Inferring Phylogenetic Relationships using the Smith-Waterman Algorithm and Hierarchical Clustering

Rafael Hidalgo[1], Anthony DeVito[1], Nesreen Salah[1], Aparna S. Varde[1], Robert W. Meredith[2]
1. Department of Computer Science, 2. Department of Biology
Montclair State University, Montclair, NJ, United States
(hidalogor2 | devitoa4 | salahn1 | vardea | meredithr)@montclair.edu

Abstract— **All biological species undergo change over time due to the evolutionary process. These changes can occur rapidly and unpredictably. Due to their high potential to spread quickly, it is critical to be able to monitor changes and detect viral variants. Phylogenetic trees serve as good methods to study evolutionary relationships. Complex big data in biomedicine is plentiful in regards to viral data. In this paper, we analyze phylogenetic trees with reference to viruses and conduct dynamic programming using the Smith-Waterman algorithm, followed by hierarchical clustering. This methodology constitutes an intelligent approach for data mining, paving the way for examining variations in SARS-Cov-2, which in turn can help to discover knowledge potentially useful in biomedicine.**

Keywords—Algorithms, biological data, clustering, data analytics, knowledge discovery, healthcare, intelligent data mining, medicine, phylogeny, SARS-Cov-2, virus, virology

I. Introduction

For all of recorded human history, and far predating the time humans have walked the Earth, viruses have existed. Viruses vary in certain characteristics such as the degree to which they are infectious and/or fatal. One underlying tendency shared among all viruses is that "Viruses may evolve at high, uneven, and fluctuating rates among genome sites" [1]. Due to this tendency, viruses that start off as benign may quickly evolve into a more sinister incarnation. A dangerous virus may cause untold damage to the health of humans and animals up to and including death. It is critical to be able to study strains and variants of a virus in order to facilitate a complete understanding of their evolutionary history. Arguably, "viruses are most easily studied using phylogenetic comparison, which requires tailored questions and research methods [2]". Phylogenetics has an extensive history in the study of viruses such as Hepatitis C [3] and the Zika Virus [4]. In data science terms, this can provide complex big data, large in *volume*, having much *variety*, and requiring *veracity* in its analysis.

In this paper, we use intelligent data mining on the *RdRp* (RNA-dependent RNA polymerase) gene, a gene omnipresent and unique to RNA viruses. We institute an approach harnessing the *Smith-Waterman algorithm* in dynamic programming, followed by hierarchical clustering with *WPMGA (Weighted Pair Group Method with Arithmetic-Mean),* using a sample real dataset. Comparisons are made by calculating the edit distance between each *RdRp* gene sequence. These comparisons can potentially reveal which strains are more closely related informing the structure of the inferred phylogeny. We build on previous work [1-4] in these problem spaces by using a subset of coronavirus gene sequences. Inferences from this analysis can set the stage for further biomedical studies to detect variants.

II. Models and Methods

A. Background on Phylogentic Trees

A phylogenetic tree depicts "the relationship between biological lineages related by common descent" [1]. Thus, a phylogeny is a specific type of tree diagram that groups particular gene sequences according to their evolutionary history. Fig. 1 is an example phylogenetic tree illustrating the evolution of a four base pair gene sequence [1]. Note that *A,T,G,C* stand for adenine, guanine, thyamine, cytosine respectively (the four bases that make up DNA). In the topmost set of branches, initially the gene sequence diverges from the global MRCA or most recent common ancestor (*AAAA* to *AGAA* then to *AGGA*). There is only one mismatch or mutation between *AAAA* and *AGAA*: at the second position in each sequence. Additionally, there is only one mutation between *AGAA* and *AGGA* at the third position.

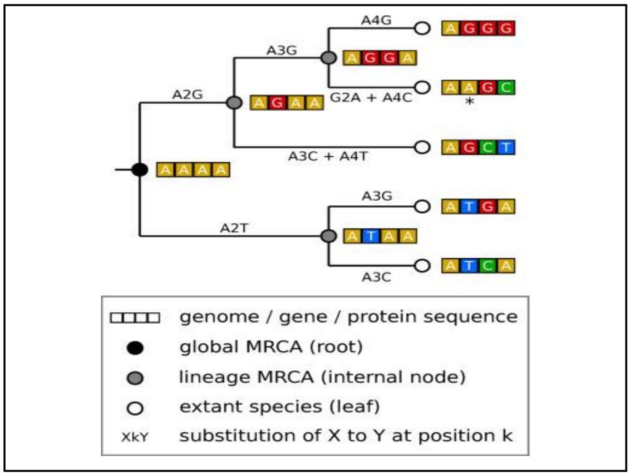


Fig. 1: An example phylogenetic tree. [1]

