Biologia de Sistemas

Ronaldo Fumio Hashimoto

ronaldo@ime.usp.br

Questions

The questions are the red marked text in Slides 24, 28, 32, 36, 40, 44.

Introdução

- Algumas pessoas por iniciativa própria estão coletando dados do novo coronavírus e fazendo suas análises. Como, por exemplo, dados de sequências genômicas que estão no repositório *SARS-CoV-2 genome sequence data analyses* (link) e também neste repositório *Covid-19: análise de genomas* (link).

Fasta Dataset

- Sequências genômicas do COVID-19 (em um total de 68) do primeiro repositório foram coletadas em um arquivo fasta pelo autor do segundo repositório e podem ser obtidas neste link.
- The fasta file contains genomic sequences that look like

```
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
CGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
TGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
```

- The next slide presents a way to transform genome sequences into feature vectors.

Análise da composição nucleotídica

$$P_{xy} = \frac{f_{xy}}{f_x f_y}$$

Where:

- f_x is the frequence of nucleotide x,
- f_y is the frequence of nucleotide y,
- f_{xy} is the frequence of dinucleotide xy.

So, we will have measurements for 16 dinucleotides!!!

And, if we aggregate the frequence of mononucleotides T, C, and G.

So, we will have 19 measurements for each sequence!!!

Nucleotide composition analysis

- In fact, this method was used in the paper (link) presented in the next slide.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition

Received: 11 January 2015
Accepted: 26 October 2015
Published: 26 November 2015

Qin Tang[1,2], Yulong Song[1,2], Mijuan Shi[1], Yingyin Cheng[1], Wanting Zhang[1] & Xiao-Qin Xia[1]

Many coronaviruses are capable of interspecies transmission. Some of them have caused worldwide panic as emerging human pathogens in recent years, e.g., severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV). In order to assess their threat to humans, we explored to infer the potential hosts of coronaviruses using a dual-model approach based on nineteen parameters computed from spike genes of coronaviruses. Both the support vector machine (SVM) model and the Mahalanobis distance (MD) discriminant model achieved high accuracies in leave-one-out cross-validation of training data consisting of 730 representative coronaviruses (99.86% and 98.08% respectively). Predictions on 47 additional coronaviruses precisely conformed to conclusions or speculations by other researchers. Our approach is implemented as a web server that can be accessed at http://bioinfo.ihb.ac.cn/seq2hosts.

Nucleotide composition analysis

On Page 7, we can find the description of the nucleotide composition analysis:

Nucleotide composition analysis. The mononucleotide frequencies and dinucleotide biases of the spike sequences were computed using our original Python scripts. Dinucleotide bias is the ratio of the observed value to the expected frequency of each of the 16 dinucleotides: $\rho_{XY} = f_{XY}/f_X f_Y$, where ρ_{XY} is the dinucleotide bias, f_{XY} is the frequency of dinucleotide XY, f_X and f_Y are the frequencies of nucleotide X and nucleotide Y[38], respectively.

In this study, we considered 19 factors, including three mononucleotide frequencies (G, C and T) and 16 dinucleotide biases. As none of the frequencies has a normal distribution, the nonparametric

Applying this method to the dataset, we obtain the following feature vectors presented in the next slide.

Genome Sequences to Feature Vectors

	G	C	T	...	TG	TC	TT
Sequences				...			
0	0.196178	0.183826	0.321194	...	1.374268	0.801048	0.796865
1	0.196274	0.183901	0.321082	...	1.376058	0.800307	0.796404
2	0.196141	0.183700	0.321216	...	1.375361	0.801415	0.796713
3	0.196144	0.183680	0.321014	...	1.375052	0.801901	0.797047
4	0.196280	0.183701	0.321133	...	1.374184	0.801912	0.796767
...	...	...	...	...	...	...	...
63	0.196231	0.183747	0.321106	...	1.375771	0.802128	0.796926
64	0.196231	0.183714	0.321139	...	1.375628	0.802190	0.797085
65	0.196231	0.183714	0.321139	...	1.375097	0.801623	0.796760
66	0.195813	0.184559	0.320583	...	1.376687	0.815239	0.789715
67	0.202000	0.187034	0.317697	...	1.358975	0.825624	0.800239

[68 rows x 19 columns]

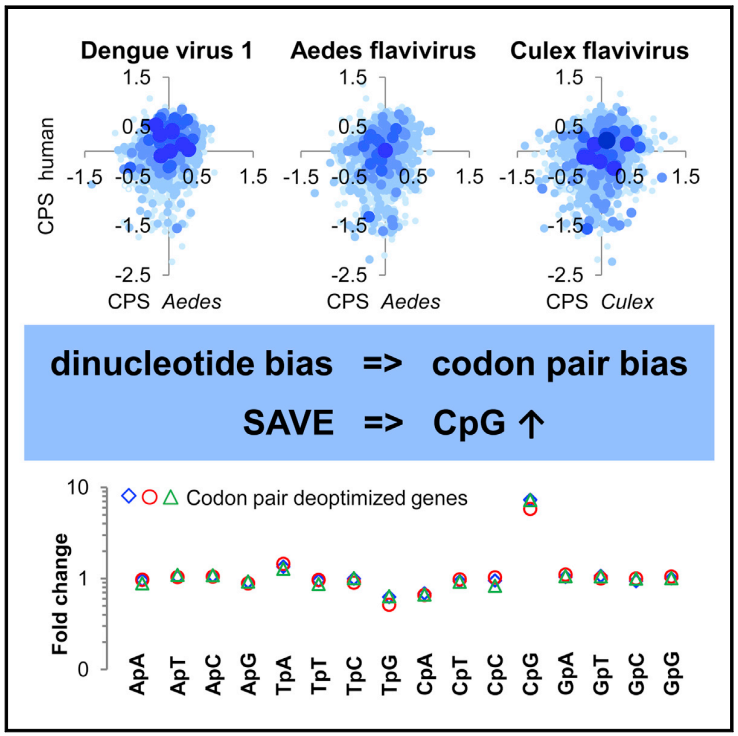
Nucleotide composition analysis

A molecular biological foundation that could explain why we can use dinucleotide biases can be found in this paper (link):

Cell Reports

Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias

Graphical Abstract



Authors

Dusan Kunec, Nikolaus Osterrieder

Correspondence

dusan.kunec@fu-berlin.de (D.K.),
no.34@fu-berlin.de (N.O.)

In Brief

Kunec and Osterrieder demonstrate that the encoding of viral proteins is not influenced by codon pair preferences of their host but that it can be influenced by host dinucleotide bias. Codon pair bias is primarily a consequence of dinucleotide bias. Attenuation by codon pair deoptimization works through an increase in CpG dinucleotides in recoded genes.

Data Standardization

- we will use a feature scaling method called standardization, which gives our data the property of a standard normal distribution.
- Standardization shifts the mean of each feature (column) so that it is centered at zero and each feature has a standard deviation of 1.
- For instance, to standardize the j-th feature (the j-th column), we can simply subtract the sample mean μ_j from every genome sequence sample and divide it by its standard deviation σ_j:

$$x_j' \leftarrow \frac{x_j - \mu_j}{\sigma_j}$$

- Applying the standardization to each column, we obtain the following feature vectors presented in the next slide.

