A Rank-Based Sequence Aligner with Applications in Phylogenetic Analysis

Dinu, Liviu P.

2014-08-18

Dinu , L P , Ionescu , R T & Tomescu , A I 2014 , ' A Rank-Based Sequence Aligner with Applications in Phylogenetic Analysis ' , PLoS One , vol. 9 , no. 8 , 104006 . https://doi.org/10.1371/journal.pone.010

http://hdl.handle.net/10138/160752
https://doi.org/10.1371/journal.pone.0104006

OPEN ACCESS Freely available online

PLOS ONE

A Rank-Based Sequence Aligner with Applications in Phylogenetic Analysis

Liviu P. Dinu[1,2·], Radu Tudor Ionescu[1*·], Alexandru I. Tomescu[3·]

1 Faculty of Mathematics and Computer Science, University of Bucharest, Bucharest, Romania, 2 Personal Genetics, Bucharest, Romania, 3 Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

Abstract

Recent tools for aligning short DNA reads have been designed to optimize the trade-off between correctness and speed. This paper introduces a method for assigning a set of short DNA reads to a reference genome, under Local Rank Distance (LRD). The rank-based aligner proposed in this work aims to improve correctness over speed. However, some indexing strategies to speed up the aligner are also investigated. The LRD aligner is improved in terms of speed by storing k-mer positions in a hash table for each read. Another improvement, that produces an approximate LRD aligner, is to consider only the positions in the reference that are likely to represent a good positional match of the read. The proposed aligner is evaluated and compared to other state of the art alignment tools in several experiments. A set of experiments are conducted to determine the precision and the recall of the proposed aligner, in the presence of contaminated reads. In another set of experiments, the proposed aligner is used to find the order, the family, or the species of a new (or unknown) organism, given only a set of short Next-Generation Sequencing DNA reads. The empirical results show that the aligner proposed in this work is highly accurate from a biological point of view. Compared to the other evaluated tools, the LRD aligner has the important advantage of being very accurate even for a very low base coverage. Thus, the LRD aligner can be considered as a good alternative to standard alignment tools, especially when the accuracy of the aligner is of high importance. Source code and UNIX binaries of the aligner are freely available for future development and use at http://lrd.herokuapp.com/aligners. The software is implemented in C++ and Java, being supported on UNIX and MS Windows.

Citation: Dinu LP, Ionescu RT, Tomescu AI (2014) A Rank-Based Sequence Aligner with Applications in Phylogenetic Analysis. PLoS ONE 9(8): e104006. doi:10.1371/journal.pone.0104006

Editor: Charles Y. Chiu, University of California, San Francisco, United States of America

Received December 11, 2013; Accepted July 9, 2014; Published August 18, 2014

Funding: The work of Alexandru I. Tomescu was supported by the Academy of Finland under grant 250345 (CoECGR). The work of Radu Tudor Ionescu was supported from the European Social Fund under Grant POSDRU/159/1.5/S/137750. The research of Liviu P. Dinu was supported by Personal Genetics. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. There are no other funding sources for this study.

Competing Interests: Liviu P. Dinu is an employee of Personal Genetics, and received salary from them. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all PLOS ONE policies on sharing data and materials.

* Email: raducu.ionescu@gmail.com

· These authors contributed equally to this work.

Introduction

Novel high-throughput sequencing technologies generate up to several millions of short DNA reads (30 to 400 nucleotides long) from random locations in the genome. Putting together these reads into a coherent whole is a significant computational challenge. The first and most expensive step of this process is aligning each read to a known reference genome. Recently, many tools designed to align short reads have been proposed [1]. Sequence alignment tools are designed to optimize the trade-off between correctness and speed, usually sacrificing correctness over speed. This leaves room for new tools for sequence alignment that can better satisfy one of (or both) the two needs, namely efficiency and accuracy. With broad applications from phylogenetic analysis to finding motifs or common patterns in a set of given DNA sequences, new alignment tools are of great interest for the entire community of computational biology researchers.

This paper proposes a method for assigning a set of short DNA reads to a reference genome, under Local Rank Distance (LRD) [2]. Local Rank Distance is an extension of rank distance [3] that is designed to work on overlapping k-mers instead of single characters as rank distance. Despite the fact that LRD was only recently introduced, it has already demonstrated its performance in phylogenetic analysis [2] and native language identification [4].

