PENGGUNAAN *N-MERS FREQUENCY* DAN ALGORITMA AGNES UNTUK PEMBENTUKAN POHON FILOGENETIK PADA VIRUS MEMATIKAN

Khoirul Umam[1)]
Riad Taufik Lazwardi[2)]
[1,2)]Fakultas Bisnis, Institut Teknologi dan Bisnis Kalbis
e-mail: khoirul.umam@outlook.com

ABSTRACT

Every organism has DNA (deoxyribonucleic acid) which carries genetic information. One of the methods for analyzing strings of DNA sequences is n-mers frequency. It is a data mining method on strings of DNA sequences that is converted into numerical data. We studied 13 deadly viruses, consisting of Rabies, HIV, Ebola, Smallpox, Marburg, Herpes B, Lujo, Avian Influenza, Spanish Flu H1N1, Dengue, HPV, SARS-CoV, and SARS-CoV-2. This study aims to establish the phylogenetic tree and find out the genetic relationship of the deadly viruses. The first method we used are collecting viral DNA sequences from the NCBI database. Afterward, the strings of DNA sequences were converted into numerical data using the n-mers frequency. After that, the dissimilarity matrix was calculated and the phylogenetic tree was established using the AGNES algorithm. Based on the phylogenetic tree, the aforementioned 13 viruses were classified into three clusters, namely cluster 1 from the realm Riboviria, cluster 2 from the realm Duplodnaviria and cluster 3 from the realm Varidnaviria. The clustering results of 13 viruses are valid because each virus is clustered based on its taxon. In addition, viruses that have the closest genetic relationship are grouped first, while viruses that have the distant genetic relationship are grouped later.

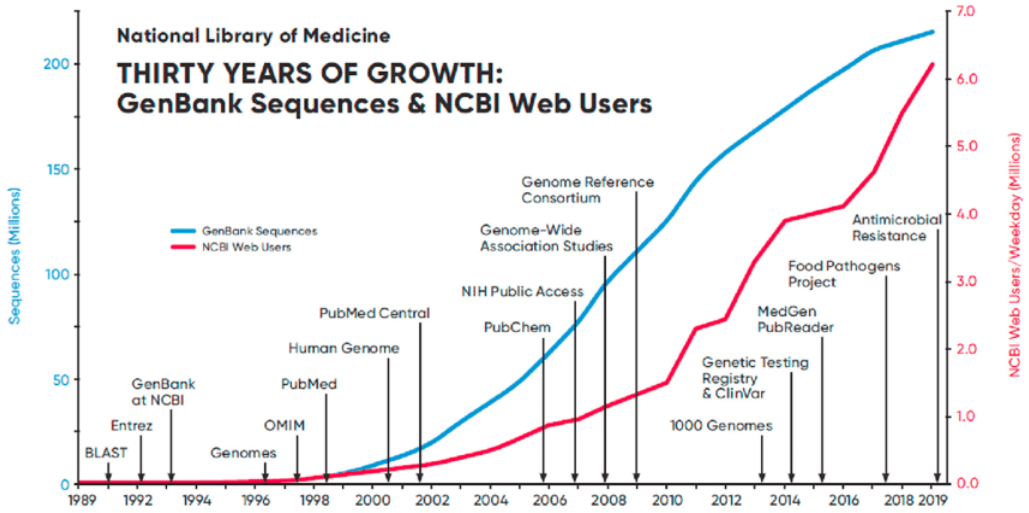
Keywords: n-mers frequency, AGNES, DNA sequences, deadly viruses

ABSTRAK

Setiap makhluk hidup mempunyai DNA (*deoxyribonucleic acid*) yang membawa informasi genetik. Salah satu metode untuk menganalisis barisan DNA yang berupa data *string* adalah menggunaan *n-mers frequency*. Metode ini merupakan metode *data mining* pada barisan DNA, dimana data barisan DNA yang merupakan data *string* diubah menjadi data numerik. Dalam penelitian ini dibahas studi genetik pada 13 virus mematikan, yakni virus Rabies, HIV, Ebola, Cacar, Marburg, Herpes B, Lujo, Flu Burung, Flu Spanyol H1N1, Dengue, HPV, SARS-CoV dan SARS-CoV-2. Penelitian ini bertujuan untuk membentuk pohon filogenetik dan mengetahui huubungan genetik dari virus mematikan tersebut. Proses awal penelitian dimulai dengan mengumpulkan barisan DNA virus dari database NCBI. Kemudian data barisan DNA diubah menjadi data numerik menggunakan *n-mers frequency*. Setelah itu, dihitung matrik disimilaritas, dilanjutkan pembentukan pohon filogenetik menggunakan algoritma AGNES (*Agglomerative Nesting*). Berdasarkan hasil penelitian ini, pohon filognetik dari 13 virus tersebut terdiri dari tiga klaster, yaitu klaster 1 yang berasal dari kelompok *Riboviria,* klaster 2 yang berasal dari kelompok *Duplodnaviria* dan klaster 3 yang berasal dari kelompok *Varidnaviria*. Hasil *clustering* 13 virus tersebut adalah valid, karena setiap virus dikelompokkan berdasarkan taksonnya. Selain itu, virus-virus yang memiliki kekerabatan paling dekat dikelompokkan terlebih dahulu, sedangkan yang kekerabatannya jauh akan dikelompokkan kemudian.

