CopperHead: A New And Simple Solution To Sequence Assembly

Omar Rosado Ramirez

Department of Computer Science, University Of Puerto Rico, Rio Piedras Campus

Abstract

In bio-informatics, Sequence Assembly faces few problems, the most persistent ones being the “bad read problem”, which creates “extra” reads that make the newly constructed DNA strand bad, and it's consequence, the “bad road problem”, which creates extra and erroneous strands in the graph representation of the assembled DNA strand. In this study, we create an assembler that reads a fixed strand of DNA and recreates the whole sequence assembly scenario with fixed errors, so we may observe the creation and assembly of the strand without change.

Objectives

In this study, our main goal and focus is to solve the bad read problem by giving it a new set of instructions that read and assemble the DNA strand in a more efficient way.

By eliminating this problem, we solve our other problem, the bad road problem(which we will not discussed in this experiment) due to it being a direct consequence of the bad read problem.

Sequence Assembly?

Imagine you have five identical books. Now, take the books and rip them appart, page by page, to pieces. Let’s take all those pieces and put them together. Finally, put the five books back together. Now try doing this with DNA strands. Yeah, cant huh? This is Sequence Assembly. In bio-technology, it is the way scientists discover DNA strands. This is possible via programs called sequencers, which separate the DNA in small pieces, called sequences, and assemblers, which do exactly that, assemble the DNA pieces back together.

What seems to be the problem?

The first problem in Sequence Assembly is the bad read problem, which creates erroneous reads in the newly formed DNA strand, which makes it “faulty” DNA.

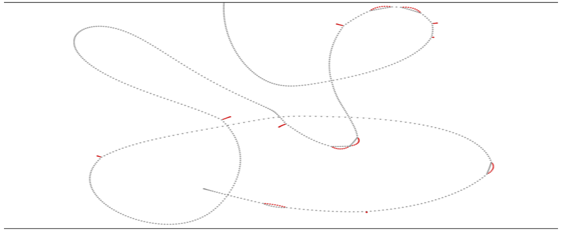
So, how do you attack this problem?

To make this problem go away, we’re implementing simple Python scripts that are governed by the NetworkX library, which controls graph-like programs. We first create an artificial DNA strand, for constant results:

1. cttaggatcctgggagttgtcagctagcatgactgactcgatcga

Then, we insert artificial errors into this strand, represented with the letter X. In the graph, the errors are represented with red dots.:

2. cttaXggatcctgggaXgttgtcaXgctagcatgactgactcgatXXcga



Finally. we sequence the strand and read each one. When an X is found, it is deleted. Since we represent the strand in the program as a list, the list element is removed and the remaining list is turned into a string and appended, creating the original strand.

3. ctta ggatcctggga gttgtca gctagcatgactgactcgat cga
4. cttaggatcctgggagttgtcagctagcatgactgactcgatcga

All of these steps together form our solution: CopperHead. Although it sounds fairly simple, it is a big step in Sequence Assembly, due to the level of simplicity and impact. The impact is that, via making artificial errors, we’re able to see how other assemblers manage and treat real ones.

Results

We now see that the manipulation of errors in a more human way is necessary to solve Sequence Assembly’s problem. Human in the sense that, the program must take the strand, “manually” take the errors out of it, and finally append it back. It is too soon to say that this solution will work for any given DNA strand, although it is still missing fundamental work and change.

Conclusion

CopperHead’s approach to the bad road problem has opened a new road for experimentation: List to DNA Conversion. Since the strand is represented as a string, we can turn it into a list and humanly manipulate it. (Playing god). In the future, the List to DNA Conversion will be applied in more extent and will hopefully solve Sequence Assembly’s problem.

Acknowledgements

The making of this study would not have been possible without my Mentor, Humberto Ortiz Zuazaga. Thank you for all your words of wisdom and your Gandalf the Gray-like personality. I would also like to thank my colleague Ricardo Lopez for approaching the problem in a different way, shedding light to this experiment. I would also like to thank every professor in the CCOM faculty for dedicating their time and love to me and the CCOM student body. Many thanks to all.

Bibliography

[1] Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman . An Eulerian Path Approach to DNA Fragment Assembly .http://www.pnas.org/content/98/17/9748 , 2001.

[2] Manfred G. Grabherr, Brian J. Hass, Moran Yassour. Trinity: Reconstructing A Full-length Transciptone Without A Genome From RNA-Seq Data. http://dx.doi.org/10.1038/nbt.1883 , 2011.