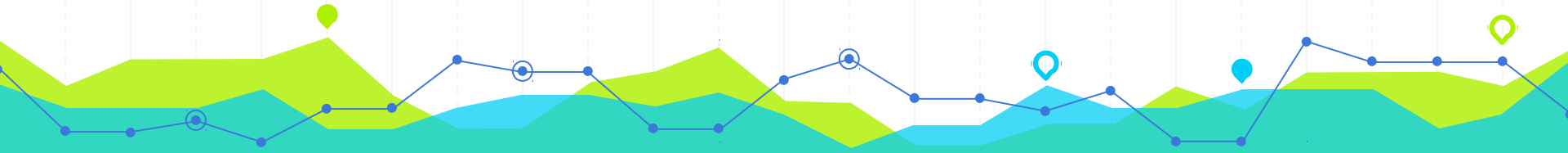
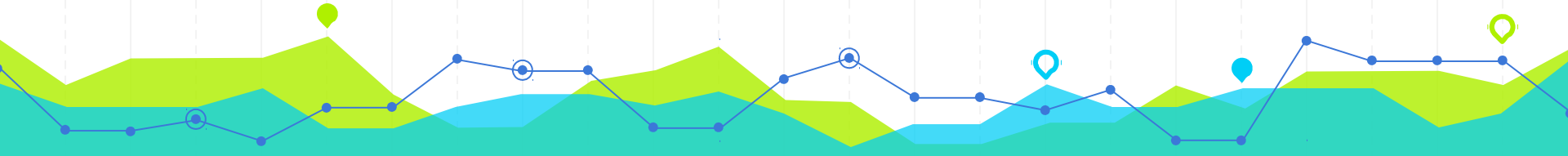
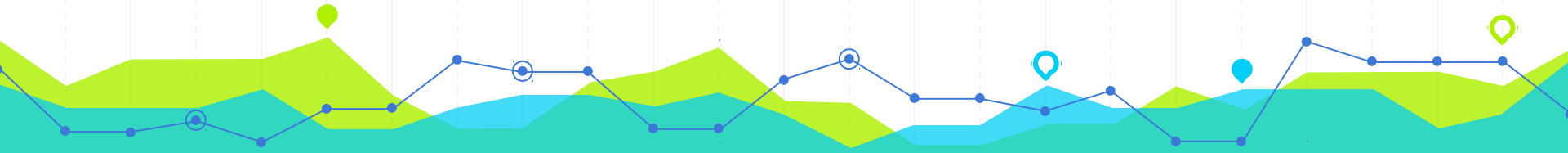


Protein Interaction Prediction

Project Lecture Structure

- Problem Definition
- Sustainability
- Development
- Conclusions





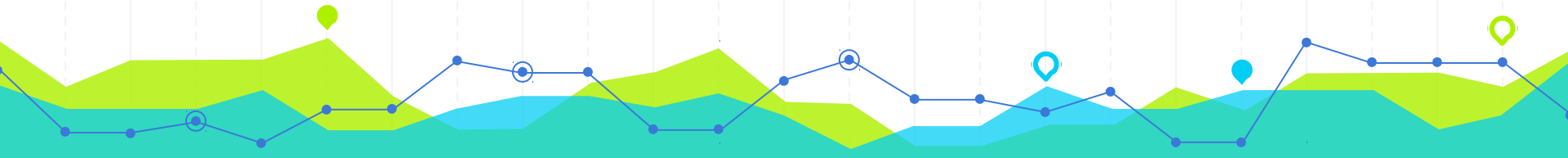
Problem Definition

Goal and scope of the project

1

Goal

The goal of this project is to, given two proteins A and B, find the path an interaction that starts in A and ends at B will most likely follow.



Scope

Cross Specialty Communication
This project required of communication between specialists of different fields.

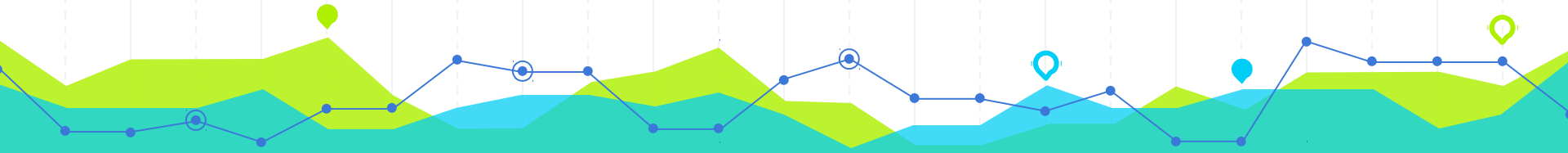
Coding
Thousands of lines of code of different programming languages are expected to be written for this project completion.

Graph Data Mining
Graph data mining was of the utmost importance in this project to find patters suitable for prediction.

Research Project Type
This project had a lot of inherent uncertainty and required of research mindset for its completion.

Computer Science
Several branches of the computer science field were used in this project (Algorithmics, Software Architecture...) .

Data Visualization Techniques
Visualization was used in this project in order to make sense of the data mining algorithm outputs .



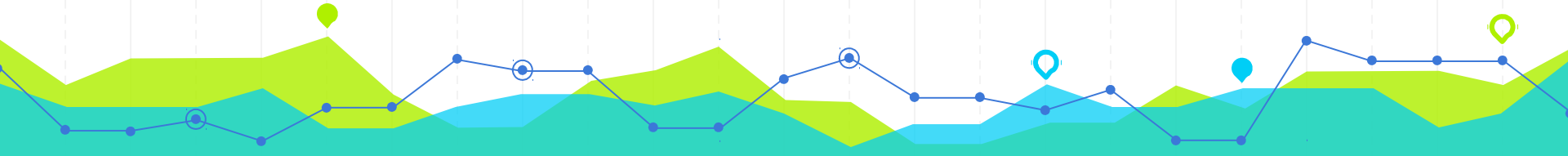
Sustainability

2

The economic, environmental and social aspects of the project

Economic Sustainability

- Most I.T. systems can be seen as investments
- Research is expensive
- The cost of this project development is inexpensive
- This project result could be used to reduce future research costs



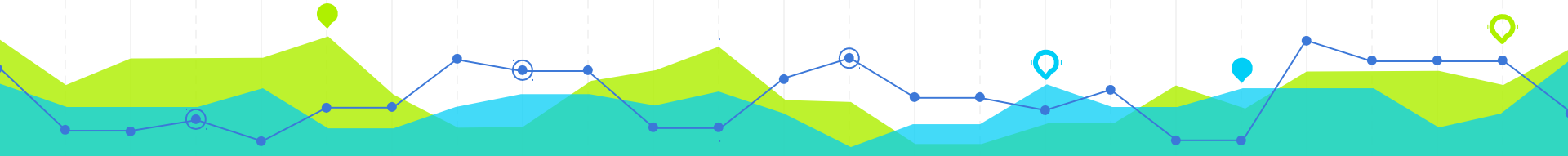
Budget Estimation

Human Resource	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
Project Manager	5	10	10	10	10	5
Data Scientist			75		75	
Software designer		25		25		
Software programmer		125		75		
Technical writer	75					75
Total	80	160	85	110	85	80

Human Resource	Cost per hour (€/hours)	Required Time (hours)	Cost(€)
Project Manager	30.0	50	1500.0
Data Scientist	17.5	150	2625.0
Software designer	15.0	50	750.0
Software programmer	12.5	200	2500.0
Technical writer	10.0	150	1500.0
Total			8875.0

Environmental Sustainability

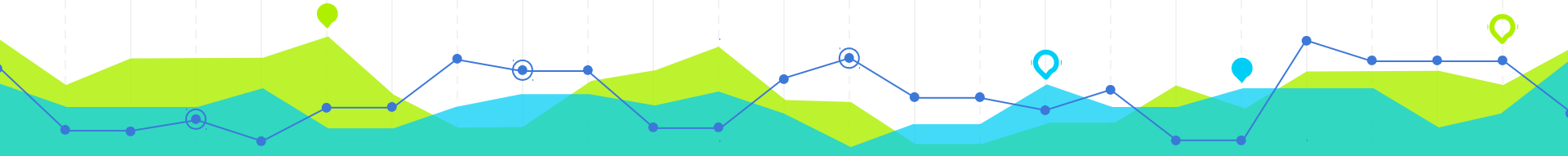
- This project didn’t require any special equipment
- By reducing the number of future research experiments the resulting ecologic footprint can be reduced

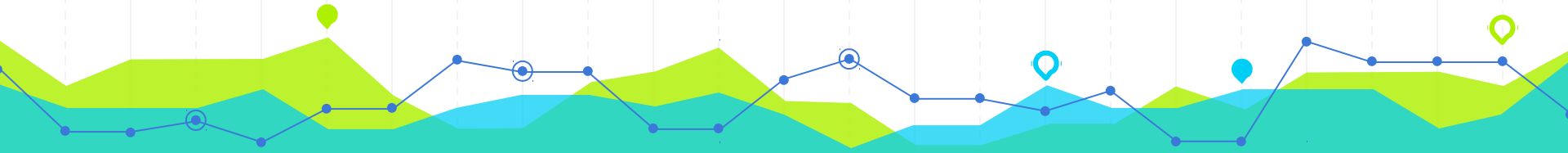


Social Sustainability

Bioinformatics can offer multiple social benefits:

- Biology Knowledge Advancement
- Medical Knowledge Advancement
- Personalized Healthcare