Kata kunci: *n-mers frequency*, AGNES, barisan DNA, virus mematikan

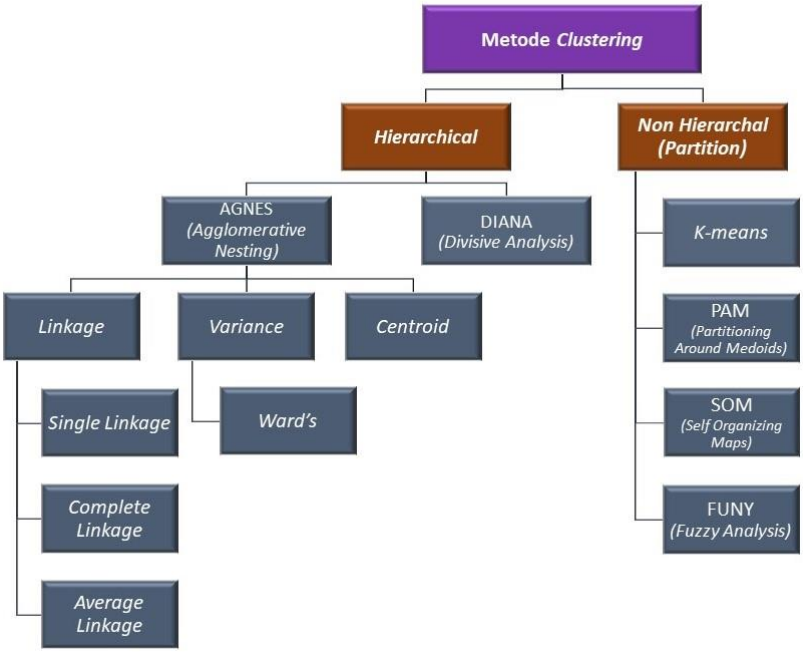
Perkembangan teknologi yang semakin canggih menyebabkan biaya *sequencing* DNA yang semakin murah. Dengan semakin murahnya biaya *sequencing* DNA mengakibatkan laju banyaknya data barisan DNA semakin tinggi (Chisholm & Wordsworth, 2016). Pada Gambar 1, menunjukkan peningkatan data barisan DNA pada database GenBank dalam 30 tahun terakhir. Pada tahun 2019 terdapat dari sekitar 215 juta barisan DNA dan terus bertambah setiap tahunnya (NLM, 2021). Pada Gambar 1 juga menunjukkan adanya peningkatan banyaknya pengguna web NCBI (*The National Center for Biotechnology Information*; https://www.ncbi.nlm.nih.gov). Hingga tahun 2019, jumlah pengguna aktif web NCBI adalah sekitar 6,2 Juta orang.



Gambar 1. Pertumbuhan data barisan DNA di Genbank dan pengguna web NCBI dari tahun 1989 sampai 2019. Sejak tahun 2000, data barisan DNA di GenBank dan pengguna web NCBI meningkat secara signifikan (NLM, 2021).

Banyaknya data barisan DNA, mendorong kegiatan *clustering* barisan DNA menjadi pekerjaan rutin dalam dunia biologi molekuler, khususnya dalam bidang terapan bioinformatika. Kegiatan *clustering* menggunakan data barisan DNA bertujuan untuk melihat pengelompokan dari objek-objek berdasarkan genomnya. Terdapat berbagai macam metode *clustering*. Berdasarkan cara membaginya, metode *clustering* terdiri dari metode *hierarchical* dan metode *partitioning*. Metode *hierarchical* digolongkan menjadi dua kelompok, yaitu AGNES *(Agglomerative Nesting)* dan DIANA *(Divisive Analysis).* Sedangkan untuk metode partitioning terdiri dari *Self Organizing Maps* (SOM), *Partitioning Around Medoid* (PAM), *Fuzzy Analysis* (FANY) dan *K-Means.* Bagan pembagian metode *clustering* dapat dilihat pada Gambar 2 (Kaufman & Rousseeuw, 1990).

Pada Gambar 2, dijabarkan bahwa pada metode AGNES dibagi menjadi tiga yaitu *Linkage, Variance* (Ward's) dan *Centroid.* Pada metode *Linkage* dibagi kembali menjadi tiga yaitu *Single Linkage, Complete Linkage*, dan *Average Linkage.* Pada penelitian ini menggunakan metode AGNES dalam melakukan *clustering*, dimana metode ini mengelompokkan dua objek yang paling mirip kedalam kelompok yang sama (Aggarwal & Reddy, 2014).



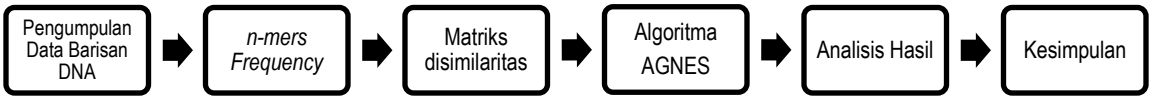
Gambar 2. Pembagian Metode *Clustering* berdasarkan cara membaginya, ada dua yaitu (1) metode *hierarchical* terdiri dari AGNES dan DIANA, dan (2) metode *nonhierarchical* terdiri dari SOM, PAM, FANY dan K-Means.

Penelitian ini melakukan *clustering* dari 13 virus mematikan. Data virus yang digunakan dalam penelitian ini berasal dari 10 virus mematikan berdasarkan artikel di cnnindonesia.com (Juniman, 2018), yaitu: (1) Rabies, (2) HIV, (3) Ebola, (4) Cacar, (5) Marburg, (6) Herpes B, (7) Lujo (Lusaka dan Johannesburg), (8) Flu Burung H5N1, (9) Flu Spanyol H1N1, dan (10) Dengue, serta 3 virus lainnya yakni: (11) HPV yang merupakan virus mematikan bagi wanita (Gollin, 2015), (12) SARS (SARS-CoV) yang mewabah tahun 2002 (Widya Putri, 2019) dan (13) Covid-19 (SARS-CoV2) yang mewabah di akhir tahun 2019 hingga saat ini (Baskara, 2020).

Pada tahap awal penelitian, salah satu metode untuk menganalisa data barisan DNA yang berupa *string* adalah menggunakan *n-mers frequency.* Metode ini mengubah data dari bentuk *string* menjadi numerik (Chor et al., 2009). Kemudian data yang sudah berupa numerik diubah menjadi bentuk matriks disimilaritas untuk dijadikan input algoritma AGNES yang outputnya berupa pohon filogenetik. Penelitian ini bertujuan untuk menggunakan *n-mers frequency* dan algoritma AGNES untuk pembentukan pohon filogenetik dari 13 virus mematikan tersebut. Pohon filogenetik yang diperoleh menggambarkan kekerabatan virus-virus tersebut berdasarkan genomnya.

METODE

Metode yang digunakan dalam penelitian ini adalah penggunaan *n-mers frequency* untuk mengubah data *string* menjadi numerik, dan penggunaan algoritma AGNES untuk membentuk pohon filogenetik. Untuk pembentukan pohon filogenetik dilakukan tahapan berikut: (1) pengumpulan data barisan DNA, (2) *n-mers frequency,* (3) matriks disimilaritas, (4) algoritma AGNES, (5) analisis hasil dan (6) kesimpulan. Tahapan tersebut dapat dilihat pada Gambar 3. Setiap perhitungan dilakukan dengan bantuan *open-source software* R.



Gambar 3. Tahapan penelitian yang dimulai dengan pengumpulan data barisan DNA, *n-mers frequency,* matriks disimilaritas, algoritma AGNES, analisis hasil dan kesimpulan.

