Published in final edited form as:

J Data Mining Genomics Proteomics. 2013 August 1; 4(3): . doi:10.4172/2153-0602.1000133.

SlopMap: a software application tool for quick and flexible identification of similar sequences using exact k-mer matching

Ilya Y. Zhbannikov^{1,*}, Samuel S. Hunter^{1,†}, Matthew L. Settles^{1,2,‡}, and James A. Foster^{1,2,§}

¹Department of Bioinformatics and Computational Biology, University of Idaho, USA

²The Institute of Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, USA

Abstract

With the advent of Next-Generation (NG) sequencing, it has become possible to sequence a entire genomes quickly and inexpensively. However, in some experiments one only needs to extract and assembly a portion of the sequence reads, for example when performing transcriptome studies, sequencing mitochondrial genomes, or characterizing exomes. With the raw DNA-library of a complete genome it would appear to be a trivial problem to identify reads of interest. But it is not always easy to incorporate well-known tools such as BLAST, BLAT, Bowtie, and SOAP directly into a bioinformatics pipelines before the assembly stage, either due to incompatibility with the assembler's file inputs, or because it is desirable to incorporate information that must be extracted separately. For example, in order to incorporate flowgrams from a Roche 454 sequencer into the Newbler assembler it is necessary to first extract them from the original SFF files.

We present SlopMap, a bioinformatics software utility that allows quickly identification similar to the provided reference reads from either Roche 454 or Illumina DNA library. With simple and intuitive command-line interface along with file output formats compatible to assembly programs, SlopMap can be directly embedded to biological data processing pipeline without any additional programming work. In addition, SlopMap preserves flowgram information needed for Roche 454 assembler.

1 Introduction

New methodologies enabled by Next Generation Sequencing (NGS) that are of particular interest to us include transcriptome analysis for RNA research [4] and mitochondrial sequencing from exome data [3]. These applications identify reads that are similar to a reference transcript or genome. It may seem trivial to identify the reads of interest among those produced by NGS hardware, using well-known general-purpose alignment or mapping tools such as Blat [6], Bowtie2 [7], BWA [8], and SOAP [5]. But even an efficient tool may be difficult to incorporate directly into a bioinformatics pipeline before the assembly stage, since it may be necessary to convert data to a different file format. For example, existing mappers usually use the SAM/BAM file format as output. None use sff format files [1] as both input and output, and none but Bowtie2 support FastQ output.

Moreover, it is difficult to use existing mappers in such cases when it needs to establish a similarity threshold, i.e. when one wants reads that are 50%, 70% or 85% similar to the reference. Relying only on input parameters such as gap penalties and seed size, which most

*Electronic address: zhba3458@vandals.uidaho.edu; Corresponding author.

†Electronic address: shunter@uidaho.edu

‡Electronic address: msettles@uidaho.edu

§Electronic address: foster@uidaho.edu

well-known aligners have, it is difficult to achieve flexible mappings with required percentage of similarity. On the other hand, it is often desirable to find reads that are at least 90% similar to the reference, and to discard the rest. Also, it is common to use 5% – 10% tolerance threshold, which means if the number of gaps in a particular read is more than 5% – 10%, the read is ignored. This is a potential problem when the DNA library contains erroneous reads, which is common with existing NGS technologies.

Another problem arises if there is insufficient data on the edge of the transcript or reference genome (Figure 1). In this situation the whole read (marked red) can potentially be discarded due to lack of data on the edge, even if a part of the read has significant similarity to the reference.

And finally, sometimes it is not easy to tune a general-purpose mapper due to many features to achieve its optimal performance.

We present SlopMap, a bioinformatics software utility that quickly and flexibly identifies similarities between a given reference sequence and sequence reads from either Roche 454 or Illumina sequence data. SlopMap can be directly embedded to biological data processing pipeline before the assembly stage without any additional programming work, since it has an intuitive user interface, maintains file formats compatibility to standard assembly programs, and preserves original information such as bases, quality metrics, and flowgrams. SlopMap uses only exact kmer matching, which we call "sloppy" mapping, without providing an alignment step. SlopMap is a simple and robust tool that can be used with the threshold up to 5% similarity (95% mismatches) when it is desired.

2 Method

SlopMap is fully implemented and optimized in C++ for efficiency. This is a command-line application with all the input parameters to be specified within the command string. SlopMap is tunable via input parameters for kmer size k , percentage of similarity t and distance between two consecutive kmers d . It also supports flexible input and output file formats: FastQ, FastA, SFF and TXT.