However, when comparing *AGGA* with the global most recent common ancestor (*AAAA*), there are two mutations at the second and third positions. Actual gene sequencs such as the *RdRp* gene are much greater in length and far more complex, constituting big data in *volume* and *variety*. Thse sequences can be compared in a manner similar to that described above, catering to *veracity*. In Fig. 1, the phylogeny is already known,

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

but there is a clear evolution marked by incremental changes to the gene sequence over time, which can be traced via edit distance. Evolutionarily closer sequences typically have fewer mismatches than more distantly related ones, as is illustrated by the gene sequences mentioned above. A phylogenetic tree “may be characterized by topology, branch lengths, and whether the tree is rooted or unrooted. Trees that share a common topology obtained using different data sets are known as congruent [1].

Branch length may be representative of either the amount of change that has occurred between nodes or the amount of time that has passed between these nodes. The former is deemed additive while the latter deemed ultrametric. The shape of the tree is indicative of the evolutionary process acting on a particular species or gene sequence. Trees may be either rooted or unrooted. Rooted trees imply directionality of change and ancestry while unrooted trees only depict the relatedness.

Phylogenetic analysis can be critically important for never-before-detected viruses, such as the SARS coronavirus and the Zika virus: in classifying the virus, determining how the virus enters the human population, and initiating the search for the virus’ natural reservoir. This natural reservoir may constitute all host organisms for a particular virus whether human or animal as well as the environments in which a virus may live to later be transmitted to a host organism (this may include: high-touch surfaces, enclosed, populated spaces, and so forth).

B. Use of Phylogenetic Trees in Coronavirus Studies

In an attempt to understand the evolution of the coronavirus, researchers can use a phylogenetic network, an example of which appears in Fig. 2. This network carries information from mitochondrial and *Y* chromosomal data to show a myriad of optimal trees. This information can then be applied in finding infectious paths that can lead to public health risks.

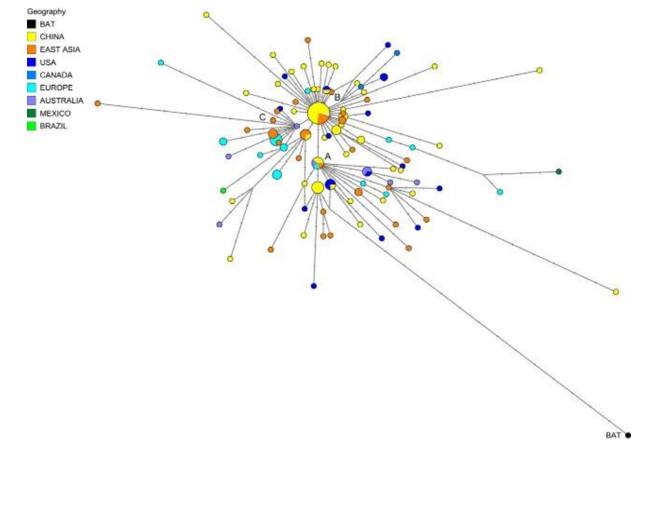


Fig. 2: Phylogenetic network of 160 SARS-CoV-2 genomes [5].

For example, the bat coronavirus has 96.2% sequence similarity to the human virus. This variant is labeled cluster A in the concerned work [5]. Two other variants are found and labeled clusters B and C. The A variant is predominant among Americans and Australians. The C variant is observed in significant proportions in Europeans. Interestingly enough, the B type remains the most prevalent variant in East Asia and does not seem to have spread. Finding the origination of three distinct variants and where they are currently located, shows that the phylogenetic network is effective in tracing the spread of infections. These findings can be critical when implementing possible treatments.

