List of suggested projects

Below are 22 different projects classified according to the topics learned in class.

It is strongly recommended that you first go through all topics and then select a specific project. Remember! You are also welcome to choose your own project but you must consult with us on time!

Projects that require programming are marked: * scripting (R markdown), ** coding

Note that the programming level and extend may vary according to the decision of the team.

Next Generation Sequencing analysis
(DNA sequencing, Gene expression, microbiome, single cell…)

* Project 1 - Analyzing RNA expression through time and space

Gene expression is a dynamic process that changes through time and space. Expression patterns of different cells or tissues may vary due to different timing of biological processes and/or due to the morphology of the organism. In this project, you will investigate a dataset of gene expression where samples represent different time points of the process or are extract from different spatial positions of a tissue/organism.

Project 2 - Identifying disease related mutations from NGS data

One of the major applications in Next Generation Sequencing (NGS) is to detect disease-causing mutation. To this end, the genomes of healthy and affected individuals are sequenced and the researchers need to identify the differences, which are correlated with the disease.

** Project 3 - Identifying Monoallelic Expression in RNA sequencing data

Monoallelic gene expression (also called allelic exclusion) is a process in which transcription occurs from only one of two homologous alleles in a diploid cell. Monoallelic expression is responsible for several biological phenomena including: X-inactivation, Genomic imprinting, Monospecificity of B lymphocytes, Regulation of olfactory receptors expression, Phenotypic diversification by random autosomal monoallelic expression. In this project, you will develop a strategy to detect monoallelic expression from NGS data (e.g disease, development etc ...)

* Project 4 - Characterizing alternative transcript distribution

Alternative splicing is a mechanism that results in a single gene coding for multiple proteins by using different combinations of exons. The different proteins may have different biological functions. In this project, you will investigate alternative splicing differences between several biological conditions.
**Project 5 - Studying the influence of sequencing coverage/replicates on differential analysis**

When designing a sequencing experiment there is many times tradeoff between sample coverage and number of experimental replicates. In this project, you will investigate the tradeoff by simulating experiments with different coverage/ number of replicates and test how these parameters affect the power to detect differentials.

**Project 6 – Studying the impact of antibiotics on the gut microbiome**

Traditional microbiology relies upon cultivated clonal cultures. Metagenomics is the study of genetic material recovered directly from environmental samples. In this project, you will study the effect of antibiotics treatment on gut microbiome, using metagenomics data-analysis tools.

*Project 7 - Single cell RNA-Seq analysis*

The advancements in NGS sequencing enabled new techniques for analyzing nucleic acid from single cells. Single cell RNA-Seq is becoming more prevalent in recent years. These techniques allows us to study the variability of cell populations. On the flip side, single cell RNA-Seq yields datasets that are sparse and have low signal-to-noise rates. In this project, you will study the properties of single-cell RNA-Seq data. You will compare single-cell data to bulk measurements, experiment with different data imputation techniques and analyze the cell-to-cell variation within a dataset.

**Project 8 - de-novo genome assembly**

Assembling an unknown genome is a complicated task that is affected from the biological complexity of the genome and the technical quality of the data. In this project, your aim is to assemble a genome from DNA-Seq data. Starting from short reads you will use the classical de-bruijn graph approach or add long reads to enhance the assembly quality.

**Project 9 - Comparing between supervised and unsupervised approaches to classify gene expression profiles of cancer patients**

As we have learnt in class, there are many different methods to classify gene expression data, these methods can roughly be classified to supervised and unsupervised methods. As the data is more complex, the problem of classification becomes harder. In this project, you will explore different methods for classifying gene expression data of cancer to find the best performing method.

For your project, you will have to choose the type of cancer you are interested to study from GEO and the classification you are interested to perform (e.g, normal vs tumor, stage of the disease etc.). You will compare the clustering methods and conclude which is best for your dataset.
**Project 10 - Predicting miss-regulated miRNA in cancer and their target genes**

MicroRNAs (miRNAs) are small non-coding RNA molecules that function in RNA silencing and post-transcriptional regulation of gene expression. In this project, you will work with miRNA expression data from cancer and normal tissues. The aim of the project is to identify the miRNA that are miss-regulated in the cancer samples and predict their target genes.

**Evolution, phylogenetic trees and structural bioinformatics**

**Project 11 - Evolution of enzymes (sequence vs structure)**

Enzymes are large biological molecules responsible for the thousands of chemical inter-conversions that sustain life. While their function is kept through evolution, the sequence and structure of enzymes differs between organisms. In this project you will study the evolution of enzyme across different organisms, representing the tree of life using sequence and structural comparison approaches.