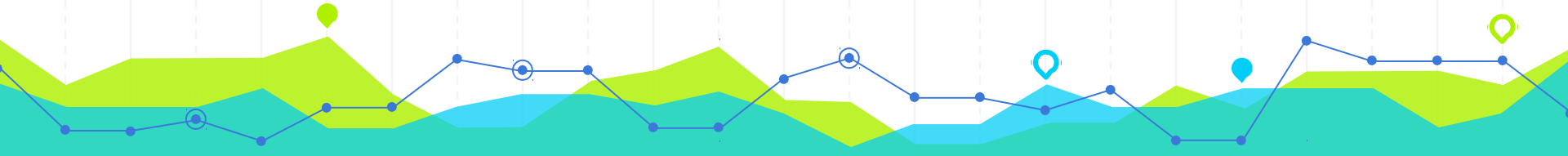
Development

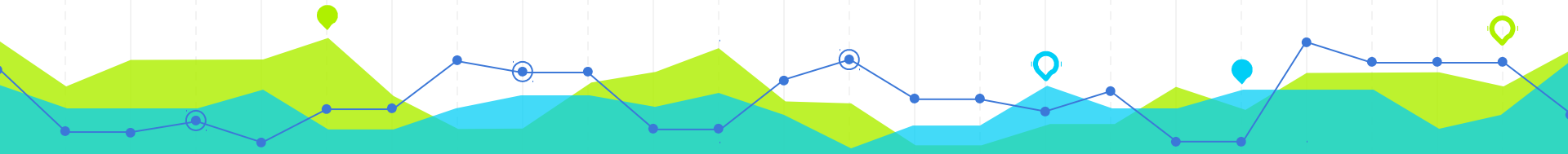
3

The process followed to complete the project

Project Task Structure

- Computer Model Implementation
- Data Exploration
- Prediction Algorithm Design and Implementation
- Prediction Adjustment
- Final Stage





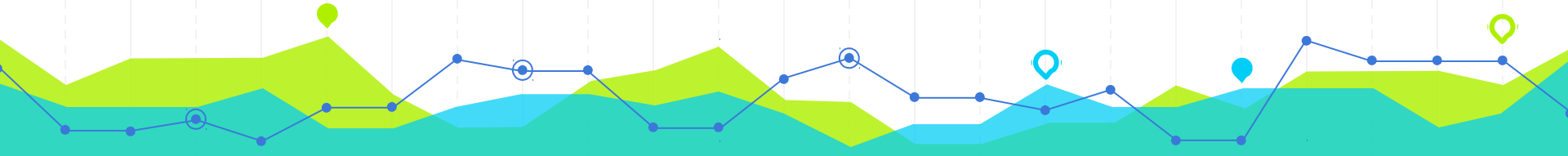
Computer Model Implementation

3.1

Build a computational model to start exploring the data

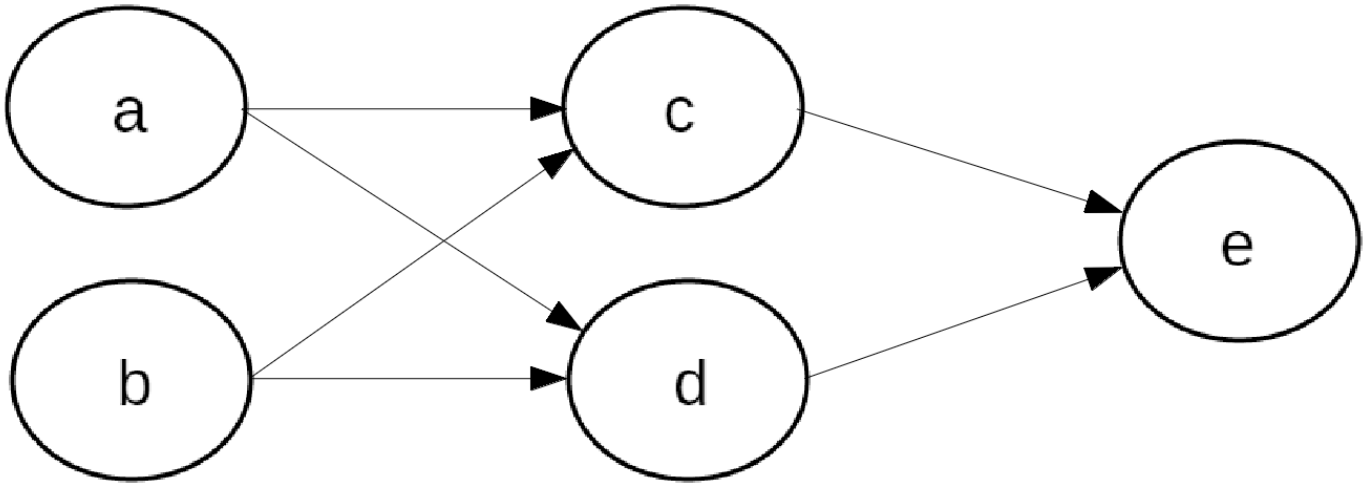
Data (I)

- A data set of relationships between proteins
- A data set of known interaction paths

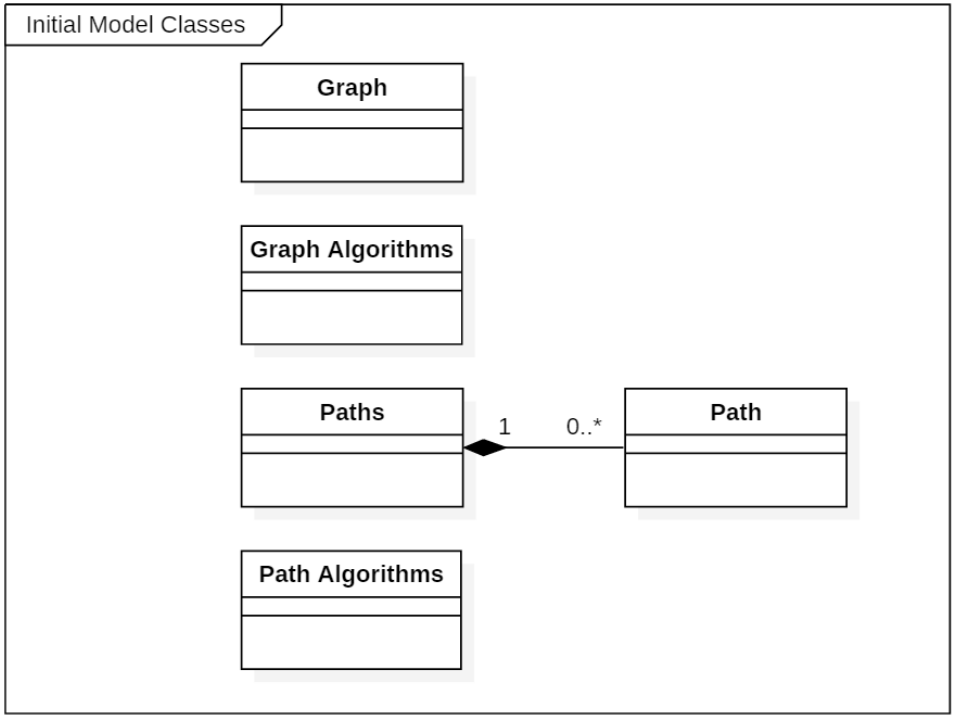


Paths Data Format

[a,b],[c,d],[e]

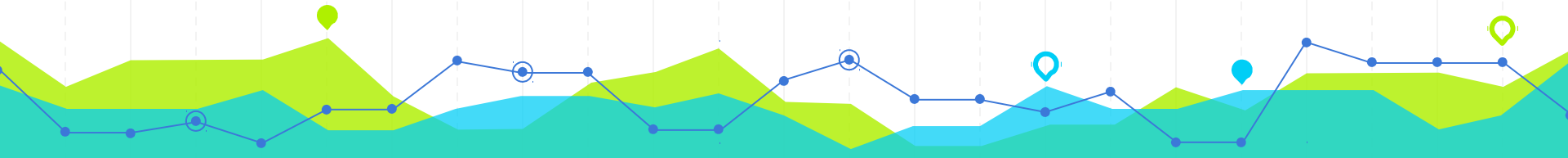


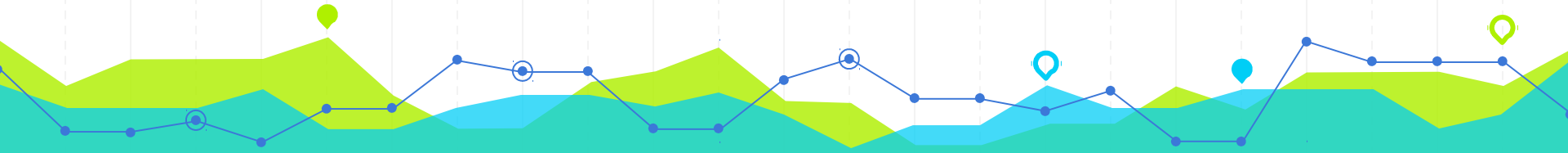
Initial Model Architecture



Involved Technologies

- C++ Programming Language (Sublime Text Editor)
- Git Technology





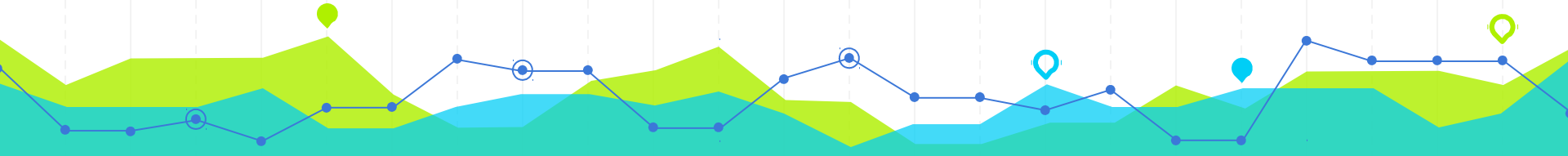
Data Exploration

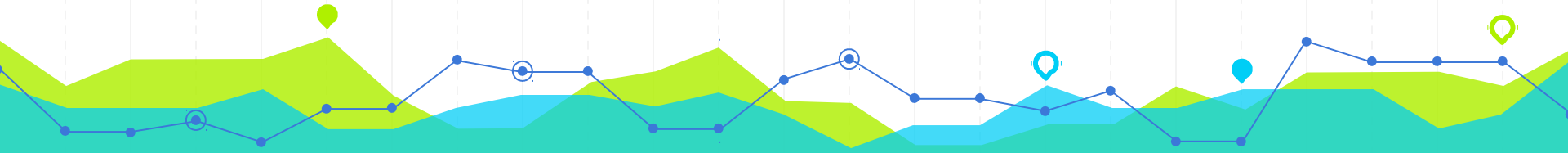
Find patters that can be exploit to make the prediction

