**APPLIED BIOINFORMATICS**

**PROJECT REPORT**



Mario L. Martin

Student ID: 130020616

BS32010: Applied Bionformatics

March 2014

During the present project, analyses of two different gene expression experiments data have been performed. The first experiment was a microarray expression profiling, and the second one was another expression profiling performed using RNA-sequencing. The data obtained from these experiments consisted in gene expression levels from 4 replicates of untreated and untreated tissues.

The different analyses consisted on, given the files for each experiment, identify, using RStudio-based programming, the most differentially expressed genes. Once these had been indentified, a search for homologous genes in other species was done and a phylogenetic analysis, including the building of a phylogenetic tree, was performed, in order to see the relationship between all the homologous genes and identify which homologues of which species were more closely related to our experimental specie (*Gallus gallus*). Finally, one of the more closely related homologues for each most differentially expressed gene was taken and then a BLAT search against the *Gallus gallus* genome and some closely related species genomes was carried out using the ENsembl online tools (BLAT) to find the relevant area of the chicken genome and the orthologous regions in the other genomes. Finally, a comparison was done between these two genome regions in order to see the differences in the chromosome localization, presence of more than one copy of the gene per genome and other relevant data.

***Experimental Data Background***

One of the most exciting discoveries in the last decades is the detection, at molecular level, of the different components of what have been called “molecular clocks”. Molecular clocks are biochemical mechanisms presents in all known cells (prokaryotes, eukaryotes and archaea) that regulate different cellular events by defining a constant rhythm in which a given process is cyclically performed. The duration of each cycle depends on the process and the molecular clock which regulates that process.

The circadian clock, for example, is a mechanism that oscillates with a period of 24 hours, and it is highly important in all organisms, from bacteria to higher mammals, as it coordinates the day-night cycles. In some photosynthetic bacteria, this is very useful in order to coordinate the expression of photosynthesis-related genes during the day and nitrogen-fixing genes during the night, in order to save energy (as photosynthesis can be done only during the day) and also avoid unwished cross-reaction between both processes (as the oxygen generated during the photosynthesis easily deactivates the nitrogenase complex used for nitrogen fixation). In higher mammals, like humans, the circadian clocks are highly important to regulate the sleeping, brain processes, metabolic pathways, body temperature and other biological activities; and its deregulation is related with a lot of different illness, such as obesity or sleeping disorders.

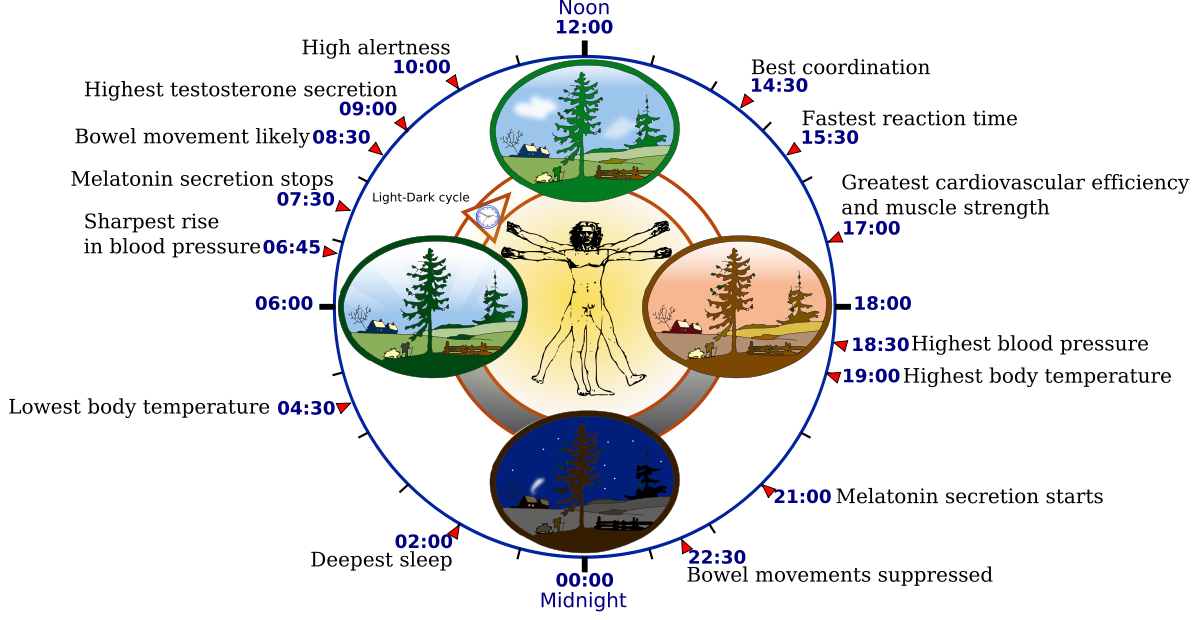


Figure 1: Skematic diagram showing some of the relevant processes regulated by the circadian cycle in humans. Source: http://en.wikipedia.org/wiki/Circadian\_rhythm

