Reflections on the course *Current Topics in Bioinformatics*

My Expectations

I was compelled to start Current Topics in Bioinformatics, because I am familiar to some Data Science techniques and Machine Learning Architectures. When studying ML and when working a job in the field, I found mostly boring use-cases: The solutions that gain money for businesses seem to rarely be compelling. So I hope to find more interesting challenges in Bioinformatics with potential benefits of improving healthcare.

I know that the human body is a complex machine that is understood only in parts, and I am aware that a lot of effort has been made to capture genomic, diagnostic and imagery data. Many biologists that I talked to in the past had little interest in Programming or Machine Learning, that's where I see myself as a supporter.

In this subject, I wanted to find out: What ML techniques are currently in use for Bioinformatics? What are open challenges in Bioinformatics? How quickly can I understand Medicine-related topics? How big is my knowledge gap? And what level of analysis do Biology people think of?

Additionally, I wanted to learn some R.

Seminar Discussions

Simon Orozco - The transposable element challenge

Before this presentation I have never heard of transposable elements. I only knew that the DNA contains non-coding elements "exons". Simon's taxonomy was hard to understand but the underlying story he told was clear: Transposable Elements are not well understood and can only be recognized because of their repeated occurrence through the genome. This implies that the data encapsuled in chromosomes can transform instead of being static, which is a new idea to me. But when I consider DNA-repairing mechanisms, protein-synthesis and epigenetic factors, the idea of static DNA is probably absurd anyway.[1]

When Simon pointed out that he used Machine Learning to detect and annotate TE's I wondered which method he used to embed the nucleotide sequences into vectors. I

[1] As a physicist, I like static systems. Chaotic systems scare me.

was intrigued to learn that he tried different methods, from simple one-hot encoding to k-mer encoding which seemed very reasonable for DNA. I wondered why he didn't use a recurrent network architecture, since all his methods were based on some kind of pooling over different sizes of potential TE's. Still, he could achieve good results, so apparently the global (long-range) structure of a TE is not relevant to the annotation problem.

He talked in detail about databases of TE's that are necessary to benchmark his results. And he explained the difficulty of annotating the full TE instead of a partial sequence.

In a later seminar by Miriam Merenciano learned that TE's can affect the expression of genes, thereby influencing resistance to pathogens. She explained that the moving nature of TE's is a source for genome variance and mutations.

Irepan Salvador-Martínez - Spatial Transcriptomics

Before the subject, I knew about advancements in Genomics, specifically about the Human Genome Project. I only had a vague idea of how this worked. Irepan compared Bulk-Sequencing, Single-Cell-Sequencing and the recent Spatial Sequencing. He took some time to explain to me again how this process works. The sample is sliced, and special care goes into the preservation of the cell structure. Then, the RNA is multiplied by PCR, then nucleotides are introduced that combine with the RNA strings and send out a light pulse when binding to the DNA sequence. Then the stop nucleotide must be removed before the process can be repeated. It was incredible to see that RNA can be captured on a sub-cellular level. I believe, that the understanding of healthy and sick cellular processes can be understood better with this technique.

I learned, that Genome Sequencing can be accelerated by slicing the DNA at random points. This creates an interesting computational challenge of recombination after the analysis. When Jose explained these challenges on the example of sequencing the Spider genome, he focused on the importance of comparison to known genomes of model organisms.

During the whole course I learned to differentiate Genomics from Transcriptomics and Proteomics. I didn't know before, that genes could have multiple different ways of transcription and different ways of Splicing to generate way more proteins than are genes in the body.

Additionally, I learned that some bases in DNA are known to be prone to mutations and current effort is being made to predict the likelihood of mutations. With methods like Alpha-Fold researchers can then evaluate what effects the mutations have on the protein's function.

Honorable Mention: Clara's Innovative visualization techniques for Clonal Evolution

Although her topic was not super interesting to me, I love her way of visual storystelling and I will use her presentation slides as reference for my presentations in the future.

Final Reflections

As a physics bachelor student, this subject was my first look into biology after high school. Working as a bioinformatician is not more than one idea among many others. Due to my sparse knowledge, I wanted to get an oversight on important topics in medicine and bioinformatics, just as this course name suggests.

I learned about Genomics, the function of Transposable Elements, Transcriptomics, Cancer Evolution, the study of mutation and mutation tracking for pathogens.

The focus on cancer in P3 was my highlight in the practical part of the course, because I was forced to understand the influence of single genes on cancer cells. Fortunately, the skin cancer that we chose was well suited as an example of typical malignant mechanisms. From all I know (which is not that much) cancer is one of the main topics in current medicine research, and I am glad to have learnt a little about how this research is pursued.

The Clementi group at my university FUB in Berlin works in Biophysics, specifically in the study of macromolecular movement. I was interested in doing my bachelors' thesis there before this course. This course supported my interest in Biophysics and Bioinformatics, seeing how many unsolved problems are to be discovered. Not only does this fit my interest, I also think that with my skillset as a physicist with a strong background in Data Science and Machine Learning I can contribute to the field.