ISSN 2348-313X (Print)
International Journal of Life Sciences Research ISSN 2348-3148 (online)
Vol. 8, Issue 2, pp: (11-13), Month: April - June 2020, Available at: www.researchpublish.com

Longest Path Algorithm and using it to Find Unknown Genome via Shotgun Sequencing

Lee, Kin Seng

Singapore, Malaysia

lksark@hotmail.com

Abstract: **This article talks about Long Path Algorithm and the potential of using it to identify unknown genome via Shotgun Sequencing method. The aim of finding the longest possible path in a graph is to find a path that can cover the longest possible travelling distance in a graph, from a known starting node to an unknown ending node. Inside the longest path, individual nodes can either re-occur or not re-occur.**

Keywords: **Longest Path, Shotgun Sequencing, find unknown genome, Overlap Graph.**

I. INTRODUCTION

In contrast to shortest path algorithm, there is longest path algorithm. Longest path algorithm is to find the possible longest path in a graph by finding the longest possible travelling distance in a graph. Therefore, we may able to use Longest Path Algorithm to identify an unknown genome via Shotgun Sequencing method.

This paper briefly talks about the working principle of the Longest Path Algorithm, without experimental figures. More information and Longest Path Algorithm coding please visit my website: http://www.algoonline.net/LongestPath/Longest_Path_algorithm.htm.

II. BRIEF DESCRIPTION OF LONGEST PATH ALGORITHM WORKING PRINCIPLE

The aim of finding the longest possible path in a graph is to find a path that can cover the longest possible travelling distance in a graph, from a known starting node to an unknown ending node. Inside a graph's longest path, individual nodes can be either occur once or multiple times. The coding between these two are only minor differences.

Longest Path Algorithm coding is somewhat similar to Shortest Path Algorithm. However Longest Path Algorithm occasionally have path travelling in closed loops problem, causing total travelling distance become infinity and program freeze. Therefore, longest path programming will need to have additional coding to check do targeted destination node re-occur inside current node's longest path. The targeted destination node will not again add inside current node's longest path if found re-occurred. Shortest Path problem on the other hands will not have such dilemmas.

Similar to shortest path, the sub-path of the longest path itself are also the longest path.

The closed loops could be nodes next to each other; the loops could be numerous numbers of nodes that linked together in big circles. The later one is more difficult to be spotted by eye.

A. Longest Path with repetitive nodes

In some of the longest path problems, the individual nodes could be repetitive. Nodes are not distinct in the longest path. Some of the nodes occur multiple times in the longest path. One of the examples is genome sequences.

When the same nodes reoccur twice in the path, it forms a closed loop path. Program can get caught inside the closed loop, keep-on circulating and getting infinity travelling distance, creating program logic error. Hence, our program needs to put these closed loop paths into separate queues to avoid program freeze.

ISSN 2348-313X (Print)
International Journal of Life Sciences Research ISSN 2348-3148 (online)
Vol. 8, Issue 2, pp: (11-13), Month: April - June 2020, Available at: www.researchpublish.com

When same nodes occur three times in the path, we have 2 closed loops. In such case, we need to manually identify which closed loops should situated in first position; which closed loops at the subsequent position; alternately we can remove both closed loops out of the longest path; or only keeping one out of two.

Hypothetically, looping paths can reoccur multiple times along the route; they can be confusingly intertwined with each other's; they can be jumping back and forth inside the longest path.

These looping nodes can be either included or excluded from the actual longest path we are looking for. Moreover, the last node in the longest path might be situated inside one of these closed loop paths.

Hence, we need to manually verify all these closed loop paths. To determine whether we ought to keep them inside the longest path.

The Longest Path Algorithm with re-occurring nodes are derived from Longest Path Algorithm without re-occurring nodes. All nodes will have an additional new queue to record any occurrence of closed loop paths to this node.

