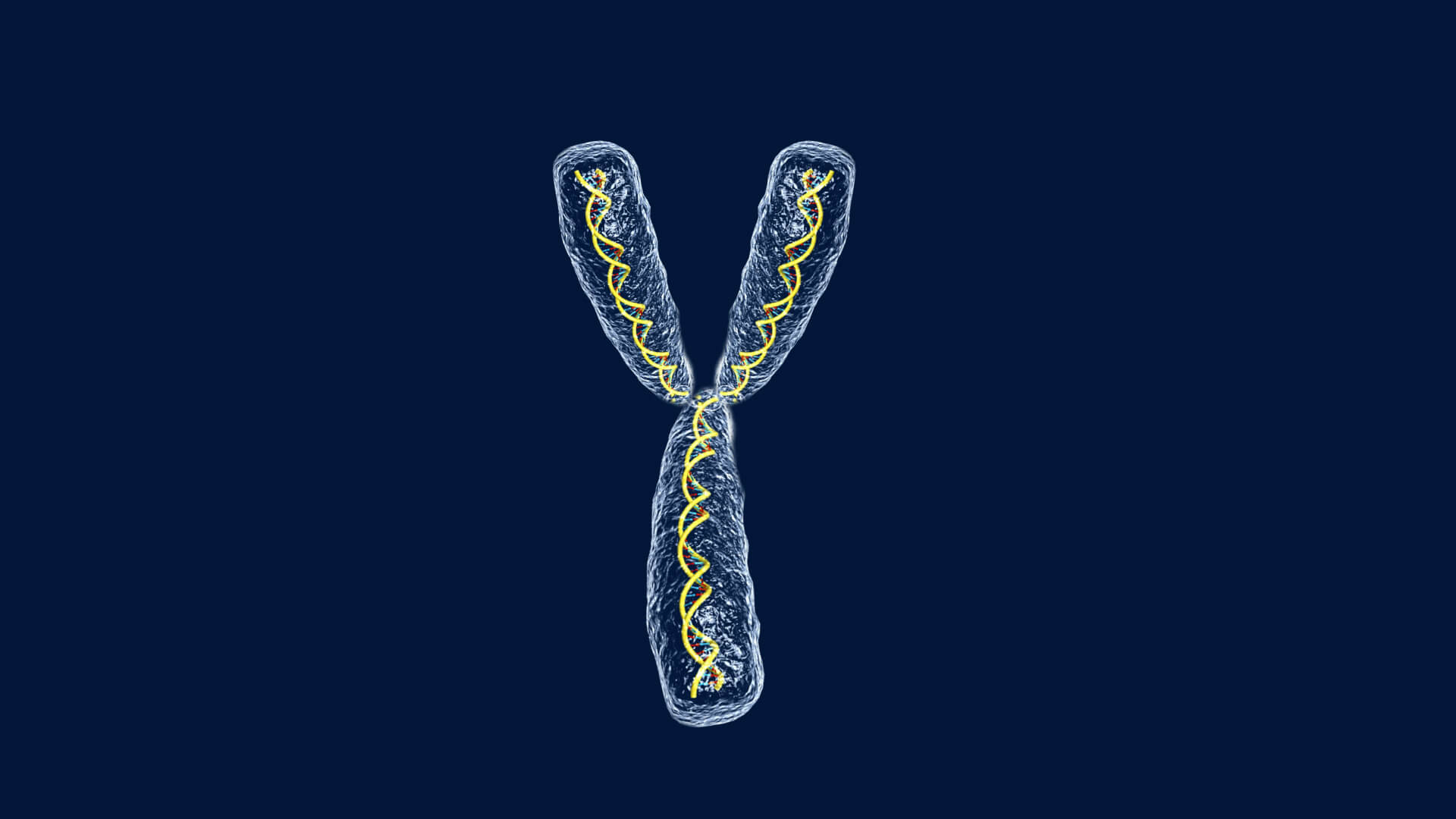
Bioinformatics 2018-2019

Date: 04 December 2018

Project report

|  |
| --- |
| Bioinformatics@Data Science A.Y. 2018-2019  Manuscript Title  Silvia Basile1, Tommaso Lanciano1  1Group no. 11 Abstract Max 150 words. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. |

1. Basic introduction about the disease/process

The phenomenon we will take into account in this work is the *Y chromosome*. Along with the X chromosome, they are responsible of sex determination of an offspring in all mammals, including humans. Most of the human cells have a couple of sexual chromosomes: XX for females, XY for males. Hence, the Y chromosome is transmitted only from a father to a son and it is composed of more than 59 million base pairs, representing approximately the 2% of the total DNA in a male cell. It is a third of the X chromosome, with almost 200 genes and at least 72 of them code for proteins and some of them are called Y-linked genes because they are present only in the Y chromosome. We can see the shape of the Y chromosome in Figure 1.1. Actually the name “Y chromosome” is not related to the shape of this particular chromosome since every chromosome normally appear as amorphous blob under the microscope and only after mitosis they reach a defined shape.

