Bioinformatic resources

Giacomo Fantoni
telegram: @GiacomoFantoni
Ilaria Cherchi
telegram: @ilariacherchi
Github: https://github.com/giacThePhantom/thesis_notes

June 25, 2022

Contents

Chapter 1

Introduction

1.1 Bioinformatics

Bioinformatics can be defined as the development of new algorithms and statistical methods that allow to establish relations between members of huge sets of data. It can also be described as the analysis and interpretation of different data types including nucleotide and amino acid sequences, protein domains and protein structures. The development and implementation of these programs allow efficient access and management of different types of information.

1.1.1 The human genome

The human genome contains around 3.2 billion base pairs. About 80% of it is associated with a biochemical function. Of particular interest is non-coding DNA, which doesn't code for proteins but is mainly involved in:

- Protection of the genome.
- Gene switches.
- Gene expression regulation.
- Transcription factor binding sites.
- Operators.
- Enhancers.
- Promoters.
- Silencers.

1.2 Involvement of computer science

Computer science plays a fundamental part in bioinformatics, providing the algorithms necessary to exploit data collected in experiment to reach a significant conclusion.

1.2.1 Databases

Databases or data banks are collections of correlated data utilized to represent a portion of the real world. They are structured in a way to allow data organization and management in terms of:

- Insertion.
- Update.
- Search.
- Deletion.

1.2.2 Program

A program codifies an algorithm into a programming language. It is used to test and realize a proposed solution. Computer science can be defined as the science of the automatic elaboration of information, with algorithm as its central focus.

1.2.3 Algorithm

Algorithms (from the name of the Persian mathematician *Muhammad ibn Musa al-Khwarizmi*) can be defined as a system of well-defined rules and procedures that lead to the solution of a problem with a finite number of steps. They can be described in pseudo-code.

1.2.3.1 Substring search algorithm

A substring search algorithm is an algorithm that searches the occurrence of a string in another, allowing to understand if the former is contained in the latter. A naïve implementation is described in figure 1.1.

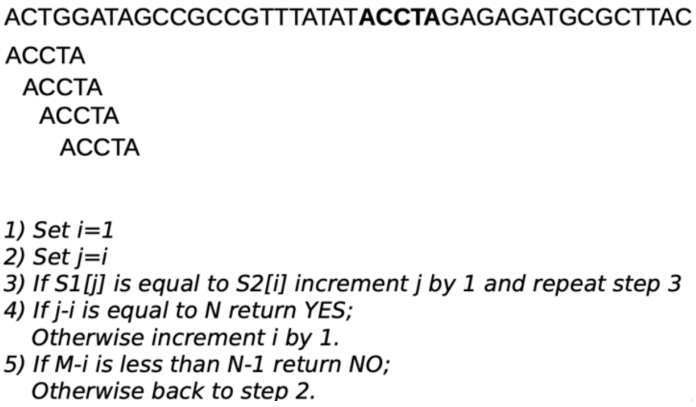

Figure 1.1: Naïve implementation of a substring search algorithm

Chapter 2

Scientific literature

2.1 Literature sources

All bioinformatics works are based on literature. Different sources of literature can be found.

2.1.1 Primary literature

Primary literature is defined as original materials. It is authored by researchers, contains original research data and is usually published in a peer-reviewed journal. Primary literature works can be:

- Journal articles or conference proceedings, which are usually the first formal appearance of a result.
- Original articles: the original research conducted by the authors, including methods and resources used.
- Letters or communications: short reports of original research focused on an outstanding finding whose importance means that it will be of interest to scientists in other fields.

2.1.2 Secondary literature

Secondary literature is the summary or review of the theories and results of original scientific research. Secondary literature works can be:

- Open letters.
- Comments.
- News.
- Reviews.
- Correspondence.
- Opinions.
- Protocols.

2.2 Structure of a scientific article

Scientific articles tend to have a well defined structure, composed, in order, of:

1. Title.
2. Abstract.
3. Keywords.
4. Introduction or background.
5. Methods or experiments.
6. Results or analysis.

7. Discussion.
8. Conclusion.
9. References or bibliography.
10. Figures and tables.
11. Supplementary materials.

2.3 Impact measures

An impact measure is used to define the goodness of a research or if it had a big impact in the community.

2.3.1 Impact of a journal

A measure of impact of a journal measure the impact of the publication of a journal. It can be measured in different ways:

- Impact factor (IF): a measure that reflects the average number of citations of articles published in a science journal. It can be biased due to self-citations, journal-forced citations and it does not take into account negative citations. It is computed as:

$$IF_y = \frac{Citations_y}{Publications_{y-1} + Publications_{y-2}}$$

- Journal of Citation reports JCR.
- Scimago Journal Rank SJR.