**Project 12 - Phylogenetic trees based on sequence and structure of ribosomal proteins**

Phylogenetic trees are usually built based on rRNA sequences. Ribosomal proteins are also found in all living organism but less commonly used for building phylogenetic trees. In this project you will construct three phylogenetic trees: (1) ribosomal proteins structure-based, (2) ribosomal proteins sequence-based, (3) rRNA sequence-based (control tree) and explore the differences to conclude what type of data is best to generate reliable trees.

**Project 13 - Phylogenetic trees based on sequence and structure of your desired protein**

The classical way to examine protein homology is via sequence comparison. However, as we learned in class the structure of a protein is usually more related to its function. In this project, you will examine the relationship between sequence and structural conservation by building a phylogenetic tree based on sequence/structural alignments. You will construct different phylogenetic trees: (1) structure-based, (2) sequence-based, and compare to rRNA classical trees (control tree).

**Project 14 - Predicting gene fusion events in evolution**

In evolution genes evolved by fusion of protein domains, however there are also examples of genes that have evolved by simple fusion of two Open Reading Frames (ORFs). Fusion of genes can be beneficial in bacteria and viruses that have selective pressure to keep their genome small. In this project your aims is to identify fused genes in an organism of interest and find their different origins.
Project 15 – Predicting the structural and functional features of hypothetical proteins.

Bioinformatics tools are used to predict features of newly discovered proteins and to give the researchers a clue regarding the possible function of the protein. Your aim in the project is to use existing tools to explore metagenomics data. You will select hypothetical proteins from the data and attempt to characterize them based on their predicted structures.

**Motif search**

**Project 16 - Predicting RNA motifs in folded mRNAs**

RNA is a single strand nucleic acid sequence however it is usually found in the cell in a folded structure. While there are several known RNA structures which relate to functional elements such as tRNA and miRNA, the structure of most of the RNAs in the cell is unknown. The aim of this project is to translate RNA structures to a graph representation in order to find common RNA secondary structures in RNA data.

**Project 17 - Developing an algorithm for predicting motif in RNA given the sequence environment**

Motifs are short sequences that are highly represented in the data and are expected to relate to functional regions. Motifs of RNA binding proteins are highly degenerative and thus it is very hard to differentiate between functional and non-functional sites. In this project, you will develop an algorithm to identify functional RNA motifs, combining information from the motif environment.

**Biological Networks**

**Project 18 - Using graph representation to identify co-expressed genes based on gene expression data**

As we learned, similar gene expression patterns across samples can indicate different functional relationships between genes. In class, we learnt different approaches to cluster gene expression data in order to identify relationships between genes. In this project, you will use a graph (network) approach to identify gene similarities. For your project you will have to choose a gene expression dataset from public databases (such as GEO), and build a network of genes based on their gene expression similarity.

**Project 19 - Generating transcriptional regulatory networks**

One of the major challenges in biology is to be able to generate networks of interactions between the different components of the cell. A very important type of regulatory interaction required in the genes expression pathway is the interaction between Transcription Factors (TFs) and the genes they regulate. In order to generate regulatory maps of TFs-gene interactions we first have to detect the targets of the TFs. The common experimental approach to date to detect the targets of TF is ChIP-seq - chromatin immunoprecipitation followed by parallel sequencing. In this project, you will use chip-seq data for a specific TF of your choice, and generate and analyze the regulatory network of the TF.
**Project 20 - Predicting transcription regulation based on expression data**

Transcription regulation is a fine tuned mechanism, which is achieved by many transcription factors (TF) which usually bind to the promoter regions of the gene and activate or repress RNA transcription. It is well established that gene which have common expression patterns are regulated by the same transcription factors. Based on this concept it is possible to predict interactions between transcription factors and their target genes and generate cellular regulatory networks. In this project, you will build regulatory networks between TFs and target genes based on gene expression data.

**Project 21 - Constructing lncRNA-miRNA interactions network**

MicroRNAs (miRNA) are a class of small noncoding RNA (~22 nts) that play a central role in posttranscriptional regulation of gene expression by mRNA cleavage, translational repression or mRNA destabilization. Long noncoding RNAs (lncRNAs) is a large and diverse class of noncoding transcripts usually longer than 200nt. The function of many lncRNAs is unknown. Previous studies have shown that lncRNAs act as miRNA sponges, reducing their regulatory effect on mRNAs. Your aim is to find the lncRNAs that are targeted by miRNAs based on miRNA and lncRNA expression data and miRNA binding prediction algorithms and construct a regulatory network.

**Project 22 - Generating protein-RNA regulatory networks**

One of the major challenges in biology is to be able to generate networks of interactions between the different components of the cell. A very important type of regulatory interaction required in the genes expression pathway is the interaction between RNA binding proteins and the mRNA. In order to generate regulatory maps of protein-RNA interactions we first have to detect the targets of the RNA-Binding Proteins. In the last decade, several experimental approaches were developed to find the targets of RNA-binding proteins (such as CLIP). The further challenge is to identify the real targets, define the binding motifs and generate the interaction network.