Long Reads Alignment

Giulia Guidi[1,2], Aydın Buluç[1]
{gguidi, abuluc}@lbl.gov
[1]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, USA
[2]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

April 27, 2017

1 Problem Statement

Sequencing technologies produce a large amount of redundant data with respect to the real genomic sequence to be assembled. Furthermore, the Pacific Bioscience techonology, considered in this work, is prone to an error rate equal to 15%.
The ultimate goal is to efficiently exploit long reads produced by Pacific Bioscience technology to *de novo* assemble a genome. To achieve this scope, the first problem to be addressed is the determination of a reliable kmer sub-set to be used as a vector during the alignment phase of the genome assembly. This report briefly explains the preliminaries analysis computed to identify the correct approach to the problem.

2 Reliable k-mers

The first computed analysis regards the identification of a reliable set of k-mers (RKS). The reads were generated starting from the Escherichia Coli genome, using the PacBio reads simulator. Then, using Jellyfish software, we generated different k-mers dataset from these reads, varying the value of k, we took into account values from 15 to 29.
The k-mers belonging to the RKS set are choosen looking at their occurrence among the generated reads. So, from Jellyfish output, we chose k-mers that occur in the range [depth/2 and 2*depth][1].
The following plots (Figure 1, 2, 3 and 4) show the percentage of RKS over the total number of generated k-mers per each k-mer occurrence, taking into account different k-mers length.
For the following analysis, we decided to take into account k values equal to 15, 17 an 19.

3 Existence and unique existence of k-mers in the genome

Here, we computed some statistics con the previously selected dataset of k-mer. The goal was to identify which groups of k-mers, based on their occurrences in the initial dataset (that means the occurrence in the reads generated from the PacBio simulator), presented the highest percentages of unique existence in the genome. To do that, we found the matches between each k-mers of the dataset and the genome sequence.
The *GroupID* variable represents the occurrence of a k-mer in the initial dataset. *State A*, *State B* and *State C* indicate the number of k-mers (within a group, that means with the same occurence in the initial dataset) with *real* occurence in the genome equal to 0, equal to 1 and greater than 1,

[1]The depth is equal to 30.

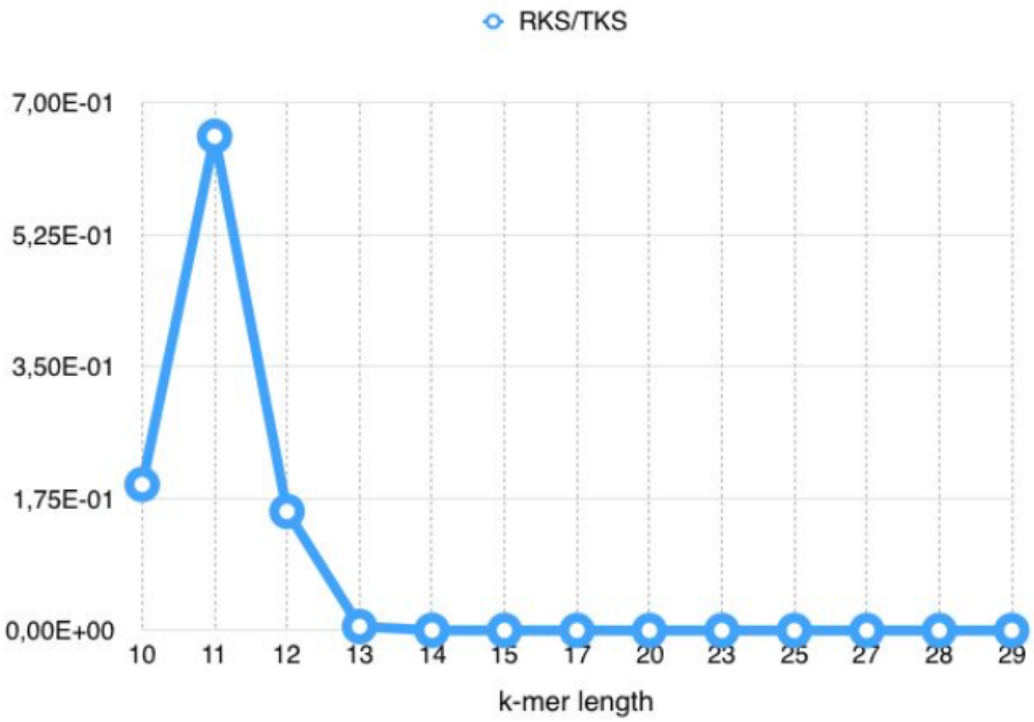

Figure 1: Percentage of RKS over the total number of generated k-mers per each k-mer occurrence.

