Conventional Data Science Techniques to Bioinformatics and Utilizing a Grid Computing Approach to Computational Medicine

By Andrew M. K. Nassief

The Lonero Foundation

18th, January 2020

Abstract:

Conventional data visualization software have greatly improved the efficiency of the mining and visualization of biomedical data. However, when one applies a grid computing approach the efficiency and complexity of such visualization allows for a hypothetical increase in research opportunities. This paper will present data visualization examples presented in conventional networks, then go into higher details about more complex techniques related to leveraging parallel processing architecture. Part of these complex techniques include the attempt to build a basic general adversarial network (GAN) in order to increase the statistical pool of biomedical data for analysis as well as an introduction to the project utilizing the decentralized-internet SDK. This paper is meant to show you said conventional examples then go into details about the deeper experimentation and self contained results.

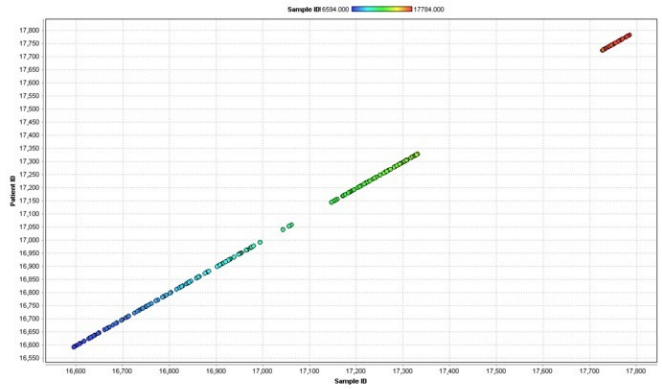
1.0 Background

The rise of big data have led to advancements in how data scientist can be leveraged in the medical field. Softwares such as RapidMiner Studio, Neo4j, and GraphXR have become database oriented approaches to the visualization and categorization of data. In the intersection of it all is the rise of bioinformatics and standardized formatting for genomic variants including VCF (Variant Cell Format) files. In this paper, case by case examples will be shown on how to utilize RapidMiner Studio and other tools for the visualization and tagging of genomic data, as well as approaches in which the efficiency of such techniques may greatly improve through parallel processing. We will then look at progress towards a generative adversarial network utilizing the decentralized-internet SDK for biostatistical analysis.

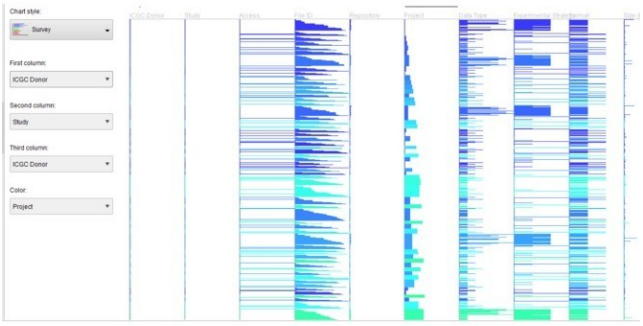
2.0 Examples

This section will look at examples of conventional data science techniques for mining, visualization, and categorization.

1.0 Rapid Miner



This figure above is an example of being able to see the genomic mutation view of cancer genomic donors all at once utilizing the mining and data visualization capabilities of Rapid Miner Studio.

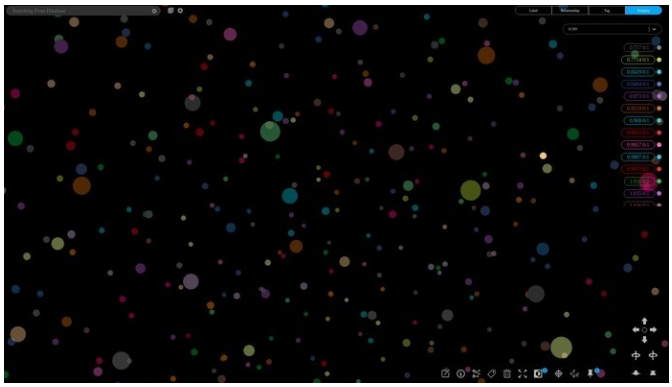


This figure above is the sorting view in Rapid Miner Studio that allows one to be able to sort the experimental data for ICGC donors.

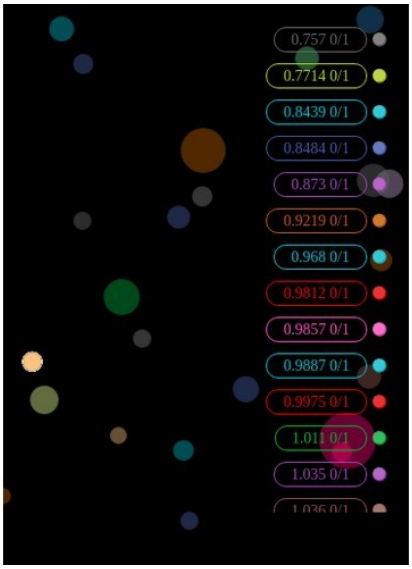
These examples are to show the capabilities of utilizing big data mining and visualization to quicker garnish conclusive results and/or view statistical outliers for said data.

