Sampling and Clustering of Phenotypic and Genotypic Soybean Dataset

A PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the degree

of
BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

Submitted by:
Mohit Mohta, Ankit Gaur and Suryaveer

Guided by:
Dr. Kapil Ahuja, Associate Professor, Discipline of CSE, IIT Indore

INDIAN INSTITUTE OF TECHNOLOGY INDORE
December 2018

Declaration of Authorship

We hereby declare that the project entitled "Sampling and Clustering of Phenotypic and Genotypic Soybean Dataset" submitted in partial fulfilment for the award of the degree of Bachelor of Technology completed under the supervision of Dr. Kapil Ahuja, Associate Professor, Computer Science and Engineering, IIT Indore is an authentic work.

Further, I/we declare that I/we have not submitted this work for the award of any other degree elsewhere.

Signed:

Mohit Mohta Ankit Gaur Suryaveer Singh

_____ _____ _____

Certificate

This is to certify that the B.Tech Project entitled, "*Sampling and Clustering of Phenotypic and Genotypic Soybean Dataset* " and submitted by Mohit Mohta, Ankit Gaur and Suryaveer Singh in partial fulfillment of the requirements of B.Tech Project embodies the work done by them under my supervision.

Supervisor

Dr. KAPIL AHUJA
Associate Professor,
Indian Institute of Technology Indore
Date:

Acknowledgements

It is our privilege to express our gratitude to several persons who helped us directly or indirectly to conduct this research project work. We express our heart full indebtedness to our BTP guide **Dr. Kapil Ahuja** for his sincere guidance and inspiration in completing this Project.

We are extremely thankful to **Mr. Aditya Anand Shastri** for his coordination and cooperation and for his kind guidance and encouragement.

We also thank our friends who have more or less contributed to the making of this project.

This study has indeed helped us to explore more knowledgeable avenues related to this topic and we are sure it will help us in future.

INDIAN INSTITUTE OF TECHNOLOGY INDORE

Abstract

Department of Computer Science and Engineering

Bachelor of Technology

Sampling and Clustering of Phenotypic and Genotypic Soybean Dataset

Soybean forms an important cash crop for the Indian and French economy. It is the third largest cultivated crop in India. In France, the amount of area being used for Soybean farming is going up from 122,000 hectares in 2015 to about projected 200,000 hectares in 2020.

In the Indian context, we aim to develop a species of Soybean that is drought and heat resistant (because of the erratic monsoons over this past decade) and in the French context, we need to develop the species of Soybean that are resistant to flooding/too much rainfall and extreme cold (because of increased frequency of arctic blasts happening over the past many years). We break down our work in two parts. The first part deals with the phenotypic data and finds correlations between different phenotypic factors. Using this phenotypic data, we also cluster plants with similar phenotypic features together and then finally, take the best species out of each cluster. We have then also applied pivotal sampling techniques to reduce the dimensionality of the data.

Basically, this project can be divided into the following parts:

- Correlations between several phenotypic factors
- Spectral clustering to cluster plants based on their phenotypic properties
- Sampling of the whole genome sequence of soybean to reduce the dimensionality of the data using Pivotal Sampling
- Clustering of sampled whole genome sequences using Spectral Clustering

Contents

List of Tables

List of Figures

Part I

Study and clustering based on Phenotypic data for Soybean

Chapter 1

Introduction

1.1 About Phenotypic Data

A phenotype is the composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical or physiological properties, behavior, and products of behavior (such as a bird's nest). In context of Soybean, we have been provided a phenotypic data set that has observable characteristics of many different species of Soybean plant. We were given several properties mapped to a specific species. For eg, early plant vigour, Hypocotyl color, stem determination, days to 50% flowering, flower color, leaf shape, leaflet color, number of leaflets, seed yield per plant, 100 seed weight and many more.

1.2 About some phenotypic properties of Soybean plant

- Early plant vigour: The strength of plant in its early days.
- Hypocotyl color: The part of the stem of an embryo plant beneath the stalks of the seed leaves or cotyledons and directly above the root is Hypocotyl.
- Days to 50% flowering
- Flower Color
- Leaf Shape and color
- Number of Leaflets
- Pubescence: Fine short hair on plant stem - present or not. It's color and density, type (erect, semi-appressed etc) are other related properties.
- Plant height, Number of Primary and Secondary Branches.
- Lodging: the displacement of stems or roots from their vertical and proper placement.

- Pod color, Seeds per pod, Number of Pods per plant, Days to pod initiation
- Days to 80% maturity
- Seed coat color
- Hilum color: the scar on a seed marking the point of attachment to its seed vessel
- 100 seed weight (g)
- Seed yield per plant (g)
- Shattering Score: Estimated percent of pods open 2 weeks after harvest. All are NaN in the given data.

1.3 Aims to achieve with the phenotypic data

- Finding correlation between different phenotypic factors. This will give us a very good understanding about what factor is related to what other factor and how strong is the correlation! Moreover, this can also help to understand what factors can be dropped during clustering. For instance, if two of the parameters are highly correlated then in that case, we can drop one of the parameters and that will be fine because the features were strongly correlated and so taking measure of any one of them will also enclose the information about the other within.
- Finding similarity between different species. This will give us a better understanding on how two species are close to each other depending on the phenotypic factors. In order to find the similarity for the given data, application of a clustering algorithm was best suited. Find similarity based on phenotypic factors will help us understand the best species based on yield, if all other phenotypic factors are the same. So, before clustering, we dropped yield as a factor and so within each cluster we can assume that all the species have more or less similar physical properties and so now we can easily select the best yielding plant amongst them.

Chapter 2

Literature Review

There has been a lot of research ongoing in the field of genomics and a lot of efforts are being made to attain better accuracy to map the phenotypic and the genomic characteristics. As in this part, we dealt only with the phenotypic data so the chapter discusses literature pertaining to previously known methods of finding correlations among several factors and about spectral clustering.

2.1 Study and Correlations of Phenotypic Properties of Soybean

Correlation is simply the degree of association between two variables. The Pearson's correlation coefficient is the measure of the linear association between the two variable for which it is being calculated. When the x variable is a random covariate to they variable, that is, x and y vary together (continuous variables), we are more interested in determining the strength of the linear relationship than in prediction, and the sample correlation coefficient, r_{xy}, is the statistics employed for this purpose. [1]

The Pearson (Product–Moment) correlation r was developed by Pearson (1896) and was based on the work of others, including Galton (1888), who first introduced the concept of correlation [2] [3]. As a matter of fact, correlation charts, also known as scatter diagrams is one of the seven basic tools of statistical quality control.[4]. Although non-linear relationships are fundamental to most physical and statistical phenomena, r is not appropriate in these cases and may provide false results for non-linear relationships. In such situations, you might need to perform data transformations in order to linearize your variables first.

Correlation and covariance have a very vital role in performing clustering; correlation is therefore taken as a measure for calculating the similarity between pairwise objects.

Factor analysis, behavioural genetic models, structural equations models and other related methodologies use the correlation coefficient as the basic unit of data. [5]

There are a number of different correlation coefficients to handle the special characteristics of such types of variables as dichotomies, and there are other measurements of association for nominal and ordinal variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

The value of the Pearson's correlation coefficient lie between -1 to 1. Correlations equal to 1 or 1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation). This means, that if the value of the correlation coefficient is -1, the variables are highly negatively correlated, i.e if one increases other decreases and if 1, then the variables are highly positively correlated i.e if one increases the other decreases.

If two random variables x and y are statistically independent, their correlation coefficient is zero. However, the converse is not true; i.e., if r = 0, this does not necessarily imply that x and y are statistically independent. The correlation coefficient is thus an estimate of association between the variables and is valid only when the observations are randomly drawn. Many statistical software packages include a program for such calculation and the correlation coefficient r, is routinely printed out in connection with other statistical parameters.

We will explain about how we have made used of this important statistical parameter to draw out correlations between the phenotypic properties later.

2.2 Spectral Clustering

Clustering is one of the most widely used techniques for exploratory data analysis, with applications ranging from statistics, computer science, biology to social sciences or psychology. In virtually every scientific field dealing with empirical data, people attempt to get a first impression on their data by trying to identify groups of "similar behavior" in their data. Compared to the "traditional algorithms" such as k-means or single linkage, spectral clustering has many fundamental advantages. Results obtained by spectral

clustering often outperform the traditional approaches, spectral clustering is very simple to implement and can be solved efficiently by standard linear algebra methods. [6]

The spectral clustering method can basically be divided in three parts:

2.2.1 Constructing the similarity graph

Constructing the similarity graph for spectral clustering is not a trivial task, and little is known on theoretical implications of the various constructions.

