Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw586
Advance Access Publication Date: 7 September 2016
Application Note

OXFORD

Genome analysis

stringMLST: a fast k-mer based tool for multilocus sequence typing

Anuj Gupta[1,2], I. King Jordan[1,2,3] and Lavanya Rishishwar[1,2,3,*]

[1]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA, [2]Applied Bioinformatics Laboratory, Atlanta, GA 30332, USA and [3]PanAmerican Bioinformatics Institute, Cali, Valle del Cauca 760043, Colombia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 22, 2016; revised on August 17, 2016; accepted on September 5, 2016

Abstract

Rapid and accurate identification of the sequence type (ST) of bacterial pathogens is critical for epidemiological surveillance and outbreak control. Cheaper and faster next-generation sequencing (NGS) technologies have taken preference over the traditional method of amplicon sequencing for multilocus sequence typing (MLST). But data generated by NGS platforms necessitate quality control, genome assembly and sequence similarity searching before an isolate's ST can be determined. These are computationally intensive and time consuming steps, which are not ideally suited for real-time molecular epidemiology. Here, we present stringMLST, an assembly- and alignment-free, lightweight, platform-independent program capable of rapidly typing bacterial isolates directly from raw sequence reads. The program implements a simple hash table data structure to find exact matches between short sequence strings (k-mers) and an MLST allele library. We show that stringMLST is more accurate, and order of magnitude faster, than its contemporary genome-based ST detection tools.

Availability and Implementation: The source code and documentations are available at http://jordan.biology.gatech.edu/page/software/stringMLST.

Contact: lavanya.rishishwar@gatech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Sequence typing of bacterial pathogens is essential to molecular epidemiology. The original gene-based methods for typing bacterial isolates, such as multilocus sequence typing (MLST) (Maiden *et al.*, 1998), relied on Sanger sequencing of amplicons from a small set of loci. With the advances in next-generation sequencing (NGS) technology, it has become faster, cheaper and more useful to sequence whole bacterial genomes for typing purposes, rather than a handful of individual amplicons (Hyytia-Trees *et al.*, 2007; Jackson *et al.*, 2015; Maiden *et al.*, 2013). Whole genome-based typing methods, such as core genome MLST (cgMLST) (Jolley *et al.*, 2012), whole genome MLST (wgMLST) (Cody *et al.*, 2013) and average nucleotide identity (ANI) (Goris *et al.*, 2007), can allow for even finer typing resolution than MLST. Nevertheless, researchers and epidemiologist continue to rely on traditional MLST schemes due to the presence of legacy sequence type (ST) information that has accumulated from numerous surveys over the years, and determining the ST often remains the first pass in the analysis of bacterial isolates in the NGS era (Desoubeaux *et al.*, 2016; Katz *et al.*, 2009).

Contemporary methods that analyze whole genome sequence data to perform gene-based typing of bacterial isolates are computation- and time-intensive. Here, we describe stringMLST, a k-mer based method for the rapid gene based characterization of bacterial isolates directly from genome sequence reads. stringMLST has the advantages of being assembly- and alignment-free as well as having a small memory footprint, a minimalist code base and straightforward installation. It can be used on existing MLST schemes or on user-designed custom typing schemes, including larger-scale schemes that use scores (rMLST) or hundreds (cgMLST) of loci. We ran stringMLST on a large dataset of bacterial genome sequence reads

Table 1. Comparative performance comparison and accuracy testing

Comparative test

Tool name	Type[a]	Input	% Correct		Run time[b]
			Alleles	STs	
stringMLST	K-mer	Reads	100.0	100.0	45
CGE/MLST	BLAST	Reads	99.6	97.5	2922
SRST2	Mapping	Reads	98.6	92.5	1887
SRST	BLAST	Assembly	95.0	77.5	2386
Offline CGE	BLAST	Assembly	96.1	80.0	170

Accuracy test (stringMLST; $k = 35$)

#Isolates[c]	#Alleles[d]	#Correctly predicted		Run time[b]	Mem[e]
		STs	Alleles		
1002	7014	1000	7012	40.7	0.67

Larger-scale schemes (stringMLST versus BLAST)

#Isolates[c]	#Alleles[d]	#Correctly predicted		RTR[f]	Sch[g]
		Alleles	%		
20	1060	1009	95.2	516.7	rMLST
20	31 919	28 976	90.8	43.0	cgMLST

[a]Algorithmic paradigm implemented by the tool.
[b]Average runtime per sample (in seconds).
[c]Total number of isolates tested.
[d]Total number alleles tested.
[e]Peak memory usage (in GB).
[f]Run time rate or the rate of processing sequence read files as kb/s.
[g]Typing scheme.

with known ST information to validate its accuracy and performance.

2 Materials and methods

2.1 Algorithm overview

stringMLST relies on exact pattern matching using sequence substrings or k-mers, short DNA sequences of length k (see Supplementary Information). Isolates are characterized by finding the specific allele for each locus in the typing scheme that shows the maximum k-mer hits, based on a k-mer to loci relationship database. This austere algorithmic design allows stringMLST to rapidly process sequence read files with a small memory footprint.

2.2 Database

Database construction requires a profile definition file for the typing scheme along with allele sequences for each locus in the scheme; this file can be created by the user or retrieved them from the PubMLST database (Jolley and Maiden, 2010). stringMLST k-merizes each locus-specific allele sequence and records the corresponding allele and loci for each k-mer (Supplementary Fig. S1 and S2).

