SplitThreader: Exploration and analysis of rearrangements in cancer genomes

Vicente Machaca Arceda

Universidad Nacional de San Agustín, Arequipa-Perú,
vmachacaa@unsa.edu.pe

Abstract. Phylogenetics analysis is a very important task in Bioinformatics, we could learn about evolution, the relation between specimens. Nevertheless, a phylogenetics tree depends on the similarity analysis performed before. This similarity analysis is based on sequence alignment methods like BLAST and CLUSTALW, but they are too slow and we need other algorithms to process similarity between sequences. In this work, we present an analysis of four alignment-free algorithms based on the image texture computed from a sequence. We compared first-order statistics, gray level co-occurrence matrix, local binary patterns, and multi-resolution local binary patterns. Moreover, we used several mapping functions for each base. Then, we compared which of these algorithms were more similar to CLUSTALW. Finally, we got that first-order statistics is the method that is more likely to CLUSTALW with the advantage of having a low computational cost.

Keywords: Similarity analysis, Phylogentics trees, alignment-free methods, image textures.

1 Introduction

Genomics data has growth up exponentially, for example in 2009, they reached about 0.8 ZB, moreover in 2020 they reached about 40 ZB [1]. Furthermore, cancer related data are generated from: gene expression data (Microarray), NGS data, protein–protein interaction (PPI), pathway annotation data y gene ontology (GO). These data are important for research in cancer diagnosis and treatment. Big data resources allow researchers to observe large retrospective, and heterogeneous data of cancer patients [2].

For instance, the human genome is made of approximately 3.2 billions bp of DNA [3]. The HIV-1 genome is made of 20k bp of DNA, meanwhile the COVID-19 is made of 32k bp [4]. Additionally, there are approximately 19000 to 25000 genes (no one knows for sure) [3]. Finally, human genes have dozens of introns, each of which can be tens of thousands of nucleotides. Distinguishing exons from introns and other forms of non-coding DNA is challenging [3]. This lack of information, makes difficult the research in cancer genomics.

Moreover, genomic instability is one of the hallmarks of cancer [5,6], resulting in a widespread copy number changes, structural variants and chromosome-scale rearrangements [7]. Furthermore, copy number variants and gene fusions are common drivers in cancer [8,9]. In this context, it is very important to detect these structural variants, but the available algorithms for identifying gene fusions do not have perfect specificity (false positive rate) and they require a joint analysis of genomic and transcriptomic data. Moreover, rearrangements variants are difficult to study because of the sheer complexity of rearrangements, which often include adjacencies between distant regions of a chromosome or even between unrelated chromosomes [7].

In this work, we reviewed and replicated the tool SplitThreader [7]. It is an open source interactive web application for analysis and visualization of genomics rearrangements and copy number variation in cancer genomes.

2 Concepts

In this section, we present the most relevant concepts related to bioinformatics and cancer genomics.

2.1 Genomics data

"DNA is abbreviation of deoxyribonucleic acid, organic chemical of complex molecular structure that is found in all prokaryotic and eukaryotic cells and in many viruses. DNA codes genetic information for the transmission of inherited traits" [10]. Moreover, the term genomics is used to refer the sum total of DNA in cells [3]. For instance, in Figure 1, we present a piece of COVID-19 DNA (genome) in FASTA format. DNA data is just a string of four characters: A, C, G, T that represent the nitrogen bases Adenine, Cytosine, Guanine and Thymine respectively. Unfortunately, this data is not 100% accurate and for a single human genome, it could reaches about 4 GB (3.2 billion bases).

```
>gb:MN988668|Organism:Wuhan seafood market pneumonia virus|Strain
Name:2019-nCoV WHU01|Segment:null|Host:Human
TTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAAC
GAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACT
AATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGT
TGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCC
CTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACG
TGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGC
TTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGATG
CTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCG
TAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTT
CTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAG
GCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGT
TACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGC
CCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGT
CCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTGC
TTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAAA
TTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAAC
CAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACC
```

Fig. 1: A piece of COVID-19 DNA.

2.2 Structural variants

According to the National Center for Biotechnology Information (NCBI): "Structural variation (SV) is generally defined as a region of DNA approximately 1 kb and larger in size and can include inversions and balanced translocations or genomic imbalances (insertions and deletions), commonly referred to as copy number variants (CNVs)" [11]. In other words, this variations represent mutation in DNA, this mutations could be: insertions, deletions, inversions and translocations. In Figure 2, we present some examples.