C. Other Virus Studies

The Zika virus was first isolated in 1947, and for the next 20 years the isolate was primarily obtained from East and West Africa. The Zika virus is in the genus *Flavivirus* and is transmitted by mosquitoes [4]. During that time, it was thought that infection by the Zika virus was sporadic at best. Now it is theorized that epidemics that were attributed to the dengue viruses might have been incorrect and should have been attributed to the Zika virus. The confusion may have been because of the similar clinical symptoms both infections present. Symptoms of someone infected with the Zika virus include fever, malaise, headache, maculopapular rash, and conjunctivitis. In 2007, Zika virus outbreaks began to be reported. Different genotypes of the Zika virus discovered include are: West African, East African, and Asian genotypes, as seen in Fig. 3 here [4]. In 2015, Zika virus outbreaks were reported in northern Brazil, Rio Grande State, Guatemala and Puerto Rico. Through sequence analysis and generating phylogenetic trees, researchers learned that the virus from those cases are most closely related to the Asian genotype.

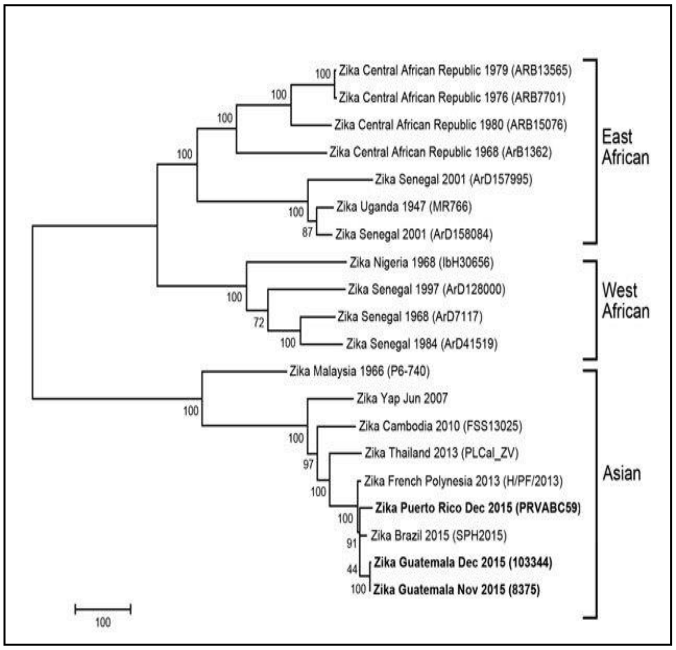

100 Zika Central African Republic 1979 (ARB13565)
100
Zika Central African Republic 1976 (ARB7701)
100
Zika Central African Republic 1980 (ARB15076)
100
Zika Central African Republic 1968 (ArB1362)
Zika Senegal 2001 (ArD157995)
100 Zika Uganda 1947 (MR766)
87 Zika Senegal 2001 (ArD158084)
East African
Zika Nigeria 1968 (IbH30656)
100
Zika Senegal 1997 (ArD128000)
72
Zika Senegal 1968 (ArD7117)
100 Zika Senegal 1984 (ArD41519)
West African
Zika Malaysia 1966 (P6-740)
Zika Yap Jun 2007
100
Zika Cambodia 2010 (FSS13025)
100
Zika Thailand 2013 (PLCal_ZV)
97
Zika French Polynesia 2013 (H/PF/2013)
100
Zika Puerto Rico Dec 2015 (PRVABC59)
100
Zika Brazil 2015 (SPH2015)
91
Zika Guatemala Dec 2015 (103344)
44
100 Zika Guatemala Nov 2015 (8375)
Asian
100

Fig. 3: Phylogenetic tree with Zika Virus isolates [4].

D. Models in Studying Viral Pathogens

There are other researchers who also use some computational methods to compare and visualize relationships of coronavirus to various pathogenic taxa [2]. This can be important when discussing the threat viral pathogens pose, especially now with coronavirus. Researchers often choose one gene from different viral sequences to build the phylogenetic tree. Sequence data can be newly generated or downloaded from the *National Center for Biotechnology Information* (NCBI database). These gene sequences are then aligned. With the alignments, phylogenetic methods such as maximum parsimony and maximum likelihood are employed to infer phylogenetic trees. This process is invaluable in understanding origins of viral outbreaks.

III. Proposed Approach: Smith-Waterman and Hierarchical Clustering

In this paper, we institute an intelligent data mining approach that harnesses the Smith-Waterman algorithm followed by hierarchical clustering for inferring phylogenetic relationships.

