*Project Specification draft*

**Project:** Analysis of Ovarian Cancer Single-cell RNA-seq data

**Supervisor:** Sascha Ott

**Student:** Ava Spataru

**Abstract**

Different cell types of a multicellular organism carry the same DNA, yet fulfil very different functions. In part this is possible because the DNA is chemically modified to affect its function, in particular as to which region of DNA is active or inactive. Epigenetic sequencing data sets provide insight into these chemical states. By analysing such data we can answer questions about differences between cell types, or differences between healthy and unhealthy cells, for example.

**Main project idea**

The main goal of this project is to analyze cell data from multiple patients suffering of ovarian cancer and be able to classify cells as parts of different types, as well as to understand how the identified types transform after treatment.

**Motivation behind project**

There are multiple reasons why attempting to classify cells is important. Besides this being a next step towards understanding how organisms work, it could have significant impact in how treatment is conducted against cancer. If the classification is a success on even a small set of data, the approach can then be generalized to different types of tissue.

One in five patients suffers through chemotherapy without any positive outcome, because sometimes the tumor isn’t affected in any way by the treatment. Currently, the doctors can’t determine whether the treatment will reduce the size of the tumor or have no significant impact. If the cells could be clearly separated into types, then they would be able to associate similar tumors and determine whether the most beneficial treatment is chemotherapy or the surgery.

If the cells could be clearly separated in types and subtypes, then that could be the beginning of designing medical treatment specific to each patient. If the doctors could see the types of cells and how each type responds to different medication, then they could start creating different medical treatments based on the combination of cell types presented.

**Personal motivation**

After a discussion with my supervisor, it was decided that the project is feasible, despite the lack of previous knowledge in the domain. I was really interested in learning more about the world of computational biology. The domain is quite broad and requires a lot of prior understanding in order to conduct research on it, which is why even though the title was decided on in the early stages of the project, the more specific details of what the project consists of were discussed after reading more about the area and getting a grasp of what the research involves. This allowed me to focus on a more specific part of the research domain, that I considered more interesting for a 3rd year project.

**Challenges of the project**

One of the main challenges with this project is the lack of prior knowledge in the domain of computational biology, as well as little experience with data analysis. However, in the early stages of the project, the supervisor suggested a number of papers and online resources that proved very useful to filling in the knowledge gaps.

The classification of cells if a very difficult task, mostly because it involves analyzing a large amount of complex data, based on various criteria. The size of the data set isn’t the only challenging part of the project. Another difficulty with classifying cells is the limited understanding of what each bit in the transcript actually represents. The transcripts contain a lot of information, that sometimes may seem similar for two cells, but carries very different meanings. This is because there exists a vague understanding of what information should be carried in each cell, but not enough to be able to quickly identify errors.

Multiple experiments have been conducted before in this area, attempting to classify cells of patients before and after chemotherapy. Unfortunately, the results are hard to interpret. In the beginning there were four types of cells identified and after the treatment, cells were grouped into five types. Out of these five groups, only one can be mapped to an initial group. This could be interpreted in multiple ways, for example: perhaps some of the groups of cells changed beyond recognition. Another possibility is that some groups of cells divided into smaller groups, but it is also equally possible that the initial groups of cells completely disappeared and were replaced by new ones.

“Testing” is another challenge with this project, because there is no straight-forward way of knowing if the identified types are actual correct types. Currently there is no knowledge of how many cell types should be identified, which means that there is no easy way to determine if two types are actually just one group or should be divided into smaller groups. Some cell data may differ for each person, meaning that there is a possibility the results of this project aren’t widely applicable.

The current sets of data have been maintained in a petri dish, which means that they may have developed slightly different than the ones inside a human body. For this reason, the actual analysis might be slightly different than the data of an actual patient. However, there exists a chance that a new set of data will be made available for analysis during the project, which will be recently extracted and therefore more accurate.

**Background information**

Every cell in the human body has different types of RNA. The RNA is a polymeric molecule specific to each cell that will be different based on several factors, including the function of the cell. A type of RNA (mRNA) goes from the nucleus of the cell to different parts of the cell or outside, to transmit information. Drop-seq is a strategy that analyses these “messages” (mRNA transcripts) of each cell and identifies similar transcripts as belonging to the same cell population. This classification is done based on the idea that similar transcripts, mean that the cells fulfill similar functions and therefore are similar.