During the formation of the vertebrate body, at the same time the mesoderm, ectoderm and endorderm are being formed; the paraxial mesoderm (which is how the mesoderm at the side of the neural tube is called) starts to separate into bilateral blocks called **somites**, which will give rise to skeletal muscle, dermis, tendons, cartilage and endothelial cells.

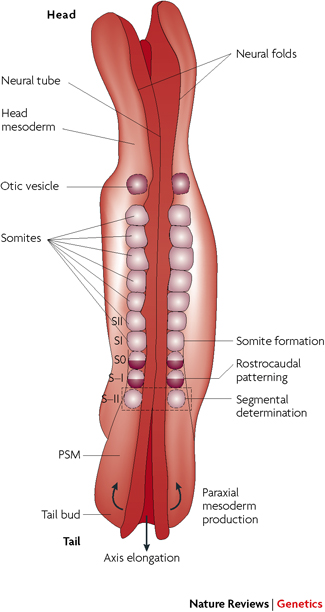


Figure 2: Skematic Dorsal view of a 4-week-old human embryo showing somites and the presomitic mesoderm (PSM) forming the paraxial mesoderm that flanks the axial neural tube.

## Source: Mary-Lee Dequéant & Olivier Pourquié. (2008) Segmental patterning of the vertebrate embryonic axis. *Nature Reviews Genetics* 9, 370-382

It has been shown, from previous experiments in chicken (*Gallus gallus*), that the formation of each somites (somitogenesis) it is a highly regulated process that follows a 90 minutes cycle. This hour and a half oscillation relies in the negative feedback of mRNA and proteins of the different genes involved directly (as a part of the somitogenesis) or indirectly (as regulators) in the process, as well as the short half lives of the key components (1).

The research where the data comes from is focused in one of the key components which regulate the timing of this cycle, a CdK (Cyclase Dependent Kinase) inhibitor called roscovitine. CdKs are the core of the cell cycle regulation and important transcriptional regulators. Roscovitine inhibits these kinases, altering the expression and regulation of other key cell components and therefore altering the duration of the cell cycle (1). The activity of this molecule plays an important role during the morphogenesis of the vertebrate body and the somitogenesis. Roscovitine, as cell cycle disruptor, has also been shown to play key roles in apoptosis, and that is why it is being studied as a drug candidate for the treatment cancerous cells in different types of cancer.

In the present study, 8 chicken early-stage embryos were treated or untreated with roscovitine (4 replicates x treated + untreated samples). After that, two different techniques were used to obtain the gene expression profiles: a microarray (Affymetric GeneChip array) and RNA sequencing. The data from these analyses was the starting point for the present project.

***Overview of the project***

The aim of the present project is integrate the different bioinformatics techniques learned and used during the module in a practical exercise using real data from real experiments where all the skills learned can be applied.

The first part of the project will consist in, given the data from the microarray/RNA Sequencing experiments, use R programming to write the proper scripts that will allow to obtain the most differentially expressed genes between the treated and untreated samples for each set of data. For each technique set of data, different packages specifically designed for this kind of bioinformatical analysis, which will be explained in detail in the corresponding sections.

From these analyses, two lists of genes will be obtained, one for each technique, corresponding to the most differentially expressed. As the techniques used were different, as well, as the RStudio packages, normalization processes and R scripts, the genes obtained can be very different between the two techniques and even between two analyses of the same technique data.

Two genes from these lists were selected for the next steps, in order to work with more than just one variable and be able to work with more than one point of view, but at the same time using a reasonable low number of genes, as some of the scripts in the next steps took a long time to run. In future projects or real-work analyses, more genes should be selected (as many as possible) in order to have a wide point of view of the results, as well several different points of view to compare.