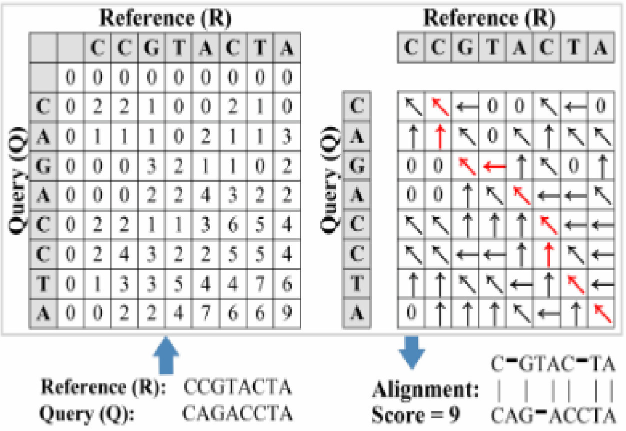
The *Smith-Waterman* algorithm in dynamic programming has been created with respect to the edit distance problem for the specific cases of comparing strings of nucleic acid sequences or protein sequences [6]. It conducts local sequence alignment; i.e. to find similar regions from two strings of nucleic acid/protein sequences. Rather than examining full sequences, it compares smaller segments of all possible lengths, for optimization [6].

In our work, the *Smith-Waterman* algorithm is adapted for data mining on a sample of the complex biological data. Nucleic acid sequences of the RNA-dependent RNA polymerase (*RdRP*) gene are used as they are uniquely identifiable for RNA based viruses such as SARS coronavirus. In our adaptation, this performs local sequence alignment on the *RdRP* gene where it examines segments of every possible length. Its advantage over global sequence alignment is that genes of varying sizes can be more seamlessly and efficiently compared. Algorithm 1 presents the pseudocode for the Smith-Waterman algorithm, adapted for our work from classical sources in Algorithms [7].

Algorithm 1: Adaptation of Smith-Waterman for Phylogeny of SARS-Cov-2

```
Input: Sequences of length M and N, Initialize the Alignment Score S=0
$S[0,0] = 0$
$for\ i = 1\ to\ M\ do$:
    $S[i, 0] = 0$
$for\ j = 1\ to\ N\ do$:
    $S[0, j] = 0$
    $for\ i = 1\ to\ M\ do$:
        $S[i,j] = MAX \begin{cases} 0 \\ S[i-1, j-1] + \delta(x_i, y_j) \\ S[i-1, j] + \delta(x_i, -) \\ S[i, j-1] + \delta(-, y_j) \end{cases}$
$return\ S[M, N]$
Output: Alignment Score S
```

This algorithm applies a scoring scheme to find the optimal matching subsequence. Adding gaps is allowed to maximize the alignment score. The scoring scheme consists of increasing the alignment score when there is a match, and decreasing it when there is a mismatch or a gap. Given this scoring scheme, a score matrix and a corresponding trace-back matrix is calculated. Once the tables are completed, the highest alignment score is recognized, and traced back to its point of origin, providing us with the sequence having the maximum alignment score. See Fig, 4 for a general graphical example of this concept [8].



		C	C	G	T	A	C	T	A
	0	0	0	0	0	0	0	0	0
C	0	2	2	1	0	0	2	1	0
A	0	1	1	1	0	2	1	1	3
G	0	0	0	3	2	1	1	0	2
A	0	0	0	2	2	4	3	2	2
C	0	2	2	1	1	3	6	5	4
C	0	2	4	3	2	2	5	5	4
T	0	1	3	3	5	4	4	7	6
A	0	0	2	2	4	7	6	6	9

	C	C	G	T	A	C	T	A
C	↖	↖	←	0	0	↖	←	0
A	↑	↑	↖	0	↖	↑	↖	↖
G	0	0	↖	←	↑	↖	0	↑
A	0	0	↑	↖	↖	←	←	↖
C	↖	↖	↑	↑	↑	↖	←	←
C	↖	↖	←	←	↑	↑	↖	←
T	↑	↑	↖	↖	←	↑	↖	←
A	0	↑	↑	↑	↖	←	↑	↖

Reference (R): CCGTACTA
Query (Q): CAGACCTA

C-GTAC-TA
Alignment: | | || | ||
Score = 9 CAG-ACCTA

Fig. 4: Tabular representation of adapting Smith-Waterman [8].