After the data is sequenced, the result is a set of objects with multiple dimensions of information to be analyzed. There are multiple methods to investigate highly dimensional data, such as scatterplots and heat maps, but the one that will be primarily used in this project is t-SNE (t-Distributed Stochastic Neighbour Embedding). T-SNE is a fairly new approach to representing highly-dimensional data in a space with lower dimensions, that can be more easily interpreted. The algorithm employs machine learning to adjust to each set of data and has 2 parameters: perplexity and epsilon. Even tough the outcome of the algorithm is very useful for analysis, it can be erroneous. There is an online paper (REF) that explains how t-SNE sometimes might be misleading. For example, the perplexity parameter seems to be the most reliable between 5 and 50, however a small perplexity leads to the domination of local variation and a perplexity larger than the number of points doesn’t usually reach stability. There are a multitude of factors, such as cluster size, distance between clusters etc. with unclear behaviour, that are relevant to understand how t-SNE works.

**Project goals**

The main aim of the project is to analyze the existing sets of data (from 2 patients) and successfully classify the cells as part of different types. The secondary aim of the project is to develop a deeper understanding of which genes modify during chemotherapy and how this affects the type classification.

Since the aim of the project is hard to reach, there are a couple of milestones that the success of the project could be measured against. However, these could change as the project advances, to accommodate for a different, more appropriate approach.

A total number of 130 hours has been assigned to each task. Since the project itself was estimated to 300 hours, the rest of 170 hours will be allocated to reading, researching, report writing and improving the analysis. The estimations will probably change, because as the project progresses, different objectives may appear.

|  |  |  |  |
| --- | --- | --- | --- |
| Goal ID | Goal description | Approach | Hours |
| G1 | Use t-SNE to separate the cell data into types. | This can be done after some practice with applying t-SNE and installing all the tools needed. | 20 |
| G2 | Determine what the parameters for t-SNE should be and why. | “Tweaking” parameters and observing the behavior of the data set on multiple runs with different values. When observing, the points made on the webpage *“How to use t-SNE effectively”* (REF) should be considered. | 5 |
| G3 | Compare similar behaviors on t-SNE between these two sets of data and other already existing sets of data. | Done by simple observation, and the aim is that similar behaviors on the same algorithm might help identify common properties. | 10 + (50 stretch goal) |
| G4 | Compare the groups identified in each set of data with the other ones, while emphasizing on the most common genes in each. | Adjust certain parts of the data to evidentiate different genes. For example, observe how the data set acts if 100 more cells which are mostly made up of gene A are added. | 30 |
| G5 | Identify which cells have changed from the initial sample and which cells haven’t, as well as determine if the remaining cells are unwanted or which unwanted cells did the treatment have effect on. | Done by observation and reasoning about the behavior of certain genes. | 5 |
| G6 | Identify genes that completely disappeared from the cell types or genes that transformed. | This step involves an in-depth analysis of each initial type and the expected behavior of the composing genes after treatment. | 5 |
| G7 | Roughly estimate the number of cells in each tissue that completely disappeared | The aim of this stage is to estimate a “percentage of change” for separate parts of the data set and determine to what extent the treatment has equivalent effect upon different datasets. | 5 |
| G8 | Try and create a mapping based on different assumptions of which gene changed to which gene. Repeat the assumptions until a stable result is obtained. | Create a set of assumptions, such as: If gene A in this cell changes to gene C, then gene B must have changed to gene D etc. | 20 |
| G9 | Attempt to classify the cells by prioritizing different genes. | Experiment with different ways of classification by accentuating certain differences. For example, classify based on the existence of a certain gene and the percentage of that gene within each cell. | 30 |

# Stretch goal

The project is primarily a research project, that doesn’t seem to be, in the early stages, involving a lot of implementation. For this reason, there is a possible stretch goal that would involve developing a tool that would run the t-SNE multiple times in parallel, on the same set of data, while also changing parameters and flagging any noticeable behavior. The tool could also identify the run which presents the most meaningful visualization of the data, based on the observations from the previously mentioned webpage: *“How to use t-SNE effectively”* (REF).