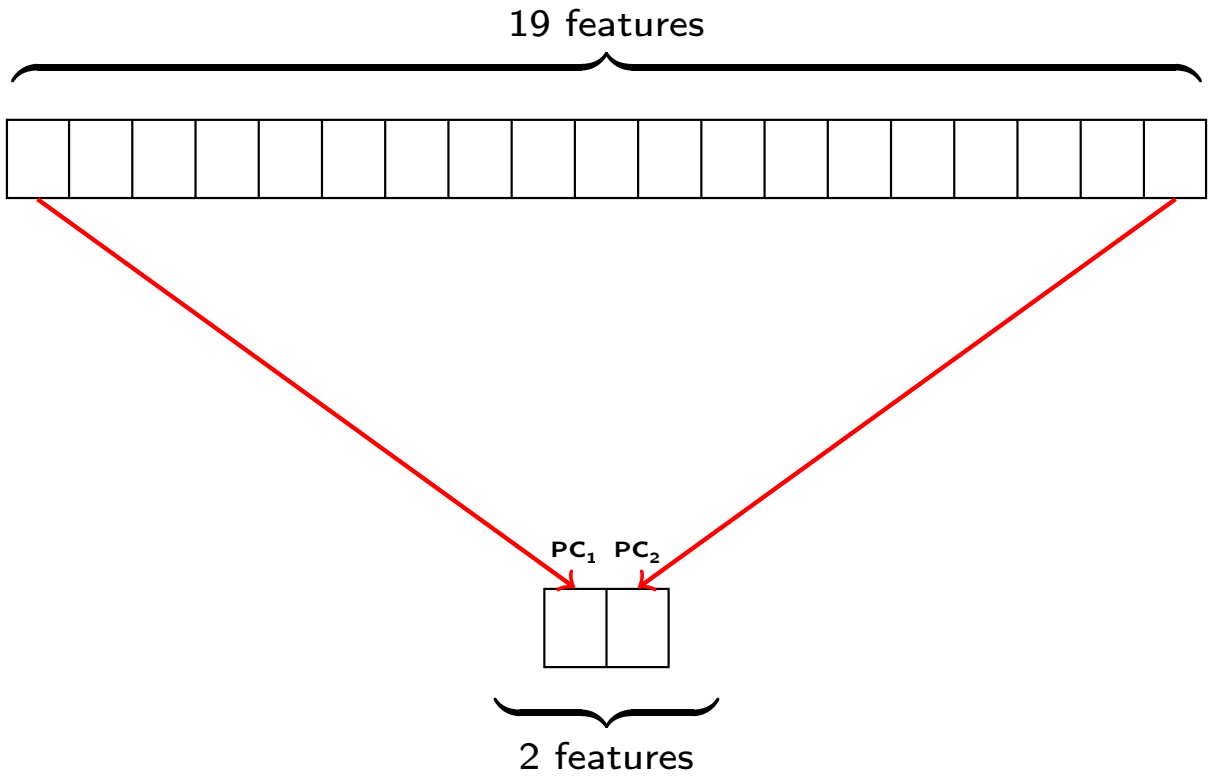
Data Standardization

Sequences	G	C	T	...	TG	TC	TT
				...			
0	0.110955	0.123225	0.138062	...	-0.136203	-0.147440	-0.119659
1	0.122963	0.133381	0.129856	...	-0.115236	-0.161420	-0.128469
2	0.106397	0.105939	0.139651	...	-0.123396	-0.140515	-0.122565
3	0.106796	0.103231	0.124896	...	-0.127016	-0.131336	-0.116185
4	0.123622	0.106033	0.133609	...	-0.137185	-0.131115	-0.121534
...	...	...	...	...	...	...	...
63	0.117619	0.112387	0.131586	...	-0.118594	-0.127054	-0.118493
64	0.117619	0.107810	0.134037	...	-0.120274	-0.125874	-0.115468
65	0.117619	0.107810	0.134037	...	-0.126496	-0.136584	-0.121665
66	0.065824	0.223365	0.093287	...	-0.107866	0.120489	-0.256189
67	0.831982	0.561952	-0.118050	...	-0.315355	0.316534	-0.055242

[68 rows x 19 columns]

Dimensionality Reduction - PCA

Principal Component Analysis is a technique for data dimensionality reduction



Principal Component Analysis

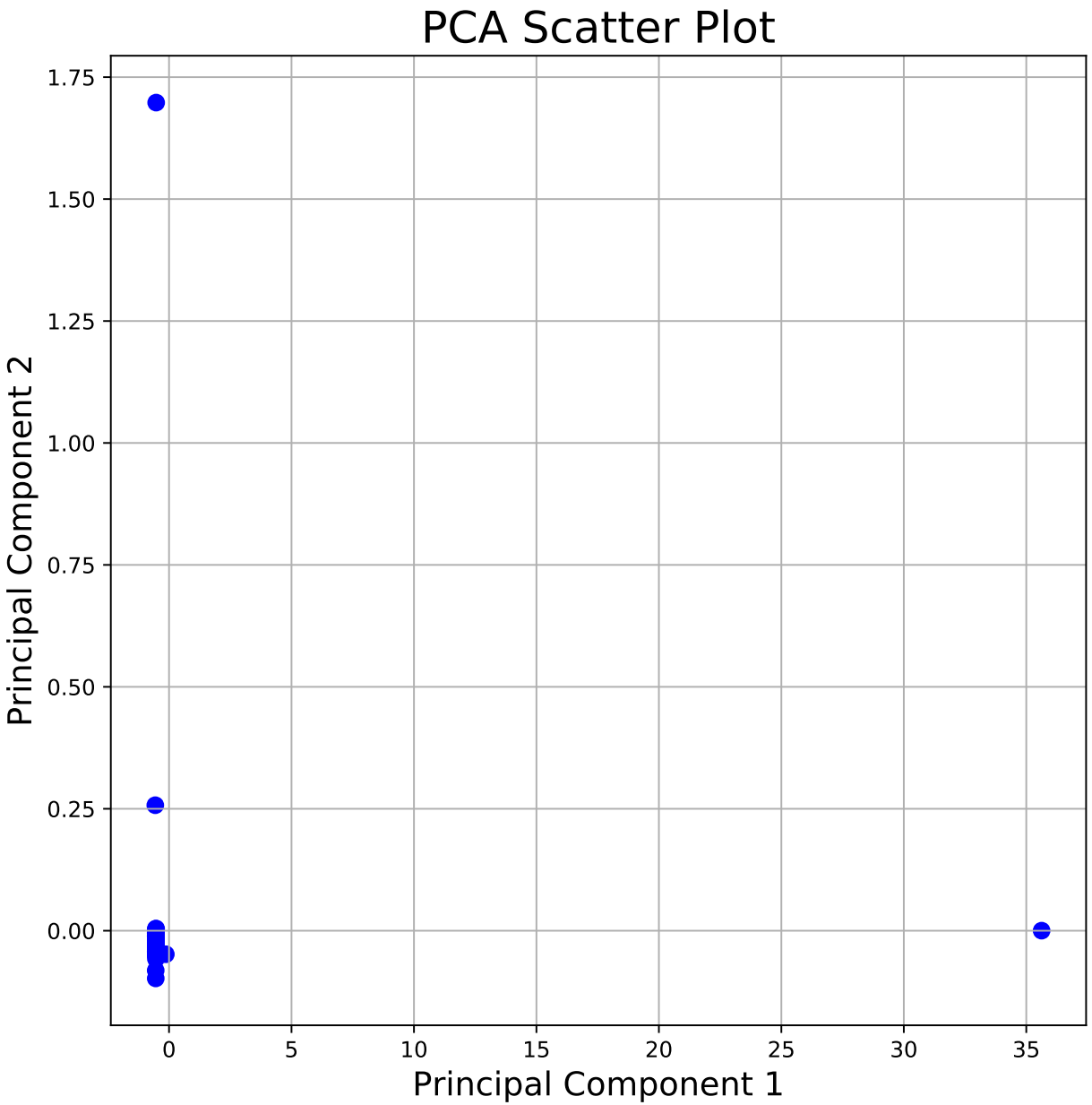
Applying the Principal Component Analysis on the standardized data, we obtain the following plotting using the first two principal components:

Sequences	PC 1	PC 2
0	-0.532224	-0.007329
1	-0.545619	-0.042707
2	-0.537864	-0.029685
3	-0.529753	-0.038280
4	-0.537378	-0.024534
...	...	...
63	-0.540701	-0.042126
64	-0.539434	-0.045830
65	-0.535194	-0.026688
66	-0.561362	0.257311
67	-0.525135	1.697806

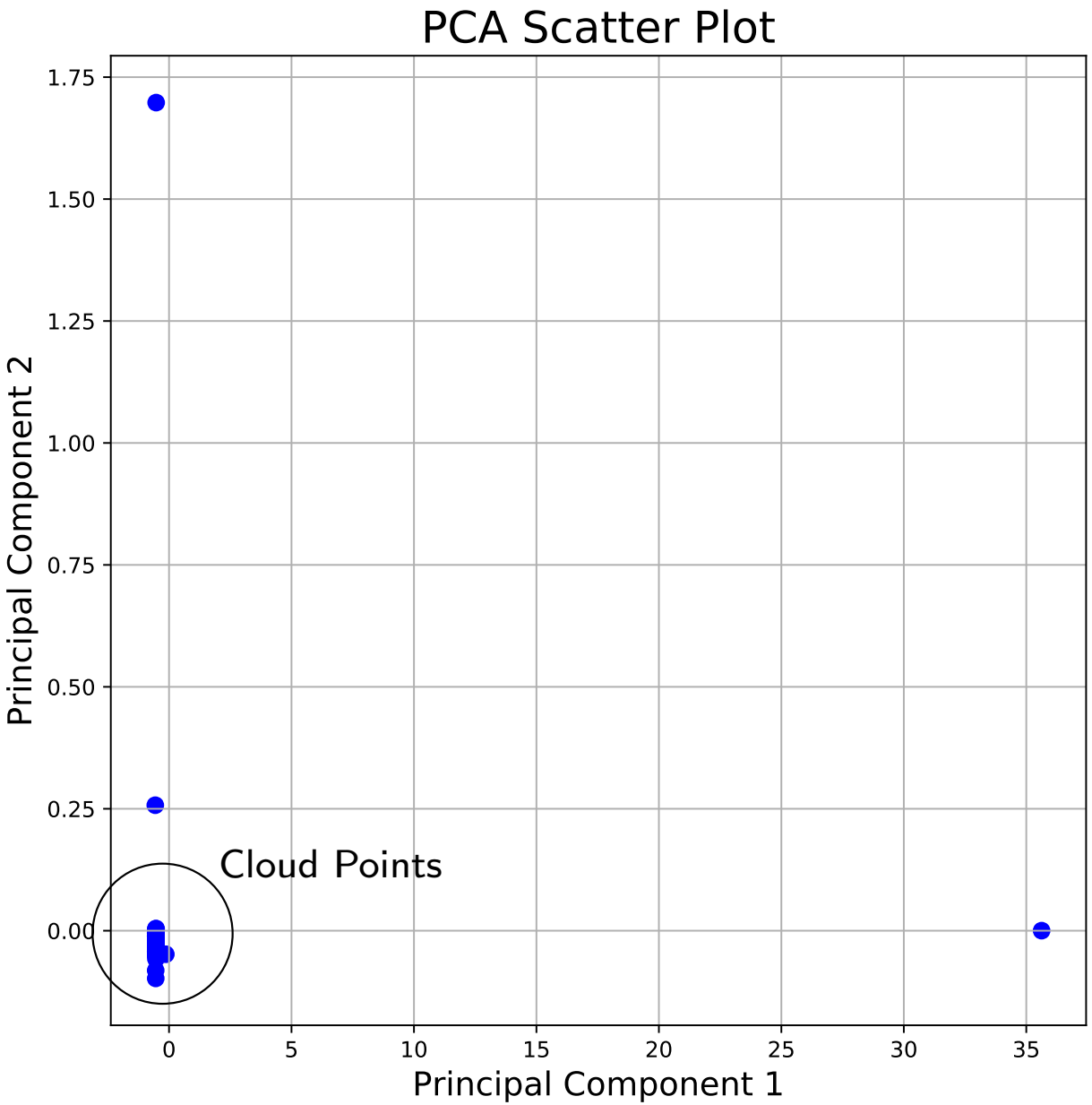
[68 rows x 2 columns]

The scatter plot of these points is presented in the next slide.

