Bioinformatics / Computational Biology

Genes and metabolic diseases - Finding places to reproduce DNA for the gene therapy of metabolic diseases - Analysis of gene expression data from DNA microarrays and practical training in the field of Nutritional Genomics

1. Purpose of Exercise

The purpose of the exercise is to familiarize students with concepts of Bioinformatics and Computational Biology, as well as to inform them about the basic analysis techniques used. The theoretical part of the exercise includes introductory data, description of the role of specific genes in the occurrence of metabolic diseases, theoretical data for finding DNA replication sites, description of technologies for measuring gene expression (gene expression profiling) with emphasis on DNA microarrays, as well as basics steps of data analysis from microarray experiments. One of the main goals of the analysis is to find genes that present the so-called differential expression (differential
expression) between states of the experiment. Techniques for data normalization, statistical controls applied to the data, and data analysis by clustering are described. In the practical part of the exercise, students have the opportunity to practice in basic steps of biological data analysis. The analysis is performed with the MATLAB programming environment as well as with the R language, which is very widespread in the field of Data Science, as well as with appropriate web analytics tools and databases, offered by organizations such as the National Center for Biotechnology Information (NCBI).

2. Exercise preparation

Understanding exercise requires basic knowledge of Biology, about the organization of genetic information in the cell and its transmission. You can refer to suitable sources on the internet for terms such as DNA, RNA, proteins, transcription, translation, central doctrine of Molecular Biology, genome, transcript, etc. As shown in Figure 1, based on the central tenet of Molecular Biology, the encoded information in the genes is transferred to the proteins in the following flow: DNA -> RNA -> proteins. The synthesis of an RNA copy from DNA is called transcription, while the synthesis of RNA-based protein is called translation.

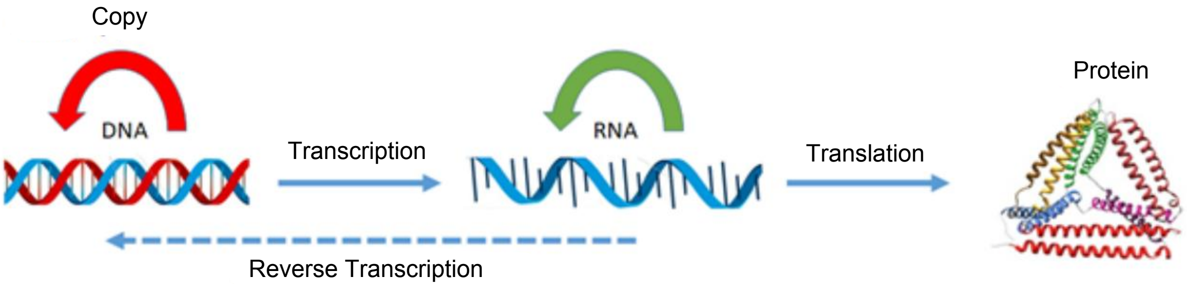

Figure 1: The central doctrine of Molecular Biology

3. Genes and metabolic diseases

3.1 Quotations

DNA is contained in the nucleus of cells, contains genes and is the model for the synthesis of proteins as well as the means of transmitting information inherited from generation to generation. DNA is a macromolecule, made up of nucleotides. Each nucleotide consists of a sugar molecule, deoxyribose, attached to a phosphate group and a nitrogenous base. In DNA nucleotides the nitrogenous base can be one of: adenine (A), guanine (G), cytosine (C) and thymine (T). Each gene is a subset of DNA and consists of coding regions, the exons, and non-coding regions, the introns. Small changes in DNA, such as a single nucleotide mutation, or even large changes, such as the simultaneous mutation of multiple genes or even an abnormal number of chromosomes, combined with environmental and lifestyle factors can make a person predisposing factor or cause one or more diseases.

3.2 Biological clock genes and metabolic diseases

The internal biological clock determines circadian changes in behavior and physiology, and is regulated by light or dark in the environment, eating disorders and other hormonal signals. The term circadian comes from the Latin words "circa" and "dies", which mean "about" and "day", respectively. Circadian rhythms are present in most living organisms, from unicellular to mammalian, are around the clock and allow living organisms to synchronize with the external light-dark cycle. Several studies have shown that light, which is perceived by specific photosensitive ganglion cells of the retina that invade the suprachiasmatic nucleus, is the major coordinator of circadian rhythm in humans [1]. In addition to the sleep / wake cycle and meal / fasting cycles, the circadian system also regulates normal processes such as lipid and glucose metabolism, body temperature, and hormone secretion, allowing optimization of energy intake, use, and storage. during the day. Many parameters related to glucose metabolism, such as glucose tolerance, insulin sensitivity, and plasma glucose, glucagon, and insulin levels show circadian changes. In humans, maximal insulin secretion is observed during the day, while during the night there is a decrease in insulin secretion and an increase in glucose production. In rodents, this pattern of insulin secretion is shifted by 12 hours according to their nocturnal activity.