Pengumpulan Data Barisan DNA

Data Barisan DNA merupakan barisan data *string* dari basa-basa nukleotida, yakni *Adenine* (A), *Thymine* (T), *Guanine* (G), dan *Cytosine* (C). Terdapat beberapa format barisan DNA, antara lain GenBank oleh NCBI *(National Center for Biotechnology Information),* EMBL (*European Molecular Biology Laboratory),* GCG *(Genetics Computer Sequence),* ASN.1 *(Abstract Syntax Notation One),* NBRF *(National Biomedical Research Foundation),* FASTQ dan FASTA (Mount, 2004). Format barisan yang digunakan akan menentukan software yang digunakan untuk menganalisis barisan DNA tersebut.

Format FASTA adalah salah satu format yang paling sering digunakan di berbagai *software* analisis DNA, karena pada format FASTA hanya terdiri dari deskripsi dan kode barisan DNA saja. Selain itu, format FASTA tidak terdapat nomor atau keterangan karakter lain sehingga memudahkan untuk menduplikat barisan DNA tersebut dan menggabungkan beberapa file FASTA menjadi satu file. Setiap data dalam format FASTA diawali simbol ">", satu baris deskripsi, kemudian diikuti oleh barisan basa DNA yaitu A, C, G, dan T (Mount, 2004). Contoh barisan DNA dengan format FASTA dapat dilihat pada Gambar 4.

```
>KP212153.1 Human papillomavirus type 16 isolate CNA34, complete genome
AATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAACCGGTTAGTATAAAAGCAG
ACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCCACAGGAGCGACCCAGAAAGTTACCACAG
TTATGCACAGAGCTGCAAACAACTATACATGATATAATATTAGAATGTGTGTACTGCAAGCAACAGTTAC
TGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTATATAGAGATGGGAATCCATATGC
TGTATGTGATAAATGTTTAAAGTTTTATTCTAAAATTAGTGAGTATAGACATTATTGTTATAGTTTGTAT
GGAACAACATTAGAACAGCAATACAACAAACCGTTGTGTGATTTGTTAATTAGGTGTATTAACTGTCAAA
AGCCACTGTGTCCTGAAGAAAAGCAAAGACATCTGGACAAAAAGCAAAGATTCCATAATATAAGGGGTCG
GTGGACCGGTCGATGTATGTCTTGTTGCAGATCATCAAGAACACGTAGAGAAACCCAGCTGTAATCATGC
ATGGAGATACACCTACATTGCATGAATATATGTTAGATTTGCAACCAGAGACAACTGATCTCTACTGTTA
TGAGCAATTAAATGACAGCTCAGAGGAGGAGGATGAAATAGATGGTCCAGCTGGACAAGCAGAACCGGAC
```

Gambar 4. Contoh Barisan DNA Virus HPV (*Human papillomavirus*) tipe 16 dengan Format FASTA. Format FASTA diawali simbol ">", deskripsi "KP212153.1 Human papillomavirus type 16 isolate CNA34, complete genome", kemudian barisan basa DNA yaitu A, C, G, dan T. Barisan DNA ini diperoleh dari NCBI (*The National Center for Biotechnology Information)* (NCBI, 2015).

Data barisan DNA yang digunakan adalah 13 barisan DNA virus mematikan. Data tersebut berformat FASTA yang diperoleh dari GenBank atau NCBI, dapat diakses melalui halaman situs www.ncbi.nlm.nih.gov/nuccore (NCBI, 2020). Detail data barisan DNA dari 13 virus mematikan tersebut dapat dilihat pada Tabel 1.

Tabel 1. Detail data barisan DNA dari 13 virus mematikan yang berasal dari GenBank, meliputi: nama virus, kode GenBank, panjang barisan DNA, lokasi dan waktu koleksi

No	Nama Virus	Deskripsi	Kode GenBank	Panjang (bp)	Lokasi	Waktu Koleksi
1	Rabies	Rabies lyssavirus	EU643590.1	11923	Hunan, China	2006
2	HIV	Human immunodeficiency virus 1	MN090628.1	9056	USA	2018
3	Ebola	Zaire ebolavirus	MH425138.1	18937	Zaire	9-Aug-2014
4	Cacar	Variola virus	DQ441420.1	186293	Bangladesh	1974
5	Marburg	Marburg marburgvirus	KP985768.1	19113	Uganda	2014
6	Herpes B	Human alphaherpesvirus 1	KJ847330.1	151024	India	2011
7	Lujo	Lujo mammarenavirus	NC_012777.1	7163	South Africa	2008
8	Flu Burung H5N1	Influenza A virus H5N1	AY627898.2	2341	Thailand	2008
9	Flu Spanyol H1N1	Influenza A virus H1N1	CY087030.1	2281	New York, USA	3-Jun-2009
10	Dengue	Dengue virus	KR919820.1	10724	Brunei	2014
11	HPV	Human papillomavirus type 16	KP212153.1	7894	Barazil	18-May-2010
12	SARS-CoV	Severe acute respiratory syndrome coronavirus	AY864805.1	29751	Beijing, China	2004
13	SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2	MT079854.1	29897	Wuhan, China	22-Jan-2020

Pada software R, barisan DNA yang berupa data *string*, A, C, G dan T, dibaca menggunakan fungsi *readDNAStringSet* pada library "Biostrings". *Syntax* untuk *readDNAStringSet* pada *software* R dengan data bernama "data13.txt" adalah sebagai berikut:

```
reads = readDNAStringSet("data13.txt", format="fasta")   (1)
```

n-Mers Frequency

Setelah data dibaca menggunakan fungsi *readDNAStringSet*, tahap selanjutnya penggunaan *n-mers frequency* untuk mengubah data barisan DNA yang berupa data *string* menjadi data numerik (Han et al., 2012). Metode *n-mers frequency* digunakan untuk mengetahui banyaknya kemunculan *n* basa nukleotida *(substring)* yang sama dari suatu barisan DNA. Dikarenakan ada empat basa nukleotida pada DNA (A, C, G, T), maka banyaknya pola kemunculan adalah 4^n *substring,* dengan *n* = 3, maka diperoleh $4^3 = 64$ *subtring*, yaitu: AAA, AAC, AAT, AAG, ACA, ACC, ACT, ACG, ATA, ATC, ATT, ATG, AGA, AGC, AGT, AGG, CAA, CAC, CAT, CAG, CCA, CCC, CCT, CCG, CTA, CTC, CTT, CTG, CGA, CGC, CGT, CGG, GAA, GAC, GAT, GAG, GCA, GCC, GCT, GCG, GTA, GTC, GTT, GTG, GGA, GGC, GGT, GGG, TAA, TAC, TAT, TAG, TCA, TCC, TCT, TCG, TTA, TTC, TTG, TGA, TGC, TGT, TGG, TTT (Bustamam et al., 2017).

