JURNAL INFORMATIKA, Vol. 9 No. 2 Oktober 2022, Halaman 140~147
ISSN: 2355-6579 | E-ISSN**:** 2528-2247

Penentuan Kekerabatan Hewan Berdasarkan Struktur Protein IGF2 Menggunakan Metode K-Means dan N-Gram

Ruth Ema Febrita[1], Maghfirotul Amaniyah[2]

[1,2] Politeknik Negeri Banyuwangi
Jl. Raya Jember KM 13, Banyuwangi, Indonesia

e-mail: [1]ruthemafebrita@poliwangi.ac.id, [2]maghfirotulamaniyah@poliwangi.ac.id

Informasi Artikel	Diterima: 30-08-2022	Direvisi: 26-09-2022	Disetujui: 29-09-2022

Abstrak

Dalam ilmu Biologi, terdapat berbagai cara untuk menentukan kedekatan antar dua individu, antara lain dengan mengamati kesamaan morfologi fisik kemudian membuat dendogram dan pembuatan pohon filogeni untuk menelusur kekerabatan berdasarkan sejarah evolusi suatu makhluk hidup. Akan tetapi pendekatan ini sangat sulit untuk dilakukan apabila hewan yang akan ditentukan kekerabatannya tidak berada dalam kondisi yang hidup, sehingga sangat sulit untuk mengamati ciri-ciri fisik yang ada. Penelitian ini bertujuan untuk memberikan pendekatan yang berbeda dalam menentukan kekerabatan hewan dengan menggunakan algoritma clustering untuk mengelompokkan struktur protein IGF2. Kekerabatan dilakukan dengan menggunakan metode clustering K-Means. Untuk memudahkan dalam melakukan pengelompokkan struktur protein yang memiliki panjang sekuens yang beragam, maka teknik n-gram digunakan untuk memecah string menjadi beberapa subsekuens dengan panjang yang sama. Pengelompokkan dengan metode K-Means telah dilakukan dan mendapatkan hasil terbaik pada jumlah cluster sebanyak tujuh cluster, dengan *silhouette coeficient* rata-rata sebesar 0.331, indeks *purity*sebesar 0.735, dan *precision*sebesar 0.823 yang mengindikasikan proses clustering cukup efektif.

Kata Kunci: analisis kekerabatan, k-means, n-gram

Abstract

In Biology, there were various ways to determine the closeness between two individuals, such as by observing the similarity of physical morphologies then making a dendogram and also by making a phylogenetic tree to trace the kinship based on the evolutionary history. However, this approach is very difficult to do if the animal whose relatives are to be determined is not in a living condition, so it is very difficult to observe the existing physical characteristics. This study aims to provide a different approach in determining animal kinship using clustering algorithm to cluster the IGF2 protein structures. Kinship is determined using the K-Means clustering method. N-gram technique is used to break the sequence into several subsequences with the same length, because each sequence can have various length. Grouping with the K-Means method had been done and got the best results on the number of clusters as many as seven clusters, with an average silhouette coefficient of 0.331, a purityindex of 0.735, and a precisionof 0.823 which indicates the clustering process is quite effective.

Keywords*: animal kinship analysis, k-means, n-gram*

1. Pendahuluan

Ilmu yang mempelajari kekerabatan/ kedekatan antara satu makhluk hidup dengan makhluk hidup lainnya merupakan salah satu cabang dari ilmu Biologi. Penentuan kekerabatan antara beberapa spesies makhluk hidup sering dilakukan dengan beberapa tujuan, antara lain untuk menentukan tingkat sebaran spesies yang terdapat pada suatu daerah, hingga mengetahui silsilah evolusi suatu makhluk hidup berdasarkan kedekatan dengan nenek moyangnya. Manfaat mengetahui kekerabatan antara dua makhluk hidup adalah untuk mencari perbedaan dan persamaan antara dua spesies. Selain itu, penentuan kekerabatan dapat digunakan untuk mengetahui marker (penciri) antar spesies. Dua makhluk hidup yang memiliki kekerabatan yang dekat

dapat memiliki kemiripan dalam ciri-ciri fisik, morfologi, cara bereproduksi, manfaat, kandungan gen dalam kromosom, serta kandungan zat kimia.

Prasgi dkk (2022) melakukan analisis kekerabatan fenetik varietas tanaman gulma dengan pendekatan karakteristik morfologi antara lain: bentuk bunga, bakal buah, batang, daun, dan bentuk akar. Dalam melakukan analisis kekerabatan, analisis cluster dan indeks metriks kemiripan digunakan sehingga menemukan dua cluster utama, dengan perbedaan mendasar pada warna bunga, dan properti-properti lainnya.

