Homopolymer Compression: Multiple Alignment in the Face of Homopolymer Errors

Saksham Goel[1], Drew Beckwith[2], and Dr. Haynes[3]

[1] Undergraduate Student, Department of Computer Science and Software Engineering, Auburn University
[2] Undergraduate Student, Department of Computer Science and Software Engineering, Auburn University
[3]Assistant Professor, Department of Computer Science and Software Engineering, Auburn University

All organisms on this Earth have some kind of genetic material, with advanced ones having it in the form of 'DNA'. A DNA sequence is made of four chemical compounds, namely Adenine(A), Guanine(G), Cytosine(C), and Thymine(T) that form a double helix of a chain of base pairs.[3] Adenine only pairs up with Thymine, while Cytosine only pairs with Guanine, and vice versa. DNA sequences are important because not only they can help us solve the mysteries of evolution, but also understand and potentially fix various genetic disorders that may be caused due to random mutations, inheritance, or environmental causes such as UV rays.

Single molecule DNA sequencing is often erroneous due to the challenges of measuring DNA bases that are smaller than the wavelength of light. To get around this, PacBio circular consensus sequencing reads the same piece of DNA multiple times. From these multiple reads, a consensus sequence is determined, which gives us a close determination of what the actual sequence is--close, but not exact. These reads are most erroneous in determining how many of a particular base there is in a stretch of the same base (a homopolymer). A homopolymer stretch of a DNA sequence has only one type of base for a long period. For example, in an incredibly short DNA sequence AGGGGCTC, the GGGG part is a homopolymer. The technique that we implemented for our research aims to provide better results for DNA sequences with error-prone homopolymers. exist to enhance the educational experience, quality of the degree program, and the value to the student.

In our algorithm, we first start by compressing each DNA read into unique base pairs and the frequency of occurrence of each base pair. This can be seen in Fig. 1. Then, for these compressed DNA sequences, we gener[1]ate a consensus sequence using partial order alignment. A partial order alignment is a method of multiple sequence alignment by incrementally aligning sequences onto a growing sequence graph and using a dynamic programming backtrace to find the consensus sequence.

AGGCTCC →

A	G	C	T	C
1	2	1	1	2

Fig. 1 DNA Sequence compressed into unique base pairs and frequency of occurrence.

After this, we globally align the compressed consensus sequence to every read. A global alignment consists of the two sequences aligned, and the alignment score which reflects how well the two sequences correlate to each other. Fig. 2 showcases what a global alignment looks like.

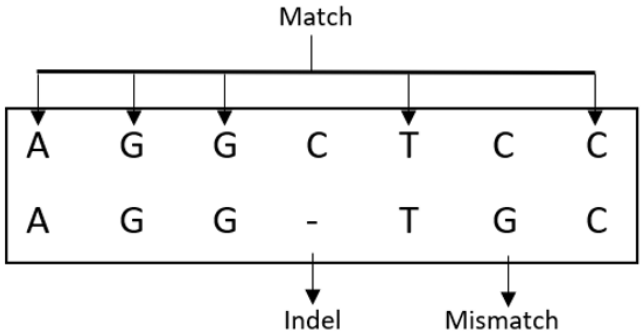

Fig. 2 Global Alignment of two sequences

Once we have all of this, it is time to expand our compressed consensus sequence using the information we have from every read (the unique base pairs and their frequencies). We iterate over every global alignment and

[1]Corresponding author: szg0112@auburn.edu

try to find the frequency of a base at that specific location (starting from index 0) and construct the expanded version of the consensus sequence piece by piece. A question that arises here is out of the different number of frequencies that we have for every read, which one do we choose? We can choose either the mean, the median, or the mode of the list of frequencies. For research purposes, we do all three and compare the results of each of them. Once the expanded sequence is ready, we can globally align it to the original sequences and find the average score of our Homopolymer Compressed consensus sequence. A higher average alignment score means that the consensus sequence better represents the raw reads.

After obtaining our expanded consensus sequences and finding the average score of each of these consensus sequences, we align all our sequences to the consensus sequence that is obtained simply by their partial order alignment (POA).[1] The final step is comparing the scores of our expanded consensus sequences to that obtained by simple POA.