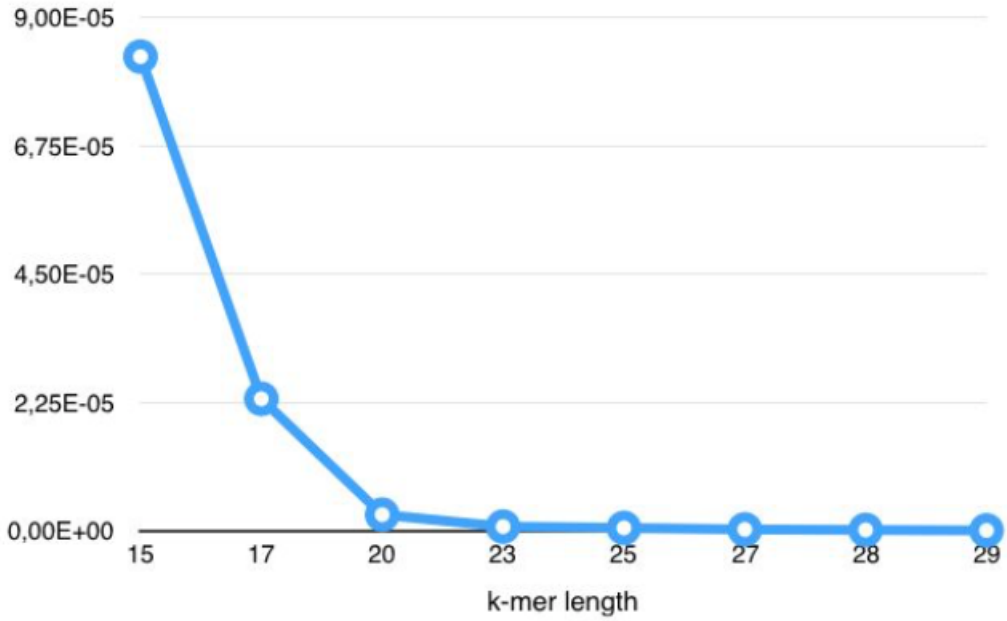

Figure 2: Percentage of RKS over the total number of generated k-mers per each k-mer occurrence.

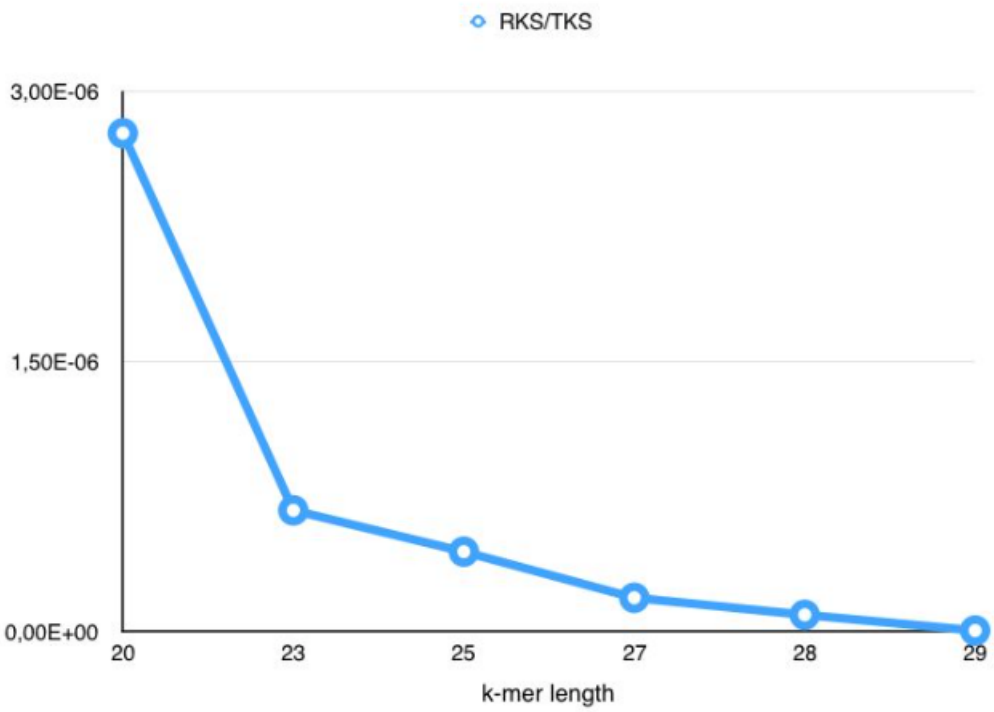

Figure 3: Percentage of RKS over the total number of generated k-mers per each k-mer occurrence.