Sementara itu, pendekatan fenetik juga digunakan oleh (Riandini & Astuti, 2020) untuk menganalisis hubungan kekerabatan tanaman pisang. Dalam melakukan analisis berdasarkan ciri-ciri fisik, peneliti menggunakan fenogram untuk menggambar kekerabatan varietas pisang yang ada. Pengelompokan menggunakan fenogram juga menghasilkan cluster terbaik sebanyak dua cluster utama.

Beberapa penelitian terdahulu telah menggunakan ciri-ciri fisik atau morfologi dalam penentuan tingkat kedekatan antar dua individu. Penelitian ini ingin bertujuan untuk melakukan pendekatan yang berbeda dalam penentuan kedekatan individu, khususnya dalam kingdom animalia (hewan), yang dilakukan berdasarkan pengelompokan berbasis analisis struktur urutan asam amino protein, serta bagaimana hasil cluster yang dihasilkan dengan menggunakan algoritma pengelompokan K-Means. Melalui algoritma clustering diharapkan dapat mengelompokkan hewan yang memiliki struktur protein yang mirip ke dalam suatu cluster yang sama. Struktur protein IGF2 digunakan sebagai fitur data yang akan diolah dengan tujuan supaya hasil penelitian dapat dimanfaatkan lebih lanjut untuk mengetahui kemiripan manfaat yang terkandung dalam suatu daging hewan dan kedekatan kandungan dalam daging hewan dalam rangkaian penelitian autentikasi produk halal (Febrita & Amaniyah, 2021).

IGF2 digunakan dalam penelitian ini karena IGF2 merupakan protein yang disekresikan, yang berperan penting dalam pertumbuhan dan perkembangan janin pada masa prenatal. IGF2 berperan dalam proses diferensiasi makhluk hidup. IGF2 diturunkan secara genetis oleh orang tua kepada anak yang dapat dijumpai dalam struktur kromosom (Baral & Rotwein, 2019; Xiang et al., 2018). Alel IGF2 paternal dilaporkan menjadi bagian yang ditranskripsi pertama kali sehingga hampir selalu terekspresi pada tiap perkembangan makhluk hidup. Identifikasi fungsi dan ekspresi protein IGF2 banyak dipelajari, terutama yang berkaitan dengan daging dan penyakit kanker (Wei et al., 2018; Criado-Mesas et al., 2019).

K-Means merupakan algoritma clustering yang banyak diterapkan karena memiliki keandalan dalam mengelompokkan data dalam jumlah yang besar dengan waktu komputasi yang cepat dan efisien. Terdapat beberapa penelitian terdahulu yang telah berhasil menerapkan K-Means dalam pengelompokan data. K-Means berhasil diterapkan dalam pengelompokan dokumen skripsi (Adhe et al., 2020). Dokumen skripsi yang digunakan sebagai dataset terlebih dahulu dilakukan prosedur preprosesing, yakni *case folding, tokenizing, filtering,* dan *stemming*. Prosedur preprosesing dilakukan agar data yang akan diolah menjadi bentuk yang paling asli dan menghilangkan noise yang disebabkan oleh kata-kata yang tidak terlalu signifikan maknanya sehingga proses komputasi dapat berlangsung dengan lebih ringan dan cepat. Setelah itu, algoritma akan membuat matriks TF-IDF yang menunjukkan nilai seberapa pentingnya sebuah kata dalam suatu dokumen dilihat dari frekuensi kemunculan kata dalam dokumen tersebut dan dokumen lainnya. Matriks TF-IDF inilah yang kemudian digunakan dalam proses clustering. Hasil cluster terbaik ditemukan saat nilai cluster n=2 dengan *silhouette score* sebesar 0.12.

K-Means juga diimplementasikan untuk mengelompokkan berita (Yudiarta et al., 2018) dan tweet pertanian (Irsyad & Pribadi, 2020). Metriks yang digunakan untuk mengukur seberapa baik hasil cluster tidak hanya menggunakan *silhouette score*, melainkan juga menggunakan *precision, recall,* dan *purity*dari hasil cluster. *Precision* merupakan suatu rasio antara data yang benar-benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Sedangkan *recall* digunakan untuk mengukur sensitifitas hasil prediksi. Sedangkan *purity* menunjukkan seberapa tercampurnya data-data yang terdapat dalam sebuah cluster (James et al., 2018). Salah satu hal yang dapat mempengaruhi hasil cluster adalah pemilihan centroid awal. Pemilihan centroid awal secara acak dapat menghasilkan ketidakakuratan hasil cluster, sehingga perlu dilakukan optimasi pemilihan inisialisasi centroid.

