Genomic and Proteomic data exploration and pattern mining

Submitted by
Sreehari P V
(Reg No: 31018019)

In partial fulfillment of the requirements for the award of Master of Science in Computer Science with Specialization in Machine Intelligence
Of

Cochin University of Science and Technology, Kochi
Conducted by

Indian Institute of Information Technology and Management-Kerala
Technopark Campus
Thiruvananthapuram-695 581

April 2020

BONAFIDE CERTIFICATE

This is to certify that the project report entitled ”**Genomic and proteomic data exploration and pattern mining**” submitted by **Sreehari P V (31018019)** in partial fulfillment of the requirements for the award of Master of Science in Computer Science with Specialization in Machine Intelligence is a bonafide record of the work carried out at “I**ndian Institute of Information Technology and Management-Kerala”** under our supervision.

Supervisor	**Course Coordinator**
Mr. ANOOP V.S.	Dr. Asharaf S
Senior Scientist & Head (Research & Training	AssociateProfessor
Kerala Blockchain Academy	IIITM-K
IIITM-K	

Head of Institution

Prof.Saji Gopinath

Director

IIITM-K

DECLARATION

I, **Sreehari P V**, student of **Master of Science in Computer Science** with specialization in **Machine Intelligence** , hereby declare that this report is substantially the result of my own work , except where explicitly indicated in the text, and has been carried out during the period December 2019 to April 2020.

Place:Thiruvananthapuram
Date:15-4-2020 Student's signature

ACKNOWLEDGEMENT

First and foremost I thank God Almighty for his blessings for this project. I take this opportunity to express my gratitude to all those who helped me in completing this project successfully.

With deep respect, I express my gratitude to Mr. ANOOP V.S.,my supervisor and Dr. Asharaf S, my course coordinator and Professor at Indian Institute of Information Technology and Management- Kerala (IIITMK) ,for their constant encouragement throughout this project.

I also like to thank all other faculties of our institute,for their valuable comments and guidance and the technical staff for providing the technical support, for the completion of this project.

I would like to thank all my friends and classmates of IIITMK who have been supporting me for the completion of my work.

PROJECT PROFILE

Title	:	Genomic and Proteomic data exploration and pattern mining
Type	:	Academic Project.
Objective	:	The objective of the project is to do genomic and Proteomic analysis of a wide variety of gene data using different tools
Organization	:	IIITM-K
Project Guide	:	Mr.Anoop V.S Senior Scientist & Head (Research & Training) KBA IIITM-K

ABSTRACT

At a time, when the world is facing a major financial and humanitarian crisis due to an ongoing pandemic caused by a deadly virus, this project is a small humble attempt to showcase the power of bioinformatics by carrying out genomic and proteomic analysis of two different gene data: one involving cancer cells that have been causing for a long time, one of mankind's deadliest disease and the other being the gene of a virus that dominates news of today and is responsible for the above mentioned crisis. Attempts are then made to uncover patterns in the viral gene so as to understand its behaviour by making use of the most advanced tools in AI.

Table of Contents

List of figures

Chapter 1

INTRODUCTION

This first part of this project aims to use unsupervised learning techniques in order to determine whether there are groups or clusters among different cancer cells based on their gene expression measurements. This is a difficult problem as there are thousands of gene expression measurements for each cancer cell making it very difficult to visualize the data. Hence Principal Component Analysis is carried out to reduce the data into a few dimensions namely the principal components followed by hierarchical clustering to determine the clusters whose results are then compared with that obtained by K-Means.

This is followed by extensive genomic analysis and minor proteomic analysis of the SARS-CoV-2 viral genome so as to understand and explain its behaviour.

Lastly this project attempts to leverage the power of OpenCog, an open source Artificial General Intelligence framework to gain more insights from the viral genome.

Chapter 2

LITERATURE REVIEW

2.1 Global genomic and proteomic analysis identifies biological pathways related to high-risk neuroblastoma

- *Qing-Rong Chen*,*Young K Song,1,Li-Rong Yu,3,Jun S. Wei,1 Joon-Yong Chung, Stephen M. Hewitt,Timothy D. Veenstra, and Javed Khan*

Neuroblastoma (NB) is a heterogeneous pediatric tumor. To better understand the biological pathways involved in the development of high-risk neuroblastoma, we performed parallel global protein and mRNA expression profiling on NB tumors of stage4 *MYCN*-amplified (4+) and stage1 *MYCN*-not-amplified (1-) using isotope-coded affinity tags (ICAT) and Affymetrix U133plus2 microarray respectively. A total of 1461 proteins represented by 2 or more peptides were identified from the quantitative ICAT analysis, of which 433 and 130 proteins are up- or down-regulated respectively in 4+ tumors compared to the 1- tumor. Pathway analysis of the differentially expressed proteins showed the enrichment of glycolysis, DNA replication and cell cycle processes in the up-regulated proteins and cell adhesion, nervous system development and cell differentiation processes in the down-regulated proteins in 4+ tumor; suggesting a less mature neural and a more invasive phenotype of 4+ tumor. Myc targets and ribosomal proteins are over represented in the 4+ tumors as expected; functional gene sets reported to be enriched in neural and embryonic stem cells are significantly enriched in the 4+ tumor, indicating the existence of a stemness signature in *MYCN*-amplified stage 4 tumor. In addition, protein and mRNA expression are moderately correlated ($r = 0.51$, $p < 0.0001$), as approximately half of the up-regulated proteins in 4+ tumor have elevated mRNA level (n=208), and 1/3 of down regulated proteins have lower mRNA expression (n=47). Further biological network analysis revealed that the differentially expressed proteins closely interact with other proteins of known networks; the important role of MYCN is confirmed and other transcription factors identified in the network may have potential roles in the biology of NB tumor. We used global genomic and proteomic analysis to identify biologically relevant proteins and pathways important to NB progression and development that may provide new insights into the biology of advanced neuroblastoma.

