Succinct Colored de Bruijn Graphs Supplement

August 28, 2017

Contents

1 Validation on Six E. coli Genomes

In order to validate our data structure and test the accuracy of the bubble-calling method of Vari during development, we compared the bubbles found by running the bubble calling algorithm on the *E. coli* reference dataset using Cortex and Vari. The bubbles outputted by each method were compared by identifying the flank preceding each bubble. Both Vari and Cortex identified 465 bubbles across all six *E. coli* K-12 substrains. This number accounts for the reverse complement bubbles found by Vari. The methods agree on 98.5% (458 / 465) of the bubbles. Thus, Vari found seven bubbles that were not identified by Cortex, which were shown to be valid, and Cortex found seven bubbles not identified by Vari. Nonetheless, our validation shows that 98.5% of the variation determined by Cortex and Vari is identical.

Accession Number	Sub-strain	Genome Size
AP009048	W3110	4,646,332 bp
CP009789	ER3413	4,558,660 bp
CP010441	ER3445	4,607,634 bp
CP010442	ER3466	4,660,432 bp
CP010445	ER3435	4,682,086 bp
U00096	MG1655	4,641,652 bp

Table 1: Characteristics of the substrains of *E. coli* K-12 used to test the performance and accuracy of Vari

.

2 Validation on AMR Genes and Simulated Sample

An additional experiment used during development comprised a set of 54 antimicrobial resistance (AMR) genes and a simulated metagenomics sample containing seven of these 54 AMR genes and four AMR genes not contained in this set,

We first compiled a database of known AMR genes based on sequences in the databases CARD [McArthur *et al.*(2013)McArthur *et al.*], Resfinder [Zankari *et al.*(2012)Zankari *et al.*] and ARG-ANNOT [Gupta *et al.*(2014)Gupta *et al.*] — each of these AMR-specific databases are actively curated and contain the genetic sequences for a large variety of AMR genes. This database contains all known AMR genes, their drug resistance, and mechanism conferring resistance. We selected 54 beta-lactamase genes from this database that are known to have very high clinical and public health importance, and simulated 26,516,559 paired-end 120 bp reads from seven of the 54 beta-lactamase genes (Accession numbers AFQ67211, CAJ19612, AEX08599, CAC33434, CAB82578, ADM26831, AAK63223, and AFI61435), as well as four additional AMR genes that were not included in this set of 54 genes (Accession numbers AAA27471, ADD83116, NP_387454, and AAB24797). These latter four genes were tetracycline-resistant genes. Tetracyclines are a group of broad-spectrum antibiotics and hence, their resistance is also clinically important. This AMR dataset was used not only in the memory and time performance but also used to test the ability of Vari in identifying beta-lactamase genes from a typical metagenomic sample containing a variety of AMR genes.

We validated the ability of Vari to correctly identify the AMR genes contained in a metagenomics sample using a set of reference genes. Vari constructed the colored de Bruijn graph from the set of 54 beta lactamases and the simulated metagenomics sample. Hence, there were 55 unique colors in the graph because there exists one color for the metagenomic sample and one unique color for each of the 54 beta-lactamase genes. Next, for each of the 54 genes, the unique k-mers were identified and the total number of these k-mers that were contained in the simulated sample was determined.

The shared k-mer fraction for each of the 54 genes ranged from 0.41 to 1 with a mean of 0.62. All of the seven beta-lactamase genes that were contained in the simulated sample had a shared k-mer fraction of 1, whereas none of the remaining 47 genes did. Of the 47 beta-lactamase genes that were not contained in the simulated sample, two had a shared k-mer fraction 0.98 and 0.95, however, these genes had 97% and 95% sequence similarity to one of the seven genes contained in the sample. All the remaining 45 genes had a shared k-mer fraction between 0.79 and 0.41. Hence, this demonstrates (on a small scale) that this use of the colored de Bruijn graph and our match color algorithm is a viable method to identify AMR genes in a metagenomics sample.

For metagenomic experiments, we count the number of k-mers in common between the metagenomic sample and each of the AMR genes. If each AMR gene is available as a separate color, the set operations can be done with a linear traversal of the intermediate file containing the uncompressed C matrix. In this case, loading and traversing the colored de Bruijn graph is not necessary. We refer to this matrix scan algorithm as *match color*.