2.1 Input Files

2.1.1 Reference—The reference library file, also known as a “database”, is a FastA formatted file that contains one or multiple records. Each record consists of two parts: a header and a sequence string. The header must contain a name which is a unique identifier of the record. The sequence string is a DNA sequence specifies a reference genome or any other DNA of interest.

2.1.2 DNA Query Library Files—The DNA query library files are data from the NGS machines. SlopMap can take either Roche 454 SFF or FastQ formatted files, or Illumina paired- or single-end reads. SlopMap computes the similarity of each read in a query library file to record in a reference library.

2.2 Search procedure

SlopMap uses a multi-kmer approach with single-base minimum overlap between two consecutive kmers in order to quickly determine the similarity to the reference record. The algorithm is simple and straightforward:

1. Compute a kmer dictionary from the given reference library.
2. For each query sequence (“read”) in the DNA query library:

- Compute a set of consecutive kmers.
- Calculate the read's similarity.

3. Output summary statistics and files.

2.2.1 Dictionary—The first step is to build a dictionary indexed with kmers by sampling the given reference library sequences (Figure 2). The kmers are short (usually 9–15 bases long) substrings that represent the contiguous reference library record string. The reference library sequences are sampled with the following predefined parameters: *kmer size* and a distance representing a *constant overlap between consecutive kmers*. These parameters remains unchanged throughout program execution. All kmers are hashed and associated with the offset position in the reference string and a reference record id. By default SlopMap uses kmer size of 11 bases. Google Dense Hash Map that allows short data retrieval time and memory efficiency is used as a data structure for the kmer dictionary. [2]

2.2.2 Kmer matching—Query strings are sequentially sampled, so that each query string in the given DNA query library is handled individually. Kmer size and distance between two consecutive kmers remains the same for all reads in query library. For each read taken from the query library, a dictionary search is performed and similarity between the read and the reference is calculated as follows:

$$S = \frac{\text{number shared bases}}{\text{Min_length}(\text{query string}, \text{reference string})}, 0 \leq S \leq 1$$

Where *number shared bases* is the total length of kmers shared between reference and a query string, *Min length(query string, reference string)* is the minimum length, either of query or reference. Those reads that meet the pre specified similarity threshold (parameter “*t*” in SlopMap, by default it is set to 11 bases) are then saved in output files, others are discarded. The lowest value of *S* is 0 and in this case the read does not share anything with the reference. In opposite to this, *S* equal to 1 means the read is 100% identical to the reference.

2.3 Output files

The output files are:

- A report file that contains all information about the reads, including similarity, positions of first and last match, bases and quality scores.
- One or two FastQ formatted file(s), depending on whether the data are from Roche 454 or Illumina machines.
- Optionally, an SFF formatted file, which contains only those Roche 454 sequences similar to the provided reference.

3 Validation

We compared SlopMap to several alternative DNA mapping tools: Bowtie2, BWA, Blat, on two different query DNA libraries: 621,578 Roche 454 *Escheria coli* K12 W583 reads; and 3,875,453 Illumina *Yeast Saccharomyces cerevisiae* W303-K6001 reads. For the reference sets, we randomly chose ten genes with various lengths from both genomes, each of which has over 4000 genes: *thrA*, *thiQ*, *cydD*, *ycgB*, *dhaR*, *alkA*, *yfgF*, *yphE*, *mseS*, *parC* (*E-coli*, GenBank accession number: U00096.2); *YNL095C*, *YNL094W*, *YNL093W*, *YNL092W*,

YNL091W, YNL090W, YNL089C, YNL088W, YNL087W, YNL085W (Yeast, GenBank accession number: AF458977.1).

We estimate the number of reads found by SlopMap with various thresholds and kmers in order to:

- Estimate the effect of threshold to the number of reads found by SlopMap. Suggest the values of t and k for a solution close to optimal.
- Estimate the effect of various values of d (distance between two consecutive kmers) to the number of reads found by SlopMap against various threshold values. Determine the range of optimal values for d .
- Compare the number of reads found by SlopMap to the number of reads found by other tools. In particular, to answer a question: what are the threshold values when read set found by other tools are still subsets of reads found by SlopMap.

