De Bruijn Graph Assembler

Kevin Boehme Shaun Miller Ken Reese

October 16, 2014

1 Methodology

Our error handling approach was quite simple. Once the reads were read in from the FASTA format and our algorithm broke them into the specified kmer length, we traversed the list of kmers looking for kmers that only appeared once. This effectively checked the support for each kmer in our list. If a kmer was found only once it would suggest a genotyping error and as such would be thrown out. If on the other hand, the kmer was supported multiple times, this suggests that it was correctly genotyped and was subsequently kept. Our algorithm currently does not attempt to bridge non-branching nodes.

2 Quality Analysis

In order to analyze the quality of our assembly we relied on a few, informative metrics including average contig size, N50, number of contigs, and the maximum contig size. These metrics a provided a solid foundation for determining the quality of our assembly as we adjusted the kmer length as well as providing a tractable means of comparing our assembly to other De Bruijn graph assemblers (such as Velvet).

Real Dataset, Small, klen = 42

Mean contig size:	382
N50:	1015
Number of Contigs:	3
Largest Contig Size:	1015

Real Dataset, Large, klen = 57

Mean contig size:	149.513
N50:	232
Number of Contigs:	251
Largest Contig Size:	1558

3 Comparison

4 BLAST Results

5 Assembler Improvements

Appendix A — Charts and Tables

Appendix B — Notes

Kmer Size	Number of Contigs	N50	Max Length
10	15122	2	36
11	7510	10	91
12	7510	10	91
13	2669	31	235
14	2669	31	235
15	1479	63	732
16	1479	63	732
17	1137	76	1112
18	1137	76	1112
19	1021	84	1275
20	1021	84	1275
21	872	103	1275
22	872	103	1275
23	776	112	884
24	776	112	884
25	645	135	1279
26	645	135	1279
27	521	239	1474
28	521	239	1474
29	456	237	1474
30	456	237	1474
31	347	374	**1477**

Kmer Size	Number of Contigs	N50	Max Length
10	7170	11	73
11	3879	15	136
12	2285	23	211
13	1554	31	432
14	1227	41	863
15	1068	47	878
16	962	54	1326
17	898	61	1327
18	835	71	1328
19	789	82	1329
20	739	93	1330
21	682	103	1331
22	645	111	1332
23	637	111	1333
24	619	114	1366
25	592	115	1368
26	568	119	1370
27	525	144	1372
28	504	148	1374
29	483	166	1376
30	455	177	1378
31	424	190	1380