Bioinformatic resources

Giacomo Fantoni

telegram: @GiacomoFantoni

Ilaria Cherchi

telegram: @ilariacherchi

Github: https://github.com/giacThePhantom/thesis_notes

June 24, 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.


```
ACTGGATAGCCGCCGTTTATATACCTAGAGAGATGCGCTTAC
ACCTA
 ACCTA
  ACCTA
   ACCTA


1) Set i=1
2) Set j=i
3) If S1[j] is equal to S2[i] increment j by 1 and repeat step 3
4) If j-i is equal to N return YES;
   Otherwise increment i by 1.
5) If M-i is less than N-1 return NO;
   Otherwise back to step 2.
```

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

Chapter 4

Motif analysis

Chapter 5

Expression analysis

Chapter 6

Gene set enrichment

Chapter 7

Network analysis

Chapter 8

Genome wide association studies