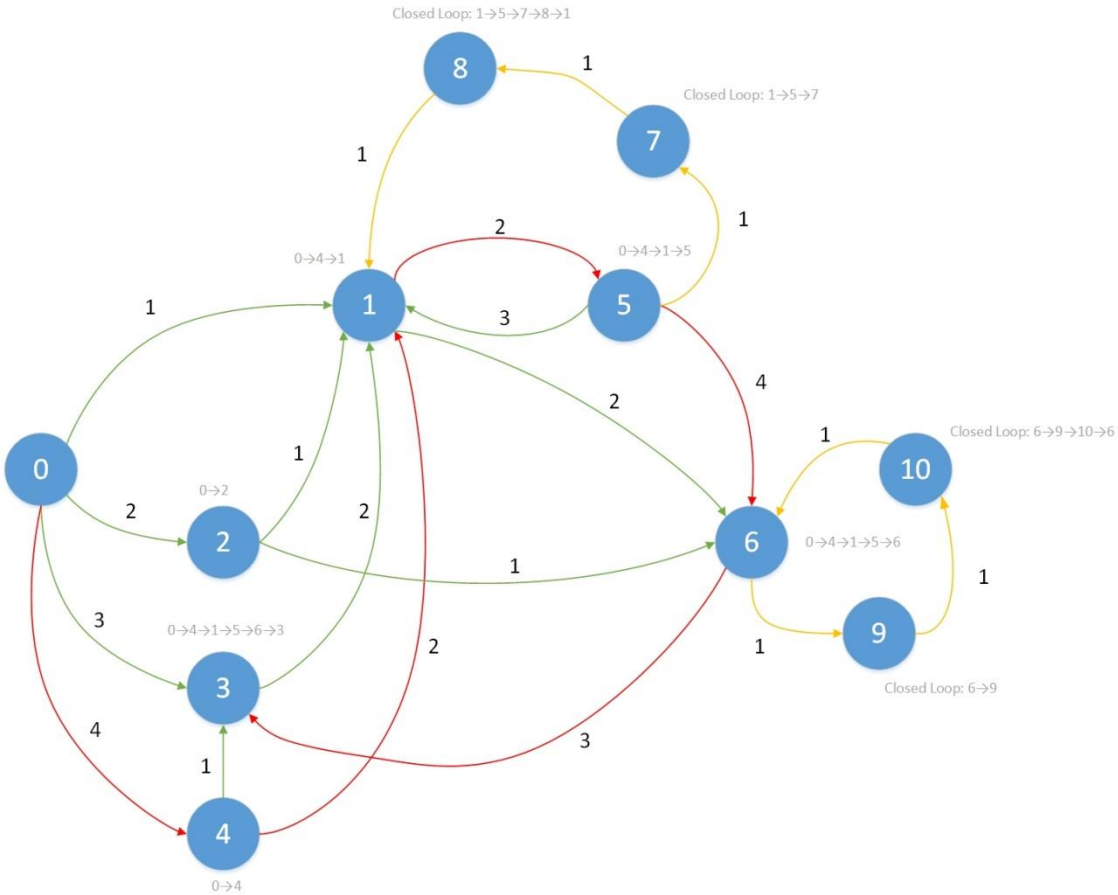

Fig. 1: Example of Longest Path graph: Red colour path are the longest path. Orange colour paths are the closed loops

B. Longest Path Algorithm to be used in identifying unknown Genomes

In the Longest Path Algorithm, computer will compute to get the longest possible path in a graph, cover as much nodes as possible. This working principle might help to identify unknown genomes.

Shotgun sequencing method is used to sequence an unknown genome. In shotgun sequencing, DNA is fragmented up randomly into numerous small segments, sequenced to get reads. Multiple overlapping reads for the target DNA are acquired by performing several rounds of fragmentation and sequencing. Machine will save the reads result into 'FASTQ' text-based format file. Computer programs then use these different reads to produce Overlap graph via setting fix characters length of reads segment as the *k*-mer. And then computer compute to get the longest path from the Overlap graph to obtain a continuous sequence. [2]

ISSN 2348-313X (Print)
International Journal of Life Sciences Research ISSN 2348-3148 (online)
Vol. 8, Issue 2, pp: (11-13), Month: April - June 2020, Available at: www.researchpublish.com

By making 'k-mer' nodes' substring length longer characters, we can reduce the occurrences of closed loops, reduce total number of nodes, making the graph smaller and lesser computing efforts. But on other hands, if there are sequencer machine errors on individual reads, the longest path result we are getting might not be accurate.