For each of the two genes, the mRNA (the reason why mRNA sequences were used will be explained below in the corresponding section) was searched in NCBI Nucleotide database and a MEGABLAST search against the same Nucleotide database was performed. Repetitive sequences (such as different splicing versions of the same gene) were removed and the rest of the sequences were downloaded in a FASTA file.

The sequences of each gene were aligned using two different multiple sequences alignment algorithms: MUSCLE and Clustal Omega. These alignments were phylogenetically analyzed using RStudio and the proper packages and scripts in order to build two different phylogenic trees (one for each alignment) for each gene. Each phylogenetic analysis included fitting each tree with the most suitable nucleotide substitution model, as well as a bootstrapping in order to know how good the final fitted tree for each alignment of each gene was.

Finally, the trees generated were compared with the ones that the NCBI and ENSEMBL generate by themselves, in order to observe the main differences and similarities between them and discuss them and their causes.

Lastly, to close the project, a close homologue of each of the two selected genes was selected, and a BLAT search against some closely related to chicken genomes, as well as the own chicken genome, in the ENSEMBL database was performed. This retrieved a genomic map for each of the genomes indicating regions with homology for our sequence (the closely related homologue gene), as well as the region in the genome were the gene’s homologues are located. A comparison between the different genomes was performed in order to see if the location of the selected genes was conserved between the different species selected (syntenia) or not, and discuss these results.

A summary map of the project can be seen in figure 3.

