Finding Breakpoints and Similarity of Viral Genomes using recan

Abinu Ajithan Jyothini

August 6th 2020

1 Introduction

In many different group of viruses recombination of genetic material is an important evolutionary process that generates much of genetic diversity.The patterns of recombination evident within the genomes of viruses can reveal lots of detail about their evolution and biology.Recombination is responsible for the evolution of viruses in response to selective forces present in a host environment [1]. A better understanding of recombination events taking place in the genomes of viruses are useful for gaining in-sites about the biology of the viruses.The patterns of sequence exchange between viruses in different species can reveal otherwise undetectable ecological links between some species and barriers between others.[2]. The breakpoints of genomes can reveal details about mechanistic and biochemical processes regarding the recombination process.[2].This project makes use of the python package 'recan'[3](recombination analyzer) for constructing genetic distance plots and exploring them.Also, we make use of bio-python[4] package for plotting fasta sequences.

2 Methods:

2.1 recan Python Package:

recan is a python package developed for the recombination analysis and it provides a mean for constructing genetic distance plots and exploring them interactively.recan is based on python packages like Biopython, Pandas, Matplotlib, and Plotly libraries.recan requires the input to be in the form of alligned fasta sequence.There are options for changing the window size,adjusting the sequence of interest(for example; sequence where the breakpoints occurs), changing the method of distance calculations etc.

2.2 Input Data:

The fasta sequence of viral genomes can be obtained from any of the protein data banks. RCSB protein data bank [5](https://www.rcsb.org/) is most popular among bioinformatic scientists. We can directly obtain the fasta sequence of the genome from RCSB.Since the recan only takes aligned fasta format of the genome, the fasta sequence downloaded from the RCSB protein data bank needs to be alligned.This can be achieved by using NCBI Multiple SEquence Allignment Viewer(https://www.ncbi.nlm.nih.gov/projects/msaviewer/).In this project, i chose lumky skin disease virus's genome [6] for finding out the breakpoints. I have also used the genome sequence of BCRA1 genes, which are the genes that produce tumor suppressor proteins, for demonstrating the visualization of the fasta sequence.