Once the maximum alignment score for every combination of genes is calculated, the scores are used to build a phylogenetic tree by the *Weighted Pair Group Method with Arithmetic-Mean (WPGMA)* hierarchical clustering technique [9]. The WPGMA technique infers *ultrametric* trees, which can be plotted as following a time axis so as to make the tips line up. It builds a *rooted tree* having edge lengths such that all leaves are equidistant from the root [9, 10]. WPGMA's rooted tree (dendrogram) corresponds to a pairwise similarity matrix. In each iteration, the two closest clusters, e.g. C_x and C_y, are united to construct a higher-level cluster, i.e. $(C_x \cup C_y)$. Its distance to another cluster C_k is the arithmetic mean of the average distances between members of (C_k, C_x) and (C_k, C_y) as stated in Equation 1 where Δ denotes distance.

$$\Delta((C_x \cup C_y), (C_k)) = \frac{\Delta(C_k, C_x) + \Delta(C_k, C_y)}{2} \quad (1)$$

This is deployed in our work as follows. Through WPGMA, we take the maximum alignment score of each combination of genes to build a similarity matrix. We take the maximum value within the matrix, and pair up the genes that are associated with the maximum alignment into a dendrogram. The pair of genes then undergo clustering in a similarity matrix. The alignment score between the clustered genes and all other genes is obtained by taking the arithmetic mean of alignment scores of clustered genes. This process is then repeated until a full dendrogram is produced, analogous to other studies in the literature on clustering [10].

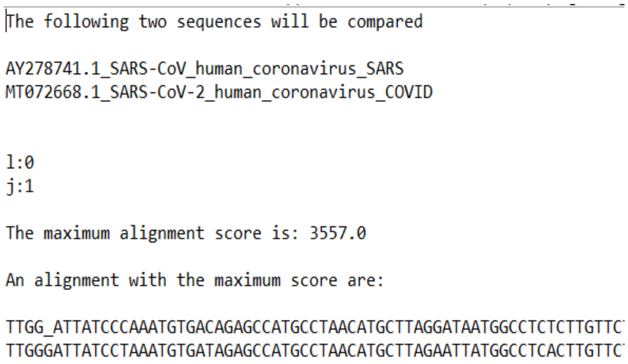
IV. Experiments and Observations

We present a synopsis of our results utilizing 12 representative DNA sequences that comprise the RNA-dependent RNA polymerase *RdRP* gene emanating from different coronavirus strains. These sample strains are presented in Listing 1.

Listing 1: Strains Discovered from Phylogeny Analysis

0: AY278741.1_SARS-CoV_human_coronavirus_SARS
1: MT072668.1_SARS-CoV-2_human_coronavirus_COVID
2: KM888179.1_hare_coronavirus
3: LC469028.1_Mi-CoV-1_bat_coronavirus
4: HQ728484.1_Miniopterus_bat_coronavirus
5: AB918718.1_IFB2012-17F_bat_coronavirus
6: NC_014470.1_BM48-31/BGR/2008_bat_coronavirus
7: KY417142.1_As6526_bat_coronavirus
8: MT084071.1_MP789_pangolin_coronavirus
9:KJ713299.1_MERS_camel_coronavirus_KSA-CAMEL-376
10: JX869059.2_MERS_human_coronavirus
11: NC_009224.1_Botryotinia_fuckeliana_totivirus

The DNA sequences have been downloaded from the *NCBI* data base as fasta files. A snippet of the output comparing the first two organisms in the fasta file can be seen in in Fig 5. Once the *Smith-Waterman* algorithm runs, the maximum alignment score is produced. The higher the score, the closer the two pairs of sequences are to one another. Afterwards, the maximum alignment score is taken and applied with WPGMA hierarchical clustering to build a phylogenetic tree. Fig. 6 illustrates the inferred phylogenetic tree from this data sample.



```
The following two sequences will be compared

AY278741.1_SARS-CoV_human_coronavirus_SARS
MT072668.1_SARS-CoV-2_human_coronavirus_COVID


l:0
j:1

The maximum alignment score is: 3557.0

An alignment with the maximum score are:

TTGG_ATTATCCCAAATGTGACAGAGCCATGCCTAACATGCTTAGGATAATGGCCTCTCTTGTTC
TTGGGATTATCCTAAATGTGATAGAGCCATGCCTAACATGCTTAGAATTATGGCCTCACTTGTTC
```

Fig. 5: Snippet of output comparing the first two organisms in the fasta file.

