Partitional and Hierarchical Clustering in Python

Members

- Mehul Agrawal – 2017A7PS0054H
- Abhishek Dhanwani – 2017A7PS0161H
- Shah Vishakh Rakesh – 2017A7PS1445H

Dataset

Amino Acid Sequences (amino.fasta)

Similarity Metric

- Used the Global Sequence Alignment metric for similarity.

K-Means

- **K=3**

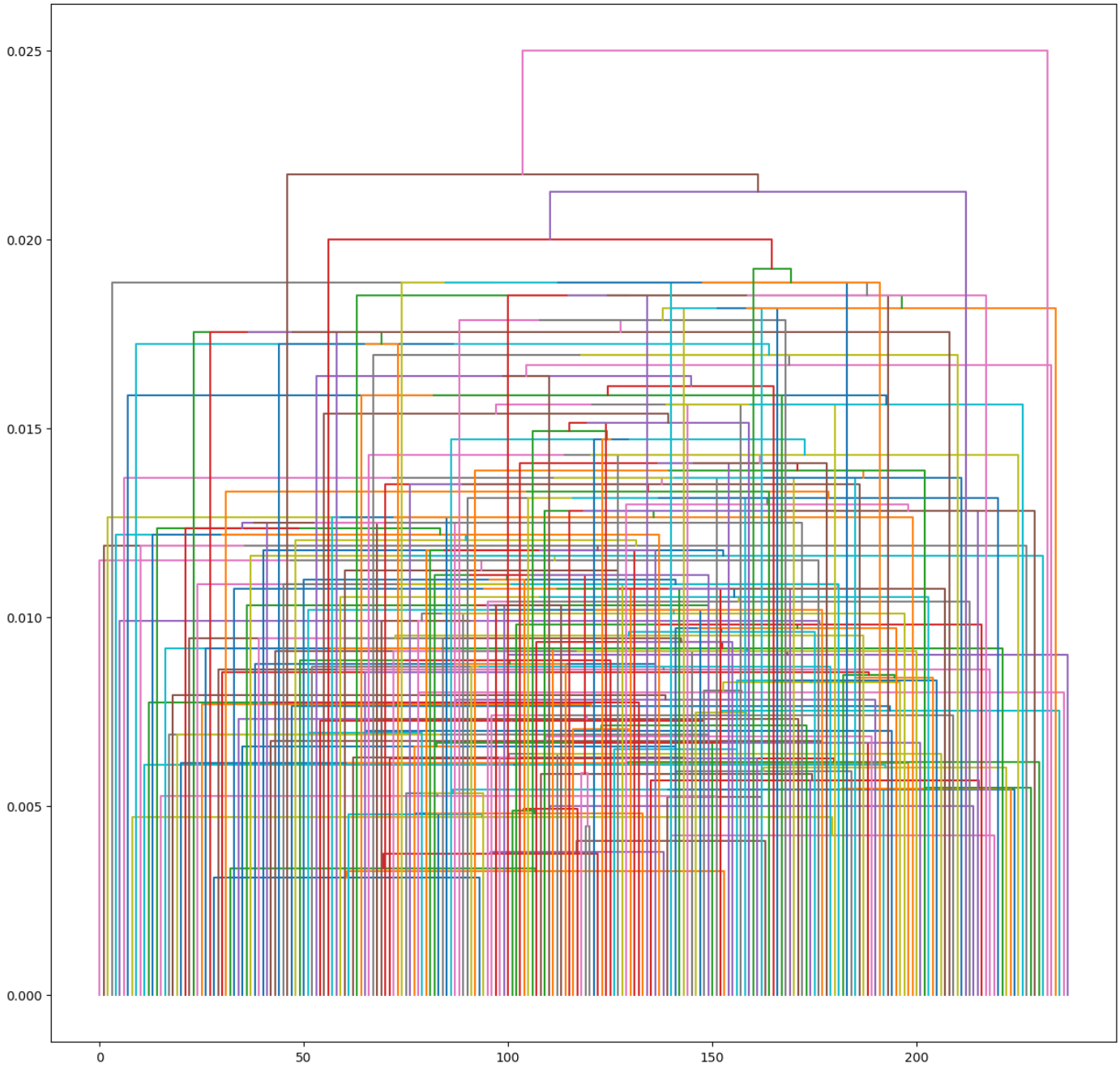
```
1 th Cluster
0, 3, 5, 7, 8, 9, 10, 11, 16, 18, 20, 21, 25, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36, 38, 44, 45, 50, 51, 53, 54, 55, 56, 57,
58, 59, 61, 62, 63, 64, 65, 66, 68, 71, 73, 77, 79, 80, 81, 83, 85, 86, 89, 91, 92, 93, 96, 97, 100, 101, 102, 103, 104, 105,
106, 112, 113, 114, 115, 116, 117, 118, 121, 122, 123, 125, 126, 128, 132, 133, 135, 138, 139, 140, 141, 146, 147, 151, 153, 1
54, 157, 158, 162, 163, 164, 167, 170, 173, 176, 177, 179, 180, 182, 183, 185, 186, 190, 192, 194, 195, 196, 198, 199, 200, 20
1, 203, 204, 206, 208, 210, 211, 213, 214, 215, 216, 219, 220, 221, 223, 224, 228, 229, 231, 233, 234, 235,

2 th Cluster
1, 4, 12, 13, 15, 17, 19, 24, 37, 39, 40, 49, 52, 60, 67, 69, 70, 72, 75, 78, 84, 88, 94, 98, 99, 108, 110, 111, 119, 120, 12
4, 127, 131, 136, 137, 144, 148, 149, 150, 156, 161, 165, 168, 171, 174, 175, 178, 181, 184, 187, 188, 197, 205, 207, 212, 21
7, 218, 222, 225, 227, 232,

3 th Cluster
2, 6, 14, 22, 23, 33, 41, 42, 43, 46, 47, 48, 74, 76, 82, 87, 90, 95, 107, 109, 129, 130, 134, 142, 143, 145, 152, 155, 159, 1
60, 166, 169, 172, 189, 191, 193, 202, 209, 226, 230,
```