The rank-based sequence aligner works as follows. Given a set of reads that need to be aligned against a reference genome, the aligner determines the position of each read in the reference genome that gives the minimum Local Rank Distance. The proposed aligner will be referred to as the LRD aligner through the rest of this paper. Some strategies of optimizing the search for the best positions of reads are also proposed and investigated. The LRD aligner is improved in terms of speed by storing k-mer positions in a hash table for each read. An approximate LRD aligner that works even faster is obtained through the following strategy. The approximate aligner considers only the positions in the reference that are likely to give the minimum distance, by previously counting the number of k-mers from the read that can be found at every position in the reference.

The LRD sequence aligner is designed to work with genomic data produced by Next-Generation Sequencing technologies. These high-throughput technologies are able to produce up to 200 million DNA reads of length between 30 and 400 base pairs in a single experiment. Despite this abundance of reads, their short

length makes the problem of assembling them into the originating genome a difficult one in practice. Therefore, methods for finding the class, the order, the family or even the species of an unknown organism, given only a set of short Next-Generation Sequencing DNA reads originating from its genome, are of interest. A method that can be used to solve this phylogenetic analysis task is proposed in this work. The method works as follows: given a collection R of short DNA reads, and a collection G of genomes, it finds the genome G [G that gives a minimum score. This method serves two purposes. First, the method can be used to determine the place of an individual in a phylogenetic tree, by finding the most similar organism in the phylogenetic tree. This can be achieved by using only a set of short DNA reads originating from the genome of the new individual. Second, the method is used to evaluate the performance level of the rank-based aligner and to compare it with other state of the art alignment tools, such as BWA [5], BOWTIE [6], or BLAST [7]. Experimental results on simulated reads were obtained under two scenarios: low and high error rate. In the former scenario, all the aligners besides BWA have full precision. In the latter scenario, the LRD aligner is the only one that attains full precision. It seems that the LRD aligner gives the most accurate results, while being more computationally expensive than the other aligners.

A set of experiments are conducted to determine the precision and the recall of the proposed LRD aligner, in the presence of contaminated reads. The task is to align reads sampled from several mammals on the human mitochondrial DNA sequence genome. The goal is to maximize the number of aligned reads sampled from the human genome (true positives), and to minimize the number of aligned reads sampled from the other mammals (false positives). Again, the LRD aligner seems to have the best performance, followed closely by BOWTIE and BLAST.

The proposed aligner is also tested on three human vibrio pathogens with results that point towards the same conclusion of [8,9]. In all the experiments presented in this work, the rank-based aligner shows results that are better than the state of the art alignment tools, in terms of accuracy. The results obtained in this work can be considered as a strong argument in favor of using rank-based distance measures for computational biology tasks, in order to obtain results that are more accurate from a biological point of view.

It is important to point out that the main focus of the experiments is on the alignment accuracy of the aligner based on LRD. Therefore, the simple strategy of assigning each read to the genomic sequence with the best LRD distance was used. However, in other biological problems, these alignments can be fed to other more elaborate methods. For example, in profiling bacterial species from a metagenomics sample, various tools, such as the MG-RAST server [10], MEGAN [11] and metaBEETL [12], align the reads to a reference taxonomy, but report as hit the Lowest Common Ancestor node of a set of significant hits in this taxonomic tree.

Related Work

Similarity Measures Between Genomes. Since most DNA variations between organisms of the same species consist of point mutations like single nucleotide polymorphisms, or small insertions or deletions, edit distance is the standard string measure in many biomedical analyses, such as the detection of genomic variation, genome assembly [13], identification and quantification of RNA transcripts [14–16], identification of transcription factor binding sites [17], or methylation patterns [18].

In the case of genomic sequences coming from different related species, other mutations are present, such as reversals [19], transpositions [20], translocations [21], fissions and fusions [22]. For this reason, there have been a series of different proposals of similarity between entire genomes, including rearrangement distance [23], k-break rearrangements [24], edit distance with block operations [25].

Some of the other popular distance measures for recent computational biology techniques are the Hamming distance [26,27] and the Kendall-tau distance [28], among others [29]. Rank distance [3] is another such measure of similarity, having low computational complexity, but high significance in phylogenetic analysis [30,31] and in finding common patterns in DNA sequences [32].

Sequence Aligners. One of the most widely used computational biology programs is BLAST [7]. Compared to the previously developed techniques based on dynamic programming [33], BLAST increases the speed of alignment by reducing the search space. An interesting remark is that BLAST calculates the statistical significance for each sequence alignment result.

