ANALISIS CLUSTERING VIRUS MERS-CoV MENGGUNAKAN METODE SPECTRAL CLUSTERING DAN ALGORITMA K-MEANS

Septian Wulandari [1], Dian Novita [2]
Program Teknik Studi Informatika, Universitas Indraprasta PGRI
septian.pmb09@rocketmail.com[1], dyan.novita@gmail.com[2]

Submitted November 16, 2020; Revised March 27, 2021; Accepted March 28, 2021

Abstrak

Virus MERS-Cov telah menyebar hingga ke negara-negara lain di luar Arab Saudi. Hal ini dikarenakan virus MERS-CoV dapat bermutasi dengan cepat sehingga dikhawatirkan dapat mengancam kesehatan masyarakat bahkan kesehatan dunia. Virus ini berkembang dan menjadi penyakit pernapasan akut serta angka kematian mencapai 30% di antara 536 kasus. Salah satu cara untuk mengelompokkan virus MERS-CoV adalah dengan mengelompokkan susunan DNA pada virus MERS-CoV yang memiliki kesamaan sifat dan fungsinya. *Spectral clustering* merupakan salah satu metode pengelompokkan yang dapat mengidentifikasikan ekspresi gen DNA. Metode ini juga mampu mempartisi data DNA dengan struktur yang lebih rumit dibandingkan dengan metode partisi *clustering*. Tujuan dari penelitian ini adalah menganalisis clustering virus MERS-CoV menggunakan metode *spectral clustering* dan algoritma *k-means.*Penelitian ini menggunakan pendekatan literature deskriptif kuantitatif. Hasil penelitian menunjukkan bahwa hasil *clustering* dengan metode *spectral clustering* dan algoritma *k-means* dihasilakan tiga *cluster* dan lebih homogen dibandingkan dengan *clustering* hanya menggunakan *k-means* saja.
Kata Kunci : *K-Means, Spectral Clustering*, Virus Mers-CoV

Abstract

The MERS-Cov virus has spread to other countries outside Saudi Arabia. This is because the MERS-CoV virus can mutate rapidly so it is feared that it could threaten public health and even world health. This virus develops and becomes an acute respiratory disease and the mortality rate reaches 30% among 536 cases. One way to classify the MERS-CoV virus is by grouping the DNA sequences of the MERS-CoV virus which have similar characteristics and functions. Spectral clustering is a grouping method that can identify DNA gene expression. This method is also able to partition DNA data with a more complex structure than the partition clustering method. The purpose of this study was to analyze the MERS-CoV virus clustering using the spectral clustering method and the k-means algorithm. This study used a quantitative descriptive literature approach. The results showed that the results of clustering using the spectral clustering method and the k-means algorithm produced three clusters and were more homogeneous than clustering using k-means only.
***Key Words** : K-Means, Spectral Clustering, Mers-CoV Virus*

1. PENDAHULUAN

Middle East Respiratory Syndrome (MERS) adalah suatu penyakit pada saluran pernapasan yang disebabkan karena virus corona jenis baru dan disingkat dengan nama MERS-CoV. MERS-CoV pertama kali ditemukan di negara Arab Saudi pada tahun 2012. Virus ini berkembang dan menjadi penyakit pernapasan akut serta angka kematian mencapai 30% di antara 536 kasus yang dilaporkan pada 12 Mei 2014 [1]. Pada sejak April 2012 hingga akhir November 2019, terdapat total 2494 kasus sindrom pernapasan Timur Tengah (MERS) yang dikonfirmasi pada laboratorium, termasuk 858 kematian terkait (dengan kasus tingkat kematian: 34,4%) [2].

Virus MERS-Cov telah menyebar hingga ke negara-negara lain di luar Arab Saudi. Hal ini dikarenakan virus MERS-CoV dapat bermutasi dengan cepat sehingga dikhawatirkan dapat mengancam kesehatan masyarakat bahkan kesehatan dunia. Virus MERS-CoV merupakan spesies beta coronavirus yang dapat menginfeksi manusia. Struktur genom pada virus MERS-CoV mengandung dipeptil-peptidase 4 (DPP4, atau CD26) yang diidentifikasikan sebagai reseptor *host-sel* untuk *entry sel* [1]. Karakteristik struktur genom pada virus MERS-CoV dapat diketahui dengan mengidentifikasikan susunan nukleotida pada DNA. Deoxyribonucleic acid (DNA) merupakan polinukleotida untai ganda yang memiliki karakteristik komponen penyusun antara lain gula deoksiribosa, gugus fosfat dan basa nitrogen (adenin, guanin, timin dan sitosin) [3]. Perbedaan penyusunan keempat nukleotida itulah yang menjadi perbedaan karakteristik pada semua makhluk hidup.

Sejak tahun 2012 orang yang terjangkit virus MERS-CoV tidak hanya ditemukan di negara Arab Saudi, namun virus ini telah mewabah ke berbagai negara di luar Arab Saudi seperti Jerman, Yunani, Prancis, Italia, Belanda, Inggris (UK), Mesir, Malaysia, Filipina, dan Tunisia [4]. Oleh karena itu, perlu dilakukan penelitian terhadap virus MERS-CoV mengelompokkan penyebaran virus MERS-CoV serta untuk meminimalisir penyebarannya ke negara lain dan mencegah penularannya kepada turis asing yang berkunjung ke negara-negara yang terpapar virus MERS-CoV.