The circadian rhythm can be deregulated due to the abnormal operation of the biological clock or due to the synchronization between the suprachiasmatic nucleus and the external environment or due to the synchronization between the suprachiasmatic nucleus and the peripheral clocks. Disorder of the circadian rhythm can lead to diseases such as Metabolic Syndrome, Obesity and Type 2 Diabetes. Today's lifestyle and habits, such as working and eating at night, exposure to artificial light during the night, and modified sleep schedules are the most important factors in deregulating the circadian rhythm. Shift workers, who are the prime example of circadian rhythm deregulation, experience changes in the pancreatic ÿ-cell response and lipid and glucose metabolism, as well as an increased risk of developing metabolic syndrome, heart attack, heart disease, and Type 2 Diabetes. People with modified or reduced sleep patterns have an increased Body Mass Index, decreased glucose tolerance, and increased insulin resistance. Obese people or people with Type 1 or Type 2 Diabetes have circadian rhythm disorders, insulin secretion and glucose tolerance. Some of

These disorders have been observed in people with diabetes who have been deprived of sleep. Experiments performed on experimental animals with modified biological clock genes have revealed that the circadian rhythm plays a central regulatory role in glucose metabolism and homeostasis, and in particular that abnormal function of the biological clock genes may lead to pre- or postoperative diabetes. Studies of the interaction of biological clock genes with glucose metabolism in humans have shown that genetic variants of the CLOCK gene are associated with increased susceptibility to Obesity and Metabolic Syndrome. Genetic variants

(Genetic variants) of the BMAL1 gene are associated with Hypertension, Gestational Diabetes and SDT2, while mononucleotide polymorphisms of the PER2 gene have been associated with high levels of fasting glucose and abdominal obesity. Genetic variants of the CRY2 gene are associated with SDT2. A genetic polymorphism of the melatonin receptor, a substance that plays a key role in determining circadian rhythms, is also associated with decreased insulin secretion, gestational diabetes, and SDT2. In summary, there are many studies that suggest a close link between biological clock gene dysfunction and metabolic diseases, such as Metabolic Syndrome, Obesity, and SDT2 [2].

3.3 Gender Genes and Diabetes

Diabetes mellitus (SD) is a major global health problem, with the World Health Organization warning that more than 400 million people will develop SD in 2030. high blood glucose levels due to either insufficient secretion of the hormone insulin by the pancreas or tissue insulin resistance. SDT2 is the result of a complex interaction of factors related to the lifestyle of the individual and the genome he has inherited. Diabetic environmental factors affect people with different genetic backgrounds differently. Thus, for the greatest success of the prevention and treatment of SDT2, it is necessary to understand the mechanisms of gene-gene interaction and gene-environment. The study of the emergence and development of SDT2 in Pima Indians in Arizona has contributed significantly to our understanding of diabetes. The Pima Indians, like many other indigenous peoples, were displaced by their 1902 Law (Reclamation

Act). The change from the rural way of life to the economically and nutritionally poorer urban way of life has led to an increase in the incidence of obesity and STD2 in this population. The very high incidence of SDT2 (> 50%) and its very early onset, even in adolescence, in Pima Indians as opposed to other populations in the area of European descent and similar lifestyle, indicate the catalytic effect of genes in the appearance of SDT2 [3].

3.4 Practical part

Question 3.1: Find information about the INS and CLOCK genes using appropriate web resources (http://www.genenames.org/ and http://omim.org/) and confirm that Type 2 Diabetes is linked to the INS gene.

Question 3.2 .: Find out on which chromosome and in which genomic region the CLOCK gene is located using appropriate online sources (http://genome.ucsc.edu/).

Question 3.3: Retrieve the DNA sequence of the INS gene and the encoding protein, as well as the rational genes in the chimpanzee and mouse.

In your report you must answer all the above Questions 3.1. to 3.3., as well as include the necessary screenshots that justify your answers.

4. Finding DNA replication sites for gene therapy metabolic diseases

4.1 Quotations

Genome reproduction is one of the most important functions of the cell. Before the cell divides, the genome reproduces so that each of the daughter cells inherits its own copy of the genome. Watson and Crick in their 1953 study described the process of replication of the genome as follows: The two strands of the original DNA molecule are unwound and each strand is a template for the synthesis of the new molecule for each of the daughter cells.

DNA replication begins in specific genomic regions called replication origin (ori) regions and is carried out by molecular copy machines called DNA polymerases.

4.2 Finding DNA replication sites

Finding breeding sites is of particular importance not only for understanding the function of cell reproduction but also for resolving important biomedical problems with genomic therapy methods. Genomic therapy methods use genetically modified mini-genomes, called viral vectors vectors) and can penetrate cell membranes, like real viruses. In 1990, genomic therapy was first applied to humans to treat Severe Combined Immunodeficiency in a four-year-old girl.

The central idea of genomic therapy is to infect a person with a deficiency in a critical gene with a virus vector containing an artificial gene that encodes the therapeutic protein for the disease under study. To make sure that the vector of the virus replicates properly inside the cell, biologists need to know the exact regions of DNA replication.

