1 Introduction

Genomic data science allows researching about life and decoding its unknown mechanisms computationally. Its challenge is to gain knowledge from a vast amount of biological *sequencing* data. Sequencing is the process of turning biological genetic information, encoded in deoxyribonucleic acid (DNA) molecules, into machine-readable information. This work is about the algorithmic interpretation of data from medical experiments and modeling of biological systems *in silico*.

The first successful sequencing technique was invented by Frederick Sanger [72] in 1977. Only after almost thirty years in 2005, private companies started to improve this technique by parallelizing the process and therefore reducing sequencing costs [74]. The thereby developed de-facto standard until today is called as *next-generation sequencing*. DNA as well as ribonucleic acid (RNA) (Sect. 2.1) can be sequenced with these techniques.

The interpretation of the output of the sequencing machines lies in the field of bioinformatics. The major challenge is to draw correct conclusions on the underlying biology, while, unlike computer-based algorithms, biology and the bio-chemical sequencing process seem to follow non-deterministic behavior. Since many biological processes are still unknown, modeling nature with computational methods is very difficult. Experiments are conducted on biological molecules following individual processes (protocols) and contain various unknown factors, the so-called *bias*.

Nevertheless, with scientists all over the world using more or less the same technologies, thousands of terabytes of comparable data were created over the years. The *recycling* of this data, which means re-analyzing and comparing it under different and new hypotheses, saves further sampling and sequencing costs and allows computer scientists to contribute to this field of research without practical training in molecular biology.

Furthermore, the fast-growing field of machine learning and artificial intelligence profits from ever growing data sources and provides opportunities for better and deeper biological analysis.

In this thesis the focus lies on RNA sequencing (RNA-seq) data. The first part specifically focuses on small RNA (sRNA), which is a type of RNA, that is rather short and does not code for proteins but has regulatory functions instead. The goal is to provide a rich resource of information and tools around existing sRNA-seq experiments, use this resource to discover new biological and medical insights and provide researchers all over the world the opportunity to do the same.

In the second part of the thesis, contrarily, primarily protein coding RNA is considered. Experimental processes contain unintended side-effects, which are usually filtered out computationally. A new analysis algorithm is developed for gaining additional information from these neglected parts of the data.

The biological hypothesis in this work relates to neurodegenerative diseases. In all projects, neurological examples are chosen and analyzed as representative for any kind of medical condition. The hypothesis is that an infection of human brain tissue with viruses or bacteria is possible and that these pathogens relate to the disease.

This proposition received increasing attention over the last years [3, 14, 81]. The human brain is well protected by the blood-brain-barrier and therefore should not contain any other molecules than the ones necessary for brain functions. Nevertheless, for most neurodegenerative diseases the causing mechanisms are still unknown. While the genetic and environmental risk factors have been studied in depth in recent years, new angles have to be considered for gaining new insights. For example, the microbiome of the gut as well as the gut-brain-axis are being studied and found to influence various neurological diseases, such as Alzheimer's disease (AD) and Parkinson's disease (PD) [65].

Nonetheless, before starting to design expensive experiments for proving new hypotheses, strong evidence should be available. This thesis aims to provide resources

and tools for making use of already existing data and proposing novel biological theorems. Like this, no resources for *in vitro* experiments need to be wasted for finding the desired evidence.

The major contributions of this thesis are the following:

1. A publicly available and search-able database is provided for sRNA related data from various world-wide publications in this field.

2. Over 4000 sRNA samples were analyzed and the results are presented online alongside further analysis options. While the system supports ten organisms and many different tissues and diseases, the focus of the discussion in this thesis lies on human neurodegenerative diseases.

3. sRNA samples from neurologically diseased patients were analyzed and evaluated for pathogenic organisms in order to formulate medical infection related hypotheses.

4. An algorithm has been developed for identifying contamination and infection in RNA-seq samples and its performance in precision compared to current solutions.

5. An analysis pipeline is proposed for evaluation of the developed algorithm from point 4.

6. Based on points 4 and 5 a hypothesis is generated about an infection of brains of patients with frontotemporal dementia (FTD).

The chapters of this thesis are organized as follows.