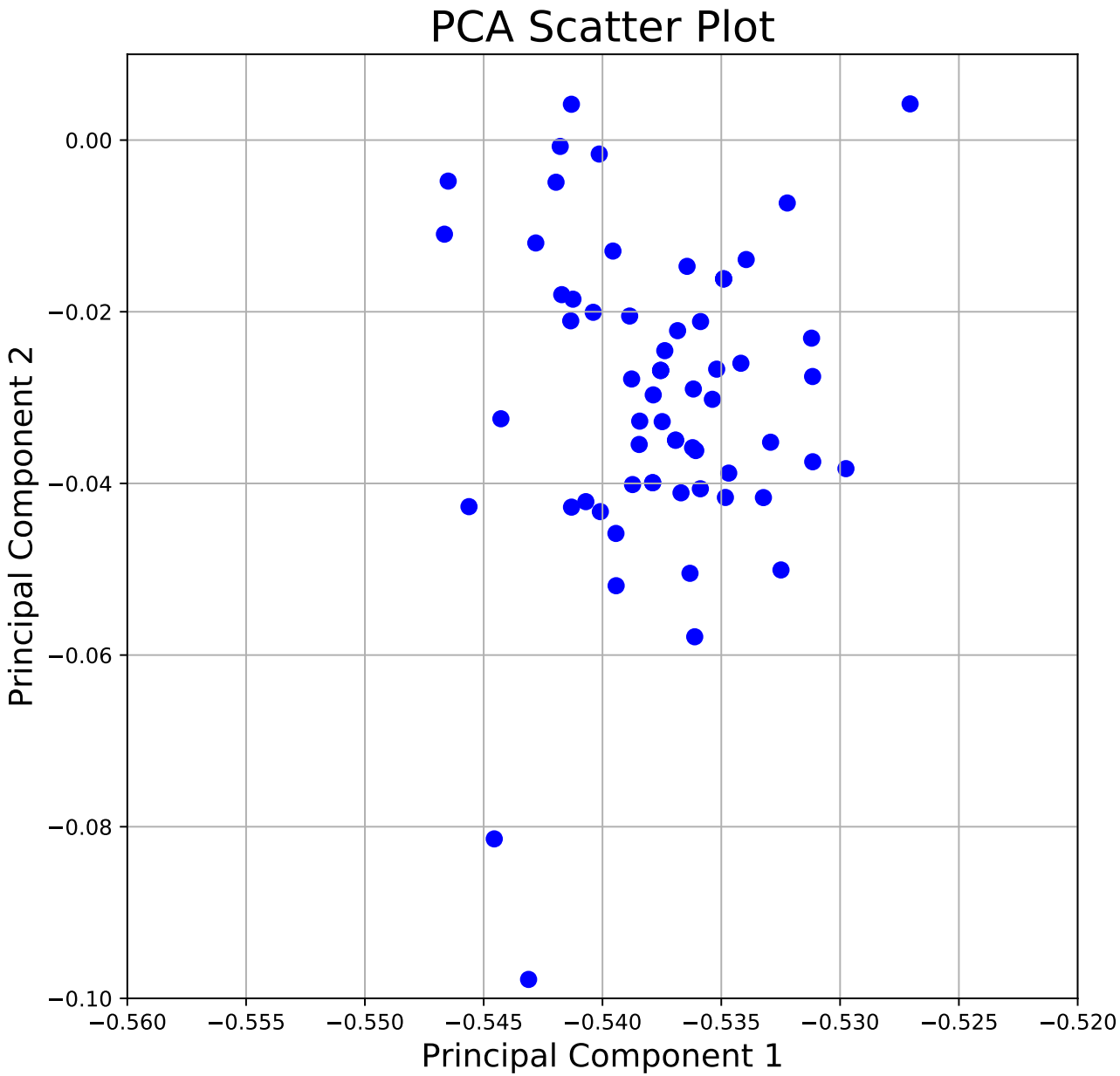
PCA - Scatter Plot



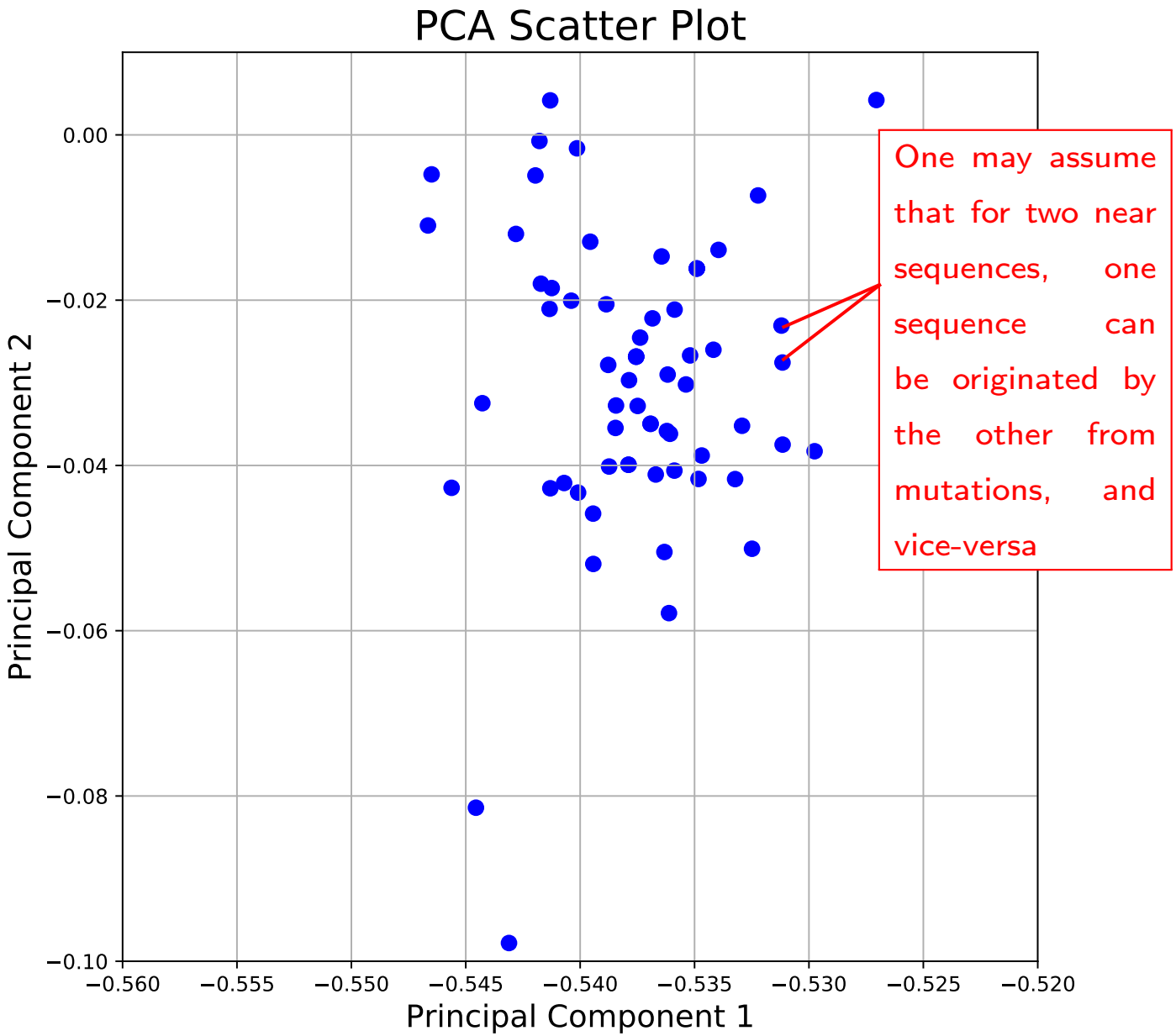
PCA - Scatter Plot



PCA of the cloud points



PCA of the cloud points



Distance Matrix

- We can compute the distance matrix of each sequence (point) pair using all the 19 features.
- So, we are not going to use the PCA components. These are just for visualization purpose.
- Given two sequences i and j, with 19 features $(s_1^{(i)}, s_2^{(i)}, \ldots, s_{19}^{(i)})$ and $(s_1^{(j)}, s_2^{(j)}, \ldots, s_{19}^{(j)})$, respectively, we can compute the Euclidian distance between them in the following way:

