Bioinformatic resources

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

June 26, 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.

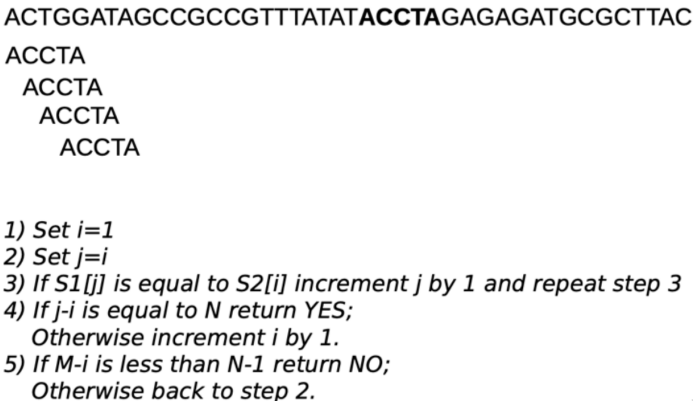

Figure 1.1: Naïve implementation of a substring search algorithm

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

4.1 Introduction

4.1.1 Definition

A DNA motif is a pattern of nucleotide sequences. They are usually associated to DNA-protein binding site and so to regulatory regions. They are a small pattern, usually between 5 and 30bp that can recur many times in the genome and many times in the same gene. Motifs can be:

- Standard.
- Palindromes.
- Gapped.

4.1.2 Functions

DNA motif functions include:

- Sequence specific binding sites reached by transcription factors, nucleases and ribosomes.
- mRNA processing:
  - Splicing: exonic splicing enhancer ESE.
  - Editing: protospacer adjacent motif PAM, a DNA sequence that immediately follows the target DNA sequence of the Cas9 nuclease in the CRISPR system.
  - Polyadenilation.
  - Transcription termination.

4.1.2.1 Degenerate motifs

Motifs in regulatory regions are often similar but variable: they are degenerate. Transcription factors are often pleiotropic, meaning that they regulate a lot of genes, but they need to be expressed at different levels. Degenerate motifs cause non-specific binding: a protein can bind genomic position different with respect to the one corresponding to the expected functional state.

4.2 Motif search

The objectives of motif search are to identify:

- Over-represented motifs in the genome.
- Motifs conserved in ortholog sequences.
- Sequences that can be candidates for transcription factor binding.

Motifs can be represented as a consensus sequence or as profiles like positional matrices or HMMs.

4.2.1 Consensus sequence

A consensus sequence represents the result of multiple sequence alignments with the goal of finding recurrent motifs across the sequences. This sequence can be potentially different from all input sequence: it presents only the most conserved sequences for each position. It is built such that it minimizes the distance from each input sequence at each position. It can be also written following a IUPAC notation.

4.2.2 Positional matrix

A positional matrix is an alternative way to represent a motif than the consensus sequencs. The elements in the matrix represent all possible bases at each position. Example of these matrices are:

- Position frequency matrix PFM or PSWM.
- Position weight matrix PWM or PSSM.
- Position probability matrix PPM or PFM.

4.2.2.1 Populating a position frequency matrix

A position frequency matrix is computed as:

$$M_{k,j} = \sum_{i=1}^{N} \delta(X_{i,j} = k)$$

Given k the set of all symbols in the alphabet. N the number of aligned sequences. j iterates over the length of the sequence.

4.2.2.2 Populating a position probability matrix

A position probability matrix is very similar with respect to the position frequency matrix, with the exception that each cell represent the probability that in that sequence position a particular base will be found. Its cells are computed as:

$$M_{k,j} = \frac{1}{N} \sum_{i=1}^{N} \delta(X_{i,j} = k)$$

4.2.2.3 Assessing the probability that a sequence belong to a PPM

To assess the probability for a sequence to belong to a PPM the probabilities for each base i found at each position j are multiplied:

$$P(seq \in PPM) = \prod_{j=1}^{R} M_{seq_j,j}$$

4.2.2.4 Correcting PPMs

4.2.2.4.1 Laplace smoothing Laplace smoothing introduces pseudocounts to allow to estimate probabilities in case of too few observations. A pseudocount is an amount added to the number of observed cases in order to change the expected probability.