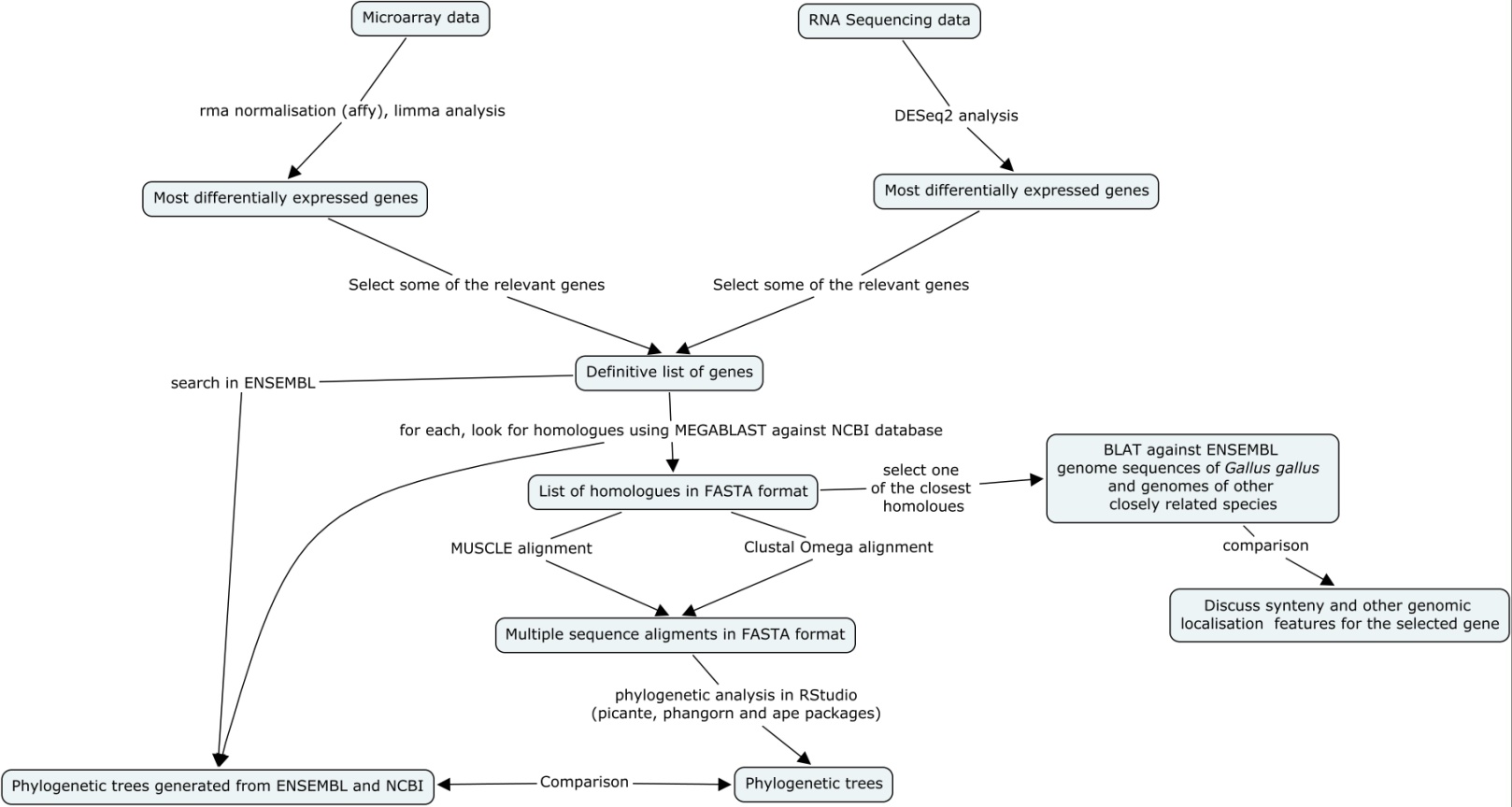


Figure 3: Summary map of the Project. Includes all the relevant points and a summary of the analyses performed in each step.

***Identification of the most differentially expressed genes***

The first step for the data analysis was identifying the most differentially expressed genes (between treated and untreated samples) for each set of samples from each technique (microarray and RNA Sequencing). For both set of samples, programming in RStudio was used to extract the desired data but, as the two techniques are pretty different, as it is the format of the output files for those, different scripts and packages were used for each technique data, which will be described below in the corresponding sections.

RStudio (2) is a free open source software that provides an accessible environment for programming using R language (Figure 4). It is used for statistical computing and graphics, being very useful for its intuitive interface, flexibility and the high number of different packages available to perform almost any statistical analysis in any area, from economy to biology.

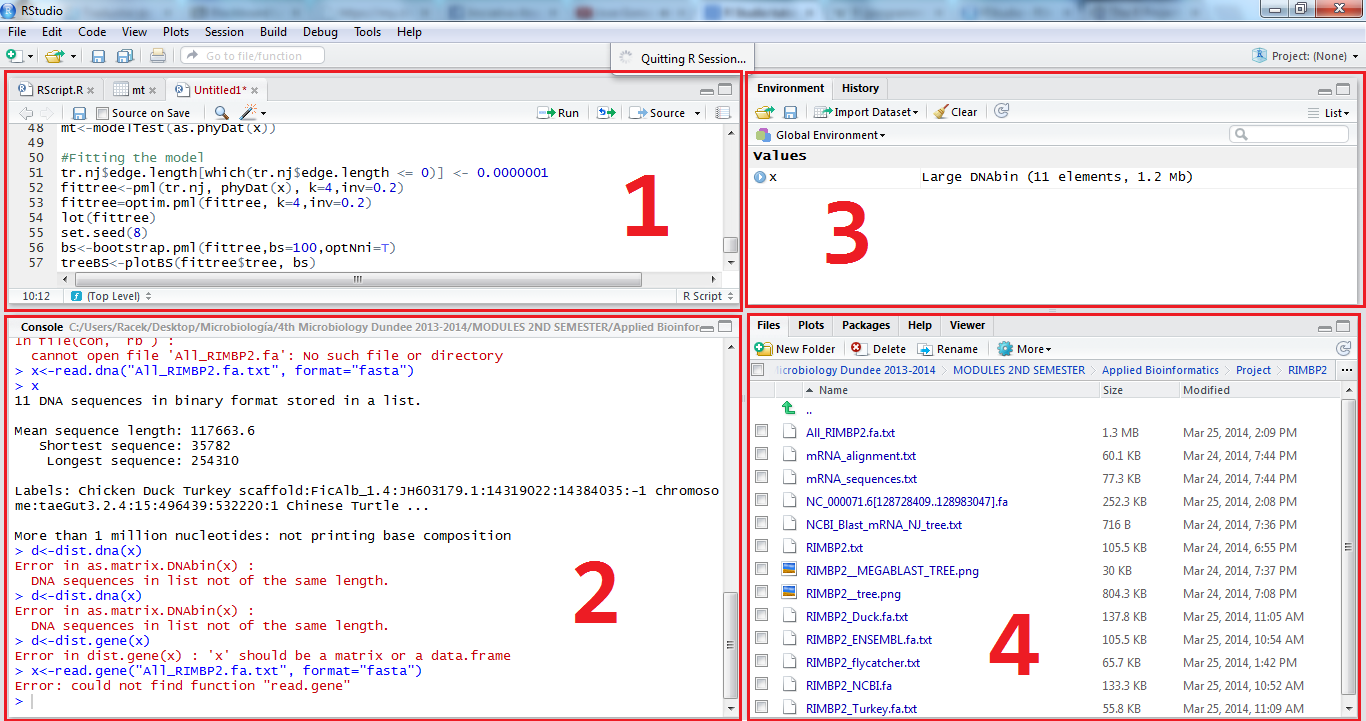
R (3) is a free software programming language specially developed for statistical computing and graphics. It allows the user an easy data manipulation, calculation and graphical display, and can be easily expanded using the different packages developed by different users around the world, that provides specific tools for specific data analyses.