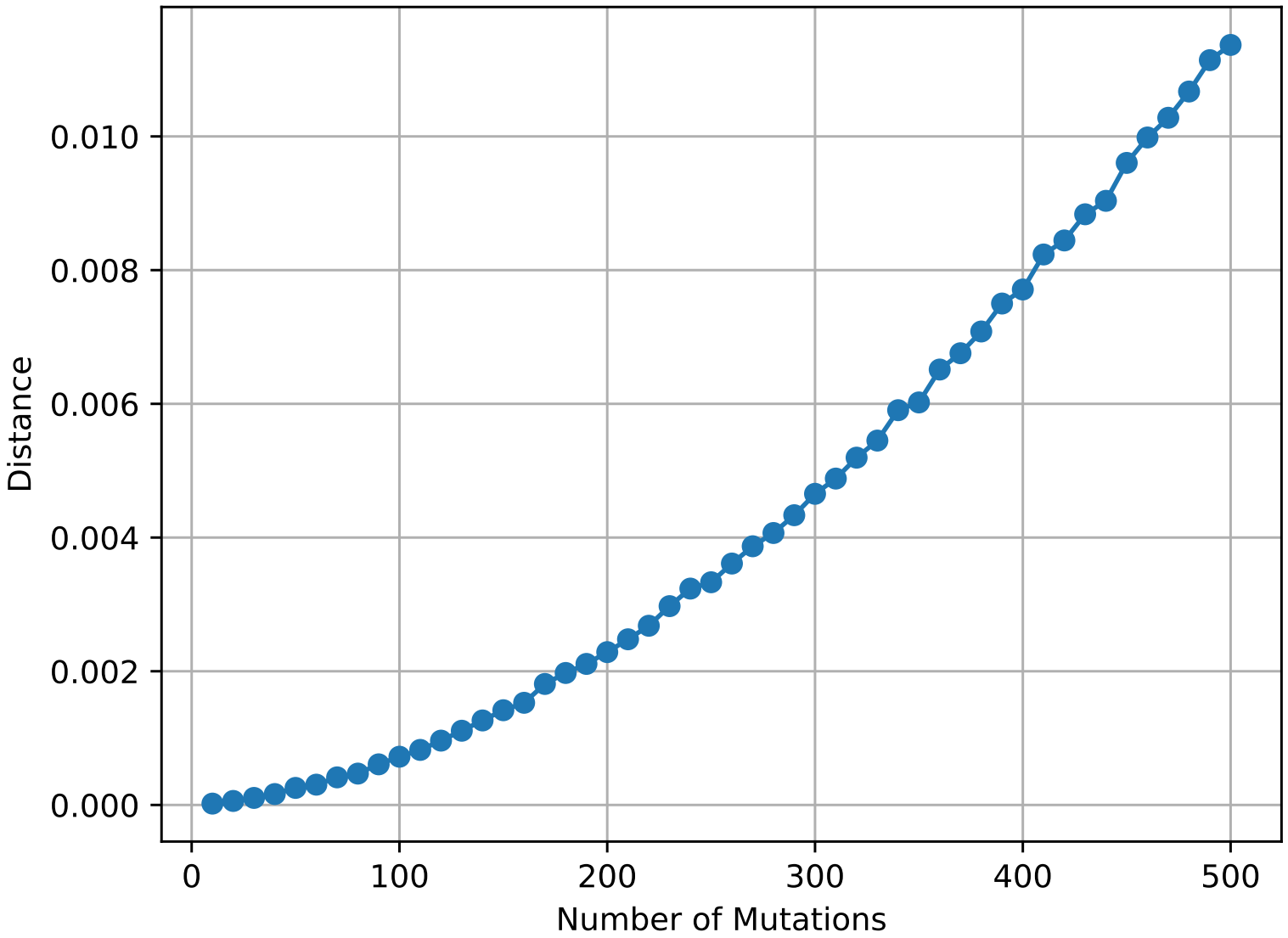
$$Dist[i,j] = \sqrt{\sum_{k=1}^{19} \left(s_k^{(i)} - s_k^{(j)} \right)^2}$$

- The distance matrix is presented at Slide 22.

Distance x Mutation

- One may ask what is the relation between mutations and distances.
- We can access how much distances are related from mutations by simulation.
- For example, given a genomic sequence S, we can perform m random mutations, obtaining a new sequence S' so that we compute the Euclidian distance between them using their 19 features. Do this a certain number of times (say, for example, 1000 times) and take the average.
- Varying m from 10 to 500, we obtained the following plot presented in the next slide. As we can see, there may be a relation between mutations and distances.

Distance x Mutation



Distance Matrix

Sequences	0	1	2	...	65	66	67
0	0.000000	0.064360	0.050939	...	0.043107	0.743996	1.707907
1	0.064360	0.000000	0.065604	...	0.055902	0.753940	1.743686
2	0.050939	0.065604	0.000000	...	0.037401	0.737778	1.730918
3	0.080873	0.073888	0.070527	...	0.063006	0.730722	1.740518
4	0.055615	0.064383	0.049107	...	0.029120	0.724162	1.726228
...	...	...	...	...	...	...	...
63	0.075372	0.062219	0.061336	...	0.041770	0.735061	1.743579
64	0.080543	0.068652	0.071840	...	0.049038	0.743404	1.746996
65	0.043107	0.055902	0.037401	...	0.000000	0.736055	1.727458
66	0.743996	0.753940	0.737778	...	0.736055	0.000000	1.636569
67	1.707907	1.743686	1.730918	...	1.727458	1.636569	0.000000

[68 rows x 68 columns]

Distance Matrix to Adjacency Matrix

- From the distance matrix, we can obtain an adjacency matrix by setting "near" points (sequences) i and j to 1, and 0, otherwise.
- This can be achieded by setting less than or equal to a certain given threshold θ

$$\text{Adj}[i,j] = \begin{cases} 1, & \text{if } i \neq j \text{ and } \text{Dist}[i,j] \leq \theta \\ 0, & \text{otherwise.} \end{cases}$$

Sequence of thresholds

- Consider a sequence of increasing thresholds $\theta_1 = 0.018$, $\theta_2 = 0.036$, $\theta_3 = 0.055$, and $\theta_4 = 0.01$, and $\theta_5 = 2.0$.
- So, we can compute a sequence of adjacency matrices Adj_1, Adj_2, Adj_3, Adj_4 and Adj_5, by using the previous threshold sequence $\theta_1 = 0.018$, $\theta_2 = 0.036$, $\theta_3 = 0.055$, and $\theta_4 = 0.01$, and $\theta_5 = 2.0$.
- Using the adjacency matrices Adj_1, Adj_2, Adj_3, Adj_4, and Adj_5, we can draw the distance graphs G_1, G_2, G_3, G_4, and G_5.
- Using the distance graphs G_1, G_2, G_3, G_4 and G_5, we can draw its degree distribution and find the sequences that have the biggest degree.
- Can we assume that the sequences found in the the previous item are the most "important" ones? Why?

Adjacency Matrix + Distance Graph + Degree Distribution

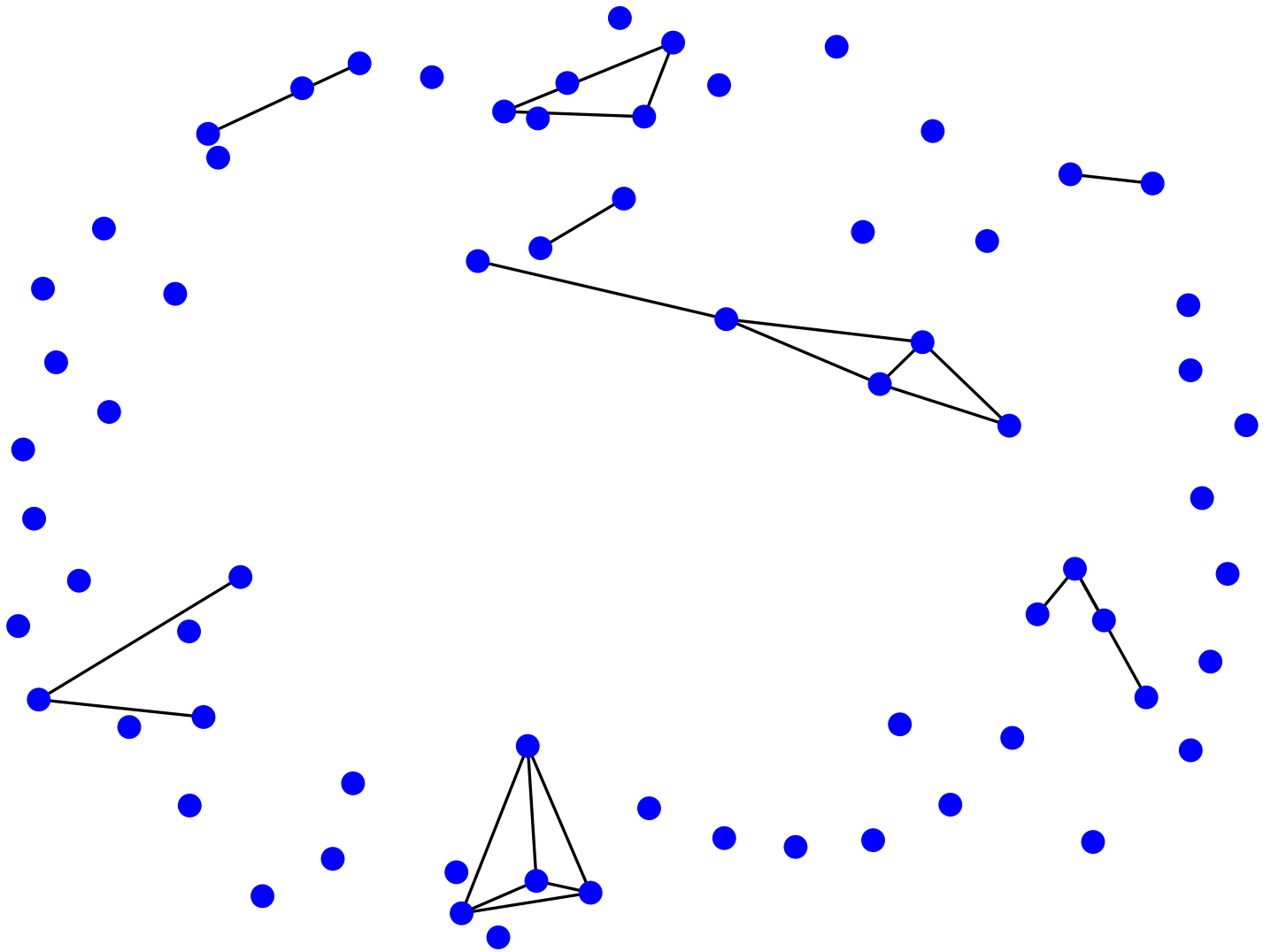
$$\theta_1 = 0.018$$

Sequences	0	1	2	3	4	...	63	64	65	66	67
						...					
0	0	0	0	0	0	...	0	0	0	0	0
1	0	0	0	0	0	...	0	0	0	0	0
2	0	0	0	0	0	...	0	0	0	0	0
3	0	0	0	0	0	...	0	0	0	0	0
4	0	0	0	0	0	...	0	0	0	0	0
...	..	..	..	..	..	...	..	..	..	..	..
63	0	0	0	0	0	...	0	1	0	0	0
64	0	0	0	0	0	...	1	0	0	0	0
65	0	0	0	0	0	...	0	0	0	0	0
66	0	0	0	0	0	...	0	0	0	0	0
67	0	0	0	0	0	...	0	0	0	0	0