4.3 Practical part

In the following exercise we will focus on the problem of finding breeding sites in the bacterial genome. There should be a "hidden message" in the play area stating that the playback process should start there. We do know that the start of reproduction is accomplished with the help of DnaA, a protein that binds to a short segment within the region of reproduction called the DnaA box. Based on the assumption that DNA is a language, we will look for common "words" within the breeding area of the bacterium genome. We believe that these common "words" may be the binding sites of the DnaA protein. For example the "word" ACTAT

appears very often in the following symbol:

ACAACTATGCATACTATCGGGAACTATCCT

Question 4.1 .: Implement a function in the MATLAB programming environment that finds how many times a given "word" appears in a genome section, and then try to use this function to calculate how many times the ATC "word" appears in the following genome section.

```
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
```

```
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctattttttacggaagaatgatcaagctgctgctcttgatcatcgtttc
```

Question 4.2: Implement a function in the MATLAB programming environment that calculates which are the most frequently displayed k-character "words" in a genome section, and then try to find the most common 9-character "words" in the upper part of the bacterium genome.

In your report you should include your answers to the above Questions 4.1. and 4.2., as well as the code in MATLAB that you implemented.

5. Analysis of gene expression data from DNA microarrays and practical training in the field of Nutritional Genomics

5.1 Quotations

The study of the genome and the way of its coding and organization are the subject of Genomics. The field of so-called "omics" also includes
Transcriptomics, which studies molecules that result from the transcription of a genome, such as mRNA. The study of proteins, as functional molecules produced on the basis of RNA, is the subject of the field of Proteomics.

The genes contained in an organism's genome are not permanently activated. Each cell of the organism, at any given time and under specific conditions, produces different combinations and amounts of proteins, through the expression (transcription) of the relevant genes. It is practical and at the same time reliable to quantify it
mRNA level expressed in cells, instead of proteins. For the study of transcriptome, large-scale experiments are used, in which the expression of genes, possible changes in it (activation / repression of genes) and the biological significance of these changes are studied.

5.2 Technologies for recording gene expression

The main technologies used to measure expressed mRNA on a large scale (genome-wide survey of the transcriptome) are mainly DNA microarrays and RNA sequencing. Each of them has advantages and disadvantages. RNA sequencing seems to have prevailed over DNA microarrays, however DNA microarrays are a mature technology for which there is extensive experience and knowledge in data analysis. In addition, there is a wealth of data available for analysis in public repositories. RNA sequencing is still an expensive technology and the analysis of the resulting data is demanding. On the other hand, RNA sequencing

provides great flexibility in locating sequences. It is worth noting that despite him
different type of data produced by the two technologies, the analysis approach for drawing biological conclusions has common features.

5.3 Brief introduction to microarray technology

DNA microarray technology is used to simultaneously measure a large number of RNA transcripts in a cell sample. It will then be roughly explained how gene expression is quantified. The presentation of more technical details goes beyond the purposes of the exercise. The function of microarrays is based on the property of DNA to hybridize, i.e. single-stranded nucleic acids can interact with complementary sequences to form a double-stranded complex. Microarrays are essentially micro-plates on which certain nucleic acid molecules are arranged in a lattice arrangement. Each tile is designed so that includes sequences from characteristic genes of the organism under study. These strands of DNA immobilized on the plate are called probes. To measure the expressed mRNA, the following is done: first the mRNA is isolated from the cell sample and then the complementary DNA (cDNA) is generated, by reverse transcription,

which is indicated by a suitable fluorescent dye. Then solution with the cDNA
passes over the plate and hybridizes to the corresponding complementary probes on the microarray. When the hybridization has taken place, fluorescence is activated and by measuring the fluorescence level we can quantify the expressed mRNA [4].

The results of the measurements in a gene expression experiment with microarrays have the two-dimensional array format, as shown in Figure 2. The array lines correspond to the different microarray detectors, while the array columns correspond to the different samples counted with the same microarray type in the experiment. Each cell of the table contains the measurement recorded for the respective probe (gene)

and for the corresponding sample of the experiment.

ID_REF	GSM162954	GSM162956	GSM162957	GSM162958	GSM162959
1007_s_at	218.27	255.46	137.36	254.60	121.07
1053_at	92.81	121.42	51.46	119.53	85.78
117_at	53.72	64.33	49.31	59.87	42.92
121_at	225,12	208.08	279.43	208.61	246.80
1255_g_at	11.71	11.62	12.69	12:15	13,14
1294_at	243.61	277.53	170.93	312.98	316.07
1316_at	52,19	52.38	54.02	49.59	51.79
1320_at	50.09	51.02	43.81	44.33	51.06
1405_i_at	112.19	87.33	114.67	405.02	136.76
1431_at	12325,61	11801,22	6845.59	10051.77	11614,58
1438_at	55.71	48,14	69.34	47.41	50.05
1487_at	337.21	441.02	222,22	382.95	328.81
1494_f_at	8941.16	7506.00	3975.62	6156,13	8837.92
1552256_a_at	954,65	1291.16	717.29	1354.82	1614,45
1552257_a_at	133,61	243.35	127.39	148.38	187.68

Figure 2: Part of gene expression experiment measurements with DNA microarrays (after relative pre-processing of the data)

5.4 Analysis of a gene expression experiment with DNA microarrays

Roughly speaking, the steps to follow in a DNA microarray experiment include: a) defining the biological query under consideration, b) designing an appropriate experiment to investigate the biological query, and determining the type of microarray that will be needed for the measurements. , (c) appropriate sample processing and hybridization; (d) reading the measurements from the microarray to derive the expression level based on fluorescence values; and (finally) (e) processing and analyzing data using appropriate methods conclusions. In the context of this laboratory exercise, the focus is on the stage of data analysis.