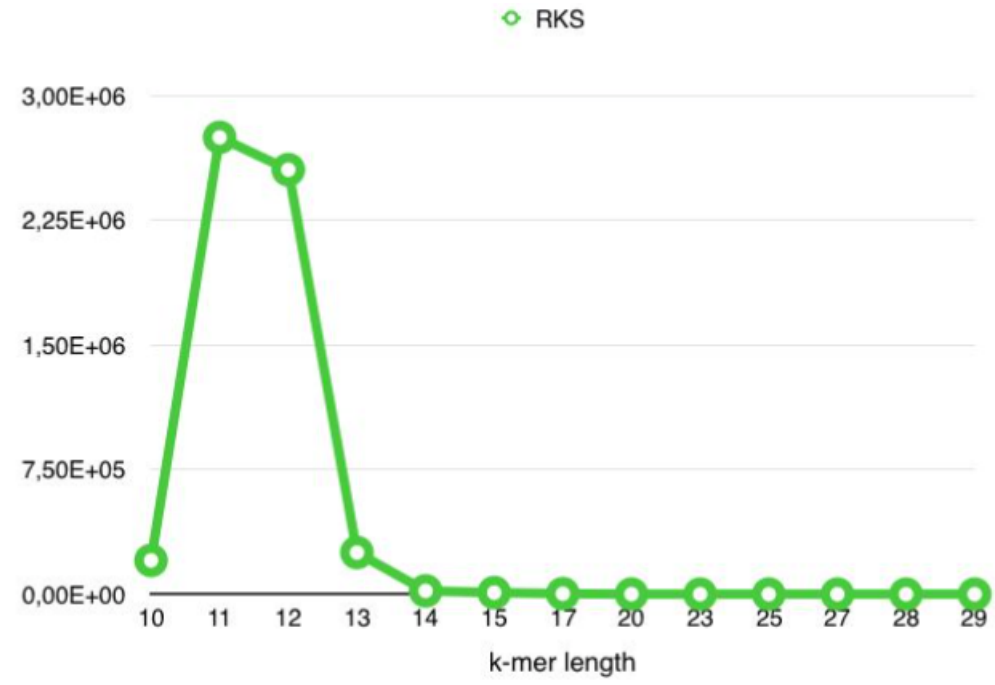

Figure 4: Percentage of RKS over the total number of generated k-mers per each k-mer occurrence.

respectively. Regarding the presented plots, for all of them the x-axis is based on the *GroupID*, while the y-axis represents both the percentage of k-mers existence and unique existence of k-mers in the genome (Figure 5, 6 and 7).
The percentage of existence (Eq.1) and unique existence (Eq.2) are defined as:

$$Existence = \frac{StateB + StateC}{StateA + StateB + StateC} \cdot 100 \tag{1}$$

$$Unique\,existence = \frac{StateB}{StateA + StateB + StateC} \cdot 100 \tag{2}$$

Figure 5 shows the obtained results with the k-mer length equal to 15, the *GroupID* goes up to 268. The greater unique existence percentages, drawn with red dashed line, are associated with *GroupID* equal to 5, 6 and 7 (with a maximux of 47.84% for *GroupID* equal to 6).
Figure 6 shows the obtained results with the k-mer length equal to 17, in this case the *GroupID* goes up to 166. The greater unique existence percentages, drawn with red dashed line, are associated with *GroupID* equal to 4 and 5 (with a maximux of 48.01% for *GroupID* equal to 4).
Figure 7 shows the obtained results with the k-mer length equal to 19, in this case the *GroupID* goes up to 118. The greater unique existence percentages, drawn with red dashed line, are associated with *GroupID* equal to 3 and 4 (with a maximux of 46.69% for *GroupID* equal to 4).
In order to compare results with different k-mer length, a maximum *GroupID* equal to 90 is considered for all the x-axis[2].

4 Proposed approach

The ultimate goal of the previous analysis consists in obtaining a [k-mer -by- read] matrix (where A_{ij} is the occurrence/absence of k-mer i in read j). This matrix is used as starting point for the construction of a feature vector in finding alignments among the reads.
From the previous statistics we decided to take into account the k value equal to 15 and consider the k-mers that occur in the range [4,8] as this range provides the highest percentages of unique existence in the *Escherichia Coli* genome.
Firsly, our approach consists in the creation of a dictionary containing all the k-mers belonging to the defined range. Then, we construct a [kmer -by- read] sparse matrix, where the $\{i, j\}$ cell is a *pair* data structure. The first value of the pair correspond to the identification number of the k-mer i contained in the read j, while the second value is a *vector* data structure where all the positions of that k-mer in the considered read are saved.
Once creating the matrix, we compute its transpose [read -by- k-mer] in order to multiply them and obtain a [read -by- read] matrix. We implement the calculation to obtain as final cell $\{i, j\}$ a *map* data structure organized as follow. The *keys* correspond to the identification numbers of the shared k-mers between the two reads, while the *values* are pairs of vectors containing the k-mers positions on the two reads.
To do: filter on the matrix to identify reads pairs sampled from the same region of the genome, implement an Apply() function to compute the delta-pos among k-mers belonging to the same read and compare k-mers delta-pos between reads sharing the same k-mers.

[2]The depth is equal to 30.