All tests were performed on a Linux server with Dual-Core AMD Opteron 8216 2.4 GHz processors (32 processors total) and 1 TB of shared memory and a laptop with single Intel Core i3 processor (four cores) with 4GB of memory.

4 Results and discussion

As a result of validation and testing, we calculated the number of found reads against various threshold value and kmer sizes. Also the plots that shows number of reads against various distances between two consecutive kmers were built. And finally, Venn diagrams for kmer sizes 10, 12 and 15 were built. These diagrams shows the comparison between SlopMap and other tools. We also provided recommendation for optimal values of parameters k and d .

4.1 Number of reads found by SlopMap against different threshold values and kmer lengths

In Figure 3 the plot shows the number of reads found by SlopMap against different kmer sizes and threshold values. From this plot its easy to see that for kmer values k of 7, 8, 9 along with threshold value less than 0.1 (10% of similarity), there is large amount of reads found. These values of k should be used carefully and only in such situations when threshold value t is above 0.1. Values when $t = 0.1$ are not suitable for kmer lengths less than 9 bases and can not be considered for common practical applications.

Those values of k that are within a range 10...15 bases can be considered as optimal values. For Roche 454 libraries we would not recommend to use those kmer sizes that are above 20, since with these kmer sizes it is very difficult to achieve enough coverage. This may not be true for Illumina reads due to their high quality. For Illumina libraries, it is possible to use k above 20 bases.

One important remark can be made in concluding this subsection: due to erroneous sequences (base-call errors, misleading homo-polymers) and such artefacts as adapters (TruSeq adapters for Illumina and RLMIDs along with Badapter in Roche 454 libraries) within a read, it is not always possible to achieve enough coverage on the output due to noise introduced to sequences by errors and/or adapters. In such cases when query library contain many errors it is highly recommend to conduct quality trimming (also called "quality cleaning") in order to remove low-quality bases or apply lower threshold values.

4.2 Number of reads found by SlopMap against various threshold values and different distances between two consecutive kmers

In order to examine how different values of parameter d (a distance between two consecutive kmer in a read) impact on the number of reads found by SlopMap and propose the optimal value for d , we provided a set of tests with kmer size $k = 11$ and threshold values $t = 0.1 \dots 1.0$ (i.e. from 10% to 100% similarity). This is shown in Figure 4. As expected, the number of reads degrades while t raises. Our recommendations for the parameter d value to be no more than kmer length. Otherwise there may not be sufficient coverage. The values for the d from 1 to 5 are optimal for $k = 11$ bases, since they give number of reads significantly higher than other tools within $t = 0.5$ ($t = 0.5$ is set by default in SlopMap).

4.3 Sensitivity test: comparison the number of reads reported by SlopMap and other tools

Larger gaps are those when the part of the transcriptome aligns to one part of the read and another part aligns close to the first part but with a gap of length less than 50 bases. Such gaps are usually well recognized by widely used aligners. From the other side, alignment tools (except Blat) unable to recognize those reads that partially lie at the edges of gene or transcriptome (Figure 1). SlopMap can find such reads and thereby identify more similar sequences than general-purpose alignment tools. In order to examine and prove that we conducted several tests in order to compare the reads found by SlopMap to reads found by other tools (BWA, Blat, Bowtie2). We are interested what is the cut-off point when reads found by alignment tools are still subsets of reads found by SlopMap. We can roughly say that the set of reads reported by one application is a subset of reads reported by another application if there is more 95% overlap between these two sets. For two kmer lengths (10 and 15 bases) and threshold values (0.1...1 with step of 0.1) we computed Venn diagrams that show overlap sets reported by SlopMap and other tools. From these diagrams we conclude that for threshold values below 0.3 (30% similarity) and for both kmer sizes (10 and 15), the reads found by Blat is a subset of number of reads found by SlopMap. Threshold values when reads reported by Bowtie 2 and BWA are still subsets of number of reads reported by SlopMap are 0.7 and 0.3 (70% and 30% similarity) for kmer size 10 and 15 respectively. Results are shown in Figures 5 and 6.

When threshold value t is 100%, SlopMap does not find any reads similar to the reference genome. This is expected because reads can contain base-call errors, homopolymers and other artefacts that introduce noise into sequences.

