JAMBURA JOURNAL OF MATHEMATICS
Jambura J. Math. Vol. 2, No. 2, pp. 73-86, July 2020
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jjom
DOI: https://doi.org/10.34312/jjom.v2i2.4320

Penggunaan *N-mers Frequency* pada Analisis Barisan DNA

Khoirul Umam[1*], Rahmat Sagara[2]

[1,2] Program Studi Matematika Bisnis, Fakultas Bisnis, Institut Teknologi dan Bisnis Kalbis, Jl. Pulomas Selatan Kav. 22, Jakarta Timur 13210, Jakarta, Indonesia

** Penulis Korespondensi. Email: khoirul.umam@kalbis.ac.id*

ABSTRAK

Salah satu metode untuk menganalisis barisan DNA adalah menggunaan *N-mers Frequency*. *N-mers Frequency* termasuk metode *data mining* pada barisan DNA, dimana barisan DNA yang merupakan data *string "ACGT"* akan diubah menjadi data numerik. *N-mers Frequency* pada tulisan ini menggunakan $N = 3$. Hal ini disebabkan karena pada proses sintesis protein, tRNA akan membawa tiga basa *nekleotida* (anti kodon) yang akan dipasangkan dengan tiga basa *nekleotida* (kodon) pada pita mRNA. Dalam hal ini mRNA dibentuk dari duplikasi barisan DNA. Studi ini dilakukan untuk mengetahui akurasi dari penggunaan *N-mers Frequency*. Untuk menghitung Akurasi penggunaan *N-mers Frequency*, dilakukan tahapan seperti berikut: (1) pengumpulan data barisan DNA, (2) *N-mers Frequency*, (3) matriks jarak, (4) pengelompokan menggunakan algoritma *K-means++*, PAM, AGNES, dan DIANA, (5) menghitung akurasi, dan (6) kesimpulan. Akurasi dari Penggunaan *N-mers Frequency* pada penelitian ini adalah 100%, dengan menggunakan data 100 barisan DNA yang telah diketahui jenisnya, yaitu: virus HPV, virus Ebola, virus Marburg, dan virus Zika.

Kata Kunci:
N-mers Frequency; *K-means++*; *Data Mining*; Barisan DNA

ABSTRACT

N-mers Frequency is a method for analyzing DNA sequences. It is a data mining method in the DNA sequence. The DNA sequence which is the string data, "ACGT". It will be converted into numerical data. The N of N-mers Frequency is determined that N = 3. It is because the tRNA protein synthesis process will carry three nucleotide bases (anti-codons) which will be paired with three nucleotide bases (codons) in the mRNA band which is formed by duplicating rows of DNA codes. This study, conducted to determine the accuracy of the use of N-mers Frequency. For calculate of the accuracy, steps are performed as follows: (1) collection of DNA sequences data, (2) N-mers Frequency, (3) distance matrix, (4) grouping using K-means++, PAM, AGNES and DIANA algorithms, (5) calculate the accuracy, and (6) conclusions. The accuracy of the use of N-mers Frequency in this study is 100%, with using data from 100 DNA sequences which known their types, i.e., HPV virus, Ebola virus, Marburg virus, and Zika virus.

Keywords:
N-mers Frequency; k-Mean++; Data Mining; DNA Sequences

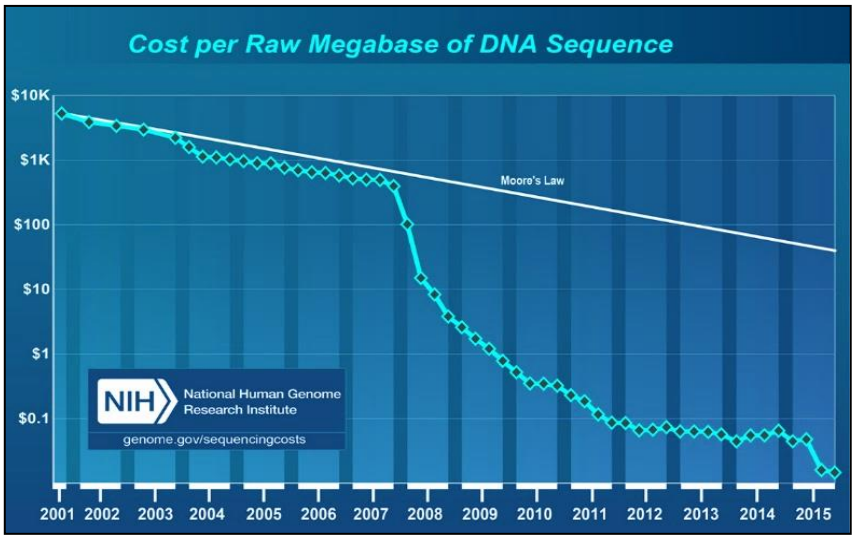
Diterima: 30 Desember 2019 | Disetujui: 4 Maret 2020 | Online: 08 Maret 2020

Format Sitasi:

K. Umam and R. Sagara, "Penggunaan *N-mers Frequency* pada Analisis Barisan DNA," *Jambura J. Math.*, vol. 2, no. 2, pp.73-86, 2020.

1. Pendahuluan

Seiring berkembangnya teknologi dan semakin murahnya biaya *sequencing* DNA, banyaknya data barisan DNA menjadi tak terbendung. Data dari National Human Genome Research Institute pada Gambar 1 menunjukkan bahwa biaya untuk per Megabase barisan DNA pada tahun 2015 sekitar $0.01. Biaya ini merupakan hal sangat murah dibandingkan biaya pada tahun sekitar tahun 2001 yang biayanya hampir mencapai $10.000 [1].