Thinking of a similarity function Before we move on to construct a similarity graph, we need to define a similarity function on the data. As we are going to construct a neighborhood graph later on, it is important for us to ensure that the local neighborhoods induced by this similarity function make sense. That is why, we need to be sure that points which are considered to be "very similar" by the similarity function are also closely related in the application the data comes from. For example, to construct a similarity function between text documents it makes sense to check whether documents with a more similarity indeed belong to the same text category. The global "long-range" behavior of the similarity function is not so important for spectral clustering — it does not really matter whether two data points have similarity score 0.01 or 0.001, say, as we will not connect those two points in the similarity graph anyway. Ultimately, the choice of the similarity function depends on the domain the data comes from, and no general advice can be given.

Choosing the similarity graph The next choice one has to make concerns the type of the graph one wants to use, such as the k-nearest neighbor or the ϵ-neighborhood graph. The following figure illustrates the behavior of the different graphs:

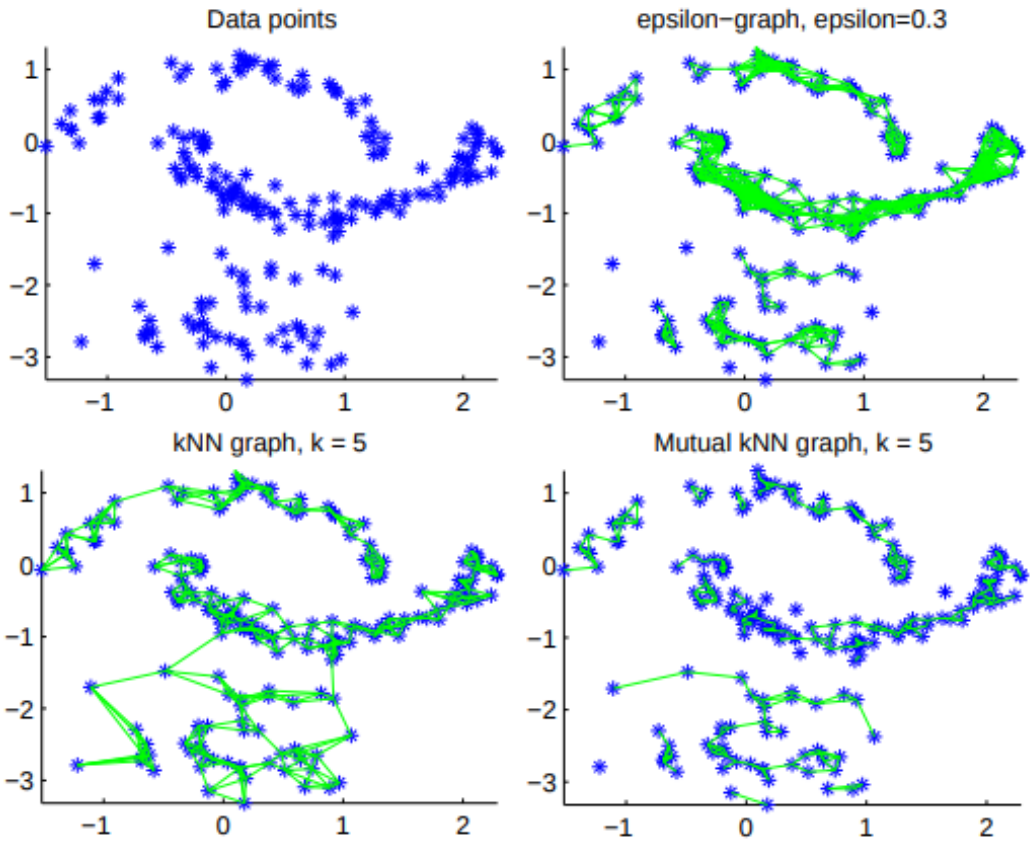

FIGURE 2.1: Different similarity graphs

In the ϵ-neighborhood graph, we can see that it is difficult to choose a useful parameter ϵ. With $\epsilon = 0.3$ as in the figure, the points on the middle moon are already very tightly connected, while the points in the Gaussian are barely connected. This problem always occurs if we have data "on different scales", that is the distances between data points are different in different regions of the space.

The k-nearest neighbor graph, on the other hand, can connect points "on different scales". We can see that points in the low-density Gaussian are connected with points in the high-density moon. This is a general property of k-nearest neighbor graphs which can be very useful. We can also see that the k-nearest neighbor graph can break into several disconnected components if there are high density regions which are reasonably far away from each other. This is the case for the two moons in this example.

The mutual k-nearest neighbor graph has the property that it tends to connect points within regions of constant density, but does not connect regions of different densities with each other. So the mutual k-nearest neighbor graph can be considered as being "in

between" the ϵ-neighborhood graph and the k-nearest neighbor graph. It is able to act on different scales, but does not mix those scales with each other. Hence, the mutual k-nearest neighbor graph seems particularly well-suited if we want to detect clusters of different densities.

2.2.2 How to compute the eigenvectors?

Now, let k be the number of clusters that we have decided to make. To implement spectral clustering in practice one has to compute the first k eigenvectors of a potentially large graph Laplace matrix. Luckily, if we use the k-nearest neighbor graph or the ϵ-neighborhood graph, then all those matrices are sparse. Efficient methods exist to compute the first eigenvectors of sparse matrices, the most popular ones being the power method or Krylov subspace methods such as the Lanczos method[7]. The speed of convergence of those algorithms depends on the size of the eigengap (also called spectral gap). The larger this eigengap is, the faster the algorithms computing the first k eigenvectors converge.

Note that a general problem occurs if one of the eigenvalues under consideration has multiplicity larger than one. For example, in the ideal situation of k disconnected clusters, the eigenvalue 0 has multiplicity k. As we have seen, in this case the eigenspace is spanned by the k cluster indicator vectors. But unfortunately, the vectors computed by the numerical eigensolvers do not necessarily converge to those particular vectors. Instead they just converge to some orthonormal basis of the eigenspace, and it usually depends on implementation details to which basis exactly the algorithm converges.

2.2.3 The K-means step

K-means is the last step for the sprectral clustering algorithm. This step is used to take out the final partition from the real-valued matrix of eigenvectors obtained in the step above.

While it is somewhat arbitrary what clustering algorithm exactly one chooses in the final step of spectral clustering, one can argue that at least the Euclidean distance between the points y_i is a meaningful quantity to look at. We have seen that the Euclidean distance between the points y_i is related to the "commute distance" on the graph, and in Nadler, Lafon, Coifman, and Kevrekidis (2006) [8] the authors show that the Euclidean distances between the y_i are also related to a more general "diffusion distance".

Chapter 3

Analysis of different phenotypic factors

3.1 Pearson Correlation Coefficients for Phenotypic Traits

- The Pearson correlation coefficient is just one of many types of correlation coefficients in the field of statistics.
- In order to determine how strong the relationship is between two variables, a formula must be followed to produce what is referred to as the coefficient value.
- The coefficient value can range between -1.00 and 1.00. If the coefficient value is in the negative range, then that means the relationship between the variables is negatively correlated, or as one value increases, the other decreases. If the value is in the positive range, then that means the relationship between the variables is positively correlated, or both values increase or decrease together.
- To find correlation between x and y:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

Where x and y are any two columns and i is the row number.