This Rapid Miner example is minimal given further developments have been made utilizing an API, and a Rapid Miner extension have been developed. Integrative research similar to this have been done as well, including the creation of Cancer@Home, which utilizes parallel processing.

2.0 GraphXR and Neo4j



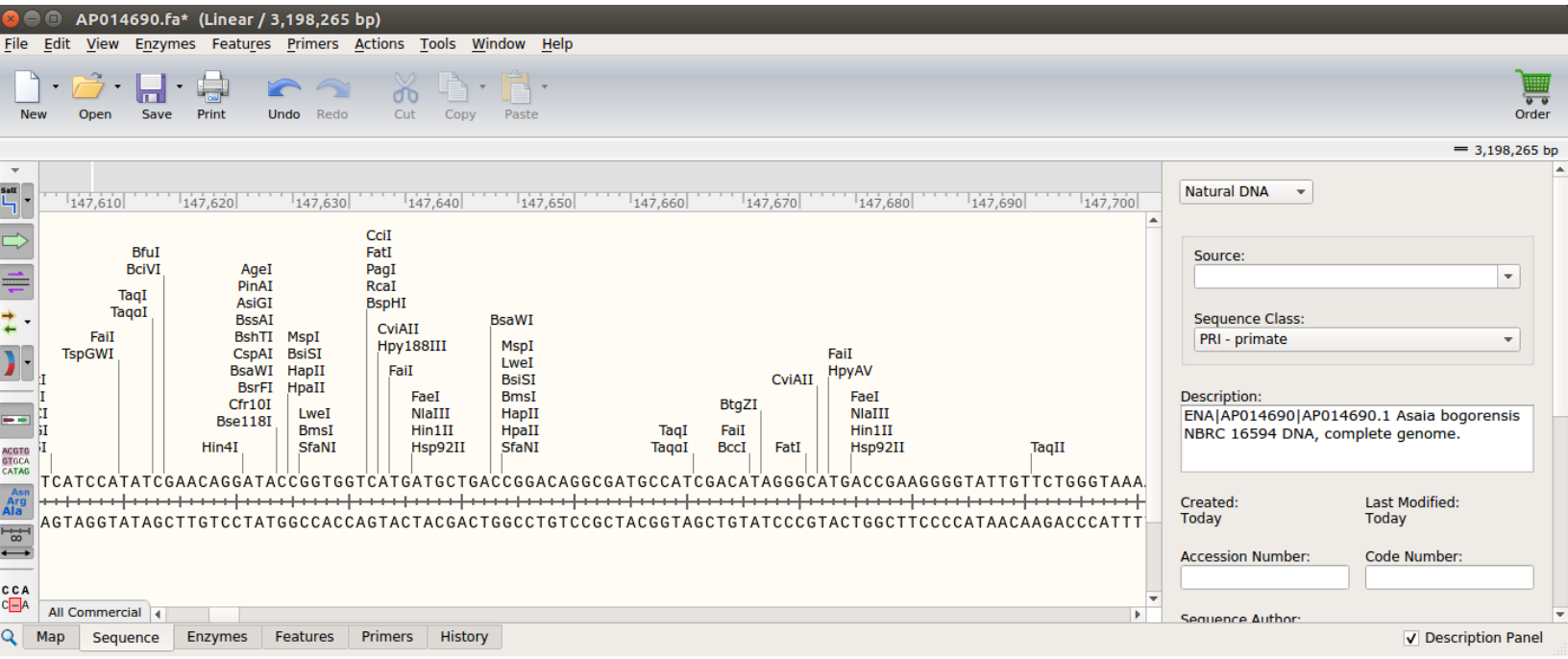
Shown above is a default dashboard in GraphXR for a project that is based off the visualization of pre-existing breast cancer data.



This image above shows the actual color coding view in GraphXR that was set to allow you to utilize Neo4j and Neo4j Desktop for the tagging of such data. This is just a tiny fraction of what is possible with modern data visualization. This also gives a sense of the possible roles a data scientist may have in the future of medicine with regards to data organization.

