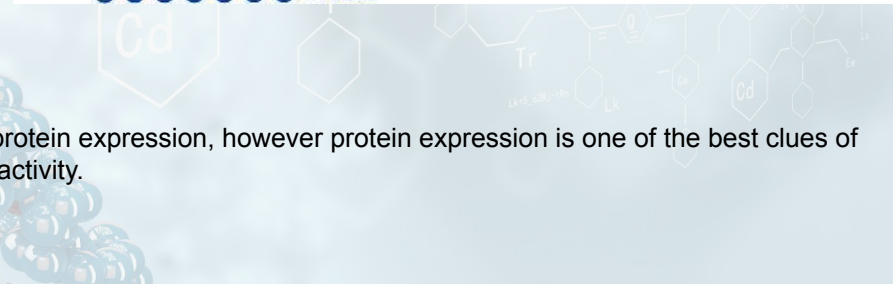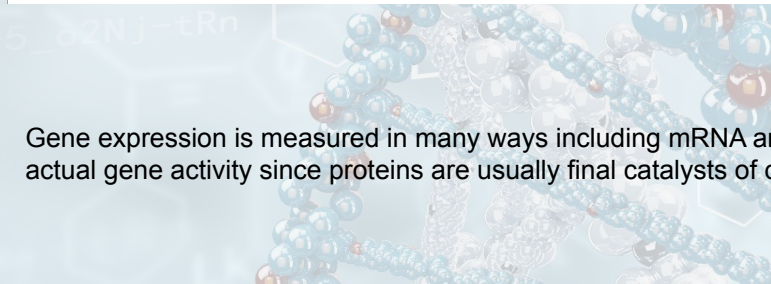

Bioinformatics can be defined as the application of computer technology to the management of biological information. Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules.

Sequence Analysis

Position 321: A → G
A: Adenine
G: Guanine
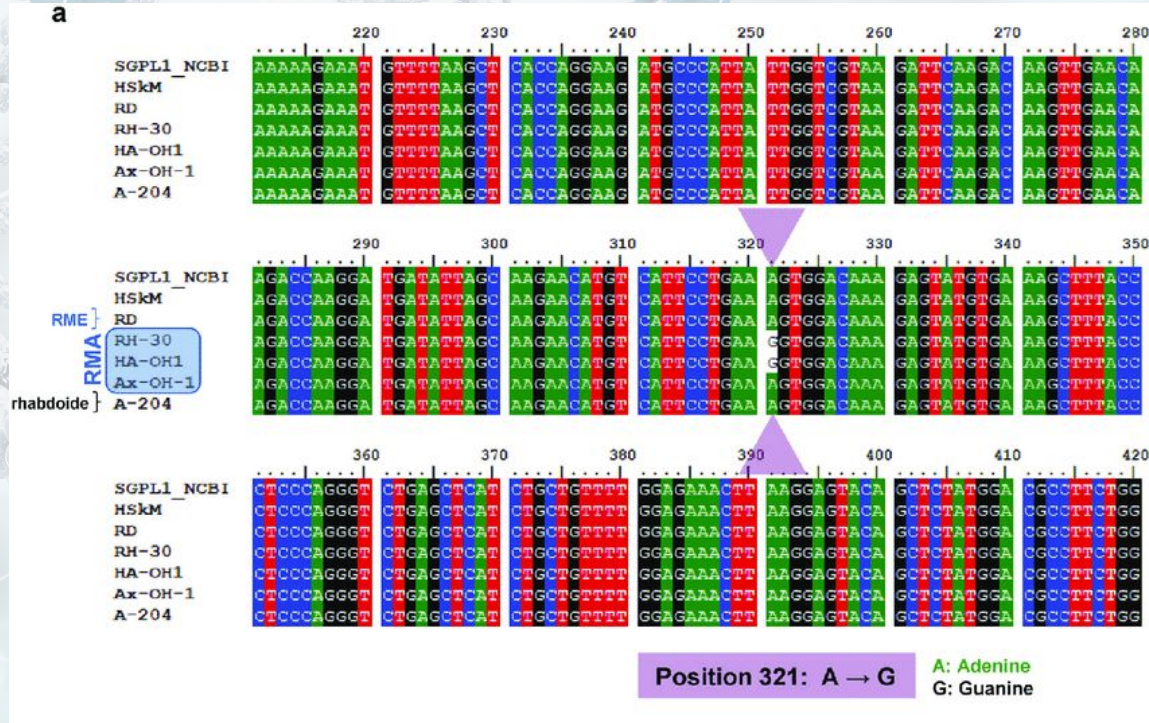
Sequence analysis is the most primitive operation in computational biology. This operation consists of finding which part of the biological sequences are alike and which part differs during medical analysis and genome mapping processes.