TABLE 3.1: Pearson Correlation Coefficients for Numerical Phenotypic Traits

	Days to 50% flowering	Plant height	Number of primary branches	Number of pods per plant	Days to 80% maturity	Number of seeds per pod	100 seed weight (g)	Seed yield per plant (g)	Days to pod initiation	Number of node per plant	Number of seeds per plant
Days to 50% flowering	1.000000	0.469065	0.236088	0.247986	0.473682	-0.068382	-0.585372	-0.255696	0.921882	0.514428	-0.050327
Plant height	0.469065	1.000000	-0.104849	0.035972	0.229953	-0.147207	-0.330333	-0.182647	0.436083	0.688646	-0.064142
Number of primary branches	0.236088	-0.104849	1.000000	0.616234	0.103023	0.197647	-0.158245	0.212007	0.252345	0.244131	0.268456
Number of pods per plant	0.247986	0.035972	0.616234	1.000000	0.082830	0.101657	-0.157316	0.332864	0.245497	0.389958	0.420420
Days to 80% maturity	0.473682	0.229953	0.103023	0.082830	1.000000	-0.066647	-0.379114	-0.248644	0.510874	0.232718	-0.139851
Number of seeds per pod	-0.068382	-0.147207	0.197647	0.101657	-0.066647	1.000000	0.111992	0.265555	-0.056746	-0.011394	0.226616
100 seed weight (g)	-0.585372	-0.330333	-0.158245	-0.157316	-0.379114	0.111992	1.000000	0.419121	-0.580759	-0.378506	0.000335
Seed yield per plant (g)	-0.255696	-0.182647	0.212007	0.332864	-0.248644	0.265555	0.419121	1.000000	-0.237122	-0.061864	0.849714
Days to pod initiation	0.921882	0.436083	0.252345	0.245497	0.510874	-0.056746	-0.580759	-0.237122	1.000000	0.503426	-0.038355
Number of node per plant	0.514428	0.688646	0.244131	0.389958	0.232718	-0.011394	-0.378506	-0.061864	0.503426	1.000000	0.081939
Number of seeds per plant	-0.050327	-0.064142	0.268456	0.420420	-0.139851	0.226616	0.000335	0.849714	-0.038355	0.081939	1.000000

3.2 Visualization of Correlation Matrix using Heat-map

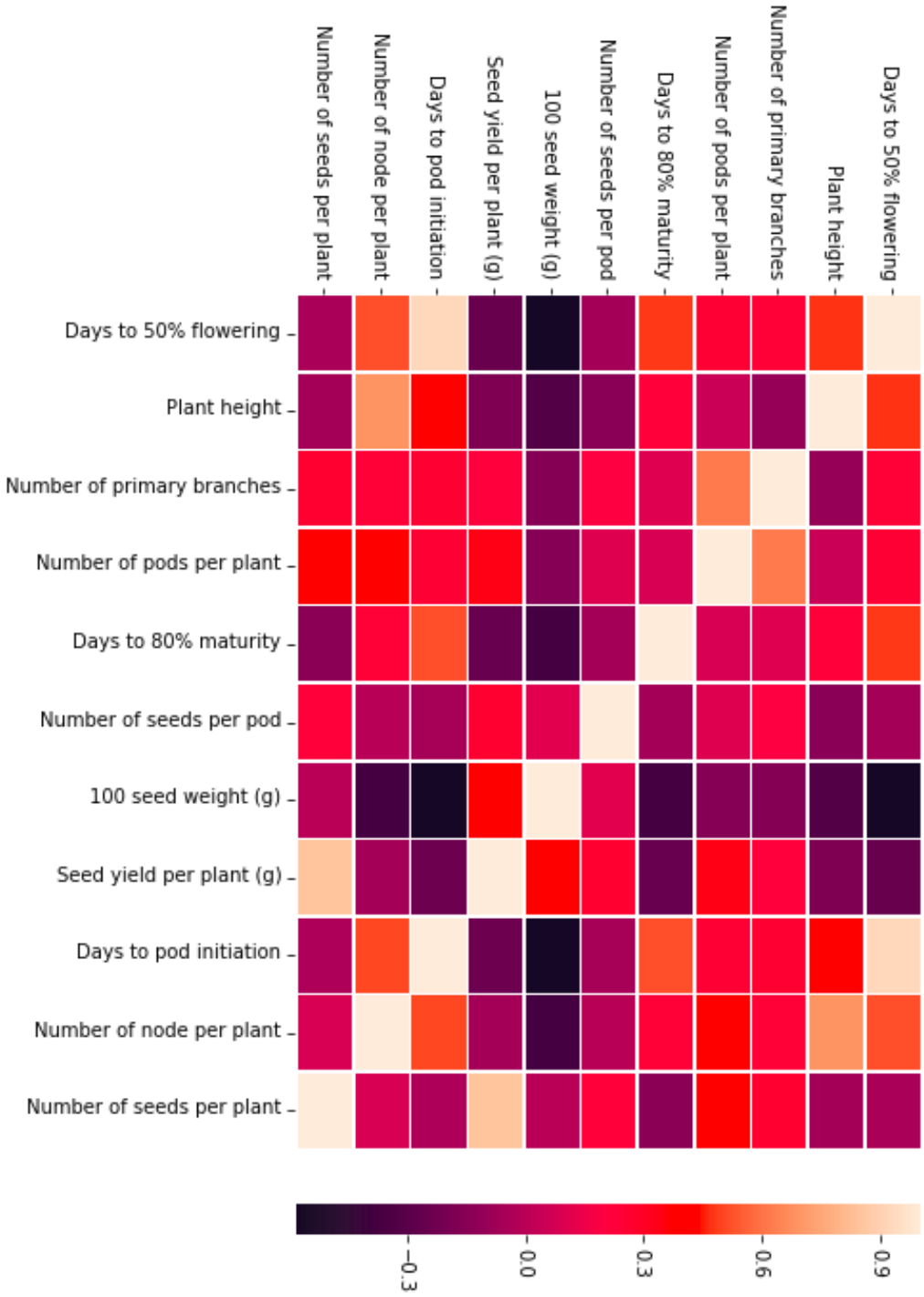

FIGURE 3.1: Visualisation of Correlation Matrix using Heat Map

3.3 Scatter Plots of Highly Correlated Features

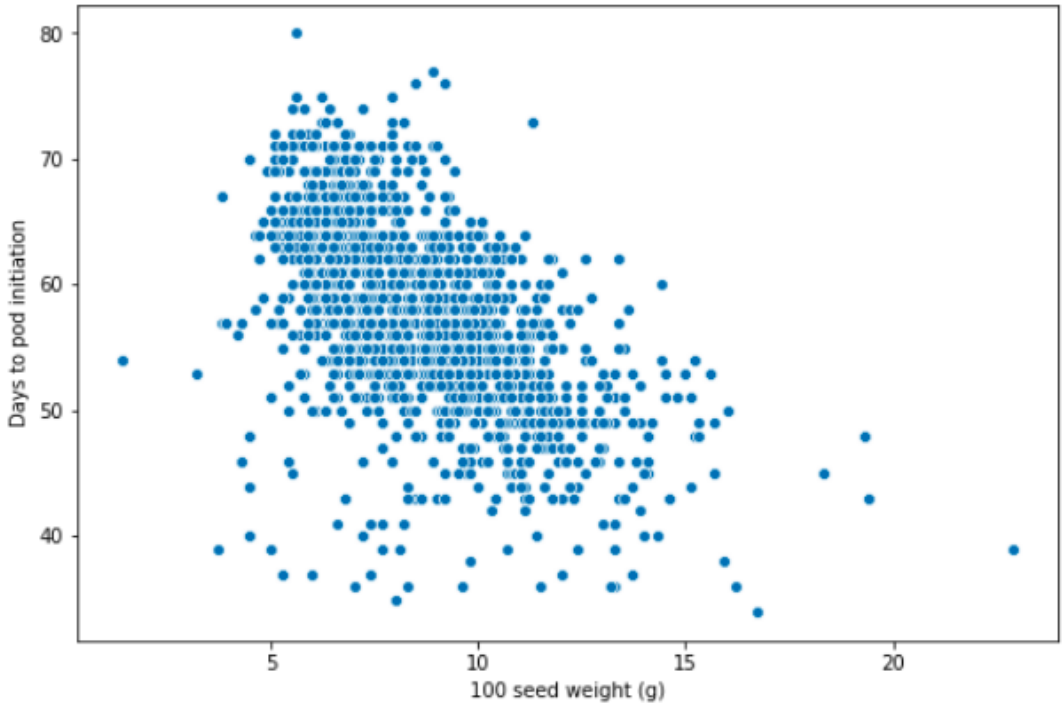

FIGURE 3.2: Days to Pod Initiation vs 100 seed weight (g)