The idea behind this tool is that it would reduce the number of human misinterpretations, caused by the misleading behavior of t-SNE, while also allowing fast multiple runs and store the information from previous runs. By expanding goal G3, described above, there is a possibility that similar sets of data will act in a similar way in t-SNE, meaning that if the tool could also compare runs of t-SNE and identify similar ones, then it could potentially identify similar sets of data, thus indicating that the same treatment might be appropriate.

Another addition to the tool could be presenting multiple ways to analyze the same set of data. For example it could include, besides t-SNE, a number of different methods of high-dimensional data visualization, such as the ones described in the paper *“High-dimensional visualizations” (REF)*

The motivation behind this tool is that it would provide a storage of previous runs of t-SNE on the same data set, as well as give a high-level overview of the differences between runs and highlight unexpected behaviours. This will make it easier and faster for researchers to visualise and compare their set of data with previous ones, while logging a history of all previous runs.

This tool is only listed as a stretch goal because it has only been briefly mentioned in a supervision meeting and a more detailed discussion is needed to determine whether it would be achievable and meaningful as a part of a 3rd year project.

**Personal goals**

In terms of personal development, this project should help me understand more about the world of computational biology and give me a good example of what real-life research projects are and how they are conducted.

**Timeline**

**Phase 1: initial reading & software**

Before term 1: Initial reading

Week 1-2, term 1: Work on the specification

Installing necessary software packages

Friday, week 2, term 1 (12/10): Submit specification

Week 2-3, term 1: Familiarize with software, by doing exercises and tutorials

Wednesday, week 3, term 1 (17/10): DROP-seq training

**Phase 2: apply on real data set**

Week 3-4, term 1: Use the software on the actual project data and observe behavior.

Monday, week 4, term 1 (22/10): Meeting with a PhD student researching in the area. To help shape the project and discuss current progress.

Week 5-7, term 1: Perform the main data analysis on the data and formulate findings

Understand the behavior of the data set and speculate about transformations on genes

Week 8-9: Work on the progress report

Discussing current findings and possible future action.

Tuesday, week 9, term 1: Submit progress report

**Phase 3: finializing the project & writing final report**

Week 1-7, term 2: Further development on the analysis.

Compare with existing findings.

Reasoning about the impact of the project.

Week 8-9, term 2: Work on the presentation.

Week 9-10, term 2: Presentation

Week 10, term 2 – Week 1, term 2: Work on the final report

Week 2, term 2: Adjusting the final report to feedback.

Week 2, term 3: Submit final report

The structure of the project is primarily Waterfall, since most steps require the previous ones to have been completed. However, this may change because most of the goals could be adapted based on new findings.

**Ethical considerations**

The data to be analyzed throughout the project is real data from patients at the Birmingham City Hospital, which involves some ethical considerations. However, the patients are entirely anonymized and have given their consent by signing an ethical approval, which is how they have volunteered to be part of the research.

Some of the analyzed data will eventually be made public, when the results of the research are published, but the patients have all consented to this, which means there are no more further considerations for the scope of the project.

**Risk assessment**

Risk 1: Software not functioning as expected. (Main software to be used: RStudio, Seurat)

Solution 1: Alternative software packages (Monocle, SC3, MAGIC) have been made available and could potentially be used throughout the project.

Risk 2: Incomplete understanding of the project or problems using the software.

Solution 2: Maintaining a good relationship with the project supervisor, which will allow for me to ask for help and support with any such issues.

Risk 3: Illness.

Solution 3: The work on the project is aimed to finish before the deadline, such that the schedule can be easily deferred in case of unexpected events, such as illness.

**References**

* *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.*

# *Introduction to t-SNE (Google Tech Talk)*

# *How to use t-SNE effectively*

* -current ways to visualize data (see video)
* ***https://pdfs.semanticscholar.org/43f7/66c06e2a7770d9f37dcd9cfff5bd5dcfc22f.pdf***