Chapters 2, 3 and 4 introduce the background in biology, bioinformatics and computer science. Chapter 2 includes an introduction to the central dogma of molecular biology, to the molecules of interest for this thesis and the sequencing technology. Furthermore, introductions to microbiology and neurodegenerative diseases are given. Chapter 3 introduces the state-of-the-art concepts and tools in bioinformatics for analyzing sequencing data, such as alignment, differential expression analysis and metagenomic analysis. Chapter 4 finalizes the introductions with concepts related to computer science. These are, on the one hand, approaches and frameworks for internet-based technologies and, on the other hand, selected topics in the fields of data science and machine learning.

The Chapters 5 and 6 form the first part of the research chapters surrounding sRNA data. Chapter 5 introduces the architecture and implementation of the online available Small RNA Expression Atlas (SEA). Chapter 6 presents a selection of biological discoveries based on the SEA system.

The Chapters 7 and 8 form the second half of research chapters surrounding transcriptomic sequencing data. Chapter 7 presents a new algorithm for identifying viruses or bacteria in transcriptomic data of a human or other host. In Chapter 8 an analysis pipeline is proposed for interpreting the outcome of Chapter 7 and based on this, a biological case study is presented.

The conclusion of this thesis is ultimately drawn in Chapter 9.

2
Biological background

The origin for the analyses in this thesis, as well as the goals lie in the fields of biology and medicine, even though the actual work is fully performed *in silico*. In this chapter, the different biological topics are introduced. First, the data source is explained, i.e. what ribonucleic acid (RNA) is and how it is represented digitally via sequencing. Then, necessary for making assumptions during the workflow development, an introduction to microbiology is given. Lastly, required for interpreting the results and forming a hypothesis, neurodegenerative diseases are presented and the current state of research concerning infections in those.

2.1 Ribonucleic acid (RNA)

The central dogma of molecular biology [13] (Fig. 2.1) describes the mechanisms of life itself. It characterizes a two-step process of how proteins are created from genetic material. While deoxyribonucleic acid (DNA) can be named as our genetic code, RNA can be understood as a compiled version of this code. While the genetic code is in general identical in all cells of the same organism, its "compilation" differs from cell to cell. This "compilation" is called *transcription* and the "compiler" is an enzyme called *RNA-polymerase*.

The DNA molecules are shaped as a double helix, composed by two strands of

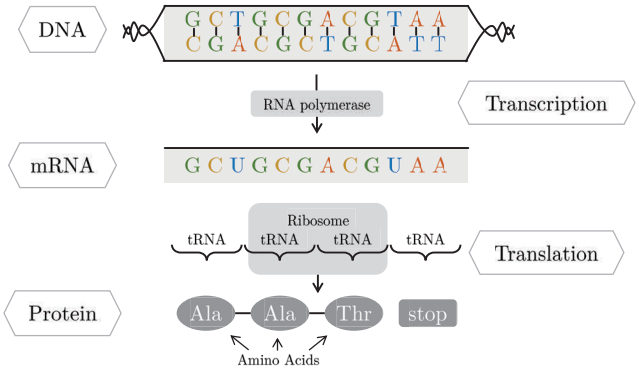

Figure 2.1: The central dogma of molecular biology: The machinery of life existent in every cell is a two-step process turning DNA into mRNA and then into proteins.

nucleic acids. Its four different nucleic acids are adenine (A), cytosine (C), guanine (G) and thymine (T). The first step for making proteins and cells eventually, is to open up this double helix and creating a copy of one of the strands. The regions of DNA where this is done are called genes. The copy is made from RNA, which is, in comparison to DNA, only a single stretch of nucleotides A, C, G and uracil (U).

Some proteins called transcription factors, regulate this process by interacting with the RNA-polymerase. The high diversity of cells in an organism is created from the same set of DNA by the selection of different genes for transcription. A complex, yet not fully discovered system involving epigenetic factors like histones and DNA methylation is additionally changing transcription for every individual cell.

Two types of RNA are synthesized: *coding* and *non-coding* RNA. Only coding RNA, called messenger RNA (mRNA), is defining the actual structure of the prospective proteins. The role of non-coding RNA on the other hand, is not fully explained but some kinds, for example micro RNA (miRNA), interact with mRNA molecules and thereby change the genetic compilation (Sect. 2.2).

The ribosome is responsible for the second step of the central dogma: the translation from mRNA to amino acids which are forming the cell's proteins. It is made from RNA itself, ribosomal RNA (rRNA) and synthesizes amino acids with the help of another type of RNA, the transfer RNA (tRNA). This tRNA recognizes triplets of nucleotides, named as codons, and carries one out of 20 possible amino acids to the corresponding position. Every codon has exactly one target amino acid.