Salah satu cara untuk mengelompokkan virus MERS-CoV adalah dengan mengelompokkan susunan DNA pada virus MERS-CoV yang memiliki kesamaan sifat dan fungsinya. *Spectral clustering* merupakan salah satu metode pengelompokkan yang dapat mengidentifikasikan ekspresi gen DNA. Jika matriks kesamaan pada ekspresi gen DNA yang dibangun dapat mendekati matriks ideal, metode *spectral clustering* akan memiliki kinerja pengelompokan yang lebih baik [5]. Disamping itu *spectral clustering* juga merupakan salah satu metode *clustering* yang pernah dikembangkan dalam memperbaiki akurasi regresi [6].

Penelitian terdahulu yang menggunakan analisis clustering pada DNA virus MERS-CoV dilakukan oleh Alhadi Bustamam, *et, all* pada tahun 2017 dengan menggunakan metode k-mer [4]. Pada penelitian tersebut digunakan 20 urutan DNA MERS-CoV dan memiliki hasil untuk beberapa penderita MERS-CoV yang berasal dari suatu negara belum tentu terinfeksi dari negara asal yang sama. Sedangkan, penelitian yang dilakukan oleh A. Chin, et, all pada tahun 2015 dengan melakukan spectral clustering pada ekspresi gen untuk mengidentifikasi jenis atau subjenis kanker menghasilkan sebelas data ekspresi gen pada spectral clustering mengungguli keenam metode clustering lainnya [7]. Pada metode *spectral clustering*, algoritma partisi yang umumnya digunakan adalah menggunakan algoritma *k-means*. Algoritma *K-Means clustering* merupakan algoritma yang berperan penting dalam bidang *data mining* serta sederhana untuk diimplementasikan dan dijalankan [8]. Sehingga metode *spectral clustering* diharapkan mampu mengelompokkan susunan DNA pada virus MERS-CoV.

Tujuan dari penelitian ini adalah menganalisis clustering virus MERS-CoV menggunakan metode *spectral clustering* dan algoritma *k-means*. Pada penelitian ini diharapkan mampu memberikan informasi penyebaran virus MERS-CoV melalui kesamaan karakteristik DNA virus MERS-CoV di berbagai negara yang terjangkit.

2. METODE PENELITIAN

Pada penelitian ini digunakan pendekatan studi literatur deskriptif kuantitatif. Pendekatan studi literatur dengan mengumpulkan referensi uantuk mendukung penyelesaian penelitian ini. Kemudian, pendekatan deskriptif kuantitatif dilakukan dengan mengolah, menganalisa, dan menginterpretasikan data sesuai dengan kebutuhan peneliti.

Data yang digunakan pada penelitian ini adalah 100 data barisan DNA virus MERS-CoV yang didapatkan dari *National Center for Biotechnology Information* (NCBI) pada *website* www.ncbi.nlm.nih.gov. Data barisan DNA virus MERS-CoV yang diperoleh berupa format FASTA. Barisan DNA yang digunakan berjenis *complete genome*, karena berjenis *complete genome* diharapkan data tersebut mampu memberikan informasi genetik virus MERS-CoV yang lebih banyak. Data yang digunakan pada penelitian ini masih berbentuk string atau barisan DNA Virus MERS-CoV. Dikarenakan data yang digunakan masih dalam bentuk string, maka data tersebut perlu diubah menjadi bentuk numerik. Tahap pertama yang dilakukan pada penelitian ini adalah mengubah data dalam bentuk *string* menjadi numerik dilakukan dengan ekstraksi ciri menggunakan *n-mers frequency*. *N-mers frequency* atau *k-mers frequency* merupakan perhitungan frekuensi dari panjang *string* yang diberikan pada proses membaca kumpulan barisan [9]. *The number of features for any sequence DNA is thus the total number of possible k-mers* [8]. Hal ini menunjukkan bahwa *k-mers* merupakan jumlah ekstraksi ciri yang mungkin untuk setiap urutan barisan DNA.

Nilai yang diperoleh dari ekstraksi ciri sangat bervariasi mulai dari data yang terlalu besar sampai dengan data yang terlalu kecil, sehingga perlu adanya normalisasi atau standarisasi untuk menyeragamkan interval pada data. Tahap kedua yang dilakukan pada penelitian ini adalah normalisasi data dengan menggunakan *min-max normalization*. Berikut merupakan rumus *min-max normalization* [10]:

$$v' = \frac{v-min}{max-min}(new_{max} - new_{min}) + new_{min} \quad (1)$$

Keterangan: v : nilai elemen hasil ekstraksi *n-mers frequency* sebelum dinormalisasi
v': nilai elemen hasil ekstraksi *n-mers frequency* setelah dinormalisasi
min : nilai minimum matriks
max: nilai maksimum matriks
new_{max}: nilai range maksimum terbaru
new_{min}: nilai range minimum terbaru

Setelah dilakukan normalisasi, maka langkah ketiga yang dilakukan adalah proses *clustering* dengan menggunakan metode *spectral clustering* dengan menggunakan algoritma *k-means*. *Spectral clustering* merupakan salah satu metode pengelompokkan yang mudah digunakan dan metode pengelompokkan yang cukup cepat, terutama untuk data yang renggang hingga data beberapa ribu.