2.2 Features, Evaluation and Treatment Coronavirus (COVID-19)

- *Marco Cascella; Michael Rajnik; Arturo Cuomo; Scott C. Dulebohn; Raffaela Di Napoli.*

According to the World Health Organization (WHO), viral diseases continue to emerge and represent a serious issue to public health. In the last twenty years, several viral epidemics such as the severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002 to 2003, and H1N1 influenza in 2009, have been recorded. Most recently, the Middle East respiratory syndrome coronavirus (MERS-CoV) was first identified in Saudi Arabia in 2012.

In a timeline that reaches the present day, an epidemic of cases with unexplained low respiratory infections detected in Wuhan, the largest metropolitan area in China's Hubei province, was first reported to the WHO Country Office in China, on December 31, 2019. Published literature can trace the beginning of symptomatic individuals back to the beginning of December 2019. As they were unable to identify the causative agent, these first cases were classified as "pneumonia of unknown etiology." The Chinese Center for Disease Control and Prevention (CDC) and local CDCs organized an intensive outbreak investigation program. The etiology of this illness is now attributed to a novel virus belonging to the coronavirus (CoV) family.

On February 11, 2020, the WHO Director-General, Dr. Tedros Adhanom Ghebreyesus, announced that the disease caused by this new CoV was a "COVID-19," which is the acronym of "coronavirus disease 2019". In the past twenty years, two additional coronavirus epidemics have occurred. SARS-CoV provoked a large-scale epidemic beginning in China and involving two dozen countries with approximately 8000 cases and 800 deaths, and the MERS-CoV that began in Saudi Arabia and has approximately 2,500 cases and 800 deaths and still causes sporadic cases.

This new virus seems to be very contagious and has quickly spread globally. In a meeting on January 30, 2020, per the International Health Regulations (IHR, 2005), the outbreak was declared by the WHO a Public Health Emergency of International Concern (PHEIC) as it had spread to 18 countries with four countries reporting human-to-human transmission. An additional landmark occurred on February 26, 2020, as the first case of the disease, not imported from China, was recorded in the United States.

Initially, the new virus was called 2019-nCoV. Subsequently, the task of experts of the International Committee on Taxonomy of Viruses (ICTV) termed it the SARS-CoV-2 virus as it is very similar to the one that caused the SARS outbreak (SARS-CoVs).

The CoVs have become the major pathogens of emerging respiratory disease outbreaks. They are a large family of single-stranded RNA viruses (+ssRNA) that can be isolated in different animal species.[1] For reasons yet to be explained, these viruses can cross species barriers and can cause, in humans, illness ranging from the common cold to more severe diseases such as MERS and SARS. Interestingly, these latter viruses have probably originated from bats and then moved into other mammalian hosts — the Himalayan palm civet for SARS-CoV, and the dromedary camel for MERS-CoV — before jumping to humans. The dynamics of SARS-Cov-2 are currently unknown, but there is speculation that it also has an animal origin.

The potential for these viruses to grow to become a pandemic worldwide seems to be a serious public health risk. Concerning COVID-19, the WHO raised the threat to the CoV epidemic to the "very high" level, on February 28, 2020. Probably, the effects of the epidemic caused by the new CoV has yet to emerge as the situation is quickly evolving. On March 11, as the number of COVID-19 cases outside China has increased 13 times and the number of countries involved has tripled with more than 118,000 cases in 114 countries and over 4,000 deaths, WHO declared the COVID-19 a pandemic.

World governments are at work to establish countermeasures to stem possible devastating effects. Health organizations coordinate information flows and issues directives and guidelines to best mitigate the impact of the threat. At the same time, scientists around the world work tirelessly, and information about the transmission mechanisms, the clinical spectrum of disease, new diagnostics, and prevention and therapeutic strategies are rapidly developing. Many uncertainties remain with regard to both the virus-host interaction and the evolution of the epidemic, with specific reference to the times when the epidemic will reach its peak.

At the moment, the therapeutic strategies to deal with the infection are only supportive, and prevention aimed at reducing transmission in the community is our best weapon. Aggressive isolation measures in China have led to a progressive reduction of cases in the last few days. In Italy, in geographic regions of the north, initially, and subsequently throughout the peninsula, political and health authorities are making incredible efforts to contain a shock wave that is severely testing the health system.

2.3 Coronavirus origins: genome analysis suggests two viruses may have combined

- *Misha Ketchell, Editor, The conversation*

The SARS-CoV-2 genome was rapidly sequenced by Chinese researchers. It is an RNA molecule of about 30,000 bases containing 15 genes, including the S gene which codes for a protein located on the surface of the viral envelope (for comparison, our genome is in the form of a double helix of DNA about 3 billion bases in size and contains about 30,000 genes).

Comparative genomic analyses have shown that SARS-CoV-2 belongs to the group of *Betacoronaviruses* and that it is very close to SARS-CoV, responsible for an epidemic of acute pneumonia which appeared in November 2002 in the Chinese province of Guangdong and then spread to 29 countries in 2003. A total of 8,098 cases were recorded, including 774 deaths. It is known that bats of the genus *Rhinolophus* (potentially several cave species) were the reservoir of this virus and that a small carnivore, the palm civet (*Paguma larvata*), may have served as an intermediate host between bats and the first human cases.