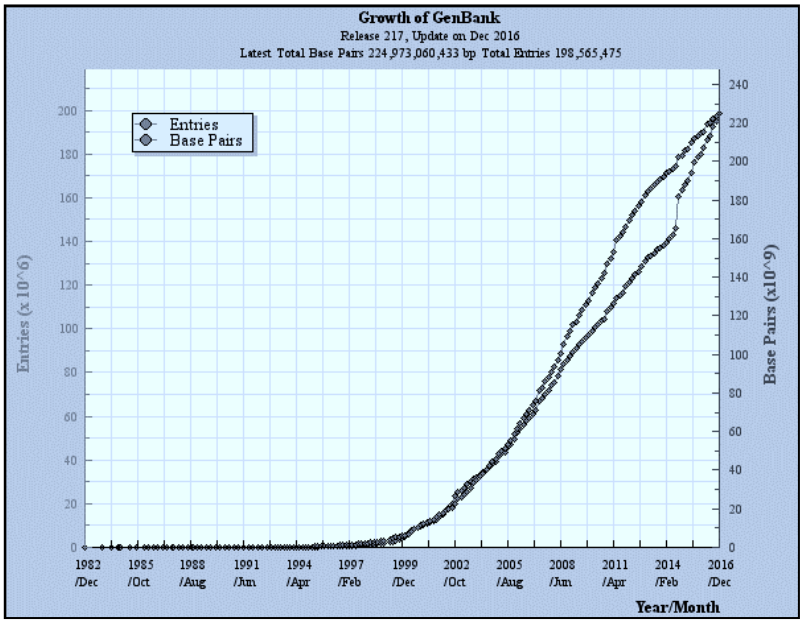
Gambar 1. Biaya per megabase barisan DNA

Pada laboratorium modern dengan mesin *Polymerase Chain Reaction* (PCR) yang mahal dapat menganalisis genom manusia sebanyak 3 miliar pasangan basa DNA dalam beberapa jam. Di tambah lagi penemuan mini PCR oleh Kraves pada tahun 2014, yang bisa dikatakan sebagai alat tes DNA personal yang bisa jalan dengan baterai solar. Penemuan ini membuat biaya *sequencing DNA* menjadi semakin murah. Semakin murahnya biaya *sequencing* DNA menjadikan laju banyaknya data barisan DNA sangat tinggi.

Pada Gambar 2 ditunjukkan bahwa data dari *National Center for Biotechnology Information* (NCBI) pada bulan desember 2016, terdapat sekitar 224 milyar *base pairs* DNA dari sekitar 200 juta entri barisan DNA. Jika ditambahkan dengan DNA data base sejenis seperti *European Molecular Biology Laboratory* (EMBL), *DNA Data Bank of Japan* (DDBJ) dan *National Health & Medicine Big Data (Nanjing) Center,* banyaknya *base pairs* DNA akan jauh melebihi 224 milyar [2].

Data tersebut menunjukkan bahwa studi tentang *bioinformatics* berkembang sangat pesat. Pada 2017, *Nanjing Center* sedang membangun *gigantic DNA database.* Melalui kerja sama dengan *The State-Owned Yangzi Group*, Southeast University and Nanjing

Medical University, *Nanjing Center* menargetkan akan merekam informasi kesehatan dan medis dari 80 juta orang melalui *sequencing* DNA. Sampel barisan DNA tersebut akan diambil bertahap sekitar 400.000 sampai 500.000 pertahun. *Database* tersebut dibangun untuk mempelajari mutasi genetik yang terkait dengan penyakit mematikan, mencari dampak interaksi antara gen dan masalah lingkungan pada kesehatan manusia dan memberikan dukungan statistik untuk diagnosis dan pengobatan penyakit mematikan [3].

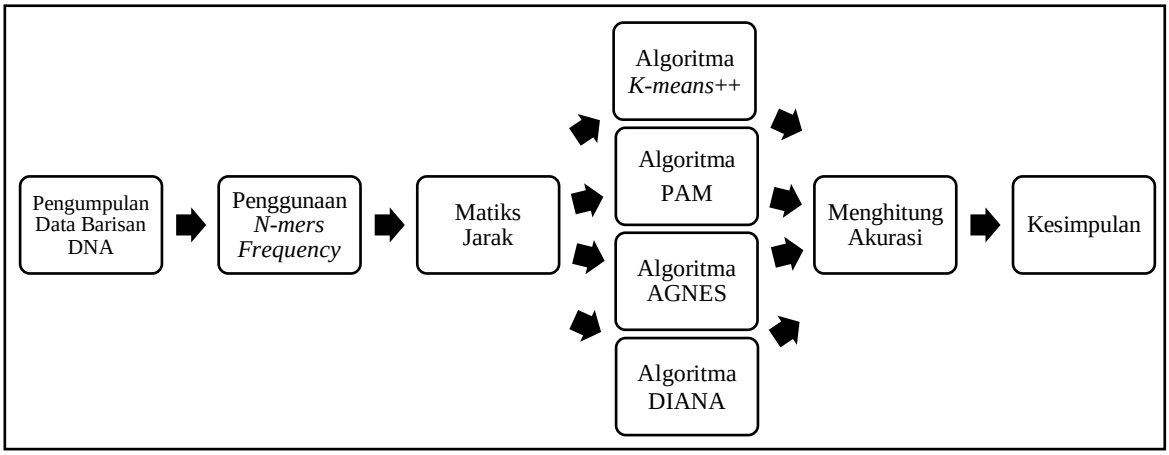


Gambar 2. Pertumbuhan data barisan DNA dari GenBank.

Untuk menganalisis barisan DNA, terutama yang berhubungan dengan kesehatan manusia diperlukan keakuratan yang tinggi. Salah satu metode *data mining* barisan DNA adalah metode *N-mers Frequency* [4][5]. Tulisan ini bertujuan untuk melihat akurasi penggunaan *N-mers Frequency* dalam analisis barisan DNA. Untuk melihat akurasi penggunaan *N-mers Frequency,* data barisan DNA yang sudah diketahui kelompoknya akan diacak, kemudian akan dikelompokkan kembali menggunakan algoritma *K-mean*++, PAM *(Partitioning Around Medoid)*, AGNES *(Agglomerative Nesting)* dan DIANA *(Divisive Analysis)*.

