# DESKGEN TASK REPORT

JEROME JAVELLE

## 1. Overview of the Project

In order to establish a link correlation between guide sequences and their activities at different times, with or without the presence of the drug, we suggest the following approach :

— cluster the data
— look for patterns in the guide sequences inside each cluster

## 2. Features and Implementation Choices

2.1. **Language and Framework.** In the purpose of having a reusable and maintainable code, we decide to implement our project using `C++` mainly. Although the development time is longer, the execution speed reachable with this language is several order of magnitudes faster than scripting languages or VM interpreted languages. This choice is also motivated by the many executions we would run in order to test several parameters for the machine learning algorithms used. We also use a python script to process the input data file once (no need for huge optimization since the task is of linear complexity and is executed only once).

2.2. **Clustering.** First, we must select a training set from the whole data set. We use a python script `selectData.py` to select a training set which presents noticeable bias by ignoring data points with similar guide counts in the different environments. Then, after reading the articles provided, we decide to use a variant of SVMs for unsupervised clustering called "Support Vector Clustering". We base our program on the algorithm found in the paper "Support Vector Clustering" by Ben-Hur, Horn, Sieglemann and Vapnik. The clustering algorithm goes as follows :

— Create the kernel matrix $Q = (k(x_i, x_j))_{i,j}$ where $k = e^{-q||x_i - x_j||_2^2}$ is the Gaussian kernel function of parameter $q$.
— Solve the optimization problem suggested in the SVC paper based on the Wolfe dual form $\sum_i k(x_i, x_i)\alpha_i - \sum_{i,j} k(x_i, x_j)\alpha_i\alpha_j$ using `dlib`.
— Compute the radius $R$ of the smallest enclosing sphere in the kernel space.
— Build graph of data points where 2 points are connected whenever the whole segment between both points are inside the sphere (20 sampled points).
— Find connected components of the graph that will constitute the clusters using `boost`.

2.3. **Cluster Analysis.**

## 3. Complexity Analysis

## 4. Results

## 5. Future Improvements

Next steps in this project would be :

- analyze the impact on the parameters of the clustering algorithm (gaussian kernel, soft margin)
- change the selection process for the training set
- consider other functions of the raw guide counts (other than activities)
- improve the performance of the cluster assignment step
- add more features to the statistical analysis of sequences
- change the way we chose the training set

In relation to this project, it may be useful to maintain internally a "clustering engine" with metrics to check how good the clustering is. We may even consider several clustering algorithms and have them compete with each other for 2 purposes :

- take the result which gives the best metrics (might depend on the initial data set)
- give more credit to a discovered pattern if it has been found by several fundamentally different algorithms

I would be curious to try clustering using "augmented" Minimum Spanning Tree because it would not be affected by the shape of the clusters, and it would allow to change the "granularity" of the clustering without having to run the algorithm again. Moreover, the complexity does not grow exponentially with the dimension of the data space but stays linear.