3.2

Data Exploration Approaches

- Preliminary Exploration
- Transcription Factor Identification
- Graph Visualization
- Minimum Path
- Path Position Analysis





Preliminary Exploration

First Approach

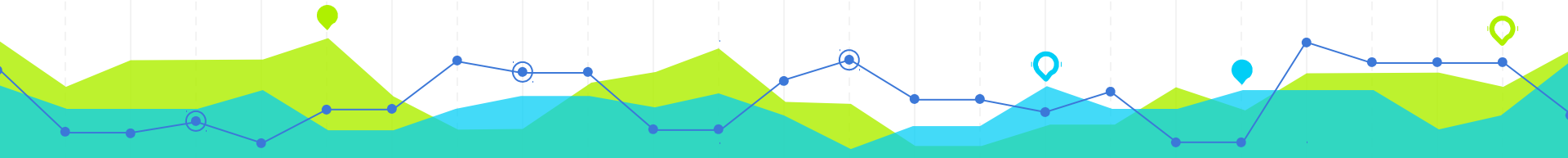
3.2.1

Measured Graph Indicators

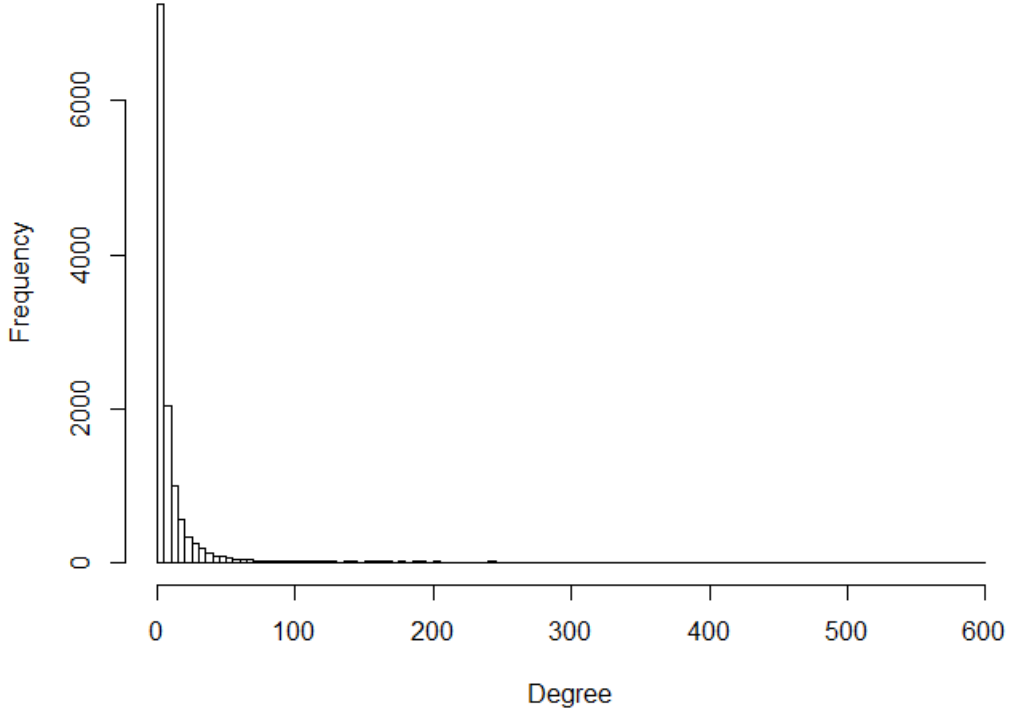
Degree
Number of edges incident to the vertex.

Betweenness Centrality
The number of times a node acts as a bridge along the single shortest path between two other nodes.

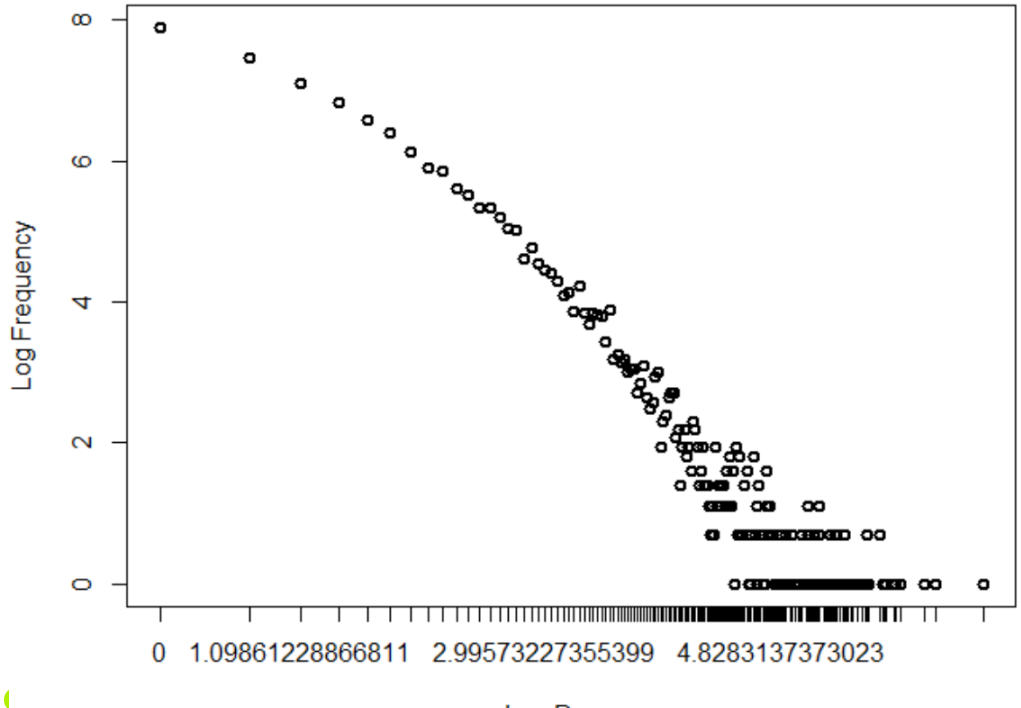
Closeness Centrality
The length of the shortest path between the node and all other nodes in the graph. The non normalized version is the sum of all the shortest path lengths.



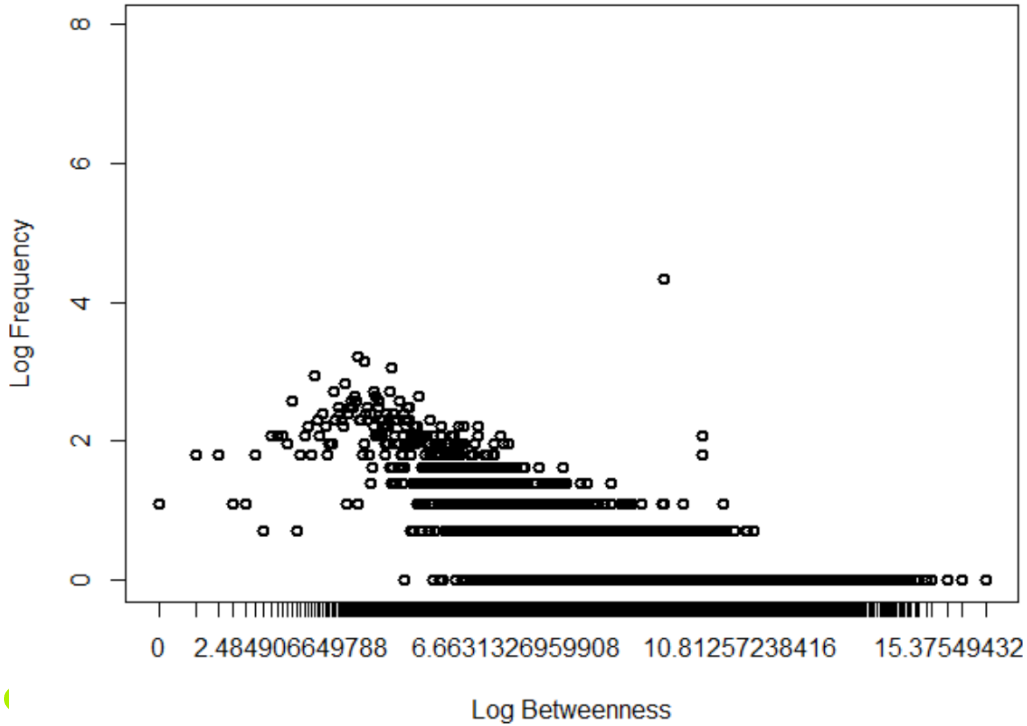
Unscaled Degree Hist.



