Pell, Jason; **Hintze, Arend**; **Canino-Koning, Rosangela**; **Howe, Adina**; **Tiedje, James M.**; **Brown, C. Titus**

Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. (English)

Zbl 1256.68052
Proc. Natl. Acad. Sci. USA 109, No. 33, 13272-13277 (2012).

Summary: Deep sequencing has enabled the investigation of a wide range of environmental microbial ecosystems, but the high memory requirements for de novo assembly of short-read shotgun sequencing data from these complex populations are an increasingly large practical barrier. Here we introduce a memory-efficient graph representation with which we can analyze the k-mer connectivity of metagenomic samples. The graph representation is based on a probabilistic data structure, a Bloom filter, that allows us to efficiently store assembly graphs in as little as 4 bits per k-mer, albeit inexactly. We show that this data structure accurately represents DNA assembly graphs in low memory. We apply this data structure to the problem of partitioning assembly graphs into components as a prelude to assembly, and show that this reduces the overall memory requirements for de novo assembly of metagenomes. On one soil metagenome assembly, this approach achieves a nearly 40-fold decrease in the maximum memory requirements for assembly. This probabilistic graph representation is a significant theoretical advance in storing assembly graphs and also yields immediate leverage on metagenomic assembly.

MSC:

68P05 Data structures
05C80 Random graphs (graph-theoretic aspects)
05C90 Applications of graph theory
92-08 Computational methods for problems pertaining to biology
92D10 Genetics and epigenetics

Keywords:

deep sequencing; complex populations; memory-efficient graph representation; k-mer connectivity

Software:

ALLPATHS; GAGE

Full Text: DOI

References:

[1] Briefings in Bioinformatics 10 (4) pp 354– (2009) · doi:10.1093/bib/bbp026
[2] Genome Research 22 (3) pp 557– (2012) · doi:10.1101/gr.131383.111
[3] Qin, Nature; Physical Science (London) 464 (7285) pp 59– (2010) · doi:10.1038/nature08821
[4] Hess, Science 331 (6016) pp 463– (2011) · doi:10.1126/science.1200387
[5] Wooley 6 (2) pp e1000667– (2010) · doi:10.1371/journal.pcbi.1000667
[6] Gans, Science 309 (5739) pp 1387– (2005) · doi:10.1126/science.1112665
[7] Science 304 (5667) pp 66– (2004) · doi:10.1126/science.1093857
[8] Mackelprang, Nature; Physical Science (London) 480 (7377) pp 368– (2011) · doi:10.1038/nature10576
[9] Pevzner, PNAS 98 (17) pp 9748– (2001) · Zbl 0993.92018 · doi:10.1073/pnas.171285098
[10] Miller, Genomics 95 (6) pp 315– (2010) · doi:10.1016/j.ygeno.2010.03.001
[11] Compeau, Nature biotechnology 29 (11) pp 987– (2011) · doi:10.1038/nbt.2023
[12] Bioinformatics 27 (4) pp 479– (2011) · Zbl 05891125 · doi:10.1093/bioinformatics/btq697
[13] PNAS 108 (4) pp 1513– (2011) · doi:10.1073/pnas.1017351108
[14] Kelley, Genome biology 11 (11) pp R116– (2010) · doi:10.1186/gb-2010-11-4-116

[15] CACM 13 pp 422– (1970) · Zbl 0195.47003 · doi:10.1145/362686.362692
[16] Shi, Journal of computational biology : a journal of computational molecular cell biology 17 (4) pp 603– (2010) · doi:10.1089/cmb.2009.0062
[17] Bioinformatics 26 (13) pp 1595– (2010) · Zbl 1183.68146 · doi:10.1093/bioinformatics/btq230
[18] Melsted, BMC bioinformatics [electronic resource] 12 pp 333– (2011) · doi:10.1186/1471-2105-12-333
[19] Liu, BMC bioinformatics [electronic resource] 12 pp 85– (2011) · Zbl 05889789 · doi:10.1186/1471-2105-12-85
[20] Genome Research 18 (5) pp 821– (2008) · doi:10.1101/gr.074492.107
[21] Genome Research 19 (6) pp 1117– (2009) · doi:10.1101/gr.089532.108
[22] Bioinformatics 27 (13) pp i94– (2011) · Zbl 1263.92015 · doi:10.1093/bioinformatics/btr216
[23] Grabherr, Nature biotechnology 29 (7) pp 644– (2011) · doi:10.1038/nbt.1883
[24] PHYS REP 54 pp 1– (1979) · doi:10.1016/0370-1573(79)90060-7
[25] Gilbert 3 (3) pp 243– (2010) · doi:10.4056/sigs.1433550
[26] Gilbert 3 (3) pp 249– (2010) · doi:10.4056/aigs.1443528
[27] ZHANG, Cold Spring Harbor Symposia on Quantitative Biology 68 (0) pp 205– (2003) · doi:10.1101/sqb.2003.68.205
[28] Price, Bioinformatics 21 (suppl_1) pp i351– (2005) · doi:10.1093/bioinformatics/bti1018
[29] Iqbal, Nature genetics 44 (2) pp 226– (2012) · doi:10.1038/ng.1028
[30] 1 pp 485– (2004) · Zbl 1090.68515 · doi:10.1080/15427951.2004.10129096
[31] PHYS REV E 66 pp 011907– (2002) · doi:10.1103/PhysRevE.66.011907
[32] TRANS AM MATH SOC 54 pp 426– (1943) · doi:10.1090/S0002-9947-1943-0012401-3