2. Metode Penelitian

Penelitian ini menggunakan metode K-Means untuk melakukan clustering pada sekuens protein yang telah dipotong menjadi subsekuen dengan menggunakan n-gram. Prosedur text-mining akan dilakukan dalam sekumpulan subsekuen protein tersebut yaitu dengan pembuatan matriks Term Frequency-Invers Document Frequency (TF-IDF) yang

dapat merepresentasikan bobot setiap subsekuen dengan memperhatikan frekuensi munculnya subsekuens tersebut pada setiap sekuens protein yang terdapat dalam dataset. Matriks TF-IDF kemudian diubah dalam bentuk vektor dan digunakan untuk melakukan proses clustering dengan menggunakan K-Means, dimana hipotesis yang digunakan adalah sekuens yang berada pada satu cluster yang sama akan memiliki tingkat kekerabatan yang cukup dekat. Adapun prosedur yang lebih mendetail akan dijelaskan pada sub bab berikut.

1. Dataset

Dataset yang digunakan merupakan sekuens protein hewan berbentuk IGF2 yang diambil dari Universal Protein Resource, yang merupakan suatu portal yang menyediakan informasi mengenai protein dan anotasinya (Bateman et al., 2021). Jumlah data yang digunakan sebanyak 34 sekuens dari kingdom animalia yang terdiri dari 12 spesies, yang terdiri dari kelompok mamalia (manusia, babi, kerbau, anjing, tikus, kelinci, kuda, sapi, kambing, domba) dengan pembanding spesies dari kelompok di luar mamalia berupa ayam dan ikan. Penggunaan beberapa jenis spesies dimaksudkan untuk mengetahui efektifitas metode clustering yang diterapkan, dengan asumsi bahwa metode disebut efektif apabila mampu mengelompokkan spesies yang mirip dalam satu cluster. Dalam satu spesies digunakan beberapa sekuens dengan nomor akses yang berbeda. Sekuens protein yang digunakan diberi label sesuai jenis spesiesnya. Label ini nantinya tidak akan digunakan dalam proses clustering, melainkan dalam pengujian hipotesis penelitian.

2. Pembuatan Vektor TF-IDF

Setiap sekuens protein akan diubah menjadi subsekuen menggunakan n-gram dengan n=10. Adapun contoh pemotongan sekuen protein dengan n=5 dijelaskan dalam Tabel 1.

Tabel 1. Ilustrasi Pemecahan Sekuens dengan N-Gram

Sekuens	Subsekuens (n=5 gram)
MVSPTSQIIVVAPETELLA	['MVSPT', 'VSPTS', 'SPTSQ', 'PTSQI', 'TSQII', 'SQIIV', 'QIIVV', 'IIVVA', 'IVVAP', 'VVAPE', 'VAPET', 'APETE', 'PETEL', 'ETELL', 'TELLA']

Subsekuen tersebut dijadikan dasar dalam pembuatan matriks TF-IDF. *Term frequency* (tf) digunakan untuk menghitung bobot sebuah subsekuen berdasarkan frekuensi kemunculan subsekuen tersebut pada suatu string protein (sekuen). Semakin sering suatu subsekuen muncul pada suatu string protein, maka semakin tinggi nilai bobotnya.

Inverse Document Frequency (IDF) menghitung seberapa bernilainya suatu subsekuen pada suatu database string protein dengan menghitung kemunculan subsekuen pada kumpulan string protein yang ada. Semakin jarang dokumen yang memiliki subsekuen tersebut, maka dapat dikatakan subsekuen tersebut merupakan penciri pada suatu sekuen dan sangat bernilai. Adapun formula dari IDF adalah:

$$Idf_x = \log\left(\frac{\sum sekuen\ dalam\ database}{\sum sekuen\ yang\ mengandung\ subsekuen\ x}\right) \quad (1)$$

Dengan demikian, TF-IDF merupakan perkalian antara nilai TF dan IDF suatu subsekuen.

3. Pengelompokan dengan K-Means

Matriks TF-IDF kemudian akan diubah dalam bentuk vektor. Vektor inilah yang digunakan dalam proses clustering. Adapun tahapan-tahapan dari proses clustering adalah sebagai berikut:

1. Menentukan jumlah cluster (n) yang nanti akan dibentuk selama proses clustering berlangsung. Dalam hal ini, akan dilakukan proses clustering dengan nilai n beragam, yaitu n=[2,9] untuk mendapatkan cluster yang terbaik.
2. Penentuan n buah centroid secara acak dari objek vektor tf-idf yang telah didapatkan.
3. Perhitungan jarak antara vektor lainnya dengan centroid. Vektor akan menjadi anggota cluster pada centroid terdekat. Dalam penelitian ini digunakan 3 jenis perhitungan jarak berbeda, yaitu: *euclidean distance* dan *manhattan distance* guna memberikan perspektif yang berbeda dalam mengevaluasi hasil cluster. Adapun perhitungan *euclidean distance* akan dijelaskan pada persamaan (2) dan *manhattan distance* dijelaskan pada persamaan (3).

$$D_{(x,y)} = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \quad (2)$$

$$D_{(x,y)} = \sum_{i=1}^{n}|y_i - x_i| \quad (3)$$

Dimana D(x,y) adalah jarak antara dua vektor (x dan y) yang akan dihitung, sedangkan i adalah fitur ke-i pada vektor.