This basic introduction to *gene expression* left out many details and additional processes such as introns and splicing. In this work though, the focus lies primarily on mRNA, and secondly on small RNA (sRNA), which are introduced in the following section. How gene expression is measured and made machine-readable is described in Section 2.3.

2.2 Small RNA

Next to coding mRNA, there is also a collection of non-coding RNA which are not translated into proteins. They have different purposes such as regulating gene expression. Nevertheless, many of their functions are still unknown.

Small RNA are a group of non-coding RNA. They are called "small" because they are comprised of only about 19 to 25 nucleotides [2]. The major type of sRNA are miRNA, which are typically responsible for gene silencing. They bind to mRNA and thereby stop the translation into proteins. MiRNA are derived from hairpin precursor RNA, that has its name because of its shape with imperfect binding to itself. A similar mechanism to interfere with the gene regulatory network is carried out by small interfering RNA (siRNA). They generally bind to all kinds of RNA molecules and thereby cleave and degrade them. A depiction of both types can be seen in Figure 2.2.

The binding sites can be predicted since the sRNA's nucleotides pair up with the open nucleotide strands of other RNA molecules. C and G connect to each other as well as A and U. Heavily simplified, it can be imagined that once an sRNA "sits on" one end of mRNA, the so called promoter region, it blocks the ribosome from attaching as well and thereby stops the amino acid synthesis [32].

Due to their shape (Fig. 2.2), miRNAs have thousands of (binding) targets and

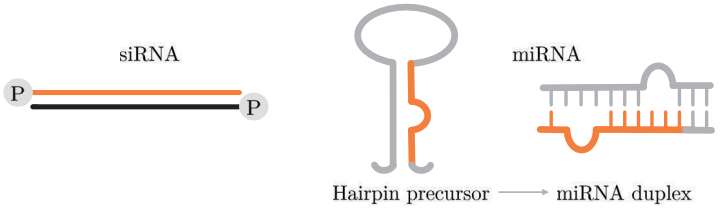

Figure 2.2: Two different types of small RNA: miRNA and siRNA. **LEFT** siRNA induce mRNA degradation by complementary binding to it. **RIGHT** miRNA are derived from hairpin precursors, destabilize mRNA and inhibit its translation. (Adapted from [41].)

potentially control the whole transcriptome [19]. Because of their functions, sRNA play an important role in disease. A vast amount is still unknown, like which sRNA do exist and what their exact role in the mighty gene regulatory system is. Furthermore, sRNA can serve as biomarkers since some of them express aberrantly in the condition they are marking, for example a disease.

2.3 Sequencing

Thanks to modern sequencing techniques, researchers, especially computer scientists have the possibility to examine the processes of molecular biology and genetics *in silico*. Sequencing is the effort of encoding the physical sequence of nucleic acids into a sequence of literal letters. Different techniques have evolved since the beginning of DNA sequencing in the 1970s. The current state-of-the-art technique is known as next generation sequencing (NGS). In the following, the process for RNA sequencing (RNA-seq) is outlined according to the steps in Figure 2.3 as described by Kukurba et al. [40].

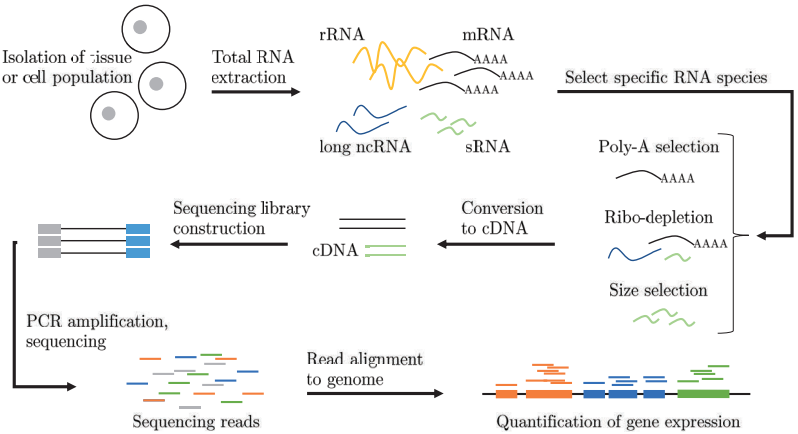

Figure 2.3: The next generation sequencing process for RNA (adapted from [40]): From cells extracted RNA may be sub-selected by its type and converted to complementary DNA. After constructing the sequencing library, the molecules are amplified through PCR and sequenced, resulting in sequencing reads, which are finally computationally aligned to the genome.

