Bioinformatics, 33(6), 2017, 926–928
doi: 10.1093/bioinformatics/btw742
Advance Access Publication Date: 23 November 2016
Applications Note

OXFORD

Sequence analysis

Training alignment parameters for arbitrary sequencers with LAST-TRAIN

Michiaki Hamada,[1,2,3,*] Yukiteru Ono,[4] Kiyoshi Asai[3,5] and Martin C. Frith[2,3,5,*]

[1]Department of Electrical Engineering and Bioscience, Faculty of Science and Engineering, Waseda University, 55N–06–10, 3–4–1, Okubo Shinjuku-ku, Tokyo 169–8555, Japan, [2]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 169–8555, Japan, [3]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135–0064, Japan, [4]IMSBIO Co, Ltd, Tokyo 170–0013, Japan and [5]Graduate School of Frontier Sciences, University of Tokyo, Chiba 277–8562, Japan

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on October 9, 2016; revised on November 4, 2016; editorial decision on November 17, 2016; accepted on November 18, 2016

Abstract

Summary: LAST-TRAIN improves sequence alignment accuracy by inferring substitution and gap scores that fit the frequencies of substitutions, insertions, and deletions in a given dataset. We have applied it to mapping DNA reads from IonTorrent and PacBio RS, and we show that it reduces reference bias for Oxford Nanopore reads.
Availability and Implementation: the source code is freely available at http://last.cbrc.jp/
Contact: mhamada@waseda.jp or mcfrith@edu.k.u-tokyo.ac.jp
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The classic approach to pair-wise sequence alignment is to seek alignments that maximize a score, which is a sum of substitution and gap scores. This is equivalent to seeking alignments with maximum likelihood, using a statistical model with probabilities for each kind of substitution (e.g. c→t), insertion, and deletion. This approach was developed several decades ago, mainly for proteins, but also for nucleotide sequences (Chiaromonte *et al.*, 2002; States *et al.*, 1991). It is arguably *least* suited to homology search, because different homologs of one protein have different levels of divergence, so that one set of parameters cannot be optimal for all homologs.

Here, we are interested in aligning nucleotide sequences that differ mainly by sequencing error. Compared to homology search, it is more likely that a single set of substitution and gap probabilities will be a universal good fit, for one version of one sequencing technology, applied to one type of DNA. On the other hand, these probabilities may be quite different for different technologies, and even for different versions of the same technology. Moreover, these probabilities will differ for unusual types of DNA, such as 80%-AT Plasmodium genomes or PAR-CLIP data (Kerpedjiev *et al.*, 2014). Thus, it would be useful to have a tool that automatically determines suitable parameters for a given dataset.

Although the score/model-based approach to alignment is well-known and classic, it has been surprisingly neglected in recent high-throughput DNA aligners (Kerpedjiev *et al.*, 2014). It is likely that accuracy is maximized by using scores that fit the substitution and gap frequencies in the data.

In this study, we introduce a novel tool, LAST-TRAIN, to train alignment parameters from sequence data. We use it to train parameters for PacBio RS, IonTorrent and Nanopore. Finally, we show that it mitigates reference bias (haplotypes appearing in the reference genome tend to be over-estimated) for Oxford Nanopore reads.

2 Methods

LAST-TRAIN's input is query (e.g. DNA reads) and reference (e.g. a genome) sequence datasets. It uses a standard iterative approach: it

Table 1. Results of haplotype phasing with Nanopore long reads (NA12878)

		GraphMap		BLASR		LAST				LAST + LAST-TRAIN			
						Manual ($q = 1$)		Manual ($q = 2$)		Training		Training+LAMA	
Haplotype	Polymorphism	Count	Freq	Count	Freq	Count	Freq	Count	Freq	Count	Freq	Count	Freq
TT: CT	CYP2D6*4	207	11.9%	227	18.4%	340	20.1%	182	21.2%	327	27.3%	343	27.4%
TT: CC	(reference bias)	225	13.0%	329	26.6%	326	19.3%	164	19.1%	160	13.4%	134	10.7%
T−: CT		70	4.0%	31	2.5%	65	3.8%	36	4.2%	75	6.3%	78	6.2%
T−: CC	CYP2D6*3	226	13.0%	281	22.8%	232	13.7%	199	23.1%	199	16.6%	217	17.3%
Other		1006	58.1%	367	29.7%	726	43.1%	279	32.4%	436	36.4%	480	38.3%
Total		1734	100.0%	1235	100.0%	1689	100.0%	860	100.0%	1197	100.0%	1252	100.0%

In the first column, TX:CY indicates the phased haplotype where the 1st position (rs35742686) is 'X' ('T' in the reference genome) and the 2nd position (rs3892097) is 'Y' ('C' in the reference genome). See also Supplementary Table S14. The high frequency for TT:CC (the identical haplotype to the reference genome) is known as *reference bias* (Laver *et al.*, 2016). The values for 'BLASR' were computed from the mapping results in Ammar *et al.* (2015), where BLASR was used for mapping Nanopore reads to the reference genome. The column 'training + LAMA' shows the results of probabilistic alignment (Hamada *et al.*, 2011) using forward scores with the trained parameters by LAST-TRAIN. See Supplementary Materials S7 for the detailed command line options for every tool.

first aligns the sequences using some initial score parameters, then infers better score parameters from the alignments, then re-aligns and repeats until the parameters stop changing (Durbin *et al.*, 1998). It achieves adequate speed by an X-drop heuristic (Altschul *et al.*, 1997; Zhang *et al.*, 1998), it depletes paralogs using LAST-SPLIT (Frith and Kawaguchi, 2015), and it allows different insertion and deletion parameters and non-strand-symmetric substitution parameters. Details are in the Supplementary Material S1.