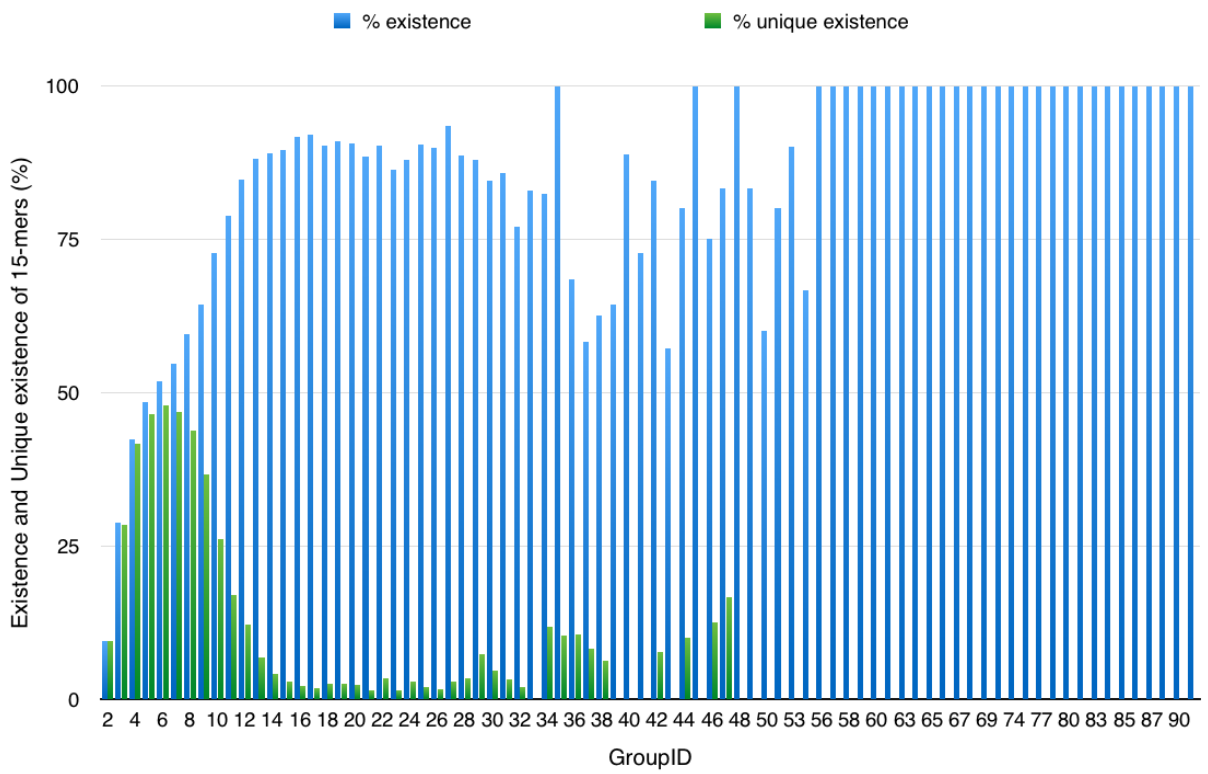

Figure 5: Existence and Unique existence of 15-mers

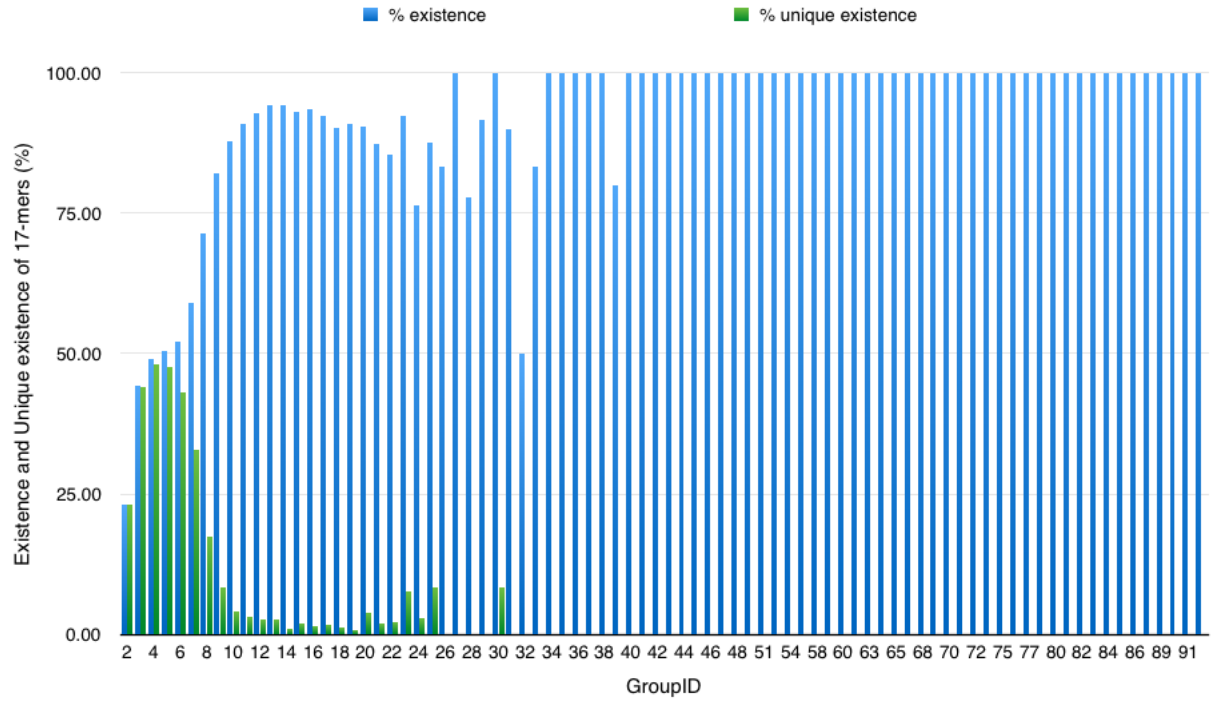

Figure 6: Existence and Unique existence of 17-mers

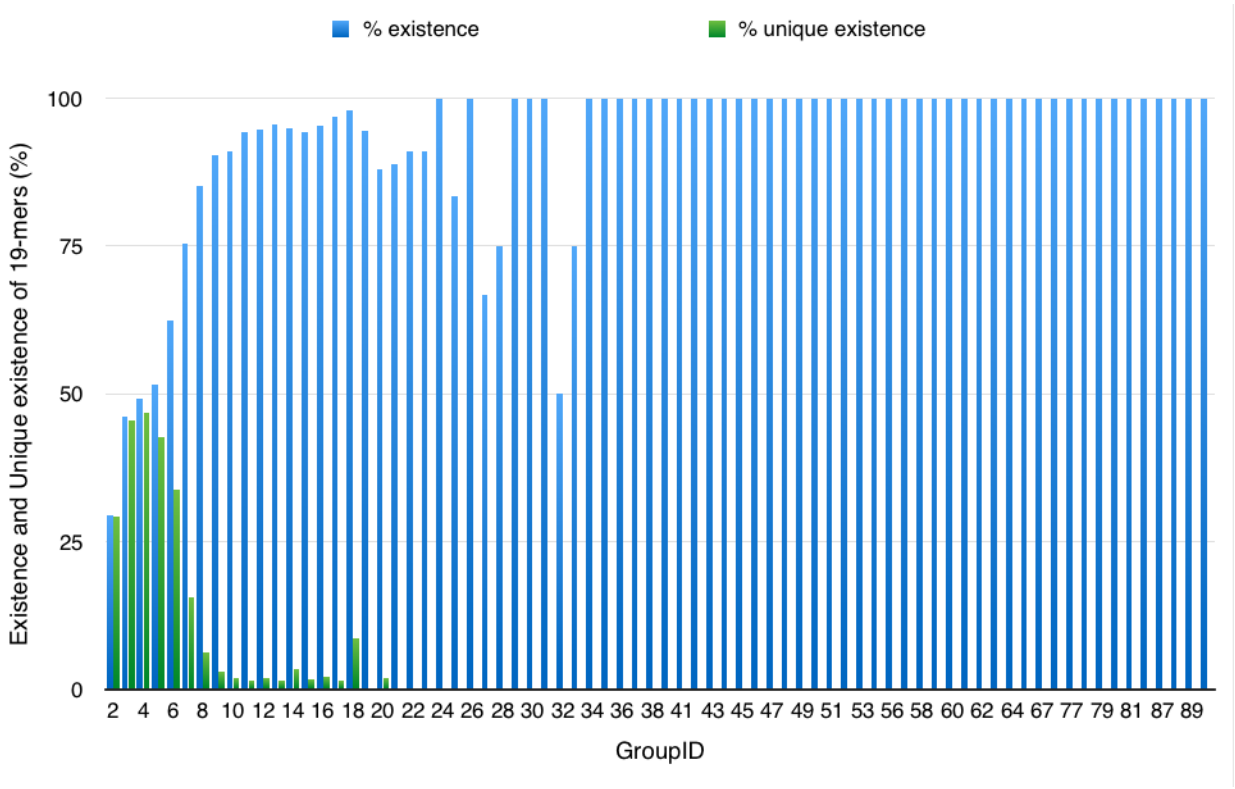

Figure 7: Existence and Unique existence of 19-mers

5 Statistics about reads pairs

A Tables

Table 1: Existence and Unique existence of 15-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
2	11054964	1163913	6137	9.57	9.52
3	2067194	828871	8706	28.83	28.53
4	630466	455824	9450	42.46	41.60
5	230524	207638	8646	48.41	46.47
6	83858	83304	6986	51.85	47.84
7	29699	30686	5231	54.74	46.77
8	10289	11145	4045	59.62	43.74
9	3953	4078	3081	64.43	36.70
10	1621	1555	2756	72.67	26.21
11	847	681	2466	78.79	17.05
12	456	364	2161	84.70	12.21
13	285	162	1941	88.07	6.78
14	217	84	1684	89.07	4.23
15	158	43	1321	89.62	2.83
16	106	27	1148	91.73	2.11
17	79	18	901	92.08	1.80
18	79	21	708	90.22	2.60
19	55	15	534	90.89	2.48
20	49	12	462	90.63	2.29
21	47	6	356	88.51	1.47
22	34	12	301	90.20	3.46
23	39	4	242	86.32	1.40
24	29	7	204	87.92	2.92
25	19	4	176	90.45	2.01
26	18	3	157	89.89	1.69
27	11	5	153	93.49	2.96
28	13	4	98	88.70	3.48
29	13	8	87	87.96	7.41
30	13	4	67	84.52	4.76
31	9	2	52	85.71	3.17