4. Menghitung kembali posisi centroid dengan menghitung rata-rata jarak tiap fitur vektor yang menjadi anggota suatu cluster dengan menggunakan persamaan (4).

$$centroid = \frac{\sum_{i=1}^{n} x_i}{n}; i = 1,2,3, \dots n \ (4)$$

Dimana n adalah banyaknya anggota cluster, dan x_i merupakan obyek ke-i.

5. Apabila posisi centroid berubah, maka ulang kembali langkah 3 dan 4. Centroid yang tidak mengalami perubahan menandakan bahwa proses clustering sudah mengelompokkan setiap data dalam cluster terdekat.

4. Evaluasi

Evaluasi merupakan tahap untuk menilai seberapa baik hasil cluster yang telah terbentuk. Dalam penelitian ini akan digunakan beberapa metode yaitu: *silhouette coeficient, precision, recall,* dan *purity*. Pada perhitungan *silhouette coeficient*, hasil cluster yang baik apabila jarak antar anggota suatu cluster semakin kecil dan jarak antara anggota suatu cluster dengan cluster yang berbeda semakin besar (Haviluddin et al., 2021). Perhitungan *silhouette coeficient* akan ditampilkan pada persamaan (5)

$$SC_{(i)} = \frac{b_i - a_i}{\max(a_i, b_i)}, i = 1,2,3, .. n \ (5)$$

Dimana b_i merupakan jarak terkecil antar cluster dan a_i merupakan rata-rata jarak antar data pada suatu cluster. Nilai *silhouette coeficient* terletak dalam range [-1,1], dimana hasil cluster efektif apabila nilai *silhouette* semakin mendekati 1, dan sangat tidak efektif apabila nilai *silhouette* mendekati -1. Cluster disebut overlap apabila nilai silhouette = 0 (Febrita et al., 2019).

Precision dan *recall* digunakan untuk mengukur ketepatan hasil cluster berdasarkan label yang telah dimiliki oleh setiap sekuens protein. Perhitungan *precision* akan ditampilkan dalam persamaan (6).

$$precision = \frac{TP}{TP+FP} (6)$$

Dimana TP merupakan data yang diprediksi termasuk dalam kelas positif dan memang ada dalam kelas positif secara fakta. FP merupakan data yang diprediksi termasuk dalam kelas positif, namun secara fakta berada di kelas negatif.

Purity merepresentasikan tingkat kemurnian suatu hasil cluster dengan menghitung berapa banyak anggota cluster yang cocok dalam suatu cluster, dimana label cluster ditentukan dari frekuensi terbanyak label yang muncul dalam cluster tersebut. *Purity* dapat dihitung menggunakan persamaan (7).

$$Purity = \frac{1}{N}\sum_{i=1}^{j} max_j|c_i \cap t_j| \ (7)$$

Dimana N merupakan jumlah sekuens protein yang ada dalam keseluruhan cluster, j merupakan jumlah cluster yang terbentuk, $max_j|c_i \cap t_j|$ merupakan jumlah maksimum data yang cocok pada suatu label cluster j.

Penentuan tingkat kedekatan antara dua sekuens dapat dilakukan dengan menghitung *cosine similarity* antara dua vektor sekuens. Hasil perhitungan *cosine similarity* menunjukkan besaran sudut antara dua vektor, dimana jika similarity = 0, maka kedua vektor memiliki sudut perbedaan 90 derajat, yang artinya tidak memiliki kemiripan. Sedangkan apabila nilai similarity semakin mendekati nilai 1, maka jarak/ kemiripan antara dua vektor semakin dekat (Dwi et al., 2019). *Cosine similarity* dapat dihitung menggunakan persamaan 8.

$$cosine_{(x,y)} = \frac{x \,.y}{\|x\|.\|y\|} \ (8)$$

Dimana x.y adalah hasil perkalian antara vektor x dan y, $\|x\|$ dan $\|y\|$ adalah panjang vektor.

3. Hasil dan Pembahasan

Metode K-Means clustering telah diterapkan dalam mengelompokkan sekuens protein IGF-2, dengan data pengujian yang akan disajikan pada Tabel 2.

Tabel 2. Hasil dan Pengujian

ID	Obyek	Len	Label cluster saat N=							
			2	3	4	5	6	7	8	9
1	Oryctolagus cuniculus	181	0	2	2	0	0	0	0	1
2	Homo sapiens	236	0	2	2	0	0	0	0	0
3	Homo sapiens	180	0	2	2	0	0	0	0	0
4	Mus musculus	180	0	0	0	1	4	1	1	8
5	Rattus norvegicus	180	0	0	0	1	4	1	1	8
6	Rattus norvegicus	197	0	0	0	1	4	1	1	8
7	Rattus norvegicus	191	0	0	0	1	4	1	1	8
8	Ovis aries	179	0	2	2	2	3	2	2	6
9	Ovis aries	235	0	2	2	2	3	2	7	6
10	Ovis aries	179	0	2	2	2	3	2	2	6
11	Ovis aries	179	0	2	2	2	3	2	2	6
12	Bos taurus	178	0	2	2	4	0	4	4	4
13	Bos taurus	179	0	2	2	4	0	4	4	4

