BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 5 2008, pages 713–714
doi:10.1093/bioinformatics/btn025

Sequence analysis

SOAP: short oligonucleotide alignment program

Ruiqiang Li[1,2], Yingrui Li[1], Karsten Kristiansen[2] and Jun Wang[1,2,*]

[1]Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China and [2]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, DK-5230, Denmark

Received on November 10, 2007; revised on December 20, 2007; accepted on January 14, 2008

Advance Access publication January 28, 2008

Associate Editor: Keith Crandall

ABSTRACT

Summary: We have developed a program SOAP for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The program is designed to handle the huge amounts of short reads generated by parallel sequencing using the new generation Illumina-Solexa sequencing technology. SOAP is compatible with numerous applications, including single-read or pair-end resequencing, small RNA discovery and mRNA tag sequence mapping. SOAP is a command-driven program, which supports multi-threaded parallel computing, and has a batch module for multiple query sets.

Availability: http://soap.genomics.org.cn

Contact: soap@genomics.org.cn

The new DNA sequencing technologies, which have been developed and implemented recently, have significantly improved throughput and dramatically reduced the cost compared with the capillary-based electrophoresis systems (Shendure *et al.*, 2004). In a single experiment using one instrument, the Illumina-Solexa system using sequencing-by-synthesis (SBS) can determine up to 40 million sequences of up to 50 bases in length, whereas the ABI-SOLiD system using ligation technology allows determination of 3 Gb mappable data. The ultra high throughput and short read length make these technologies particularly suitable for large-scale resequencing of large cohorts of individuals with known reference for studies of genetic variations (Bentley, 2006). Traditional sequence alignment software like blast (Altschul *et al.*, 1997) and blat (Kent, 2002) are unable to cope efficiently with the huge amount of reads generated in such applications, while SSAHA (Ning *et al.*, 2001) is optimized to find long alignments and fails in practice on most short queries. To our knowledge, there have been several programs developed or under developing to match the new sequencing technologies. ELAND, an alignment tool integrated in Illumina-Solexa data processing package, can do ungapped alignment for reads with size up to 32 bp (Cox, unpublished). Maq is another program for ungapped alignment, which implemented sophisticated probability models to measure alignment quality of each read using sequence quality information (Li, unpublished). Here, we present a new program SOAP, which can do both ungapped and gapped alignment, and has special modules for alignment of pair-end, small RNA and mRNA tag sequences.

*To whom correspondence should be addressed.

SOAP will allow either a certain number of mismatches or one continuous gap for aligning a read onto the reference sequence. The best hit of each read which has minimal number of mismatches or smaller gap will be reported. For multiple equal-best hits, the user can instruct the program to report all, or randomly report one, or disregard all of them. Since the typical read length is 25–50 bp, hits with too many mismatches are unreliable which are hard to distinguish with random matches. By default, the program will allow at most two mismatches. Between two haplotype genome sequences, occurrence of single nucleotide polymorphism is much higher than that of small insertions or deletions, so ungapped hits have precedence over gapped hits. For gapped alignment only one continuous gap with a size ranging from 1 to 3 bp is accepted, while no mismatches are permitted in the flanking regions to avoid ambiguous gaps. The gap could be either insertion or deletion in the query or the reference sequence. As the intrinsic character of the sequencing technology, errors will accumulate during the sequencing process. Reads always exhibit a much higher number of sequencing errors at the 3′-end, which sometimes make them unalignable to the reference sequences. To deal with the problem, SOAP can iteratively trim several basepairs at the 3′-end and redo the alignment, until hits are detected or the remaining sequence is too short for specific alignment.

Pair-end sequencing means to sequence both ends of a DNA fragment. So the two reads belonging to a pair will always have the settled relative orientation and approximate distance between each other on the genome. The technology can significantly improve the accuracy of resequencing mapping, and is a powerful method for detection of structural variants including copy number variations (CNVs), rearrangements, inversions and etc. SOAP is able to align a pair of reads simultaneously. A pair will be aligned when two reads are mapped with the right orientation relationship and proper distance. Similar filter as single-read alignment, a certain number of mismatches are allowed in one or both reads of the pair. For gapped alignment, gap is only permitted on one read, and the other end should match exactly.

Apart from genome resequencing, The high throughput sequencing technology lends itself to numerous applications. For some applications (ex. ChIP-Seq), the data analysis process is essentially identical to that of resequencing. Additionally, SOAP provides special modules for small RNA discovery and mRNA tag profiling analysis. Small RNAs have a size between

18 to 26 bp. According to the experimental protocol, the 3′-end of RNA sequence will be flanked by adapter sequences. SOAP will filter adapter sequence, and then align the remaining candidate small RNA to the reference sequence. A small RNA will be annotated if an adapter sequence is detected and the insert sequence match well with the reference sequence. Considering sequencing errors, one or two mismatches can be allowed insider either the adapter or the candidate RNA region according to user settings. On mRNA tag sequencing, there are two types of restriction enzyme digestion: (i) DpnII, which will specially recognize the site 'GATC' and cuts a 16 bp tag after the site; (ii) NlaIII, which exclusively recognizes the site 'CATG' and cuts 17 bp downstream. SOAP checks and trims off the 3′-end adapter sequence according to the enzyme type. Aligned hits should contain the enzyme site, and have at most one mismatch in the tag region.