Modest inferences can be drawn from this analysis, and pave the way for work on a much larger scale (see Fig, 6). The *RdRP* gene of coronavirus strain 6: *NC_014470.1_BM48-31/BGR/2008_bat_coronavirus* is most similar to the *RdRp* gene of the *7:KY417142.1_As6526_ bat_coronavirus.* This correlates with the relatively high alignment score of 9352. Likewise, another *RdRP* gene of strain 9, i.e. *KJ713299.1_MERS_camel_coronavirus_KSA-CAMEL-376* is most similar to the *RdRp* gene of coronavirus strain 10: *JX869059.2_MERS_human_coronavirus.* This correlates with the relatively high alignment score of 11820. Finally, the most divergent sequence analyzed is strain 3, *LC469028.1_Mi-CoV-1_bat_coronavirus*. Strain 3, *LC469028.1_Mi-CoV-1_bat_coronavirus* has the lowest alignment scores ranging from 587 to 1627.

V. Complexity and Tractability

Given that we harness a classical algorithm from dynamic programming and apply it in the context of big data mining, it is useful to discuss complexity and tractability, since these are important aspects of algorithms [7], especially as data gets bigger in volume, variety, etc. Since the dynamic programming algorithm used in this study is configured to address the issue of edit distance, we can deduce that the problem has an overall complexity of $O(m \times n)$ where m is the number of characters in the first string (phylogeny sequence) being compared to the number of characters, n, in the second string. The algorithm compares every gene sequence to every other, two gene sequences at a time. This is useful for complex biological data, especially as much larger datasets are analyzed, following the same procedure as presented with the sample data in this paper.

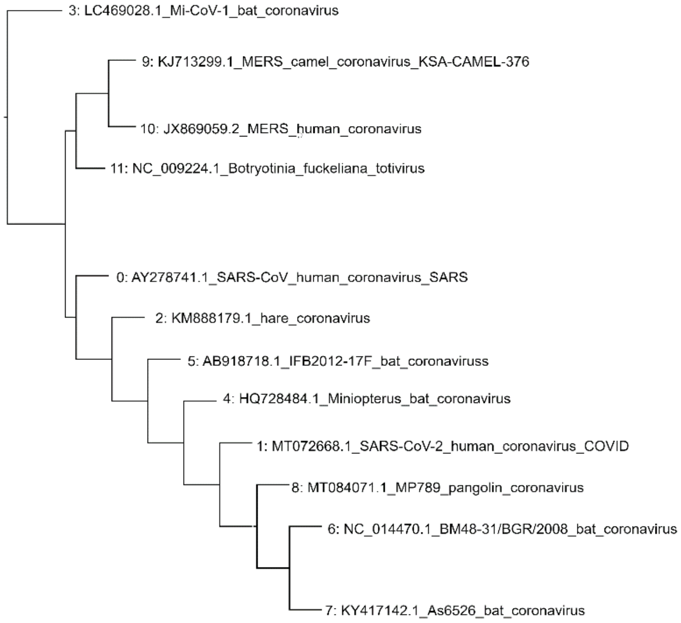

3: LC469028.1_Mi-CoV-1_bat_coronavirus
9: KJ713299.1_MERS_camel_coronavirus_KSA-CAMEL-376
10: JX869059.2_MERS_human_coronavirus
11: NC_009224.1_Botryotinia_fuckeliana_totivirus
0: AY278741.1_SARS-CoV_human_coronavirus_SARS
2: KM888179.1_hare_coronavirus
5: AB918718.1_IFB2012-17F_bat_coronavirus
4: HQ728484.1_Miniopterus_bat_coronavirus
1: MT072668.1_SARS-CoV-2_human_coronavirus_COVID
8: MT084071.1_MP789_pangolin_coronavirus
6: NC_014470.1_BM48-31/BGR/2008_bat_coronavirus
7: KY417142.1_As6526_bat_coronavirus

Fig. 6: Phylogeny tree of 12 sample *RdRp* genes from different coronaviruses generated in this study.