32	11	1	36	77.08	2.08
33	7	0	34	82.93	0.00
34	6	4	24	82.35	11.76
35	0	3	26	100.00	10.34
36	6	2	11	68.42	10.53
37	5	1	6	58.33	8.33
38	6	1	9	62.50	6.25
39	5	0	9	64.29	0.00
40	1	0	8	88.89	0.00
41	3	0	8	72.73	0.00
42	2	1	10	84.62	7.69
43	3	0	4	57.14	0.00
44	2	1	7	80.00	10.00

Table 1: Existence and Unique existence of 15-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
45	0	0	6	100.00	0.00
46	2	1	5	75.00	12.50
47	1	1	4	83.33	16.67
48	0	0	4	100.00	0.00
49	1	0	5	83.33	0.00
50	2	0	3	60.00	0.00
51	1	0	4	80.00	0.00
53	1	0	9	90.00	0.00
54	1	0	2	66.67	0.00
56	0	0	1	100.00	0.00
57	0	0	5	100.00	0.00
58	0	0	3	100.00	0.00
59	0	0	4	100.00	0.00
60	0	0	1	100.00	0.00
62	0	0	2	100.00	0.00
63	0	0	3	100.00	0.00
64	0	0	1	100.00	0.00
65	0	0	1	100.00	0.00
66	0	0	1	100.00	0.00
67	0	0	1	100.00	0.00
68	0	0	1	100.00	0.00
69	0	0	1	100.00	0.00
72	0	0	1	100.00	0.00
74	0	0	2	100.00	0.00
75	0	0	1	100.00	0.00
77	0	0	1	100.00	0.00
79	0	0	1	100.00	0.00
80	0	0	3	100.00	0.00
81	0	0	1	100.00	0.00
83	0	0	1	100.00	0.00
84	0	0	1	100.00	0.00
85	0	0	2	100.00	0.00
86	0	0	1	100.00	0.00
87	0	0	1	100.00	0.00
88	0	0	1	100.00	0.00
90	0	0	3	100.00	0.00
92	0	0	1	100.00	0.00
93	0	0	2	100.00	0.00
94	0	0	1	100.00	0.00
97	0	0	1	100.00	0.00