2. Metode

Barisan DNA merupakan barisan data *string*. Salah satu cara untuk menganalisis barisan DNA adalah mengubah data *string* menjadi data numerik. Pada tulisan ini diusulkan penggunaan *N-mers Frequency* dalam analisis barisan DNA. Tulisan ini fokus pada perhitungan akurasi penggunaan *N-mers Frequency*. Untuk menghitung akurasi penggunaan *N-mers Frequency* dilakukan tahapan penelitian berikut: (1) pengumpulan data barisan DNA, (2) *N-mers Frequency*, (3) matriks jarak, (4) pengelompokan menggunakan algoritma *K-means*++, PAM, AGNES, dan DIANA, (5) menghitung akurasi, dan (6) kesimpulan. Tahapan tersebut dapat dilihat pada Gambar 3.



Gambar 3. Tahapan penelitian

Tahapan-tahapan penelitian yang tercantum pada Gambar 3, dipaparkan secara gambling pada sub-sub bab berikut.

2.1. Pengumpulan Data Barisan DNA

Data yang digunakan sebanyak 100 barisan DNA yang berbeda yang terdiri dari 25 barisan DNA virus *Human Papillomavirus* (HPV) [6], 25 barisan DNA virus Ebola [7], 25 barisan DNA virus Marburg [8], dan 25 barisan DNA virus Zika [9]. Pemilihan data virus tersebut didasarkan pada sifat virus yang berbahaya bagi manusia, bahkan ada yang menyebabkan kematian.

Data barisan DNA 4 kelompok virus tersebut diperoleh dari *National Center for Biotechnology Information* (NCBI). Barisan-barisan DNA tersebut berupa data *complete genome* dengan format FASTA [9]. Data barisan DNA sebanyak 100 dikumpulkan dalam format FASTA yang disimpan dalam format file "txt", kemudian diberi nama "data100.txt". Format FASTA diawali symbol ">", satu baris deskripsi, kemudian diikuti oleh barisan DNA yaitu A, C, G, dan T. Panjang barisan DNA pada masing-masing barisan DNA berkisar antara 7.000 bp sampai 20.000 bp. Contoh barisan DNA dengan format FASTA dapat dilihat pada Gambar 4.

```
>NC_027779.1 Human papillomavirus isolate SE379, complete genomeATGGCCACTGAGCATCCAAGAACATTC
ATATTAAAAGCCCGCAGCAATGCGTTGCAGATTTGAGTCCGAGACTTGAAGCAGTTAAAATTTCTGTGGAAGGAAAAAGTAAAAGACGATT
TGTATGGTTATTTGGGGACCTCCCGATACTGGAAAATCATTATTCTGTTTCAGCTTGTTAAATTTCTTGAAAGGTAAAGTCATTTCTTTCA
GACAAATATGGAGTGACAAAAGAATGGACTGTGCATTATAGAGATACTACTATTGTCTCCTCTTCCAGCTCCAGCAGGAGGGTCGCAGACT
GTCCCCGGCCAGTAAATCCACGGGGTCCAGCTATTGTACCTCTTAGTGAAGATGGACTACCAGAACCTGCAATAATTGGGGGCGGTGCAAC
AGACTCAGTACCAAAGTTTTTTTCTAATGACTACTTTTCCAAATTTTCAGTAATTAACAGTAACGAATCTCCTCCTATGTCAGTAGATGAA
TGGAGATGCTATGCCCGATGGAACAGTCAATCAGGATATGTTTTATTACAGTCCTGATCGTTCTCAAGGCGCACTACCTCAGAACAAGCTA
ATCAATAAATCAATATAGACATCACACCTGGCAGAGCTGTCAACCGTTTCTGGTAAGAAACTTGGCGCCAACGGCTATTGCATACCGGTTC

>KT698168.1 Human papillomavirus type 158 isolate GC23, complete genomeATGGCAAATGGACGACCTAC
AAACGAAAGTACATTAAAAGTCCCCAACAGACTGTTGCTGACTTAAGTCCACAATTGCAAGCTGTTAGAATAACACCTGAAAGAGTTTCTA
ACCCAAAAAAAAATTGTTTATTGATTCATGGTCCACCTGATACCGGAAAATCCATGTTTTGCTATTCTTTAGTTACATTCTTAAGAGGTAAA
CACAGATGCAAGCAGATATAGTGAAAGTGGAACTTGGACTGTGCATTATAAACAACAAGTAATTTCCTCCTCTGTTGTCAGCTCTTCAAAC
CTTATCAGTGGTAGATGATCTAACTACTAATACTAATATCATAACTCCAGAATTAACTACTCCTGATTTAGACATTTTGGTAGACCCAGAT
TTAAGAGTATTTGTGGATGATATAGGACAGGATTTATTTGTTTCTTATCCTGAAACAACACTTATTCCTGCATCTGGCATTCCTATAGAGC
ACCGTTGGTGATGGTATTCCAGACCCATTTGAAGCTACTTCAGACTTCTTAATTACTCCGAAAAATGACCAGAATCAGAGTAAACTAGGTT
AGTTCATTTGTACCGGTTATGGTTCACTTTGTCTGGAGGCGAAGATACGAATCTGGCTGTCTGCCTTTGTACCGGGAGTGGTGCAGAAGA
```

Gambar 4. Contoh barisan DNA dengan format FASTA

2.2. Penggunaan *N-mers Frequency*

Data barisan DNA dalam format FASTA berupa data *string*, sehingga perlu untuk melakukan *data mining* dengan mengubah data string tersebut kedalam bentuk numerik. Salah satu cara mengubah data dari bentuk string menjadi numerik dilakukan dengan menggunakan *N-mers Frequency* [4]. *N-mers Frequency* digunakan untuk menghitung banyaknya pola kemunculan N pasangan basa *nukleotida* yang sama dari suatu barisan DNA, dengan pola kemunculan sebanyak 4^N dengan $N \geq 1$. Dalam hal ini, angka 4 adalah banyaknya basa *nekleotida* pada DNA, yaitu *Adenine* (A), *Cytosine* (C), *Guanine* (G) dan *Thymine* (T) [10].