Penentuan *n* = 3 didasari oleh proses pembentukan asam amino pada sintesis protein, barisan DNA dibaca dalam kelompok tiga nukleotida yang disebut kodon, kodon pada pita mRNA (*messenger ribonucleic acid*) hasil duplikasi dari barisan DNA dipasangkan dengan anti kodon yang dibawa tRNA (*transfer RNA*) untuk membentuk asam amino (Umam & Sagara, 2020). Menurut penelitian terdahulu penggunaan *n-mers frequency* dengan *n* = 3 mempunyai akurasi yang tinggi, yaitu 100 % (Umam & Sagara, 2020) dan 95% (Mahdiyah et al., 2019).

Pada software R, perhitungan *n-mers frequency* menggunakan library "Biostrings", yaitu fungsi *oligonucleotideFrequency*. *Syntax* untuk *n-mers frequency* dengan *n* = 3 dan 13 data barisan DNA virus mematikan adalah sebagai berikut:

```
eks = oligonucleotideFrequency(reads[1:13], width=3)
```
(2)

Matriks Disimilaritas

Setelah dilakukan *n-mers frequency*, tahapan selanjutnya adalah membentuk matriks disimilaritas atau matriks jarak. Matriks disimilaritas merupakan matriks simetri yang elemen matriksnya adalah jarak antar data (d_{ik}). Untuk mengitung d_{ik} digunakan persamaan *euclidean distance* (Kaufman & Rousseeuw, 1990), yaitu:

$$d_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{kj})^2} \tag{3}$$

Dengan d_{ik} = jarak data ke-i *dan ke-*k, m = dimensi data, x_{ij} = nilai dari data ke-i pada dimensi j, dan x_{kj} = nilai dari data ke-k pada dimensi j.

Dikarenakan data yang digunakan merupakan data barisan DNA, d_{ik} juga bisa disebut jarak genetik dari data ke-i *dan ke-*k. *Syntax* untuk menghitung matriks disimilaritas pada *software* R adalah sebagai berikut:

```
md = dist(eks, method="euclidean")
```
(4)

Algoritma AGNES

Setelah membentuk matriks disimilaritas, maka matriks tersebut digunakan sebagai input untuk algoritma AGNES *(Agglomerative Nesting)*. Algoritma AGNES mengelompokkan dua objek yang paling mirip kedalam kelompok yang sama (Aggarwal & Reddy, 2014). Algoritma AGNES dimulai dengan setiap objek dalam satu klaster yang terpisah, artinya banyaknya klaster awal sama dengan banyaknya objek, kemudian objek yang paling mirip (jaraknya paling dekat) dikelompokkan. Selanjutnya, pengelompokan terjadi terus menerus sampai semua subkelompok digabungkan menjadi satu kelompok (Bustamam et al., 2017). Hasil dari pengelompokkan menggunakan algoritma AGNES adalah pohon filoginetik atau dendogram.

Terdapat beberapa algoritma AGNES, yaitu *Single Linkage, Complete Linkage*, *Average Linkage, Variance* (Ward's) dan *Centroid* (Kaufman & Rousseeuw, 1990). Pada penelitian ini, algoritma AGNES yang dipilih adalah algoritma *average linkage* dimana algoritma ini melakukan perhitungan rata-rata jarak antara semua objek sehingga mempunyai hasil yang relatif tepat dibandingkan dengan algoritma AGNES yang lain (Kassambara, 2019). Langkah-langkah algoritma AGNES-*Average Linkage* dapat dilihat pada Tabel 2 (Bustamam et al., 2017).

Tabel 2. Algoritma AGNES-*Average Linkage* dengan *input* matriks disimilaritas dan *output* Dendogram

Algoritma 3. Algoritma AGNES-*Average Linkage*

Input : Matriks disimilaritas

Output : Dendogram

Steps:

1) Gabungkan dua kelompok yang terdekat, misalkan jarak kelompok U dan V adalah yang paling dekat, namakan dengan $d_{(UV)}$.
2) Gabungkan kelompok U dan V, namakan kelompok baru tersebut dengan (UV).
3) Hitung jarak antara (UV) dan kelompok baru (W) dengan persamaan berikut:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \quad (5)$$

Keterangan:

Σ_i = penjumlahan jarak objek *i* dalam kelompok *(UV)*

Σ_k = penjumlahan jarak objek *k* dalam kelompok *W*

d_{ik} = jarak antara objek *i* dalam kelompok *(UV)* dan objek *k* dalam kelompok *W*,

N_{UV} = banyaknya data dalam kelompok *(UV)*

N_w = banyaknya data dalam kelompok *W*

4) Ulangi langkah 1) sampai 3) sampai memperoleh klaster yang dikehendaki atau semua data telah berada dalam kelompok tunggal.

Syntax untuk menghitung algoritma AGNES-*Average Linkage* pada *software* R adalah sebagai berikut:

```
A = agnes(data, metric = "euclidean", method = "average")    (6)
```

Analisis Hasil

Setelah diperoleh pohon filogenetik dari algoritma AGNES, selanjutnya dilakukan analisis secara deskriptif virus mana sajakah yang mempunyai similaritas secara genetik. Kemudian, dilihat juga klasifikasi dan taksonomi dari virus yang mempunyai similaritas tersebut.

HASIL DAN PEMBAHASAN

Data Penelitian

Tiga belas data barisan DNA virus mematikan yang berupa data *string* dibaca menggunakan *software* R dengan fungsi *readDNAStringSet* pada library "Biostrings". Hasil *readDNAStringSet* berupa tabel yang memisahkan panjang barisan DNA *(width),* barisan DNA *(seq)* dan deskripsi data *(name).* Hasil *readDNAStringSet* dapat dilihat pada Tabel 3.

Tabel 3. Hasil *readDNAStringSet* dari 13 data barisan DNA virus mematikan yang berupa panjang barisan DNA *(width),* barisan DNA *(seq)* dan deskripsi data *(name)*.