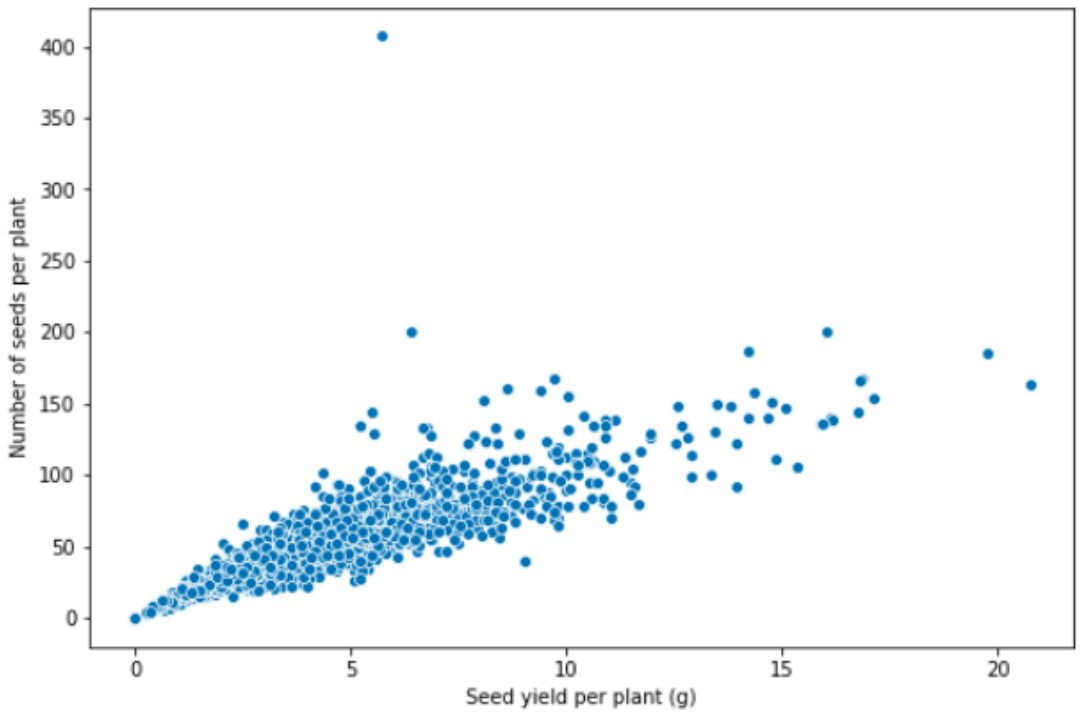

FIGURE 3.3: Number of seeds per plant vs Seed yield per plant (g)

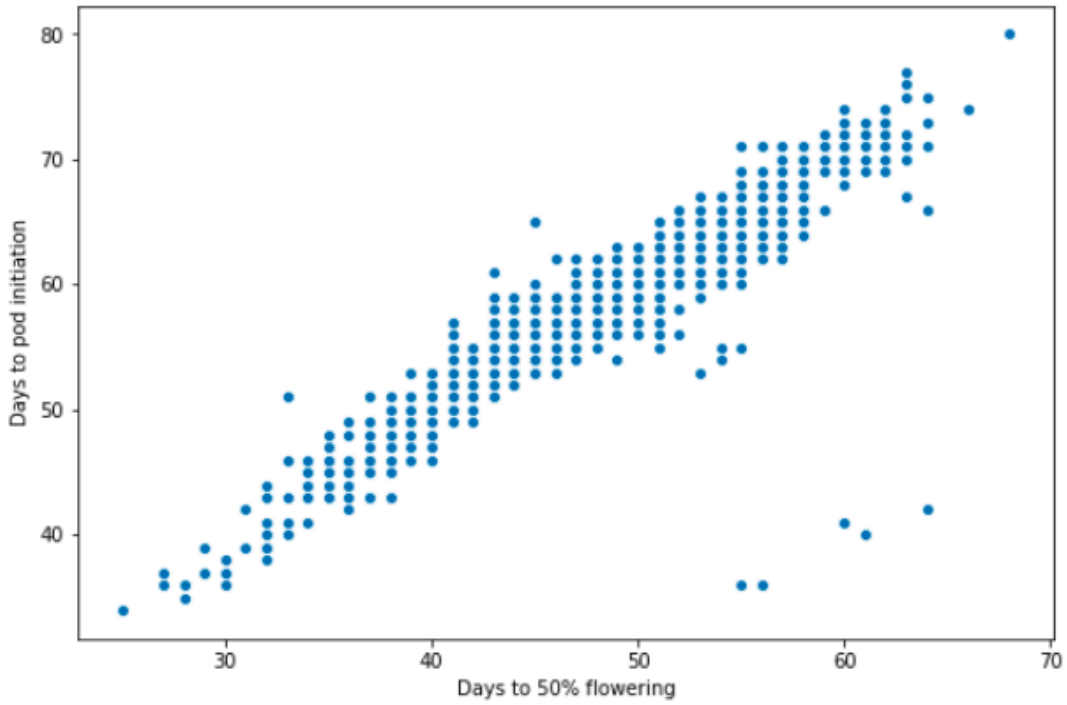

FIGURE 3.4: Days to pod initiation vs Days to 50% flowering

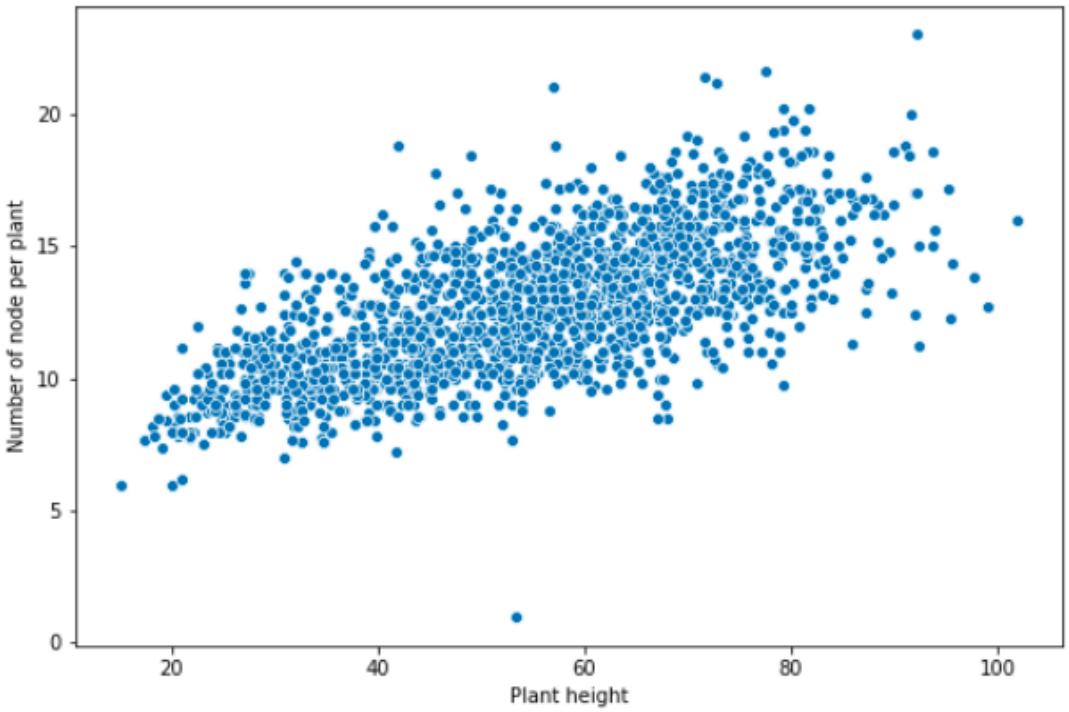

FIGURE 3.5: Number of node per plant vs Plant height

Chapter 4

Clustering Species Based on their Phenotypic Properties

4.1 Need for clustering

To get the best of all species with not so similar phenotypic factors we apply spectral clustering on the phenotypic traits dataset. We clustered the species together based on their phenotypic (observable physical) factors. So, when we take the best from each cluster, we have a collection of species with different physical properties. And, so we can have a best diverse set of species, for which, we will be requesting genetic data for further analysis!

4.2 Approach

We first drop the features of number of seeds per plant and seed yield per plant (g) and then cluster the species based on the other phenotypic factors. Then, we take the best species from each cluster.
The data had a mixture of numeric and categorical data and so applying simpler techniques like K-Means won't serve the purpose. So, we modified our categorical data using Hot Encoding and Label Encoding.

- Label Encoding: Category values like "bad", "good", "better", "best" can logically be assigned numeric values like 0, 1, 2, 3 respectively.
- Hot Encoding: Category values like "red", "blue", and "green" can be extracted as several other features like "color_red", "color_blue", "color_green". Then, normalization using standard scalar.

4.2.1 Calculating Optimal clusters for our use-case

We have used elbow method to calculate the optimal number of K.
For k = k_min to k = k_max, calculate sum of distortion in each cluster and plot it on a graph of K vs Distortion:

$$J(c, \mu) = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_j^{(i)} - \mu_{c^{(i)},j})$$

There will be a point where the marginal reduction in distortion will become very less on marginal increase in value of K. Here, we can conclude to be an optimal value of K.