**Figure 1.1** Y chromosome

1. Seed genes

As we already said, the Y chromosome has about 200 genes, most of them are located in the MSY (male specific portion of Y), a region quite in the middle of the chromosome. We will consider 45 initial genes (*seed genes*) and we will refer to them with the official gene symbols taken from HGNC (HuGO Gene Nomenclature Committee). No misinterpretation have been found in this phase. Those genes can be recognized not only through official symbols but also with the *Entrez Gene ID* (from the National Center for Biotechnology Information) and the *Uniprot Accession Number,* an alphanumeric identifier taken from UniProt Knowledgebase, considering only humans and only records coming from reviewed version (Swiss-Prot). In the phase of collecting data from different sources, we found some differences listed below:

* *Uniprot Accession Number*: 37 unique entries. Repetitions might be due to the merging or the splitting of two genes. In both cases, the accession numbers from all entries are kept. Just to make an example of this case, the genes BPY2 and BPY2B have the same AC because they are paralog, so they derive from the same ancestral genes and they evolve new functions;
* *Entrez Gene ID*: 44 unique entries. The only pair of genes with same id is (TSPY4, TSPY8), both of them are testis specific proteins and this analogy is totally reasonable because they derive from the same ancestral gene.

All data about seed genes have been collected and saved in a dataset so it will be easier to get the information we need.

1. Summary on interaction data

Once we gather initial information about the seed genes, we can procede collecting all binary protein-protein interactions (PPI). For this point the references are:

* *BioGRID:* Biological General Repository for Interaction Datasets, version 3.5.167;
* *Integrated Interactions Database* (IID), selecting all human tissues from experimental results. For this database we will compare the outputs obtained considering queries on both the gene symbol and the Uniprot Accession Number.

For each dataset we downloaded the last released version and, after selecting the interactions of all seed genes, we also included the interactions among non-seed genes interacting with at least one seed gene. We drop duplicates where necessary and, at the end, the main results are summarized in Table 3.1.

**Table 3.1.**Summary of interaction data**.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **BioGRID** | **IID** | **IID (AC)** |
| **Initial interactions** |  | 1676 | 1621 |
| **Number of seed genes** | 29 | 25 | 25 |
| **Seed genes interactions** | 363 | 313 | 258 |
| **Number of non-seed genes** | 300 | 218 | 218 |
| **Non-seed genes interactions** | 2908 | 1360 | 1363 |
| **All interactions (without duplicates)** | 2130 | 161 | 1621 |
| **Total number of genes** | 329 | 243 | 243 |

Starting from this table we can lead to some observations.

What we can clearly notice is the different number of interactions selected in BioGRID dataset and in IID one. There are four more seed genes and almost one hundred of other genes that contribute to increase the number of interactions in BioGRID dataset. The main reason according to which a bunch of genes have interactions in a database and not in the other is maybe because in IID database we considered only the PPI experimentally detected from some databases (including BioGrid), restricting the results.

Another observation is related to the difference between the interactions obtained in the Integrated Interactions Database (IID). Comparing the two outputs (Table 3.1, second and third columns), at the beginning we have a different number of interactions but at the end, after removing duplicates, the number of interactions turn out to be the same as well as the number of genes involved. This is due to the fact that, as we said before, sometimes some genes have same AC because they refer to the same protein or have been merged or splitted during some experiments even if the Uniprot AC is kept the same.

As last thing, we started with 45 initial seed genes but at the end, only 29 seed genes in BioGRID and 25 in IID are involved. There are some genes (like BPY2B, RBMY1B or TSPY4) that do not interact with others, even if other genes with similar functions do. On the other hand, we realized that some genes (CDY1 or DAZ1) interact with themselves or, more reasonably, with a copy of themselves.

# Interactomes data

# (tommy, scrivi una bozza e poi la aggiusto!!)

# Starting from the interactive data frame that we stored before, we can now build some interactome tables to

# - 4.1: sono 13 interazioni, solo 3 fra geni diversi, tutte le altre fra stesso gene. 5 da BioGrid

## ui. 597 interazioni totali, 339 provenienti da Biogrid, 31 sono seed genes 341 non-seed

## ii. 1157 inter tot, 22 sono seed genes 199 non-seed

# Enrichment analysis

Explain briefly the methods you followed to carry out the enrichment analysis and add the related table.

## This is Heading 2 style this is heading 2 style, if necessary

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

### 3.1.1 This is heading 3 style, if necessary

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

1. Sample numbered list.
2. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.
3. The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

* Sample bullet list.
* The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.
* The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

**Table 1.**Sample table. This should be the table format, add/remove columns and rows according to the data to be shown.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |S| | Predicted cost | Timing | Predicted speed | Speed |
| 1 | S219.20(100%) | 68m43s | 1.00 | 1.00 |
| 2 | 29.10+219.10(~50%) | 35m13s | 2.00 | 1.95 |
| 4 | 219.20(100%) | 68m43s | 1.00 | 1.00 |
| 10 | 29.10+219.10(~50%) | 35m13s | 2.00 | 1.95 |
| 20 | 219.20(100%) | 68m43s | 1.00 | 9.5 |

This is table foot note sample text This is table foot note sample text This is table foot note sample text

# Interactomes data

Explain briefly the methods you followed to build the intersection interactome and add the related table.

# Enrichment analysis

Explain briefly the methods you followed to carry out the enrichment analysis and add the related table.

# Notes and comments

References (if any)

Alexandrescu,A. (2001) Modern C++ Design: Generic Programming and Design Patterens Applied. Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.

Alexandrescu,A. (2001) *Modern C++ Design: Generic Programming and Design Patterens Applied.* Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.