Project Plan

A visualization interface for spatial pathway regulation data

YANG ZHANG

Stockholm 2018-3-13

Industrial and Environmental Biotechnology (TIMBM)

1 Objective

The main objective of this degree project is to create a web browser-based data visualizer that is able to combine the information from the pathway databases with the data derived from *Spatial Transcriptomics*. The project will be performed with 20 week period, including the course' objectives of project planning, lab working, scientific reporting and oral presentation.

1.1 Datasets and tools preparation

This project is an extension of work from a study of previous Spatial Transcriptomics, which provided a data set to quantitative differences of pathway genes presented in multiple tissue sections. The first objective of the project is to study the background and formation of those pathway datasets, understand knowledge on how the ST data was being structured from previous studies.

Secondly, how to code in d3.js and JavaScript need to be learnt as the *d3.js* libraries are highly dependable for data visualization and JavaScript(Python) can be helpful for interface development.

1.2 Data visualization

A major objective for this thesis project is to present the Spatial Transcriptomics data in a more intuitive way. The idea is to first create an image that illustrates the level of pathway expression of a certain tissue. This type of image should be semi-transparent so it can overlaid with its histological section. Secondly, a sunburst diagram should be plotted which represents the hierarchical nature of the pathways as well as displays information on how relevant those pathways are.

1.3 Interface Development

To present the previous data visualization works in an easy-navigate way, we aim at designing a web-browser-based interface by using JavaScript and Python. We will embed the D3 code into HTML's script part. The requirement is also to develop the tool to be easy in interact, hopefully facilitating the user's exploration of ST data in order to increase the usefulness of Spatial Transcriptomics information [5].

1.4 Write a master's thesis and hold a presentation

The purpose of thesis and oral presentation are to convey the project into a form that is communicable to others. The master's thesis and the presentation should provide a full and correct picture of the project and its derivatives.

2 Background

A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. [4] Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move.

In any organism, there is a large number of biological pathways , i.e. sets of of related biomolecules, which together conducts different functions of its cells. There are multiple pathway databases such as *KEGG* [3], *Gene Ontology* [1] and *Reactome* [2]. These pathway databases are normally formed with a map of molecualr interaction/reaction network diagram, which representing knowledge on molecular interaction, reaction and relation networks of different organisms [3] (e.g. Metabolism, Human Diseases, Drug Development).

Although the definition of pathways may vary among databases, the concept that genes can control the activities among different pathways has been established. However when the study focus on specific tissue sections and their cells, the questions has been raised: is the transcriptoms in the section yield the result of regulation of pathways? Based on those pathways databases, correlated with the method *Spatial Transcriptomics* is the way to solve the problem.

Spatial Transcriptomics is a method that allows visualization and quantitative analysis of the transcriptome in individual tissue sections. By probing the histological sections at different locations, arrayed with oligonucleotides containing positional bar-codes, it is possible to generate cDNA libraries with precise positional information for RNA-seq. This provides transcriptome data in a versatile format for bioinformatics analyses of gene expression within the tissue context. In this case, the Spatial Transcriptomics generate the result shows the qualities of genes regulating the pathways in tissues. Those results will be valuable for further data visualization studies [5].

Javascript (JS): is a interpreted programming language, which also characterized as dynamic, weakly typed, prototype-based language. JavaScript commonly used to make webpages interactive and provide online programs [6]. The majority of websites employ it, and all modern web browsers support it without the need for plug-ins, thus it fit the usage to create web-browser base interface for this project.

D3.js is a JavaScript library for manipulating documents based on data. D3 is able to bring data to life using HTML and SVG. D3 is emphasis on web standards, which gives the full capabilities of modern browsers to combine visualization components and a data-driven approach to Document Object Model(DOM) manipulation, in order to apply data-driven transformations to the document. For example, one use D3 to generate an HTML table from an array of numbers. Or, use the same data to create an interactive SVG bar chart with smooth transitions and interaction.

3 Work Breakdown Structure

A work breakdown structure (WBS) is a hierarchical breakdown of large activities into smaller and more comprehensible units. The WBS of the degree project is presented in Figure 1.