Analysis of Gene - Protein expression

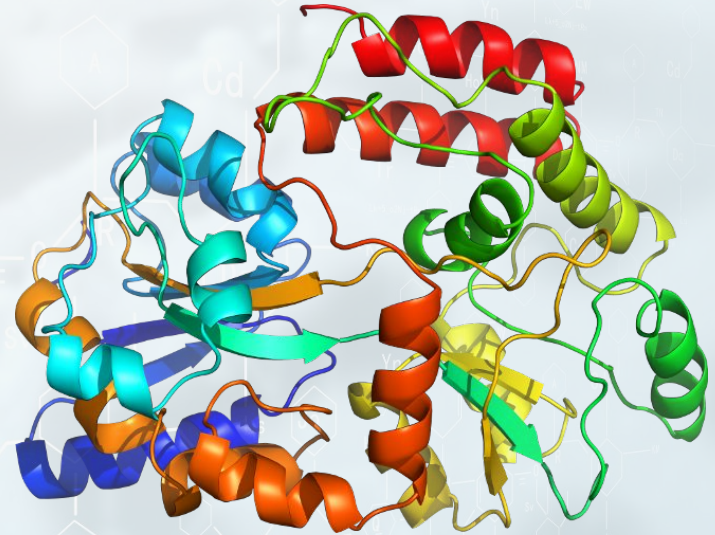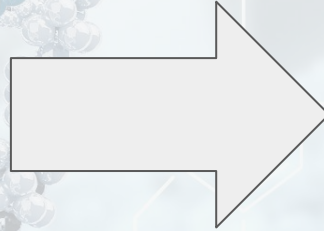Gene expression is measured in many ways including mRNA and protein expression, however protein expression is one of the best clues of actual gene activity since proteins are usually final catalysts of cell activity.

Protein structure prediction

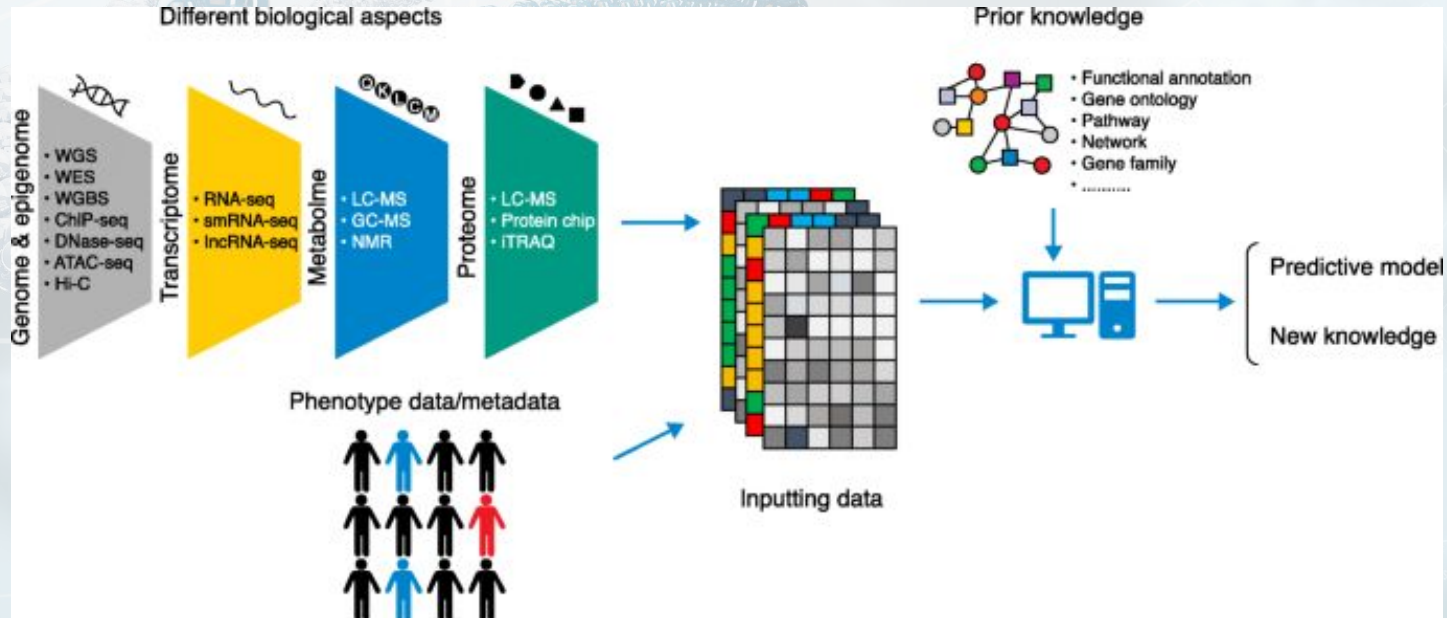β-globin gene

...G T G C T G G C C C A T...

...G T G C C G G C C C A T...

point mutation

The amino acid sequence of a protein (so-called, primary structure) can be easily determined from the sequence on the gene that codes for it. In most of the cases, this primary structure uniquely determines a structure in its native environment.

High-throughput Image Analysis

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical images.

Challenges of Data Mining

- Mining various and new kinds of knowledge.

- Mining knowledge in multidimensional space.

- Handling uncertainty, noise or incompleteness of data.

- Presentation and visualization of data mining results.

- Efficiency and scalability of data mining algorithms.

- Cloud computing and cluster computing.

Bibliography

- Khaliz Raza: Application of data mining in bioinformatics,

- Yixue Li, Luonan Chen : Big Biological Data: Challenges and Opportunities,

- George Tzanis : Biological and Medical Big Data Mining,

- Wikipedia

Thank You

Questions?