ID	Obyek	Len	Label cluster saat N=							
			2	3	4	5	6	7	8	9
14	Bos taurus	235	0	2	2	4	0	4	7	4
15	Bos taurus	179	0	2	2	4	0	4	4	4
16	Sus scrofa	213	0	2	2	0	5	0	0	5
17	Sus scrofa	237	0	2	2	0	5	0	0	5
18	Sus scrofa	184	0	2	2	0	5	0	0	5
19	Sus scrofa	181	0	2	2	0	5	0	0	5
20	Sus scrofa	181	0	2	2	0	5	0	0	5
21	Gallus gallus	226	1	1	1	3	1	3	3	4
22	Gallus gallus	187	1	1	1	3	1	3	3	4
23	Danio rerio	197	0	2	2	0	0	6	6	3
24	Capra hircus	179	0	2	2	2	3	2	2	6
25	Bubalus bubalis	235	0	2	2	4	0	4	7	4
26	Canis lupus familiaris	184	0	2	3	0	2	5	5	2
27	Canis lupus familiaris	238	0	2	3	0	2	5	5	2
28	Canis lupus familiaris	304	0	2	3	0	2	5	5	5
29	Canis lupus familiaris	185	0	2	3	0	2	5	5	2
30	Canis lupus familiaris	280	0	2	3	0	2	5	5	5
31	Equus caballus	181	0	2	2	0	0	0	0	7
32	Pan troglodytes	180	0	2	2	0	0	0	0	0
33	Macaca mulatta	244	0	2	2	0	0	0	0	0
34	Rattus norvegicus	180	0	0	0	1	4	1	1	8

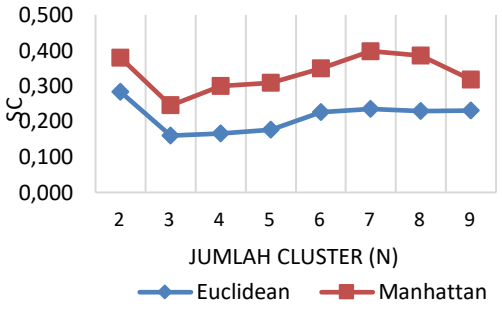
Tabel 2 menyajikan data pengujian yang telah dilakukan dalam penelitian saat parameter jumlah cluster diset dalam berbagai nilai. Pada Tabel 2, ID merupakan pembeda antara string protein yang satu dengan yang lainnya. Hal ini dapat dibuktikan dari panjang sekuen (Len.) yang berbeda walaupun memiliki kategori obyek yang sama. Dalam tabel tersebut telah dipetakan hasil pelabelan cluster pada tiap-tiap data saat dilakukan pengaturan nilai parameter N yang berbeda. Saat proses clustering diset pada N=2, semua string protein hewan dikelompokkan dalam cluster yang sama kecuali gallus gallus (ayam). Walaupun clustering merupakan sebuah proses unsupervised learning, dimana tidak diperlukan label dalam dataset yang diolah, namun dapat disaksikan bahwa saat n=2, proses clustering belum dapat mengelompokkan string protein apabila dilihat dari label objek yang diberikan. Pada saat n=3, ayam tetap dikelompokkan dalam cluster tersendiri dan berhasil membentuk sebuah cluster baru yaitu cluster Rattus norvegicus dan Mus musculus (tikus), sementara sekuens yang lainnya disimpan dalam cluster yang sama, dimana cluster ini masih bersifat sangat luas dan beragam. Saat n=4, cluster yang dihasilkan cenderung sama, namun berhasil membentuk cluster baru yaitu cluster Canis lupus familiaris (anjing).

Berdasarkan penjelasan di atas, dapat dicermati bahwa clustering dengan metode K-Means mampu menghasilkan hasil cluster yang cukup stabil, dalam hal kemampuan metode untuk memberikan label yang sama pada data yang mirip. Sebagai contoh data Bos taurus (sapi) dengan empat sampel sekuens mampu diidentifikasi dalam label cluster yang sama untuk hampir semua nilai N, kecuali pada N=7. Akan tetapi pada sampel sekuens Sus scrofa (babi), kelima sampel mampu diidentifikasi dalam label yang sama untuk semua nilai N. Dengan demikian dapat dikatakan bahwa metode K-Means clustering cukup reliabel dalam melakukan proses clustering.