3 Results

3.1 Mitigating *reference bias* in haplotype phasing

Long-read DNA sequencing is a promising way to determine *phasing* between DNA variants. Most human cells have two copies, maternal and paternal, of each chromosome (except Y). Suppose that a patient has two variants in different exons of one gene, where each variant destroys the gene's function but is present in just one chromosome (maternal or paternal). It is important to know whether they are in the *same* chromosome.

A previous study attempted to determine the haplotypes of CYP2D6, a gene that affects metabolism of clinical drugs, in human sample NA12878, by PCR amplification of the relevant genomic region followed by Oxford Nanopore sequencing (Ammar *et al.*, 2015). This sample is known to have the two haplotypes CYP2D6*3 and CYP2D6*4, shown in Supplementary Table S14 (Numanagi *et al.*, 2015; Twist *et al.*, 2016), however the original study found a prominent third haplotype. A later re-analysis (Laver *et al.*, 2016) suggested two reasons for this. First, chimeric crossovers between the two variants appeared during PCR amplification, producing two false haplotypes. Second, because one of the false haplotypes matches the reference genome, it was prominently detected after aligning the DNA reads to the reference. The latter phenomenon is termed reference alignment bias.

We reasoned that, if our trained parameters produce more accurate alignments, they should reduce reference bias. Following Ammar *et al.* (2015), we used high-quality (2D) reads (SRA1748415): 7540 reads with average length 3486. First, we trained alignment parameters using these reads and human reference genome hg19, leading to the following parameters; the substitution score matrix is

	A	C	G	T
A	7	−11	−8	−16
C	−7	5	−8	−7
G	−5	−8	5	−8
T	−19	−12	−13	7

and the gap costs are $12 + 3k$ for a length-k deletion and $15 + 3k$ for a length-k insertion.

Second, the haplotypes for two target polymorphism sites (Supplementary Materials S14) were predicted in the following (direct and simple) manner. (i) The best alignment was taken for each read after mapping the read to the reference genome (hg19). (Note that multiple maps are expected because there is a paralog of CYP2D6, whose similarity is about 94%.) (ii) Among the obtained alignments, alignments covering both of the target polymorphism sites in CYP2D6 were taken and then the count and frequency of each haplotype were computed from those alignments.

The results are shown in Table 1, indicating that reference bias is lessened by aligning with trained parameters, compared to GraphMap (v0.3.0) (Sovic *et al.*, 2016), BLASR (Chaisson and Tesler, 2012) and LAST with manually-determined parameters. Specifically, the frequency of the chimeric reference haplotype is reduced, whereas the frequency of the chimeric non-reference haplotype is increased.

In addition, we performed probabilistic alignment (Hamada *et al.*, 2011) with trained parameters, because probabilistic alignment tends to estimate columns in the alignment more accurately than conventional alignment. Specifically, we used LAMA alignment (lastal option '-j6') with $\gamma = 2$ (Hamada *et al.*, 2011). In this case, we choose the best alignment using 'forward' scores (from summing the probabilities of all alignments in the X-drop algorithm) instead of conventional scores (from the single best alignment). Table 1 suggests that trained parameters with LAMA alignment improve the results further (lower frequency of TT:CC).

3.2 Further results

We applied LAST-TRAIN to PacBio, IonTorrent, and Oxford Nanopore DNA reads using several available datasets (Supplementary Materials S2). It successfully recovers known

features, such as PacBio having more insertions than deletions (Supplementary Materials S4), and we established that a query sample size of 1–10 million bases is sufficient (Supplementary Materials S3). Moreover, evaluation on simulated datasets indicates that trained parameters slightly improved alignment accuracy (Supplementary Materials S5). Notice that all the trained parameters and their statistics are shown in supplementary information (Supplementary Materials S9). Finally, the Supplement has discussion of: last-train versus MarginAlign (Jain *et al.*, 2015), resisting the temptation of over-alignment, and use of sequence quality data (Supplementary Materials S8).

Acknowledgements

We thank Toshiyuki Sato (Mizuho Information and Research Institute, Inc) for supporting the implementation of LAST-TRAIN. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Funding

This work was supported by MEXT KAKENHI [grant numbers JP24680031, JP16H05879, JP16H01318 to M.H.; JP26700030 to M.C.F.; JP25240044 to M.H. and K.A.; JP221S0002 to K.A.].

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ammar,R. *et al.* (2015) Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Res*, **4**, 17.

Chaisson,M.J. and Tesler,G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.

Chiaromonte,F. *et al.* (2002) Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.*, 115–126. pages

Durbin,R. *et al.* (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Frith,M., and Kawaguchi,R. (2015) Split-alignment of genomes finds orthologies more accurately. *Genome Biol.*, **16**, 106–106.

Hamada,M. *et al.* (2011) Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. *Bioinformatics*, **27**, 3085–3092.

Jain,M. *et al.* (2015) Improved data analysis for the MinION nanopore sequencer. *Nat. Methods*, **12**, 351–356.

Kerpedjiev,P. *et al.* (2014) Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinformatics*, **15**, 100.

Laver,T.W. *et al.* (2016) Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci. Rep.*, **6**, 21746.

Numanagi,I. *et al.* (2015) Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data. *Bioinformatics*, **31**, 27–34.

Sovic,I. *et al.* (2016) Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.*, **7**, 11307.

States,D.J. *et al.* (1991) Improved sensitivity of nucleic acid database similarity searches using application specific scoring matrices. *Methods*, **3**, 66–70.

Twist,G.P. *et al.* (2016) Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. *Npj Genomic Med.*, **1**, 15007.

Zhang,Z. *et al.* (1998) Alignments without low-scoring regions. *J. Comput. Biol.*, **5**, 197–210.