98	0	0	2	100.00	0.00
99	0	0	2	100.00	0.00
100	0	0	3	100.00	0.00
101	0	0	2	100.00	0.00
102	0	0	2	100.00	0.00

Table 1: Existence and Unique existence of 15-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
103	0	0	2	100.00	0.00
105	0	0	1	100.00	0.00
106	0	0	1	100.00	0.00
108	0	0	1	100.00	0.00
109	0	0	2	100.00	0.00
111	0	0	1	100.00	0.00
114	0	0	1	100.00	0.00
115	0	0	1	100.00	0.00
116	0	0	1	100.00	0.00
117	0	0	3	100.00	0.00
118	0	0	3	100.00	0.00
121	0	0	1	100.00	0.00
122	0	0	5	100.00	0.00
123	0	0	3	100.00	0.00
124	0	0	1	100.00	0.00
125	0	0	1	100.00	0.00
126	0	0	3	100.00	0.00
127	0	0	3	100.00	0.00
128	0	0	4	100.00	0.00
129	0	0	4	100.00	0.00
130	0	0	1	100.00	0.00
132	0	0	2	100.00	0.00
133	0	0	1	100.00	0.00
134	0	0	1	100.00	0.00
135	0	0	1	100.00	0.00
137	0	0	2	100.00	0.00
138	0	0	4	100.00	0.00
139	0	0	1	100.00	0.00
140	0	0	1	100.00	0.00
141	0	0	1	100.00	0.00
143	0	0	1	100.00	0.00
144	0	0	1	100.00	0.00
145	0	0	2	100.00	0.00
146	0	0	1	100.00	0.00
148	0	0	1	100.00	0.00
152	0	0	1	100.00	0.00
156	0	0	1	100.00	0.00
157	0	0	3	100.00	0.00
158	0	0	2	100.00	0.00
159	0	0	4	100.00	0.00
161	0	0	1	100.00	0.00
163	0	0	1	100.00	0.00
164	0	0	2	100.00	0.00
166	0	0	2	100.00	0.00
168	0	0	2	100.00	0.00