Pada saat n=5, algoritma berhasil terbentuk dua cluster baru dari cluster n=3, yaitu cluster Ovis aries-Capra hircus (domba-kambing) dan cluster Bos taurus- Bubalus bubalis (sapi-kerbau). Akan tetapi saat n=5, cluster anjing kembali berbaur dengan cluster besar. Saat n=6, cluster sapi-kerbau menghilang, namun cluster anjing berhasil dimunculkan kembali dan terbentuk sebuah cluster baru yaitu Sus scrofa (babi). Saat n=7, cluster babi kembali menyatu dengan cluster utama walaupun berhasil membentuk kembali cluster sapi-kerbau dan berhasil membuat cluster baru yaitu Danio rerio (zebra fish). Sampai n=7, clustering berhasil membentuk cluster yang cukup efektif dan mendeteksi suatu pencilan kemudian dikelompokkan ke dalam cluster yang baru. Akan tetapi saat n=8, berhasil terbentuk suatu cluster baru yang merupakan campuran antara satu protein sapi, satu protein kambing, dan satu protein domba. Hal ini terjadi karena clustering mempertimbangkan juga panjang karakter sekuens, sehingga mengelompokkan panjang yang serupa menjadi cluster yang sama, walaupun sebetulnya merupakan obyek yang berbeda. Hal ini juga terjadi saat n=9, walaupun saat n=9 berhasil membentuk cluster Oryctolagus cuniculus (kelinci).

Dalam melakukan evaluasi hasil cluster, terdapat beberapa metode pengukuran yang digunakan. Pengukuran yang pertama adalah dengan *silhouette coefficient*. *Silhouette*

coeficient hasil cluster akan ditampilkan pada Gambar 1.

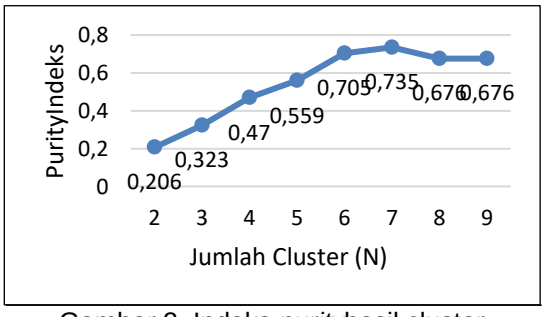


Gambar 1. *Silhouette coeficient* hasil cluster

Dalam menghitung *silhouette coeficient*, terdapat dua metode pengukuran yang digunakan untuk melihat memberikan pandangan yang berbeda terkait penafsiran hasil cluster. Euclidean *distance* mengukur jarak antara dua vektor seakan menarik garis lurus dan dihitung menggunakan rumus pitagoras. Sementara Manhattan *distance* menghitung jumlah selisih jarak untuk setiap fitur/ dimensi data yang diperhitungkan.

Pada Gambar 1, Manhattan selalu menghasilkan nilai yang lebih besar dari Euclidean karena pada Manhattan menghitung semua selisih setiap dimensi antara kedua vektor, sehingga sangat mungkin menghasilkan nilai yang lebih besar dibandingkan dengan Euclidean. Akan tetapi baik Euclidean dan Manhattan selalu menghasilkan nilai *silhouette coeficient* yang positif, yang berarti hasil cluster tidak menunjukkan tendensi overlap maupun kesalahan dalam mengelompokkan anggota kedalam cluster yang salah. Berdasarkan perhitungan *silhouette coeficient*, cluster terbaik didapatkan saat n=2, dengan rata-rata *silhouette coeficient* sebesar 0.331. Hasil cluster terbaik kedua diperoleh saat n=7 dengan rata-rata *silhouette coeficient* sebesar 0,317.

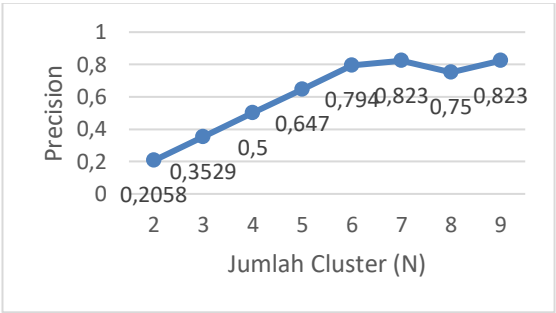
Selain *silhouette coeficient*, indeks *purity*juga digunakan untuk mengevaluasi hasil cluster. Indeks *purity*akan ditampilkan pada Gambar 2.



Gambar 2. Indeks *purity*hasil cluster

Berdasarkan Gambar 2, Indeks *purity* cenderung meningkat seiring dengan bertambahnya jumlah cluster. Hal ini disebabkan karena dengan semakin bertambahnya jumlah cluster berpotensi untuk mengelompokkan sekuens yang mirip ke dalam cluster yang berbeda, yang menunjukkan efektifitas hasil cluster. Akan tetapi, terjadi penurunan nilai indeks *purity* saat n=8 dan n=9 yang disebabkan karena cluster yang benar dan sesuai yang telah terbentuk dipecah kembali menjadi cluster yang berbeda sehingga berbaur dengan anggota dari cluster yang lain berdasarkan panjang sekuens. Dengan demikian indeks *purity* terbaik diperoleh saat n=7.