We used this experiment for validation purposes, choosing to focus on beta-lactamase genes since they have a critical role in public health. In particular, our experiment on the simulated AMR dataset validates Vari's ability to correctly identify AMR genes from a metagenomics sample, which is of paramount importance as regulatory agencies increasingly look to metagenomics as a method to monitor the spread and evolution of AMR across different environments FAOActionPlan2016. Our simulation based experiment and results focus on beta-lactamases, which include genes that confer

			Cortex		Vari	
Dataset	No. of k-mers	Colors	Memory	Time	Memory	Time
AMR genes and sample	9,348,365	55	7.08 GB	2m 55s	0.718 GB	29m 21s

Table 2: Comparison between the peak memory and time usage required to store all the k-mers and run bubble calling on the data in Cortex and Vari. The peak memory is given in megabytes (MB) or gigabytes (GB). The running time is reported in seconds (s), minutes (m), and hours (h).

resistance to a class of antibiotics that are considered to be the last resort for infections from multi-drug-resistant bacteria [McKenna(2013)McKenna, Queenan and Bush(2007)Queenan and Bush]. This experiment found that all beta-lactamases were correctly identified and only two of the remaining 47 genes were identified to be in the sample, which had 97% and 95% sequence similarity to one of the beta-lactamases in the sample.

AMR Gene	**Resistance Type**	**Accession Number**
AmpH	beta-lactamase	AFQ67211
OKP-B-4	beta-lactamase	CAJ19612
NDM-6	beta-lactamase	AEX08599
MAL-1	beta-lactamase	CAC33434
MOX-2	beta-lactamase	CAB82578
TLA-1	beta-lactamase	ADM26831
SED-1	beta-lactamase	AAK63223
TEM-1	beta-lactamase	AFI61435
TET-X	Tetracycline	AAA27471
TET-X(1)	Tetracycline	ADD83116
TET-C	Tetracycline	NP_387454
TETR-G	Tetracycline	AAB24797

Table 3: List of AMR genes used to generate the simulated sample. The first seven genes were included in the the 54 beta-lactamase genes we considered for this experiment, and the remaining four were tetracycline genes. Each of the genes were approximately 1,000 bp in length and had varied GC content.

Table 4 shows the results of the validation on AMR dataset. The set of beta lactamase genes is listed as well as the number of k-mers in these genes. For each gene, we also count the number k-mers shared with the sample and then divide the shared k-mer count by the k-mer count to calculate the fraction. The first seven genes are in the sample so as expected their fraction is 1.