Since then, many *Betacoronaviruses* have been discovered, mainly in bats, but also in humans. For example, RaTG13, isolated from a bat of the species *Rhinolophus affinis* collected in China's Yunan Province, has recently been described as very similar to SARS-CoV-2, with genome sequences identical to 96%. These results indicate that bats, and in particular species of the genus *Rhinolophus*, constitute the reservoir of the SARS-CoV and SARS-CoV-2 viruses.

Chapter-3

TOOLS AND TECHNOLOGIES

4.1 R software

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, data mining surveys, and studies of scholarly literature databases show substantial increases in popularity; as of February 2020, R ranks 13th in the TIOBE index, a measure of popularity of programming languages.

A GNU package, the official R software environment is written primarily in C, Fortran, and R itself(thus, it is partially self-hosting) and is freely available under the GNU General Public License. Pre-compiled executables are provided for various operating systems. Although R has a command line interface, there are several third-party graphical user interfaces, such as RStudio, an integrated development environment, and Jupyter, a notebook interface.

4.2 Python

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library. Python was conceived in the late 1980s as a

successor to the ABC language. Python 2.0, released in 2000, introduced features like list comprehensions and a garbage collection system capable of collecting reference cycles. Python 3.0, released in 2008, was a major revision of the language that is not completely backward-compatible, and much Python 2 code does not run unmodified on Python 3.The Python 2 language, i.e. Python 2.7.x, was officially discontinued on 1 January 2020 (first planned for 2015) after which security patches and other improvements will not be released for it.With Python 2's end-of-life, only Python 3.5.x and later are supported.

Python interpreters are available for many operating systems. A global community of programmers develops and maintains CPython, an open source reference implementation. A non-profit organization, the Python Software Foundation, manages and directs resources for Python and CPython development.

4.3 Biopython

Biopython is a set of freely available tools for biological computation written in Python by an international team of developers.

It is a distributed collaborative effort to develop Python libraries and applications which address the needs of current and future work in bioinformatics. The source code is made available under the Biopython License, which is extremely liberal and compatible with almost every license in the world. It contains classes to represent biological sequences and sequence annotations, and it is able to read and write to a variety of file formats. It also allows for a programmatic means of accessing online databases of biological information, such as those at NCBI. Separate modules extend Biopython's capabilities to sequence alignment, protein structure, population genetics, phylogenetics, sequence motifs, and machine learning. Biopython is one of a number of Bio* projects designed to reduce code duplication in computational biology.

4.4 BLAST-Basic Local Alignment Search Tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to

sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

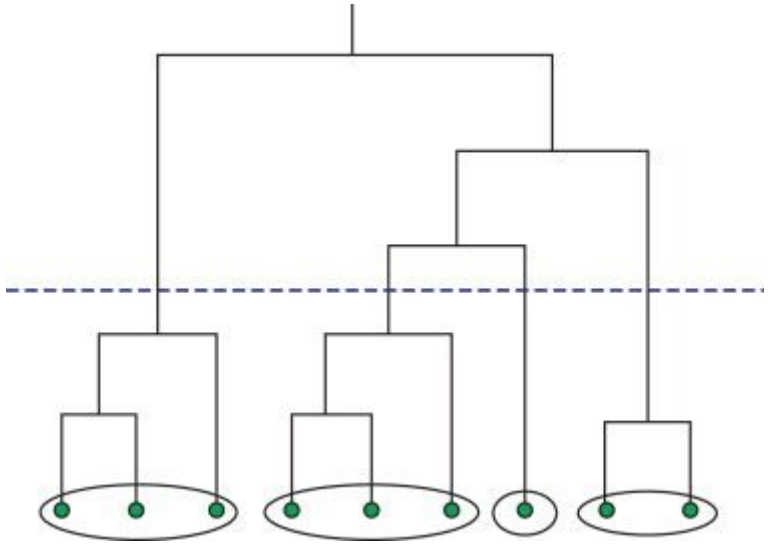
4.5 Principal component analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing *n* observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It is often used to visualize genetic distance and relatedness between populations. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after a normalization step of the initial data. The normalization of each attribute consists of *mean centering* – subtracting each data value from its variable's measured mean so that its empirical mean (average) is zero. Some fields, in addition to normalizing the mean, do so for each variable's variance (to make it equal to 1); see Z-scores.] The results of a PCA are usually discussed in terms of *component scores*, sometimes called *factor scores* (the transformed variable values corresponding to a particular data point), and *loadings* (the weight by which each standardized original variable should be multiplied to get the component score). If component scores are standardized to unit variance, loadings must contain the data variance in them (and that is the magnitude of eigenvalues). If component scores are not standardized (therefore they contain the data variance) then loadings must be unit-scaled, ("normalized") and these weights are called eigenvectors; they are the cosines of orthogonal rotation of variables into principal components or back.

4.6 Hierarchical Clustering

One potential disadvantage of K-means clustering is that it requires us to pre-specify the number of clusters K. Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K. Hierarchical clustering has an added advantage over K-means clustering in that it results in an attractive tree-based representation of the observations, called a dendrogram. In this project, bottom-up or agglomerative clustering. bottom-up agglomerative is used. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk. How can we interpret this dendrogram? Each leaf of the dendrogram represents one of the observations. However, as we move up the tree, some leaves begin to fuse into branches. These correspond to observations that are similar to each other. As we move higher up the tree, branches themselves fuse, either with leaves or other branches. The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other. On the other hand, observations that fuse later (near the top of the tree) can be quite different. In fact, this statement can be made precise: for any two observations, we can look for the point in the tree where branches containing those two observations are first fused. The height of this fusion, as measured on the vertical axis, indicates how different the two observations are. Thus, observations that fuse at the very bottom of the tree are quite similar to each other, whereas observations that fuse close to the top of the tree will tend to be quite different. For identifying clusters on the basis of a dendrogram, we make a horizontal cut across the dendrogram, as shown in Figure 4.1(below). The distinct sets of observations beneath the cut can be interpreted as clusters.