Terdapat dua pengukuran lainnya yang digunakan dalam mengevaluasi hasil cluster yang terbentuk, yaitu dengan menggunakan *precision*. Hasil *precision* akan disajikan pada Gambar 3.



Gambar 3. Nilai *precision* hasil cluster

Berbeda dengan *purity* yang hanya menghitung berapa banyak data yang sesuai dengan clusternya, *precision* menghitung juga berapa banyak data yang tidak seharusnya berada pada suatu cluster. Berdasarkan nilai *precision* yang ada pada Gambar 4, dapat diketahui bahwa cluster terbaik dihasilkan saat n=7 dan n=9. Saat n=9 memiliki nilai cluster yang baik karena mampu mendeteksi data pencilan dalam suatu cluster dan dipisahkan dalam cluster tersendiri.

Berdasarkan beberapa metode evaluasi hasil cluster, maka dapat disimpulkan bahwa cluster terbaik terbentuk saat n=7, dengan keanggotaan cluster yang akan ditampilkan pada Tabel 3.

Tabel 3. Cluster Terbaik

ID Cluster	Anggota Cluster
0	*Oryctolagus cuniculus* (kelinci eropa)
	Homo sapiens (manusia)
	Sus scrofa (babi)
	Equus caballus (kuda)
	Pan troglodytes (simpanse)
	Macaca mulatta (monyet resus)
1	Rattus norvegicus (tikus cokelat)
	Mus musculus (tikus hitam)

ID Cluster	Anggota Cluster
2	Ovis aries (domba)
	Capra hircus (kambing)
3	Gallus gallus (ayam)
4	Bos taurus (sapi)
	Bubalus bubalis (kerbau)
5	Canis lupus familiaris (anjing)
6	Danio rerio (ikan zebra)

Setelah cluster terbaik ditemukan, maka tingkat kedekatan antar anggota cluster dapat dihitung dengan menggunakan cosine similarity. Cosine similarity digunakan untuk menghitung derajat kemiripan dua buah vektor. Pada ID Cluster 1, Rattus norvegicus dan Mus Musculus memiliki nilai cosine similarity sebesar 0.6092, yang berarti besaran sudut kedua vektor sekuens sebesar 52.4 derajat. Pada ID Cluster 2, Ovis aries dan Capra hircus memiliki kedekatan antara 0.439 – 0.686, yang berarti memiliki kedekatan sebesar 46.7 – 63.9 derajat. Pada ID Cluster 4, Bos taurus dan Bubalus bubalis memiliki kedekatan antara 0.450-0.688, yang berarti memiliki besaran sudut yang berkisar antara 46,5 – 63.2 derajat. Pada ID Cluster 0, tingkat kedekatan antar spesies anggotanya akan dijabarkan pada Tabel 4.

Tabel 4 Tingkat Kedekatan Spesies pada ID Cluster 0

Obyek	Kelinci Eropa	Manusia	Babi	Kuda	Simpanse	Monyet Resus
Kelinci Eropa	1	0.009 – 0.015	0.032 – 0.061	0.077	0.015	0.008
Manusia	0.009 – 0.015	1	0.025 – 0.065	0.042 – 0.071	0.59	0.166 – 0.268
Babi	0.032 – 0.061	0.025 – 0.065	1	0.010 – 0.017	0.037 – 0.065	0.043 – 0.076
Kuda	0.077	0.042 – 0.071	0.010 – 0.017	1	0.072	0.038
Simpanse	0.015	0.59	0.037 – 0.065	0.072	1	0.268
Monyet Resus	0.008	0.166 – 0.268	0.043 – 0.076	0.038	0.268	1

Berdasarkan Tabel 4, sekuens yang memiliki jarak paling dekat adalah manusia dan simpanse dengan cosine similarity sebesar 0.59 atau sudut vektor sebesar 53.8 derajat. Dengan demikian dapat dikatakan bahwa ID Cluster 0 masih memiliki anggota cluster yang belum sesuai.

4. Kesimpulan

Algoritma K-Means dapat diterapkan dalam melakukan pengelompokan sekuens protein IGF2 yang telah diolah menggunakan n-gram dengan cukup efektif. Hasil clustering terbaik diperoleh saat N=7 yang menghasilkan nilai *silhouette coeficient* rata-rata sebesar 0.331, indeks *purity* sebesar 0.735, dan *precision*sebesar 0.823 yang mengindikasikan proses clustering cukup efektif.