No	*width*	*Seq*	*Name*
1	11923	ACGCTTAACAACCAGATCAAAGAAGAAGCAGA…	EU643590.1 Rabies virus
2	9056	GACCTGAAAGCGAAAGAGAAACCAGAGGAGC…	MN090628.1 HIV-1
3	18937	GAAAGAAGAATTTTTAGGATCTTTTGTGTGCGA…	MH425138.1 Zaire ebolavirus
4	186293	GTGTCTAGAAAAAAATGTGTGACCCACGACTG…	DQ441420.1 Variola virus
5	19113	AGACACACAAAAACAAGAGATGATGATTTTGT…	KP985768.1 Marburg
6	151024	CAGCCCGGGCCCCCCGCGGGCGCGCGCGCG…	KJ847330.1 Human herpes virus 1
7	7163	GTTAATTGGACTACTTTTCTAGTACCTCACTTC…	NC_012777.1 Lujo virus
8	2341	AGCGAAAGCAGGTCAAATATATTCAATATGGA…	AY627898.2 Influenza A virus H5N1
9	2281	TTGAATGGATGTCAATCCGACTCTAATTTTCCT…	CY087030.1 Influenza A virus H1N1
10	10724	AGTTGTTAGTCTACGTGGACCGACAAGAACAG…	KR919820.1 Dengue virus
11	7894	AATAATTCATGTATAAAACTAAGGGCGTAACCG…	KP212153.1 Human papillomavirus type 16
12	29751	ATATTAGGTTTTTACCTACCCAGGAAAAGCCAA…	AY864805.1 SARS-CoV
13	29897	CGAACCTGTAAAACAGGCAAACTGAGTTGGAC…	MT079854.1 SARS-CoV-2

Hasil *n-mers Frequency*

Untuk mengubah 13 data barisan DNA virus mematikan tersebut yang berupa data *string* menjadi data numerik digunakan *n-mers frequency* dengan *n* = 3. Hasil *n-mers Frequency* berupa matriks yang berukuran 13 x 64, dapat dilihat pada Gambar 5.

	AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT	AGA	AGC	AGG	AGT	ATA	ATC	ATG	A
1	266	181	270	192	211	152	71	213	379	132	252	174	183	249	252	
2	384	159	281	267	277	120	16	121	356	198	234	174	225	100	170	
3	629	417	405	532	497	278	129	324	411	243	296	290	363	372	337	
4	7099	3446	2851	6758	3755	1503	1848	2978	4175	1253	1400	3025	7884	4368	3940	
5	686	413	438	586	488	237	101	329	418	211	279	288	428	377	363	
6	1399	1636	1266	595	1635	2695	3011	834	1282	2271	2704	872	691	1300	1370	
7	164	127	139	179	148	106	23	148	160	94	68	118	103	163	138	
8	71	48	67	68	55	25	16	32	75	38	59	37	40	42	57	
9	88	49	69	74	74	31	10	31	85	27	40	35	44	37	70	
10	384	237	277	210	300	141	82	167	326	165	255	147	158	135	292	
11	299	166	120	203	266	102	56	142	138	74	109	131	265	69	194	
12	762	537	562	653	781	389	156	656	537	350	416	448	396	335	775	
13	894	617	580	760	809	376	165	675	604	300	329	509	470	339	724	

Showing 1 to 13 of 13 entries, 64 total columns

Gambar 5. Hasil *n-mers frequency* dari 13 data barisan DNA virus mematikan yang membentuk matriks berukuran 13 x 64.

Hasil Perhitungan Matriks Jarak

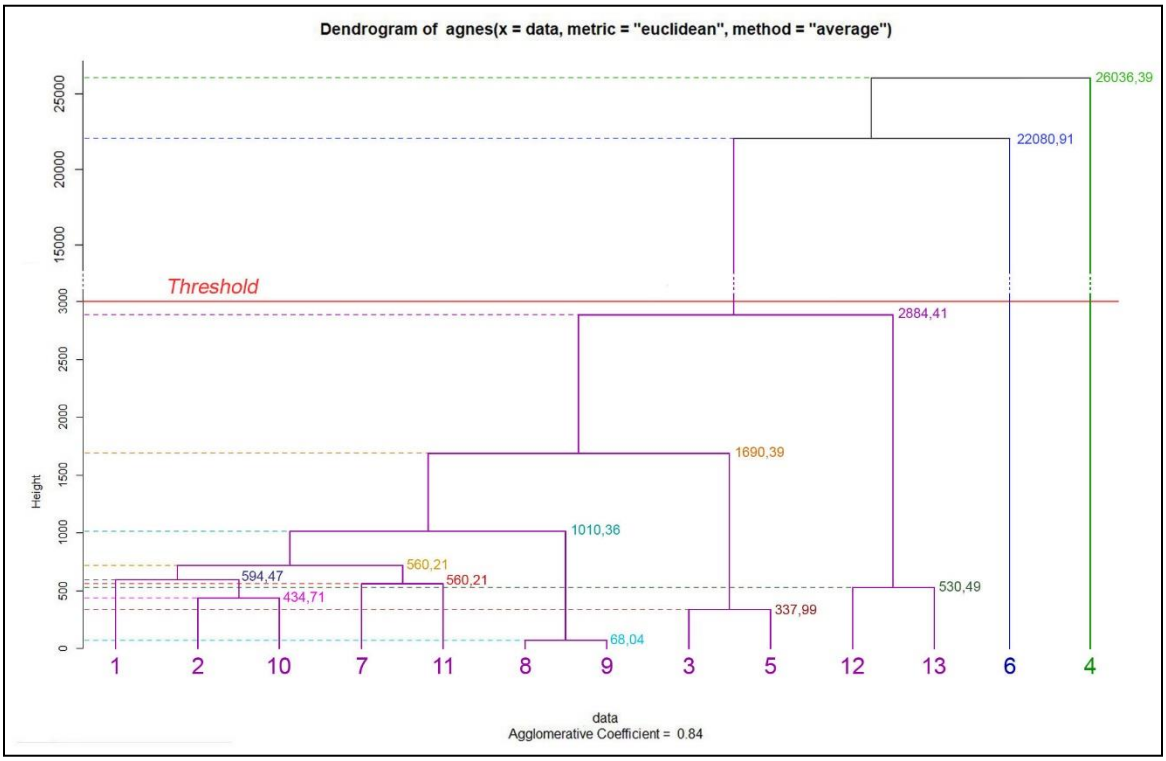
Untuk membentuk matriks disimilaritas digunakan persamaan *Euclidean distance.* Pada Gambar 6 merupakan hasil perhitungan matriks jarak dari 100 barisan DNA menggunakan *software* R, diperoleh matriks berukuran 13 x 13.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.0000	682.5504	1134.369	26319.85	1271.832	22062.15	707.6779	1287.86723	1292.22328	506.3862	853.8296	2633.5349	2812.9678
2	682.5504	0.0000	1403.865	26562.69	1524.579	22355.71	679.1384	1004.31121	1003.33444	434.7091	580.5067	2943.0289	3075.2088
3	1134.3694	1403.8650	0.000	25328.10	337.997	21614.20	1617.7620	2241.96209	2240.07433	1255.5983	1540.9293	1727.2782	1881.0428
4	26319.8504	26562.6905	25328.097	0.00	25161.888	26485.67	26792.1151	27396.16077	27388.99995	26472.0042	26588.1524	24087.0299	23858.8053
5	1271.8317	1524.5790	337.997	25161.89	0.000	21717.20	1716.0559	2348.39264	2344.40462	1410.2939	1615.3083	1667.7122	1755.5438
6	22062.1539	22355.7120	21614.200	26485.67	21717.203	0.00	22536.1398	22893.33870	22911.46464	22144.0725	22522.3345	20991.3211	21142.0781
7	707.6779	679.1384	1617.762	26792.12	1716.056	22536.14	0.0000	719.05772	714.52432	739.3707	560.2062	3113.3371	3255.1025
8	1287.8672	1004.3112	2241.962	27396.16	2348.393	22893.34	719.0577	0.00000	68.03675	1167.5277	875.0126	3753.2128	3890.6562
9	1292.2233	1003.3344	2240.074	27389.00	2344.405	22911.46	714.5243	68.03675	0.00000	1169.8547	870.0448	3752.4103	3887.2691
10	506.3862	434.7091	1255.598	26472.00	1410.294	22144.07	739.3707	1167.52773	1169.85469	0.0000	749.1181	2755.5477	2925.3620
11	853.8296	580.5067	1540.929	26588.15	1615.308	22522.33	560.2062	875.01257	870.04483	749.1181	0.0000	2992.5482	3097.7019
12	2633.5349	2943.0289	1727.278	24087.03	1667.712	20991.32	3113.3371	3753.21276	3752.41029	2755.5477	2992.5482	0.0000	530.4903
13	2812.9678	3075.2088	1881.043	23858.81	1755.544	21142.08	3255.1025	3890.65624	3887.26909	2925.3620	3097.7019	530.4903	0.0000