4.7 NCBI- National Center for Biotechnology Information website

As a national resource for molecular biology information, NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. More specifically, the NCBI has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules.

NCBI assumed responsibility for the GenBank DNA sequence database in October 1992. NCBI staff with advanced training in molecular biology build the database from sequences submitted by individual laboratories and by data exchange with the international nucleotide sequence databases, European Molecular Biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ). Arrangements with the U.S. Patent and Trademark Office enable the incorporation of patented sequence data.

In addition to GenBank, NCBI supports and distributes a variety of databases for the medical and scientific communities. These include the Online Mendelian Inheritance in Man (OMIM), the Molecular Modeling Database (MMDB) of 3D protein structures, a Gene Map of the Human Genome, the Taxonomy Browser, and the Cancer Genome Anatomy Project (CGAP), in collaboration with the National Cancer Institute.

Entrez is NCBI's search and retrieval system that provides users with integrated access to sequence, mapping, taxonomy, and structural data. Entrez also provides graphical views of sequences and chromosome maps. A powerful and unique feature of Entrez is the ability to retrieve related sequences, structures, and references. The journal literature is available through PubMed, a Web search interface that provides access to over 11 million journal citations in MEDLINE and contains links to full-text articles at participating publishers' Web sites.

BLAST is a program for sequence similarity searching developed at NCBI and is instrumental in identifying genes and genetic features. BLAST can execute sequence searches against the entire DNA database in less than 15 seconds.

Additional software tools provided by NCBI include: Open Reading Frame Finder (ORF Finder), Electronic PCR, and the sequence submission tools, Sequin and BankIt. All of NCBI's databases and software tools are available from the WWW or by FTP. NCBI also has email servers that provide an alternative way to access the databases for text searching or sequence similarity searching.

4.7 OpenCog

OpenCog is a unique and ambitious open-source software project whose vision is to create an open source framework for Artificial General Intelligence, intended to one day express general intelligence at the human level and beyond. The OpenCog project provides key components and a platform for AI R&D. OpenCog is a unique and ambitious open-source software project whose vision is to create an open source framework for Artificial General Intelligence, intended to one day express general intelligence at the human level and beyond. The OpenCog project provides key components and a platform for AI R&D.

4.7.1 AtomSpace The OpenCog AtomSpace is a knowledge representation (KR) database and the associated query/reasoning engine to fetch and manipulate that data, and perform reasoning on it. Data is represented in the form of graphs, and more generally, as hypergraphs; thus the AtomSpace is a kind of graph database, the query engine is a general graph rewriting system, and the rule-engine is a generalized rule-driven inferencing system. The vertices and edges of a graph, known as Atoms, are used to represent not only "data", but also "procedures"; thus, many graphs are executable programs as well as data structures. These Atoms, which are permanent and immutable, can be assigned fleeting, changing Values to indicate the truth or likelihood of that atom, or to hold other kinds of transient data. The AtomSpace enables flow-based programming, where Atoms represent the pipes, and Values are what flows through the pipes. More info at https://wiki.opencog.org/w/AtomSpace.

4.7.2 Meta-optimizing semantic evolutionary search (MOSES) is a new approach to program evolution, based on representation-building and probabilistic modeling. MOSES has been successfully applied to solve hard problems in domains such as computational biology, sentiment evaluation, and agent control. Results tend to be more accurate, and require less objective function evaluations, than other program

evolution systems, such as genetic programming or evolutionary programming . Best of all, the result of running MOSES is not a large nested structure or numerical vector, but a compact and comprehensible program written in a simple Lisp-like mini-language. MOSES performs supervised learning, and thus requires either a scoring function or training data to be specified as input. As output, it generates a Combo program that, when executed, approximates the scoring function. MOSES uses general concepts from evolutionary search, in that it maintains a population of programs, and then explores the neighborhood of modified, "mutated" programs, evaluating their fitness. After some number of iterations, the fittest program found is output.More info https://wiki.opencog.org/w/Meta-Optimizing_Semantic_Evolutionary_Search

https://wiki.opencog.org/w/OpenCogPrime:Probabilistic_Evolutionary_Learning_Overview