For this purpose, the quantified measurements resulting from the microarray must be properly pre-processed, including "data cleaning", checking
quality and normalization. Basic analysis techniques can then be applied

microarray data, such as identifying genes that exhibit differential expression and identifying groups of genes with common expression patterns.

Data normalization

The normalization of microarray data is very important. With normalization we can convert the data from all the samples of an experiment to a common scale, so that they can be compared. This removes systematic errors that may or may not have occurred for technical reasons. Initial measurements from a microarray experiment (fluorescence measurements) are transformed by applying a logarithm (usually a logarithm based on 2). The initial fluorescence intensity values show a large dispersion, while with normalization the values are evenly distributed so the different samples of the experiment can be compared with each other. If we want to compare data from different experiments, then more "advanced" forms of normalization must be used.

[4].

Box Plots

Graphs are used in statistics to graphically present the summary measures of a distribution and help to compare distributions with each other. In the experiments gene expression we use the graphs to determine if the data need normalization or if they have been normalized successfully. For each sample of the experiment, a graph of the distribution of the measured data for all genes in the sample is drawn. What is expected is that for normalized data the graphs will be comparable for all samples. A large deviation in the graphs is an indication that the data are not balanced.

Figure 3 shows the graphs obtained from a gene expression experiment before and after data normalization. Each rectangle is marked with a rectangle, the lower side of which shows the first quadrant and the upper side shows the third quadrant of the distribution. The horizontal line inside the rectangle shows the median. In addition, the minimum and maximum value of the distribution are noted (calculated after excluding the extreme values - outliers). If we choose it, the graphs can also show the extreme values. If the values through the distributions in all samples are at the same level, this is an indication that the data are normalized and therefore the samples can be compared with each other.

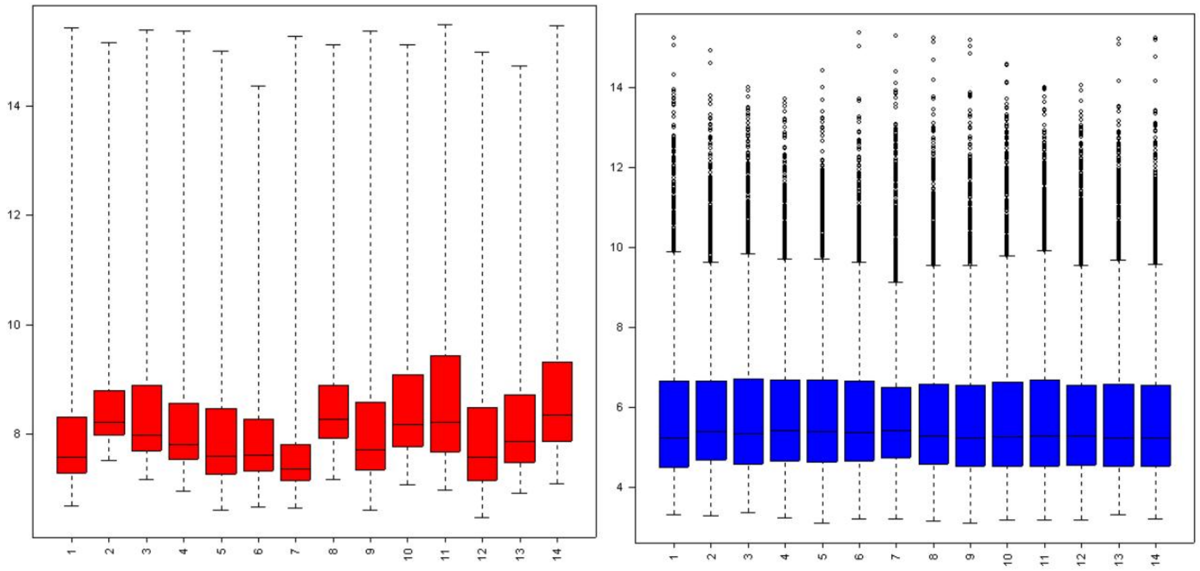

Figure 3: Diagrams for each sample of the experiment before normalization (left image) and after normalization (right image) of the measurements [5]

Identification of genes with differential expression and control of statistical significance

It then examines how genes that exhibit differential expression between states of an experiment are defined and how statistical analysis is used to test whether such differential expression is statistically significant. Different conditions in an experiment can be defined, for example: samples of normal cells versus pathological, cell samples before and after the effect of a factor (eg in Nutrition studies, this factor may be a nutrient), different phases of the cell cycle of the same cell sample etc. Determination of differentially expressed genes is particularly useful. For example, in experiments in which the studied conditions are "normal tissue" and "cancerous tissue", finding statistically significant differential expression between the conditions may indicate genes as features that can be used diagnostically using classification techniques. . Note that statistical hypothesis testing is an appropriate analysis approach to address peculiarities of microarray data, such as random noise. When a difference is statistically significant, it means that it is not due to chance, but has a biological background.