- **K=5**

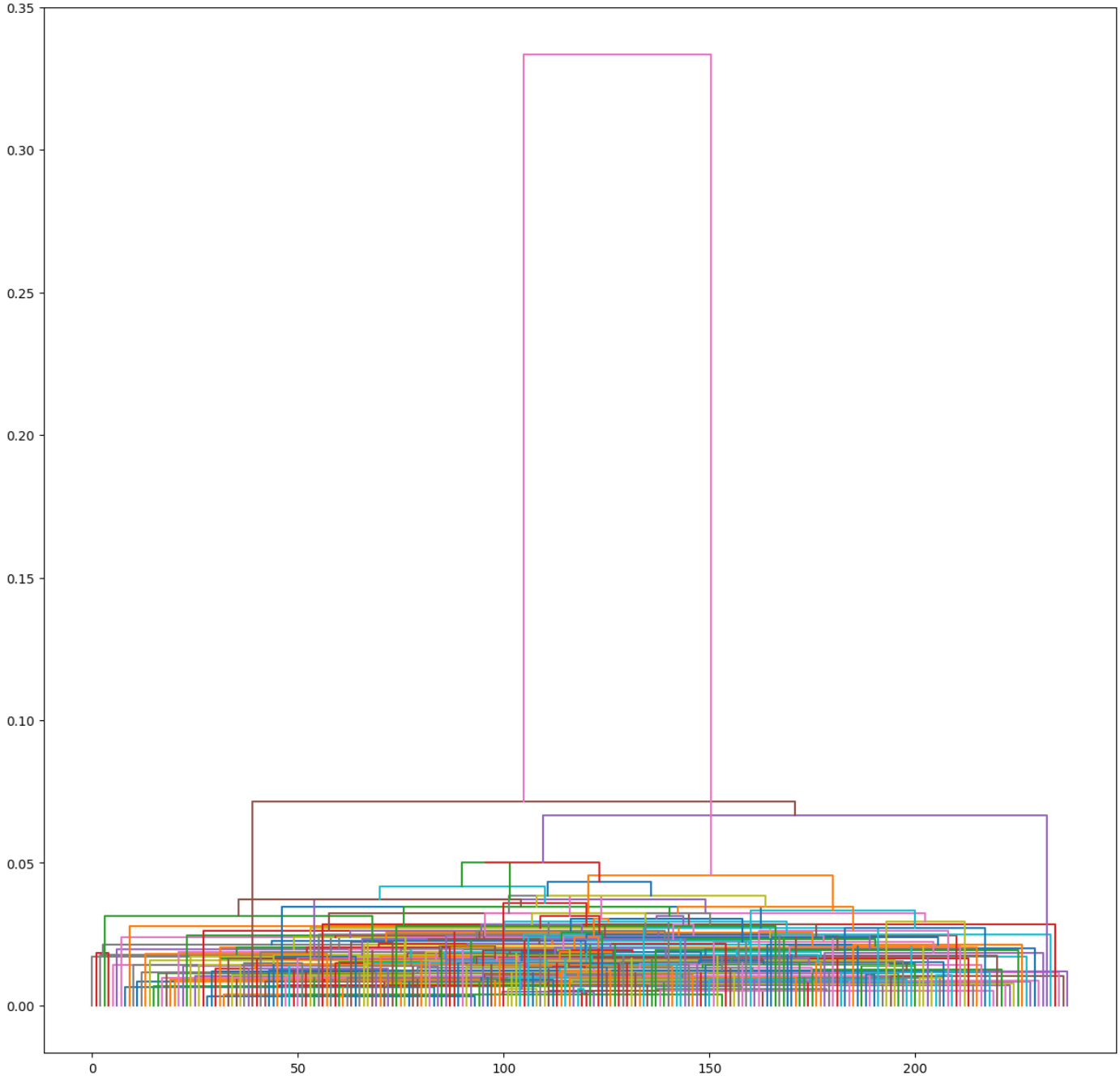
```
1 th Cluster
0, 5, 7, 8, 9, 10, 11, 16, 18, 21, 25, 26, 28, 29, 30, 31, 32, 34, 35, 36, 38, 45, 50, 51, 53, 54, 55, 57, 58, 61, 62, 64, 68,
71, 77, 79, 81, 83, 85, 86, 89, 91, 93, 97, 101, 102, 103, 104, 105, 106, 112, 113, 114, 115, 116, 117, 118, 122, 123, 125, 12
6, 128, 132, 133, 138, 139, 140, 141, 147, 151, 153, 154, 157, 158, 162, 163, 164, 167, 170, 173, 176, 177, 180, 182, 183, 18
5, 186, 190, 192, 194, 195, 196, 198, 199, 200, 201, 203, 204, 206, 208, 210, 211, 213, 214, 215, 216, 219, 220, 221, 223, 22
4, 228, 229, 231, 233, 234, 235,

2 th Cluster
1, 12, 13, 15, 17, 19, 24, 37, 39, 40, 49, 52, 60, 67, 69, 70, 75, 78, 84, 88, 94, 98, 99, 108, 110, 111, 119, 120, 124, 127,
131, 136, 137, 144, 149, 150, 156, 161, 165, 168, 171, 174, 175, 178, 181, 184, 187, 188, 197, 205, 207, 217, 218, 222, 225, 2
27,

3 th Cluster
2, 6, 14, 22, 23, 33, 41, 42, 43, 46, 47, 48, 76, 82, 87, 90, 95, 107, 109, 129, 130, 134, 142, 143, 145, 152, 155, 159, 160,
166, 169, 172, 189, 193, 202, 209, 226, 230,

4 th Cluster
3,

5 th Cluster
4, 20, 27, 44, 56, 59, 63, 65, 66, 72, 73, 74, 80, 92, 96, 100, 121, 135, 146, 148, 179, 191, 212, 232,
```