4.2.2.4.2 Adding a background model Another way to correct PPMs is to add a background model. With this a new matrix is computed as:

$$M_{k,j} = \log_2 \frac{M_{k,j}}{b_k}$$

Where b represent a background model and can vary across nucleotides for organisms with high GC content. It is typically computed as:

$$b_k = \frac{1}{|k|}$$

And so is 0.25 for nucleotides and 0.05 for amino acids.

4.2.3 Hidden Markov model

A Markov chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. The possible future states are fixed and not based on how the process arrived at its present state. The state is directly visible to the observer: state transition probabilities are the only parameter. The state remain transparent, while the output is easily obtainable. A HMM of the first order is defined as:

- A finite set of states S.
- A discrete alphabet of symbols.
- A matrix of transition probabilities $T = P(i|j)$, the probability of transition from state j to $i>$
- A matrix of emission probabilities (the probability of skipping a state) $T = P(X|i)$, the probability of X emission in state i.

4.2.3.1 Assessing the probability that a sequence is generated by a HMM

The probability that a sequence is generated by a HMM can be computed as:

$$P(S|w) = \sum_{\pi} P(S, \pi|w)$$

Where S is the sequence, w are the probabilities parameters, π are all possible computationally inefficient paths. Efficient algorithm to compute this probability are Forward-Backward and Viterbi.

4.2.4 Sequence logos

Sequence logos are visual representation of positional matrices and simple HMM profiles. The height of each character in a sequence logo is proportional to its information content: 2 bit if 1 base occurs in all input sequences, 1 if two bases occur and 0 if all bases occur equally. The higher the variability, the lower the height of a specific base. In particular the height of base b at position l is computed as:

$$f(b,l)R_{sequence}(l)$$

Where:

$$R_{sequence}(l) = 2 - (H(l) + e(n))$$

Such that $H(l)$ is the Shannon entropy and is computed as:

$$H(l) = -\sum_{b=a}^{t} f(b,l)\log_2 f(b,l)$$

And

$$e(n) = \frac{1}{\ln 2} \cdot \frac{4-1}{2n}$$

4.3 Motif identification

There are two types of motif identification: pattern matching and pattern discovery.

4.3.1 Finding known motifs - pattern matching

Pattern matching is the problem of finding known motifs, for example seeing if a binding of a protein to an upstream region of a gene is significant. In order to find out whether a transcription factor matches a promoter the PFM matrix is used to compute a score for each sliding window. This scores can be plotted against a threshold, so as to identify regions able to support a putative binding.

4.3.1.1 Total binding affinity

Total binding affinity TBA is a cutoff-free method. The TBA is a method used to describe the affinity of a DNA sequence for a transcription factor described by a PFM with a single score. It takes into account binding sites of all possible affinities and considers the whole sequence, keeping into account both high and low affinity sites.

4.3.2 Finding de novo motifs - pattern discovery

Pattern discovery if the problem of finding de novo motif, for example finding the motifs upstream of a specific gene. Given a set of sequences, the objective is to find the most represented motifs. Using the MEME suit it is possible to identify new sequences and through Jaspar they can be compared to already characterized transcription factors. Methods can be:

- Exact: give optimal solution given specific parameters.
- Approximated: give suboptimal solution decreasing the computational burden. They are MULTIPROFILER, CONSENSUS, MEME, Gibbs sampler and MotifSampler for example.

4.3.2.1 Distance between a real motif and the consensus

The distance between a real motif and the consensus is generally less that that for two real motifs. The consensus sequence must be guess and a scoring function to compare different guesses and choose the best one must be chosen.