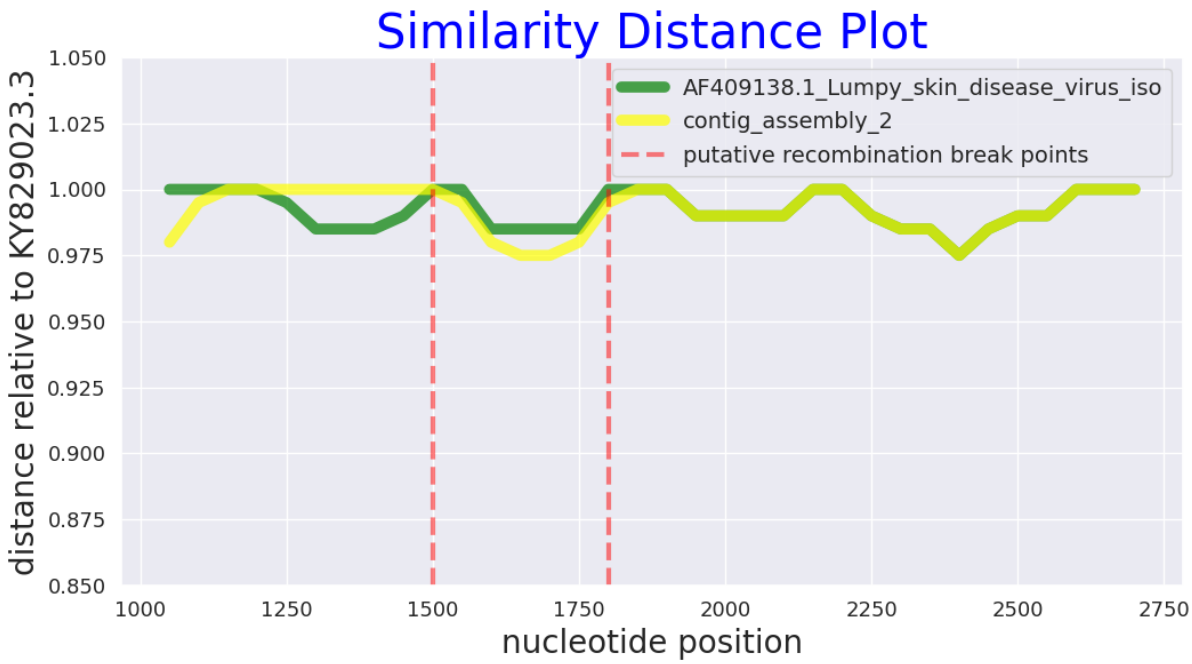


Figure 1: Distance Plot

3 Results and Discussion:

3.1 Breakpoints:

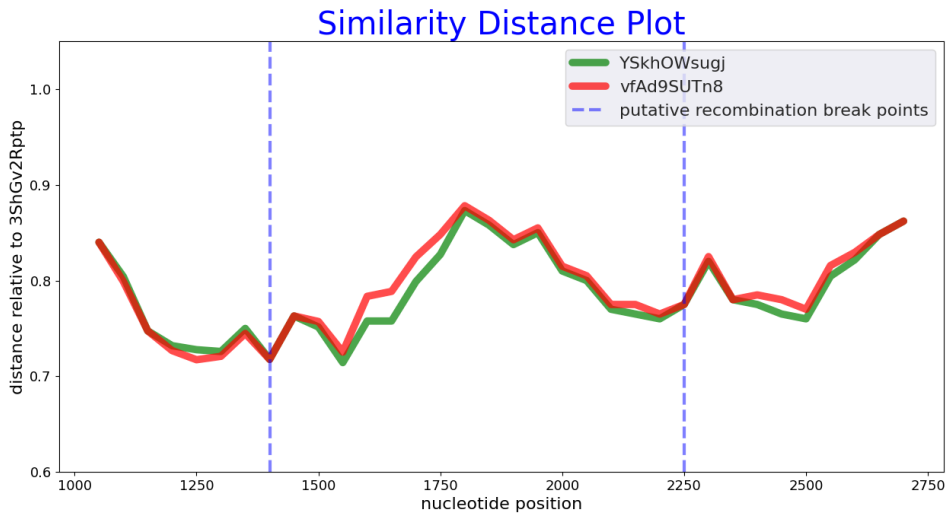
In genetics, the rearrangement of chromosomes is called mutation.It is an abnormality in the chromosome involving a change in the native chromosome [7].Theses changes may involve several events such as, duplication, inversion, deletion and translocation.These events are usually caused by a breakage in the helices of DNA and rejoining of the broken ends to form a new gene,different from the gene order of the chromosomes before they were broken [8].Some regions of chromosomes are more susceptible for rearrangements than others and thus are sources of genetic diseases and cancer.The Figure1 shows the breakpoints of lumky skin disease viral genome. We can see that, the recombination breakpoints are at around nucleotide number 1500 and at 1800. This is the clear evidence that recombination is taking place and the genome is undergoing mutation.

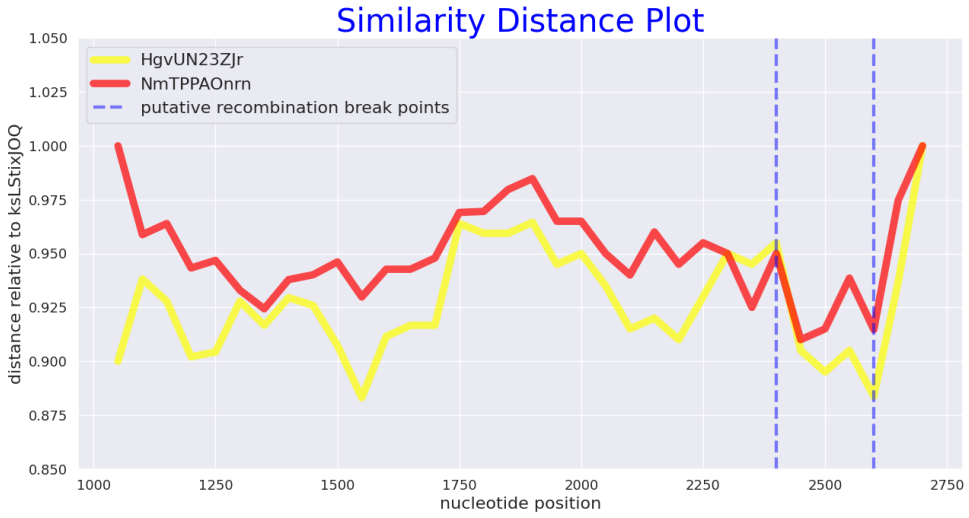
3.1.1 Breakpoints of HIV sequences:

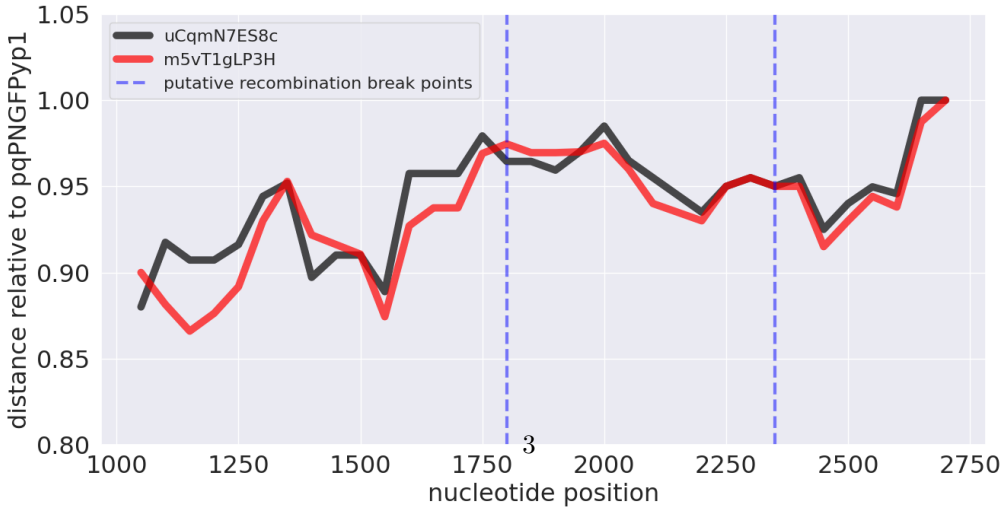
I plotted the distance plots of 18 sequences of HIV virus and found out the breakpoints.The subplots shown in Figure2 the breakpoints of nine compared sequences.The breakpoints obtained are summarised in Table3.1.2

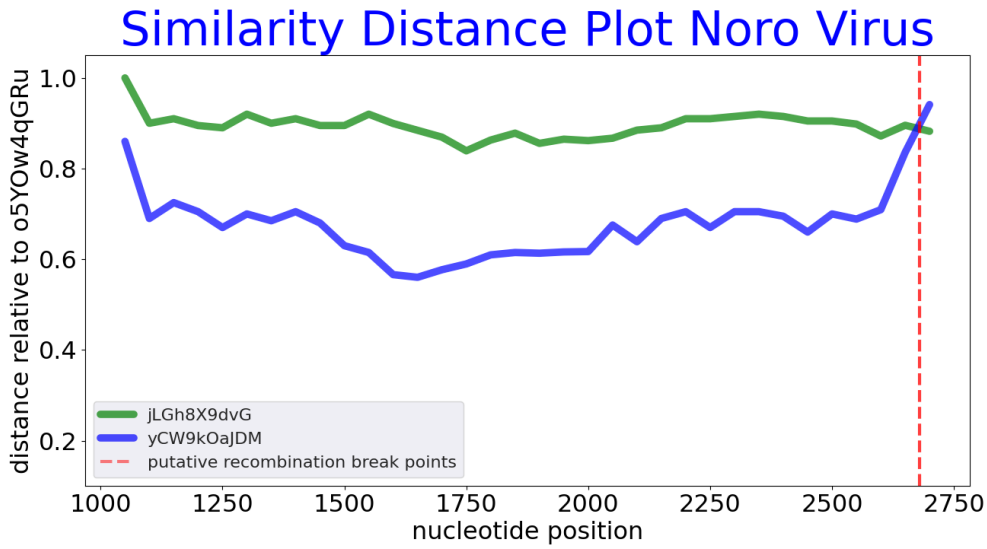
3.1.2 Breakpoints of NoroVirus sequences:

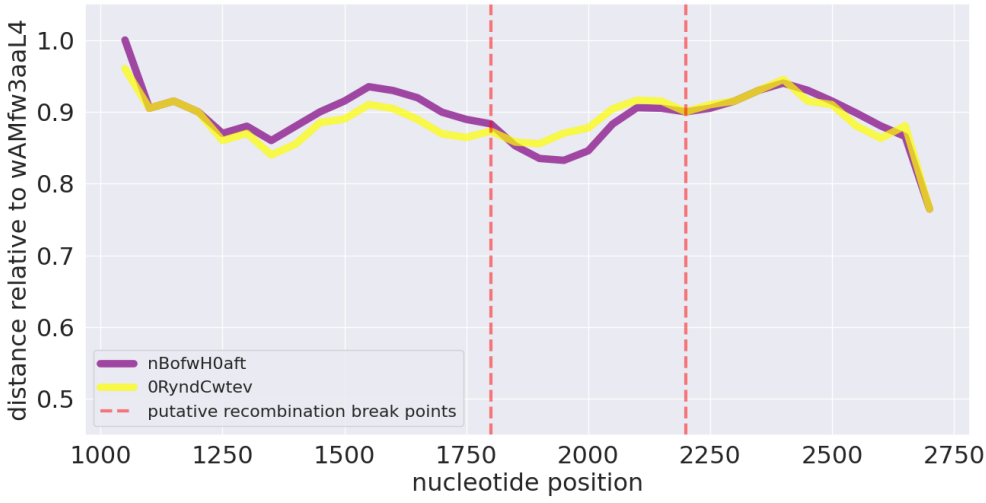
I have also plotted the distance plots of noro virus genome sequence for finding the breakpoints.Figure3 shows the distance plots for noro virus sequence. The breakpoints obtained are given in Table3.1.2

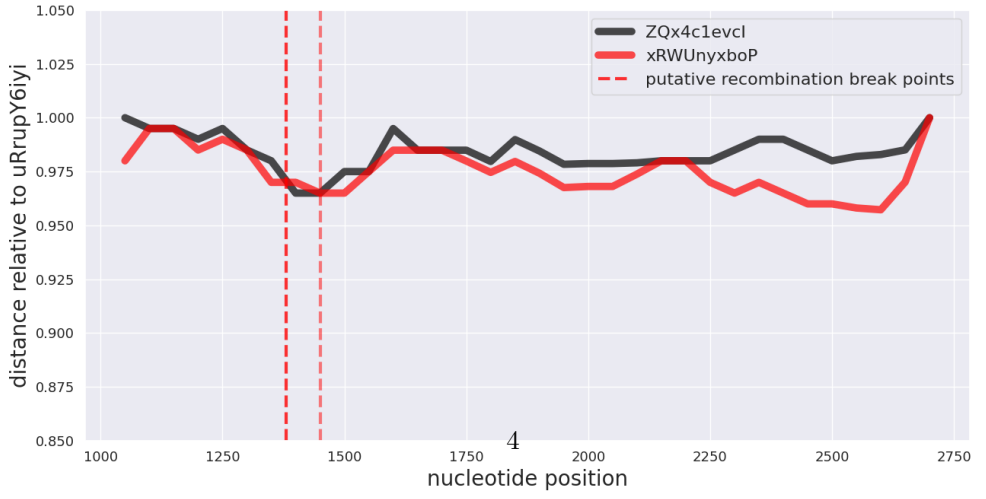

Similarity Distance Plot
YSkhOWsugj
vfAd9SUTn8
putative recombination break points
distance relative to 3ShGv2Rptp
nucleotide position

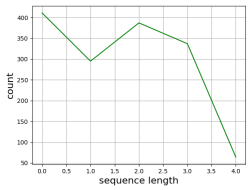

Similarity Distance Plot
HgvUN23ZJr
NmTPPAOnrn
putative recombination break points
distance relative to kslStixJOQ
nucleotide position


uCqmN7ES8c
m5vT1gLP3H
putative recombination break points
distance relative to pqPNGFPyp1
nucleotide position

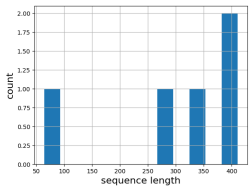

Similarity Distance Plot Noro Virus
distance relative to o5YOw4qGRu
1.0
0.8
0.6
0.4
0.2
jLGh8X9dvG
yCW9kOaJDM
putative recombination break points
1000
1250
1500
1750
2000
2250
2500
2750
nucleotide position


distance relative to wAMfw3aaL4
1.0
0.9
0.8
0.7
0.6
0.5
nBofwH0aft
0RyndCwtev
putative recombination break points
1000
1250
1500
1750
2000
2250
2500
2750
nucleotide position


distance relative to uRrupY6iyi
1.050
1.025
1.000
0.975
0.950
0.925
0.900
0.875
0.850
ZQx4c1evcI
xRWUnyxboP
putative recombination break points
1000
1250
1500
1750
2000
2250
2500
2750
nucleotide position



