BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 3 2015, pages 426–428
doi:10.1093/bioinformatics/btu656

Sequence analysis

Advance Access publication October 4, 2014

CompMap: a reference-based compression program to speed up read mapping to related reference sequences

Zexuan Zhu[1], Linsen Li[1], Yongpeng Zhang[1], Yanli Yang[1] and Xiao Yang[2,*]

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China and [2]Infectious Disease Initiative, The Broad Institute, Cambridge, MA 02142, USA

Associate Editor: John Hancock

ABSTRACT

Summary: Exhaustive mapping of next-generation sequencing data to a set of relevant reference sequences becomes an important task in pathogen discovery and metagenomic classification. However, the runtime and memory usage increase as the number of reference sequences and the repeat content among these sequences increase. In many applications, read mapping time dominates the entire application. We developed CompMap, a reference-based compression program, to speed up this process. CompMap enables the generation of a non-redundant representative sequence for the input sequences. We have demonstrated that reads can be mapped to this representative sequence with a much reduced time and memory usage, and the mapping to the original reference sequences can be recovered with high accuracy.

Availability and implementation: CompMap is implemented in C and freely available at http://csse.szu.edu.cn/staff/zhuzx/CompMap/.

Contact: xiaoyang@broadinstitute.org

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

Received on July 22, 2014; revised on August 31, 2014; accepted on September 29, 2014

1 INTRODUCTION

In pathogen discovery (Kostic *et al.*, 2011) or metagenomic classification applications (Francis *et al.*, 2013; Lindner and Renard, 2013), curating databases that correspond to a certain taxonomic rank (e.g. species or genus) and assign next-generation sequencing (NGS) reads to these taxa are fundamental tasks. The number of redundant reference sequences (strains or contigs) corresponding to different taxa varies largely as the result of different a priori research interests. For example, there are >39 000 sequences corresponding to the species of *Salmonella enterica* that causes salmonellosis whereas only one reference genome is available for *Bradyrhizobium sp. ORS 278*. Generating a non-redundant representation of each taxon would alleviate this bias and by aligning reads to this non-redundant sequence, memory usage and runtime could be largely reduced. In particular, in applications where exhaustive read placements to highly similar strains or species are required and the multi-mapped reads could reach >99% of the overall input (Francis *et al.*, 2013), read mapping time dominates the entire application. GenomeMapper (Schneeberger *et al.*, 2009) was an early attempt to reduce redundant read mapping to multiple homogenous sequences relying on the existing knowledge of polymorphism information. Without the external information and the assumption of homogeneity, we developed a program CompMap that relies on compression. Although compression techniques have been widely used on NGS data transmission and storage (Zhu *et al.*, 2013), there is a lake of bioinformatic programs like CompMap that can directly use the compressed data and recover the original results if necessary.

CompMap is a reference-based compression program designed to reduce redundancies in a given set of related reference sequences, either heterologous or homogenous, by identifying, recording and eliminating repetitive subsequences. A reduced sequence was obtained to serve as a non-redundant representation of the original input. NGS reads are then mapped to this reduced sequence, and if needed, the original mapping results can be recovered. CompMap shows good performance in the experiment of NGS read mapping to >5000 bacterial genomes.

2 WORKFLOW AND IMPLEMENTATIONS

The workflow of CompMap is illustrated in Figure 1. CompMap consists of three stages.

First, one or multiple sequences, denoted as $\boldsymbol{R}$, are selected from the input by specification or based on length or similarity. An index table **INDEX** is created to store the positions of *k*mers in $\boldsymbol{R}$ with predefined prefixes. These *k*mers serve as the seeds for downstream local alignment. By default, we set prefix to be 'CG' and 'AT' without using the entire 16 combinations for speed and memory considerations, as these are the most commonly occurring dimers in the DNA sequences. *K*mers are stored in binary format with symbols encoded as 'A' = 00, 'C' = 01, 'G' = 10 and 'T' = 11, and other rarely occurring symbols (such as 'W', 'M', 'N', etc.) randomly converted to A, C, G or T.

Second, the non-reference sequences are concatenated to form a sequence $\boldsymbol{M}$, which is then compared against $\boldsymbol{R}$. The comparison starts by identifying in $\boldsymbol{M}$ any *k*mer K_i that is present in the **INDEX**. Then, the occurrences of K_i in $\boldsymbol{R}$ are retrieved. Next, a local alignment is attempted with each matching *k*mer and extends base by base on both ends of the corresponding loci on $\boldsymbol{M}$ and $\boldsymbol{R}$. If a mismatch was encountered, the alignment would be generated for the next N base window. The window would be included with the existing alignment if the mismatch rate remains below e (by default 5%). We only consider a repeat with length at least L, and for such a repeat, we record its length as well as alignment positions in both $\boldsymbol{M}$ and $\boldsymbol{R}$. The locations of

*To whom correspondence should be addressed.