[68 rows x 68 columns]

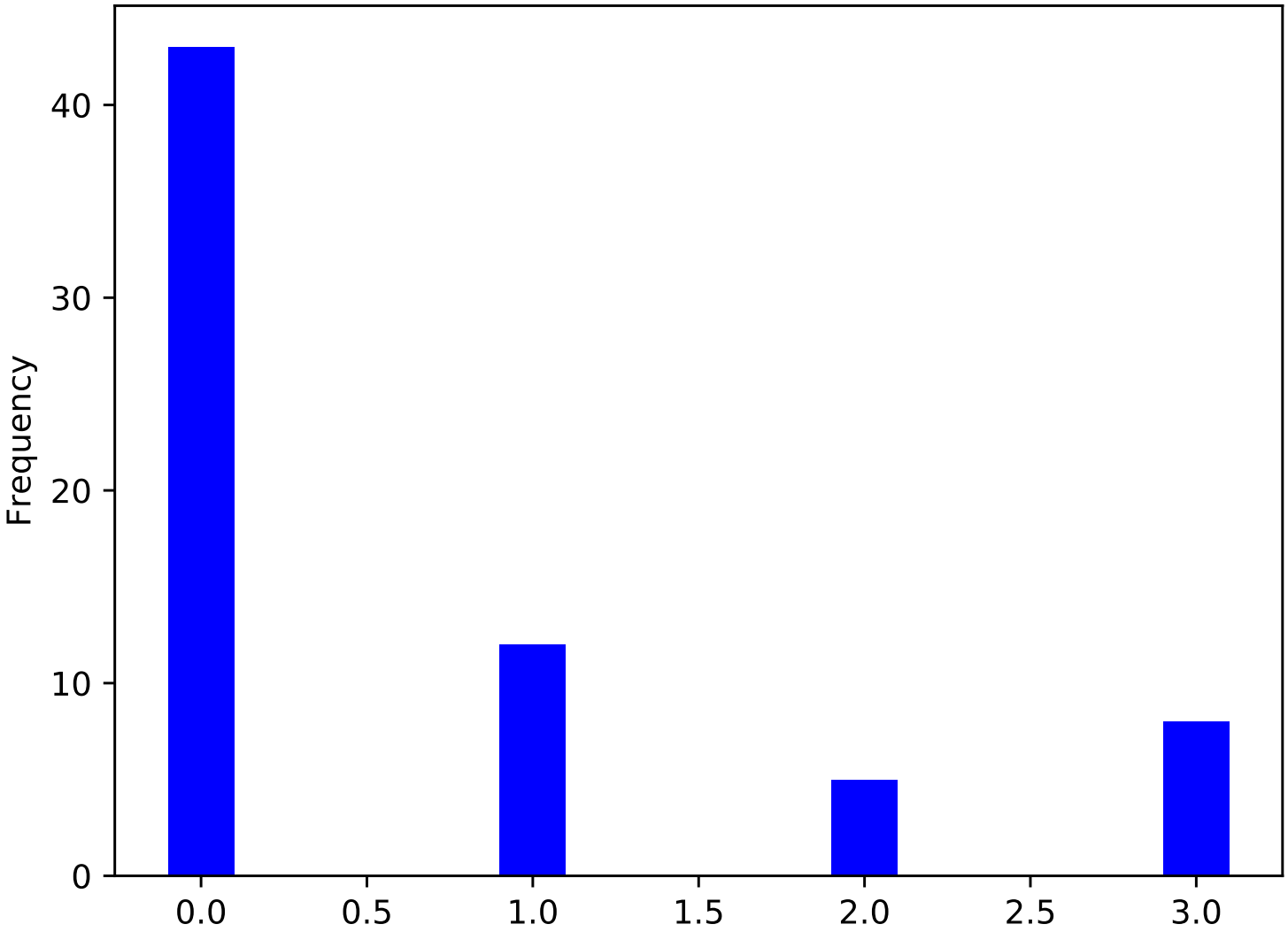
Adjacency Matrix + Distance Graph + Degree Distribution

$\theta_1 = 0.018$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_1 = 0.018$$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_1 = 0.018$$

- We have 8 nodes (sequences) with the highest degree = 3
- What are these nodes?

Adjacency Matrix + Distance Graph + Degree Distribution

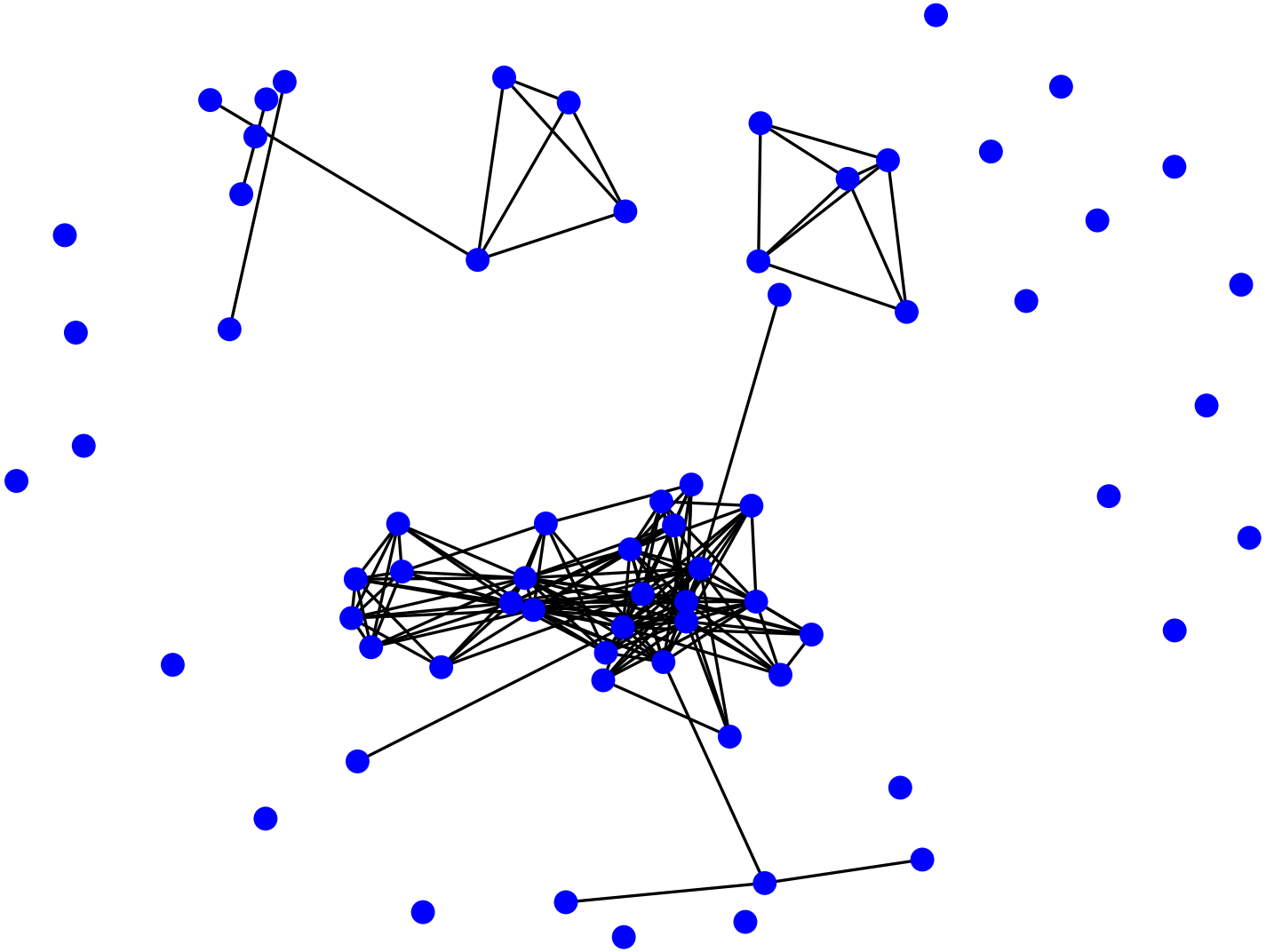
$$\theta_2 = 0.036$$

Sequences	0	1	2	3	4	...	63	64	65	66	67
						...					
0	0	0	0	0	0	...	0	0	0	0	0
1	0	0	0	0	0	...	0	0	0	0	0
2	0	0	0	0	0	...	0	0	0	0	0
3	0	0	0	0	0	...	0	0	0	0	0
4	0	0	0	0	0	...	0	0	1	0	0
...	..	..	..	..	..	...	..	..	..	..	..
63	0	0	0	0	0	...	0	1	0	0	0
64	0	0	0	0	0	...	1	0	0	0	0
65	0	0	0	0	1	...	0	0	0	0	0
66	0	0	0	0	0	...	0	0	0	0	0
67	0	0	0	0	0	...	0	0	0	0	0