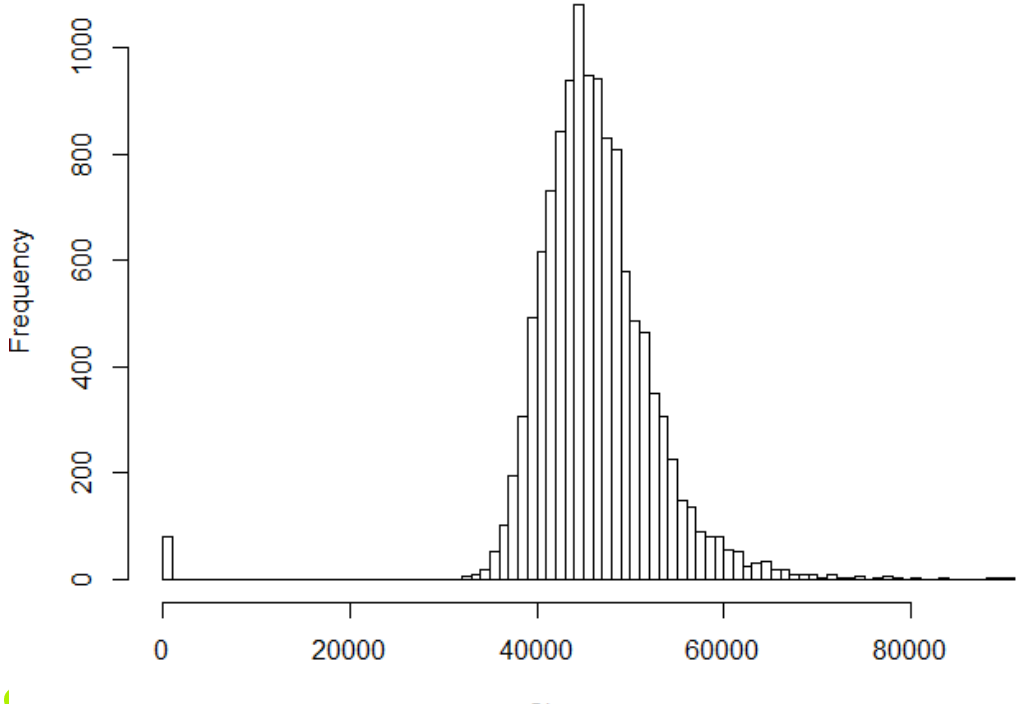
Frequency & Degree Plot



Freq. & Betweenness Plot

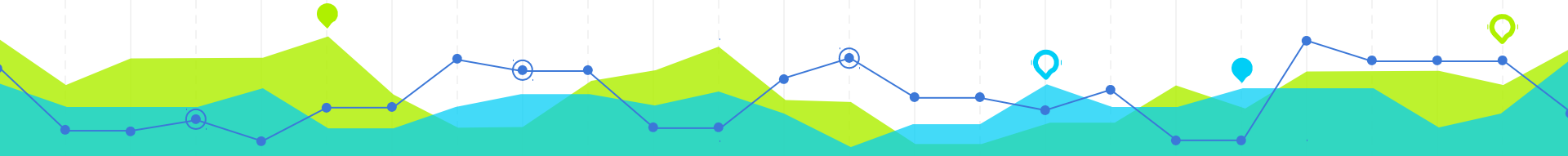


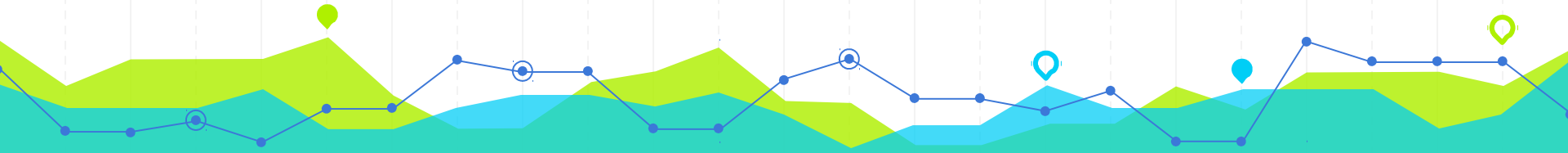
Closeness Histogram



Transcription Factor

A protein that controls the rate of transcription of genetic information from DNA to messenger RNA.





Trancription Factor Identification

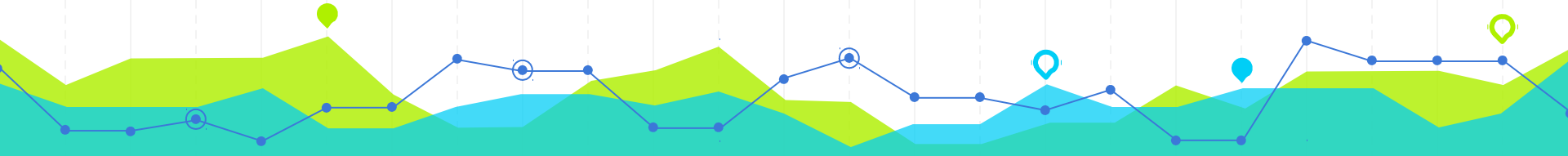
Second Approach

3.2.2

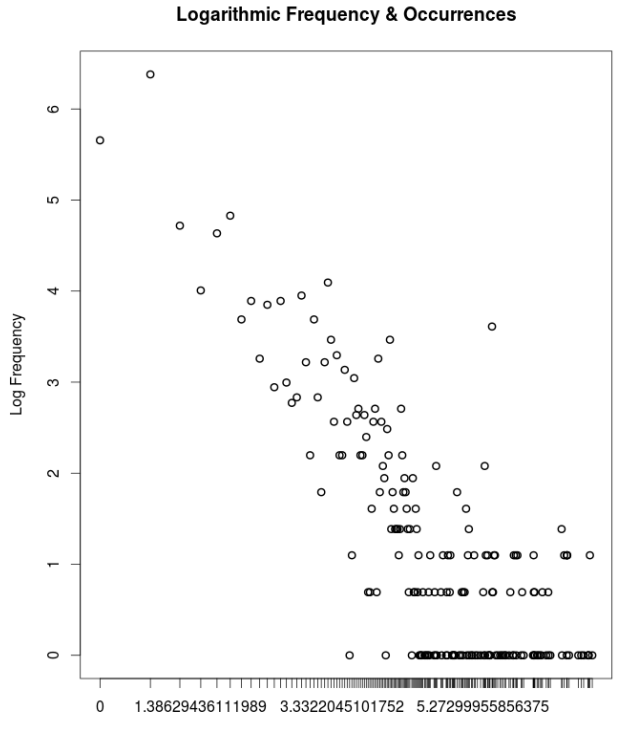
Measured Path Indicators

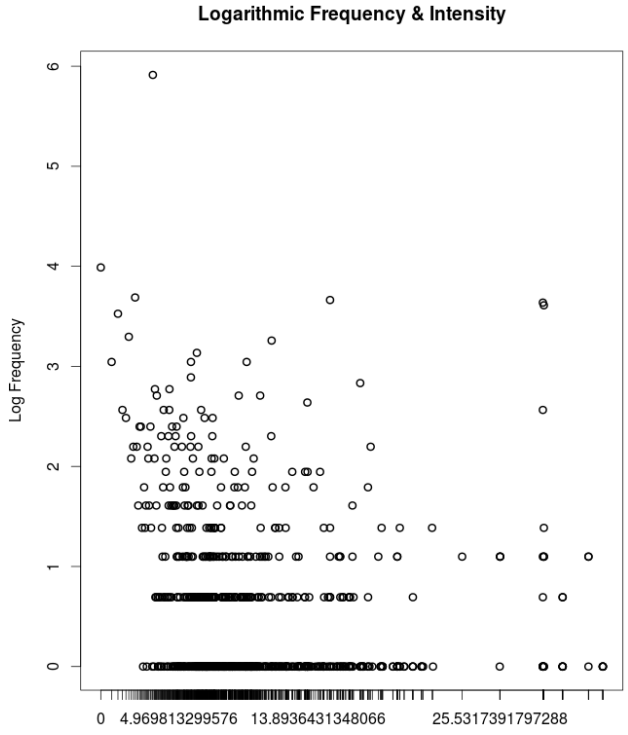
Ocurrences
The number of compressed paths where the protein is present. The combinatorics of the path are not regarded. (For example, lets measure the occurrence of 'a' in the path [a,b][c,d][e]: occurrence([a,b][c,d][e]) = 1)

Intensity
The number of uncompressed paths where the protein is present. The combinatorics of the path are taken into account (For example, lets measure the intensity of 'a' in the path [a,b][c,d][e]: intensity([a,b][c,d][e]) = 2x2x1 = 4)



Path Indicators Plots





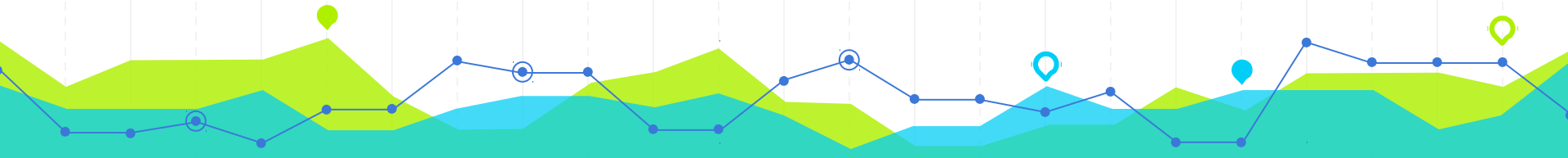
TF Candidate Lists

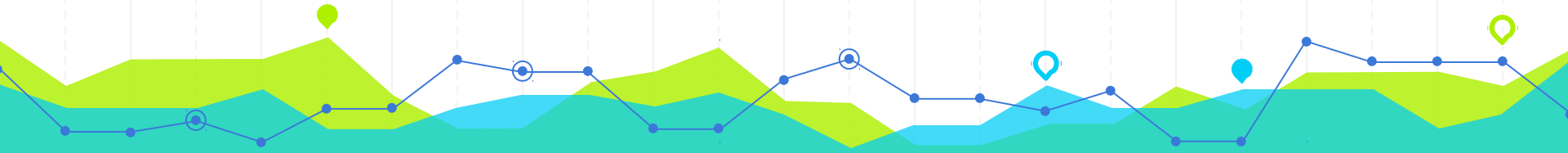
Top 50 outliers of the degree, betweenness centrality and closeness centrality of:

- Original Graph
- Paths Graph
- Fused Graph

The results were…

- Original Graph Candidates – No clear match with TFs
- Paths Graph Candidates – No clear match with TFs
- Fused Graph Candidates – No clear match with TFs