3. Probabilistic logic networks

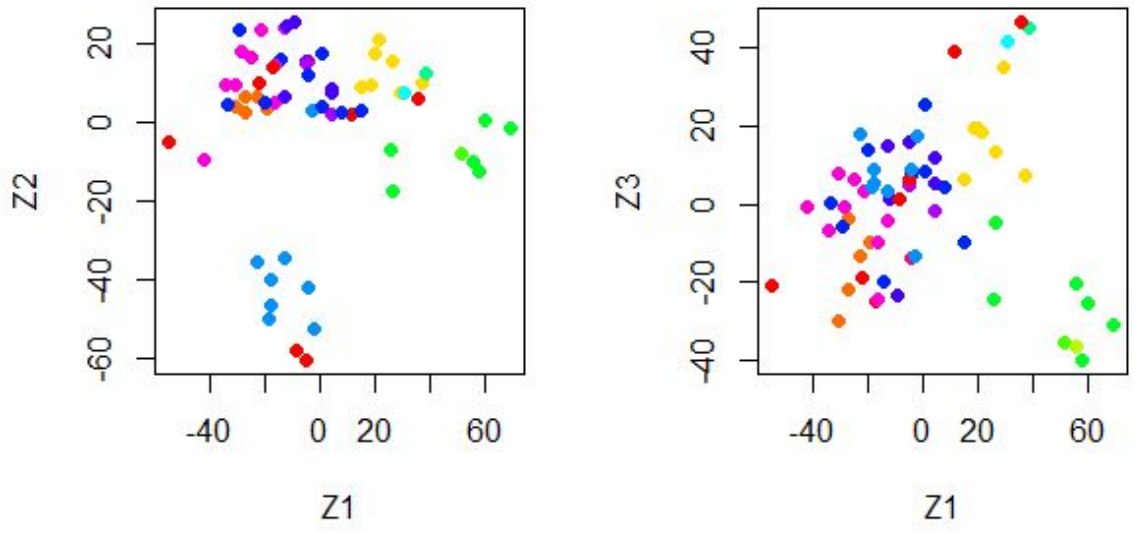
PLN is a novel conceptual, mathematical and computational approach to uncertain inference. In order to carry out effective reasoning in real-world circumstances, AI software must robustly handle uncertainty. However, previous approaches to uncertain inference do not have the breadth of scope required to provide an integrated treatment of the disparate forms of cognitively critical uncertainty as they manifest themselves within the various forms of pragmatic inference. Going beyond prior probabilistic approaches to uncertain inference, PLN is able to encompass within uncertain logic such ideas as induction, abduction, analogy, fuzziness and speculation, and reasoning about time and causality.

https://wiki.opencog.org/w/Probabilistic_logic_networks

Chapter-4

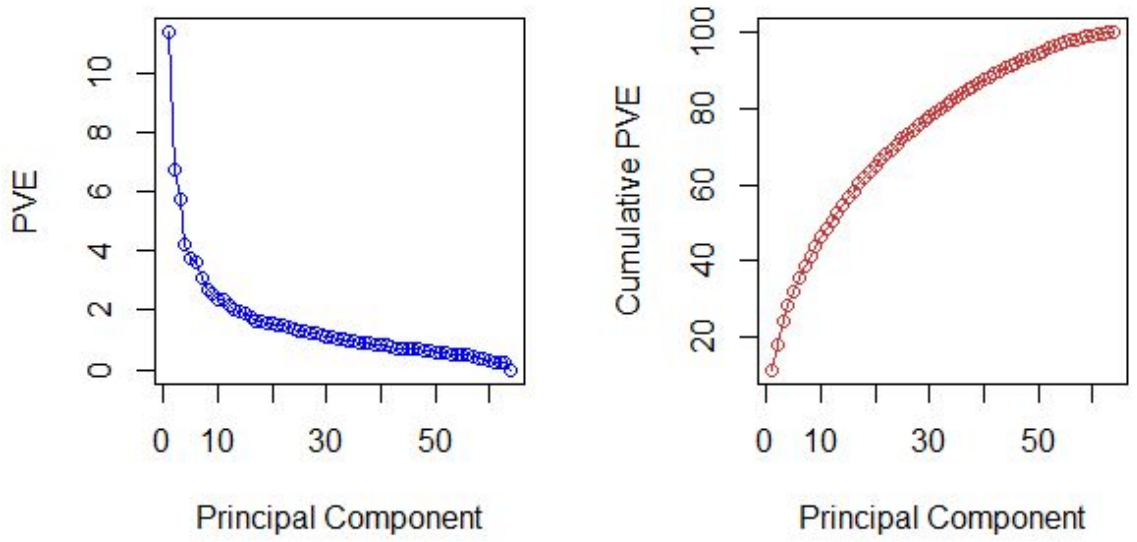
Clustering Cancer Cells Based on Gene Measurements

Here,the NCI60 data set, which consists of 6,830 gene expression measurements for each of 64 cancer cell lines, is used. There are labels associated with each cancer cell line (or each row of data) which identifies the type of cancer such as “Breast”,”Leukemia” etc. The goal is determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements. This is a difficult question to address, in part because there are thousands of gene expression measurements per cell line, making it hard to visualize the data. Hence Principal Component Analysis is carried out to reduce the data into a few dimensions namely the principal components.The figure given below shows the representation of the NCI60 gene expression data set in a two-dimensional space, Z1 and Z2 and Z1 and Z3. Each point corresponds to one of the 64 cell lines. There appear to be four groups of cell lines, which have been represented using different colors.



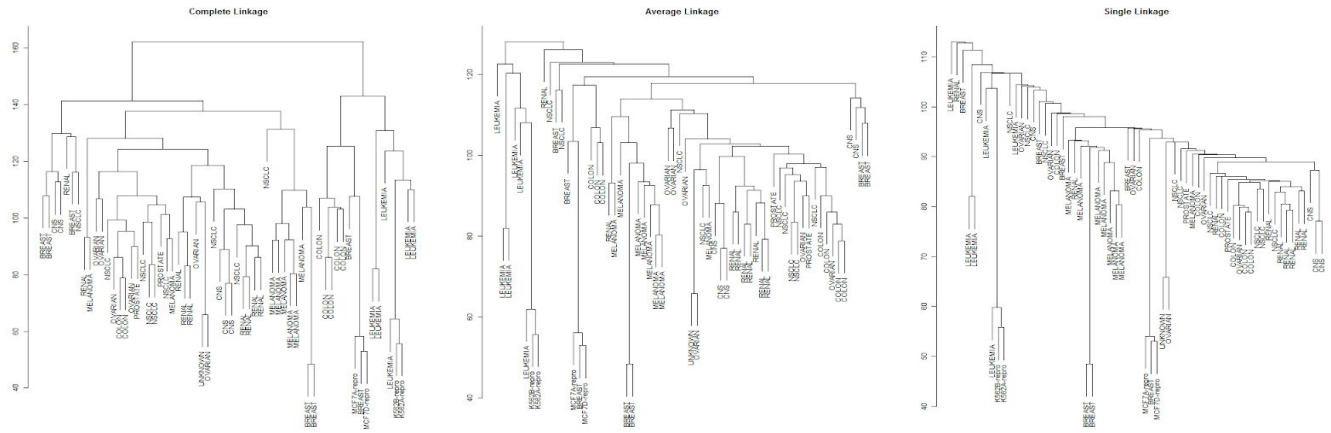
From the above figure, it is clear that cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.