To calculate the differential expression of a gene between two states (conditions) of the experiment, the logarithm (based on 2) of the ratio of the expression values of the gene in the two comparable conditions (also known as logFC - log fold change) can be used. . Using statistical terms, one of the conditions is called the test condition, and the expression value in this condition is placed in the numerator of the ratio, while the other is the control condition, and the expression value in this condition is placed to the denominator. Depending on the value of the logarithm we determine the relative changes in expression e.g. activation, suppression of gene expression [4].

There are various statistical analysis techniques used to find genes with statistically significant differential expression between the states of the experiment. In order to determine the genes that show differential expression between two conditions of an expression experiment (study condition / control condition), the statistical hypothesis test can be used, using the t test (Student's t-test or simply t-test). For

For example, a t-test can analyze an experiment in which we study gene expression in tissue before and after taking a particular nutrient. For more than two comparable conditions in the experiment, the Variance Analysis (ANOVA) can be used.

- Analysis of Variance).

The procedure for checking the statistical hypothesis with t-test includes: the definition of the statistical hypothesis, the determination of the zero and the alternative hypothesis, the calculation of the statistical hypothesis t and finally, the decision to reject or not the zero hypothesis for selected level of importance. In hypothesis testing, we choose the null hypothesis that represents that there is no change in the parameter we are studying. For example, in the analysis of gene expression experiments, we can consider as a null hypothesis (H0) the hypothesis that the mean value of gene expression under the two comparable conditions does not change. The alternative hypothesis (Ha) on the other hand, is the logical opposite of the null hypothesis, that is, the mean expression value of the gene under the two conditions is different. When the alternative hypothesis is formulated in this way, the statistical hypothesis test is called a bilateral test. If the alternative hypothesis is defined as the hypothesis that the mean expression value of the gene in the study condition is only higher or only less than the mean expression value of the gene in the control condition, the control is called unilateral.

Following the statistical case audit process, the audit statistic is applied
t-test and the value of quantity t is calculated, based on the samples we have and assuming that the null hypothesis holds. Based on the value of the statistical control t calculated, a probability p can be found from suitable statistical tables.

(probability), known as p-value. The p-value is defined as the probability that the value of the statistical control t is equal to or more "extreme" than the value calculated on the assumption that the null hypothesis is valid (the definition of what constitutes an extreme value depends on whether the hypothesis test is unilateral or bilateral). If the p-value is less than or equal to the selected significance level, this indicates that the data observed in the samples are unlikely to be in agreement with the null hypothesis and thus, the null hypothesis can be rejected. Therefore, we accept the alternative hypothesis (in our example, we accept that there is a difference in the mean expression value of the gene between the two conditions studied). Otherwise, the null hypothesis cannot be rejected.

In experiments that study the differential expression between two states of the experiment, "volcano plots" diagrams are often used, which are so named because of their characteristic shape. In these diagrams, each point refers to a specific gene and has coordinates, on the horizontal axis, the logarithm of the expression ratio mentioned above (log2FC), and on the vertical axis, the negative decimal logarithm of the p-value (-log10p- value). The higher the (absolute) value of the intersection of a point, the greater the differential expression of this gene, while the higher the value of the ordinate, the more statistically significant this differential expression is. Based on the literature, there are limits that are considered acceptable as limits of differential expression and statistical significance (for example, acceptable limits may be

| log2FC |> 1.5 and p-value <= 0.05). We can thus consider them as differentially expressed (differentially expressed genes) the genes that meet both conditions with respect to these limits [4].