4.3.2.2 Elements of the problem

The problem of finding de novo motifs can be formalized considering the following elements:

- n the length of each sequence.
- DNA, an array of size $t \times n$.
- l, the length of the motif or l-mer.
- s_i, the starting position of an l-mer in sequence l.
- $s = (s_1, s_2, \ldots, s_t)$, an array of motifs starting position.

If the starting positions s are given, finding the consensus is easy. When those are not given, finding the best motif is solving the median string problem.

4.3.2.3 The median string problem

Given a set of t DNA sequences the objective is to find a pattern that appears in all t sequences with the minimum number of mutations. The Hamming distance is used, such that:

$$d_h(v, w) = \#\ nucleotide\ pairs\ that\ do\ notmatch\ when\ v\ and\ w\ are\ aligned$$

Then, for each DNA sequence I, all $d_h(v, x)$ are computed, where x is an l-mer with starting position s_i. Then the minimum $d_h(v, x)$ among all l-mers of the sequence. The $TotalDistance(v, DNA)$ is the sum of the minimum Hamming distances for each DNA sequence I, so

$$TotalDistance(v, DNA) = \min_s d_h(v, s)$$

Where s is the set of starting positions.

Chapter 5

Expression analysis

5.1 Introduction

5.1.1 Expressed genes

The expressed genes are those genes that have been transcribed. A gene expression profile of a cell is the snapshot of which genes are expressed in that cell at the time the sample was taken. Knowing which genes are expressed in a cell allows the identification of new genes or transcripts and the comparison of expression profiles between samples. Variability in gene expression is mainly due to alternative splicing and different regulation. It can be analysed to uncover characteristics of diseases, development and dynamic responses to stimuli.

5.1.2 Differential gene expression

During a differential gene expression experiment the expression profile of genes is compared between samples. Comparison can be done between:

- Different cells.
- Different tissues.
- Different disease states.
- Different developmental stages.
- Different culture conditions.

Negative and positive controls and the range of variability within samples must be taken into account.

5.1.2.1 Differential gene expression workflow

A typical differential gene expression analysis workflow consists of:

1. Formulation of the biological question.
2. Experimental design: choice of platform, control and replicates.
3. Running the experiment.
4. Image processing done by a machine.
5. Low-level analysis: data pre-processing with normalization.
6. High-level analysis.
7. Obtaining biological conclusions and interpretation of results.

5.1.2.2 High throughput methods

To perform differential gene expression analysis high throughput methods can be used. Their pros and cons are described in table 5.1.

Pros	Cons
Fast	Difficult to filter non coding RNA
Comprehensive (entire genomes)	Not enough attention to design
Easy	Artefacts
Getting cheaper	Cannot afford controls or replicates

Table 5.1: High throughput methods pros and cons

5.1.3 Databases

Repositories of array and NGS data mainly contain expression data. All of these databases can be interrogated with Bioconductor packages in R.

5.1.3.1 Gene expression omnibus

The gene expression ombinbus or GEO is a public repository for the archiving and distribution of gene expression data submitted by the scientific community. It is a curated, online resource for gene expression data browsing, query, analysis and retrieval. It is convenient for the deposition of gene expression data as required by funding agencies and journals. Submitted data needs to include:

- Platform.
- Sample.
- Series.
- Dataset.

GEO is connected to repositories specifically tailored to store raw data like BioProject or SRA.

5.1.3.2 ArrayExpress

ArrayExpress or EBI si another online repository of array expression data.

5.1.3.3 Gene expression Atlas

The gene experssion atlas GXE NCBI provides information on gene expression patterns. The raw data is re-analyzed with common pipleines.

5.2 Microarrays

5.2.1 Introduction

Microarrays have been introduced at the beginning of the 2000s and are the first high throughput technology. They are useful to investigate, for example:

- Genomic profiles.
- Transcriptomic profiles.
- The methylome.
- DNA-protein interactions.