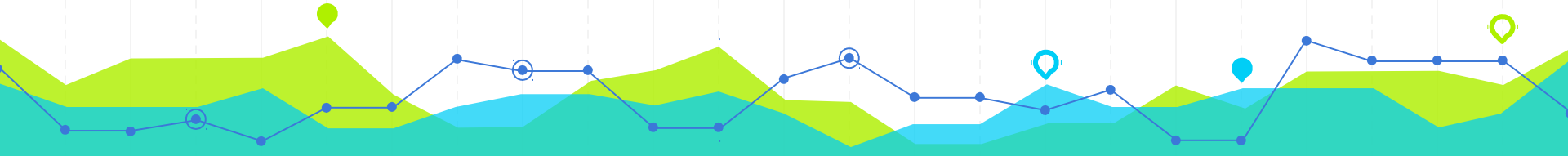
Graph Visualizatiom

Third Approach

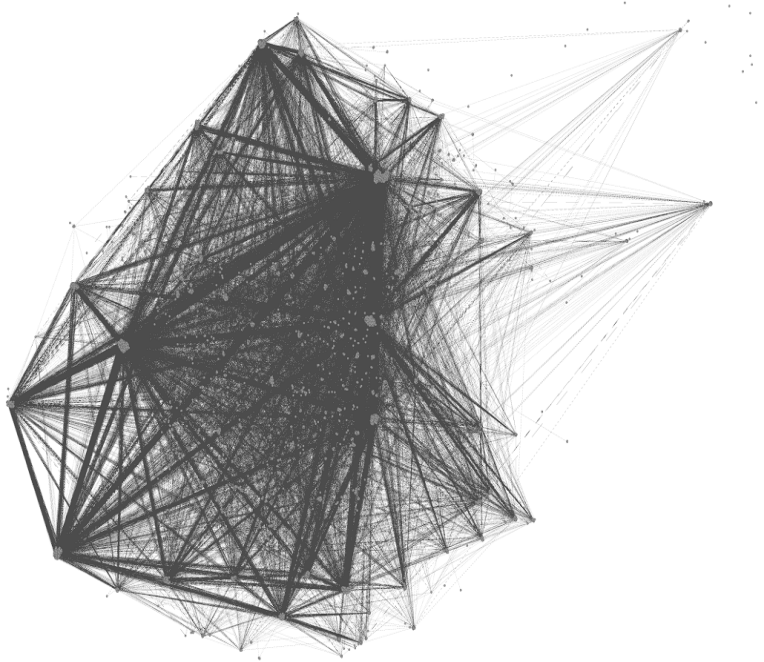
3.2.3

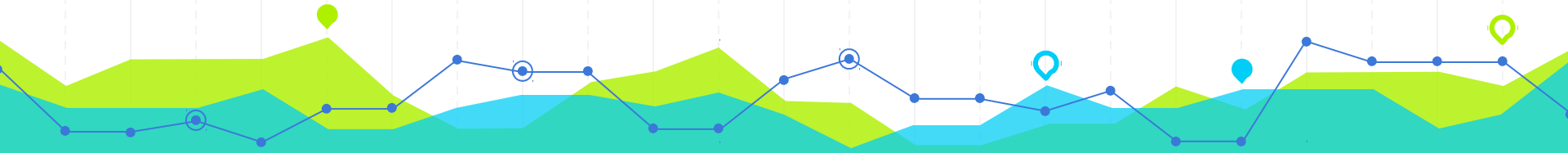
Graph Visualization Alg.

- Fruchterman-Reingold force-directed graph drawing alg.
- Yifan Hu Multilevel force-directed graph drawing alg.
- Forceatlas I
- Forceatlas II