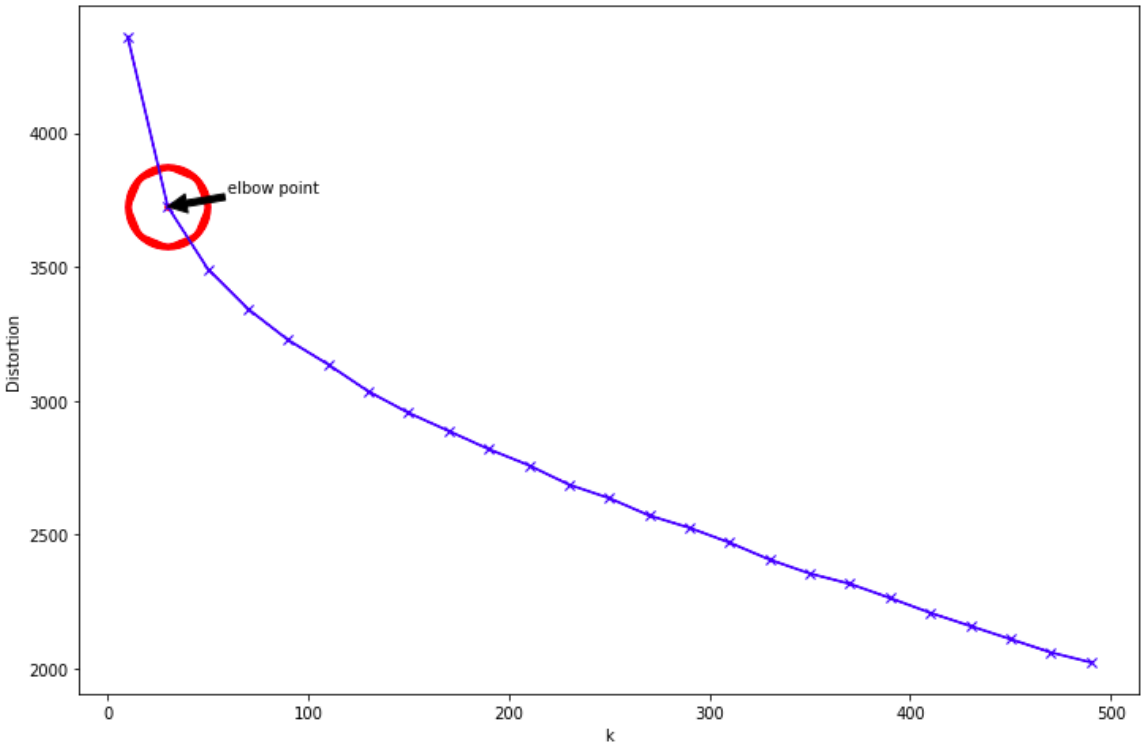

FIGURE 4.1: Determination of Optimal Clusters using Elbow Method

Value of K	Distortion Value	Value of K	Distortion value
10	4.359595	250	2.636826
30	3.724966	270	2.571894
50	3.491997	290	2.526217
70	3.342943	310	2.470833
90	3.227233	330	2.407244
110	3.135637	350	2.355372
130	3.035186	370	2.316924
150	2.955084	390	2.264052
170	2.887565	410	2.208787
190	2.819608	430	2.159514
210	2.759461	450	2.110721
230	2.687013	470	2.061156

TABLE 4.1: Variation of distortion with value of K

4.3 Results from Spectral Clustering

Cluster_Id	Cluster_Yield	No_Of_Species	Max_Yield	Best_Species	Avg_Yield
29	1781.88	374	20.74	CAT 2808	4.76438
16	99.94	21	16.82	CAT 530	4.7590
32	139.74	24	16.74	CAT 2875	5.8225
14	327.76	49	16.12	JSM 232	6.6889
5	370.8	61	16.02	RKS 54	6.0786
24	233.84	40	15.06	2006 M	5.846
27	93.08	18	13.96	JS 20-49	5.1711
33	147.84	31	11.74	CAT 818	4.7690
12	176.92	32	11.06	CAT 2383	5.52875
23	485.72	98	10.92	CAT 1733	4.9563
22	111.44	28	10.72	G 2130	3.98
3	466.08	94	10.58	CAT 1266	4.9582
6	185.96	44	10.56	JS 20-37	4.2263
28	96.94	23	10.48	JSM 302	4.2147
19	123.36	28	10.48	CAT 1705	4.4057
1	187.26	40	10.28	EC 33940	4.6815
21	178.92	42	10.26	8116-21 D	4.26
0	304.88	84	10.06	JS 93-37	3.6295
34	71.86	14	9.82	CAT 1502	5.1328
4	313.06	121	9.7	UPSM 1034	2.5872
20	68.74	16	9.56	EC 457464	4.296
2	202.12	44	9.42	JSM 152	4.5936
11	157.22	42	8.92	CAT 2162	3.7433
8	113.86	25	8.88	CAT 2117 B	4.5544
15	73.84	26	8.58	CAT 3379	2.84
13	92.44	25	8.14	JS 20-56	3.6976
9	136.64	33	7.9	CAT 1241	4.1406
18	144.26	56	7.62	CAT 2667	2.5760
25	58.782	13	7.54	CAT 164	4.5216
26	80.66	19	7.22	PCR 3229	4.2452
7	195.08	59	7	CAT 999	3.3064
10	95.6	39	6.18	UPSM 670	2.4512
30	32.66	10	5.78	CAT 1740	3.266
17	49.02	12	5.78	PS 1471	4.085
31	97.72	41	5.02	CAT 1410	2.383415

TABLE 4.2: Results of spectral clustering on phenotypic data

Part II

Sampling and clustering techniques applied on Plant Genome

Chapter 1

Introduction

1.1 About Genotypic Data

- Whole Genome Sequence (WGS) is the thread like chain of nucleotides; A (Adenine), T (Thymine), G (Guanine), C (Cytosine) that make up an organism.
- Each WGS sequence is contained in a FASTA format file.
- A sequence contains approximately 16 billion characters as each sequence has around 170 million records and length of each records is 90 chars.

1.2 Aims to achieve with the genotypic data

- The need to sample: We need to find similarities between sequences or make clusters of species. Now, as in previous work spectral clustering was found to give very good results for construction of phylogenetic trees. But, it is computationally impossible to apply spectral clustering on such a huge data. And, so there is a need to sample. With sampling we aim to reduce a sequence of 16 billion characters to a few millions.
- Moreover, we also then aim to cluster the sampled genomic data and then compare the clusters obtained from phenotypic data to that of the genotypic clusters. However, since at this time we don't have the corresponding genotypic data for the phenotypic species we clustered and vice versa, so we just tested our clustering algorithm with a few of other genotypic sequences and when we have the corresponding ones we'll also try to do the mappings.

Chapter 2

Literature Review

Continuing the previous work, in this chapter we now deal with the genotypic data only. There have been numerous methods and approaches to sample and cluster genomic sequence. We went through a lot of papers. And, when there was a point when we were stuck on to how to assign the probabilities for pivotal sampling, we luckily hit a goldmine with some really exciting works done in the field. We will briefly discuss about them in this section.