Table 1: Existence and Unique existence of 15-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
169	0	0	1	100.00	0.00
171	0	0	2	100.00	0.00
173	0	0	1	100.00	0.00
179	0	0	1	100.00	0.00
182	0	0	1	100.00	0.00
183	0	0	2	100.00	0.00
185	0	0	1	100.00	0.00
186	0	0	1	100.00	0.00
190	0	0	1	100.00	0.00
191	0	0	3	100.00	0.00
192	0	0	1	100.00	0.00
198	0	0	1	100.00	0.00
199	0	0	1	100.00	0.00
201	0	0	2	100.00	0.00
202	0	0	1	100.00	0.00
205	0	0	1	100.00	0.00
208	0	0	1	100.00	0.00
210	0	0	1	100.00	0.00
212	0	0	1	100.00	0.00
219	0	0	1	100.00	0.00
220	0	0	1	100.00	0.00
221	0	0	1	100.00	0.00
232	0	0	1	100.00	0.00
261	0	0	1	100.00	0.00
268	0	0	1	100.00	0.00

Table 2: Existence and Unique existence of 17-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
2	3652095	1099555	3952	23.20	23.12
3	668265	529021	4477	44.39	44.02
4	206144	194009	3977	48.99	48.01
5	60569	58288	3571	50.53	47.61
6	16525	14935	3104	52.19	43.21
7	4352	3505	2774	59.06	32.97
8	1419	872	2673	71.41	17.57
9	612	289	2537	82.20	8.41
10	335	111	2327	87.92	4.00
11	207	68	1983	90.83	3.01
12	137	50	1704	92.76	2.64
13	81	37	1319	94.36	2.57
14	65	11	1045	94.20	0.98
15	60	16	777	92.97	1.88
16	42	10	594	93.50	1.55

Table 2: Existence and Unique existence of 17-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
17	35	8	417	92.39	1.74
18	31	4	284	90.28	1.25
19	25	2	245	90.81	0.74
20	17	7	156	90.56	3.89
21	13	2	88	87.38	1.94
22	13	2	75	85.56	2.22
23	5	5	55	92.31	7.69
24	8	1	25	76.47	2.94
25	3	2	19	87.50	8.33
26	3	0	15	83.33	0.00
27	0	0	11	100.00	0.00
28	4	0	14	77.78	0.00
29	1	0	11	91.67	0.00
30	0	1	11	100.00	8.33
31	1	0	9	90.00	0.00
32	1	0	1	50.00	0.00
33	1	0	5	83.33	0.00
34	0	0	7	100.00	0.00
35	0	0	3	100.00	0.00
36	0	0	6	100.00	0.00
37	0	0	2	100.00	0.00
38	0	0	1	100.00	0.00
39	1	0	4	80.00	0.00
40	0	0	2	100.00	0.00
41	0	0	2	100.00	0.00
42	0	0	2	100.00	0.00
43	0	0	4	100.00	0.00
44	0	0	3	100.00	0.00
45	0	0	1	100.00	0.00
46	0	0	3	100.00	0.00
47	0	0	3	100.00	0.00
48	0	0	1	100.00	0.00
50	0	0	4	100.00	0.00
51	0	0	4	100.00	0.00
53	0	0	4	100.00	0.00
54	0	0	4	100.00	0.00
57	0	0	3	100.00	0.00
58	0	0	2	100.00	0.00
59	0	0	3	100.00	0.00
60	0	0	2	100.00	0.00
61	0	0	4	100.00	0.00
63	0	0	5	100.00	0.00
64	0	0	2	100.00	0.00
65	0	0	2	100.00	0.00
66	0	0	1	100.00	0.00