Dalam metode *spectral clustering* memperlakukan pengelompokan data sebagai grafik dan mempartisi data tanpa membuat asumsi dalam bentuk *cluster* data. *Spectral Clustering* akan dibentuk sebuah graf dari data yang ada. Dimana verteks dari graf tersebut merupakan setiap record pada data. *Edge*-nya berupa hubungan antar data yang biasanya bernilai jarak dari dua *record* yang berhubungan [6].

Langkah-langkah dalam *spectral clustering* yaitu [6]:

1. Kontruksi graf similaritas dari dataset training. *Verteks* pada graf tersebut merupakan representasi dari setiap

record pada data training. Bobot dari tiap *edge* merupakan jarak antara satu *verteks* dengan *verteks* lainnya. Perhitungan jarak antar *verteks* menggunakan persamaan jarak *exponential* yang tertulis pada persamaan 2.

$$w_{ij} = exp\frac{-\|s_i - s_j\|^2}{2\sigma^2} \quad (2)$$

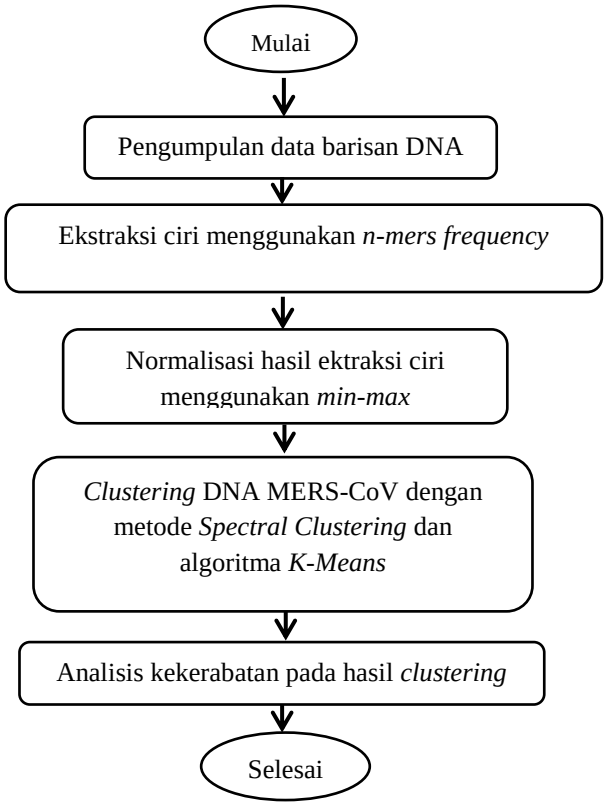
Setelah itu, bobot dari setiap edge yang ada dibentuk menjadi matriks *weight*. Dengan begitu matriks *weight* merupakan representasi graf similaritas dari dataset.

2. Dari matriks *weight* dihitung derajat dari setiap *verteks* dengan menjumlahkan bobot dari edge yang terhubung pada *verteks* yang bersangkutan. Dari derajat *verteks* tersebut dapat dibentuk matriks *degree* yang merupakan matriks diagonal yang berisi bobot setiap *verteks*.
3. Dibentuk normalisasi matriks *Laplacian* dengan menggunakan matriks *weight* (W) dan matriks *degree* (D) yang telah dihitung sebelumnya. Perhitungan matriks Laplacian (L) dengan rumus pada persamaan (3).

$$L = D - W \quad (3)$$

4. Dihitung k *eigenvector* pertama dari matriks *Laplacian*, dimana k merupakan parameter jumlah *cluster*. Maka terbentuklah matriks *k-eigen* yang merupakan k *eigenvector* pertama dari matriks *Laplacian*. Matriks *k-eigen* berukuran $n\,x\,k$, dengan variabel n merupakan jumlah *record* pada data masukan.
5. Normalisasi data dengan matriks *k-eigen* sehingga akan terbentuk k kolom yang merepresentasikan setiap nilai normalisasi *eigen* pada setiap kolomnya.
6. Hasil dari data normalisasi kemudian di *cluster* dengan menggunakan algoritma partisi *k-means*. Data normalisasi mewakili masukan data latih. Data latih ke-i akan dimasukkan pada suatu *cluster* jika dan hanya jika data hasil normalisasi ke-i masuk pada *cluster* yang sama.