2.1 Pivotal Sampling

The pivotal method is based on splitting the vector of inclusion probabilities into two parts. Only two inclusion probabilities are modified, and the method consists of selecting two units that will be denoted by i and j.
If $\pi_i + \pi_j > 1$, then $A = (1 - \pi_j)/(2 - \pi_i - \pi_j)$,

$$\pi_k^{(1)} = \begin{cases} \pi_k & \text{if } k \in \mathrm{U} \setminus \{\mathrm{i}, \mathrm{j}\}, \\ 1 & \text{if k = i}, \\ \pi_i + \pi_j - 1 & \text{if } k = j \end{cases}$$

$$\pi_k^{(2)} = \begin{cases} \pi_k & \text{if } k \in \mathrm{U} \setminus \{\mathrm{i}, \mathrm{j}\}, \\ \pi_i + \pi_j - 1 & \text{if k = i}, \\ 1 & \text{if } k = j \end{cases}$$

On the other hand, if $\pi_i + \pi_j < 1$, then $A = \pi_i/(\pi_i + \pi_j)$,

$$\pi_k^{(1)} = \begin{cases} \pi_k & \text{if } k \in \mathrm{U} \setminus \{\mathrm{i}, \mathrm{j}\}, \\ \pi_i + \pi_j & \text{if k = i}, \\ 0 & \text{if } k = j \end{cases}$$

$$\pi_k^{(1)} = \begin{cases} \pi_k & \text{if } k \in \mathrm{U} \setminus \{\mathrm{i}, \mathrm{j}\}, \\ 0 & \text{if k = i,} \\ \pi_i + \pi_j & \text{if } k = j \end{cases}$$

In the first case, a one is allocated to only one inclusion probability. In the second case, a zero is allocated to only one inclusion probability. The problem is thus reduced to a population of size N - 1. In at most N steps, a solution is obtained. This method is interesting for its extreme simplicity, and it can be implemented by means of a strictly sequential procedure, i.e. by a single scan through a data file.[9]

2.2 Spectral Clustering of genomic sequences

Clustering and sampling are very important techniques of unsupervised learning , and these have been exhaustively researched. Thus, huge amount of information is available on these subjects. Hence, here we do not attempt to give a review of works done in these fields, rather we only present literature regarding usage of SC and sampling techniques in the field of plant genome.[10]

The problem of clustering data items into related groups based on similarity is an extremely common problem arising in a variety of disciplines and applications, and clustering algorithms for various applications have been studied for decades. Such algorithms depend upon the knowledge or acquisition of similarity information to relate data items to each other, e.g. the affinity between points in Euclidean space. Spectral techniques, which make use of information obtained from the eigenvectors and eigenvalues of a matrix, have attracted increasing research attention with respect to clustering in recent times. [11]

Spectral Clustering can be performed in two ways; recursive and non-recursive. Bouaziz [12] in 2012 used this method in a recursive way for genetic studies. However, we use a common non-recursive way [13], [6] since it is simpler and cheaper . It also gives tight and compact clusters.

Li [14] in 2010 used SC for clustering gene sequences (which are a subset of WGSs) where they construct ed the similarity matrix by Cosine Similarity. We use other basic techniques like Alignment Score, Jukes Cantor and Pairwise Distance as these capture the similarity between the genome sequences in a better way. Lawson [15] in 2012 used advance techniques of constructing the similarity matrix as mentioned above.

We have already discussed about papers that helped us in researching about applications of Spectral Clustering in Part 1. Here we have applied spectral clustering technique for clustering of genome sequence data. The points in the data used represent nucleotide, considered as genome sequences. Whole genome sequence is made up of billions of these genetic letters.

Chapter 3

Pivotal Sampling based on Local Pattern Histograms of Binary Images

3.1 Introduction to Pivotal Sampling

- Pivotal Sampling is a probability sampling technique to sample across the length of sequence.
- Each of the unit in population has a assigned probability of getting selected in the sample.
- In each iteration, either one unit will be sampled or will not be sampled depending upon the probability value assigned to it, and probability of that sampled/non-sampled will be updated.
- Thus, after each iteration, this procedure is repeated on the n 1 remaining units and with updated probabilities.

3.2 How do we decide the probabilities?

- We were stuck at this point and then got a breakthrough as we came across a paper[16] on Similarity Estimation based on Local Pattern Histograms of Binary Images.
- It is a new method of feature extraction of DNA sequences represented by binary images.

- The proposed method had linear time complexity for the length of DNA sequences, which is practical even when long sequences such as Whole Genome Sequences (WGS) are compared.

- The method generated binary images for a genome sequence. In the paper, they applied it on some mammalian species. So, we proceeded on to plan this for Soybean genome. But, we needed to decide probability of sampling with this information. And, this was again something to research on. We will elaborate more on this later.

3.3 Generating a binary image from a genome sequence

- Firstly, we assign 2D numerical vectors on xy-plane, which are perpendicular or in opposite directions to each other.

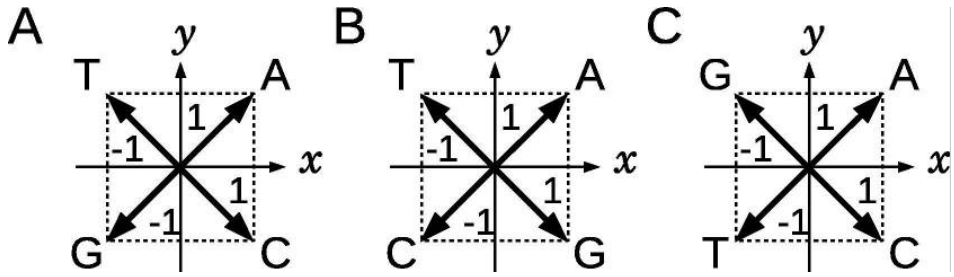

FIGURE 3.1: Three independent assignments of vectors on the xy-plane to individual nucleotides. Four nucleotides A, T, G, and C are arranged counterclockwise on the xy-plane "ATGC" (A), "ATCG" (B), and "AGTC" (C). Assignment A is used throughout our work.

- For example, generating a binary image of sequence "ACATATG"

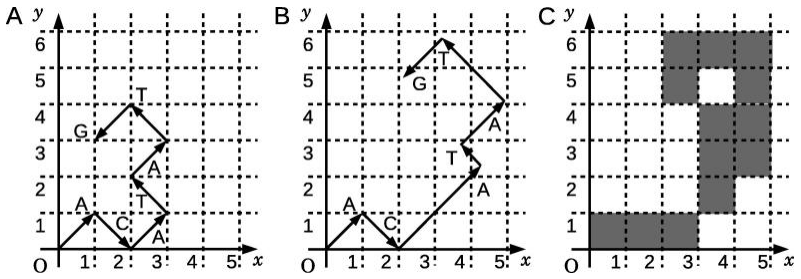

FIGURE 3.2: A. The primary graphical representation. B. The graphical representation modified with weighting factors. C. The generated binary image. Each grid represents an individual pixel of a binary image.

[1]Image Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4880953
[2]Image Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4880953

- To extract potential information conveyed by individual nucleotides, we use weighting factors, based on Markov Chain Model.
- A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.
- So, according to the second order Markov Chain, we define the probability that a nucleotide z occurs after a pair of nucleotides xy is calculated using

$$P(z|xy) = \frac{N_{xyz}}{\sum_{s\in[A,T,G,C]} N_{xys}}$$

where N_{xyz} and $N_{xys}(z, s \in [A, T, G, C])$ are the numbers of occurrence of triplets xyz and xys till the analyzed point.

- To emphasize rare patterns that appear in genome sequences, we used self-information I(E), the amount of information that is received when a certain event E occurs, as the weighting factor.
- Let P(E) be the probability that event E occurs, I(E) is defined as I(E) = $\log_2$P(E) in bit units. A trajectory for each genome sequence in a 2D plane is drawn as follows:

$$R_i = \sum_{k=1}^{i} w_k V_k$$

where R_i is the coordinate of the i^{th} point on the trajectory, V_k is the vector assigned to the k^{th} nucleotide of the genome sequence, and wk is the corresponding weighting factor I(E).