2.3 ST discovery

The process of ST discovery can conceptually be broken down into three stages—filtering, counting and reporting. In the filtering stage, stringMLST discards a sequence read if the k-mer situated at the middle of the sequence read does not have a match in the stringMLST database. Sequence reads whose middle k-mers have a match are k-merized in the counting stage. Each k-mer is then searched in the database and for each match, the corresponding allele and loci are recorded; a counter is incremented for each allele whose constituent k-mer was matched. Once all the sequences have been processed, stringMLST identifies the allele at each locus with the maximum counter value to generate an allelic profile and corresponding ST call.

2.4 Implementation

stringMLST is implemented in Python and is designed to be platform-independent and lightweight.

3 Performance evaluation

The accuracy and runtime of stringMLST were evaluated in three ways: (i) a comparative performance test against existing genome-based MLST detection tools, (ii) an accuracy test using a set of samples with known ST information and (iii) a test of its utility for larger-scale typing schemes. A total of 1042 samples from four species were obtained from the PubMLST/EBI ENA database for these tests (Supplementary Table S1 and Supplementary Information). For the comparative test, 10 samples each from 4 species (*Neisseria meningitidis*, *Streptococcus pneumoniae*, *Campylobacter jejuni* and *Chlamydia trachomatis*) were tested against 4 commonly used genome sequence read based ST determination tools: CGE/MLST (Larsen *et al.*, 2012), SRST2 (Inouye *et al.*, 2014), SRST (Inouye *et al.*, 2012) and an offline implementation of the CGE/MLST by Pritchard, L. (referred to here as offline CGE; https://github.com/widdowquinn/scripts/tree/master/bioinformatics). stringMLST correctly predicted the allelic profile and ST of all 40 samples tested with an average runtime of 45 s per sample (Table 1). The next best performing tool was the web-based CGE/MLST, which incorrectly predicted a single allele but took considerably longer time per sample (2922 s or ~50 min). A large part of CGE/MLST's runtime is the time spent in uploading the sequence to the server. The online-only nature of the tool also makes it hard to incorporate it in large-scale data analysis pipelines. SRST2 was able to correctly identify 276/280 alleles (three incorrect ST predictions) followed by offline CGE (eight incorrect STs) and SRST (nine incorrect STs). stringMLST had the shortest runtime of 45 s which was 3.7-fold faster than the next fastest tool (offline CGE). However, offline CGE requires assembled sequence reads which adds substantial additional overall runtime for ST determination.

For the large-scale accuracy test, stringMLST was run on all 1002 *N. meningitidis* samples available on the PubMLST/EBI ENA database (October 15, 2015) with known ST information (Supplementary Table S2). The program was run for a range of different k-mer values ($K = 15$, 21, 31, 35, 45, 55 and 66). Of the 1002 samples tested, stringMLST correctly predicted 1000 samples (99.8% accuracy). Eleven samples were initially reported as incorrectly predicted, but manual inspection revealed that 10 of these samples were actually mis-annotated in the PubMLST database (Supplementary Table S3). Out of the 10 mis-annotated samples, stringMLST detected 9 of them correctly. For the two incorrectly predicted samples, stringMLST failed to correctly predict one allele each in each case, meaning that the correct clonal complex was still identified for both samples. The average runtime of stringMLST was nearly a minute or less on all the samples with an average memory consumption of less than 1 GB (Table 1 and Supplementary Table S2).

The utility of stringMLST for larger-scale typing schemes was evaluated using ribosomal MLST (rMLST) on 53 loci and core

genome MLST (cgMLST) on 1605 loci (see Supplementary Information). stringMLST was able to correctly predict 95.2% of alleles for the rMLST scheme and 90.8% of alleles for cgMLST (Table 1). Use of stringMSLT for rMLST and cgMLST also resulted in accurate reconstruction of the phylogenetic relationships among known STs and comparable performance to whole genome phylogenetic analysis with ANI (Supplementary Fig. S6). stringMLST's fast and reliable performance, together with its simple underlying algorithm and platform-independence, make it a suitable tool for genome-based bacterial typing on machines of any size.

Funding

IHRC-Georgia Tech Applied Bioinformatics Laboratory to A.G., I.K.J. and L.R.

Conflict of Interest: none declared.

References

Cody,A.J. *et al*. (2013) Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. *J. Clin. Microbiol*., **51**, 2526–2534.

Desoubeaux,G. *et al*. (2016) Epidemiological outbreaks of *Pneumocystis jirovecii* pneumonia are not limited to kidney transplant recipients: genotyping confirms common source of transmission in a liver transplantation unit. *J. Clin. Microbiol*., **54**, 1314–1320

Jackson,B.R. *et al*. (2015) Notes from the field: listeriosis associated with stone fruit—United States, 2014. *MMWR Morb. Mortal. Wkly. Rep*., **64**, 282–283.

Goris,J. *et al*. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol*., **57**(**Pt 1**), 81–91.

Hyytia-Trees,E.K. *et al*. (2007) Recent developments and future prospects in subtyping of foodborne bacterial pathogens. *Future Microbiol*., **2**, 175–185.

Inouye,M. *et al*. (2012) Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC Genomics*, **13**, 338.

Inouye,M. *et al*. (2014) SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med*., **6**, 90.

Jolley,K.A. *et al*. (2012) Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid web-based analysis methods. *J. Clin. Microbiol*., **50**, 3046–3053.

Jolley,K.A., and Maiden,M.C. (2010) BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, **11**, 595.

Katz,L.S. *et al*. (2009) Meningococcus genome informatics platform: a system for analyzing multilocus sequence typing data. *Nucleic Acids Res*., **37**, W606–W611.

Larsen,M.V. *et al*. (2012) Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol*., **50**, 1355–1361.

Maiden,M.C. *et al*. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA*., **95**, 3140–3145.

Maiden,M.C. *et al*. (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol*., **11**, 728–736.