Data interpretation is subject to specific computational analyses. Microarrays monitor thousands of genes in parallel. Each spot contains multiple and identical DNA probes and thousands of spots are disposed as a matrix on a solid support.

5.2.2 Fabrication

Microarrays can be fabricated using different technologies. Probes can be:

- Oligonucleotides.
- cDNA.
- Small PCR fragments related to specific mRNA.

The probes are synthesized and placed on the support and can have different lengths, usually between 25 and 60nt. Moreover microarrays can have different numbers of channels:

- 2 channels: test and control samples are labelled with different fluorophores.
- 1 channel: one sample is loaded per time.

Depending on the technology microarrays can capture, for example:

- Exons.
- Genes.
- 3′-ends.

5.2.3 Reading signal

A scanner allows to read the fluorescence light emitted by the fluorophores. The information is stored in 2 images for channels arrays at 16 bit resolution. The image in grey scale is represented in a red-green scale that represents the light emitted by the two fluorophores. The resulting colour will be proportional to the quantity of test and control DNA.

5.2.4 Image analysis

After having obtained the images, these have to be analyzed. This is performed by a specific technology like Affymetrix. The first step is a segmentation analysis: the shape and patterns inside data is analysed to assess the signal quality for each spot. The background and foreground are identified to correct for noise generated by the former. One of the standard method is to create a signal model and fit the data to it in order to evaluate the quality of the spot, for example computing the interquartile range on the distribution in order to find feature, exclusion zone and background. Then the fluorescence of each spot is estimated and the relative expression for each gene is interpreted thanks to annotation information.

5.2.5 Batch effect

For microarrays is complex to compare different technologies as different probes and methods are used. It is always preferable to avoid performing an integrative analysis.

5.2.6 Data pre-processing

Data pre-processing is needed to reduce errors introduced during the experimental process. It consists typically of:

- Background subtraction: eliminates background noise.
- Normalization: all samples are brought into a similar range of distribution, to reduce the effect of:
  - Unequal quantity of starting sample.
  - Differences in labelling efficiency.
  - Differences in detection efficiency.
  - System biases.
- Summarization: summary of information from several spots into a single measure for each gene.
- Statistical quality control: removes low quality samples and probe sets.

5.2.6.1 Two channels array

The pre-processing pipeline for 2 channels array consists of different steps.

5.2.6.1.1 Background correction During background correction signal R_s and G_s and background estimates R_b and G_b are separated. Then the background corrected estimates R_c and G_c are computed as:

$$R_c = R_s - R_b \qquad \wedge \qquad G_c = R_s - R_b$$

Or as:

$$R_c = \max(R_s - R_b, 0) \qquad \wedge \qquad G_c = \max(G_s - G_b, 0)$$

5.2.6.1.2 Summarization and transforms Log-ratios estimates relative expression as"

$$\log \frac{R_c}{G_c}$$

5.2.6.1.3 Normalization Normalization is useful to identify systematic intensity-dependent bias in the data. The ratio of signal might depend on the average signal density measured across different channels. The function of dependance can be fitted to a polynomial regression like Loess to obtain normalization to make the plot more informative.

5.2.6.2 One channel array

Many methods have been developed to pre-process Affymetrix one channel arrays:

- Advanced methods: GCRMA, PLIER.
- Popular mehtods: RMA and MAS5.
- Rudimentary methods: MAS4, LOESS.

5.2.6.3 Robust multi array average

Robust multi array average is a pre-processing methods that consists of three steps.

5.2.6.3.1 Background correction Background correction removes local artefacts and nois. The probe measure data is assumed as a combination of background noise in a normal distribution and signal in an exponential distribution. Assuming strictly positive distribution of signal background, the corrected signal is positively distributed. Background correction is performed on each array separately using the observed distribution of PMs.

5.2.6.3.2 Normalization Normalization is used to remove array effects, making all distributions the same. Quantile normalization is used to correct for array biases, as it compares the expression levels between arrays for various quantiles. It protects against outliers.