3.3.1 Graphical Representation of few Mammalian Mitochondrial Genomes without Weights

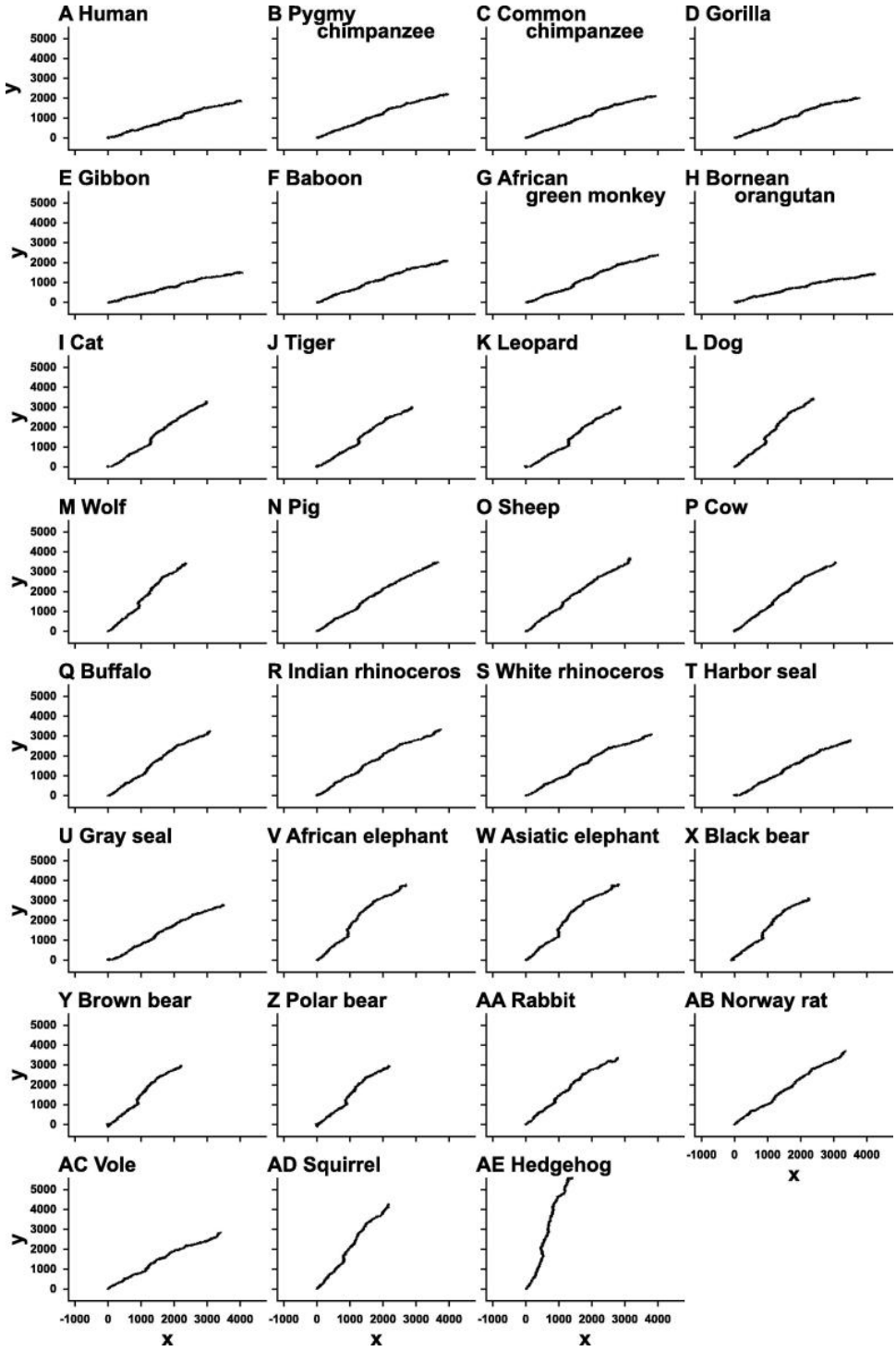

FIGURE 3.3: Graphical Representation of few Mammalian Mitochondrial Genomes without Weights

[3]Image Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4880953

3.3.2 Graphical Representation of few Mammalian Mitochondrial Genomes with Weights

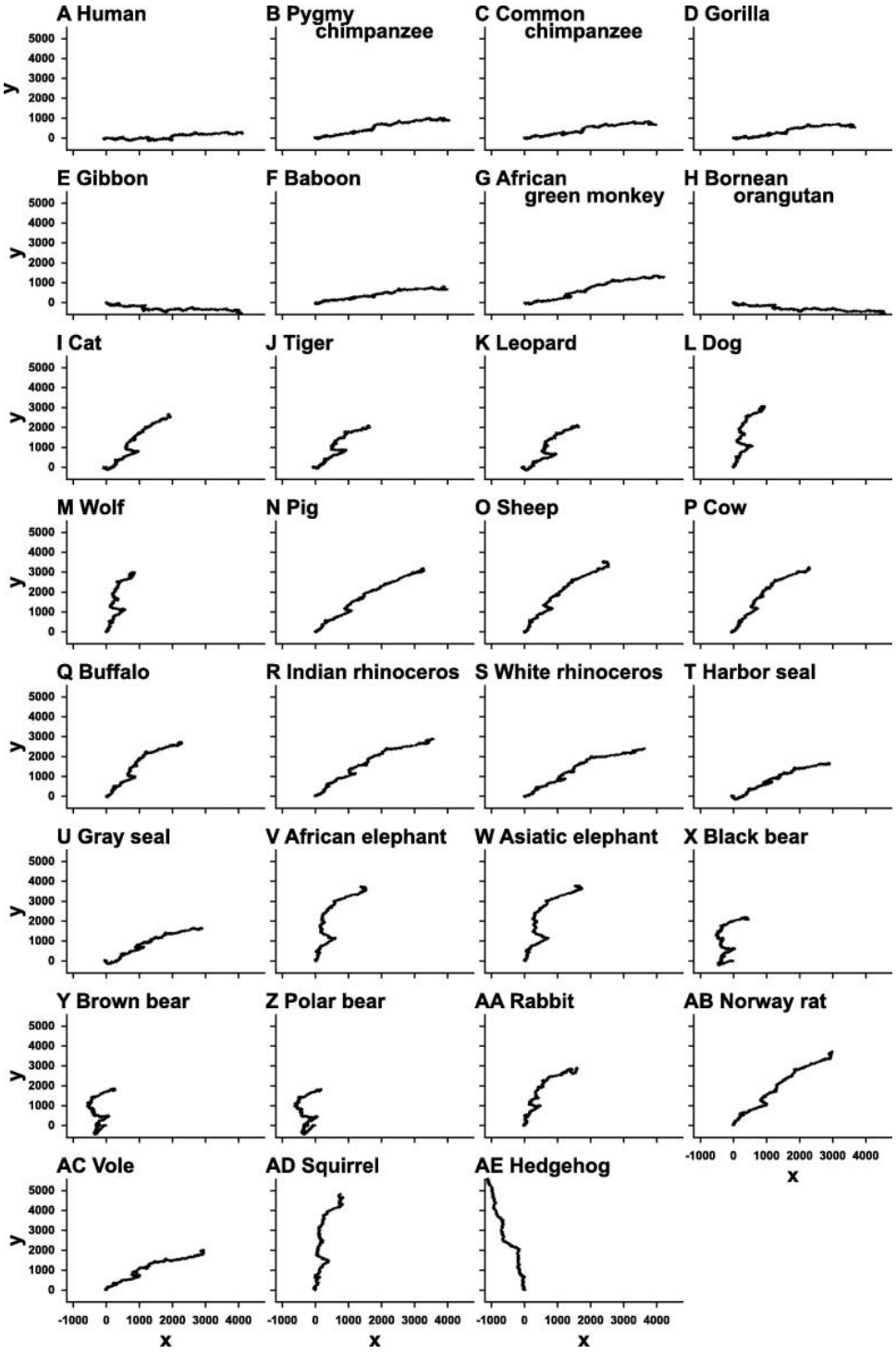

FIGURE 3.4: Graphical Representation of few Mammalian Mitochondrial Genomes with Weights

[4]Image Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4880953

3.4 Our Approach

We use the fact that the segments with higer weights contribute more in the shape of the curve than those with the lower weights. Also, adding the weighing factors really helped in highlighting the difference between closely related species as evident from the figures above.

So, in order to sample, we design our approach in a way that the segments which have more weight in the curve have higher probability of selection. This is explained in a logical manner below:

- First pass: we compute the total weight of all the nucleotides in a sequence.
- Then, in the second pass, compute the individual weight of the nucleotide and normalize by dividing with total weight. Let us call this as P_i.
- Note that sum of all P_i will be 1. We will multiply each P_i by the number of samples we need to select as in pivotal sampling. (In pivotal sampling, the sum of probabilities of all P_i should be equal to the number of samples we need to select)
- Then, we simply perform pivotal sampling (algorithm already discussed) over the data to get the sampled sequence.

3.5 Results

- We successfully sampled 20 sequences with 16 billion characters in each sequence to reduce it to 1 million characters each.
- That is, it was a reduction from 16 x 10^9 to 10^6 nucleotides per sequence. We can adjust the number of samples we need to select just by changing the multiplication factor as suggested in the approach above.
- Now, with the sampled data, spectral clustering which was practically impossible on the original data can be easily applied on the sampled data.

Chapter 4

Application of Spectral Clustering on the Sampled Genomic Data

4.1 Need for clustering

- In previous work, we have sampled the number of characters in the sequence, i.e., sampled a sequence of a few billions to millions.
- Now, in order to find species similar based on their genomic data, we move on to clustering for the rescue.
- With clustering, we can then compare clusters obtained from phenotypic data to those obtained from genotypic data.
- This way we can make useful correlations between genotypic properties and phenotypic properties and map specific physical properties with specific records in the genomic data.
- However, right now since we don't have corresponding genotypic data for the species of whose phenotypic we worked, all what we could do at the moment in this part was to prepare the clustering algo and test it on few samples, and then when in future we have the corresponding genomic data, apply the same on it.