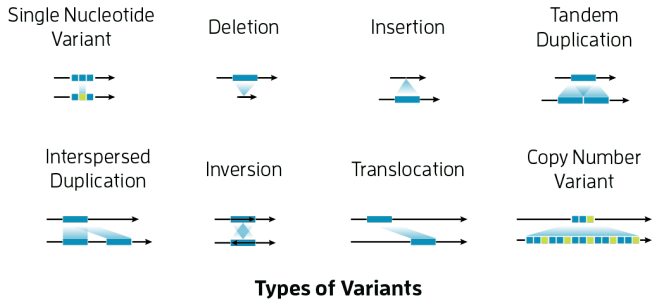

Fig. 2: Example of structural variants. Source: [12]

Copy number variants .- According to the National Human Genome (NIH): "A copy number variation (CNV) is when the number of copies of a particular gene varies from one individual to the next" [13]. For example in Figure 3, we present some examples of CNV, we could see how the number of genes varies individual 2 to 6. Additionally, it is recognized that some cancer diseases are associated to CNV [7–9, 13].

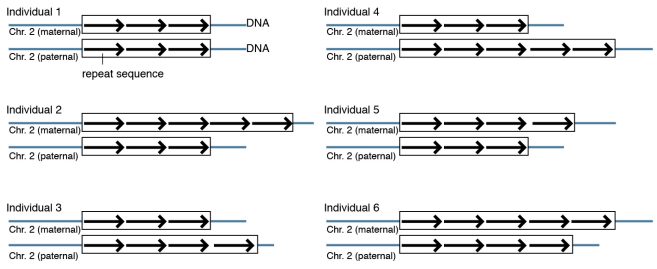

Fig. 3: Example of copy number variation. Source: [13]

Gene fusions .- Gene fusion is a gene made by two or more genes [14]. For example, the first gene fusion discovered in cancer was BCR/ABL (related to leukemia), it is resulted from a fusion of chromosomes [15].

3 Related work

4 Proposal

In this work, we replicated the results of SplitThreader

5 Results

6 Conclusions

References

1. Archana Prabahar and Subashini Swaminathan. Perspectives of machine learning techniques in big data mining of cancer. In *Big Data Analytics in Genomics*, pages 317–336. Springer, 2016.
2. Jules J Berman. *Principles of big data: preparing, sharing, and analyzing complex information*. Newnes, 2013.
3. John M Archibald. *Genomics: A Very Short Introduction*, volume 559. Oxford University Press, 2018.
4. Gurjit S Randhawa, Maximillian PM Soltysiak, Hadi El Roz, Camila PE de Souza, Kathleen A Hill, and Lila Kari. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. *bioRxiv*, 2020.
5. Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.
6. Philip J Hastings, James R Lupski, Susan M Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564, 2009.
7. Maria Nattestad, Marley C Alford, Fritz J Sedlazeck, and Michael C Schatz. Splitthreader: Exploration and analysis of rearrangements in cancer genomes. *bioRxiv*, page 087981, 2016.
8. Adam Shlien and David Malkin. Copy number variations and cancer. *Genome medicine*, 1(6):1–9, 2009.
9. Felix Mitelman, Bertil Johansson, and Fredrik Mertens. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, 7(4):233–245, 2007.
10. Britannica definitions. Dna. https://www.britannica.com/science/DNA, 2021. Accessed: 2021-05-07.
11. NCBI. Overview of structural variation. https://www.ncbi.nlm.nih.gov/dbvar/content/overview/, 2021. Accessed: 2021-05-07.
12. PacBio. Two review articles assess structural variation in human genomes. https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-genomes/, 2021. Accessed: 2021-05-07.

13. National Human Genome. Copy number variation (cnv). https://www.genome.gov/genetics-glossary/Copy-Number-Variation, 2021. Accessed: 2021-05-07.
14. National Cancer Institute. Gene fusion definition. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/fusion-gene, 2021. Accessed: 2021-05-07.
15. Kees Stam, Nora Heisterkamp, Gerard Grosveld, Annelies de Klein, Ram S Verma, Morton Coleman, Harvey Dosik, and John Groffen. Evidence of a new chimeric bcr/c-abl mrna in patients with chronic myelocytic leukemia and the philadelphia chromosome. *New England Journal of Medicine*, 313(23):1429–1433, 1985.