Berikut merupakan diagram alur pada penelitian ini:



Gambar 1. Diagram Alur Penelitian

3. HASIL DAN PEMBAHASAN

Pengumpulan data pada penelitian ini diambil pada laman NCBI yaitu 100 barisan DNA Virus MERS-CoV dengan format FASTA atau string. Kemudian analisis *clustering* virus MERS-CoV menggunakan metode *spectral clustering* dan algoritma *k-means* dilakukan dengan langkah-langkah sebagai berikut:

1. Langkah pertama yang dilakukan adalah mengubah data dalam bentuk *string* menjadi numerik dilakukan dengan

ekstraksi ciri menggunakan *n-mers frequency.*

	atg	att	caa	cac	cag	cat	cca
2	721	744	567	426	421	498	383
3	721	744	567	426	421	498	383
4	700	730	560	427	420	472	377
5	700	730	560	427	420	472	377
6	701	728	560	427	419	472	377
7	703	729	560	426	420	473	377
8	701	726	561	427	420	471	378
9	700	727	561	426	419	472	377
10	702	727	561	426	420	472	378
11	700	729	560	427	419	471	377
12	697	729	554	424	414	477	371
13	706	737	562	428	416	479	374
14	704	735	565	421	421	483	375
15	699	723	558	417	417	478	371
16	702	726	560	418	418	473	371
17	722	751	568	421	420	499	383

Gambar 2. Potongan Ekstraksi Ciri Barisan DNA MERS-CoV

Gambar 2 menunjukkan potongan ekstraksi ciri 100 barisan DNA MERS-CoV yang diperoleh dengan pola kemunculan empat basa yaitu kombinasi A, C, G, T dan protein memiliki tiga basa sehingga pola kemunculan menjadi adalah $4^3 = 64$ kombinasi kemunculan protein atau asam amino yang menjadi variable pada penelitian ini.

2. Langkah kedua adalah melakukan normalisasi dari hasil ekstraksi ciri untuk membuat rentang data menjadi 0 sampai dengan 1 digunakan rumus min max normalization pada persamaan (1) dan diperoleh hasil pada Gambar 3. Pada Gambar 3 terlihat bahwa hasul ekstraksi ciri yang memiliki rentang nilai yang berjauhan menjadi normal setelah dilakukan normalisasi sehingga rentang data barisan DNA MERS-CoV berada pada 0 sampai dengan 1.

acc	acg	act	aga	agc	agg
0,3927765	0,2234763	0,7155756	0,4909707	0,3848758	0,3961625
0,6388262	0,3927765	0,2234763	0,7155756	0,4909707	0,3848758
0,6252822	0,3871332	0,2257336	0,7054176	0,4808126	0,3837472
0,6252822	0,3871332	0,2257336	0,7054176	0,4808126	0,3837472
0,6241535	0,3882619	0,2257336	0,7054176	0,479684	0,3837472
0,6241535	0,3882619	0,2257336	0,7031603	0,4819413	0,3837472
0,6241535	0,3882619	0,2257336	0,7054176	0,4808126	0,3837472
0,6241535	0,3882619	0,2257336	0,7054176	0,4785553	0,3837472
0,6241535	0,3882619	0,2257336	0,7054176	0,4808126	0,3837472
0,6241535	0,3871332	0,2257336	0,7076749	0,4808126	0,3837472
0,6320542	0,3860045	0,2223476	0,7009029	0,479684	0,3803612
0,6286682	0,3893905	0,2268623	0,7121896	0,4887133	0,3848758
0,6264108	0,3882619	0,227991	0,7121896	0,4887133	0,3860045
0,6207675	0,3837472	0,2268623	0,7088036	0,4774266	0,3814898
0,6218962	0,3848758	0,2268623	0,7076749	0,4785553	0,3826185
0,6343115	0,3927765	0,2291196	0,7121896	0,493228	0,3848758
0,6343115	0,3905192	0,2291196	0,7121896	0,4920993	0,3848758

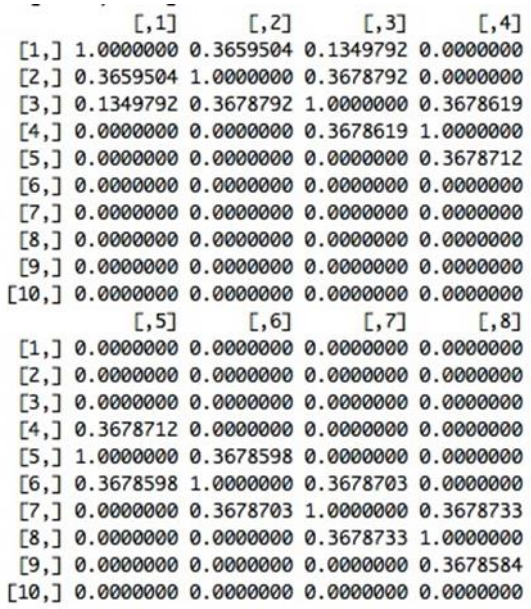
Gambar 3. Potongan Normalisasi Barisan DNA MERS-CoV