5.2.6.3.3 Summarization Summarization combines probe intensities across arrays to get a single intensity value tor each gene or probeset. In median polishing each chip is normalized to its median and each gene normalized to its median. The procedure is repeated until medians converge. A maximum of 5 iteration is allowed to prevent infinite loops.

5.2.7 Gene expression microarray

An expression microarray experiment is used to test differences in gene expression between two or more conditions that could be for example cancer versus normal or different treatments. Each condition can be represented by one or more samples. The null hypothesis is that there exists no difference between the gene expression in the two conditions. The comparison is done using the ratio between the test and the control samples. It should not differ in case of null hypothesis validity. These ratios are also defined as fold changes:

$$FC = \begin{cases} Ratio & Ratio > 1 \\ -\frac{1}{Ratio} & Ratio < 1 \end{cases}$$

Because ratios are not symmetric with respect to 1 the statistics are not easy to analyse, so the log-ratio is often used. The log ratio of the null hypothesis should be 0.

5.2.7.1 Replicates

Replicates are needed considering the noise of microarray data. They can be distinguished between:

- Technical replicates: experiments on more RNA samples obtained from the same biological source.
- Biological replicates: experiments on more biological sources belonging to the same condition.

Ideally each condition should be represented by more biological replicates in order to perform a statistical test. They can also be summarized as mean for each gene.

5.2.7.2 Statistical tests

Microarray correlation can be exploited to identify differentially expressed genes. A gene is called differentially expressed through a Z statistics. To find significantly differentially expressed genes a test statistic for each gene should be used. A low p-value is interpreted as evidence that the null hypothesis can be false and so a gene is differentially expressed.

5.2.7.2.1 T-Test The T-test is a parametric test to check the difference between the mean of two groups. It assumes that the variance of those two groups is the same. Is computed as:

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

5.2.7.2.2 Walch t-test Walch t-test considers different variance between two groups. It is the default implementation fo *R* $t.test()$ function. Is computed as:

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}$$

5.2.7.2.3 Wilcoxon test The Wilcoxon test is a non parametric test to check the equality of two distributions.

5.2.7.2.4 Permutation test The permutation test generates a null distribution on an observation of interest by changing the group labels. It compares the values observed in data and the values in the generated null distribution. Is computed as:

$$p = \frac{\#\{b : |T_b| \geq T_{obs}\}}{B}$$

5.2.7.2.5 ANOVA analysis ANOVA analysis is used to control different treatments. It allows to test the null hypothesis that the differences within and between at least 3 groups are the same on average.

5.2.7.2.6 2-way ANOVA 2-way ANOVA compares the mean differences between groups that have been split on two independent variables called factors. It understand if there is an interaction between the two independent variables on the dependent one.

5.2.7.2.7 From p-value to probability of significant results Given a p-value and n the number of hypothesis to test, P the probability of having al least one significant result is computed as:

$$P = 1 - (1 - pvalue)^n$$

5.2.7.3 Correction methods

Correction methods are used to correct p-values in the presence of multiple null hypothesis. Some examples are:

- Bonferroni: very conservative, it has a significance threshold of $\frac{\alpha}{N}$. It reduces false positive while introducing false negatives.
- Benjamini-Hochenberg: it tunes false positives and false negatives.
- False discovery rates: it checks if the kth ordered p-value is larger than $\frac{k \cdot a}{N}$

5.2.7.4 Receiver operating characteristic

The receiver operating characteristic find genes that better discriminate between two conditions.

- $AUC = 1$: perfect classification.
- $AUC = 0.5$: random classification.

5.2.7.5 Clustering

The objective in differential gene expression is to look for genes that behave differently between samples. Once having obtained then they can be clustered according to similar expression to make the outliers more obvious.