Figure 4: RStudio interface in Windows. 1.- Data window (where data can be visualized and scripts can also be opened/created and edited and then be sent to the console to be executed. 2.- Console (whercomplete definition of their content. The historye the code is executed, either from the console, from a R file or directly writing in the console). 3.- Environment/History (The environment contains all the objects, matrix, lists and values created. The history contains all the commands executed in the console from the start of the session). 4.- Files (shows the files present in the current working directory), plots (contains all the plots created during the session), packages (all the packages being used) and help (where help about the different packages/commands can be found).

*Microarray*

The data obtained from the microarray was stored in eight files, one for each replica (four) and treatments (treated with roscovitine or untreated).

*RNA Sequencing*

***Homologues search***

Once the lists of most expressed genes for each technique were obtained and the two genes to perform the rest of the analyses were selected (MUC6 and MMEL1), the next step was identifying, for each of them, homologous genes to perform a phylogenetic analysis which includes a tree building to visualize the phylogenetic relationship between all the homologues.

*First approaches and problems*

The first idea to identify the homologues was perform a search using the gene symbol in the Nucleotide NCBI database. Nucleotide is a huge database that includes sequences from several different sources, including GenBank, RefSeq, TPA (Third-Party Annotation) and others. In other words, any kind of DNA and RNA sequences in all the databases associated to NCBI. The aim of this search was find the DNA sequence for MUC6 and MMEL1 genes to then perform a MEGABLAST search with that sequence against the own Nucleotide database to find homologues.

The problem came with the results for this search, as there was no sequence for the gene, only the whole genome sequence for the *Gallus gallus* chromosome 21 and 5 (where MMEL1 and MUC6 are located, respectively), and the only sequence that could be obtained from them was the whole loci sequence, which implies working with sequences of more than 100.000 base pairs, which make the posterior analyses (alignments, tree building) virtually impossible to perform in the available computers, as well as retrieving different results than the ones that would be found working only with the gene sequence, as the whole loci may contain unrelated regions that can change more during the evolutionary process, changing the final results from the ones that would be obtained working only with the gene sequence. The other sequence found with this search was the mRNA predicted sequence for each gene, but that was initially discarded, as the main objective was working with the original gene DNA sequences.

The second idea was search the gene in the ENSEMBL database, which does contain the DNA gene sequence. The sequences for both genes were downloaded and, using them MEGABLAST search against Nucleotide was performed (just in case that the sequences were in the database, but with different names as they could have not been classified yet), but, as expected, the only sequences found were the *Gallus gallus* whole genome chromosomes sequences (as the loci for each gene in those sequences obviously matched with the MMEL and MUC6 sequences from ENSEMBL) and mRNA sequences for the MUC6 and MMEL1 of *Gallus gallus* and other related species.

As none of them were DNA gene sequences, a second search was performed in the ENSEMBL database. The entry for each gene (in our case, MUC6 and MMEL1) contains a lot of information about the sequence, location, variants, expression and also phylogenetic relationships with homologues genes found in the ENSEMBL database (Figure 5). Using this information, some of the homologues were searched and picked up from the ENSEMBL database and stored in a FASTA file in order to perform the following steps.