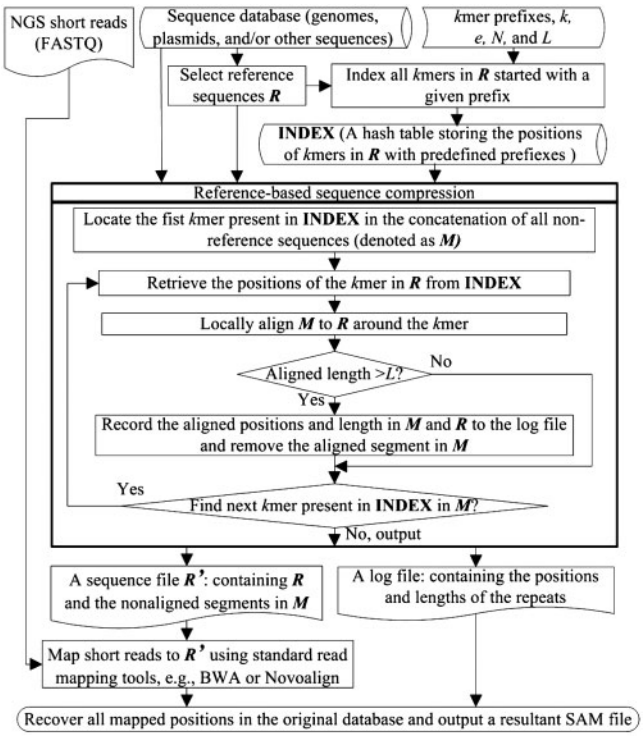

Fig. 1. The flowchart of CompMap

unaligned regions are also recorded. They are further concatenated with ***R*** to form a representation sequence ***R'*** of the input sequences. The unaligned regions shorter than input reads are abandoned.

Finally, NGS reads are mapped to ***R'*** via standard read mapping programs like BWA (Li and Durbin, 2009) or Novoalign (http://www.novocraft.com). Positions where the read should be aligned to the original sequences can be recovered by searching through the records generated in the previous stage. The final alignment output is in SAM format. The implementation details of CompMap are provided in the Supplementary Materials.

3 CASE STUDIES

We conducted two experiments using different datasets on a cluster running 64-bit Red Hat 4.4.4–13 with 32-core 3.1 GHz Intel(R) Xeon(R) CPU E31220. The parameters of CompMap are using the default: $k = 10$, and k mers prefixes = {'CG','AT'}, $e = 0.05$, $N = 10$ and $L = 1000$. A small e allows for a trade-off between compression ratio and the mapping precision. BWA is used as the read mapping tool. We provided guidelines for parameter selection and detailed results in Supplementary Materials.

3.1 Experiment on heterologous data

In the first experiment, we obtained the datasets and database from Francis *et al*. (2013). Particularly, SRR031601, SRR032505 and SRR032501, derived from three bacteria agents *Yersinia kristensenii*, *Yersinia ruckeri* and *Yersinia rohdei*, respectively, are mapped to 170 complete genomes of eight bacterial agents of bioterrorism identified by the U.S. Centers for Disease Control and Prevention. Because reads are derived from different genomes in the database, a low mapping rate is expected.

From the database, 17 chromosomes representing different species are selected to form the reference, based on which the database is compressed to 51.6% of the original size. CompMap mapped 53.19, 20.97 and 40.76% short reads of the three datasets to the database. The mapped ratios are close to BWA's 53.26, 21.33 and 40.89%. Meanwhile, ~95% mapped positions are consistent between BWA and CompMap, where the latter reduced runtime by 43.7% and memory by 49.2%.

3.2 Experiment on homogeneous data

In the second experiment, three *Escherichia coli* NGS datasets SRR1063349, ERR385912 and ERR231645 are mapped to a database consisting of up to 5338 genomes and plasmids of different *E.coli* strains.

Because of a higher homology among sequences in the database, a better compression ratio and more time/space saving are expected compared with heterologous data. A single genome was randomly selected from the database as a reference, and a 34.8% compression rate was achieved. BWA mapped 65.26, 99.37 and 98.96% short reads of the three NGS files to the database, whereas CompMap achieved similar rates of 65.25, 99.31 and 98.79%, but using only an average of 30.8% runtime and 36.7% memory. Compared with heterologous data, a lower percent of mapped positions (90%) is consistent between CompMap and BWA, mainly because lossy compression results in lower accuracy in particularly repetitive datasets (Supplementary Materials). In addition to that, when reference sequences are highly similar, existing aligners like BWA may not be able to identify exhaustive read mapping locations (Supplementary Materials). Nevertheless, when the goal is to identify whether any read could be aligned to any input sequence (Lindner and Renard, 2013), the consistency would substantially improve, e.g. for SRR1063349, the consistency is ~97.6% for the entire input database.

4 CONCLUSION

The case studies show that CompMap allows researchers to speed up NGS short read mapping while maintaining comparable accuracy. The saving of time and space is in proportion to the size of NGS files and the similarity among reference sequences. CompMap is simple yet efficient. It can be readily used along with other tools for NGS data analysis with no or few modifications.

Funding: National Natural Science Foundation of China (61471246 and 61205092), Guangdong Foundation of Outstanding Young Teachers in Higher Education Institutions (Yq2013141), Shenzhen Scientific Research and Development Funding Program (JCYJ20130329115450637, KQC2011 08300045A and ZYC201105170243A) and Guangdong Natural Science Foundation (S2012010009545).

Conflict of interest: none declared.

REFERENCES

Francis,O.E. *et al.* (2013) Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Res.*, **23**, 1721–1729.

Kostic,A.D. *et al.* (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.*, **29**, 393–396.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Lindner,M.S. and Renard,B.Y. (2013) Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.*, **41**, e10.

Schneeberger,K. *et al.* (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol.*, **10**, R98.

Zhu,Z. *et al.* (2013) High-throughput DNA sequence data compression. *Brief. Bioinform*, **16**, 1–15.