4 Milestones

1. **Basic tools prepared, Instructions read** - When reach this milestone, it means all pre-study works about *d3.js*, Javascript and database backgrounds' study will be finished. Also the tools preparation such as github

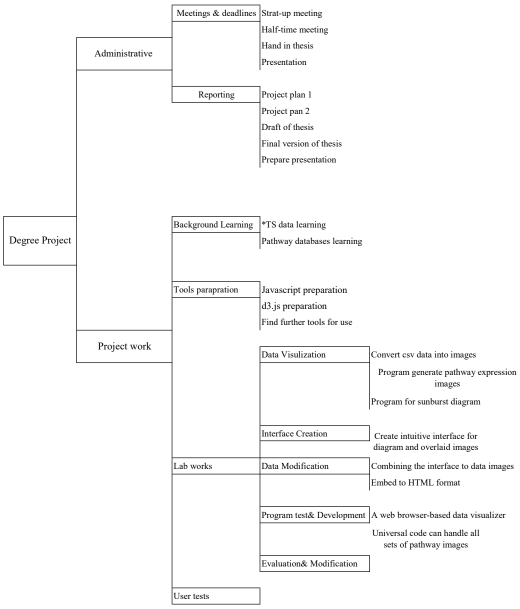

Figure 1: Work breakdown structure

and local server set up can be completed. Major coding works will begin after this milestone.

2. **Data visualization complete** - That will be a relatively loose milestone, once data visualization codes been written, it means a basic pathway expression visualizer will be created. However, code will have to modified to meet the adjust of the requirement from the subsequence working packages.

3. **Half-time meeting** - The second version of the project plan will be finished before the half-time meeting. Categorized problems encountered in the stage that need to be overcomed. There after formed a matured project structure and handled to examiner and supervisor.

4. **Interface created**

5. **Data modification finished** - main coding job of the project is completed and a primitive program should deliver.

6. **Program verified and evaluation complete** - A final version of program should be carried out, start writing master thesis afterwards.
7. **Final version of masters thesis finished** can subsequently be handed in.
8. Presentation - final milestone.

5 Time plan

A Gantt chart based on the Work breakdown structure which also including milestones and deadlines presented in Figure 2. It shows an approximated time estimates for the whole project.

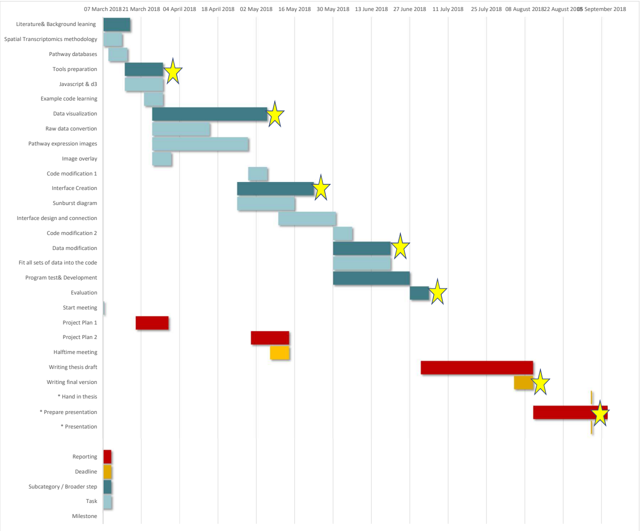

Figure 2: Gannt time table, * the date to hand up thesis and hold presentation is during week 19,20 according to the project schedule, the real data may vary due to the summer break.

6 Specification (in MoSCoW)

The MoSCoW structure (Must Have, Should Have, Could Have, Won't Have this time) is used to prioritise tasks related to the project.

Must

- Successful convert the ST data(csv) into visualized pathway expression image.
- Overlaid the pathway expression images into histological sections.

- Create a visualizer interface that fits required sets of data.
- Project planning.
- Write masters thesis.
- Opposition
- Oral Presentation

Should

- Optimize the code to fit more data sets.
- Successful handle the connection and shifts between sunburst diagram and overlaid images.
- Create a csv data auto converting program

Could

- Develop the interface to become user friendly and easy navigating.
- Try to create a images zoom-able program.

Will not

- A deep understand of Spatial Transcriptomics.
- A strategy to quantitatively judge the quality of pathway expression in tissues.

