BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 7 2009, pages 969–970
doi:10.1093/bioinformatics/btp092

Sequence analysis

MOM: maximum oligonucleotide mapping

Hugh L. Eaves[1] and Yuan Gao[1,2,*]

[1]Center for the Study of Biological Complexity and [2]Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia, USA

Received on September 22, 2008; revised on February 13, 2009; accepted on February 14, 2009
Advance Access publication February 19, 2009
Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Current short read mapping programs are based on the reasonable premise that most sequencing errors occur near the 3′ end of the read. These programs map reads with either a small number of mismatches in the entire read, or a small number of mismatches in the segment remaining after trimming bases from the 3′ end or a single base from the 5′ end. Though multiple sequencing errors most likely occur near the 3′ end of the reads, they can still occur at the 5′ end of the reads. Trimming from the 3′ end will not be able to map these reads. We have developed a program, Maximum Oligonucleotide Mapping (MOM), based on the concept of query matching that is designed to capture a maximal length match within the short read satisfying the user defined error parameters. This query matching approach thus accommodates multiple sequencing errors at both ends. We demonstrate that this technique achieves greater sensitivity and a higher percentage of uniquely mapped reads when compared to existing programs such as SOAP, MAQ and SHRiMP.
Software and Test Data Availability: http://mom.csbc.vcu.edu
Contact: ygao@vcu.edu; hleaves@vcu.edu

1 INTRODUCTION

Current Illumina-Solexa sequencing errors have been shown to be highly dependent on the position of a base within the read. The error rate is generally higher at the 3′ end, but there is also a non-negligible and sometimes significant error at the 5′ end that could be a result of bubbles or other machine or reagent related issues (Dohm *et al*., 2008).

Existing read mapping tools such as Eland (A.Cox, unpublished data) and MAQ (H.Li, unpublished data) are designed to handle errors distributed randomly throughout the read or in the case of SOAP (Li *et al.*, 2008) with its trimming capability, more frequently at the 3′ end. Reads with multiple errors at the 5′ end of the read will often be discarded entirely by either Illumina quality score filtering or by these tools, or mapped with a suboptimal alignment, even though a majority of bases in a read matched within the acceptable error parameters. Therefore, in the interest of extracting more usable data from the raw dataset, we have developed a new tool, Maximum Oligonucleotide Mapping (MOM), that handles errors at both the 3′ or 5′ ends, and provides a better alignment in such situations. With this improvement, MOM maps a larger number of reads and still outperforms some existing tools.

MOM's search sensitivity is controlled by two main parameters: the maximum number of mismatching bases allowed in a match, and the minimum allowable length of the match. In addition, seed size and seed spacing parameters can be specified, but these parameters default to reasonable values if they are not specified. MOM's algorithm returns all matches that meet or exceed the specified minimum match length and number of mismatches detectable with the given seed size and seed spacing parameters.

MOM is written in Java and is a multithreaded to take advantage of systems with multiple CPU's. MOM utilizes the fastutil Java collections library developed by Sebastiano Vigna (http://fastutil.dsi.unimi.it/), and requires a Java Runtime Environment version 6.0 or later.

2 METHODS

Like BLAST (Altschul *et al*., 1997), BLAT (Kent, 2002) and many other alignment tools, MOM is fundamentally a seed based search tool. MOM's algorithm has two stages: searching for exactly matching short subsequences (seeds) between the reference and query sequences, and performing ungapped extension on those seeds to find the longest possible matching sequence with the user specified number of mismatches. To search for matching seeds, MOM creates a hash table of subsequences of fixed length '*k*' (*k*mers) from either the reference or query sequences, and then sequentially reads the un-indexed sequence searching for matching *k*mers in the hash table.

The locations of any matches from the hash table search are passed to the local alignment algorithm for sequence extension. MOM's local alignment algorithm does not allow for indels, but does allow for a user specified number of mismatches in the matching region and unlimited mismatches in the 3′ and 5′ flanking regions. To find the longest match, MOM compares sequences in the 5′–3′ direction, starting from the first base in the query sequence regardless of the position of the seed. The sequence is extended in the 3′ direction until either the maximum number of allowable mismatches or the end of the sequences is reached. If the maximum number of mismatches is reached before the end of the sequence, MOM trims bases from the 5′ end to the first mismatching base and begins extending the alignment in the 3′ direction again. This 5′ trimming and 3′ extension occurs iteratively until the end of the sequence is reached, at which point the length of the longest alignment achieved is compared with the user specified minimum match length. If the minimum match length is met, MOM considers the match to be valid.

Multiple matches on the same read are considered to be equivalent if they have the same length and same number of mismatching bases in the read. MOM records the number of equivalent longest matches found for a given read, disregarding any shorter matches, or any matches of the same length with more mismatches. If there is only one longest match for a read, the match is considered to be unique. MOM reports the number of best matches for each query in the output file.

To test the performance and sensitivity of MOM, we generated several short read datasets using the Illumina Genome Analyzer (GAI) using the

*To whom correspondence should be addressed.

entire genome of a human subject. We then mapped the resulting reads onto the National Center for Biotechnology Information (NCBI) Human Reference Genome, build 36.3, using MOM and three other mapping programs, and compared the results. For all datasets, the median Illumina quality score was 40, corresponding to a median probability of one incorrect base call for every 10 000 bases.