3 Command lines for experiments

```
# 6 E. coli experiment
mkdir -p tmp
find . -name '*.fna' |xargs -l -i   kmc -ci0 -fm -k32 -cs300 {} {}_kmc tmp
find . -name '*.fna'|xargs -l -i  kmc_tools sort {}_kmc {}_kmc_sorted_kmc.kmc
find . -name '*.fna' |xargs -l -i echo "{}_kmc_sorted_kmc.kmc" >ecoli6_kmc2_list
cosmo-build -d ecoli6_kmc2_list >cosmo-build.stdout 2>&1 &
pack-color ecoli6_kmc2_list.colors 6 55539132 54489174 >pack-color.stdout 2>&1 &
```

```
cosmo-color -a 1 -b 2 ecoli6_kmc2_list.dbg ecoli6_kmc2_list.colors.sd_vector \
    >bubbles.stdout 2>&1 &


# count k-mers
mkdir -p kmc_temp
kmc -k32 -ci0 -t8 -fm ./GCF_000005005.1_B73_RefGen_v3_genomic.fna ./corn.kmc kmc_tmp # corn
kmc -k32 -ci0 -t8 -fm rice.fa ./rice.kmc kmc_tmp
# 'names' is a file listing arabidopsis fasta files
kmc -k32 -ci0 -t8 -fm @names ./arabidopsis.kmc kmc_tmp
# 'names' is a file listing tomato fasta files
kmc -k32 -ci0 -t8 -fm @names ./tomato.kmc kmc_tmp

# sort counted k-mers
kmc_tools sort corn.kmc -ci0 corn.sorted.kmc kmc_tmp
kmc_tools sort rice.kmc -ci0 rice.sorted.kmc kmc_tmp
kmc_tools sort tomato.kmc -ci0 tomato.sorted.kmc kmc_tmp
kmc_tools sort arabidopsis.kmc -ci0 arabidopsis.sorted.kmc kmc_tmp

# build sdbg and uncompressed color matrix
cosmo-build -d assemblies # assemblies is a file listing the sorted k-mer counts

# compress color matrix
# args: uncompressed-matrix, numcolors, totbits, setbits
pack-color assemblies.colors 4 13703494400 3420124654

# call bubbles
cosmo-color assemblies.dbg assemblies.colors.sd_vector


# 4k E. coli dataset
# This dataset can be acquired with:
# https://github.com/cosmo-team/cosmo/blob/VARI/experiments/fetch_ecoli.py
# Otherwise, download instructions are in: https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/
# Select all assemblies where organism_name is 'Escherichia coli'
# from ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/assembly_summary.txt
# k-mer size is set on the call to kmc, commands are the same for other runs
# but with -k48 or -k64 instead of -k32
# count k-mers for each file like so:
mkdir kmc_tmp
kmc -k32 -ci0 -t8 -fm ../assemblies/GCF_000005845.2_ASM584v2_genomic.fna \
  ./GCF_000005845.2_ASM584v2_genomic.fna.kmc kmc_tmp

# sort each counted set of k-mers
kmc_tools sort ../k32/GCF_000005845.2_ASM584v2_genomic.fna.kmc -ci0 \
  GCF_000005845.2_ASM584v2_genomic.fna.kmc kmc_tmp
```

```
# build sdbg and uncompressed color matrix
cosmo-build -d kmc_files

# compress color matrix
# args: uncompressed-matrix, numcolors, totbits, setbits
pack-color kmc_files.colors 3765 1198032856770 37993299892

# call bubbles where one branch is specified with bit vector 0b1 (1 in decimal) and
# the other with 0b10 (2 in decimal)
cosmo-color -a 1 -b 2 kmc_files.dbg kmc_files.sd_vector


# Beef safety

# trim
java -jar  /home/lakinsm/bin/trimmomatic-0.36.jar PE -threads 50 \
  /home/lakinsm/hmm_testing/ncbaI/${1}_R1_001.fastq.gz \
  /home/lakinsm/hmm_testing/ncbaI/${1}_R2_001.fastq.gz ${1}_forward_paired.fastq \
  ${1}_forward_unpaired.fastq ${1}_reverse_paired.fastq ${1}_reverse_unpaired.fastq \
  ILLUMINACLIP:/home/lakinsm/amr_tools/trimmomatic/adapters/TruSeq3-PE.fa:2:30:10 \
  LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

# count each of the four files per sample emitted by the trimmer
mkdir kmc_tmp
ls --color=no -1 trimmed/ |xargs -l -i ../kmc -ci0 -fq -k32 -cs65535 trimmed/{} counts/{} \
  kmc_tmp >kmc_count.log 2>&1 &

# form the union of each set of those four files
ls --color=no -1 ../union_defs |xargs -l -i ../../kmc_tools complex ../union_defs/{} \
  >../union.log 2>&1 &

cosmo-build -d kmc_files_orderby_location_then_matrix
pack-color kmc_files_orderby_location_then_matrix.colors 88 8420579985928 214918276263

cosmo-color-pd -r ../vari/tet_then_other_noN_ref_genes -b 1 -a 0 \
  kmc_files_orderby_location_then_matrix.dbg \
   kmc_files_orderby_location_then_matrix.colors.sd_vector
```

4 Data Structure Construction Statistics

	Cortex		KMC2		Vari-dBG			Vari-C	
Dataset	**CPU time**	**Mem.**	**CPU time**	**Mem.**	**CPU time**	**Int. Mem.**	**Ext. Mem.**	**CPU time**	**Mem.**
Plants	2h 25m 27s	109,579	19m 50s	4,335	1h 34m 37s	5,388	156,504	3m 09s	3,528
E. coli (k=32)	N/A	N/A	3h 15m 40s	104	9h 30m 11s	126,777	319,328	53m 54s	42,043
E. coli (k=48)	N/A	N/A	4h 35m 29s	149	10h 47m 46s	128,077	427,460	1h 02m 07s	42,100
E. coli (k=64)	N/A	N/A	5h 05m 27s	189	11h 21m 08s	127,523	522,576	1h 09m 07s	42,134
Beef safety	N/A	N/A	34h 04m 46s	11,688	82h 42m 48s	109,091	4,378,840	6h 44m 12s	217,705

Table 5: Data structure construction performance measurements. CPU time is user plus system time as reported by '/bin/time'. (Internal) memory is reported in megabytes and is the maximum resident set size. KMC2 includes both counting and sorting k-mers. Vari-dBG forms the k-mer union and builds the succinct de Bruijn graph. Vari-C compresses the color matrix.