Berdasarkan hasil clustering yang diperoleh dari pengelompokan sekuens protein hewan, didapati bahwa manusia memiliki struktur yang mirip dengan kelinci eropa, babi, kuda, simpanse, dan monyet resus. Sementara tikus cokelat memiliki kedekatan dengan tikus hitam; domba memiliki kedekatan struktur dengan kambing; sapi memiliki kedekatan dengan kerbau. Penelitian ini merupakan penelitian pertama yang menggunakan TF-IDF dan algoritma K-Means untuk mengukur kedekatan antar spesies makhluk hidup. Penelitian selanjutnya dapat dilakukan dengan memperkaya jumlah dataset yang ada sehingga dapat diukur sejauh mana efektifitas algoritma clustering dapat mengelompokkan sekuens protein secara tepat.

Referensi

Adhe, D., Rachman, C., Goejantoro, R., & Tisna, D. (2020). Implementation Of Text Mining For Grouping Thesis Documents Using K-Means Clustering. *Jurnal EKSPONENSIAL*, *11*(2), 167–174.

Baral, K., & Rotwein, P. (2019). The insulin-like growth factor 2 gene in mammals: Organizational complexity within a conserved locus. *PLoS ONE*, *14*(6), 1–23. https://doi.org/10.1371/journal.pone.0219155

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., da Silva, A., Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., … Zhang, J. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Christian, H., Seno, D., Pratama, B., & Putri, A. G. (2022). *Analisis Hubungan Kekerabatan Fenetik Varietas Portulaca oleracea dan Portulaca grandiflora di*

Desa Grogol Kelurahan Dukuh Kota Salatiga. *11*(1), 6–11.

Criado-Mesas, L., Ballester, M., Crespo-Piazuelo, D., Castelló, A., Benítez, R., Fernández, A. I., & Folch, J. M. (2019). Analysis of porcine IGF2 gene expression in adipose tissue and its effect on fatty acid composition. *PLOS ONE*, *14*(8), 1–18. https://doi.org/10.1371/journal.pone.0220708

Dwi, P., Prasetya, A., Ari, I., Zaeni, E., & Nafalski, A. (2019). *Journal Classification Using Cosine Similarity Method on Title and Abstract with Frequency-Based Stopword Removal*. *3*(3). https://doi.org/10.29099/ijair.v3i2.99

Febrita, R. E., & Amaniyah, M. (2021). Seminar Nasional Terapan Riset Inovatif (SENTRINOV) Ke-6. *Jurnal Seminar Nasional Terapan Riset Inovatif (SENTRINOVE)*, *7*(1), 260–267.

Febrita, R. E., Mahmudy, W. F., & Wibawa, A. P. (2019). High Dimensional Data Clustering using Self-Organized Map. *Knowledge Engineering and Data Science*, *2*(1), 31. https://doi.org/10.17977/um018v2i12019p31-40

Haviluddin, H., Patandianan, S. J., Putra, G. M., Puspitasari, N., & Pakpahan, H. S. (2021). Implementasi Metode K-Means Untuk Pengelompokkan Rekomendasi Tugas Akhir. *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, *16*(1), 13. https://doi.org/10.30872/jim.v16i1.5182

Irsyad, H., & Pribadi, M. R. (2020). Implementasi Text Mining Dalam Pengelompokan Data Tweet Pertanian Indonesia Dengan K-Means. *KURAWAL Jurnal Teknologi, Informasi Dan Industri*, *3*(2), 164–172. https://t.co/FXtzMcbdHp

James, B. T., Luczak, B. B., & Girgis, H. Z. (2018). MeShClust: an intelligent tool for clustering DNA sequences. *Nucleic Acids Research*, *46*(14), E83. https://doi.org/10.1093/nar/gky315

Riandini, E., & Astuti, R. R. S. (2020). *Hubungan Kekerabatan Fenetik Pisang di Kecamatan Kabawetan , Kabupaten*. *3*(2), 111–117.

Wei, C., Wu, M., Wang, C., Liu, R., Zhao, H., Yang, L., Liu, J., Wang, Y., Zhang, S., Yuan, Z., Liu, Z., Hu, S., Chu, M., Wang, X., & Du, L. (2018). Long Noncoding RNA Lnc-SEMT Modulates IGF2 Expression by Sponging miR-125b to Promote Sheep Muscle Development and Growth. *Cellular Physiology and Biochemistry*, *49*(2), 447–462. https://doi.org/10.1159/000492979

Xiang, G., Ren, J., Tang, H., Fu, R., Yu, D., Wang, J., Li, W., Wang, H., & Wan, H. (2018). Editing porcine IGF2 regulatory element improved meat production in Chinese Bama pigs. *Cellular and Molecular Life Sciences CMLS*. https://doi.org/10.1007/s00018-018-2917-6

Yudiarta, N. G., Sudarma, M., & Ariastina, W. G. (2018). Penerapan Metode Clustering Text Mining Untuk Pengelompokan Berita Pada Unstructured Textual Data. *Majalah Ilmiah Teknologi Elektro*, *17*(3), 339. https://doi.org/10.24843/mite.2018.v17i03.p06