(a) A.Lineplot



(b) B.Histogram

Compared Sequences of HIV	Breakpoints
(YSkhOWsugj and vfAd9SUTn8)	At 1400 and 2350
(HgvUN23ZJr and NmTPPAOnrn)	At 2450 and 2400
(uCqmN7ES8cand m5vT1gLP3H)	At 1350 and 1500

ComparedSequences 0f Noro Virus	Breakpoints
(jLGh8X9dvG and yCW9kOaJDM)	—-
(nBofwH0aft and 0RyndCwtev)	At 1800 and 2200
(uCqmN7ES8c and m5vT1gLP3H)	At 1380 and 1450

3.2 FASTA Sequence Plotting

The FASTA sequences of genomes can be visualized in several ways. We can make use of the Biopython package of renac for plotting the sequence of BCRA1 gene having multiple sequences.The Figure4a shows the line plot and Figure4b shows the histogram of BCRA1 gene.

One of the most commonly used analytical data to compare different sequence is GC Percentage. In genetics and molecular biology, GC content(or guanine-cytosine content) is the s the percentage of nitrogenous bases in a DNA or RNA molecule that are either guanine (G) or cytosine (C)[9].The GC percentage indicates the proportion of bases out of an implied four total bases, also including adenine and thymine in DNA and adenine and uracil in RNA.Figure5shows the GC percentage of BCRA1 gene.

4 Conclusion:

We can make use of the python package recan for finding the breakpoints of genomes, which is an important parameter for studying further about the recombination events. Here we found that, for Lumpy skin diseases virus's genome has breakpoints at nucleotide positions 1500 and around 1800. The breakpoints of selected HIV and Norovirus sequences are identified, which can be utilized for further studies of recombination in those sequences. Also we can plot the fasta sequences using the Bio-python package.

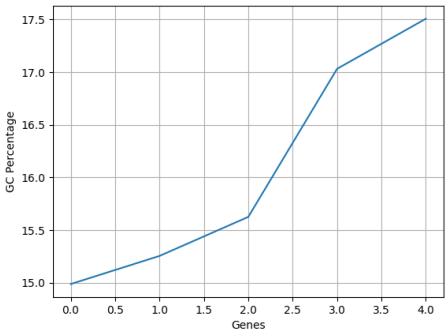


Figure 5: GC Content

References

[1] Marcos Pérez-Losada, Miguel Arenas, Juan Carlos Galán, Ferran Palero, and Fernando González-Candelas. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, 30:296–307, 2015.

[2] Darren P Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. Rdp4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution*, 1(1), 2015.

[3] Yuriy Babin. Recan: Python tool for analysis of recombination events in viral genomes. *Journal of Open Source Software*, 5(49):2014, 2020.

[4] Brad Chapman and Jeffrey Chang. Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, 20(2):15–19, 2000.

[5] Peter W Rose, Andreas Prlić, Ali Altunkaya, Chunxiao Bi, Anthony R Bradley, Cole H Christie, Luigi Di Costanzo, Jose M Duarte, Shuchismita Dutta, Zukang Feng, et al. The rcsb protein data bank: integrative view of protein, gene and 3d structural information. *Nucleic acids research*, page gkw1000, 2016.

[6] A Sprygin, E Artyuchova, YU Babin, P Prutnikov, E Kostrova, O Byadovskaya, and A Kononov. Epidemiological characterization of lumpy skin disease outbreaks in russia in 2016. *Transboundary and emerging diseases*, 65(6):1514–1521, 2018.

[7] Linlu Zhao, Elizabeth W Triche, Kyle M Walsh, Michael B Bracken, Audrey F Saftlas, Josephine Hoh, and Andrew T Dewan. Genome-wide association study identifies a maternal copy-number deletion in psg11 enriched among preeclampsia patients. *BMC pregnancy and childbirth*, 12(1):61, 2012.

[8] Anthony JF Griffiths, William M Gelbart, Richard C Lewontin, and Jeffrey H Miller. *Modern genetic analysis: integrating genes and genomes*, volume 1. Macmillan, 2002.

[9] Ekaterina Protozanova, Maxim D Frank-Kamenetskii, and Peter Yakovchuk. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. 2006.