References

[Gupta *et al.*(2014)Gupta *et al.*] Gupta, S. *et al.* (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial Agents and Chemotherapy*, **58**(1), 212–20.

[McArthur *et al.*(2013)McArthur *et al.*] McArthur, A. G. *et al.* (2013). The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, **57**, 3348–3357.

[McKenna(2013)McKenna] McKenna, M. (2013). Antibiotic resistance: The last resort. *Nature*, **499**, 394–396.

[Queenan and Bush(2007)Queenan and Bush] Queenan, A. M. and Bush, K. (2007). Carbapenemases: the versatile beta-lactamases. *Clinical Microbiology Reviews*, **7**(3), 440–458.

[Zankari *et al.*(2012)Zankari *et al.*] Zankari, E. *et al.* (2012). Identification of acquired antimicrobial resistance genes. *Antimicrobial Agents and Chemotherapy*, **67**(11), 2640–2644.

Gene Name	Total No. of k-mers	No. of Shared k-mers	Shared k-mer Fraction
AmpH	5,021	5,021	1
OKP-B-4	4,407	4,407	1
NDM-6	4,327	4,327	1
MAL-1	4,507	4,507	1
MOX-2	4,987	4,987	1
TLA-1	4,513	4,513	1
SED-1	4,475	4,475	1
TEM-1	4,321	4,321	1
TLA-1-1	4,593	4,513	0.982582
NDM-5	4,323	4,137	0.956974
BLA-I	3,463	2,763	0.797863
OKP-A-1	4,413	3,299	0.747564
Mbl	4,093	2,763	0.675055
IND-1	4,139	2,763	0.667553
CGB-1	4,157	2,763	0.664662
GIM-1	4,207	2,763	0.656763
LEN-1	4,377	2,853	0.651816
LCR-1	4,257	2,763	0.649049
Nps1	4,261	2,763	0.648439
VIM-1	4,301	2,763	0.642409
OXA-1	4,363	2,763	0.63328
MOX-1	4,987	3,153	0.632244
Z32	4,391	2,763	0.629242
SHV-1	4,423	2,783	0.629211
BEL-1	4,403	2,763	0.627527
CARB-1	4,437	2,763	0.622718
OXY-1-5	4,441	2,763	0.622157
IMI-1	4,445	2,763	0.621597
CTX-M-1	4,449	2,763	0.621038
NMC-A	4,457	2,763	0.619924
BES-1	4,459	2,763	0.619646
KPC-1	4,463	2,763	0.61909
SME-1	4,469	2,763	0.618259
CME-1	4,477	2,763	0.617154
Lut-1	4,481	2,763	0.616603
FAR-1	4,489	2,763	0.615505
VEB-1	4,499	2,763	0.614136
AIM-1	4,525	2,763	0.610608
AER-1	4,531	2,763	0.609799
ROB-1	4,535	2,763	0.609261
SFC-1-1	4,559	2,763	0.606054
cfxA	4,635	2,763	0.596116
cphA1	4,663	2,763	0.592537
CMG	4,795	2,763	0.576225
ACT-1	4,977	2,763	0.555154
MIR-1	4,977	2,763	0.555154
MOR	4,979	2,763	0.554931
Amp	4,993	2,763	0.553375
FOX-1	4,999	2,763	0.552711
PAO-1	5,089	2,763	0.542936
PENA	6,137	2,763	0.45022
PBP	6,505	2,763	0.42475
lmrD	6,675	2,763	0.413933
MECA	6,713	2,763	0.411589

Table 4: AMR gene name, number of k-mers in the colored de Bruijn graph, and number and proportion of k-mers identified in both the beta lactamase database and the simulated metagenomic sample