While BLAST remains an essential tool for biologists, the vast amount of data produced by the high-throughput sequencing technologies led to the development of faster and more accurate sequence aligners. Recently, many tools designed to align short reads have been proposed [1]. The main efforts in the design of such tools are on improving speed and correctness. Fast tools are needed to keep the pace with data production, while the number of correctly placed reads is maximized. Usually tools sacrifice correctness over speed, allowing only few mismatches between the reads and the reference genome. Tools that optimize such trade-off are BOWTIE [6] and BWA [5]. Both the BWA and the BOWTIE aligners work under the edit distance, and they use the Burrows-Wheeler Transform to efficiently align short reads against a large reference sequence, allowing mismatches and gaps. The BOWTIE2 aligner [34] combines the full-text minute index with the flexibility of hardware-accelerated dynamic programming algorithms to achieve both speed and accuracy.

The BFAST [35] tool moves towards favoring correctness over speed, allowing alignments with a high number of mismatches and indels. Another accurate tool able to align reads in the presence of extensive polymorphisms, high error rates and small indels, is rNA [27]. The experiments performed in [27] give an idea about the different approaches of such tools for optimizing the trade-off between correctness and speed. For example, in one experiment BWA is 100 times faster than BFAST, while losing about 8:00% in terms of accuracy.

Results

Data Sets

To evaluate the aligners proposed in this work, several experiments are conducted on two data sets of genome sequences. The first data set contains mitochondrial DNA sequence genomes of 20 mammals. The genomes are available for download in the EMBL database (http://www.ebi.ac.uk/ena/) using the accession numbers given in Table 1. They belong to the following biological orders: Primates, Perissodactylae, Cetartiodactylae, Rodentia, Carnivora.

Mitochondrial DNA (mtDNA) is the DNA located in organelles called mitochondria. The DNA sequence of mtDNA has been determined from a large number of organisms and individuals, and the comparison of those DNA sequences represents a mainstay of phylogenetics, in that it allows biologists to elucidate the evolutionary relationships among species. In mammals, each double-stranded circular mtDNA molecule consists of 15,000 to 17,000 base pairs. DNA from two individuals of the same species

Table 1. The 20 mammals from the EMBL database used in the phylogenetic experiments. The accession number is given on the last column.

Mammal	Latin Name	Accession No.
human	Homo sapiens	V00662
common chimpanzee	Pan troglodytes	D38116
pigmy chimpanzee	Pan paniscus	D38113
gorilla	Gorilla gorilla	D38114
orangutan	Pongo pygmaeus	D38115
Sumatran orangutan	Pongo pygmaeus abelii	X97707
gibbon	Hylobates lar	X99256
horse	Equus caballus	X79547
donkey	Equus asinus	X97337
Indian rhinoceros	Rhinoceros unicornis	X97336
white rhinoceros	Ceratotherium simum	Y07726
harbor seal	Phoca vitulina	X63726
gray seal	Halichoerus grypus	X72004
cat	Felis catus	U20753
fin whale	Balaenoptera physalus	X61145
blue whale	Balaenoptera musculus	X72204
cow	Bos taurus	V00654
sheep	Ovis aries	AF010406
rat	Rattus norvegicus	X14848
mouse	Mus musculus	V00711

doi:10.1371/journal.pone.0104006.t001

differs by only 0:1%. This means, for example, that mtDNA from two different humans differs by less than 20 base pairs. Because this small difference cannot affect the study, the experiments are conducted using a single mtDNA sequence for each mammal.

The second data set contains chromosomal DNA sequence genomes of three vibrio pathogens available in the NCBI database (http://www.ncbi.nlm.nih.gov): Vibrio vulnificus YJ106, Vibrio parahaemolyticus RIMD 2210633, and Vibrio cholerae El Tor N16961. The genomes of these three organisms consist of two circular chromosomes. Additional information about these chromosomes, including accession number and size (given in Megabase pairs), is given in Table 2. The genomic sequences of these vibrio species have been revealed by different studies [9,36,37]. Several studies report that Vibrio vulnificus shares morphological and biochemical characteristics with other human vibrio pathogens, including Vibrio cholerae and Vibrio parahaemolyticus [8,9].

Alignment in the Presence of Contaminated Reads

In this experiment, reads sampled from the genomes of several mammals are aligned on the human mtDNA sequence genome. The reads were simulated with the wgsim tool [38], using the default parameters. More precisely, the reads were generated using an error rate of 0:02, a mutation rate of 0:001, a fraction of indels of 0:15 (out of the total number of mutations) and a probability of extending an indel of 0:30.

The LRD aligner is compared to the BWA, the BOWTIE2 and the BLAST aligners, under two different scenarios. In the first scenario, 10,000 contaminated reads are sampled from the orangutan genome. In the second scenario, 50,000 contaminated reads are sampled from 5 mammals, namely the orangutan, the blue whale, the harbor seal, the donkey, and the house mouse. There are actually 10,000 reads sampled from each of the 5 mammals. In both scenarios 10,000 reads simulated from the human genome are included. The simulated reads are always 100

Table 2. The genomic sequence information of three vibrio pathogens consisting of two circular chromosomes.