The following plot gives an idea about the Proportion of Variance and the cumulative Proportion of Variance



From the above graph it is clear that the first seven principal components explain around 40 % of the variance in the data. This is not a huge amount of the variance. However, looking at the plot,it is evident that while each of the first seven principal components explain a substantial amount of variance, there is a marked decrease in the variance explained by further principal components. So, there is an elbow in the plot after approximately the seventh principal component. This suggests that there may be little benefit to examining more than seven or so principal components

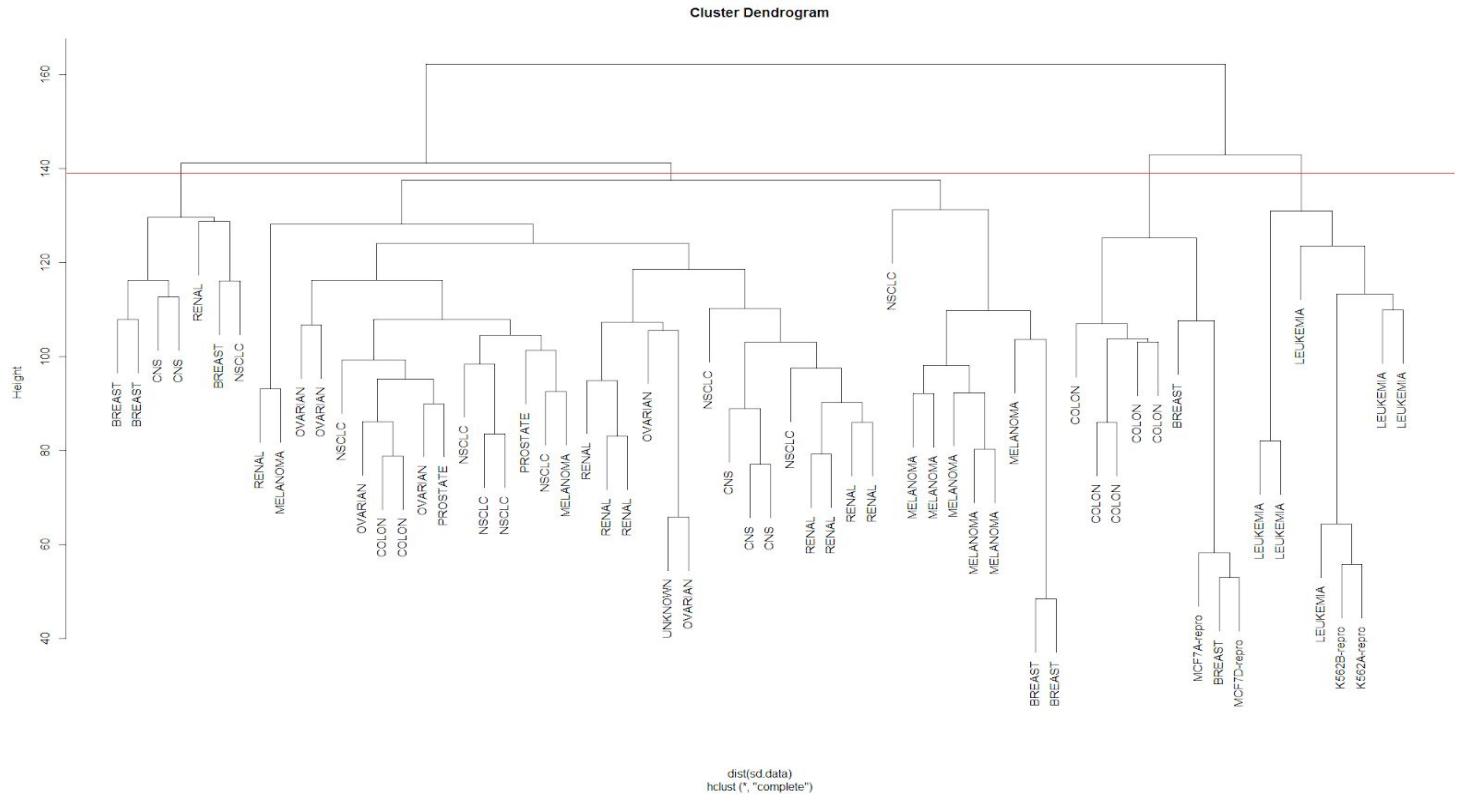
The next step is to hierarchically cluster the cell lines in the NCI60 data, with the goal of finding out whether or not the observations cluster into distinct types of cancer. First the variables to have mean zero and standard deviation one to ensure each gene to be on the same scale .The results of hierarchical clustering of the observations using complete(left), single(right), and average(center) linkage with euclidean distance is used as the dissimilarity measure is shown in the below given figure.