Table 2: Existence and Unique existence of 17-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
68	0	0	2	100.00	0.00
69	0	0	1	100.00	0.00
70	0	0	1	100.00	0.00
71	0	0	6	100.00	0.00
72	0	0	1	100.00	0.00
74	0	0	1	100.00	0.00
75	0	0	1	100.00	0.00
76	0	0	1	100.00	0.00
77	0	0	5	100.00	0.00
79	0	0	3	100.00	0.00
80	0	0	4	100.00	0.00
81	0	0	1	100.00	0.00
82	0	0	5	100.00	0.00
83	0	0	1	100.00	0.00
84	0	0	2	100.00	0.00
85	0	0	6	100.00	0.00
86	0	0	5	100.00	0.00
87	0	0	1	100.00	0.00
89	0	0	1	100.00	0.00
90	0	0	1	100.00	0.00
91	0	0	2	100.00	0.00
92	0	0	1	100.00	0.00
94	0	0	1	100.00	0.00
95	0	0	1	100.00	0.00
98	0	0	1	100.00	0.00
100	0	0	2	100.00	0.00
101	0	0	2	100.00	0.00
104	0	0	3	100.00	0.00
106	0	0	1	100.00	0.00
107	0	0	1	100.00	0.00
108	0	0	1	100.00	0.00
109	0	0	1	100.00	0.00
110	0	0	1	100.00	0.00
112	0	0	1	100.00	0.00
114	0	0	1	100.00	0.00
116	0	0	1	100.00	0.00
118	0	0	1	100.00	0.00
120	0	0	1	100.00	0.00
122	0	0	1	100.00	0.00
125	0	0	2	100.00	0.00
127	0	0	1	100.00	0.00
130	0	0	1	100.00	0.00
136	0	0	3	100.00	0.00
140	0	0	2	100.00	0.00
141	0	0	1	100.00	0.00

Table 2: Existence and Unique existence of 17-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
143	0	0	1	100.00	0.00
151	0	0	1	100.00	0.00
153	0	0	1	100.00	0.00
164	0	0	1	100.00	0.00
166	0	0	1	100.00	0.00

Table 3: Existence and Unique existence of 19-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
2	2089009	868637	4796	29.48	29.32
3	363543	307174	4804	46.18	45.47
4	88917	81730	4402	49.20	46.69
5	20364	17910	3813	51.61	42.55
6	4305	3883	3282	62.47	33.85
7	1150	732	2801	75.44	15.63
8	410	175	2182	85.18	6.32
9	197	59	1765	90.25	2.92
10	123	27	1218	91.01	1.97
11	57	15	897	94.12	1.55
12	42	15	718	94.58	1.94
13	23	8	490	95.59	1.54
14	20	13	352	94.81	3.38
15	17	5	268	94.14	1.72
16	9	4	176	95.24	2.12
17	4	2	122	96.88	1.56
18	2	8	82	97.83	8.70
19	3	0	51	94.44	0.00
20	6	1	43	88.00	2.00
21	3	0	24	88.89	0.00
22	2	0	20	90.91	0.00
23	1	0	10	90.91	0.00
24	0	0	6	100.00	0.00
25	1	0	5	83.33	0.00
26	0	0	5	100.00	0.00
27	1	0	2	66.67	0.00
28	1	0	3	75.00	0.00
29	0	0	4	100.00	0.00
30	0	0	4	100.00	0.00
31	0	0	6	100.00	0.00
32	1	0	1	50.00	0.00
33	1	0	3	75.00	0.00
34	0	0	1	100.00	0.00
35	0	0	1	100.00	0.00
36	0	0	1	100.00	0.00

Table 3: Existence and Unique existence of 19-mers

GroupID	State A	State B	State C	Existence (%)	Unique existence (%)
37	0	0	3	100.00	0.00
38	0	0	5	100.00	0.00
40	0	0	6	100.00	0.00
41	0	0	1	100.00	0.00
42	0	0	3	100.00	0.00
43	0	0	3	100.00	0.00
44	0	0	5	100.00	0.00
45	0	0	5	100.00	0.00
46	0	0	4	100.00	0.00
47	0	0	5	100.00	0.00
48	0	0	2	100.00	0.00
49	0	0	2	100.00	0.00
50	0	0	4	100.00	0.00
51	0	0	3	100.00	0.00
52	0	0	3	100.00	0.00
53	0	0	3	100.00	0.00
54	0	0	2	100.00	0.00
56	0	0	2	100.00	0.00
57	0	0	3	100.00	0.00
58	0	0	1	100.00	0.00
59	0	0	4	100.00	0.00
60	0	0	3	100.00	0.00
61	0	0	3	100.00	0.00
62	0	0	2	100.00	0.00
63	0	0	3	100.00	0.00
64	0	0	2	100.00	0.00
66	0	0	2	100.00	0.00
67	0	0	2	100.00	0.00
76	0	0	1	100.00	0.00
77	0	0	2	100.00	0.00
78	0	0	1	100.00	0.00
79	0	0	2	100.00	0.00
80	0	0	2	100.00	0.00
81	0	0	1	100.00	0.00
86	0	0	1	100.00	0.00
87	0	0	1	100.00	0.00
88	0	0	1	100.00	0.00
89	0	0	1	100.00	0.00
92	0	0	2	100.00	0.00
93	0	0	4	100.00	0.00
103	0	0	1	100.00	0.00
105	0	0	1	100.00	0.00
106	0	0	2	100.00	0.00
118	0	0	1	100.00	0.00