3.0 SnapGene Viewer

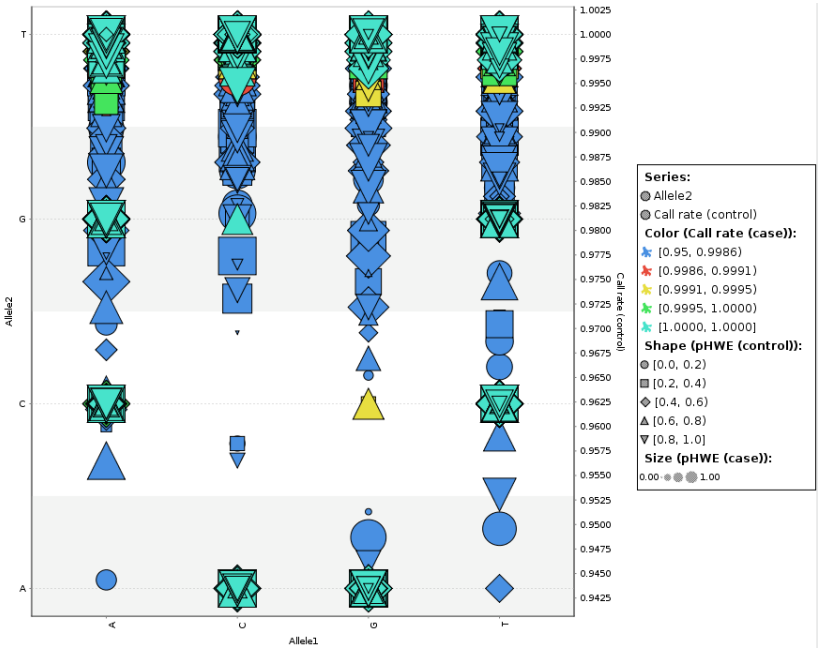
SnapGene Viewer is extensive in allowing one to view certain types of files that contain genomic information such as loading FASTA files.



Shown above is a closeup of SnapGene Viewer viewing the FASTA file of the sequence of AP014690.1. The sequence is related to the genomic analysis for the organism Asaia bogorensis NBRC 16594. SnapGene Viewer allows one to view the default sequence, commercially tag sequence classes, and change the view settings.

The allowance of such capabilities can be integrated with much larger data sets or genotypes and parallel processing can help offload latency times for creating output visualizations.

1.1 Rapid Miner

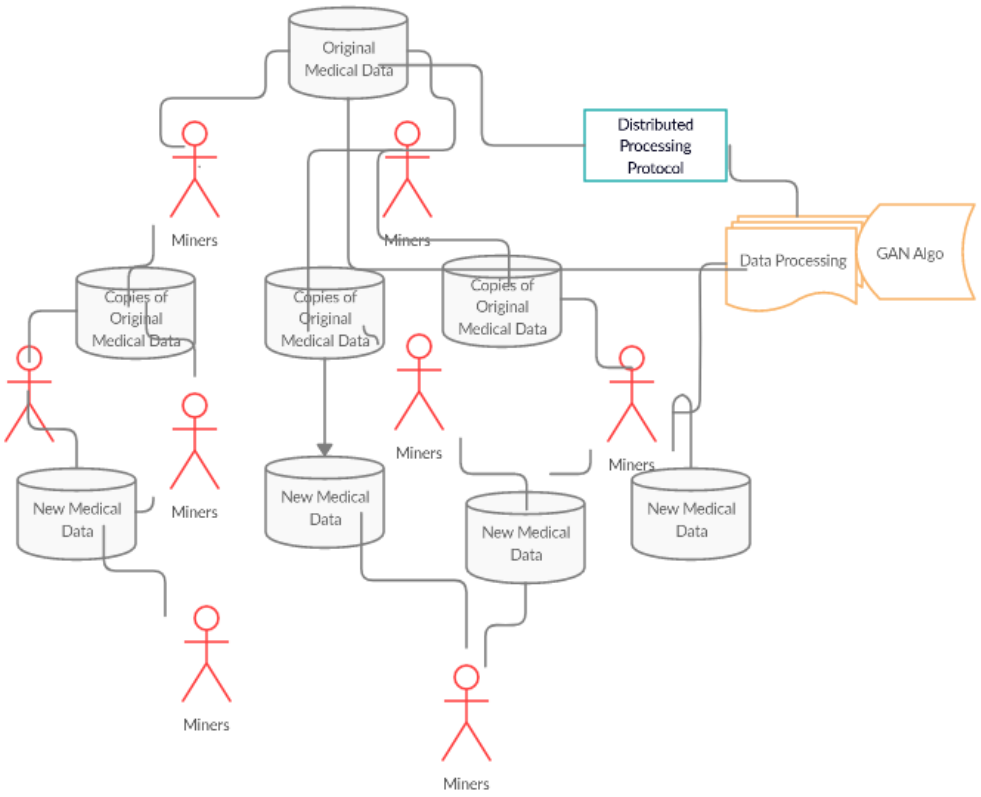


Rapid Miner has advanced data visualization capabilities. One can convert certain formats of data data files in similar ways that VCF files have headers. Once you convert specific file formats to an XML database query or CSV file, you can read in the data with Rapid Miner Studio. This is something that was experimentally done in order to categorize certain aspects of a cohort study that was loaded to this analysis. The cohort study was dbGaP Study Accession: phs000684.v1.p1 and is a simple open source data set we analyzed related to age-related macular degeneration.