2.3.2 Personal impact

The personal impact measure the impact of a researcher. I can be measured as:

- H-index: an index that attempts to measure both the productivity and the impact of the published work of a scientist or scholar. A scholar with an index of h has published h papers each of which has been cited by others at least h times. It serves as an alternative to more traditional journal impact factor metrics in the evaluation of the impact of the work of a particular researcher.
- Web of Science WOS.
- Scoups.
- Google Scholar.

2.3.3 Peer review

Peer-reviewed articles are also called refereed articles. Peer review allows to:

- Independently verify theories and assumptions.
- To screen for the works ethic.
- Asses appropriateness for publication.
- Check for transparency of research.
- Assess the quality of the research.

Depending on the journal or publisher this process can takes from weeks to months.

Chapter 3

Biological databases

3.1 Introduction

3.1.1 Classification of databases

A huge number of biological databases are available and they can be distinguished as:

- Primary databases containing sequences of nucleotides and amino acids.
- Derived and specialized databases containing protein domains and motifs, protein structures, genes, transcripts, expression profiles, variations, pathways and many other informations.

Each database is characterized by a central biological element which constitutes the object around which the principal entry of the database is constructed.

3.1.2 Data sources

Data in these databases is derived from:

- Literature.
- In-vitro and in-vivo analysis.
- In-silico analysis.

3.1.3 Nomenclature

One of tha main problems related with biological database is nomenclature. There can be different name for the same gene or different genes with the same name. To uniquely identify genes and proteins and manage the large amount of information related, primary data banks assign an accession number to each element they store.

3.1.4 Reference genome

A reference genome is a digital sequence of nucleic acids assembled to be a representative sequence for a given species. It is assembled from DNA sequencing of a set of donors. An example of reference genome is *GRCh38* from which *hg38* is derived aggregating many donor informations.

3.1.5 Popular databases

3.1.6 GenBank

GenBank contains nucleotide sequences. The aim of the database is to store and archive historically important but redundant nucleotide sequences. Data can be submitted singularly or in a batch manner.

3.1.7 RefSeq

RefSeq is a curated and non redundant collection of DNA, RNA and protein sequences. Each RefSeq entry represents a sings molecule in a particular organism. Its basis is compiled with a process of collaboration, extraction and computation from GenBank. Each molecule is annotated reporting the name of the organism, the correct gene symbol for that organism and informative names of proteins when possible.

3.1.8 UniProt

UniProt is a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. It provides protein sequences, domains and structural information like subcellular location for many species. It also includes some alignment and mapping tools.

3.1.9 Others

Other examples of derived databases are:

- dbGap.
- Structure.
- Gene.
- Biosystems.

3.1.10 Genome browser

A genome browser is a database containing reference sequence assemblies for one or more genomes. It allows to browse data at various detail levels, from chromosome to gene, down to a single exon or intron. It also allows for the comparison between species and data extraction.

3.1.10.1 UCSC genome browser

The UCSC genome browser contains the genome of about 100 species, but it does not provide a browser for all of them. It integrates informations like SNPs, sequence conservation, regulatory elements (ENCODE) and others.

3.1.10.1.1 UCSC table browser The UCSC table browser allow to extract data from the database tables without the need for a graphical interface. It can also align sequences, annotate SNPs and convert data between genome versions. It is a flexible tool that can retrieve data for one or more genes in a variety of formats. When submitting heavy task it will redirect them to Galaxy, an online workflow system.

3.1.10.2 Ensembl genome browser

The Ensembl genome browser is the European genome browser. It focuses on vertebrate genomes. It includes genomic variants, both somatic and structural, and regulatory elements data. It offers an interfcace to access data directly BioMart. All Ensemble transcripts are based on proteins and mRNAs contained in the databases:

- UniProt/Swiss-Prot (manually curated).
- UniProt/TrEMBL (not reviewed).
- NCBI RefSeq (manually curated).

3.1.10.2.1 Biomart Biomart is a data mining platform which is able to address complex queries on ENSEMBL. It is similar to the USCS table browser, while being more powerful as it can retrieve both annotation and sequences.

Chapter 4

Motif analysis

Chapter 5

Expression analysis

Chapter 6

Gene set enrichment

Chapter 7

Network analysis

Chapter 8

Genome wide association studies