RNA molecules are being extracted by breaking the cells, a process called *lysing*, and isolating RNA chemically. Depending on what is being studied, the mixture

can then be separated. For example, sRNA can be physically separated from the rest through precipitation of RNA with low molecular weight. Another way of isolating desired molecules is pulling the mRNA out from their poly-A tail with poly-T chains. Selecting the RNA of interest at this point saves the sequencing costs later in the process.

After the desired RNA is selected, it is then reverse-transcribed into complementary DNA (cDNA) for re-using DNA sequencing technology. Reverse transcription is performed by an enzyme, reverse transcriptase, which does the exact opposite of RNA-polymerase (Sect. 2.1). Then, in order to connect cDNA strands to a chip surface in the sequencing machine, sequencing adaptors are ligated to their ends. The last step before the actual sequencing in the machine is an amplification via polymerase chain reaction (PCR), which is a copying process for getting multiple replicates of strands.

The advantage of the NGS method is that it can process a high number of strands at the same time, which is why it is referred to also as *high throughput sequencing*. The most prominent sequencing machines on the market are built by the company *Illumina*. In this fluorescence-based technique, cDNA strands are attached to a surface, e.g. a chip or flow cell.

Inside the sequencer, the reverse strands are built with light emitting nucleotides while binding one by one to the original. Different wavelengths (colors) encode for the different nucleotides. This copying process is observed by a camera from above. The sequence of light is an input to the computer which is then able to generate the *reads*, a sequence of letters A,T,C and G. The copying and measuring process is repeated many times, forward and backwards the strand. This increases the quality, since the sequencing computer can compare and process presumably redundant information. Furthermore, the light signal is stronger after several cycles since binding is supposed to happen simultaneously for all existing copies at the original position.

The output of the sequencer are files containing all reads of a sample, mostly with a quality information indicating the probability per nucleotide that it was identified correctly. The file format is named *fastQ*. All subsequential processing steps from here are in the field of bioinformatics and will be described in Chapter 3.

The first steps until PCR are similar in all sequencing techniques and can be performed by most labs with the necessary equipment. NGS and its sequencers though have been industrialized due to high specialization. Still, thanks to massively parallel processing the cost per one million sequenced base pairs fell since the invention of NGS in 2005 from 1000$ to 0.01$ in 2019 [89]. This is faster

than Moore's law would predict and led to a massive amount of data with a great potential in the field of computer science.

2.4 Microbiology

Long before eukaryotic organisms like humans, animals and even plants existed on this planet, it was already inhabited by prokaryotic organisms like bacteria. While the latter are mostly individual cells, eukaryotic cells live in symbiosis together, forming an organism.

Even though bacteria do not have a nucleus, this does not mean that they are underdeveloped. They actually had much more time for evolution than eukaryotic organisms to perfectly adapt to their environment. Their interaction with the environment as well as other bacteria is what makes them interesting to study. Not only our planet is home to an uncountable number of bacteria species. Even most cells of the human body are actually not human, but primarily bacteria. These bacteria combined are called as (human) *microbiome*. While the gut microbiome might be the most famous example, also many other parts of the body, such as skin, lungs, mouth, liver and even blood are home to microorganisms. In most cases the bacteria are highly welcomed and beneficial for human health. Only when some specific bacteria get the chance to enter the body and their population grows uncontrolled, they cause infection and disease. We call these uninvited species *pathogens*.

Another kind of uninvited species are viruses. Even though their biological structure and behavior differs significantly from bacteria, they are included when the term *pathogen* is used in this thesis. In short, viruses are not cells or organisms, but capsules containing DNA or RNA material [47]. They use the machinery of organic cells to reproduce.

Organisms can be grouped together by similarity of their genome. Organisms which share a major part of their genome are categorized as the same *species*. For example, all humans belong to the species *homo sapiens*. Since bacteria are very diverse, several bacteria species are grouped together into a *genus* or an even higher grouping term. A *taxonomy tree* organizes the grouping and naming of organisms. An example for homo sapiens and the bacterium *Escheria coli* can be found in Figure 2.4.