For data, we started by simulating DNA sequences out of A, G, T, and C by randomly picking each of them 'n' times, where n is the sequence length. From this, we generated 'k' mutated sequences. These mutations were generated using a 0.1 probability of single nucleotide polymorphism (SNP) which are mismatches in simpler understanding and a 0.1 probability of insertions or deletions. Using 25 iterations of random sequence generations with 10 mutated DNA reads for each iteration, we took the average score for each iteration for all the consensus sequences and plotted that on the chart. As we can see in Fig. 3, the regular POA performs slightly better on average than all our expanded sequences.

Fig. 3 Results from Randomly Generated Base Pairs

Since we want to target homopolymer sequences, we changed our data simulation such that it would create mutations based on the Poisson distribution of the length of individual homopolymers in the original sequence to better simulate fact that errors are higher in homopolymer sequences. The results are shown in Fig. 4, and we can see that our algorithm with median or mean as the average metric tends to perform much better than the regular POA

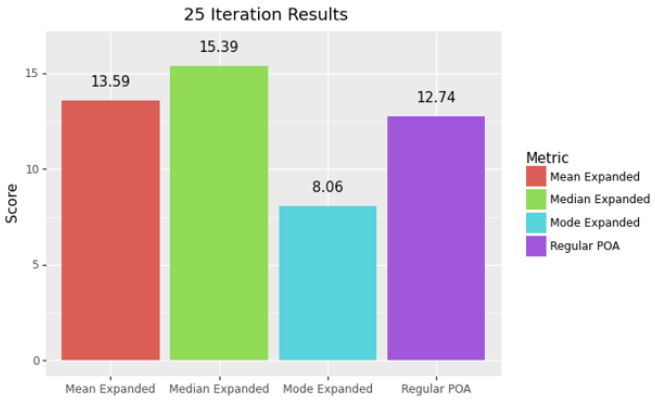

Fig. 4 Results from Data Generated using Poisson distribution of Homopolymers.

Statement of Research Advisor

This work demonstrates a proof of concept that iterative multiple sequence alignment (MSA) done in homopolymer compressed space can improve the consensus sequence quality in sequences that have higher error rates in homopolymer sequences, such as the circular consensus sequencing data from Pacbio. We are now applying this to more sophisticated Pacbio simulated data as well as real sequencing data for which we have high quality ground truth. This work may contribute to single molecule somatic variant calling to improve cancer detection and research. This work was done predominantly by Saksham and Drew.

- Dr. Haynes Heaton, Computer Science, Samuel Ginn College of Engineering

References

[1] Gao, Y., Liu, Y., Ma, Y., Liu, B., Wang, Y., Xing, Y. (2021). abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. Bioinformatics, 37(15), 2209-2211.

[2] Stöcker, B. K., Köster, J., & Rahmann, S. (2016). SimLoRD: Simulation of Long Read Data. Bioinformatics (Oxford, England), 32(17), 2704–2706. https://doi.org/10.1093/bioinformatics/btw286

[3] Watson, J. D., & Crick, F. H. (1953). A Structure for Deoxyribose Nucleic Acid. Nature, 171(4356), 737-738

Authors Biography

Saksham Goel is a senior-year student pursuing a B.S. degree in Computer Science with a minor in Statistics at Auburn University. In this project, he is responsible for the development of data generation, as well as global and partial order alignment. He is planning to move to Redmond, WA to begin his professional career as a Software Engineer upon graduation.

Drew Beckwith is a senior-year student pursuing a B.S. degree in Computer Science with a minor in Statistics at Auburn University. In this project, he is responsible for the implementation of expansion of Homopolymer consensus sequence using various metrics. Through various classes, internships, and self-study, he has found his passion in the areas of Web Development and Data Science. He is planning to move to Birmingham, AL to begin his professional career as a Software Developer upon graduation.

Dr. Haynes Heaton is an Assistant Professor in Computer Science and Software Engineering. He is responsible for the design of the project, along with mentoring the students in the development of code used in the project. He is a Computational Biologist working on methods development on emerging genomics technologies to enable the next generation of biological research. He holds a B.S. in CS and an MD from Brown University and a Ph.D. from Cambridge University.