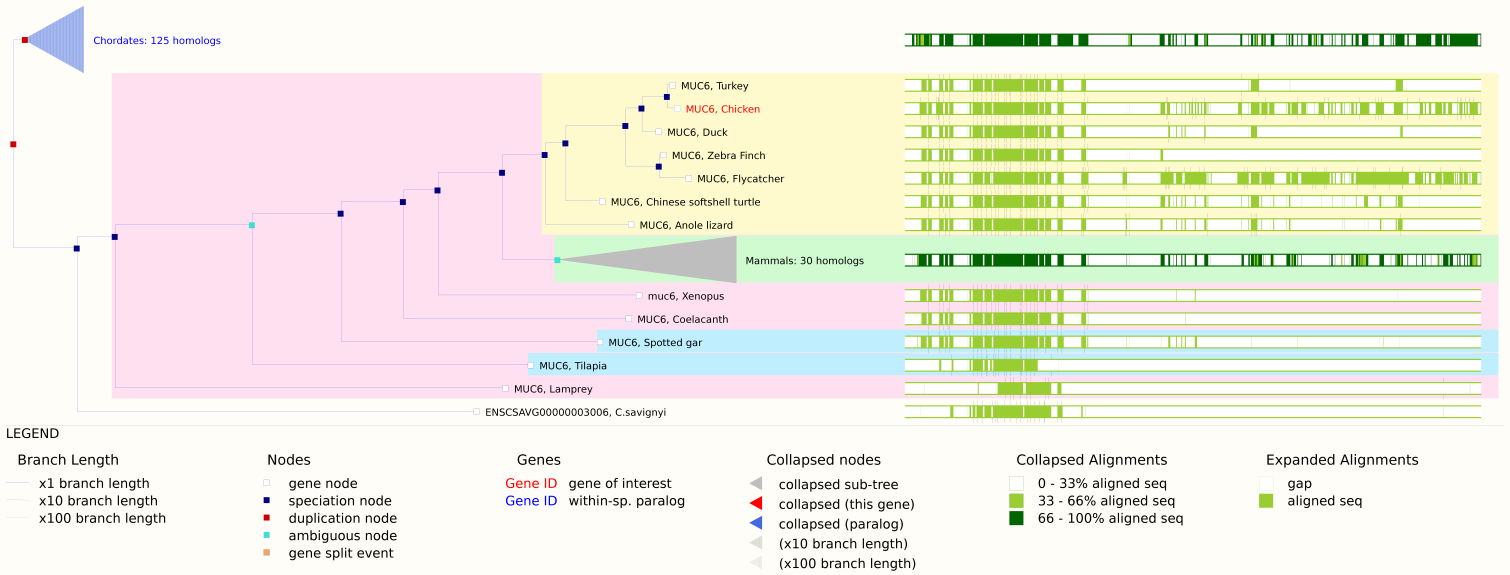


Figure 5: Phylogenetic tree built by the ENSEMBL site for MUC6 using the homologous genes contained in their database, which shows the evolutionary relationship between the. As the algorithms and techniques used to build this tree were not specified, it was going to be used only as a reference to pick homologues and then perform an own multiple sequence alignment and tree building and confirm or not these phylogenetic relationships. Unluckily, as said, the sequences in the database were not suitable for the alignment methods available.

Nevertheless, when the next step was tried to be carried out, it was seen that these sequences were not also suitable for the phylogenetic analyses. It was tried to perform a multiple alignment of the two set of homologous sequences using two different algorithms (MUSCLE and Clustal Omega, that will be detailed below), but the analyses were retrieving an error message, indicating that the process could not been completed. After some trials and errors, it was realized that some of the sequences downloaded were uncompleted (in other words, some regions in the sequence contained regions with repetitive “NNNNNNNNNNN”, which indicates that this part of the sequence was not available yet) and that was the reason the multiple alignment software crashed when they tried to align the sequences. The first idea to solve this problem was remove those sequences, but that implied working with a very low number of homologues, which would have made the analyses , under the point of view of the student, not enough statistically significant.

*Working with mRNA sequences*

The solution chosen was working with the mRNA sequences found during the first search against Nucleotide. Those sequences were all completed, with no gaps, and, as they are the RNA version of the gene sequence, the phylogenetic relationships should be highly similar, if not identical.

The mRNA sequence of MUC6 and MMEL1 was MEGABLASTed against the Nucleotide database, and a list of homologous mRNA of MUC6 and MMEL1 in other species was retrieved. The repetitive sequences from the same species were removed (as in some cases there were slightly different splice variants –but highly related between them, so all the splices would be together in the phylogenetic tree-) and the rest were downloaded in a single FASTA file for each gene. This file would be used for the following phylogenetic analyses. NCBI has an own tools to generate unfitted Neighbour–Joining (NJ) trees using the MEGABLAST data (which includes %identity, e-scores, alignment scores and other relevant data). This tree was also downloaded in order to be compared with the trees that would be generated in the following steps.