Dalam proses sintesis protein, tRNA akan membawa tiga basa *nekleotida* (anti kodon) yang akan dipasangkan dengan tiga basa *nekleotida* (kodon) pada pita mRNA yang dibentuk dari duplikasi barisan kode DNA. Oleh karena itu, pada *N-mers Frequency* ditentukan nilai $N = 3$, sehingga dimensi data akan menjadi $4^3 = 64$. Dengan kata lain, data akan memiliki 64 dimensi, 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 [11].

Syntax untuk perhitungan *N-mers Frequency* pada *software* R menggunakan fungsi *oligonucleotide Frequency* pada *library* "*Biostrings*". *Syntax* untuk *N-mers Frequency* dengan $N = 3$ pada *software* R ditunjukkan pada Gambar 5.

```
eks = oligonucleotideFrequency(reads[1:100], width=3)
```

Gambar 5. *Syntax* untuk *N-mers Frequency*

2.3. Matriks Jarak

Setelah dilakukan penggunaan *N-mers Frequency*, tahapan selanjutnya adalah membentuk matriks jarak. Matriks jarak adalah matriks simetri berukuran n x n yang elemen matriksnya merepresentasikan jarak antar data. Untuk membentuk matriks jarak digunakan persamaan *Euclidian distance* [12]. Persamaan *Euclidian distance* dapat dihitung menggunakan rumus yang ditunjukkan pada Persamaan (1),

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

dengan,

d_{ik} = jarak data ke-i dan ke-k
m = dimensi data
x_{ij} = koordinat dari data ke-i pada dimensi j
x_{kj} = koordinat dari data ke-k pada dimensi j

Syntax untuk menghitung matriks jarak pada *software* R ditunjukkan pada Gambar 6.

```
mj <- as.matrix(dist(eks, method="euclidean"))
```

Gambar 6. *Syntax* untuk menghitung matriks jarak

2.4. Algoritma Pengelompokan

Setelah membentuk matriks jarak, matriks tersebut sebagai input algoritma pengelompokan atau *clastering*. Pengelompokan ini bertujuan untuk menghitung akurasi dari penggunaan *N-mers Frequency*. Data yang akan dikelompokkan adalah data yang sudah diketahui kelompoknya masing-masing. Sebelum dikelompokkan data yang sudah diketahui kelompoknya akan dikumpulkan dan diacak, setelah itu baru dilakukan pengelompokan kembali.

Berdasarkan cara membaginya metode *clustering* terbagi menjadi dua, yaitu metode *partitioning (nonhierarchical*) dan metode *hierarchical*. Metode *hierarchical* terbagi menjadi dua yaitu, *Agglomerative Nesting* (AGNES) dan *Divisive Analysis* (DIANA). Adapun untuk metode *partitioning* terdiri dari *Self Organizing Maps* (SOM), *Partitioning Around Medoid* (PAM), *Fuzzy Analysis* (FANY) dan *K-means* [12]. Pada penelitian ini ditentukan 4 algoritma pengelompokan yang memiliki akurasi tinggi yaitu, metode *partitioning* yang terdiri atas (1) algoritma *K-means*++ dan (2) algoritma PAM, dan metode *hierarchical* yang terdiri atas (3) algoritma AGNES dan (4) algoritma DIANA.

2.4.1 Algoritma *K-means*++

Algoritma *K-means* dapat dilihat pada Algoritma 1.

Algoritma 1. Algoritma *K-means*
Input: *n* kelompok data yang akan diklaster,
Output: *K* klaster, kelompok data dengan label *n*
Steps:
1) Tentukan jumlah klaster (*K*),
2) Tentukan *centroid* (titik pusat kalster) secara *random*.
3) Hitung jarak setiap data ke *centroid*.
4) Kelompokan objek berdasarkan jarak minimum ke *centroid*.
5) Hitung *centroid* baru.
6) Ulangi langkah 3) sampai dengan 5) hingga *centroid* tidak berubah.

Salah satu kelebihan metode *K-means* adalah mampu melakukan *clustering* objek besar dengan sangat cepat sehingga metode ini sangat umum digunakan. Namun karena centroid dilakukan secara acak, memungkinkan hasil pengelompokan bersifat tidak unik (selalu berubah-ubah), sehingga keakuratannya tidak terjamin. Oleh karena itu, disarankan menggunakan algoritma *K-means*++ yang merupakan perbaikan dari algoritma *K-means*. Algoritma *K-means*++ melakukan pemilihan *centroid* secara acak sebanyak *n* kali, sedangkat algoritma *k-mean* hanya melakukan pemilihan *centroid* secara acak sebanyak 1 kali. Pemilihan *centroid* secara *random* sebanyak *n* kali pada *software* R disebut *nstart*. Dengan penambahan perhitungan *nstart* pada algoritma *K-means*++ menghasilkan hasil klaster yang unik, dengan mengambil hasil perhitungan yang mempunyai nilai minimum residual sebagai hasil klaster [13]. *Syntax* untuk Menghitung algoritma *K-means*++ pada *software* R ditunjukkan pada Gambar 7.

```
km <- kmeans(eks, 4, iter.max = 100, nstart = 100)
```

Gambar 7. *Syntax* untuk menghitung algoritma *K-means++*

2.4.2 Algoritma PAM *(Partitioning Around Medoid)*

Algoritma PAM atau *K-Medoids* dapat dilihat pada Algoritma 2.