When selecting programs to use for the comparison, we attempted to utilize a representative cross set of the currently available short read mapping programs. Most existing programs (Eland, MAQ, RMAP, etc.) perform matches based on the entire length of the read and therefore produce very similar results. However, some programs do offer alternative alignment methods such as SOAP, which can iteratively trim from the 3′ end of the read, and SHRiMP, (M. Brudno *et al*., unpublished data) which performs Smith–Waterman alignment on the reads. We therefore chose three publicly available programs to compare to MOM: the widely used MAQ program as a representative program using full read length matching, SOAP as an example of a program that can align a shorter portion of the read based on iterative trimming and SHRiMP which performs Smith–Waterman alignment. Most of our analysis focuses on SOAP, as it offers comparable performance to MOM and due to a similar algorithm, allowed measurement of the contribution of two-tailed trimming to the results.

Where possible, the programs were run with equivalent parameters to provide a fair comparison. SOAP was run using its default seed size of 10 bp, a maximum of two mismatching bases per match, and both with and without 3′ trimming enabled. MAQ was run with default settings for 40 bp reads. MOM was run twice: once using a non-overlapping seed size of 10 bp, a minimum match length of 23 bp, and a maximum of two mismatching bases per match. When using trimming, SOAP returns matches as short as 23 bp, so 23 bp was also selected as the minimum match length for MOM. In addition, MOM was run with a minimum match length of 40 with two mismatches to compare to MAQ and SOAP without trimming. The MOM default seed size of 13 bp was used for this run. Throughout this article, we will refer to the MOM run with the 23 bp minimum as MOM 23/2 and MOM run with the 40 bp minimum as MOM 40/2. SHRiMP was run using default parameters for 40 bp reads.

All tests were executed on a dedicated four CPU 2.4 GHz AMD Opteron system with 32 GB of system memory. The CPU time, elapsed time, maximum memory, the number of uniquely mapped reads and the total number of mapped reads were recorded for each run.

3 RESULTS

For all test datasets, MOM mapped the highest number of reads overall. MOM was comparable in speed to the existing tools, taking longer in some cases and less in others. These results are summarized in Table 1.

As the main difference between MOM and existing tools is the ability to identify the longest matching segment, equivalent to performing optimal two tailed trimming, we examined the contribution of two sided trimming to the results. When comparing the results from dataset A, we found SOAP was able to map 66 860 (0.9%) reads that were not found by MOM. The additional matches were primarily due to differences in the seeding algorithm of SOAP and MOM. SOAP uses the split seed algorithm originated by Eland, allowing for up to two mismatches within the seed itself, whereas MOM missed reads where there was not a mismatch free seed within the read.

MOM, however, was able to map 1 313 667 reads (18.5%) that were not mapped by SOAP. Of these, 1 309 621 were matched after 5′ trimming, a direct result of MOM's maximal segment matching algorithm. There were also 347 300 reads mapped by both SOAP and MOM where the length of the MOM match was greater than the length of the SOAP match, due to MOM's better handling of

Table 1. Performance and sensitivity comparison of short read mapping programs ordered by total reads mapped (%) and dataset

Program	Dataset[a]	Uniquely mapped reads (%)	Total reads mapped (%)	Elapsed time	RAM used (GB)
MOM 23/2	A	66.9	86.6	72 h 27 min	17.1
SOAP (w/trim)	A	59.5	74.8	118 h 27 min	19.2
MOM 40/2	A	42.5	52.0	12 h 32 min	6.3
SOAP (no trim)	A	42.5	52.0	18 h 20 min	19.2
MOM 23/2	B	77.3	96.1	40 h 58 min	16.4
SOAP (w/trim)	B	71.9	89.1	29 h 55 min	18.8
MAQ	B	63.0	77.4	17 h 01 min[b]	3.5
MOM 23/2	C	76.7	97.9	42 h 38 min	16.5
SOAP (w/trim)	C	79.0	96.7	22 h 17 min	18.8
MAQ	C	71.1	86.2	22 h 52 min[b]	3.5
MOM 23/2	D	78.9	96.5	39 min	0.5
SOAP (w/trim)	D	77.6	94.7	36 min	14.0
SHRiMP	D	55.9	92.9	11 h 36 min[b]	0.6

[a]Dataset A contained 10 601 732 40 bp raw reads without quality scores, B contained 5 909 598 40 bp reads with quality scores, C contained 5 677 142 40 bp reads with quality scores and D contained 100 000 40 bp reads with quality scores.
[b]These programs only support single threaded operation by default. Elapsed time for these programs represents execution on a single CPU.

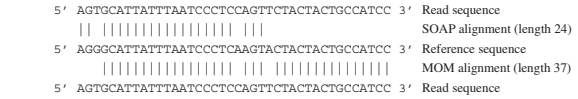

Fig. 1. Comparison of alignment from MOM and SOAP for the same read and reference sequences.

5′ errors. On average, MOM mapped an additional 3.4 bp for these reads, for a total of 1 172 766 additional bp mapped. A comparison of the alignments from MOM and SOAP for one of these reads is shown in Figure 1.

MOM's algorithm clearly finds more matches than SOAP or MAQ for a given input dataset. However, there are still several improvements we would like to make to our software, including paired read support, gap detection and Eland's split seed method. In addition, we plan to investigate the effects of the additional mapped reads on downstream processing for genome sequencing or genome to genome comparison.

ACKNOWLEDGEMENTS

We would like to acknowledge VCU students Phillip Stitcher and Gregory Smith, and the VCU Center for High Performance Computing for their assistance in this project.

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.

Dohm,J.C. *et al*. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* [Epub ahead of print, doi:10.1093/nar/gkn425].

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Li,R. *et al*. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.