

Analisis Data Biologis dalam Mengidentifikasi Gen atau Protein yang Memiliki Pola Ekspresi Serupa

Muhammad Haikal Akmal¹, Dimas Pangestu², Dzilhulaifa Siregar³, Khaila Mukti Harahap⁴, Mhd. Furqan⁵

Jurusan Ilmu Komputer, fakultas Sains dan Teknologi, Universitas Islam Negeri Sumatera Utara

¹haikalakmal315@gmail.com, ²dimaspangestu771@gmail.com, ³dzilhulaifasiregar21@gmail.com,

⁴muktiharahap321@gmail.com, ⁵mfurqan@uinsu.ac.id

Abstract

Protein expression data in biological studies are often high-dimensional and complex, making direct pattern recognition a challenge. This study applies the Spectral Clustering algorithm to uncover latent structures within protein expression datasets. The initial phase involves data cleaning through mean imputation for missing values and normalization of numerical features using StandardScaler. The dataset comprises 1,080 observations with 77 numerical attributes derived from laboratory experiments on mice. Clustering is performed using a graph-based approach, configured with four clusters and nearest neighbors affinity. To enhance interpretability, dimensionality reduction is conducted using Principal Component Analysis (PCA), resulting in a two-dimensional visual representation. The clustering outcome demonstrates clear group separations that reflect biological variation among samples. These findings suggest that unsupervised learning methods like Spectral Clustering are effective in revealing hidden patterns in protein expression data and provide a foundation for subsequent classification based on biological features.

Keywords: *unsupervised learning, PCA, protein expression*

Abstrak

Ekspresi protein dalam data biologis umumnya memiliki kompleksitas tinggi dan dimensi besar, sehingga menyulitkan pengenalan pola secara langsung. Studi ini memanfaatkan algoritma Spectral Clustering untuk mengeksplorasi struktur tersembunyi dalam kumpulan data ekspresi protein. Langkah awal mencakup pembersihan data dengan imputasi nilai hilang menggunakan metode rata-rata serta normalisasi fitur numerik menggunakan StandardScaler. Dataset terdiri dari 1.080 observasi dan 77 atribut numerik hasil percobaan pada tikus. Proses pengelompokan dilakukan dengan pendekatan berbasis graf, menggunakan parameter empat kluster dan afinitas nearest neighbors. Selanjutnya, dilakukan reduksi dimensi melalui teknik Principal Component Analysis (PCA) untuk menghasilkan representasi dua dimensi yang mudah divisualisasikan. Hasil pengelompokan memperlihatkan pemisahan yang mencerminkan perbedaan biologis antar sampel. Hal ini menunjukkan bahwa metode tak terawasi seperti Spectral Clustering efektif dalam mengungkap struktur laten pada data ekspresi protein dan dapat menjadi dasar bagi analisis klasifikasi berbasis karakteristik biologis.

Kata kunci: *unsupervised learning, PCA, ekspresi protein*

© 2025 Author
Creative Commons Attribution 4.0 International License



1. Pendahuluan

Di dalam tubuh makhluk hidup, gen dan protein bekerja secara terkoordinasi membentuk sistem biologis yang kompleks. Salah satu cara untuk memahami bagaimana sistem ini berfungsi adalah dengan menganalisis pola ekspresi gen atau protein. Gen-gen atau protein-protein tertentu sering kali menunjukkan pola ekspresi yang serupa dalam kondisi biologis tertentu, seperti tahap perkembangan, respons terhadap stres, atau keberadaan penyakit. Pola-pola ini bukan sekadar kebetulan—mereka dapat mengindikasikan bahwa gen atau protein tersebut memiliki fungsi yang saling terkait atau bahkan terlibat dalam jalur biologis yang sama [1].

Dengan kemajuan teknologi sekuensing dan proteomik, kini kita dapat mengamati ekspresi ribuan gen atau protein secara bersamaan dalam satu eksperimen. Tantangannya terletak pada bagaimana menyaring dan mengelompokkan data dalam jumlah besar ini untuk menemukan hubungan-hubungan yang bermakna. Salah satu pendekatan yang digunakan adalah mengidentifikasi kelompok gen atau protein yang menunjukkan pola ekspresi yang serupa. Analisis ini tidak hanya membantu memahami fungsi gen secara lebih mendalam, tetapi juga membuka peluang untuk menemukan biomarker baru atau target terapi potensial bagi berbagai penyakit [4].

Pemahaman terhadap pola ekspresi protein menjadi elemen krusial dalam kajian biomedis, khususnya dalam upaya mengidentifikasi mekanisme molekuler yang mendasari berbagai kondisi fisiologis dan patologis. Pendekatan ini diterapkan secara menyeluruh pada data ekspresi protein biologis dari sampel tikus, dengan mempertimbangkan proses pra-pemrosesan yang ketat mulai dari penanganan data hilang hingga normalisasi skala. Teknik ini belum banyak dikombinasikan secara sistematis dengan strategi visualisasi berbasis PCA untuk menggambarkan distribusi antar kelompok secara dua dimensi, khususnya dalam konteks analisis data protein berdimensi tinggi [2].

Ekspresi protein sendiri mencerminkan aktivitas biologis dari suatu gen dalam bentuk produk protein yang berperan penting dalam fungsi seluler. Data ekspresi protein biasanya disajikan dalam bentuk matriks berdimensi tinggi, di mana setiap baris merepresentasikan sampel biologis, dan kolom-kolom menunjukkan tingkat ekspresi dari masing-masing protein. Kompleksitas dan dimensi tinggi dari data ini menuntut pendekatan analisis yang mampu menangkap pola non-linier yang tersembunyi di dalamnya.