Algoritma 2. Algoritma PAM [14] [15]
Input: *n* kelompok data yang akan diklaster,
Output: *K* klaster, kelompok data dengan label *n*
Steps:
1) Tentukan jumlah klaster (*K*).
2) Tentukan *medoid* secara acak. *Medoid* merupakan data yang terpilih untuk mewakili klaster.
3) Hitung jarak setiap data *non medoid* ke *medoid*.
4) Kelompokan objek berdasarkan jarak minimum ke *medoid*.
5) Hitung total *distance*.
6) Ulangi langkah 2) sampai dengan 5) hingga total *distance* baru - total *distance* lama < 0.

Syntax untuk Menghitung algoritma PAM pada *software* R ditunjukkan pada Gambar 8.

```
pm <- pam(data, 4, diss = TRUE)
```

Gambar 8. *Syntax* untuk menghitung algoritma PAM

2.4.3 Algoritma AGNES *(Agglomerative Nesting)*

Algoritma AGNES menggunkan *Average Linkage* ditunjukkan pada Algoritma 3.

Algoritma 3. Algoritma AGNES [5]
Input: Matriks Jarak
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 rumus pada Persamaan 2.

$$d_{(UV)} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \tag{2}$$

dengan,
d_{ik} = jarak antara spesies *i* dalam kelompok *(UV)* dan spesies *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 pada *software* R ditunjukkan pada Gambar 9.

```
M = agnes(data, metric = "euclidean", stand = TRUE)
```

Gambar 9. *Syntax* untuk menghitung algoritma AGNES

2.4.4 Algoritma DIANA *(Divisive Analysis)*

Algoritma DIANA ditunjukkan pada Algoritma 4.

Algoritma 4. Algoritma DIANA [13][16]
Input: Matriks Jarak
Output: Dendogram
Steps:
1) Tentukan objek yang memiliki nilai rata-rata disimilaritas terbesar, objek tersebut akan dipisah menjadi klaster baru dan menjadi *splinter group.*
2) Untuk setiap objek *i* diluar *splinter group,* hitung selisih nilai antara elemen matriks *splinter group* dengan nilai rata-rata setiap sekuens yang tersisa, dengan rumus

$$D_i = \left[\text{average } d(i,j) j \notin R_{\text{splinter group}}\right] - \left[\text{average } d(i,j) j \in R_{\text{splinter group}}\right]$$

3) Tentukan objek yang memiliki nilai selisih terbesar antara elemen matriks *splinter group* dengan nilai rata-rata. Jika nilai selisih tersebut bernilai positif, maka objek yang memiliki nilai selisih terbesar bergabung dengan *splinter group.*
4) Ulangi langkah 2 dan 3 sedemikian sehingga semua nilai selisih antara elemen matriks *splinter group* dengan nilai rata-rata bernilai negatif. Maka klaster terbagi menjadi dua klaster baru.
5) Pilih klaster dengan diameter terbesar. Diameter klaster merupakan disimilaritas terbesar antara sebarang dua objek dalam satu klaster. Kemudian bagi klaster ini, mengikuti langkah 1 sampai 4.
6) Ulangi langkah 5 sampai memperoleh klaster yang dikehendaki atau semua klaster telah berisi satu objek.

Syntax untuk Menghitung algoritma DIANA pada *software* R ditunjukkan pada Gambar 10.

```
M <- diana(data, metric = "Euclidean", stand = TRUE)
```

Gambar 10. *Syntax* untuk menghitung algoritma DIANA

2.5. Menghitung Akurasi

Untuk menghitung akurasi penggunaan *N-mers Frequency*, data yang sudah diketahui jenisnya akan diacak, kemudian dikelompokkan kembali. Data yang tidak kembali pada kelompoknya akan dihitung, dan akan dihitung persentasenya. Untuk 100 data, Misalkan ada satu data yang tidak mengelompok dengan benar, maka bisa dikatakan 1% tidak akurat, atau dengan kata lain keakuratannya 99%. Persentase akurasi penggunaan *N-mers Frequency* dapat dihitung dengan menggunakan rumus pada Persamaan (3).

$$\% A = \frac{x}{100} \times 100\ \% \tag{3}$$

dengan,

$\% A$ = persentase akurasi penggunaan *N-mers Frequency*

x = banyaknya data yang mengelompok ke kelompoknya kembali.

3. Hasil dan Pembahasan

3.1. Data Penelitian

Data barisan DNA HPV 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 diskripsi data *(name)*. Hasil *readDNAStringSet* dapat dilihat pada Gambar 11.

```
> reads
  A DNAStringSet instance of length 100
      width seq                                                          names
  [1]  7353 GTCTGTAATGATAGTTGGCAACAATCATTACTTATAG...CAGCCTTTGCACCGGGAGTGGTGGAAAATAGTTTCT X70827.1 Human pa...
  [2] 18959 CGGACACACAAAAAGAAAGAAGAATTTTTAGGATCTT...TAAAAATAAATCTATTTCTTCTTTTTTGTGTGTCCA KC242799.1 Zaire ...
  [3] 19875 AGACACACAAAAACAAGAGATGATGATTTTGTGTATC...GTAACACAAAACATTTTTCATCTTTTTTGTGTGTCC MK271062.1 Mutant...
  [4] 11155 CTGTGTGAATCAGACTGCGACAGTTCGAGTTTGAAGC...GCACTCCAGCCTGGGCGACAGAGCAAGACTCCGTCT KY766069.1 Zika v...
  [5]  7746 AACGGTAAGTTGCAATTTCCTTGTACCAGGTGCGGTA...TTTGGCAATCGCATTTGGCACTGCTAAAAGACCGTT M17463.1 Human pa...
  ...   ... ...
 [96] 10808 AGTTGTTGATCTGTGTGAATCAGACTGCGACAGTTCG...ATAGCGGCGGCCGGTGTGGGGAAATCCATGGGTCTT KY415986.1 Zika v...
 [97]  7560 CCACATTCGTTCCAGCTACATTTTGGCGCCAACTCTT...CCAGAAGTGTGTTTTTGCCAAGACATTTGCCAAGTA NC_001591.1 Human...
 [98] 19897 CGGACACACAAAAAGAAAGAAGAATTTTTAGGATCTT...TAAAAATAAATCTATTTCTTCTTTTTTGTGTGTCCA KU174139.1 Mutant...
 [99] 19114 AGACACACAAAAACAAGAGATGATGATTTTTGTGTATC...TAACACAAAACATTTTTCATCTTTTTAGTGTGTCCA FJ750959.1 Lake V...
[100] 10808 AGTTGTTACTGTTGCTGACTCAGACTGCGACAGTTCG...ATAGCGGCGGCCGGTGTGGGGAAATCCATGGGTCTT KX893855.1 Zika v...
```