While trying to ascertain tractability, we classify this problem as P and NP, but not NP-Hard or NP-Complete. It is justified as follows. The problem has a poly-time algorithm, so it can be classified as P. Since P is a subset of NP, as well known in algorithmic fundamentals [7], this can also be justifiably classified as NP. Another manner in which this can be interpreted is with a poly-time certifier, which in this case is running our algorithm for a given set of gene sequences. NP-Hard problems must be intractable to the degree that no efficient algorithms exist to solve these problems. In this case, since there exists quite clearly an efficient algorithm to solve the edit distance problem as per these gene sequences, it is thus justifiable to *not* classify this as an NP-Hard problem. Finally, since NP-Complete problems must be part of both NP as well as NP-Hard problems, it is justifiable to *not* classify this problem as NP-Complete. In summary, it can be logically justified that this problem is both P and NP, but *not* NP-Hard or NP-Complete.

Note that we only address the complexity and tractability with respect to Smith-Waterman here, as this algorithm constitutes the crucial step of our problem where data *volume* matters (with other *Vs*). Once the analysis is completed using Smith-Waterman, the results are used for hierarchical clustering; hence the intermediate output passed as the input to the clustering step would obviously be smaller than the original data used by Smith-Waterman. Hence, we find it sufficient to present the complexity and tractability discussion for the dynamic programming algorithm harnessed in this study.

VI. Related Work

General Coronavirus-Related Studies: A study on chest X-ray detection with computer vision models and transfer learning [11], reveals that VGG16 and VGG19 models can detect Covid-19 vs. pneumonia vs. normal (healthy) with high accuracy, thus offering decision support in healthcare. Other efforts to combat Covid-19 in data science include creating dashboards to track and map the disease [12], and reviewing sterilization methods for PPE (personal protective equipment) [13] for SARS-Cov-2. An article on Covid-19 and social media cites numerous works as per different angles of this pandemic [14], while another one addresses veracity of postings circulated online [15]. Various mobile applications (apps) are developed that contribute to different aspects of combating the pandemic, e.g. food donation [16], recovery of small businesses [17] etc. More studies in this area [2, 5] are discussed earlier in this paper. Our work in this paper fits within the broad spectrum of such research.

Dynamic Programming and Hierarchical Clustering: Many studies, e.g. cited in [6-10] use dynamic programming or hierarchical clustering, with different applications. Other work [18] designs N-level batching with hierarchical clustering as a nonlinear integer programming method to raise efficiency via multidimensional dynamic programming, very helpful in an engineering context. Research on semantics-preserving cluster representatives is conducted [19] such that input conditions of scientific experiments constitute the data. Although this is for partition-based clustering, it can be adapted to hierarchical clustering. Another piece of work [20] finds pairwise distance of files using Smith-Waterman, and identifies relative distance of people copying them. As such duplication often occurs, this can classify proximity so that people positioned closely can be grouped by hierarchical clustering. These are examples of work on both dynamic programming and hierarchical clustering. While there are many studies in these areas, to the best of our knowledge, Smith-Waterman in dynamic programming is not prevalent with WPMGA hierarchical clustering, particularly for coronavirus studies. Thus our work modestly contributes here.

VII. Conclusions and Future Work

We deploy Smith-Waterman in conjunction with hierarchical clustering to generate a phylogenetic tree, conducing executions on real data in virus studies. Our initial work presented in this paper can potentially gain insights into a few useful inferences, e.g. the *RdRP* gene of strain 9, the *KJ713299.1_MERS_camel_coronavirus_KSA-CAMEL-376* is most similar to the *RdRp* gene of corona virus strain 10, the *JX869059.2_MERS_human_coronavirus*, and so forth.

Future research could seek to construct phylogenetic trees in a similar manner and aim to create an improved, more efficient method for studying evolutions of viruses. Comparative studies would be executed with other approaches. On the whole, our research study makes a modest contribution to computational biology through intelligent data mining.

Acknowledgments

A. Varde acknowledges NSF grant 2018575 “MRI: Acquisition of a High-Performance GPU Cluster for Research and Education”. R. Meredith acknowledges NSF grants DEB-1556701 “Collaborative Research: Advancing Bayesian phylogenetic methods for synthesizing paleontological and neontological data” and DBI-1725932 “MRI: Acquisition of a Shared Use Integrated Next-Generation Sequencing Platform”.

References

[1] A. Gorbalenya, C. Lauber, “Phylogeny of Viruses”, *Reference Module in Biomedical Sciences*, 2017, 10.1016/b978-0-12-801238-3.95723-4.