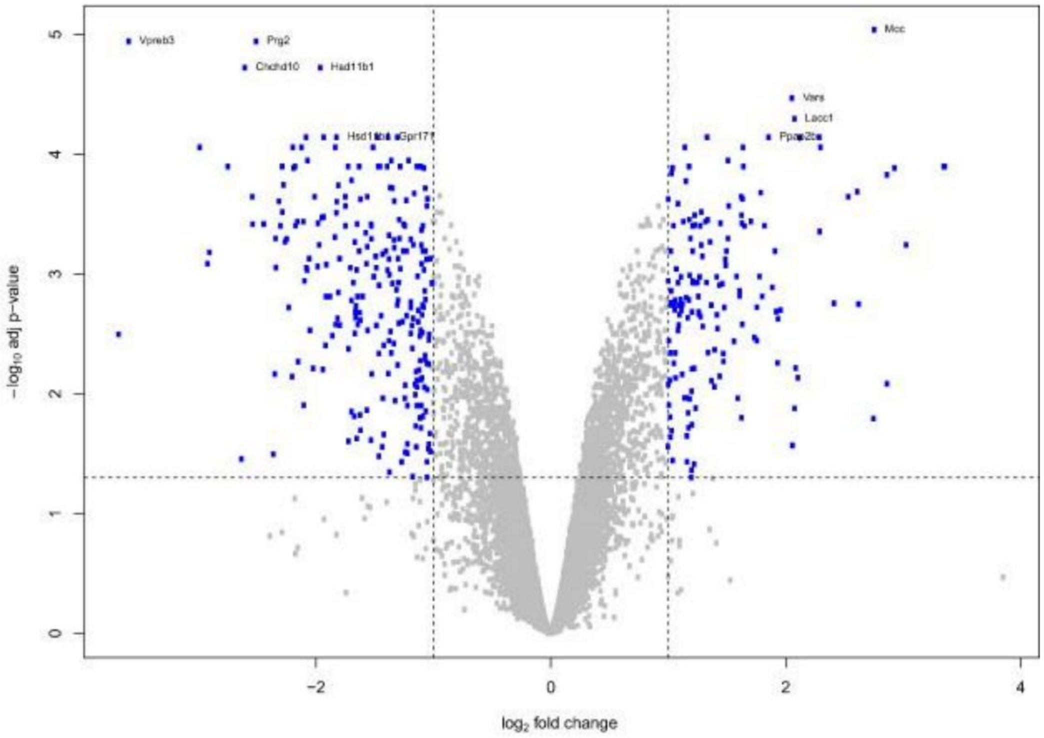

Figure 4: "Volcano crater" diagram from microarray data. The selected limits for differential expression and statistical significance are shown by the dashed lines. The genes shown with blue dots in the graph are considered important for analysis. For the ten most important genes, the gene symbol is also displayed. [6]

Cluster analysis

Grouping the gene expression data found in the format shown in Figure 2 is used to identify patterns in the data. For example, grouping can be done at the sample level (table columns) and allow us to identify unknown / new categories present in our samples (eg if the samples are from cancerous tissue, to identify different types of types cancer in gene expression data). Grouping can also be done at the gene level (table rows). This can be applied to the selection of genes as features (feature selection), which can then be used in prediction models. Still, such a grouping

can be used to identify groups of genes that exhibit co-expression [7]. As co-expression characterizes the co-expression of gene expression under different experimental conditions. Finding groups of genes that co-express is particularly important, because it can be considered as an indication of common function of genes. We can thus assess the unknown function of one gene by observing which other genes it co-expresses with, since we already know the function of the other genes (also called "guilt by association").

Very useful in clustering analysis is the graph called clustered heatmap, also known as double tree diagram, which can also be obtained by hierarchical clustering. Thermal maps are often used in the analysis and visualization of multidimensional data and are often used in large-scale experiments with microarrays, as they facilitate the finding of trends and patterns. More specifically, the thermal map is a table that in each cell depicts the level of expression of the corresponding gene in the corresponding sample, using a color scale. In addition, the columns and rows in the table are arranged so that genes and samples of similar patterns are grouped next to each other. In this visual way we can identify possible correlations between experimental states and expression patterns, as well as groups of genes with common expression patterns.

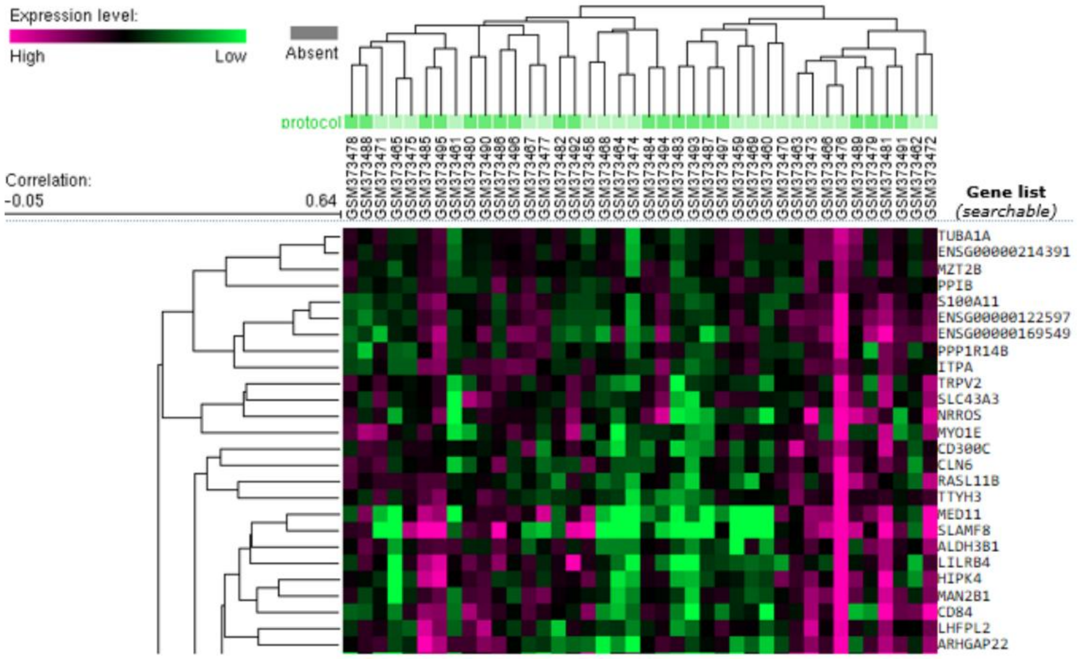

Figure 5: Detail of a thermal array thermal map obtained using NCBI GDSbrowser Web Application.