Gambar 6. Hasil perhitungan matriks disimilaritas dari 13 data barisan DNA virus mematikan yang berupa matriks berukuran 13 x 13.

Hasil *Clustering* Menggunakan Algoritma AGNES

Setelah diperoleh hasil perhitungan matriks disimilaritas, tahap selanjutnnya adalah *clustering* atau pengelompokan menggunakan algoritma AGNES. Hasil *clustering* menggunakan algoritma AGNES berupa pohon filogenetik atau dendogram yang dapat dilihat pada Gambar 7. Sebuah *threshold* ditentukan untuk membagi dendogram menjadi beberapa klaster. Bayaknya klaster ditentukan sebanyak 3 klaster, dan hal ini didasari oleh banyaknya jenis *realm* (tingkatan taksonomi virus) pada data 13 barisan virus mematikan tersebut, yaitu *realm Riboviria, Duplodnaviria* dan *Varidnaviria*.



Gambar 7. Hasil algoritma AGNES yang mengelompokkan 13 data barisan DNA virus mematikan, yakni: Rabies (1), HIV (2), Ebola (3), Cacar (4), Marburg (5), Herpes B (6), Lujo (7), Flu Burung H5N1 (8), Flu Spanyol H1N1 (9), Dengue (10), HPV (11), SARS-CoV (12), dan SARS-CoV-2 (13). Dengan nilai *threshold* sebesar 3000, dendogram terbagi menjadi 3 klaster. Klaster 1 beranggotakan 1, 2, 10, 7, 11, 8, 9, 3, 5, 12, 13; Klaster 2 beranggotakan 6; dan Klaster 3 beranggotakan 4.

Analisis Hasil

Pohon filogenetik pada Gambar 7 merupakan hasil pengelompokan dari 13 virus mematikan menggunakan algoritma AGNES-*average lingkage* dengan nilai *agglomerative coefficient* = 0.84. Nilai *agglomerative coefficient* berkisar antara 0 sampai 1, semakin dekat nilai koefisien dengan 1 artinya semakin akurat hasil pengelompokan data, sebaliknya semakin mendekati 0 artinya semakin tidak akurat hasil pengelompokannya (Kassambara, 2019). Dengan nilai *agglomerative coefficient* = 0.84, maka dendogram hasil pengelompokannya tergolong akurat.

Penggunaan *n-mers frequency* dalam penelitian ini mempunyai peran penting sehingga pengelompokan yang terbektuk akurat. Hal ini didukung oleh penelitian yang dilakukan Mahdiyah et al. yang menggunakan *n-mers frequency* dalam menganalisis barisan DNA, dengan n = 3 atau 3-*mers* diperoleh akurasi sekitar 95% (Mahdiyah et al., 2019). Hal tersebut sejalan dengan penelitian yang dilakukan Umam dan Sagara yang mengaplikasikan *n-mers frequency* dalam *clustering* data barisan DNA, yang memperoleh akurasi 100% dengan sampel 100 data barisan DNA (Umam & Sagara, 2020). Salah satu keunggulan penggunaan *n-mers frequency* dibandingkan metode klasik pensejajaran sekuens dalam analisis barisan DNA adalah *n-mers frecuency* mampu untuk menganalisis berbagai data tidak terbatas pada data yang mempunyai kemiripan saja, sehingga *n-mers frequency* dapat digunakan dalam menganalisis barisan DNA antar virus, sedangkan pensejajaran sekuens terbatas untuk data barisan DNA yang memiliki kemiripan saja, sehingga cocok untuk menganalisis kekerabatan untuk virus-virus yang sejenis.

Selain itu, penggunaan algoritma *average linkage* (Algoritma AGNES) juga berperan penting. Dibandingkan dengan algoritma DIANA *(Divisive Analysis)* ataupun algoritma AGNES lain, hasil dari algortima *average linkage* relatif lebih akurat karena algortima *average linkage* melakukan perhitungan rata-rata antar objek dalam melakukan pengelompokan (Kassambara, 2019). Hal ini juga didukung oleh penelitian yang dilakukan Bustamam et al. yang melakukan pengelompokan bakteri pada air liur menggunakan algoritma *single linkage, complete linkage,* dan *average linkage*. Berdasarkan penelitian tersebut, pengelompokan dengan algoritma *avegare lingkage* merupakan hasil pengelompokan paling akurat, dimana hal ini ditunjukkan oleh nilai indeks Davies-Bouldin yang paling kecil (Bustamam et al., 2017).