Species	Chromosome	Accession No.	Size (Mbp)
V. vulnificus YJ016	I (VV1)	NC_005139	3:4
V. vulnificus YJ016	II (VV2)	NC_005140	1:9
V. parahaemolyticus RIMD 2210633	I (VP1)	NC_004603	3:3
V. parahaemolyticus RIMD 2210633	II (VP2)	NC_004605	1:9
V. cholerae El Tor N16961	I (VC1)	NC_002505	3:0
V. cholerae El Tor N16961	II (VC2)	NC_002506	1:0

doi:10.1371/journal.pone.0104006.t002

bases long. The goal is to maximize the number of aligned reads sampled from the human genome (true positives), and to minimize the number of aligned reads from the other mammals (false positives). Unlike the other experiments presented in this paper, reverse complement reads were not included in this experiment. However, it is important to mention that the aligners are dealing with a hard task, since the contaminated reads were sampled only from organisms that are in the same class as the human. It may be that contaminated reads from other species that are not in the Mammalia class (such as viruses, for example) can be identified and discarded more easily.

The parameters of the aligners were adjusted as described next. For the BOWTIE2 aligner, two variants are evaluated. The first one uses the *local* and the *very-sensitive-local* options. The second variant uses the *end-to-end* and the *very-sensitive* options. For the BLAST aligner, the *megablast* option is used. Two variants of the LRD aligner based on 3-mers and a maximum offset between paired 3-mers of 36 are also evaluated. One is based on the exact search algorithm, while the other one uses the approximate algorithm based on hash tables that runs much faster.

To evaluate and compare the aligners, the precision and recall curve is used. Note that the precision is given by the proportion of aligned reads that are positive, while the recall is given by the proportion of true positive reads that are aligned. In order to obtain the precision-recall curve for each aligner, the idea is to vary the threshold that gives the maximum distance allowed for an aligned read. In the case of the BWA and the BOWTIE aligners, the edit distance threshold takes values from 0 to 30. The score of the BLAST aligner ranges from 185 to 100. The LRD threshold takes values from 50 to 600, for both variants of the LRD aligner. Higher precision is obtained for lower distance thresholds, while higher recall is obtained for higher distance thresholds. The only aligner that works the other way around, and gives higher precision for higher scores, and higher recall for lower scores, is the BLAST aligner.

Several statistical measures, such as the Area Under the ROC Curve (AUC), the F_1 measure, and the F_2 measure, are also presented in order to better compare the aligners. The ROC curve plots the fraction of true positive reads versus the fraction of false positive reads, at various threshold settings. The AUC score represents the area under the ROC curve. The F_1 measure (also known as the F_1 score) can be interpreted as a weighted average of the precision and recall at a certain distance threshold. The F_2 measure is similar to the F_1 measure, only that it weights recall higher than precision. For each aligner, the higher F_1 and F_2 scores can indicate the thresholds that give a good trade-off between precision and recall. The F_b measure is computed as follows:

$$F_b = (1 + b^2) \frac{\text{precision} \cdot \text{recall}}{b^2 \cdot \text{precision} + \text{recall}}. \quad (1)$$

The F_1 and the F_2 scores are immediately obtained from Equation 1, by replacing b with 1 and 2, respectively.

Human versus Orangutan Experiment. In this experiment, there are 20,000 reads to be aligned on the human mtDNA sequence. Half of them are sampled from the same human mitochondrial genome, while the other half are sampled from the orangutan mitochondrial genome. Thus, the contamination rate is 50%.

The precision-recall curves of the BWA, the BOWTIE, and the BLAST aligners together with the precision-recall curves of the two variants of the LRD aligner are presented in Figure 1. By analyzing Figure 1, it can be observed that the aligners obtain roughly similar results in terms of precision and recall. To better assess the performance of the evaluated aligners, the AUC measure and the best F_1 and F_2 scores for each aligner are presented in Table 3. In terms of the AUC, the BOWTIE and the LRD aligners attain the best results, while the other aligners fall behind. In terms of the F_1 measure, the BOWTIE aligner seems to be slightly better than the LRD aligner, while in terms of the F_2 measure, the LRD aligner achieves the best score, followed closely by the BOWTIE aligner. The BLAST aligner comes in third place after the LRD and the BOWTIE aligners. The results of the BWA aligner are also not too far from the other top scoring aligners. The results presented in Figure 1 indicate that all the aligners