3. Langkah ketiga adalah proses clustering dengan menggunakan metode *spectral clustering* dan algoritma *k-means.* Proses *clustering* dilakukan dengan mengkonstruksi graf similaritas dari dataset training. jika dua titik berdekatan maka $s_{ij} = 1$dan ketika dua titik berjauhan maka $s_{ij} = 0$. Sehingga ketika dua titik berasal dari cluster yang berbeda maka dua barisan DNA MERS-CoV memiliki jarak yang jauh. Namun, mungkin juga terdapat dua titik dari cluster yang sama juga memilki jarak yang jauh, kecuali terdapat urutan titik dari cluster yang sama yang membuat jalur diantara keduanya. Sehingga matriks S diperoleh hasil pada Gambar 4.

```
              [,1]          [,2]          [,3]
 [1,] 1.000000e+00 3.659504e-01 1.349792e-01
 [2,] 3.659504e-01 1.000000e+00 3.678792e-01
 [3,] 1.349792e-01 3.678792e-01 1.000000e+00
 [4,] 4.969783e-02 1.353319e-01 3.678619e-01
 [5,] 1.829207e-02 4.978619e-02 1.353319e-01
 [6,] 6.730480e-03 1.831527e-02 4.978580e-02
 [7,] 2.476474e-03 6.737815e-03 1.831522e-02
 [8,] 9.111811e-04 2.478709e-03 6.737816e-03
 [9,] 3.352317e-04 9.118659e-04 2.478705e-03
[10,] 1.233165e-04 3.354360e-04 9.118006e-04
              [,4]          [,5]          [,6]
 [1,] 4.969783e-02 1.829207e-02 6.730480e-03
 [2,] 1.353319e-01 4.978619e-02 1.831527e-02
 [3,] 3.678619e-01 1.353319e-01 4.978580e-02
 [4,] 1.000000e+00 3.678712e-01 1.353329e-01
 [5,] 3.678712e-01 1.000000e+00 3.678598e-01
 [6,] 1.353329e-01 3.678598e-01 1.000000e+00
 [7,] 4.978635e-02 1.353309e-01 3.678703e-01
 [8,] 1.831538e-02 4.978613e-02 1.353337e-01
 [9,] 6.737837e-03 1.831516e-02 4.978602e-02
[10,] 2.478512e-03 6.737119e-03 1.831260e-02
```

Gambar 4. Potongan Matriks S

Langkah selanjutnya adalah menghitung matriks afinitas atau matriks simetris W berdasarkan matriks W. matriks W terdiri dari nilai positif dan simetris. Menghitung matriks W dengan menerapkan k-neighboor terdekat untuk membuat representasi grafik yang menghubungkan titik data terdekat menggunakan persamaan 2. Namun, menjadi simetris jika w_{ij} dipilih sebagai tetangga terdekat begitu juga w_{ji} sehingga diperoleh hasil pada Gambar 5.

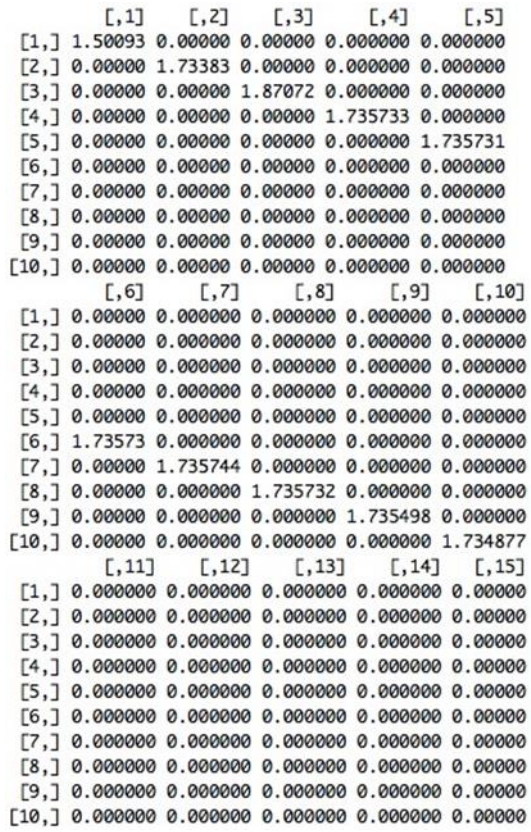


	[,1]	[,2]	[,3]	[,4]
[1,]	1.0000000	0.3659504	0.1349792	0.0000000
[2,]	0.3659504	1.0000000	0.3678792	0.0000000
[3,]	0.1349792	0.3678792	1.0000000	0.3678619
[4,]	0.0000000	0.0000000	0.3678619	1.0000000
[5,]	0.0000000	0.0000000	0.0000000	0.3678712
[6,]	0.0000000	0.0000000	0.0000000	0.0000000
[7,]	0.0000000	0.0000000	0.0000000	0.0000000
[8,]	0.0000000	0.0000000	0.0000000	0.0000000
[9,]	0.0000000	0.0000000	0.0000000	0.0000000
[10,]	0.0000000	0.0000000	0.0000000	0.0000000

	[,5]	[,6]	[,7]	[,8]
[1,]	0.0000000	0.0000000	0.0000000	0.0000000
[2,]	0.0000000	0.0000000	0.0000000	0.0000000
[3,]	0.0000000	0.0000000	0.0000000	0.0000000
[4,]	0.3678712	0.0000000	0.0000000	0.0000000
[5,]	1.0000000	0.3678598	0.0000000	0.0000000
[6,]	0.3678598	1.0000000	0.3678703	0.0000000
[7,]	0.0000000	0.3678703	1.0000000	0.3678733
[8,]	0.0000000	0.0000000	0.3678733	1.0000000
[9,]	0.0000000	0.0000000	0.0000000	0.3678584
[10,]	0.0000000	0.0000000	0.0000000	0.0000000