Berdasarkan Gambar 7, pasangan virus yang pertama kali dikelompokkan adalah Flu Burung H5N1 (8) dan Flu Spanyol H1N1 (9), yang artinya kedua virus tersebut memiliki similaritas paling tinggi secara genetik. Jarak genetik kedua virus tersebut adalah 68,04 yang merupakan jarak genetik terdekat diantara 13 virus pada penelitian ini. Semakin kecil nilai jarak genetik maka semakin similar/berkerabat, sebaliknya semakin besar nilai jarak genetik maka semakin jauh kekerabatannya. Dengan kata lain, virus Burung H5N1 (8) dan Flu Spanyol H1N1 (9) mempunyai hubungan kekerabatan paling dekat diantara 13 virus pada penelitian ini, dan hal ini didukung oleh data dari *International Committee on Taxonomy of Viruses* (ICTV) yang mengelompokkan Virus Flu Burung H5N1 (8) dan Flu Spanyol H1N1 (9) ke dalam *spesies* yang sama yaitu *Influenza A virus* (ICTV, 2020). Kesesuaian hasil pengelompokan dan data dari ICTV, menunjukkan metode yang digunakan untuk pengelompokan

pada penelitian ini adalah valid. Hal ini juga dapat dilihat pada pengelompokan selanjutnya yang mengelompokkan dari takson yang rendah ke takson yang lebih tinggi.

Pengelompokan kedua berdasarkan Gambar 7 adalah Ebola (3) dan Marburg (5) dengan jarak genetik = 337,99. Ebola dan Marburg memiliki kemiripan, keduanya menyebabkan demam hemoragik. Demam hemoragik adalah demam tinggi yang disertai pendarahan, serta dapat menyebabkan kegagalan organ, syok, dan kematian (Prasetyanto, 2020). Kedua virus ini berasal dari *family* yang sama yaitu *Filoviridae* (ICTV, 2020).

Pengelompokan ketiga pada Gambar 7 adalah HIV (2) dan Dengue (10) dengan jarak genetik = 434,71. Kedua virus ini berasal dari *realm* yang sama, yaitu *Riboviria* (ICTV, 2020). Sedangkan pengelompokan keempat adalah SARS-CoV (12) dan SARS-CoV-2 (13) dengan jarak genetik = 530,49. Kedua virus ini berasal dari kelompok yang sama, yakni *Severe acute respiratory syndrome–related coronavirus* (ICTV, 2020).

Pengelompokan kelima berdasarkan Gambar 7 adalah Lujo (7) dan HPV (11) dengan jarak genetik = 560,21. Lujo berasal dari *realm Riboviria,* sedangkan HPV berasal dari *realm Monodnaviria* (ICTV, 2020). Kemudian Pengelompokan keenam berdasarkan Gambar 7 adalah Rabies (1) yang dikempokan ke HIV (2) dan Dengue (10), sama dengan HIV dan Dengue, Rabies juga berasal dari *realm Riboviria* (ICTV, 2020). Pengelompokan selanjutnya adalah pengelompokan kembali pasangan-pasangan virus yang telah disebutkan sehingga menjadi kelompok yang lebih besar, yaitu kelompok (1, 2, 10) dikelompokkan dengan kelompok (7, 11), kemudian kelompok (1, 2, 10, 7, 11) dikelompokkan dengan kelompok (8, 9), selanjutnya kelompok (1, 2, 10, 7, 11, 8, 9) dikelompokkan dengan kelompok (3, 5) dan seterusnya. Pengelompokan akan selesai saat setiap virus menjadi satu kelompok atau dendogram yang utuh.

Kemudian ditentukan *threshold* untuk membagi dendogram menjadi 3 klaster. Untuk membagi dendogram menjadi 3 klaster ditentukan nilai *threshold* antara 2884,41 sampai 22080,91. Nilai 2884,41 merupakan nilai rata-rata dari jarak antara kelompok (1, 2, 10, 7, 11, 8, 9, 3, 5) dan (12, 13), sedangkan nilai 22080,91 merupakan nilai rata-rata dari jarak antara kelompok (1, 2, 10, 7, 11, 8, 9, 3, 5, 12, 13) dan (6). Dengan demikian, ditentukan nilai *threshold* sebesar 3000 untuk membagi dendogram menjadi 3 klaster.

Berdasarkan penelitian ini, klaster 1 berasal dari *realm Riboviria* beranggotakan: Rabies (1), HIV (2), Dengue (10), Lujo (7), HPV (11), Flu Burung H5N1 (8), Flu Spanyol H1N1 (9), Ebola (3), Marburg (5), SARS-CoV (12), dan SARS-CoV-2 (13). Pengecualian untuk HPV (11), HPV (11) berasal dari *realm Monodnaviria* yang merupakan realm baru, perluasan dari *realm Riboviria* (Walker et al., 2020). *Realm Riboviria* merupakan tingkatan taksonomi yang mencakup semua virus RNA, yakni virus yang memiliki RNA sebagai materi genetiknya. Virus pada kelompok ini memiliki enzim *RNA-dependent RNA polymerases* (RdRp) untuk proses replikasinya (Gorbalenya et al., 2018).

Sementara itu, klaster 2 berasal dari *realm Duplodnaviria* yang beranggotakan Herpes B (6). *Realm Duplodnaviria* merupakan tingkatan taksonomi yang mencakup semua virus DNA, yakni virus yang memiliki materi genetik berupa DNA dan bereplikasi menggunakan DNA polimerase. Pada siklus hidup kelompok virus ini, DNA untai ganda menyandikan protein kapsid utama HK97(HK97-MCP) dan subunit terminase kecil yang diperlukan untuk pengemasan DNA menjadi kapsid (Koonin et al., 2019a). Dan terakhir, Klaster 3 berasal dari *realm Varidnaviria* beranggotaan Cacar (6) atau *Variola virus*. *Realm Varidnaviria* merupakan tingkatan taksonomi yang mencakup semua virus DNA yang mengkode protein kapsid utama *jelly roll* vertikal (Koonin et al., 2019b).

Dari penjelasan diatas, dapat disimpulkan bahwa *clustering* 13 virus mematikan menggunakan *n-mers frequency* dan algoritma AGNES pada penelitian ini, menghasilkan pohon filogenetik yang

dibagi menjadi 3 klaster. Hasil pengelompokan menggunakan *n-mers frequency* dan algoritma AGNES ini adalah valid. Hal ini dapat dilihat dari 13 virus tersebut mengelompok berdasarkan taksonnya, dan juga virus-virus yang mempunyai kekerabatan yang paling dekat dikelompokan terlebih dahulu, sebaliknya yang mempunyai kekerabatan yang lebih jauh dikelompokkan kemudian. Hal tersebut sesuai dengan penelitian terdahulu oleh Bustamam dkk. tentang penggunaan *n-mers frequency* dan algoritma AGNES dalam pengelompokkan bakteri pada air liur (Bustamam et al., 2017). Metode pengelompokan menggunakan *n-mers frequency* dan algoritma AGNES ini bisa digunakan sebagai alternatif dalam pengelompokan virus, bakteri atau makhluk hidup lain berdasarkan genetiknya.