Gambar 11. Hasil *readDNAStringSet*

3.2. Hasil *N-mers Frequency*

Untuk mengubah data barisan DNA HPV yang berupa data *string* menjadi data numerik, digunakan *N-mers Frequency*. Dengan menentukan $N = 3$, diperoleh dimensi data sebesar $4^3 = 64$. Dengan demikian, data memiliki 64 atribut atau dimensi. Pada Gambar 12 ditunjukkan hasil *N-mers Frequency* dari 100 barisan DNA menggunakan *software* R dengan hasil Ekstraksi ciri berupa matriks yang berukuran 100×64.

	AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT	AGA	AGC	AGG	AGT	ATA	ATC	ATG	ATT	CAA
1	261	126	154	215	165	84	52	103	214	88	116	136	180	97	152	220	
2	617	415	407	533	502	281	136	316	424	247	294	289	356	386	329	521	
3	672	422	455	588	502	262	111	341	438	218	295	296	428	391	377	582	
4	244	165	287	134	239	153	85	169	320	201	263	163	108	143	257	117	
5	234	134	184	172	199	92	54	115	177	105	170	115	174	103	130	218	
6	613	410	410	528	496	278	137	324	423	243	291	296	354	390	325	520	
7	697	386	440	595	470	223	99	316	410	216	282	298	418	393	364	565	
8	234	157	284	136	257	164	75	153	324	199	273	175	106	149	282	108	
9	253	181	131	179	275	114	67	106	125	81	124	118	232	68	215	218	

Gambar 12. Hasil *N-mers Frequency*

3.3. Hasil Perhitungan Matriks Jarak

Untuk membentuk matriks jarak digunakan persamaan *Euclidian distance.* Pada Gambar 13 ditunjukkan hasil perhitungan matriks jarak dari 100 barisan DNA menggunakan *software* R, menghasilkan matriks berukuran 100×100.

Filter

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.00000	1576.921051	1733.2905	805.17079	187.64061	1573.72679	1652.292952	823.98544	349.13608	1577.075141	1652.103810	776.82752	206.55
2	1576.92105	0.000000	316.1835	1366.11713	1551.57694	37.97368	319.924991	1375.84592	1560.57041	2.000000	320.320152	1390.07806	1568.74
3	1733.29051	316.183491	0.0000	1581.37662	1723.90922	324.29308	142.975522	1591.44086	1713.00905	315.772070	143.230583	1603.17404	1745.83
4	805.17079	1366.117125	1581.3766	0.00000	721.67791	1358.93341	1531.220428	91.54234	827.78620	1366.409163	1530.978445	48.36321	706.79
5	187.64061	1551.576940	1723.9092	721.67791	0.00000	1548.11014	1646.355672	744.66570	316.55173	1551.756746	1646.265471	692.92568	103.27
6	1573.72679	37.973675	324.2931	1358.93341	1548.11014	0.00000	329.226366	1369.10263	1559.50761	38.105118	329.743840	1383.10701	1565.36
7	1652.29295	319.924991	142.9755	1531.22043	1646.35567	329.22637	0.000000	1542.51094	1638.71291	319.374388	4.795832	1551.86887	1669.16
8	823.98544	1375.845922	1591.4409	91.54234	744.66570	1369.10263	1542.510940	0.00000	841.10760	1376.139528	1542.290180	107.90273	727.69
9	349.13608	1560.570409	1713.0090	827.78620	316.55173	1559.50761	1638.712910	841.10760	0.00000	1560.694717	1638.554546	801.86470	316.97

Showing 1 to 10 of 100 entries

Gambar 13. Hasil perhitungan matriks jarak

3.4. Hasil Pengelompokan

Setelah diperoleh hasil perhitungan matriks jarak, tahap selanjutnnya adalah pengelompokan 100 barisan DNA yang sudah diacak. Pengelompokan yang dilakukan menggunakan algoritma *K-means*++, PAM, AGNES, dan DIANA. Dilakukan pengelompokan sebanyak 4 kelompok, hal ini didasarkan pada kelompok yang telah ditentukan sebelumnya, yaitu virus HPV, Ebola, Marburg, Zika. Hasil pengelompokan masing-masing ditunjukkan pada Gambar 14 s.d Gambar 17.