Gambar 5. Hasil Perhitungan Matriks W

Dengan matriks similaritas *clustering* digantikan dengan partisi-grafik, komponen-komponen grafik yang terhubung diinterpretasikan sebagai *cluster*. Grafik harus dipartisi sehingga tepi yang menghubungkan *cluster* yang berbeda harus memiliki bobot yang rendah, dan tepi dalam *cluster* yang sama harus memiliki nilai yang tinggi dan dhitung menggunakan persamaan 4 maka diperoleh matriks *D* seperti pada Gambar 6.



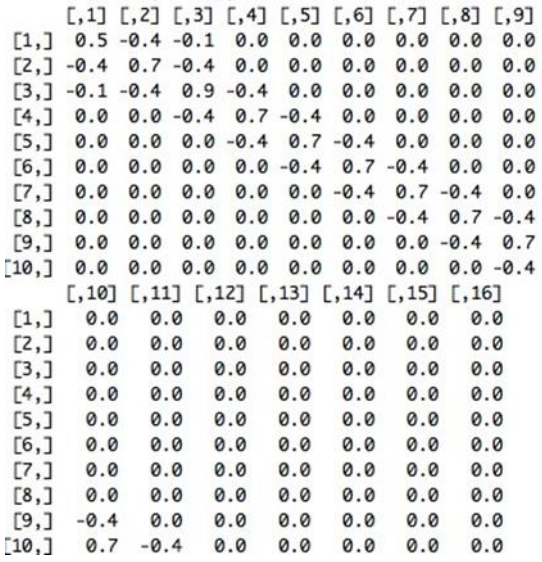
	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.50093	0.00000	0.00000	0.000000	0.000000
[2,]	0.00000	1.73383	0.00000	0.000000	0.000000
[3,]	0.00000	0.00000	1.87072	0.000000	0.000000
[4,]	0.00000	0.00000	0.00000	1.735733	0.000000
[5,]	0.00000	0.00000	0.00000	0.000000	1.735731
[6,]	0.00000	0.00000	0.00000	0.000000	0.000000
[7,]	0.00000	0.00000	0.00000	0.000000	0.000000
[8,]	0.00000	0.00000	0.00000	0.000000	0.000000
[9,]	0.00000	0.00000	0.00000	0.000000	0.000000
[10,]	0.00000	0.00000	0.00000	0.000000	0.000000

	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	0.00000	0.000000	0.000000	0.000000	0.000000
[2,]	0.00000	0.000000	0.000000	0.000000	0.000000
[3,]	0.00000	0.000000	0.000000	0.000000	0.000000
[4,]	0.00000	0.000000	0.000000	0.000000	0.000000
[5,]	0.00000	0.000000	0.000000	0.000000	0.000000
[6,]	1.73573	0.000000	0.000000	0.000000	0.000000
[7,]	0.00000	1.735744	0.000000	0.000000	0.000000
[8,]	0.00000	0.000000	1.735732	0.000000	0.000000
[9,]	0.00000	0.000000	0.000000	1.735498	0.000000
[10,]	0.00000	0.000000	0.000000	0.000000	1.734877

	[,11]	[,12]	[,13]	[,14]	[,15]
[1,]	0.000000	0.000000	0.000000	0.000000	0.00000
[2,]	0.000000	0.000000	0.000000	0.000000	0.00000
[3,]	0.000000	0.000000	0.000000	0.000000	0.00000
[4,]	0.000000	0.000000	0.000000	0.000000	0.00000
[5,]	0.000000	0.000000	0.000000	0.000000	0.00000
[6,]	0.000000	0.000000	0.000000	0.000000	0.00000
[7,]	0.000000	0.000000	0.000000	0.000000	0.00000
[8,]	0.000000	0.000000	0.000000	0.000000	0.00000
[9,]	0.000000	0.000000	0.000000	0.000000	0.00000
[10,]	0.000000	0.000000	0.000000	0.000000	0.00000

Gambar 6. Potongan Matriks Degree D

Kemudian menghitung matriks *Laplacian* pada persamaan 3. *Spectral clustering* menghasilkan kelompok *vertex* sehingga barisan DNA MERS-CoV jarang berpindah dari satu *cluster* ke *cluster* lainnya.