SIMPULAN

Berdasarkan hasil pengelompokan menggunakan *n-mers frequency* dan algoritma AGNES pada 13 virus mematikan, didapat pohon filogenetik yang terbagi menjadi 3 kluster. Klaster 1 terdiri dari 11 virus yang berasal dari *realm Riboviria*, yaitu virus Rabies, HIV, Dengue, Lujo, HPV, Flu Burung H5N1, Flu Spanyol H1N1, Ebola, Marburg, SARS-CoV, SARS-CoV-2. Sedangkan Klaster 2 terdiri dari virus Herpes B yang berasal dari *realm Duplodnaviria,* dan terakhir Klaster 3 terdiri dari virus Cacar atau *Variola virus* yang berasal dari *realm Varidnaviria*. Hasil pengelompokan ini valid, karena setiap virus mengelompok berdasarkan dengan taksonnya dan virus-virus yang mempunyai kekerabatan terdekat dikelompokkan terlebih dahulu, sedangkan yang kekerabatannya jauh dikelompokkan kemudian. Dengan demikian, penggunaan *n-mers frequency* dan algoritma AGNES dalam membentuk pohon filogenetik pada penelitian ini berpotensi untuk dapat digunakan juga dalam pengelompokkan virus atau organisme lainnya.

REFERENSI

Aggarwal, C. C., & Reddy, C. K. (2014). Data Clustering Algorithms and Applications. In *CRC Press*.

Baskara, B. (2020, April 18). Rangkaian Peristiwa Pertama Covid-19. *Kompas.Com*. https://bebas.kompas.id/baca/riset/2020/04/18/rangkaian-peristiwa-pertama-covid-19/

Bustamam, A., Fitria, I., & Umam, K. (2017). *Application of Agglomerative Clustering for Analyzing Phylogenetically on Bacterium of Saliva*. *030126*, 030126–1. https://doi.org/10.1063/1.4991230

Chisholm, S. J., & Wordsworth, S. (2016). *Annual Report of the Chief Medical Officer 2016*.

Chor, B., Horn, D., Goldman, N., Levy, Y., & Massingham, T. (2009). Genomic DNA k-mer spectra: Models and modalities. *Genome Biology*, *10*(10). https://doi.org/10.1186/gb-2009-10-10-r108

Gollin, S. M. (2015). Epidemiology of HPV-Associated Oropharyngeal Squamous Cell Carcinoma. In *Human Papillomavirus (HPV)-Associated Oropharyngeal Cancer* (pp. 1–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-21100-8_1

Gorbalenya, A. E., Krupovic, M., Siddell, S., Varsani, A., & Kuhn, J. H. (2018). *Riboviria: establishing a single taxon that comprises RNA viruses at the basal rank of virus taxonomy*. International Committee on Taxonomy of Viruses (ICTV). https://talk.ictvonline.org/taxonomy/p/taxonomy-history?taxnode_id=202007095

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques Third Edition*. Morgan Kaufmann Publishers.

ICTV. (2020). *International Committee on Taxonomy of Viruses (ICTV)*. https://talk.ictvonline.org/taxonomy/

Juniman, P. T. (2018, February 21). 10 Virus Paling Mematikan di Dunia. *CNN Indonesia*. https://www.cnnindonesia.com/gaya-hidup/20180220211154-255-277570/10-virus-paling-

mematikan-di-dunia

Kassambara, A. (2019). *Hierarchical Clustering in R: The Essentials*. https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data : an introduction to cluster analysis*. Wiley.

Koonin, E., Dolja, V., Krupovic, M., Varsani, A., Wolf, Y., Yutin, N., Zerbini, M., & Kuhn, J. (2019a). *Create a megataxonomic framework, filling all principal/primary taxonomic ranks, for dsDNA viruses encoding HK97-type major capsid proteins*. International Committee on Taxonomy of Viruses (ICTV). https://talk.ictvonline.org/taxonomy/p/taxonomy-history?taxnode_id=202007117

Koonin, E., Dolja, V., Krupovic, M., Varsani, A., Wolf, Y., Yutin, N., Zerbini, M., & Kuhn, J. (2019b). *Create a megataxonomic framework, filling all principal taxonomic ranks, for DNA viruses encoding vertical jelly roll-type major capsid proteins*. International Committee on Taxonomy of Viruses (ICTV). https://talk.ictvonline.org/taxonomy/p/taxonomy-history?taxnode_id=202008702

Mahdiyah, U., Wahyuniar, L. S., Rochana, S., Informatika, T., Teknik, F., & Kediri, K. (2019). KLASIFIKASI DNA MENGGUNAKAN FITUR N-MERS DENGAN INTEGRASI. *JOUTICA*, *4*(2), 225–228.

Mount, D. W. (2004). *Bioinformatics : sequence and genome analysis*. Cold Spring Harbor Laboratory Press.

NCBI. (2015). *Human papillomavirus type 16 isolate CNA34, complete genome*. https://www.ncbi.nlm.nih.gov/nuccore/KP212153.1/

NCBI. (2020). *Nucleotide - National Center for Biotechnology Information*. https://www.ncbi.nlm.nih.gov/nuccore

NLM. (2021). *Congressional Justification FY 2021 - Department of Health and Human Services National Institutes of Health National Library of Medicine (NLM)*. https://www.nlm.nih.gov/about/2021CJ.html

Prasetyanto, A. (2020, April 20). 12 Virus Paling Mematikan di Dunia: Corona hingga Ebola. *Kumparan.Com*. https://kumparan.com/kumparansains/12-virus-paling-mematikan-di-dunia-corona-hingga-ebola-1tFSaC1xXwG/full

Umam, K., & Sagara, R. (2020). Penggunaan N-mers Frequency pada Analisis Barisan DNA. *Jambura Journal of Mathematics*, *2*(2), 73–86. https://doi.org/10.34312/jjom.v2i2.4320

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Dempsey, D. M., Dutilh, B. E., Harrach, B., Harrison, R. L., Hendrickson, R. C., Junglen, S., Knowles, N. J., Kropinski, A. M., Krupovic, M., & Kuhn, J. H. (2020). Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020). *Archives of Virology*, *165*(11), 2737–2748. https://doi.org/10.1007/s00705-020-04752-x

Widya Putri, A. (2019, November 16). Sejarah Epidemi SARS: Bukti Wabah Virus yang Tak Pernah Berakhir. *Tirto.ID*. https://tirto.id/sejarah-epidemi-sars-bukti-wabah-virus-yang-tak-pernah-berakhir-elth