Graph Visualization Algs. Result



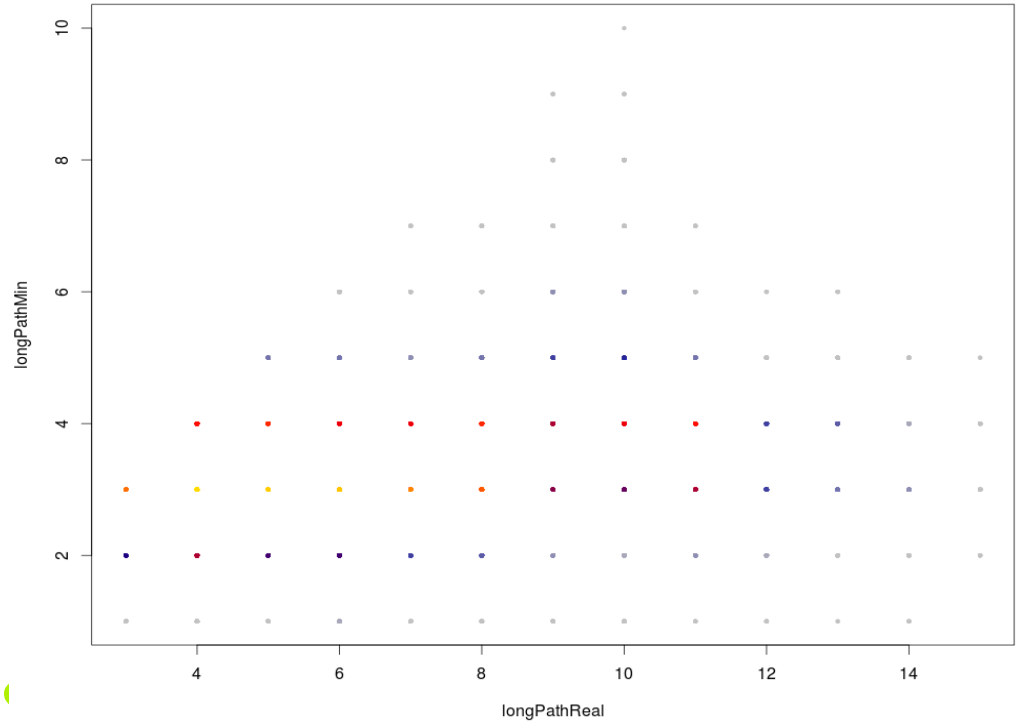


Minimum Path

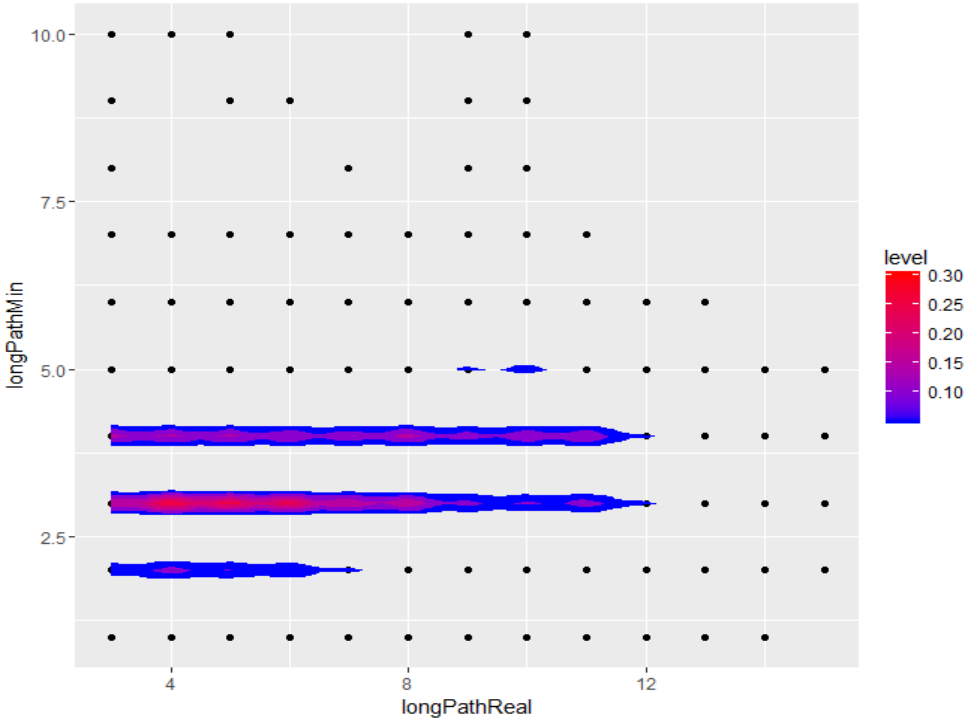
Fourth Approach

3.2.4

Heat scatter of lengths

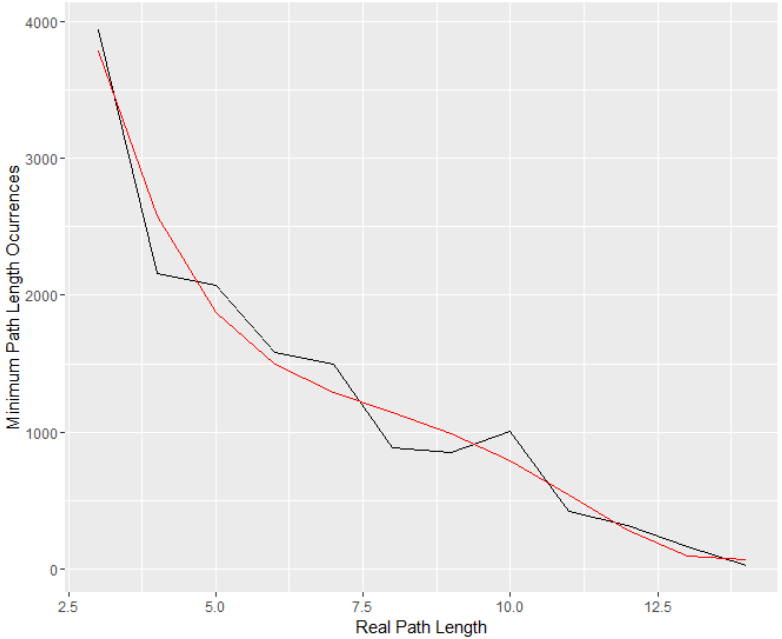


Heat map of lengths

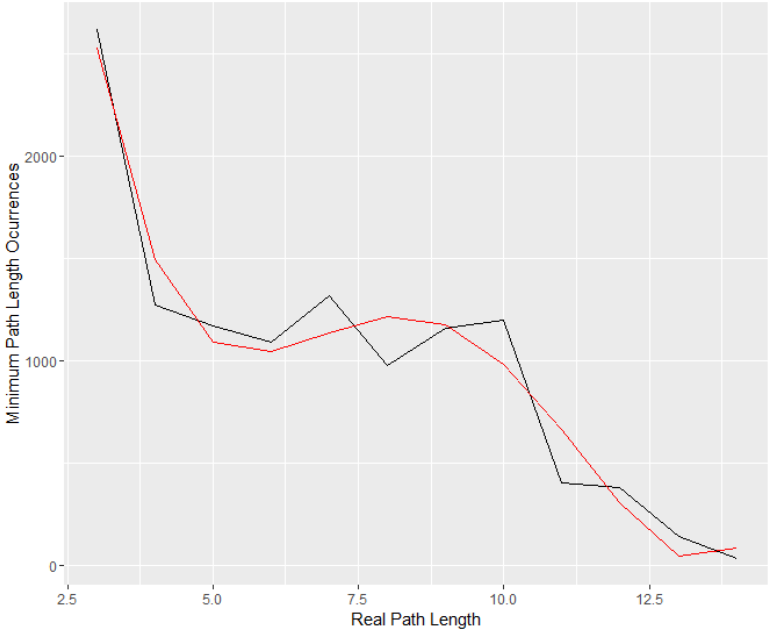


Polynomial Aprox of Path Length

Min Length 3

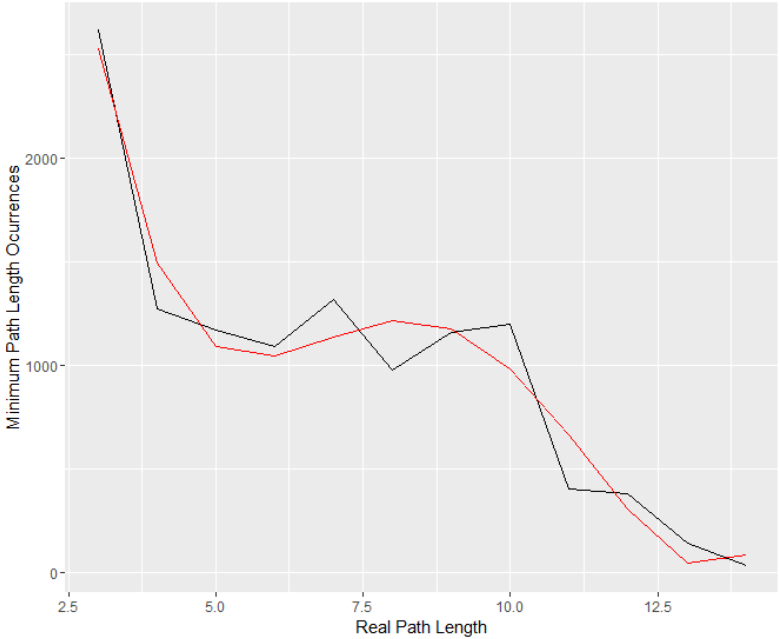


Min Length 4

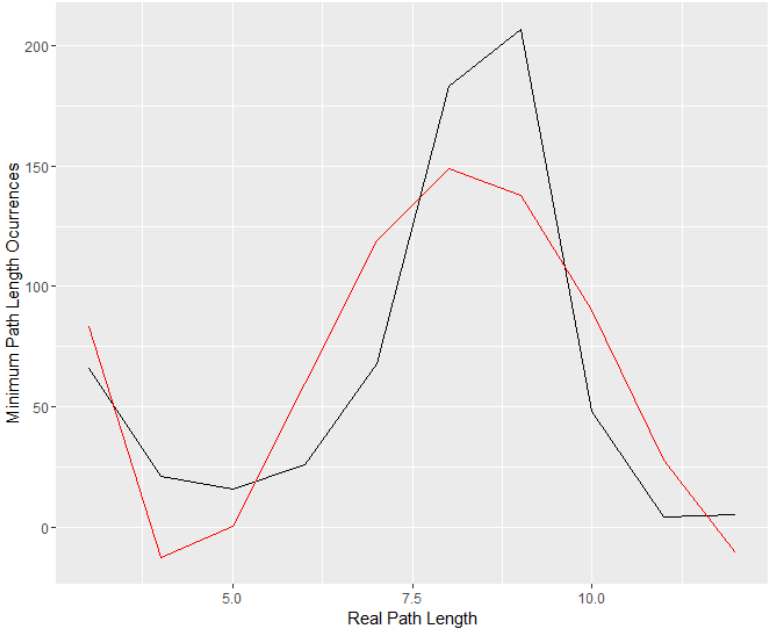


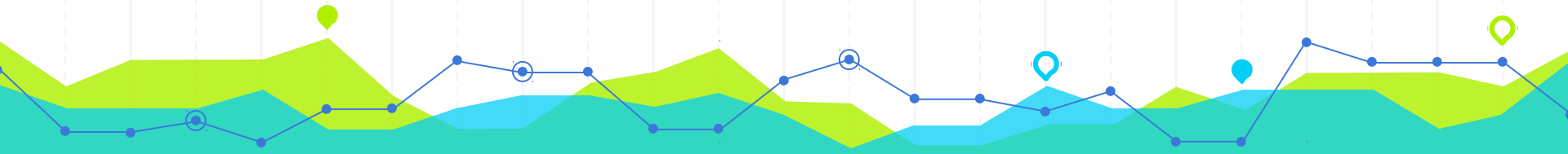
Polynomial Aprox of Path Length

Min Length 5



Min Length 6





Path Position Analysis

Fifth Approach

3.2.5

Start, Mid & End Decomposition

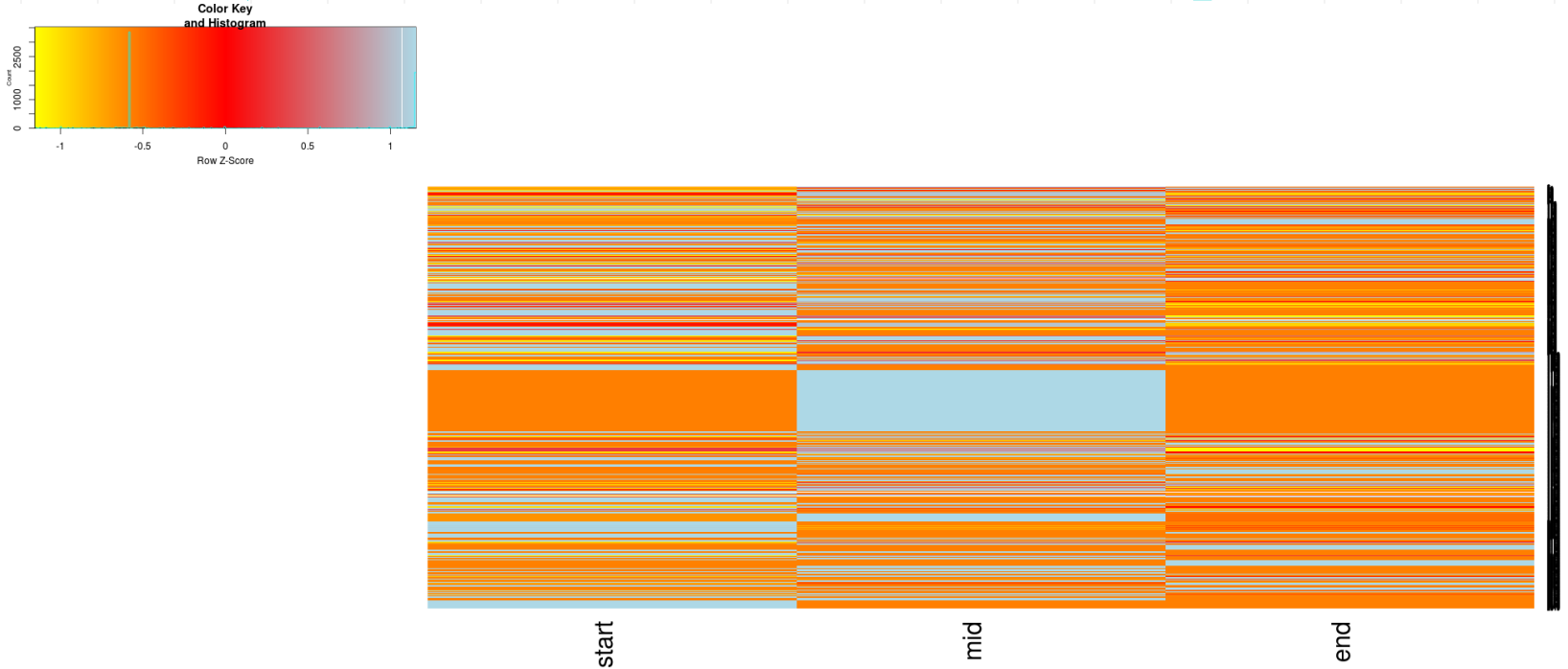
$$\text{Breaking Point } 1 = \lfloor pathSize * \tfrac{1}{3} \rfloor$$

$$\text{Breaking Point } 2 = \lceil pathSize * \tfrac{2}{3} \rceil$$

$$z = \frac{x - \mu}{\sigma}$$

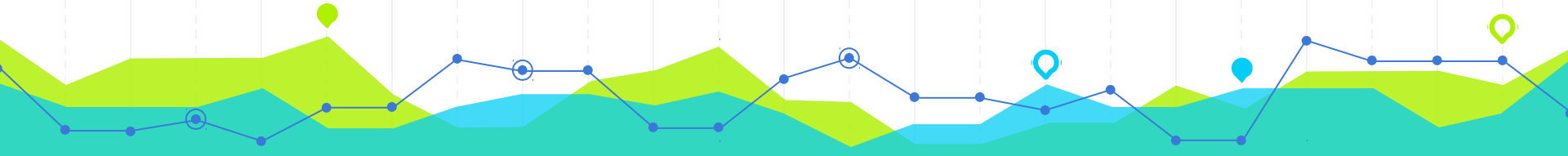
Where μ represents the mean and σ the standard deviation.

Start, Mid & End Decomposition



First Discovered Pattern

80% of the proteins accumulated more than 95% of its appearances in one position within the paths of the same length, and the other 20% accumulated more than 95% of its appearances in two position within the paths of the same length.