[2] N. Lorusso, M. Shumskaya, *Introduction to Phylogeny in Mega and Using Coronavirus SARS-Cov-2*. Tech Rep. Kean University, NJ, 2020.

[3] P. Jackowiak, K. Kuls, L. Budzko, A. Mania, M. Figlerowicz, M. Figlerowicz, “Phylogeny and molecular evolution of the hepatitis C virus”, *Infection, Genetics and Evolution*, 2014, 21:67-82.

[4] R. Lanciotti, A. Lambert, M. Holodniy, S. Saavedra, L. Signor, “Phylogeny of Zika Virus in Western Hemisphere, 2015”, *Emerging Infectious Diseases*, 2016, 22(5): 933-935.

[5] P. Forster, L. Forster, C. Renfrew, M. Forster, “Phylogenetic network analysis of SARS-CoV-2 genomes”, *National Academy of Sciences*, 2020, 117(17): 9241-9243.

[6] T. F. Smith, M.S. Waterman, “Identification of common molecular subsequences”. *Journal of Molecular Biology*, 1981, 147(1):195-197.

[7] J. Kleinberg and E. Tardos, *Algorithm Design*, Pearson Education, 2014.

[8] Y. Liao, Y. Li, N. Chen, Y. Lu, “Adaptively Banded Smith-Waterman Algorithm for Long Reads and Its Hardware Accelerator”, *IEEE Intl. Conf. on Application-specific Systems, Architectures and Processors*, 2018.

[9] R. Sokal, C. Michener, “A statistical method for evaluating systematic relationships”, *Univ. of Kansas Science Bulletin*, 1958, 38: 1409-1438.

[10] S. Landau, M. Leese, D. Stahl, B.S. Everitt, *Cluster Analysis*, Wiley, 2011.

[11] D. Karthikeyan, A. S. Varde, W. Wang, “Transfer learning for decision support in Covid-19 detection from a few images in big data”, *IEEE Big Data*, 2020, pp. 4873-4881, doi: 10.1109/BigData50022.2020.9377886.

[12] M. N. K. Boulos,, E. M. Geraghty, “Geographical tracking and mapping of coronavirus disease COVID-19”, *Intl J. of Health Geographics*, Springer Nature, 2020, 19(1): 8.

[13] A. Jinia, N. Sumbul, C. Meert. C. Miller, S. Clarke, K. Kearfott, M. Matsuzak, S. Pozzi, “Review of Sterilization Techniques for Medical and Personal Protective Equipment Contaminated With SARS-CoV-2*”, IEEE Access*, 2020, 8:111347-111354.

[14] M. Puri, Z, Dau, A. Varde “COVID and Social Media: Analysis of COVID-19 and Social Media Trends for Smart Living and Healthcare” *ACM SIGWEB*, Autumn 2021, 5:1–20, https://doi.org/10.1145/3494825.3494830

[15] J. Torres., V. Anu, A. Varde (2021) “Understanding the information disseminated using Twitter during the COVID-19 pandemic” IEEE IEMTRONICS, pp. 1-6, doi: 10.1109/IEMTRONICS52119.2021.9422523.

[16] C. Varghese, D. Pathak ,A. Varde, “SeVa: A Food Donation App for Smart Living”, IEEE CCWC, pp. 0408-0413, 10.1109/CCWC51732.2021.9375945.

[17] J. Torres, V. Anu, A. S. Varde and C. Duran, "My-Covid-Safe-Town: A mobile application to support post-Covid recovery of small local businesses," 2021 IEEE IEMTRONICS, doi: 10.1109/IEMTRONICS52119.2021.9422594.

[18] Seung-Kil Lim, June-Young Bang, Jae-Gon Kim, "Multidimensional Dynamic Programming Algorithm for N-Level Batching with Hierarchical Clustering Structure", *Mathematical Problems in Engineering*, 2017, Article ID 6021708, https://doi.org/10.1155/2017/6021708

[19] A. Varde, E. Rundensteiner, C. Ruiz, D. Brown, M. Maniruzzaman, R. Sisson (2006) “Designing semantics-preserving cluster representatives for scientific input conditions”, ACM CIKM, pp. 708-717.

[20] M. Phankokkruad, “Classification of file duplication by hierarchical clustering based on similarity relations”, Intl. Conf. on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2017, pp. 1598-1603.