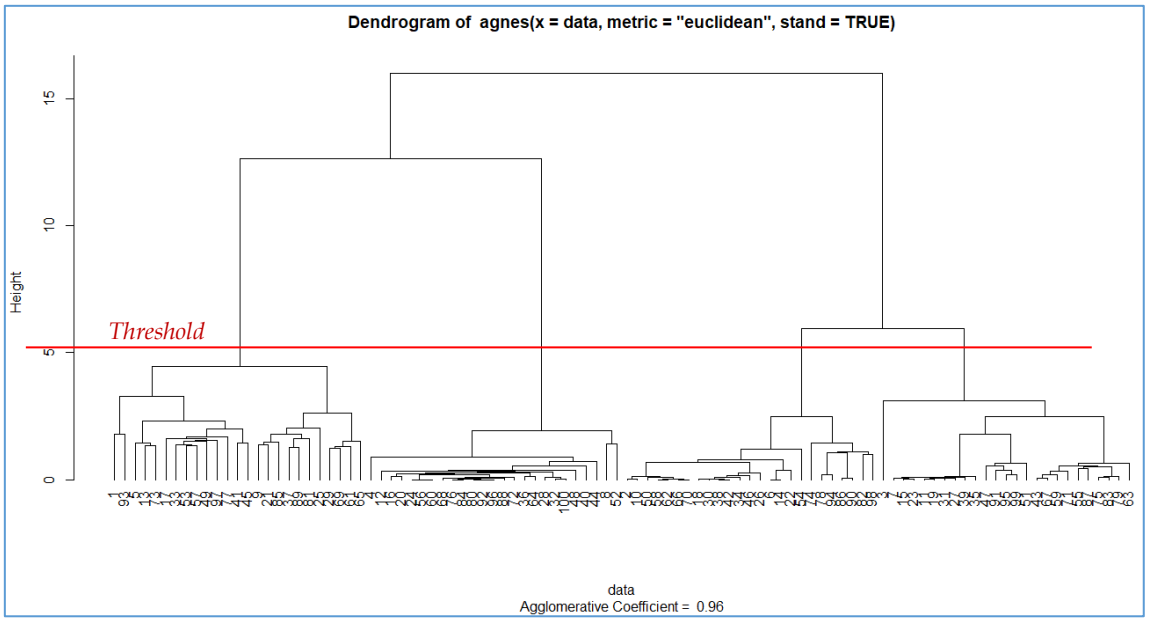
```
> source('~/a1 K-means++ 100.R')
 [1]  2  6 10 14 18 22 26 30 34 38 42 46 50 54 58 62 66 70 74 78 82 86 90 94 98
 [1]  3  7 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
 [1]  4  8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100
 [1]  1  5  9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97
 Clustering vector:
   [1] 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2
  [53] 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2 3 1 4 2

 within cluster sum of squares by cluster:
 [1] 103329.44  23223.28 688095.28 151454.96
  (between_SS / total_SS =  98.5 %)
```

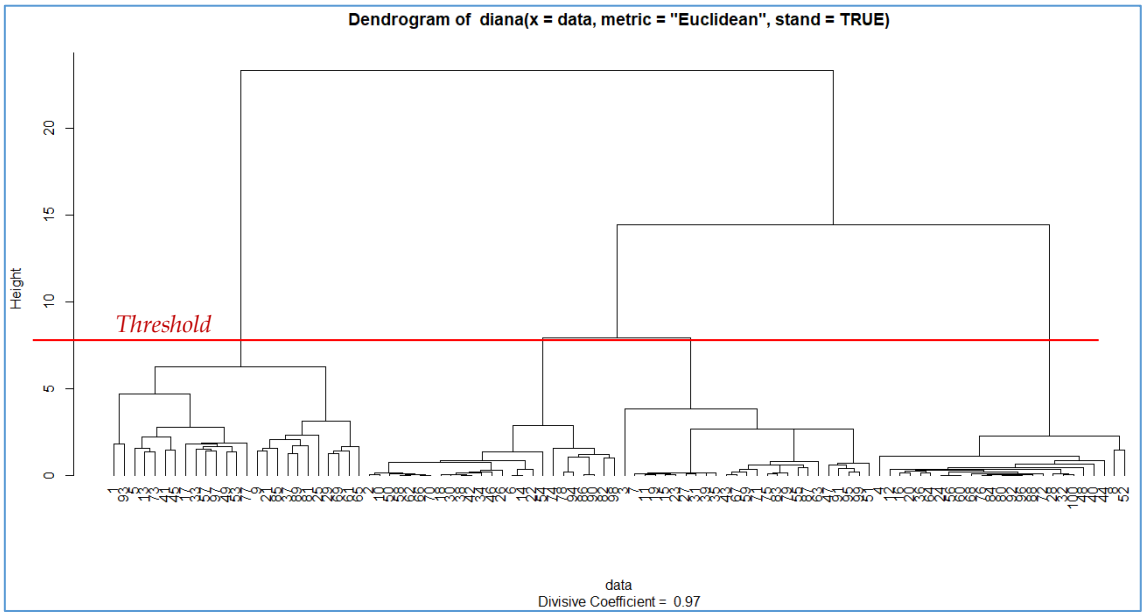
Gambar 14. Hasil pengelompokan menggunakan algoritma *K-means*++

```
> source('~/a2 pam 100.R')
 [1]  1  5  9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97
 [1]  2  6 10 14 18 22 26 30 34 38 42 46 50 54 58 62 66 70 74 78 82 86 90 94 98
 [1]  3  7 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
 [1]  4  8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100
 Medoids:
      ID
 [1,] 57 57
 [2,] 10 10
 [3,]  7  7
 [4,] 80 80
 Clustering vector:
   [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
  [53] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
 Objective function:
   build     swap
 83.20108 83.15917
```

Gambar 15. Hasil pengelompokan menggunakan algoritma PAM



Gambar 16. Dendogram hasil pengelompokan menggunakan algoritma AGNES



Gambar 17. Dendogram hasil pengelompokan menggunakan algoritma DIANA

Pada Gambar 16 dan Gambar 17 ditentukan sebuah *threshold* yang memotong Dendogram menjadi 4 kelompok dikarenakan kelompok yang ditetapkan sebelumnya adalah 4 kelompok. Dilihat dari *Agglomerative Coefficient* = 0.96 dan *Divisive Coefficient* = 0.97 pengelompokan yang terbentuk sangat baik. Hal tersebut juga dikonfirmasi oleh hasil pengelompokan algoritma *K-means*++ pada Gambar 14, bahwa persentase perbandingan *between_SS* dan *total_SS* sebesar 98,5% yang menunjukkan bahwa pengelompokan tersebut sangat baik.