5.5 Nutrigenetics / Nutrigenomics and analysis of gene expression data

Nutrigenetics is defined as the field of research that studies how different
Genetic characteristics of individuals can lead to a different response to the diet and consequently to a different phenotype. For example, people who follow the same diet may, due to different genetic material, respond differently to serum cholesterol levels [8].

The term Nutrigenomics, on the other hand, refers to the field of research that "deals with the characterization of all gene products that are affected by nutrients and their metabolic effects" [9]. In other words, it deals with how nutrient intake affects the gene response: RNA transcription,
coded proteins etc.

Nutrigenetics and Nutrigenomics (Nutrigenomics) can be used in the context of the so-called personalized diet, where the personal diet is selected based on the individual's genome, with the aim of promoting health and preventing / managing chronic diseases [8]. To do this, we need to understand how the numerous interactions between nutrients, genes, proteins and metabolic pathways affect biological disease pathways.

In the context of Nutritional Genomics, appropriate studies can be designed with experiments that examine the effect of nutritional factors on gene expression. Gene expression can be measured with DNA microarrays and the analysis can follow the steps described in section 3.4 of the exercise.

5.6 Practical part

The laboratory exercise will analyze data from DNA microarrays retrieved from the NCBI Gene Expression Omnibus (GEO) public repository.
(www.ncbi.nlm.nih.gov/geo). The analysis is performed using the R programming language and specifically through the graphical user interface (graphical user interface) RGui.
In addition, some of the tools available on the NCBI website are used. In order to answer the following questions, you will need to consult the code / comments given in the Exercise6.R file, which is located on the MyCourses exercise page.

The programming language R

The R programming language is free software / open source software that
is widely used in the field of Bioinformatics / Computational Biology, as there are a variety of packages / libraries available in R specifically for the processing of biological data.
Also, the use of R is very common in the field of Data Science.

You can download R from www.r-project.org/ . The latest version of
R is 3.5.1. There are plenty of guides for R on the internet, as well as information you can find from user manuals for each package and function. The link www.rdocumentation.org is especially useful . You can also consult the introductory guide from the official page, at cran.r-project.org/doc/manuals/r-release/R-intro.pdf . It is useful to familiarize yourself with the graphical interface and features such as:

- definition of working directory with *setwd (),.*

Installation and "loading" of packages. There are many packages available on CRAN - Comprehensive R Archive Network. The installation of a package can be done with the command *install.packages ("thepackagename") .* To be used

the package after installation, must be loaded using the command
library ("thepackagename"),

- management of vectors and arrays in R,
- use of repetition structures (loops) in R.

Analysis of gene expression experiment data

The data to be analyzed are from a gene expression experiment with a DNA microarray. This is the experiment with GEO id: **GSE7117** and entitled "Gene expression in the liver after a low-calorie diet in obese women and obese controls". This experiment is part of the field of Nutrigenomics, as it deals with how a specific diet can affect the metabolic profile and hepatic gene expression in humans. ***Data retrieval and review***

The data for the GSE7117 experiment to be analyzed is stored in the form of a microarray data file called a "Series Matrix File" (.txt file). This format is very easy to use, as the data has already been pre-processed and normalized and is ready for further analysis. Of course, for each experiment in GEO there are experimental data available in various formats, as well as the original raw data.

Question 5.1: Search the NCBI Gene Expression Omnibus website,
in the Accession Display service, the experiment with GEO accession id: GSE7117 (at www.ncbi.nlm.nih.gov/geo/query/acc.cgi). After reading the summary of the experiment (Summary) state how many samples were counted, what are the comparative situations in the experiment and how many samples belong to each situation. Also mention the type of platform
DNA microarray with which the measurements were made.

Request 5.2: Download the GSE7117_series_matrix.txt file from the MyCourses exercise page. As mentioned, the data available in the Series Matrix File format has already been normalized. Open the .txt file with an .txt file editor and provide the following information about the experiment:

i. What is the title of each sample and its Sample_geo_accession id?
ii. What kind of organism do the samples come from?
iii. To which of the two comparative conditions of the experiment does each sample belong?

Use the read.table () function in the Excercise6.R file for Request 2 to assign the part of the .txt file that contains the experimental measurements to the data frame (matrix-like object) named x.

iv. Explain what the arguments given in the code for read.table () mean. v. Use the appropriate command and list the dimensions of x.
vi. Use the appropriate command and list the names of each column in x. What do the different columns correspond to?
vii. Use the appropriate command and list the names of the first 15 lines of x. What do the different lines of x correspond to?
viii. What is the value recorded for element x [3,5] and what does this value describe?
ix. What information does the 200th row give us and what does the 7th column of the x table give us?
x. Use an appropriate R function to plot the frequency histogram for the measurements recorded for the first and third samples of the experiment.