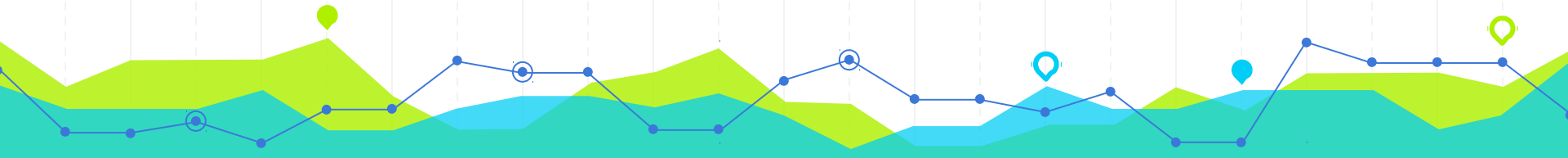
What about the paths of diff. length?

Using the Jaccard Coefficient we compared the different clusters of the different path lengths to find possible correlations..

$$J(A,B) = |A \cap B| / |A \cup B|$$

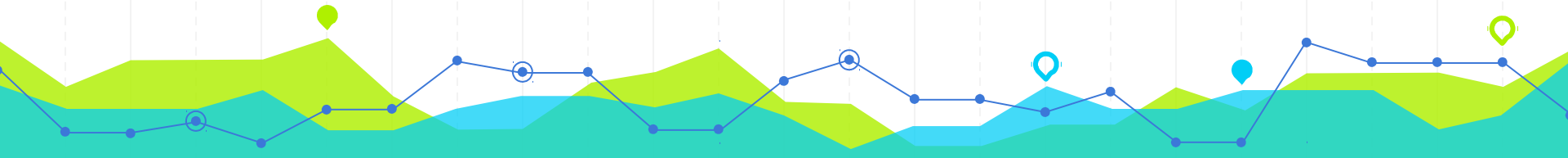
Second Discovered Pattern

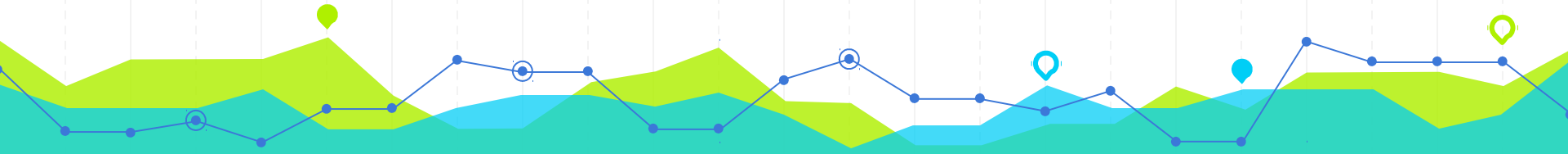
When a protein appeared more than 95% of the time in a position x within the paths of length y, in the case that this same protein also appeared in the paths of length y+1 it would appear in the same position or next to it [x-1,x,x +1] more than 90% of the time.



Involved Technologies

- C++ Programming Language (using OpenMP) written with the Sublime Text Editor)
- Rscripts (using ggplot, ggplot2 and LSD) were written using the Rstudio IDE
- Graph visualization tool Gephi
- Tested C++ Boost Graph Library





Prediction Algorithm Design and Implementation

3.3

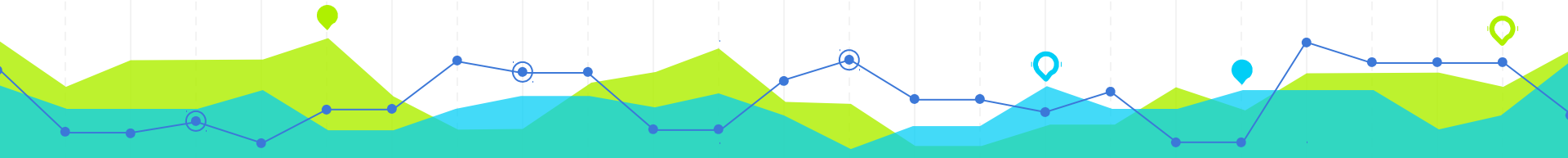
Design and Implement the prediction algorithm

Algorithm Steps

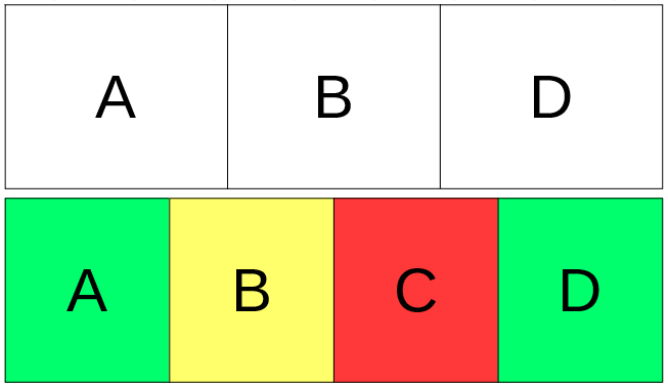
Generate candidate paths

Compute Fitness For Each Candidate Path

Sort the list from higher to lower fitness and return it



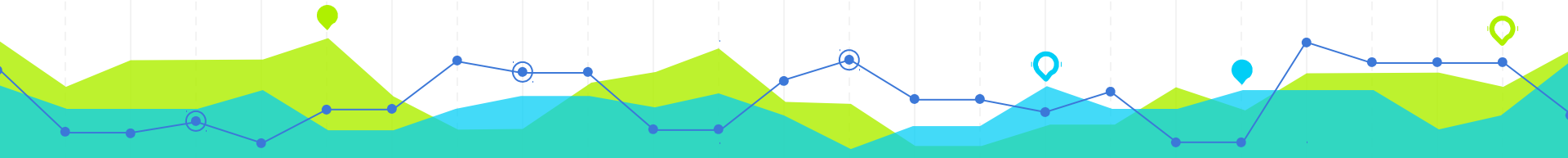
Fitness Evaluation

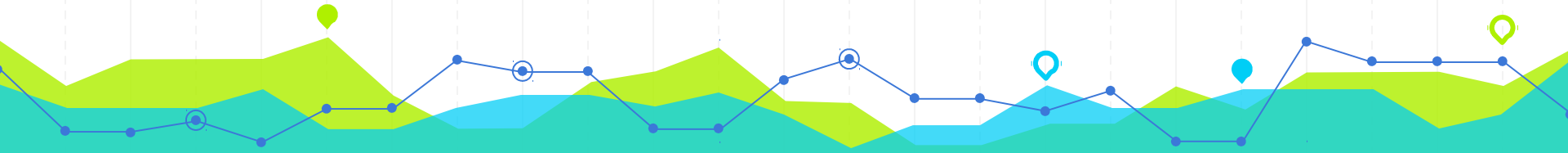


Example of path comparison scoring. A and D have a 100% match, while B has a %67 match. The total score result would be of (1.0+1.0+0.67)/4.0*10.0 .

Involved Technologies

- C++ Programming Language (using OpenMP) written with the Sublime Text Editor)
- Rscripts (using ggplot, ggplot2 and LSD) were written using the Rstudio IDE





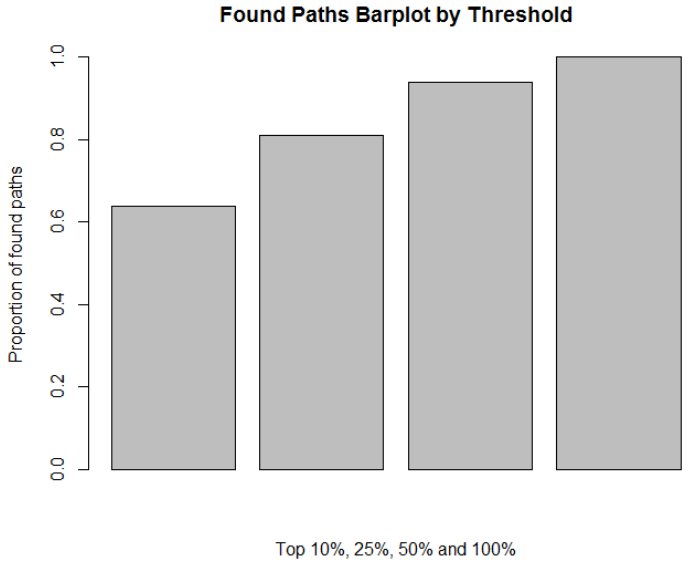
Prediction Adjustment

3.4

Validate the prediction algorithms and adjust it to increase its performance

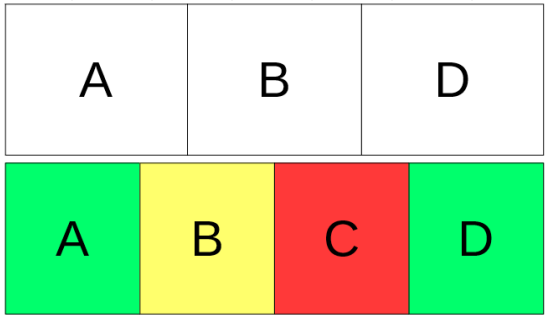
Prediction Alg. Validation

The method used was cross-validation of the path data set. The result of cross-validating (with a train/test proportion of 90%/10%) 20 times can be seen in the next figure:



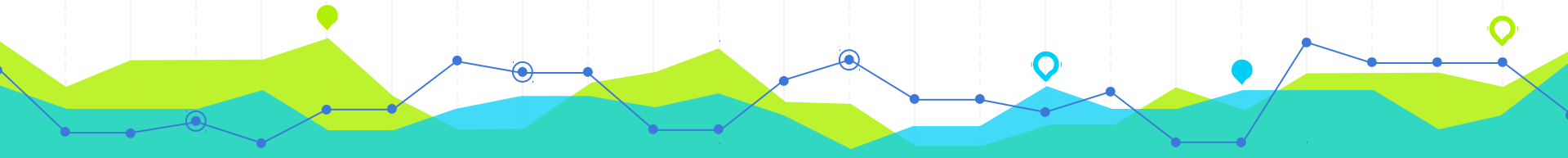
Adjustment Strategies

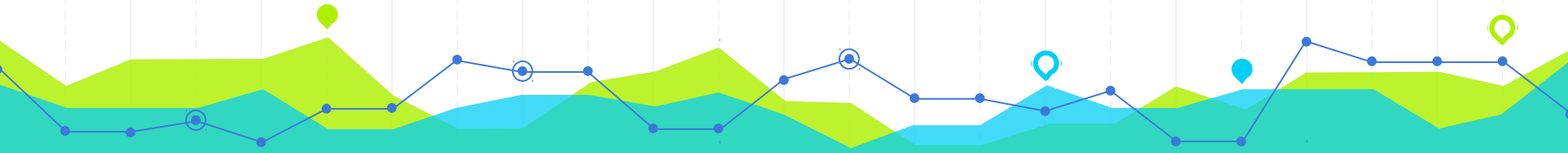
- Change the valid comparison distance. (Weighted and not weighted)
- Weighting the half-matches to increase and decrease its impact



Involved Technologies

- C++ Programming Language (using OpenMP) written with the Sublime Text Editor
- Rscripts (using ggplot, ggplot2 and LSD) were written using the Rstudio IDE



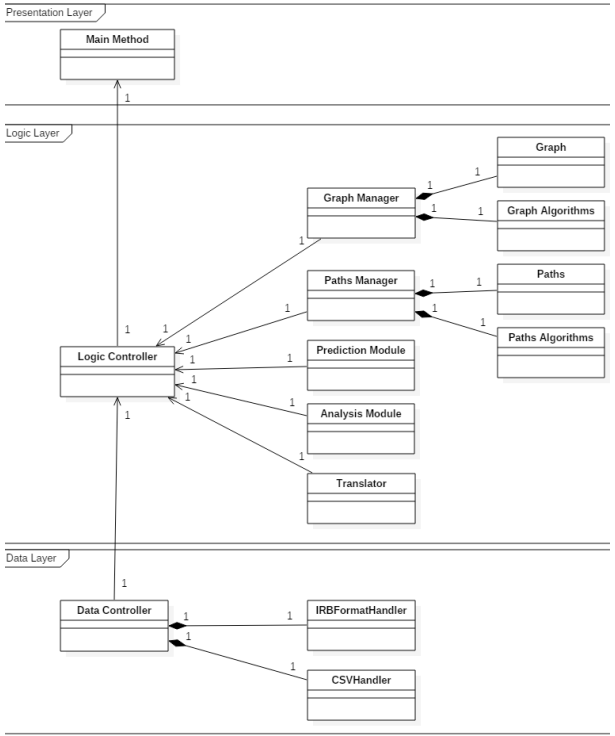


Final Stage

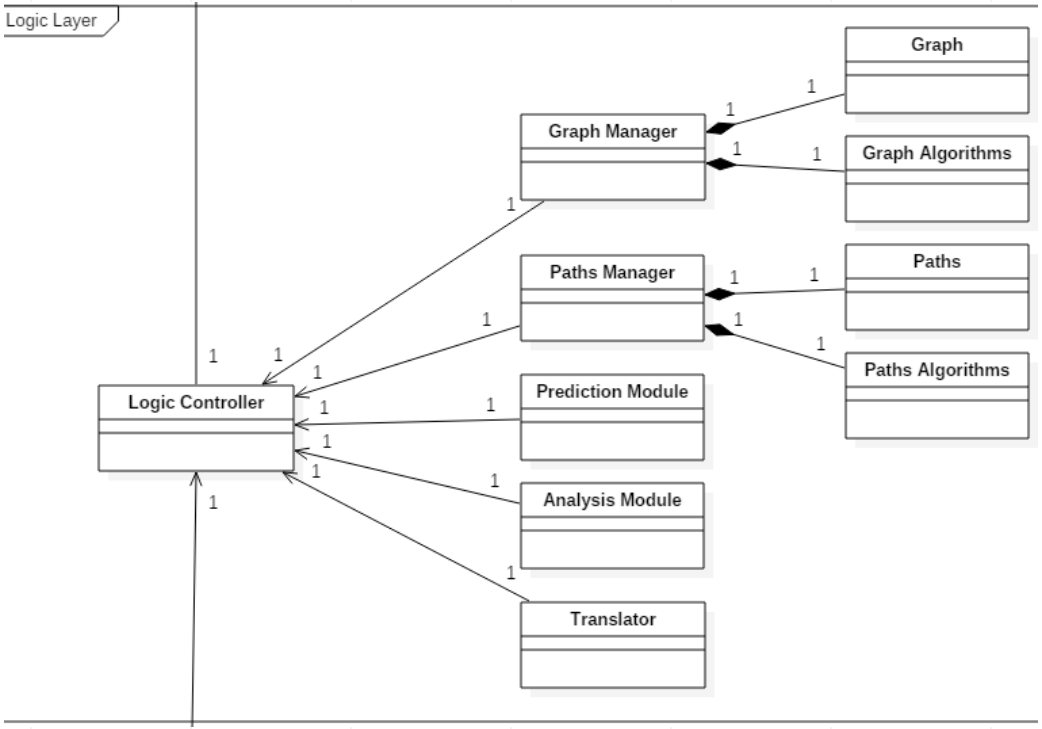
Implement the software, writte the memory
and prepare the project lecture

3.5

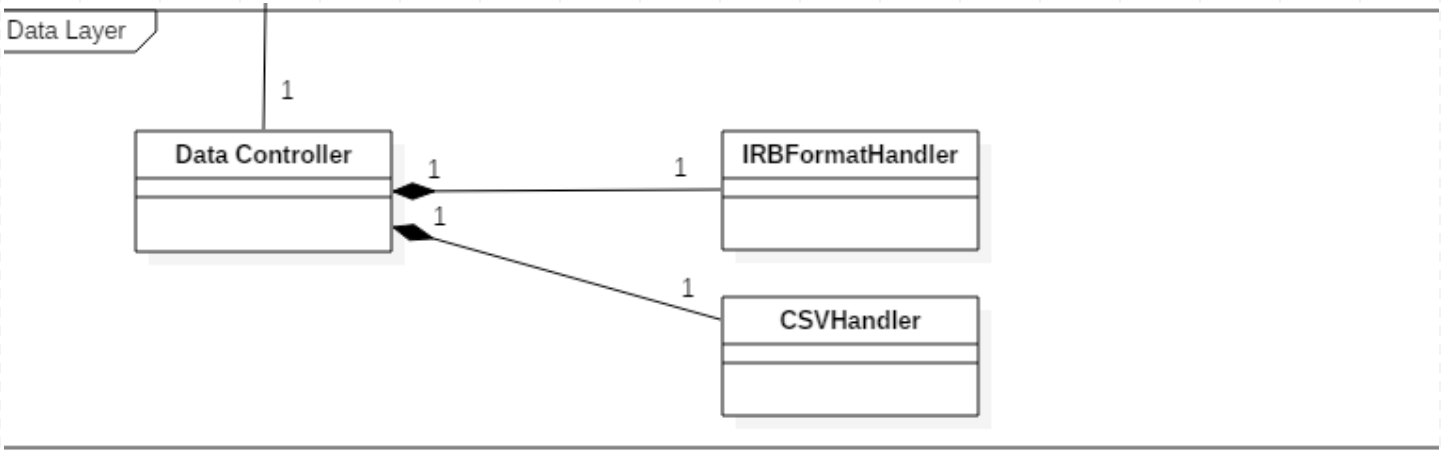
Software Architecture



Logic Layer Architecture

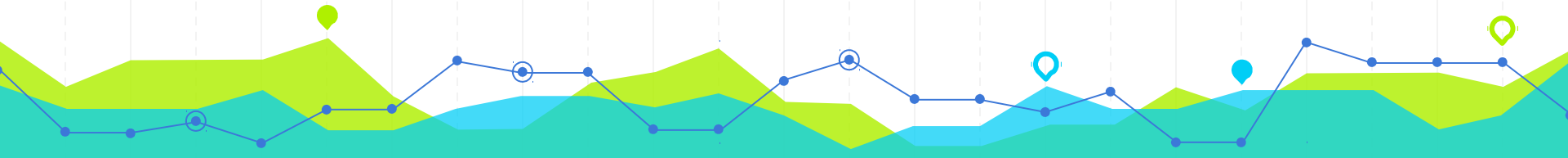


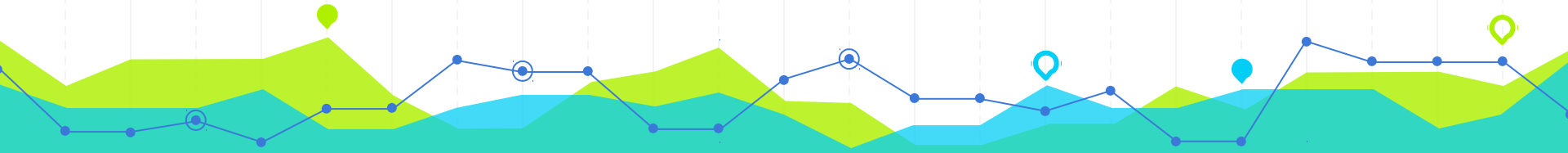
Data Layer Architecture



Involved Technologies

- C++ Programming Language (using OpenMP) written with the Sublime Text Editor
- The sharelatex web latex editor and the github git hosting services were used to write down the project memory and enable acces to the project's code
- Also, the starUML software was used to generate all the UML diagrams





Conclusions

A personal assessment of the project

4

Conclusions

Cross Specialty Communication
This project provided an opportunity to cooperate with people of a different background.

Coding
This project was a great coding practice.

Graph Data Mining
This project has served as a great introduction to graph data mining.

Research Project Type
This project encouraged a cientific mindset.

Computer Science
This project could be considered an ideal capstone project, since a wide array of C.S. techniques from different subject branches were required for its completion.

Data Visualization Techniques
This project encouraged to learn data visualization techniques beyond the degree curriculum .