	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	0.5	-0.4	-0.1	0.0	0.0	0.0	0.0	0.0	0.0
[2,]	-0.4	0.7	-0.4	0.0	0.0	0.0	0.0	0.0	0.0
[3,]	-0.1	-0.4	0.9	-0.4	0.0	0.0	0.0	0.0	0.0
[4,]	0.0	0.0	-0.4	0.7	-0.4	0.0	0.0	0.0	0.0
[5,]	0.0	0.0	0.0	-0.4	0.7	-0.4	0.0	0.0	0.0
[6,]	0.0	0.0	0.0	0.0	-0.4	0.7	-0.4	0.0	0.0
[7,]	0.0	0.0	0.0	0.0	0.0	-0.4	0.7	-0.4	0.0
[8,]	0.0	0.0	0.0	0.0	0.0	0.0	-0.4	0.7	-0.4
[9,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.4	0.7
[10,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.4

	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]
[1,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[2,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[3,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[4,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[5,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[6,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[7,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[8,]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[9,]	-0.4	0.0	0.0	0.0	0.0	0.0	0.0
[10,]	0.7	-0.4	0.0	0.0	0.0	0.0	0.0

Gambar 7. Potongan Matriks Laplacian

Selanjutnya, dihasilkan 3 *cluster*, sehingga langkah selanjutnya adalah menemukan *k-eigenvector* terkecil dan diperoleh nilai *k-eigenvector* terkecil yaitu -0,1005038. Selanjutnya, normalisasi data dengan matriks *k- eigen* sehingga akan terbentuk $k = 3$ kolom yang merepresentasikan setiap nilai normalisasi *eigen* pada setiap kolomnya. Hasil dari data normalisasi kemudian di *cluster* dengan menggunakan algoritma partisi *k-means*. Data normalisasi mewakili masukan data latih. Data latih ke-*i* akan dimasukkan pada suatu *cluster* jika dan hanya jika data hasil normalisasi ke-*i* masuk pada *cluster* yang sama. Sehingga hasil clustering dapat dilihat pada Tabel 1.

Tabel 1. Hasil Clustering 100 barisan DNA MERS-CoV

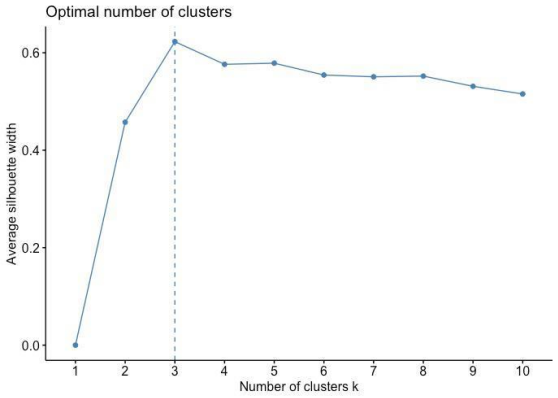
Cluster	Anggota	Jumlah
1	M28,M29,M30,M31,M32,M33,M34,M35,M36,M37,M38,M39,M40,M41,M42,M43,M44,M45,M46,M47,M48,M49,M50,M51,M52,M53,M54,M55,M56,M57,M58,M59,M60,M61,M62,M63,M64,M65,M66,M67,M68,M69,M70	43
2	M1,M2,M3,M4,M5,M6,M7,M8,M9,M10,M11,M12,M13,M14,M15,M16,M17,M18,M19,M20,M21,M22,M23,M24,M25,M26,M27,M100	28
3	M71,M72,M73,M74,M75,M76,M77,M78,M79,M80,M81,M82,M83,M84,M85,M86,M87,M88,M89,M90,M91,M92,M93,M94,M95,M96,M97,M98,M99	29

Hasil *clustering* Tabel 1 menunjukkan bahwa *cluster* 1 memiliki anggota 43 barisan DNA MERS-CoV, *cluster* 2 memiliki 28 anggota barisan DNA MERS-CoV, dan *cluster* 3 memiliki 29 anggota barisan DNA MERS-CoV.

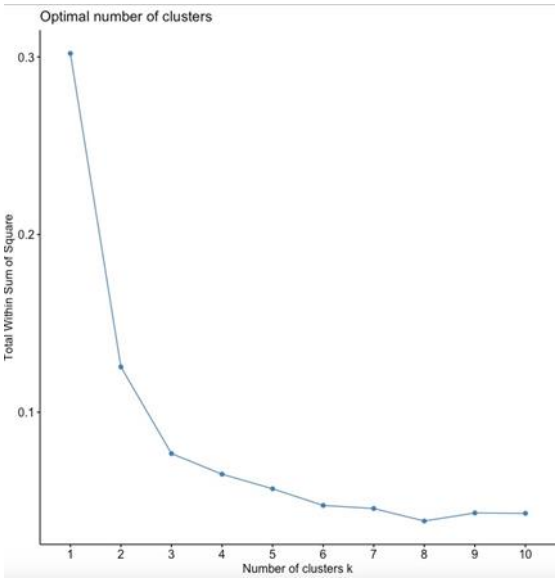
Untuk menganalisa lebih dalam lagi, dilakukan perbandingan dengan hasil *clustering* barisan DNA MERS-CoV dengan *k-means* tanpa mengunakan metode *spectral clustering*. Langkah yang dilakukan adalah mengubah data string menjadi numerik dengan *n-mers frequency*, normalisasi data menggunakan *min-max normalizati*on, kemudian barisan DNA MERS-CoV dilakukan *clustering* menggunakan k-means. Jumlah *clustering* optimal adalah 3 *cluster*. Perbandingan hasil *clustering* barisan DNA MERS-CoV dapat dilihat pada Tabel 2.