4.2 Our Approach

- Challenges: The non-numerical nature of data, the huge size of every sequence and the non-availability of data of genotypic sequences for the same species as of phenotypic were the major challenges.
- We have 10^6 character in the each sampled sequence and we represent each character as vectors on xy-plane, which are perpendicular or in opposite directions to

each other. In our case, we assigned $(1,0), (-1,0), (0,-1)$ and $(0,1)$ to A, T, G and C respectively.

- Distance Matrix: We use Manhattan distance for calculating distance matrix. We take distance of one sequence to other and create $N * N$ matrix.

$$W_{ij} = \sum_{k=0}^{K} (|x_{ik} - x_{jk}| + |y_{ik} - y_{jk}|)$$

where x_{ik}, y_{ik} is a k^{th} point in space corresponding to the i^{th} sequence.

$$W = \begin{bmatrix} W_{11} & W_{12} & ..... & W_{1n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ W_{n1} & W_{n2} & ..... & W_{nn} \end{bmatrix}$$

$$D = \begin{bmatrix} D_{11} & 0 & 0 & 0 \\ 0 & D_{22} & 0 & 0 \\ 0 & 0 & .... & 0 \\ 0 & 0 & 0 & D_{nn} \end{bmatrix}$$

$$L = D - W$$

where D is Diagonal Degree Matrix i.e all the diagonal elements represents degree of the $i(th)$ node in the graph, W is the weight matrix i.e weight between i^{th} and j^{th} node, L is the Laplacian Matrix.

All we have to do now is to calculate the eigen vectors u_j of L. Now, we can simply proceed to run the final step of spectral clustering. We run K-means on the first k eigenvectors.

4.3 Results

Since, we were not having the corresponding genotypic data for the species of whose phenotypic data we worked in the first part, we couldn't really proceed with comparing the phenotypic and genotypic data. We just applied clustering on these 20 sampled sequences just to be sure about the convergence of the algorithm in lesser time. So,

Species Name	Cluster Id	Species Name	Cluster Id
1-SRA-FASTA	0	11-SRA-FASTA	2
2-SRA-FASTA	2	12-SRA-FASTA	3
3-SRA-FASTA	2	13-SRA-FASTA	0
4-SRA-FASTA	0	14-SRA-FASTA	2
5-SRA-FASTA	1	15-SRA-FASTA	1
6-SRA-FASTA	1	16-SRA-FASTA	2
7-SRA-FASTA	1	17-SRA-FASTA	0
8-SRA-FASTA	4	18-SRA-FASTA	0
9-SRA-FASTA	2	19-SRA-FASTA	2
10-SRA-FASTA	3	20-SRA-FASTA	2
1-SRA-FASTA	0	11-SRA-FASTA	2
2-SRA-FASTA	2	12-SRA-FASTA	3

this way, for number of iterations = 300, k = 4, n = 20, d (number of features) = 10^6 we were able to do it in around 2 hours. This time can be further adjusted by changing the number of iterations. Once, we have the corresponding genotypic data for the phenotypic species we worked on, we can apply this technique on them and then move on to compare the results obtained from genotypic and phenotypic data. Thus, we may then proceed to map the phenotypic traits with specific genotypic patterns.

Part III

Conclusions and Future Work

Conclusions and Future Work

The objectives of this project were to

- Find correlations between several phenotypic factors for soybean.
- Clustering similar species based on their phenotypic factors and find the best in each cluster based on yield
- Pivotal sampling of the genomic sequence so that clustering of species based on their genomic sequences becomes less expensive
- Application of Spectral clustering on the sampled genomic data

In this project, we have achieved great feat in achieving all of the above four objectives and in the way. Not only did we learn a lot of new things in the process, but we also have established one really good, logical and intuitive method to sample the genomic data. We started with establishing correlations between several phenotypic factors. We made use of the pearson correlation coefficient and made scatter plots and heat maps for better visualization of the results.

We then proceeded on to do the clustering for the phenotypic data and the results achieved were really impressive. We also did a weighted sorting in the clusters to find out the best species in each cluster and as such our work could be of great use to people who may wish to find the best species having a set of physical properties.

We then had a meeting with Dr. Ratnaparakhe from the India Institute of Soybean Research, Indore and he was really impressed with the results. He told that the work will be indeed of great help to them to understand the traits better and also help the concerned to research more about species that perform better in particular areas.

To sample the data, we were at a dead end for a while. But, then we came across a published work, where a great work was completed on to perform similarity estimation between DNA sequences based on local pattern histograms. From that work, we learned that we can also quantitatively extract features out of DNA sequences based on a probabilistic model. We took this good from this paper to break the rock and then used

it to compute the selection probability in the pivotal sampling algorithm. The whole approach is quite intuitive and logical. Imagine a curve made of a few segments and you remove a few smaller segments. The overall figure the curve is drawing won't lose significantly because we removed just a few smaller segments and the larger segments which contribute to the majority of the curve are there as it is. So, we actually sampled a sequences of 16 billion characters to a few millions to make clustering computationally practical.

Then, we applied spectral clustering on the samples to get clusters. However, since we were not having the corresponding genotypic data for the phenotypic we clustered before, we couldn't map the genotypic clusters to the phenotypic clusters and thus couldn't proceed on to map specific phenotypic properties with the genotypic sequences.

So, in future, once we have the corresponding genotypic data for the phenotypic species we clustered or vice versa, then we will try to draw those similarities.

We can also proceed to write a paper titled, "A Quantitative Approach to Pivotal Sampling of Genomic Data" as the kind of probability assignment we used is unprecedent and is complete logical. We didn't find a paper that uses this way to sample and so what we achieved is definitely something new.

Bibliography

[1] A. Sayago A. G. Asuero and A. G. Gonzalez. "The Correlation Coefficient: An Overview". In: *Critical Reviews in Analytical Chemistry* (2006).

[2] F. Galton. "Co-relations and their measurement". In: *Proceedings of the Royal Society of London* (1888).

[3] W. A. Nicewander J. L. Rodgers. "Thirteen ways to look at the correlation coefficient". In: *The American Statistician* (1998).

[4] R. A. Nadkarni. "The quest for quality in the laboratory". In: *Analytical Chemistry* (1991).

[5] T. P. E. Auf der Heyde. "A tutorial on factor and cluster analysis." In: *Journal of Chemical Education* (1990).

[6] U. V. Luxburg. "A Tutorial on Spectral Clustering". In: *Max Planck Institute for Biological Cybernetics* (2007).

[7] K. Meerbergen. "The Lanczos method with semi-inner product". In: *BIT December* (2001).

[8] Ronald R. Coifman Ioannis G. Kevrekidis Boaz Nadler Stephane Lafon. "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems". In: *Journal of applied and computational harmonic analysis* (2001).

[9] Claude J. Deville and Tille Y. "Unequal Probability Sampling Without Replacement Through a Splitting Method". In: *Oxford University Press on behalf of Biometrika Trust* ().

[10] Milind B. Ratnaparkhe Aditya Shah Aishwary G. Anant Lal Aditya A. Shastri Kapil Ahuja. "Vector Quantized Spectral Clustering applied to Soybean Whole Genome Sequences". In: *Preprint submitted to Evolutionary Bioinformatics* ().

[11] Pentney W. Deville and Meila M. "Spectral Clustering of Biological Sequence Data." In: *Association for the Advancement of Artificial Intelligence* ().

[12] Guedj M Ambroise C. Bouaziz M Paccard C. "SHIPS:spectral hierarchical clustering for the inference of population structure in genetic studies." In: *Plosone. 2012; 7(10):e45685.* ().

[13] Weiss Y. Ng AY Jordan MI. "On spectral clustering:analysis and an algorithm." In: *NIPS. 2001;14(2):849-856.* ().

[14] Ching WK Mamitsuka H. Li L Shiga M. "Annotating gene functions with integrative spectral clustering on microarray expressions and sequences." In: *Genome informatics.2010; 22:95-120.* ().

[15] Falush D. Lawson DJ. "Similarity matrices and clustering algorithms for population identification using genetic data". In: *Annual review of human genomics.2012; 13:337-361* ().

[16] Y. Kobori and S. Mizuta. "Similarity estimation between DNA sequences based on local pattern histogram of Binary Images". In: *Genomics Proteomics Bioinformatics* ().