[68 rows x 68 columns]

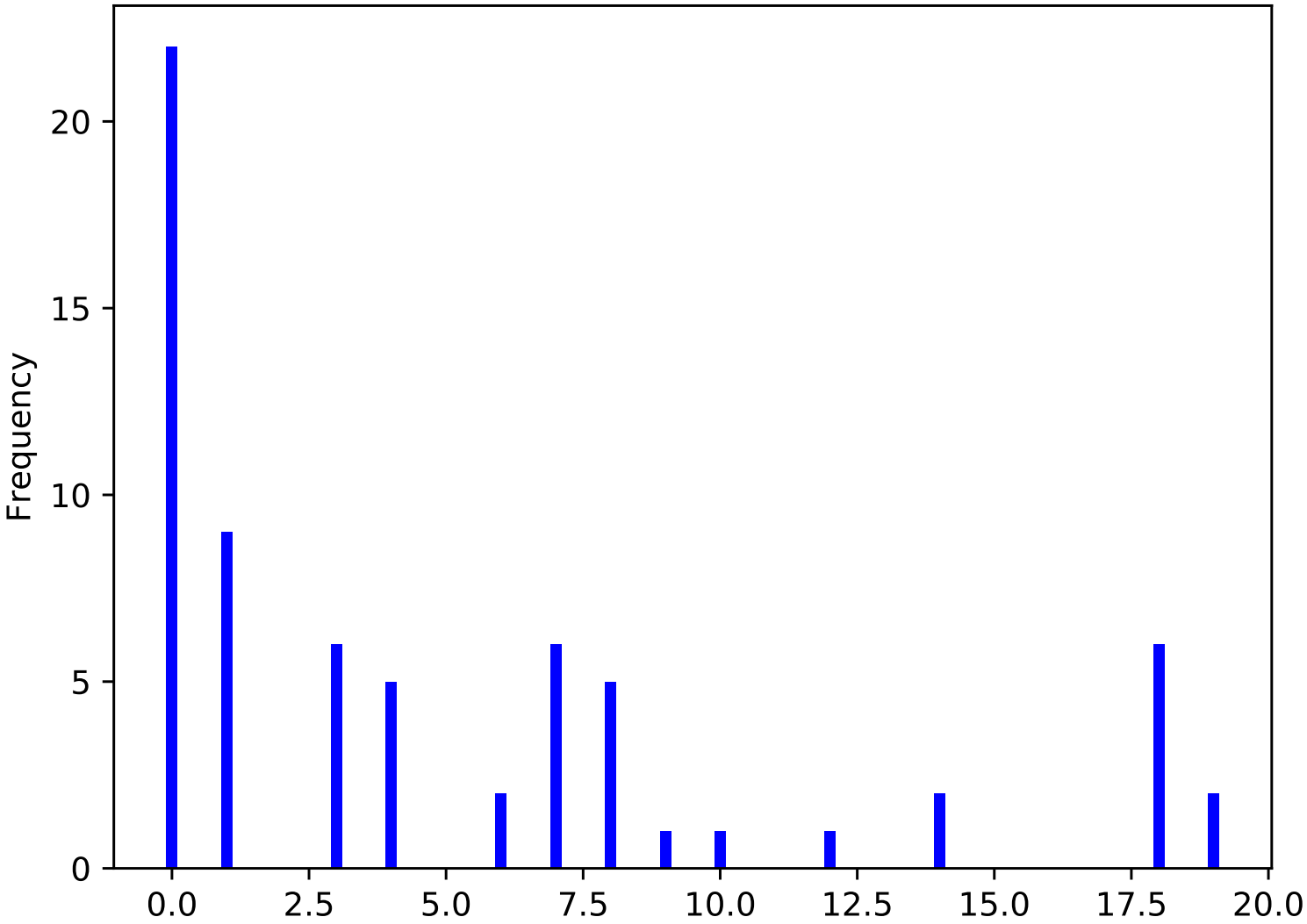
Adjacency Matrix + Distance Graph + Degree Distribution

$\theta_2 = 0.036$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_2 = 0.036$$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_2 = 0.036$$

- We have 2 nodes (sequences) with the highest degree = 19
- What are these nodes?

Adjacency Matrix + Distance Graph + Degree Distribution

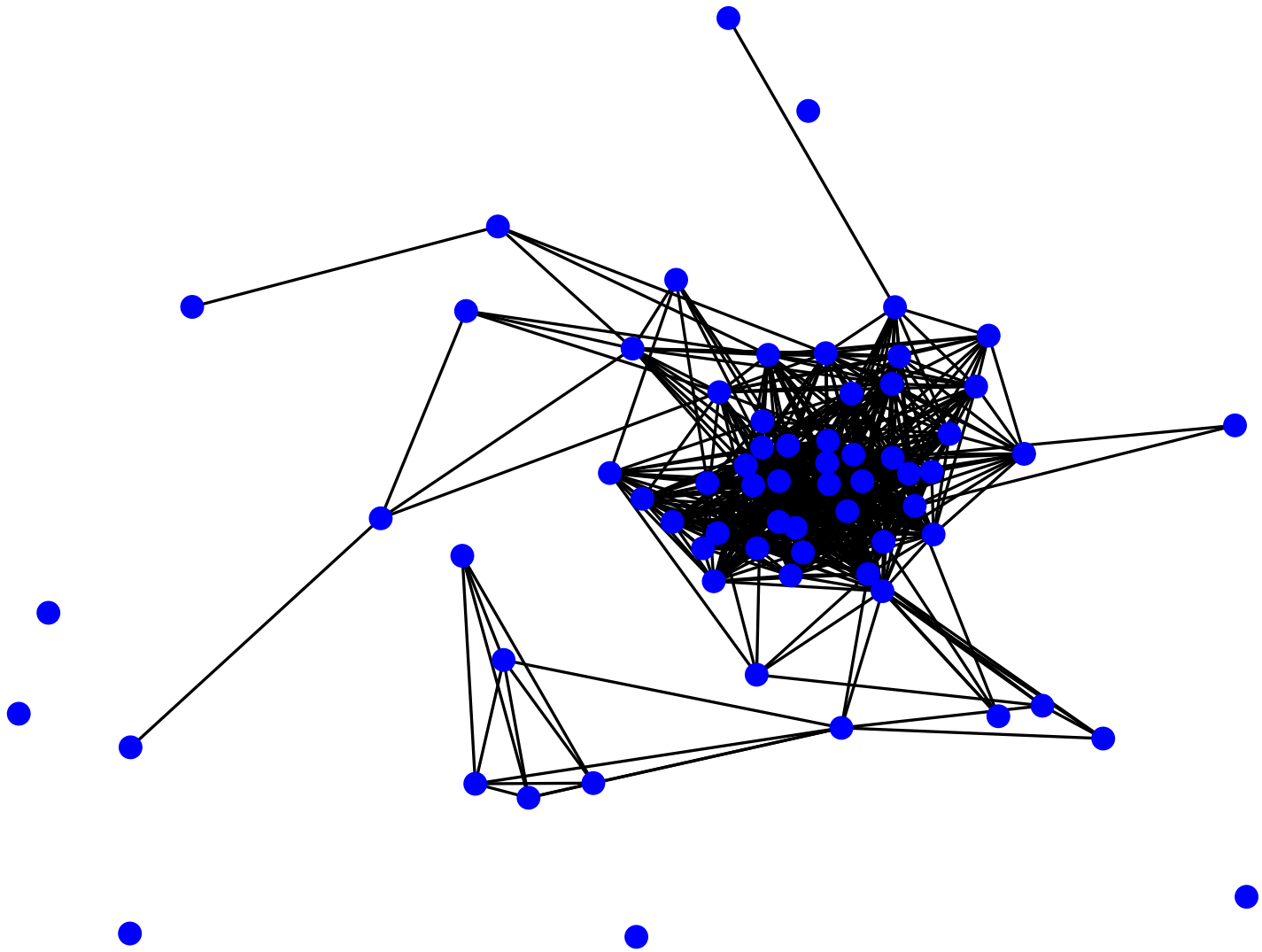
$$\theta_3 = 0.055$$

Sequences	0	1	2	3	4	...	63	64	65	66	67
						...					
0	0	0	1	0	0	...	0	0	1	0	0
1	0	0	0	0	0	...	0	0	0	0	0
2	1	0	0	0	1	...	0	0	1	0	0
3	0	0	0	0	1	...	0	0	0	0	0
4	0	0	1	1	0	...	1	1	1	0	0
...	..	..	..	..	..	...	..	..	..	..	..
63	0	0	0	0	1	...	0	1	1	0	0
64	0	0	0	0	1	...	1	0	1	0	0
65	1	0	1	0	1	...	1	1	0	0	0
66	0	0	0	0	0	...	0	0	0	0	0
67	0	0	0	0	0	...	0	0	0	0	0

[68 rows x 68 columns]