SOAP uses seed and hash look-up table algorithm to accelerate alignment. Both reads and the reference sequences are converted to numeric data type using 2-bits-per-base encoding. A read will do exclusive-OR comparison with the reference sequence. Then the value is used as suffix to check the look-up table to know how many bases are different. In order to have a tradeoff between memory usage and efficiency, SOAP uses unsigned 3-bytes data type as the table element. To admit two mismatches, a read is splitted into four fragments, the two mismatches can exist in at most two of the fragments on the same time, then if we try all six combinations of the two fragments as seed, we can however catch all hits with two mismatches (the algorithm is the same as Eland and Maq). Since mismatches are not allowed in gapped hits, SOAP used the enumeration algorithm which tries to insert a continuous gap or delete a fragment at each possible position in a read. The algorithm outputs the identical alignments as that of dynamic programming while runs much faster. Not alike Eland and Maq which load read sequences into memory and build seed index tables for reads, SOAP loads reference sequences into memory as an unsigned 3-bytes array and builds the seed index tables for all the reference sequences. Then for each read, create seeds and search the corresponding index table for candidate hits, perform alignment and report the results. The RAM required for storing the reference sequences and seed index tables can be calculated as:

$$RAM = \frac{L}{3} + (4 * 3 + 8 * 6) * 4^S + (4 + 1) * 3 * \frac{L}{4} + 4 * 2^{24}$$

where L is the total length of the reference sequences; S is seed size. For small reference like yeast, $L = 12$ Mb and selected seed size $S = 10$ bp, about 200 Mb RAM is needed; but for the whole human genome, $L = 3$ Gb and a selected seed size $S = 12$ bp, about 14 Gb RAM in total will be needed.

Evaluated on a real dataset containing 9 914 527 Illumina-Solexa single-end resequencing reads (length 32 bp), which were generated from a 5 Mb human genome region, SOAP was almost 300 (gapped) to 1200 (ungapped) times faster than blastn, while having better sensitivity (Table 1). The iterative feature of SOAP improved sensitivity. And gapped alignment can further identify hits accommodating small indels which compose only a small fraction of all hits but are a very important class of mutation. Since SOAP loads reference sequences into memory, while Eland and Maq load reads, the memory usage varies in different datasets.

Table 1. Comparison of performance and sensitivity among short oligonucleotide alignment programs

Program	Time consumed (s)	Reads aligned (%)
blastn (−F F −W 11)	165 780	85.47
blastn (−F F −W 15)	150 660	84.66
Blat (−tileSize = 8)	22 032	85.07
Eland	166	88.53
Maq	458	88.39
Soap	134	88.46
Soap iterative	161	90.9
Soap iterative + gapped	486	91.15

We used a query dataset containing 9 914 527 single-end reads (length 32 bp) generated by Illumina-Solexa Genome Analyzer. The DNA sample is a mixture of long PCR products of a 5 Mb human genome region. For blastn we tried 11(better sensitivity) and 15(faster) bp word size, and disabled DUST masking of low-complexity sequence (−F F). For blat, tileSize parameter was set at 8. SOAP used 12 bp seed, SOAP iterative will iteratively trim off 2 bp at 3′-end of read and redo alignment until hits were detected or the remaining sequence is shorter than 27 bp. Sensitivity is calculated under the same threshold by allowing at most 2 mismatches. SOAP gapped will allow one continuous insertion or deletion with size between 1 to 3 bp. After checking sequencing quality, we found the remaining unmappable reads are in low sequencing quality.

SOAP accepts FASTA format for reference, and either FASTA or FASTQ format for query reads. It's a command-driven program, which employs single command line model and batch computing model. On batch computing model, the reference sequences and hash index tables will reside in the memory and alignment procedure can be performed for multiple query datasets in a order. This model avoids I/O and time wasted on loading reference and creating hash tables multiple times, and is also suitable for real-time web service. SOAP is written in standard C++ language and runs well on Macintosh or any 64-bit Linux/Unix systems. It supports multithreaded parallel computing.

ACKNOWLEDGEMENTS

We are indebted to Wei Fan, Qibin Li, Xiaodong Fang, Zhike Lu, Guoqing Li, Junjie Qin and the other users who tested the beta version of the program for identifying bugs and proposing all kinds of improvements. We thank Shengting Li for setting up the website. The project is supported by the National Natural Science Foundation of China (30725008), and grants from the Danish Natural Science Research Council (272-05-0344 and 272-07-0196).

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.

Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.

Cox,A. (unpublished) ELAND: Efficient Local Alignment of Nucleotide Data.

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

Li,H. (unpublished) Mapping and assembly with quality. http://maq.sourceforge.net/.

Ning,Z. *et al.* (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.

Shendure,J. *et al.* (2004) Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.*, **5**, 335–344.