The actual continuous genome sequence is likely to be the longest sequence among all the possible sequences. Thus, we may use the Longest Path Algorithm to identify unknown genome sequences.

However, there are possibilities that actual genome sequences are not the longest path. These might occur if targeted genome sequence is not long enough; or there are sequencer machine errors. Hence, manual checking should always be in place.

Hypothetically, the more sequences data we collected and add into nodes' weight, the more likely we can derive our correct whole genome sequence while mitigate the impact of sequencer machine errors.

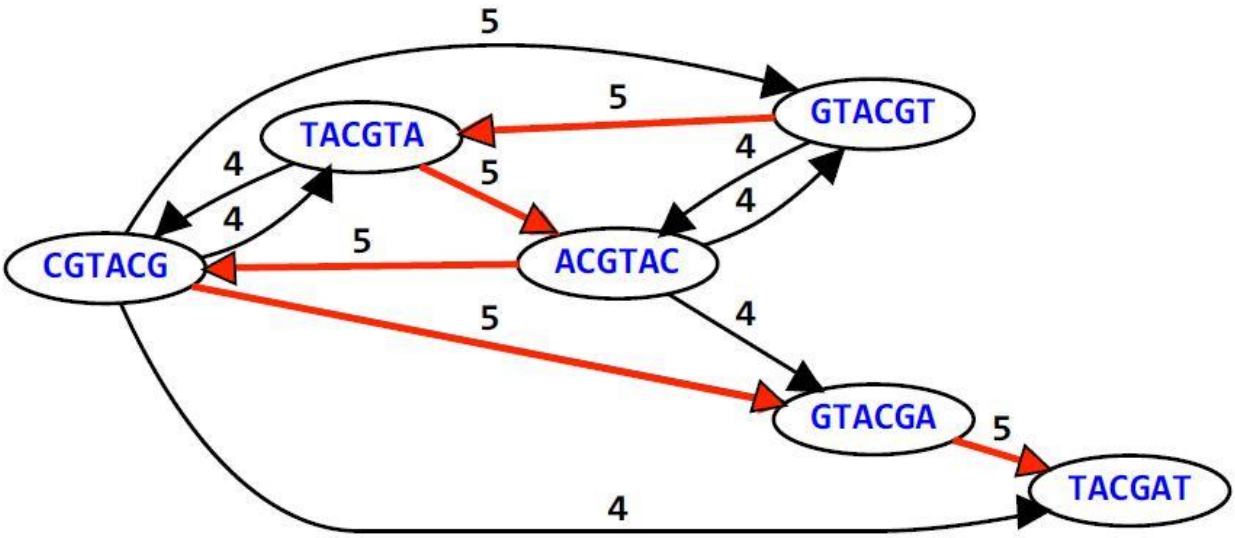

Fig. 2: Short genome's Overlap Graph by Dr. Ben Langmead and Jacob Pritt from Johns Hopkins University [1]

Nodes: all 6-mers from "GTACGTACGAT"

Edges: only select overlaps of length ≥4

III. CONCLUSION

Longest Path Algorithm has limited usages. Not many problems required Longest Path Algorithm. One of the usages might be in game engine, where the character urged to acquire as much loots (nodes) as possible. Shortest Path Algorithm in contrast is much handier, being implemented in multiple applications such as like traffic navigation, network routing etc.

In this paper, hypothesis about using Longest Path Algorithm to identify unknown genome has not undergone practical experiments. Actual experiments are needed to conclude the case. Currently I don't have the equipment and know how to prove the case.

REFERENCES

[1] Dr. Ben Langmead and Jacob Pritt from Johns Hopkins University, online learning course website Coursera "Algorithms for DNA Sequencing", 2018, Module 3, Lecture 10

[2] Wikipedia: Shotgun sequencing, Whole genome sequencing, *k*-mer