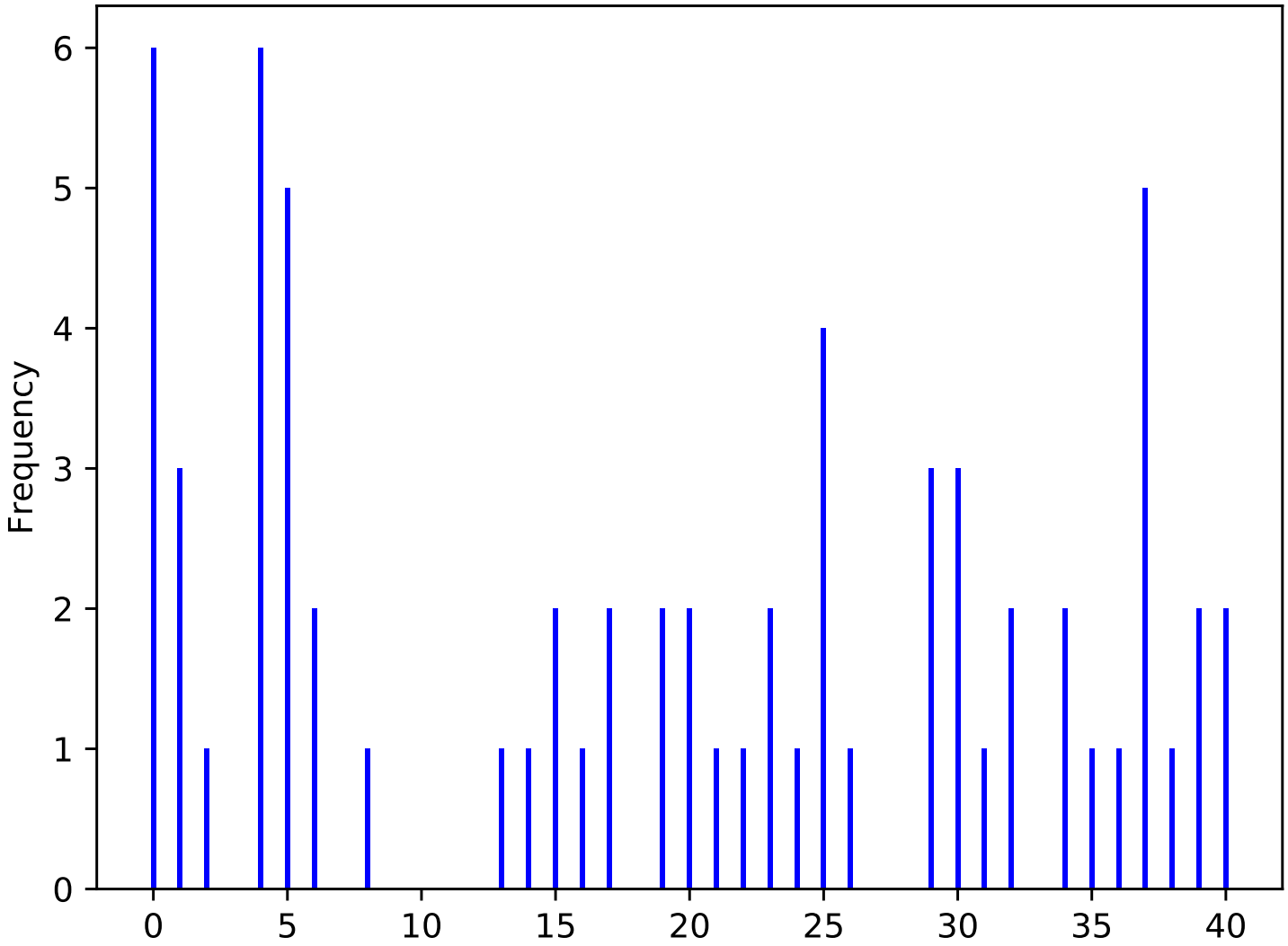
Adjacency Matrix + Distance Graph + Degree Distribution

$\theta_3 = 0.055$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_3 = 0.055$$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_3 = 0.055$$

- We have 2 nodes (sequences) with the highest degree = 40
- What are these nodes?

Adjacency Matrix + Distance Graph + Degree Distribution

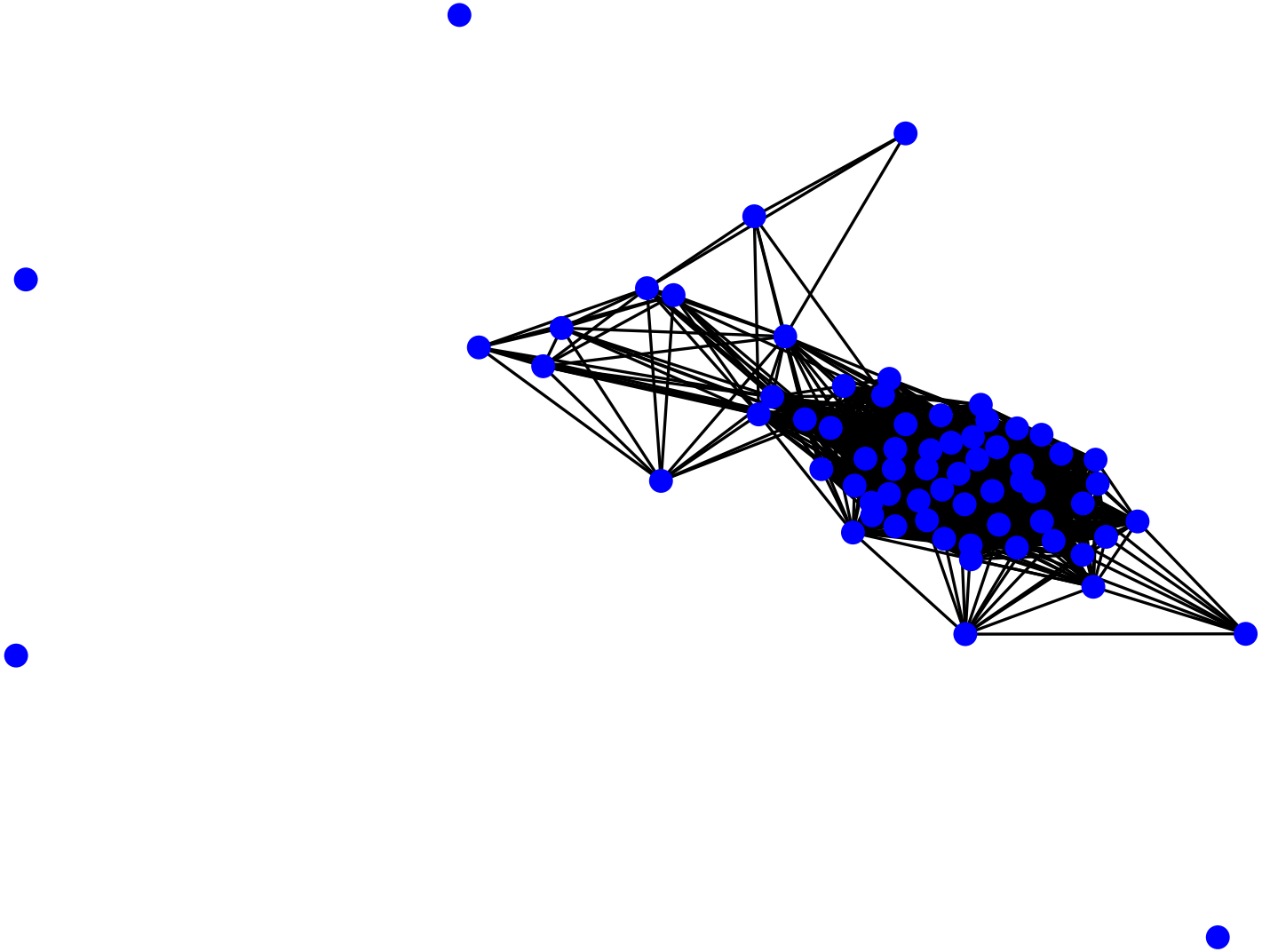
$$\theta_4 = 0.075$$

Sequences	0	1	2	3	4	...	63	64	65	66	67
						...					
0	0	1	1	0	1	...	0	0	1	0	0
1	1	0	1	1	1	...	1	1	1	0	0
2	1	1	0	1	1	...	1	1	1	0	0
3	0	1	1	0	1	...	1	0	1	0	0
4	1	1	1	1	0	...	1	1	1	0	0
...	..	..	..	..	..	...	..	..	..	..	..
63	0	1	1	1	1	...	0	1	1	0	0
64	0	1	1	0	1	...	1	0	1	0	0
65	1	1	1	1	1	...	1	1	0	0	0
66	0	0	0	0	0	...	0	0	0	0	0
67	0	0	0	0	0	...	0	0	0	0	0

[68 rows x 68 columns]