Tabel 2. Perbandingan Clustering dengan K-Means saja dan Spectral Clustering dan K-Means

Metode	Cluster	Jumlah Anggota Cluster	withinss clusters	Total sum of square
K-Means	**1**	**39**	**0,02775863**	
	2	**16**	**0,002066619**	**0,3019858**
	3	**45**	**0,02824574**	
SC-K-Means	**1**	**43**	**0,16539588**	
	2	**28**	**0,08615012**	**1,999362**
	3	**29**	**0,09468987**	



Gambar 8. Hasil rata-rata Sillhouette Metode Spectral clustering dan Algoritma K-Means



Gambar 9. Hasil rata-rata Sillhouette Metode K-Means

Tabel 2 menunjukkan perbedaan nilai *withnss clusters* pada *k-means* lebih rendah dibandingkan dengan *spectral clustering* dan *k-means* artinya pengukuran jarak rata-rata kuadrat dari semua titik pada pusat *cluster* pada *spectral clustering* dengan *k-means* masih lebih baik dibandingkan dengan *clustering* dengan *k-means* saja. *Total sum of square* metode *k-means* lebih rendah dibandingkan dengan *spectral clustering* dan *k-means* artimya penyebaran variable yang diamati pada sekitar means pada *spectral clustering* dan *k-means* masih lebih baik dibandingkan dengan *clustering* dengan *k-means* saja. Dengan kata lain, kemiripan barisan DNA dalam *clustering* dengan metode *spectral clustering* dan *k-means* masih lebih baik dibandingkan *clustering* hanya dengan *k-means* saja. Sehingga, *clustering* menggunakan metode *spectral clustering* dan algoritma *k-means* memberikan hasil *clustering* yang lebih homogen dibandingkan dengan *clustering k-means* tanpa menggunakan *spectral clustering.*

4. SIMPULAN

Pada penelitian ini, barisan DNA MERS-CoV dapat di *clustering* menggunakan metode *spectral clustering* dan algoritma *k-means*. Langkah-langkah yang dilakukan pada penelitian ini adalah megubah data menjadi numerik, kemudian dilakukan normalisasi data, *clustering* barisan DNA menggunakan metode *spectral clustering* dengan tahapan mengkonstruksi graf similartitas, kemudian menghitung matriks normalisasi *Laplacian*, hitung *k-eigen value*, serta hasil dari data normalisasi kemudian di *cluster* dengan menggunakan algoritma partisi *k-means*. Hasil penelitian menunjukkan bahwa hasil *clustering* dengan metode *spectral clustering* dan algoritma *k-means* dihasilakan tiga *cluster* dan lebih homogen dibandingkan dengan *clustering* hanya menggunakan *k-means* saja.

DAFTAR PUSTAKA

[1] N. H. Rampengan, “Middle East Respiratory Syndrome,” *J. Biomedik*, vol. 8, no. 1, pp. 17–26, 2016.

[2] World Health Organization, “Mers Situation Update November 2019,” 2019.

[3] S. Nur’aini, A. S. Mukaromah, and S. Muhlisoh, “Pengenalan Deoxyribonucleic Acid (DNA) Dengan Marker-Based Augmented Reality,” *Walisongo J. Inf. Technol.*, vol. 1, no. 2, p. 91, 2019, doi: 10.21580/wjit.2019.1.2.4531.

[4] A. Bustamam, E. D. Ulul, H. F. A. Hura, and T. Siswantining, “Implementation of hierarchical clustering using k-mer sparse matrix to analyze MERS-CoV genetic relationship,” *AIP Conf. Proc.*, vol. 1862, no. July, 2017, doi: 10.1063/1.4991246.

[5] S. Ren, S. Zhang, and T. Wu, "An Improved Spectral Clustering Community Detection Algorithm Based on Probability Matrix," *Discret. Dyn. Nat. Soc.*, vol. 2020, 2020, doi: 10.1155/2020/4540302.

[6] A. Yusuf and H. Tjandrasa, "Prediksi Nilai Dengan Metode Spectral Clustering Dan Clusterwise Regression," *J. Simatec*, vol. VIII, no. 1, pp. 39–45, 2013.

[7] A. J. Chin, A. Mirzal, and H. Haron, "Spectral clustering on gene expression profile to identify cancer types or subtypes," *J. Teknol.*, vol. 76, no. 1, pp. 289–297, 2015, doi: 10.11113/jt.v76.4036.

[8] M. Nafis Ul Alam and U. F. Chowdhury, "Short k-mer abundance profiles yield robust machine learning features and accurate classifiers for RNA viruses," *PLoS One*, vol. 15, no. 9 September, pp. 1–23, 2020, doi: 10.1371/journal.pone.0239381.

[9] S. Deorowicz, M. Kokot, S. Grabowski, and A. Debudaj-Grabysz, "KMC 2: Fast and resource-frugal k-mer counting," *Bioinformatics*, vol. 31, no. 10, pp. 1569–1576, 2015, doi: 10.1093/bioinformatics/btv022.

[10] D. F. Pramesti, Lahan, M. Tanzil Furqon, and C. Dewi, "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 723–732, 2017, doi: 10.1109/EUMC.2008.4751704.