It is clear that single linkage will tend to yield trailing clusters: very large clusters onto which individual observations attach one-by-one. On the other hand, complete and average linkage tend to yield more balanced, attractive clusters. For this reason, complete and average linkage are generally preferred to single linkages. Further analysis is done using complete linkage hierarchical clustering. The dendrogram is cut at the height that will yield a particular number of clusters, say four: This results in the following clusters.

```
hc.clusters BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
          1      2   3     2           0           0        0           0
          2      3   2     0           0           0        0           0
          3      0   0     0           1           1        6           0
          4      2   0     5           0           0        0           1
           nci.labs
hc.clusters MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
          1           0        8     8       6        2     8       1
          2           0        0     1       0        0     1       0
          3           0        0     0       0        0     0       0
          4           1        0     0       0        0     0       0
> |
```

There are some clear patterns. All the leukemia cell lines fall in cluster 3, while the breast cancer cell lines are spread out over three different clusters. The cut on the dendrogram that produces these four clusters is plotted below.



K-means clustering and hierarchical clustering with the dendrogram cut to obtain the same number of clusters can yield very different results. This is verified for the NCI60 dataset in the table below.

km.clusters \ hc.clusters	1	2	3	4
1	11	0	0	9
2	20	7	0	0
3	9	0	0	0
4	0	0	8	0

It is observed that the four clusters obtained using hierarchical clustering and Kmeans clustering are somewhat different. Cluster 2 in K-means clustering is identical to cluster 3 in hierarchical clustering. However, the other clusters differ: for eg, cluster 4 in K-means clustering contains a portion of the observations assigned to cluster 1 by hierarchical clustering, as well as all of the observations assigned to cluster 2 by hierarchical clustering.

Chapter-5

Genomic and proteomic analysis of SARS-CoV-2 Genome

5.1 The viral genome is first downloaded from the NCBI- National Center for Biotechnology Information website. The genome downloaded is Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome which was collected by chinese researchers in Wuhan in December 2020. This is the first obtained genome of cov2 virus

Using Biopython,the important components of the genome such as name, id ,description and sequence data are extracted and displayed(shown in Figure 5.1 below)

```
C:\Users\USER> cmd /C "set "DEBUGPY_LAUNCHER_PORT=55721" && C:\Users\USER\AppData\Local\Programs\Python\Python38\python.exe c:\Users\USER\.vscod
e\extensions\ms-python.python-2020.3.69010\pythonFiles\lib\python\debugpy\no_wheels\debugpy\launcher "e:\files\stdy\MACHINE LEARNING\my experime
nts\final_project\simple_eg.py" "
Id: NC_045512.2
Name: NC_045512
Description: Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
Annotations: {'molecule_type': 'ss-RNA', 'topology': 'linear', 'data_file_division': 'VRL', 'date': '13-MAR-2020', 'accessions': ['NC_045512'],
'sequence_version': 2, 'keywords': ['RefSeq'], 'source': 'Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)', 'organism': 'Severe acu
te respiratory syndrome coronavirus 2', 'taxonomy': ['Viruses', 'Riboviria', 'Nidovirales', 'Cornidovirineae', 'Coronaviridae', 'Orthocoronaviri
nae', 'Betacoronavirus', 'Sarbecovirus'], 'references': [Reference(title='A novel coronavirus associated with a respiratory disease in Wuhan of
Hubei province, China', ...), Reference(title='Direct Submission', ...), Reference(title='Direct Submission', ...)], 'comment': "REVIEWED REFSEQ
: This record has been curated by NCBI staff. The\nreference sequence is identical to MN908947.\nOn Jan 17, 2020 this sequence version replaced
NC_045512.1.\nAnnotation was added using homology to SARSr-CoV NC_004718.3. ###\nFormerly called 'Wuhan seafood market pneumonia virus.' If you
have\nquestions or suggestions, please email us at info@ncbi.nlm.nih.gov\nand include the accession number NC_045512.### Protein structures\ncan
 be found at\nhttps://www.ncbi.nlm.nih.gov/structure/?term=sars-cov-2.### Find\nall other Severe acute respiratory syndrome coronavirus 2\n(SARS
-CoV-2) sequences at\nhttps://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/\nCOMPLETENESS: full length.", 'structured_comment': OrderedDict([('A
ssembly-Data', OrderedDict([('Assembly Method', 'Megahit v. V1.1.3'), ('Sequencing Technology', 'Illumina')]))])}
Sequence Data: ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTA
TAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAG
CCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTG
TGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCG
TAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGA
CGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGG
CTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTGCTTGGTACACGGA
ACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAA
AAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCAGACGGGCGATTTTGTTAAAGC
CACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATA
CCATAATGAATCTGGCTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTAGCGCTAACATAGGTTGTAACCA
TACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACAACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAAACTTAATGAAGAGATCGCCATTATTTTGGCATCTTTTTCTGCTTC
CACAAGTGCTTTTGTGGAAACTGTGAAAGGTTTGGATTATAAAGCATTCAAACAAATTGTTGAATCCTGTGGTAATTTTAAAGTTACAAAAGGAAAAGCTAAAAAAGGTGCCTGGAATATTGGTGAACAGAAATCAATACTGAG
TCCTCTTTATGCATTTGCATCAGAGGCTGCTCGTGTTGTACGATCAATTTTCTCCCGCACTCTTGAAACTGCTCAAAATTCTGTGCGTGTTTTACAGAAGGCCGCTATAACAATACTAGATGGAATTTCACAGTATTCACTGAG
ACTCATTGATGCTATGATGTTCACATCTGATTTGGCTACTAACAATCTAGTTGTAATGGCCTACATTACAGGTGGTGTTGTTCAGTTGACTTCGCAGTGGCTAACTAACATCTTTGGCACTGTTTATGAAAAACTCAAACCCGT
CCTTGATTGGCTTGAAGAGAAGTTTAAGGAAGGTGTAGAGTTTCTTAGAGACGGTTGGGAAATTGTTAAATTTATCTCAACCTGTGCTTGTGAAATTGTCGGTGGACAAATTGTCACCTGTGCAAAGGAAATTAAGGAGAGTGT
TCAGACATTCTTTAAGCTTGTAAATAAATTTTTGGCTTTGTGTGCTGACTCTATCATTATTGGTGGAGCTAAACTTAAAGCCTTGAATTTAGGTGAAACATTTGTCACGCACTCAAAGGGATTGTACAGAAAGTGTGTTAAATC
CAGAGAAGAAACTGGCCTACTCATGCCTCTAAAAGCCCCAAAAGAAATTATCTTCTTAGAGGGAGAAACACTTCCCACAGAAGTGTTAACAGAGGAAGTTGTCTTGAAAACTGGTGATTTACAACCATTAGAACAACCTACTAG
TGAAGCTGTTGAAGCTCCATTGGTTGGTACACCAGTTTGTATTAACGGGCTTATGTTGCTCGAAATCAAAGACACAGAAAAGTACTGTGCCCTTGCACCTAATATGATGGTAACAAACAATACCTTCACACTCAAAGGCGGTGC
```