***Phylogenetic analyses***

*Multiple sequence alignment*

Once we had the homologues list in FASTA format, the main objective was performing a phylogenetic analysis for each set of sequences in order to see and discuss the evolutionary relationships between them.

The first step for this was performing a multiple sequence alignment (MSA). There are several ways to do this, including a large set of different algorithms. For the present project, two of the most popular algorithms were used: MUSCLE (multiple sequence comparison by log-expectation) and Clustal Omega. The decision of using two different algorithms was done because they are the most popular methods for each of the two major alignment approaches used: progressive alignment (Clustal) and iterative method alignment (MUSCLE). This would also useful to check if the relationship between the different homologues is “strong” or “clear” enough to be the same or very similar using both methods, or if the two methods retrieved alignments different enough to change the final phylogenetic tree (as the method used to build the phylogenetic tree would be the same).

Clustal is probably the most popular method used for multiple sequence alignments, which uses a progressive alignment construction, also known as hierarchical or tree method. This is a heuristic method that has two main stages: the building of a initial guide tree using methods such as UPGMA or NJ; and after that, the multiple alignment process, where each sequence is added to the growing MSA according to the guide tree (in other words, one sequence is used as starting point, and it is aligned with the most closely relates according to the guide tree, then the next sequence is added and aligned, and the same for the rest of them). The main problem of this method is that, as they use a guide tree as a reference, any error in the building of this tree would be extended to the rest of the alignment. The version used for this project is Clustal Omega, the newest version available, which contains secondary methods to ensure the quality of the alignment.

MUSCLE is becoming most and most popular in the last times. It is a iterative method, which means that it realigns the initial sequences at the same time new sequences are added to the MDA, which allows a better optimization of the final result. The strong point of MUSCLE is that it has a very accurate distance measurement, which makes the final relation value between the different sequences more reliable. These are the reasons why MUSCLE (as well as other iterative methods) is considered, in a general way, a more accurate method than Clustal and other progressive alignment methods.

The alignments for both set of data and both methods were performed using the EMBL-EBI website (4), as it has a very intuitive interface and supports a lot of different formats for the input and output files. It is also possible to change the default options, allowing the user to personalize the alignment. For this project, the default options were used, as it was considered that the number of sequences and the size of each of them did not required any kind of special treatment to improve the alignment.

From this step, 4 files in FASTA format were retrieved, containing the different alignments for each method and set of homologues. These files would be used to perform the phylogenetic analysis and tree building in the following steps.

*Distance matrix and tree building and optimization*

The final step

***Gene loci comparison between Gallus gallus and closely related species for MUC6 and MMEL1***

A last analysis to be performed for this project was a comparison of the location of MUC6 and MMEL1 in the chicken genome and other genomes of closely related species. The species selected for this analysis were the two of the most closely related according to all the phylogenetic trees: turkey (*Meleagris gallopavo*) and zebra finch (*Taeniogypia guttata*). The first intention was use the duck instead of the zebra finch, as it is phylogenetically closer to the chicken, but there was no karyotype available for it in ENSEMBL, which make the results more easily understandable and easy to interpret. Moreover, after the analysis, it was considered that having a species which a bit more evolutionary separated from *Gallus gallus* would be more useful to discuss the synteny between them.

The analysis was performed as follows: one the closest homologous sequence for each gene was downloaded from ENSEMBL, and then a BLAT against the three species genomes in ENSEMBL was performed. The result page contained the information that would be discussed.

*MMEL1*

# *References*

1. **Schofield, Pietà.** [Online] http://www.compbio.dundee.ac.uk/user/pschofield/Teaching/Bioinformatics/Slides/Assignment.pptx.

2. **Allaire, JJ., Kawak, Tareef and et al.** http://www.rstudio.com/. *RStudio.* [Online]

3. **Gentleman, Robert and Ihaka, Ross.** http://www.r-project.org/. *R-Project.* [Online]

4. **Dublin, UCD.** http://www.clustal.org/. *Clustal.* [Online]

5. **drive5.** http://www.drive5.com/muscle/. *MUSCLE.* [Online]

6. **Institute, European Bioinformatics.** http://www.ebi.ac.uk/Tools/msa/. *EMBL-EBI Multiple Sequence Alignment.* [Online]