A method for reconstructing DNA using a list of smaller DNA sequences

Sebastián Patiño Barrientos
Universidad Eafit
Colombia
spatino6@eafit.edu.co

Luis Miguel Arroyave Quiñones
Universidad Eafit
Colombia
larroy13@eafit.edu.co

Juan José Jaramillo Castaño
Universidad Eafit
Colombia
jujara40@eafit.edu.co

Mauricio Toro
Universidad Eafit
Colombia
mtorobe@eafit.edu.co

ABSTRACT

The idea of making optimal and each time better solutions to the problems that affect the humanity has been at the top goals of our generation, especially in the fields of computations and development of algorithms, as a record of that, we have the Needleman–Wunsch algorithm one of the first DP (dynamic programming) algorithms used in bioinformatics and computational genomics. Having said that, we can present the goal of this paper which is present an algorithmic solution to the problem of DNA sequence reconstruction using a collection of fragments and determine to which specie is it from.

1 INTRODUCTION

How do antibiotics work? Where do virus come from? What is a mutation? Can we clone people? The human kind had been studying biology for a long time but without the help of technology the revolution of the field could have been impossible. Thanks to the convergence of computation and biology, we have improved the genetic analysis discovering patterns in the genomics sequences.

The reconstruction of DNA expressions is still a significant problem in biology, and to do so firstly is needed to find all the overlapping sequences, then we can proceed to the rearrangement and reconstruction of the DNA, in order to identify from what specie do the reconstructed sequence come from.

To that end, we've come to a solution using a de Bruijn graph to represent a sequence in terms of its k-mer components and with the Eulerian path we are able find the overlapping sequences.

2 THE PROBLEM

Every organism or living being is unique, all of them having several differences from each other, but they also have some things in common that allow us to studying them better by classifying them in certain groups, so to identify them there are certain genome sequences that define the specie. The main purpose of our project is from a collections of DNA fragments resequencing and assembling them to determine the belonging specie of the organism from the given DNA sequences.

Universidad Eafit, May 2017, Antioquia, Colombia
2017. ACM ISBN xxx-xxxx-xx-xxx/xx/xx…$0.00
https://doi.org/xx.xxx/xxx_x

3 ALGORITHM

3.1 Usage

The program is executed by command line, the first step is to insert the strings and it'll keep reading until EOF (end of file).

3.2 Average time and memory

The average execution time is **30** seconds.

The memory used is **15.6** mb.

3.3 Complexity analysis

Our solution takes a time $O(N \times |S|+|E|+|G|)$.

- Being N the number of strings received.
- Being $|S|$ the average string length.
- Being $|E|$ the final amount of edges in the graph.
- Being $|G|$ the final length of the resulting genome.

4 RELATED WORKS

4.1 Longest common substring problem

The longest common substring problem is a computer science problem in which the longest substring that two or more strings have in common needs to be found.

There are many solutions for this problem. The simpler one consists in comparing one by one all of the substrings from any of the given strings to check if it is a substring of any of the other strings and then determine which one of these is the longest.

Another solution is parsing the strings with a suffix tree and *"finding the deepest internal nodes which have leaf nodes from all the strings in the subtree below it."*

This problem can also be solved by using DP (Dynamic programming), finding which of the common suffix for all pairs of prefixes of the strings has the maximum length. The maximal of these longest common suffixes of possible prefixes must be the longest common substrings.

4.2 Sequence alignment

"In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences" This problem can be solved by the application of a variety of algorithms, so this problem has no

specific solution since it can be solved by the mixture of methods and optimizations of these. An example of these algorithms and methods are DP, probabilistic or heuristic methods, etc.

4.3 String searching for literature

We usually need to find a specific word in an essay, chronicle or any other document with a big amount of text, that's why engineers always try to develop every time an efficient way to satisfy this need. Here we have an example of a searching algorithm used usually for this kind of problems. This is the Boyer–Moore string search algorithm that was developed by Robert S. Boyer and J Strother Moore at the end of the seventies,4 this algorithm doesn't search every character of the string, but uses a set of rules in order to minimize the time spend on the search.

REFERENCES

[1] 2017. Boyer–Moore string search algorithm. (Jan 2017). https://en.wikipedia.org/wiki/Boyer%E2%80%93Moore_string_search_algorithm#cite_note-1

[2] 2017. Dynamic Programming | Set 29 (Longest Common Substring). (Mar 2017). http://www.geeksforgeeks.org/longest-common-substring/

[3] 2017. Longest common substring problem. (Feb 2017). https://en.wikipedia.org/wiki/Longest_common_substring_problem

[4] 2017. Sequence alignment. (Apr 2017). https://en.wikipedia.org/wiki/Sequence_alignment#Alignment_methods

[5] Yang Chen and Jinglu Hu. 2011. Accurate Reconstruction for DNA Sequencing by Hybridization Based on a Constructive Heuristic. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 8, 4 (July 2011), 1134–1140. https://doi.org/10.1109/TCBB.2010.89

[6] Patrick Flick, Chirag Jain, Tony Pan, and Srinivas Aluru. 2015. A Parallel Connectivity Algorithm for De Bruijn Graphs in Metagenomic Applications. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '15)*. ACM, New York, NY, USA, Article 15, 11 pages. https://doi.org/10.1145/2807591.2807619

[7] Heinrich Matzinger and Angelica Pachon Pinzon. 2011. {DNA} approach to scenery reconstruction. *Stochastic Processes and their Applications* 121, 11 (2011), 2455 – 2473. https://doi.org/10.1016/j.spa.2011.04.010