Dari keseluruhan hasil pengelompokan menggunakan algoritma *K-means*++, PAM, AGNES, dan DIANA pada Gambar 14 sampai 17, diperoleh 4 klaster sama. Anggota

masing-masing klaster yang telah diurutkan disajikan pada Tabel 1.

Tabel 1. Hasil pengelompokan

Klaster	Barisan DNA ke
Klaster 1 - Ebola (25 anggota)	2 6 10 14 18 22 26 30 34 38 42 46 50 54 58 62 66 70 74 78 82 86 90 94 98
Klaster 2 - Marburg (25 anggota)	3 7 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Klaster 3 - Zika 25 anggota	4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100
Klaster 4 - HPV (25 anggota)	1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97

Dari Tabel 1, diperoleh informasi bahwa Klaster 1 merupakan pengelompokan barisan DNA virus Ebola, Klaster 2 merupakan pengelompokan barisan DNA virus Marburg, Klaster 3 merupakan pengelompokan barisan DNA virus Zika, dan Klaster 4 merupakan pengelompokan barisan DNA virus HPV.

3.5. Hasil Perhitungan Akurasi

Hasil perhitungan akurasi secara detail disajikan pada Tabel 2.

Tabel 2. Hasil perhitungan akurasi

Algoritma	Akurasi
K-means++	100 %
PAM	100 %
AGNES	100 %
DIANA	100 %
Rata-rata	100 %

Dari Tabel 2 terlihat Akurasi dari berbagai pengelompokan menggunakan algoritma *K-means*++, PAM, AGNES, dan DIANA adalah 100%, sehingga diperoleh rata-rata 100%. Dengan kata lain tidak ada satupun barisan DNA yang salah dalam penentuan klasternya. Hal ini menunjukkan bahwa *N-mers Frequency* sangat baik digunakan dalam analisis barisan DNA.

Akurasi tinggi ini dikarenakan oleh konversi dari data *string* ke data *numerik* dengan bahwa *N-mers Frequency* dengan $N = 3$ yang dilandasi oleh teori proses sintesis, dimana hasil dari *N-mers Frequency* merupakan data yang mempunyai 64 dimensi yang elemennya mewakili jumlah kodon pembentuk protein. Oleh karena itu, ketika dilakukan c*lustering* atau analisis lainnya akan diperoleh hasil yang baik.

4. Kesimpulan

Akurasi penggunaan *N-mers Frequency* dalam kasus ini menggunakan 100 barisan DNA mempunyai akurasi yang sangat baik yaitu 100%. Tidak ada satupun barisan

DNA yang salah dalam penentuan kelompoknya kembali. Klaster 1 adalah pengelompokan barisan DNA virus Ebola, Klaster 2 adalah pengelompokan barisan DNA virus Marburg, Klaster 3 adalah pengelompokan barisan DNA virus Zika, dan Klaster 4 adalah pengelompokan barisan DNA virus HPV.

Ucapan Terimakasih

Terima kasih kepada Institut dan Teknologi Bisnis Kalbis yang telah membantu pembiayaan dalam penelitian ini dan menyediakan sarana dan prasarana pada pelaksanaan penelitian.

Referensi

[1] A. Lucassen, J. Montgomery, and M. Parker, "Ethics and the Social Contract for Genomics in the NHS," 2017.

[2] NCBI, "National Center for Biotechnology Information," *U.S. National Library of Medicine*. [Online]. Available: https://www.ncbi.nlm.nih.gov. [Accessed: 15-Feb-2018].

[3] Xinhua, "China to Create Gigantic DNA Database," 2017. [Online]. Available: http://www.chinadaily.com.cn/china/2017-10/31/content_33930020.htm. [Accessed: 22-Oct-2018].

[4] B. Chor, D. Horn, N. Goldman, Y. Levy, and T. Massingham, "Genomic DNA k-mer spectra: models and modalities," *Genome Biol.*, vol. 10, no. 10, p. R108, 2009.

[5] A. Bustamam, I. Fitria, and K. Umam, "Application of Agglomerative Clustering for Analyzing Phylogenetically on Bacterium of Saliva," in *AIP Conference Proceedings*, 2017, p. 030126.

[6] S. M. Gollin, "Epidemiology of HPV-Associated Oropharyngeal Squamous Cell Carcinoma," in *Human Papillomavirus (HPV)-Associated Oropharyngeal Cancer*, D. L. Miller and M. S. Stack, Eds. Cham: Springer International Publishing, 2015, pp. 1–23.

[7] E. Mühlberger, "Genome Organization, Replication, and Transcription of Filoviruses," in *Ebola and Marburg Viruses: Molecular and Cellular Biology*, H.-D. Klenk and H. Feldmann, Eds. Winnipeg: Horizon Bioscience, 2004.

[8] S. R. da Silva, F. Cheng, and S.-J. Gao, *Zika Virus and Diseases*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2018.

[9] NCBI, "Nucleotide - National Center for Biotechnology Information," *U.S. National Library of Medicine*. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore. [Accessed: 15-Feb-2018].

[10] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, 2nd ed. Tucson: Cold Spring Harbor Laboratory Press, 2004.

[11] K. Umam, A. Bustamam, and D. Lestari, "Application of hybrid clustering using parallel k-means algorithm and DIANA algorithm," in *AIP Conference Proceedings*, 2017, p. 020024.

[12] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1990.

[13] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*, 1st ed. CRC Press, 2013.

[14] D. Arthur and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[15] R. D. Cahyaningrum, A. Bustamam, and T. Siswantining, "Implementation of spectral clustering with partitioning around medoids (PAM) algorithm on microarray data of carcinoma," in *AIP Conference Proceedings*, 2017, p. 020007.

[16] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham: Morgan Kaufmann Publishers, 2012.