7 Stakeholder analysis

The stakeholders for this project are categorized and described in the tables below:

Table 1: Stakeholder analysis table

Stakeholder	Role	Interest	Power	Approach
Primary				
Student Yang Zhang	- Study and work on build up a visualizer for the project. - Write and present the results in the form of a Master thesis.	-Pass the project course and graduate from master program. - Produce a good quality result that could be intuitive for further studies.	Decision making during executing main process of project(coding)	
Examiner Peter Savolainen	Evaluate and grade the project.	Making sure the whole graduate project follow the KTH requirement.	Making sure that the master thesis is conducted within the course boundaries	- Start up meeting -Half time meeting - Hand up Project plan - Thesis hand up
Supervisor Lukas Kll	- Give advice and supervision for the project. - Provide a full and correct picture of the project.	- A well-made visulization interface that are able to navigate further related studies.	-High power on the project execution - Evaluate my performance	- Daily working on the office - Weekly group meeting
Secondary				
Statistical Biotechnology Research Group	The office place and the group I stay during the whole project.		- Give practical suggestion for the coding part of project	- Weekly group meeting

8 Business case

Because of the academic property of this master project, the content of business analysis is limited in some aspects.

From the ethical aspect, the project is aiming at facilitate the user navigation and exploration of Spatial Transcriptomics data during their research process.

From the society and environmental aspect, biological pathway is strong related to the human health care and disease research. This project is currently treat the human breast cancer as one of its database, hopefully the final version of this visualizer can be apply into more medical research field, also increase the usefulness of Spatial Transcriptomics technology.

As for the financial aspect, the overall function of the visualization project is rather academic related than business related. the commercial usage of this visualizer is currently unclear.

9 SWOT analysis

Table 2: SWOT analysis table

	Helpful	Harmful
Internal	Strengths - Interdiscipline Knowledge from both Biology and computer science to help me have a better understanding of project. - The basic idea and structure of how visulizer should be form have been given by supervisor.	Weaknesses - Limited knowledge about d3.js and JavaScript beforehand. - Though the overall picture is clear, decisions need to be made when choose to use which method can achieve the final target.
External	Opportunities - Increase the usefulness of spatial pathway regulation data. - To test the usage of data visulization in data analysis studies.	Threats - Strict time limits of project course that might effect of produce satisfactory results.

10 Risk evaluation assessment

The risk evaluation in Table 3 is based on the Mini-risk method where risks throughout the project are identified, analysed and evaluated.

The risks are ranked by a grading system of 1-20 based on the multiplication product of the probability of the risk event (1-5) and the estimated effect the event will have on the project (1-5). This identifies the most severe risks (20 is the most severe, and 1 has the lowest risk)

Table 3: Risk evaluation table

Risk	Possible cause	Proba -bility	Effect	Rank	Comment/ Measure
Desktop Breakdown	Misoperation during coding	3	3	9	Always bring laptop as back up.
Database lost	desktop breakdown	2	2	4	Save current code to Github as frequent as possible.
Loss of time due to wrong choices of coding language	The choice of use Python or JavaScript can be conflict in some case.	4	3	12	Continuous communication with supervisor in order to verify the choices.
Delay of work	Healthy issue	1	2	2	Be tough
Lack of time to finish the project	Underestimation of workload	1	1	1	Plan the project in detail to avoid being behind.
Miscommunication or Misunderstanding between supervisor and me	Supervisor may keep busy during certain period.	3	2	6	Try to present as much as possible during weekly meeting.

11 Summary

The priority objective for this master project is to create a interactive visualization interface which combine pathway regulation information and Spatial Transcriptomics information (a sunburst diagram together with pathway expression plot) . The visualizer should also present the pathway expression level and histological issue section images in an intuitive way.

The data of pathway information is mainly refer from Reactome database. While the source of Spatial Transcriptomics information is from Spatial Transcriptomics research group at Science for Life Laboratory.

The final product of the project will be carried out in HTML format. The project will be written in python and/or Javascript, with dependencies on certain packages such as d3.js.

Most of the implications of the results are to facilitate the user navigation and exploration of such data in order to increase the usefulness of Spatial Transcriptomics information

References

[1] Gene Ontology Consortium. The gene ontology.

[2] Reactome Pathway Database. Reactome pathway.

[3] KEGG. Kegg pathway database.

[4] National Human Genome Research Institute (NHGRI). Biological pathways fact sheet.

[5] Spatial Transcriptomics Research. Spatial transcriptomics technology.

[6] Wikipedia. Javascript.