4.4 Non-consecutive matches

When we compute the similarity of reads we do not assume that kmer matches are consecutive. Non-consecutive matching is common case and discussed below. In Figure 7, the reads and a reference are shown along with kmers shared between them. In this situation, kmer 1, 2 and 3 match corresponding kmers in a reference but in different order (non-consecutive). This is still valid situation, i.e. the read can be still considered as similar to the reference, because of variations in DNA replication.

4.5 Timing considerations

Figure 8 displays the execution time required to complete each search. We compare execution times for various threshold values of SlopMap (other parameters were set to default) with other tools.

SlopMap is faster than Bowtie2 and BWA, except Blat, which is fastest. However, in order to use output of Blat program and construct FastA or FastQ from PSL formatted file, one needs to perform additional programming work. Post-mapping conversion work is also

required for BWA, except Bowtie2, with its ability to output mapped sequences in FastQ file format.

4.6 Memory requirements

SlopMap was fast and required very little memory (2–20MB, depending on reference size). The memory consumption of SlopMap during searching grows linearly with the number of sequences and also depends on the kmer size defined by the user. For example, when the *E. coli* data set containing sequences with a mean length of 450 characters was indexed on 15-mers, 50 kB of memory was utilized for every 20,000 bases.

5 Conclusion and future work

SlopMap provides researchers with a high-throughput choice for comparing large sequence sets to find similarity. The software presented is faster than open-source well-known aligners, sensitive to low-similarity matches when desired, and flexible to allow similarity comparison for DNA (and potentially RNA and proteins). SlopMap is specifically designed for matching queries against large (more than 500,000 sequences) query sets. Three of SlopMaps beneficial attributes are its speed, flexibility and ease of use. It is capable of reporting significant hits faster than Bowtie2, BWA. In future work we are going to explore parallel realization of its algorithms. A Bloom Filter can also be considered as a data storage for kmers since it offers minimal space and access time.

We believe that the biological research community will benefit from using SlopMap.

Acknowledgments

This publication was made possible by NIH Grant P20 RR16454 from the IN-BRE Program of the National Center for Research Resources.

References

1. 454 Life Sciences et al. Standard flowgram format (sff).
2. Craig, Silverstein, et al. Sparsehash: An extremely memory-efficient hash map implementation. 2012
3. Guo Y, et al. Mitoseek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. Oxford Journal of Bioinformatics (Application Note). 2013 Mar.
4. Huang, Lulin, et al. The first illumina-based de novo transcriptome sequencing and analysis of safflower flowers. PLOS ONE. 2012 Jun.7(6)
5. Li R, et al. Soap2: an improved ultrafast tool for short read alignment. Oxford Journal of Bioinformatics. 2009
6. Kent, W James. Blat - the blast-like alignment tool. Genome Research. 2002; 12(4):656–664. [PubMed: 11932250]
7. Salzberg S, Langmead B. Fast gapped-read alignment with bowtie 2. Nature Methods. 2012; 9:357–359. [PubMed: 22388286]
8. Durbin R, Li H. Fast and accurate short read alignment with burrows-wheeler transform. Oxford Journal of Bioinformatics. 2009; 25:1754–1760.

Figure 1. Situation when red reads on the edge of the exome can be discarded by the general-purpose mapper. Green sequences can still be reported.

Figure 2. Sampling the reference with constant pre-defined kmer size k and a distance d between two consecutive kmers.

Figure 3.

Number of reads found by SlopMap versus various threshold values t and different kmer lengths k . At the left side of the plot there is large amount of reads found for $k = 9$ and $t = 0.2$. Roche 454 *E. coli* K12 W583 DNA library (621,578 reads) was used as a query library for this test.

Figure 4. Number of reads found by SlopMap against various threshold values and different distances between two consecutive kmers. Roche 454 *E. coli* K12 W583 DNA library (621,578 reads) was used as a query library for this test.

Figure 5. Number of reads reported by SlopMap and other tools depending on threshold values t and kmer sizes $k = 10$ bases. When $t = 0.3$ (the top-right figure), many of the reads reported by Blat do not overlap with SlopMap; reads reported by BWA and Bowtie2 are subsets of reads reported by SlopMap until $t = 0.7$. These tests were conducted on Roche 454 *E. coli* K12 W583 DNA library used as a query set.

Figure 6. Number of reads reported by SlopMap and other tools depending on threshold values t and kmer sizes $k = 15$ bases.

Figure 7.
Number of reads found by SlopMap versus various threshold values and different distances between two consecutive kmers

Figure 8.
Search duration. Comparison of search duration between search tools.