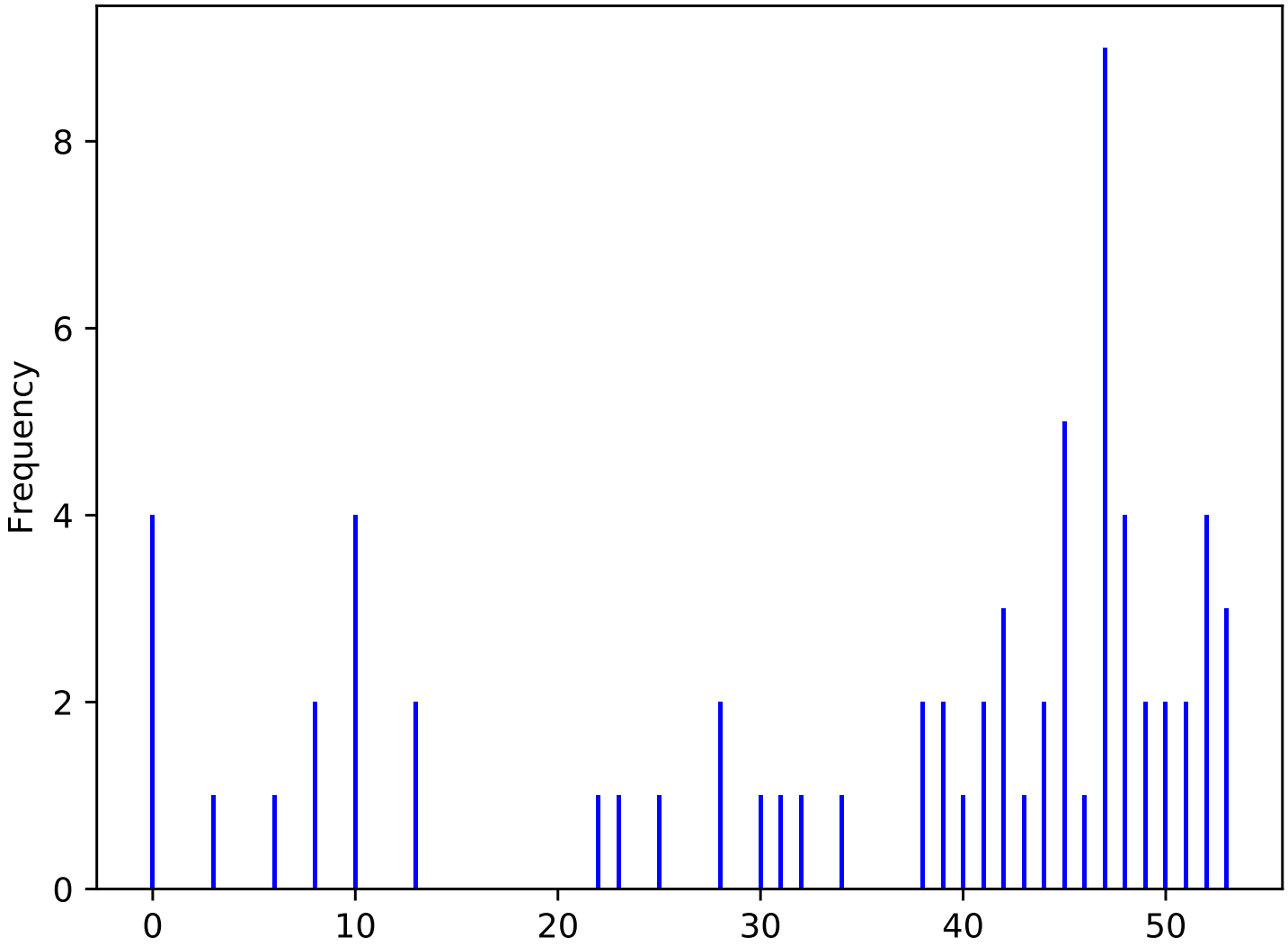
Adjacency Matrix + Distance Graph + Degree Distribution

$\theta_4 = 0.075$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_4 = 0.075$$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_4 = 0.075$$

- We have 3 nodes (sequences) with the highest degree = 53
- What are these nodes?

Adjacency Matrix + Distance Graph + Degree Distribution

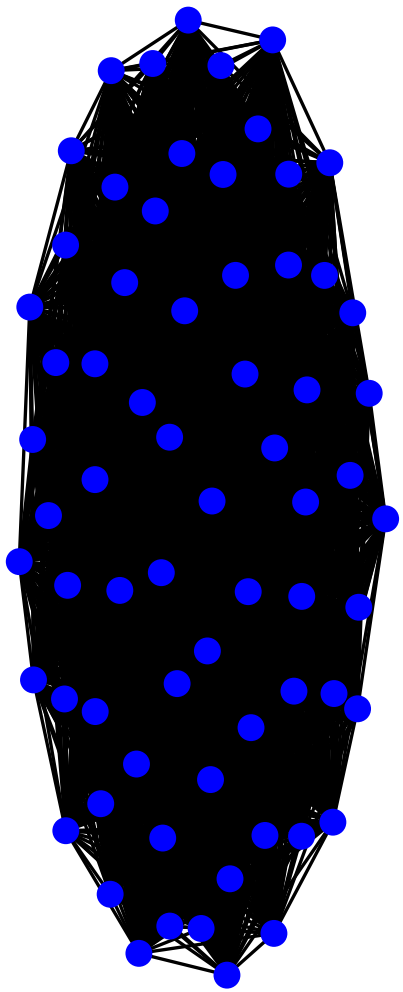
$$\theta_5 = 2.0$$

Sequences	0	1	2	3	4	...	63	64	65	66	67
						...					
0	0	1	1	1	1	...	1	1	1	1	1
1	1	0	1	1	1	...	1	1	1	1	1
2	1	1	0	1	1	...	1	1	1	1	1
3	1	1	1	0	1	...	1	1	1	1	1
4	1	1	1	1	0	...	1	1	1	1	1
...	..	..	..	..	..	...	..	..	..	..	..
63	1	1	1	1	1	...	0	1	1	1	1
64	1	1	1	1	1	...	1	0	1	1	1
65	1	1	1	1	1	...	1	1	0	1	1
66	1	1	1	1	1	...	1	1	1	0	1
67	1	1	1	1	1	...	1	1	1	1	0

[68 rows x 68 columns]

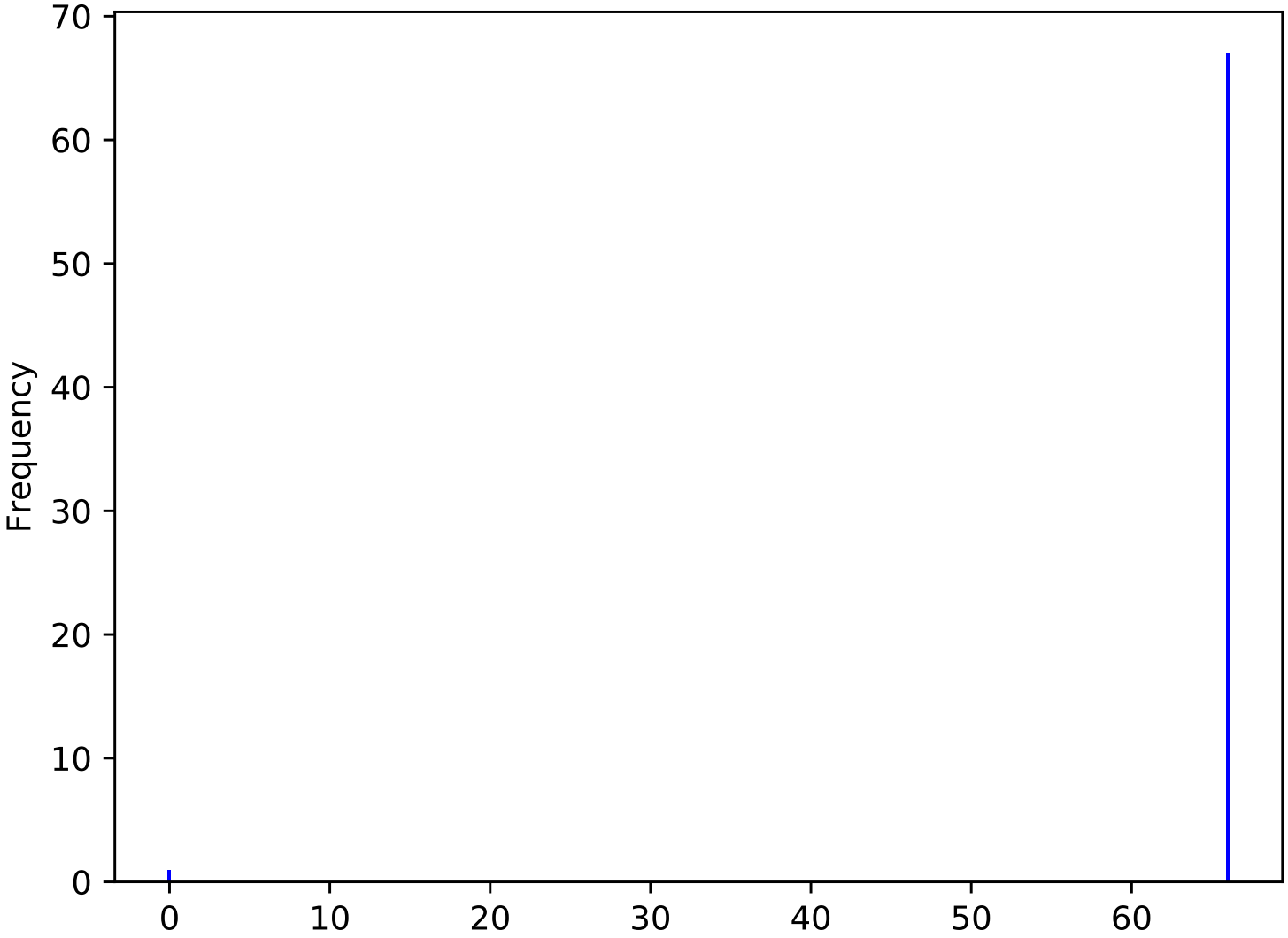
Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_5 = 2.0$$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_5 = 2.0$$



Adjacency Matrix + Distance Graph + Degree Distribution

$$\theta_5 = 2.0$$

- We have 67 nodes (sequences) with the highest degree = 66
- What is the isolated node?