Data normalization

Requested 5.3 :

i. Execute the command in the code for Request 3-i and list the result. What does the result show? Explain the functions and arguments used in this line of code. ii. Check that the data is indeed normalized by creating boxplots for the measurements of each sample. Consult the comments provided in the code for this request. List the resulting diagram and comment.

Determination of differentially expressed genes

Question 5.4: In this question you will do a statistical hypothesis test with t-test to determine the differentially expressed genes between the samples without diet intervention (controls) and with diet intervention (diet intervention), filling in the code for the question.

i. Properly fill in the function c (), which assigns to the vector "Xsamplelabels" values that give the experimental condition for each sample. Use 0 for control samples and 1 for dietary samples. ii. Use the t.test () function of R and write code to test t-test case for each detector for the two experimental conditions reported. Calculate the p-value value for each probe (gene) and list in your report the p-value values for the first 15 microarray detectors. iii. Write code and list the detectors for which the p-value you calculated with the t-test was less than 0.001. Also, find the detector with the lowest p-value.

Question 5.5: Perform differential expression analysis this time using NCBI's GEO2R service at www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE7117. Instructions for use can be found on the GEO2R page (link www.youtube.com/watch?v=EUPmGWS8ik0). First determine which samples correspond to the two states of the experiment: without diet intervention / control intervention group, and identify the first 250 most important genes that are differentially expressed, using the relevant service offered.

i. List all the results extracted for the first two most important genes resulting. What do you have to say about the p-value and logFC values they have, based on what is mentioned in the theoretical part? Is the expression of both genes suppressed or enhanced? Select on the line for the first probe (gene) and list the graph shown. What exactly does it show?

ii. Provide some brief information about the gene that you found to be differentially statistically significant with GEO2R by searching for the Gene link https://www.ncbi.nlm.nih.gov/gene. Look for information about the human gene (Homo Sapiens). in service of NCBI, in

Data grouping

Requirement 5.6 .: Create a heatmap with the data from the experiment GSE7117, using the NCBI Dataset Browser service at www.ncbi.nlm.nih.gov/sites/GDSbrowser (Cluster Analysis option).

i. How many rows and how many columns does the complete thermal map have?

Focus on the top of the thermal map to see the details, following the instructions on the website.

ii. List part of the detail of the resulting thermal map to show the basic information it contains. What is displayed in each row and column and what additional information is given? iii. Based on the color code given, how would you characterize the expression value in

first gene in the list for the first sample?

iv. Based on the resulting grouping, can you make any observations about possible correlations between the states of the experiment and the expression patterns of gene groups? Justify your reasoning.

In your report you should include the answers to all the above Requests 5.1. to 5.6. and their sub-questions, the relevant graphs requested, as well as the completed code of the Exercise6.R file.

Bibliography

[1] Tsaousoglou, M., Beri, D., Vgontzas, A., Chrousos, G., "Molecular mechanisms of circadian rhythms: a study in experimental animals and the first indications in humans", Bulletins of the 1st Pediatric Clinic of the University of Athens, no. 53, 2006.

[2] Vieira, E., Burris, T., Quesada, I., "Clock genes, pancreatic function, and diabetes", Trends in Molecular Medicine, vol. 20, no. 12, 2014.

[3] Franks, P., Merino, J., "Gene-lifestyle interplay in type 2 diabetes", Current Opinion in Genetics & Development, no. 50, pp.35-40, 2018.

[4] Nikolaou, Ch., Houvardas, P., "Computational biology", Athens: Association of Greek Academic Libraries, 2015. [electr. book] Available at: http://hdl.handle.net/11419/1577

[5] www.bioconductor.org, "MalariaLifeCycle - Statistical Microarray Analysis using affylmGUI "Available from: www.bioconductor.org/packages/devel/bioc/vignettes/affylmGUI/inst/doc/LifeCycle/Malaria LifeCycle.html # RawIntensityBoxPlot

[6] Xie, Ping & Moore, Carissa & R. Swerdel, Mavis & Hart, Ronald, "Transcriptomic profiling of splenic B lymphomas spontaneously developed in B cell-specific TRAF3-deficient mice", Genom Data. 2. 386-388. 10.1016 / j.gdata.2014.10.017, 2014.

[7] Cancer Research UK Cambridge Institute, "Microarray-analysis, Materials on the analysis of microarray expression data; focus on re-analysis of public data ", 2018. [online] Available from: http://bioinformatics-core-shared-training.github.io/microarray-analysis/

[8] Simopoulos AP, "Nutrigenetics / Nutrigenomics," Annu. Rev. Public Health, vol. 31, no. 1, pp. 53–68, 2010.

[9] Ordovas JM, Ferguson LR, Tai ES, Mathers JC, "Personalized nutrition and health," BMJ, vol. 361, pp. 1–7, 2018.

[10] University of Barcelona, “Introduction to Microarray and Next Generation Sequencing Data Analysis, Practicals ”, 2018. [online] Available from: http://www.ub.edu/stat/docencia/bioinformatica/microarrays/ADM/