5.2.7.5.1 Clustering pipeline The algorithm creates a dendrogram based on on a definition of distance measure between observation pairs. The idea is:

- Starting with N observation and a distance measure for all $\frac{N(N-1)}{2}$ pairs. Each observation is a cluster.
- For $i = n$ to 2:
  - Examine all inter-clusters distances and fuse the cluster with lower distance. The distance between fused clusters represents the height of the bar in the dendrogram.
  - Calculate new inter-cluster distances between the remaining $i - 1$ clusters.

5.2.7.5.2 Linkage Linkage defines the distance between clusters.

5.2.7.5.2.1 Complete linkage In complete linkage maximal intercluster dissimilarity is reached. All pairwise dissimilarities between the observations in cluster A and in cluster B are computed and the largest is recorded.

5.2.7.5.2.2 Single linkage In single linkage minimal interclsuter dissimilarity is reached. All pairwise dissimilarities between the observations in cluster A and in cluster B are computed and the smallest is recorded. It can result in extended, trailing clusters in which single observations are fused one at a time.

5.2.7.5.2.3 Average linkage In average linkage te mean intercluster dissimilarity is reached. All pairwise dissimilarities between the observations in cluster A and in cluster B are computed and the average is recorded.

5.2.7.5.2.4 Centroid linkage In centroid linkage the dissimilarity between the centroid for cluster A (a mean vector of length p) and for cluster B is computed and recorded. It can result in undesirable inversions.

5.3 RNA-sequencing

5.3.1 Introduction

RNA sequencing is a next-generation sequencing approach that sequences the cDNA from the mRNA component. The whole transcriptome can be compared against the whole transcriptome of another sample. It is very cheap, sequencing $\sim$ 400 gigabases per flow cell, but it has an error rate up to 1%, issues with AT and GC regions and long sequencing items.

5.3.2 Pipeline

A RNA-sequencing pipeline consists of different steps.

5.3.2.1 Sample preparation

In sample preparation RNA is extracted and sheared into 300-600np fragments through sonication or enzimatic digestion.

5.3.2.2 Library preparation

During library preparation adapter sequences are ligated to the fragments. Barcoding is possible allowing for sample multiplexing.

5.3.2.3 Cluster generation

During cluster generation the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters.

5.3.2.4 Sequencing

Sequencing is done by synthesis:

- Paired ends are used for duplicates, splicing analysis and discovery of novel isoforms.
- Single ends are used fo gene expression analysis.

5.3.2.5 Alignment

Reads are aligned to a reference genome or transcriptome or to a genomic region. Splice-aware aligner such as TopHat or STAR should be used. The alignments are refined according to coding sequences using known and predicted splice junctions.

5.3.2.6 Quantifying reads

Reads per gene are quantified. non-coding RNA are filtered out. Alternative splicing, overlapping genes and pseudogenes are dealt with. Different type of counts can be employed:

- Standard count: number of reads for transcript.
- CPM: counts scaled by the number of fragments sequenced N times one million. It allow to compare each transcript across different smples.
- TPM: measurement of the proportion of transcripts in the pool of RNA. It takes into account the length of the transcript. It is not used for direct comparison across samples, but for intra-sample normalization.
- RPKM: reads per kilobase of exon per million erads mapped. It is called FPKM for fragments.

5.3.2.7 Normalization

To perform inter-sample normalization quantile normalization is used to make the distribution identical. The best normalization methods for differential expression are coupled with sophisticated approaches as very low expressing genes are tricky, for example those with $FPKM < 1$.

Chapter 6

Gene set enrichment

6.1 Introduction

6.1.1 Functional groups characterized by gene expression change

Functional groups characterized by gene expression change aer:

- Gene sets: the set is scored depending on the expression level of its member genes.
- Network: modules satisfying some joint gene expression and topology requirement are identified.
- Pathways: they are scored exploiting gene expression and topology.

6.1.2 Gene-set enrichment analysis

Gene-set enrichment analysis is the breakdown of cellular functions into gene sets. Every set of gene is associated to a specific cellular:

- Function.
- Process.
- Component.
- Pathway.