Agglomerative Hierarchical Clustering

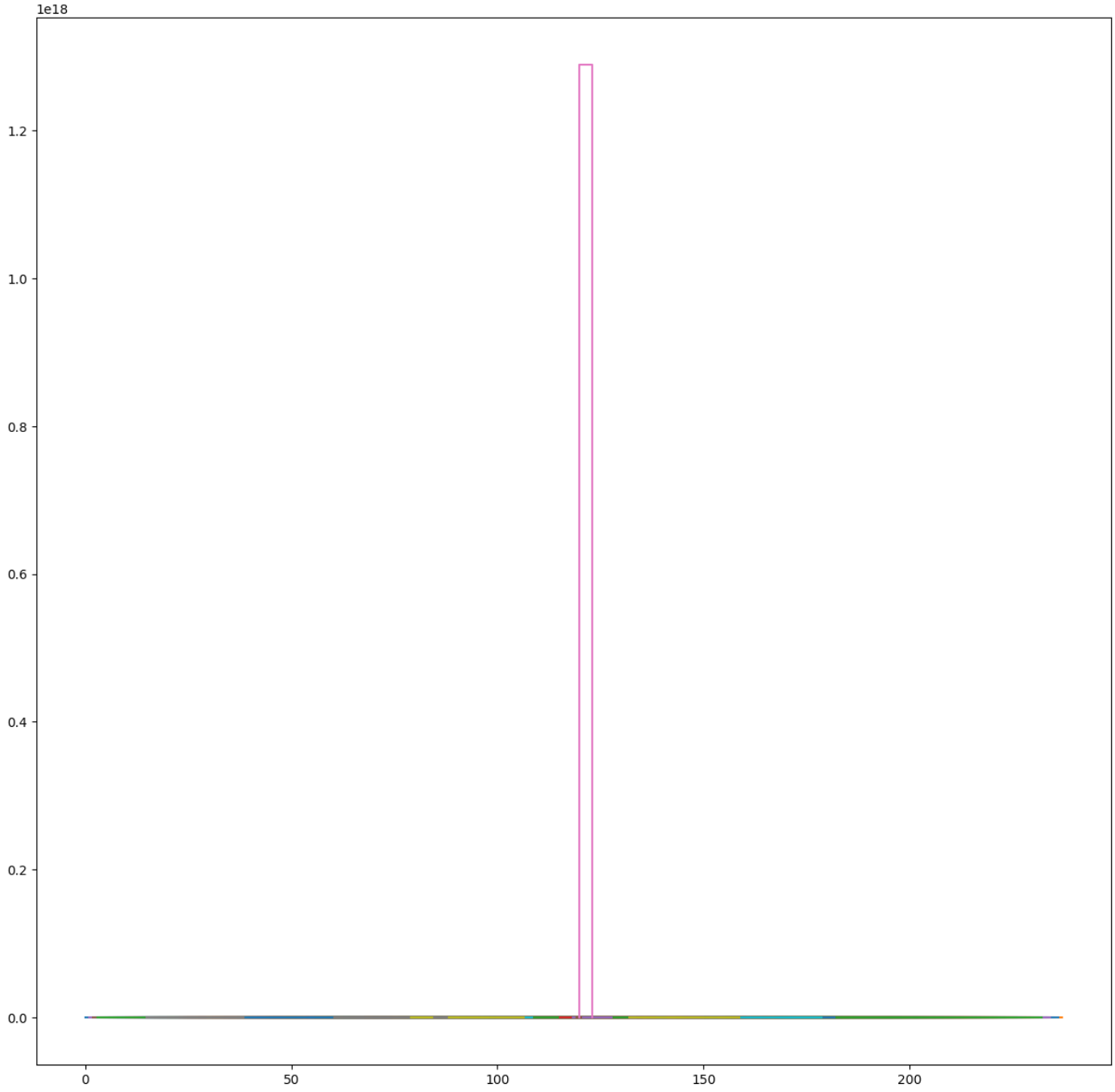
- Single-Linkage



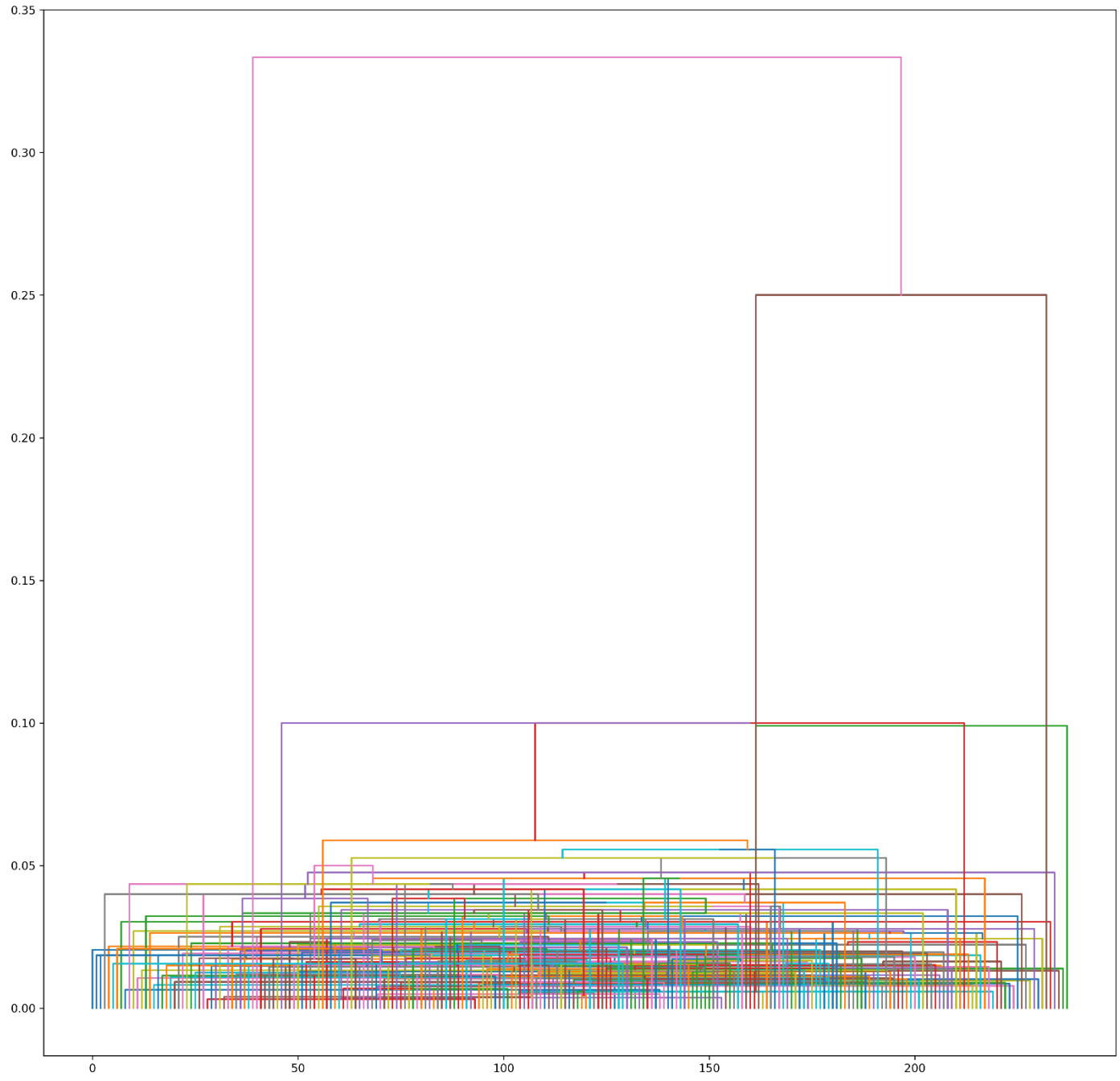
- Complete-Linkage



- Average-Linkage



Divisive Hierarchical Clustering (DIANA)



Comparison of K-Means and Hierarchical Clustering

As we can see, in K-Means each element is exclusive to a single cluster since it is partition-based clustering (unlike both approaches of hierarchical). Also, the number of clusters is user-defined.