Preliminary analysis done include using python functions in finding the length of sequence as 29,903 which agrees with what the world now knows about SARS-COV2 being an RNA molecule of about 30,000 bases.

5.2 Using Blast-Basic Local Alignment Search Tool which finds regions of similarity between biological sequence

The goal is to search for similar viral sequences in publicly available nucleotide databases. By performing an online Blast search using biopython the following search results were obtained for the above sequence.

```
query: No definition line
match: gi|1798174254|ref|NC_045512.2| Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, c
match: gi|1819735426|gb|MT121215.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/SH01
match: gi|1805293633|gb|MT019531.1| Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/I
match: gi|1820247323|dbj|LC529905.1| Severe acute respiratory syndrome coronavirus 2 TKYE6182_2020 RNA, com
match: gi|1818244627|gb|MT135044.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/235/
match: gi|1818244605|gb|MT135042.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/231/
match: gi|1818244594|gb|MT135041.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/105/
match: gi|1805293611|gb|MT019529.1| Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/I
match: gi|1802633808|gb|MN996528.1| Severe acute respiratory syndrome coronavirus 2 isolate WIV04, complete
match: gi|1818244616|gb|MT135043.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/233/
match: gi|1808633715|gb|MT049951.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/Yunn
match: gi|1805293644|gb|MT019532.1| Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/I
match: gi|1821109024|gb|MT192772.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/nCoV
match: gi|1820097838|gb|MT163719.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/WA7-
match: gi|1820097825|gb|MT163718.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/WA6-
match: gi|1819735733|gb|MT159721.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735711|gb|MT159719.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735656|gb|MT159714.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735645|gb|MT159713.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735623|gb|MT159711.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735612|gb|MT159710.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1821109035|gb|MT192773.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/nCoV
match: gi|1800455117|gb|MN988668.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU01
match: gi|1807860439|gb|MT039890.1| Severe acute respiratory syndrome coronavirus 2 isolate SNU01, complete
match: gi|1805293655|gb|MT019533.1| Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/I
match: gi|1820472829|gb|MT184912.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735667|gb|MT159715.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735567|gb|MT159709.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735506|gb|MT159708.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735484|gb|MT159707.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735473|gb|MT159706.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1812779165|gb|MT118835.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1812103009|gb|MT106053.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1800242661|gb|MN975262.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV_HKU-S
match: gi|1802471970|gb|MN994468.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1820097812|gb|MT163717.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/WA4-
match: gi|1819735744|gb|MT159722.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735700|gb|MT159718.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735689|gb|MT159717.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1806553209|gb|MT027064.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735722|gb|MT159720.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735634|gb|MT159712.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1819735443|gb|MT159705.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1815410662|gb|MT123290.1| Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/IQTC
match: gi|1806553187|gb|MT027062.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-C
match: gi|1800408777|gb|MN985325.1| Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-W
match: gi|1805293622|gb|MT019530.1| Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/I
match: gi|1803016604|gb|MT007544.1| Severe acute respiratory syndrome coronavirus 2 isolate Australia/VIC01
```

These search results were obtained after setting a low threshold for comparison. Hence by observing some of these search results, it is clear that the same viral genome was collected and uploaded by
Chinese Researchers on 5th February 2020(MT123290 https://www.ncbi.nlm.nih.gov/nuccore/1815410662/),
US Researchers on 17th February 2020(MT159705 https://www.ncbi.nlm.nih.gov/nuccore/1819735443/),
Japan on January 2020(LC529905 https://www.ncbi.nlm.nih.gov/nuccore/1820247323/),
Australia on 25th January 2020 (MT007544https://www.ncbi.nlm.nih.gov/nuccore/1803016604/) and also
HongKong on 11-Jan-2020 (MN975262 https://www.ncbi.nlm.nih.gov/nuccore/1800242661/)

Since all of these sequences are near perfect matches to the original sequence used for comparison which was collected from Wuhan in December 2019, it is possible to conclude from these search results, that the same virus observed first in Wuhan, China has been spreading across the world over the past few months. This rules out the possibility of any significant mutations in the virus, since it began its propagation through humans and hence suggests that a vaccine to the disease is both viable and possible.

5.3 Using BLAST to determine the origin of this virus

(not complete)

CONCLUSION

A

REFERENCES

- Document Cluster Mining on Text Documents- *Twinkle Svadas , Jasmin Jha*
- A Comparison of Document Clustering Techniques -Michael Steinbach, George Karypis, Vipin Kumar
- Using Word2Vec to process big text data-Long Ma ; Yanqing Zhang
- https://fasttext.cc/docs/en/supervised-tutorial.html