Microarray or RNA-seq data can be related to gene sets in order to mine its functional meaning, to find which gene sets summarize at best gene expression patterns.

6.1.3 Ontology

Ontology normally represents knowledge as a set of concepts within a domain and the relationships among those concepts. It can be used to reason about the entities within that domain and may be used to describe the domain.

6.1.4 Controlled vocabulary

A controlled vocabulary provides a way to organize knowledge for subsequent retrieval but does not allow reasoning about the entities.

6.1.4.1 Ontology and controlled vocabulary uses

Ontologies and controlled vocabularies are heavily used in biological databases as they allow the organization of data within a database, providing a meaningful link between database structure and search queries.

6.2 Gene ontology

A gene ontology is a way to capture biological knowledge for individual gene products in a written and computable form.

6.2.1 Concepts hierarchy

A set of concepts and their relationships to each other is arranged as a hierarchy.

6.2.1.1 Molecular function

The molecular function describes activities that happen at a molecular level like catalytic or binging activity. This category includes the activities rather than the entities that are involved in an action and do not specify where, when or in which context the actions happen. Molecular functions can be executed by single gene products or complexes of gene products. Some examples are:

- Catalytic activity.
- Transport activity.
- Binding.
- Toll receptor binding.

6.2.1.2 Biological process

A biological process is a series of events resulting from multiple ordered groups of molecular functions. A biological process is different from a pathway, as gene ontology does not report the dynamics or the dependences that are required to describe a pathway. Some examples are:

- Cellular physiological processes.
- signal transduction.
- Metabolic process of pyrimidine.
- Glucose transport.

6.2.1.3 Cellular component

A cellular component is linked to a component of a cell with the condition that is part of a larger object. Some examples are:

- Ribosome.
- Nucleus.
- Neuron parts.
- Internal nuclear membrane.

6.2.2 Ontology structure

An ontology is structured as an acyclic graph where terms can have more than one parent. Terms are linked by directed relationships like:

- Is part of.
- Regulates.

6.2.3 The gene ontology project

Originally in the gene ontology or GO project the hierarchies were completely independent, without links between them. From 2009 biological processes and molecular functions are linked, as biological processes are ordered assemblies of molecular functions. GO is required as there are inconsistencies in the human language: different concepts can have the same name. Furthermore it enables to interpret quickly large datasets. The aim of the GO project is to:

- Compile ontologies.
- Annotate gene products using ontology terms.
- Provide a public resource of data and tools for annotation.

A gene ontology annotation is a statement that a gene product has a particular molecular function, determined by a particular method and described in a reference.

6.3 Gene sets

6.3.1 Sources and types

Other than gene ontology there are other sources and types of gene-sets:

- Pathways (KEGG).
- Protein families and domains (PFAM).
- Predicted target of regulators like mRNA and transcription factors (MSigDB-c3).
- Protein-protein interaction modules.
- Gene expression:
  - Up and down regulation after treatment or in relation to disease (MSigDB-c2).
  - Co-expression across many conditions (MSigDB-c4).
- Genotype-phenotype association (DiseaseHub).
- Genomic position (MSigDB-c1).

6.3.1.1 Main resources

The main resources for these type of data are:

- Bioconductor.
- DiseaseHub.
- MSigDB.
- PathwayCommons.
- WhichGenes.

6.3.2 Differences between pathways and processes

From a biological perspective the difference between pathways and processes is philosophical. It is still worth speculating in a bioinformatics perspective because a gene is annotated for a GO biological process if the curators deem it significantly contributes to the process according to a number of evidences. Pathway include the wiring of genes and gene products, hence they rely on a more

intensive curation process. Some pathways include large ubiquitous actors such as the proteasome that may confound enrichment analysis, whereas the are usually absent from GO processes.

6.3.3 Enrichment test

6.3.4 Whole-distribution - GSEA enrichment

6.3.5 Gene set filter

6.3.6 Redundancy problem

Chapter 7

Network analysis

Chapter 8

Genome wide association studies