3.0 Challenge

A current experimental challenge is creating accurate computational simulations from pre-existing nucleotide sequences. This is where the generative adversarial network comes into play. Numerous of resources are available for bioinfomatics researchers wanting to read in and output data. However, little resources are available for researchers who want to create synthetic data that is systematically or statistically similar to pre-existing data with a high confidence level. This is why generative adversarial networks is useful. An experiment I tried was the generation of a VCF created with freebayes, vcfy, and the decentralized-internet sdk. The data I used for the VCF was the FASTA file of the AP014690.1. The result from my experimentation is here. Given the output of the VCF, the VCF haven't been annotated. This means that for the generative adversarial network's pipeline, a way to efficiently annotate the input data for the VCF or standardize acceptable data formats may need to come into fruition. Right now progress have been at the early stages of development for the GAN given these implications needed for the development process. Currently, progress is being made into statistical analysis techniques and a technical contribution that may be further implemented may involve techniques such as the Hardy-Weinberg principle.

4.0 Proposal



The above UML diagram shows the likely beneficial proposal for an efficient generative

adversarial network utilized for biostatistical analysis. One would want to make copies of the pre-existing biomedical data with different but statically similar variance. The original copies of the original data are copied into distributed databases, and newly statistically similar files from those copies are being generated. Over time latency improves and so does processing capacity in the network. Technology like this allows the analyst with more data. A layman term's example would be, "if one was to have 50 donor sets, and they could turn it into 500 donor sets with a high confidence interval wouldn't they?" This makes not just data for clinical trials or cohort studies useful but can be applied to other forms of data as well.

Conclusion

Conventional data visualization software allows users the ability to mine pre-existing data files. However, this approach can likely be taken a step further. Given the custom tagging capabilities of such software, one is able to change the way data sets are formatted as well as efficiently categorize and analyze applicable data. This has noticeable advantages not just visually, but statistically. An advancement of this is creating an API pipeline or an extension for outputting the data in various ways. This pipeline can hypothetically be utilized for generative adversarial networks that may be able to make similar data files to pre-existing ones. Parallel Processing techniques are part of such integrations for faster offloading the data, and hence making commercially applicable biostatistics applications more viable.

References

1. M.K. Nassief, A. (2019). A Distributed Architecture Proposal for Regressional Generative Adversarial Networks for Biostatistical Analysis Modeled after the Decentralized-Internet SDK and BOINC. [online] Authorea. Available at: https://www.authorea.com/users/289895/articles/416340-a-distributed-architecture-proposal-for-regressional-generative-adversarial-networks-for-biostatistical-analysis-modeled-after-the-decentralized-internet-sdk-and-boinc [Accessed 18 Jan. 2020].
2. Kamal, A. Data Mining Cancer Genomics Case Studies. Devpost (2017). Available at: https://devpost.com/software/data-mining-cancer-genomics-case-studies. (Accessed: 23rd December 2019)
3. Cancer@Home. Stark Drones - OS (2019). Available at: https://www.starkdrones.org/home/os#h.p_-bdQF7aLVCfp. (Accessed: 23rd December 2019)
4. Kamal, A. (2019). *Visualizing Breast Cancer Data with Neo4j and GraphXR*. [online] Medium. Available at: https://medium.com/neo4j/visualize-cancer-1c80a95f5bb4?source=----------2---------- [Accessed 18 Jan. 2020].
5. Y., A. and M., K. (2020). *Sequence: AP014690.1*. [online] Ebi.ac.uk. Available at: https://www.ebi.ac.uk/ena/data/view/AP014690 [Accessed 18 Jan. 2020].
6. Ncbi.nlm.nih.gov. (2020). *Age related Macular Degeneration (AMD) - Michigan, Mayo, AREDS, Pennsylvania (MMAP) Cohort Study: Association and Sequencing Studies*. [online] Available at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000684.v1.p1&phv=196861&phd=&pha=2890&pht=3643&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1 [Accessed 18 Jan. 2020].
7. M. Kamal, A. (2020). *Data Scientists, Software Engineers And The Future of Medicine*. [online] Hackernoon.com. Available at: https://www.hackernoon.com/data-scientists-software-engineers-and-the-future-of-medicine-4i4832tz [Accessed 18 Jan. 2020].
8. En.wikipedia.org. (2020). *Hardy–Weinberg principle*. [online] Available at: https://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle [Accessed 18 Jan. 2020].