Salah satu pendekatan yang relevan untuk mengatasi tantangan tersebut adalah teknik klasterisasi, yakni metode pengelompokan data berdasarkan kemiripan karakteristik tanpa memerlukan label awal. Teknik

ini telah banyak diterapkan dalam konteks biomedis untuk klasifikasi sel, jaringan, maupun ekspresi gen dan protein. Namun, metode klasterisasi konvensional seperti K-Means sering kali tidak optimal dalam mengenali struktur data yang kompleks dan tidak linier [3].

Untuk itu, penelitian ini mengusulkan penggunaan algoritma Spectral Clustering yang berbasis teori graf, karena kemampuannya dalam mendeteksi relasi non-linier antar data. Algoritma ini bekerja dengan membentuk graf kemiripan antar sampel, menghitung matriks Laplacian, dan mengekstrak informasi melalui eigenstruktur untuk memproyeksikan data ke ruang dimensi rendah sebelum dilakukan proses klasterisasi. Pendekatan ini terbukti efektif dalam mengungkap klaster tersembunyi dalam data berdimensi tinggi yang sulit dijangkau oleh metode lain [5].

Agar hasil analisis lebih mudah ditafsirkan secara visual, penelitian ini juga menerapkan Principal Component Analysis (PCA), yaitu teknik reduksi dimensi yang merangkum variansi terbesar dalam data ke dalam beberapa komponen utama. Dengan transformasi ini, distribusi klaster yang diperoleh dari Spectral Clustering dapat divisualisasikan secara dua dimensi, sehingga memudahkan dalam mengevaluasi dan memahami struktur kelompok dalam data ekspresi protein [6].

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksploratif untuk mengelompokkan data ekspresi protein menggunakan algoritma Spectral Clustering. Dataset yang digunakan yang terdiri dari 1080 baris data dan 77 fitur ekspresi protein bertipe numerik (float64). Data ini berasal dari sampel tikus hasil eksperimen biologis yang sudah tersedia secara publik.

2.1. Pra-pemrosesan Data

Pra-pemrosesan data dilakukan dengan mengekstraksi seluruh fitur numerik dari dataset dan menangani nilai yang hilang (missing values) menggunakan metode imputasi rata-rata (mean imputation) melalui SimpleImputer dari pustaka scikit-learn. Dataset ini terdiri dari 1.080 baris data biologis yang mencakup ekspresi dari 77 protein, dengan tambahan tiga label kategorikal yaitu kelas sampel (class), genotipe (Genotype), dan perlakuan (Treatment). Sebelum dilakukan analisis lebih lanjut, seluruh fitur numerik dinormalisasi menggunakan StandardScaler agar semua variabel berada dalam skala yang sebanding dan tidak mendominasi proses klasterisasi.

Pra- pemrosesan

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
```

```
# Membaca data dan memilih fitur numerik
df = pd.read_csv('dataset_protein.csv')
expr_data = df.select_dtypes(include=['float64'])

# Imputasi nilai hilang
imputer = SimpleImputer(strategy='mean')
expr_imputed = imputer.fit_transform(expr_data)

# Normalisasi
scaler = StandardScaler()
expr_scaled = scaler.fit_transform(expr_imputed)
```

2.2. Teknik Klasterisasi

Teknik klasterisasi yang digunakan dalam penelitian ini adalah Spectral Clustering, yang dipilih karena kemampuannya dalam mengidentifikasi kluster non-linier yang tersembunyi. Algoritma ini dijalankan dengan parameter `n_clusters=4`, `affinity='nearest_neighbors'`, dan `random_state=42`, di mana algoritma membentuk graf kedekatan antar titik data, menghitung matriks Laplacian, lalu mengekstrak eigenvector-nya untuk memetakan data ke ruang berdimensi lebih rendah sebelum proses pengelompokan dilakukan.

Teknik Klasterisasi

```
from sklearn.cluster import SpectralClustering

# Menjalankan Spectral Clustering
spectral = SpectralClustering(n_clusters=4,
                              affinity='nearest_neighbors', random_state=42)
clusters = spectral.fit_predict(expr_scaled)

# Menyimpan hasil kluster ke DataFrame
df['Cluster'] = clusters }
```

2.3. Reduksi Dimensi dan Visualisasi

Untuk meningkatkan interpretabilitas hasil klasterisasi, diterapkan teknik reduksi dimensi Principal Component Analysis (PCA), yang mengubah data berdimensi tinggi menjadi dua komponen utama [7]. Hasil transformasi ini divisualisasikan dalam bentuk plot sebar dua dimensi (scatter plot), di mana masing-masing titik diberi warna berdasarkan label kluster hasil Spectral Clustering, sehingga memudahkan dalam mengevaluasi pola distribusi antar kelompok.

Reduksi Dimensi dan Visualisasi

```
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns

# PCA dan visualisasi
pca = PCA(n_components=2)
pca_result = pca.fit_transform(expr_scaled)

plt.figure(figsize=(10, 6))
sns.scatterplot(x=pca_result[:, 0], y=pca_result[:, 1],
               hue=clusters, palette='tab10')
plt.title('Clustering Protein Expression Data (Spectral Clustering + PCA)')
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.legend(title='Cluster')
```

```
plt.grid(True)
plt.tight_layout()
plt.show()
```

2.4. Penyimpanan output

Hasil akhir disimpan dalam bentuk file CSV agar dapat dianalisis lebih lanjut.

Reduksi Dimensi dan Visualisasi

```
# Menyimpan hasil ke file
df.to_csv('clustered_protein_expression_spectral.csv',
         index=False)
```

2.5. Replikasi dan Volume Data

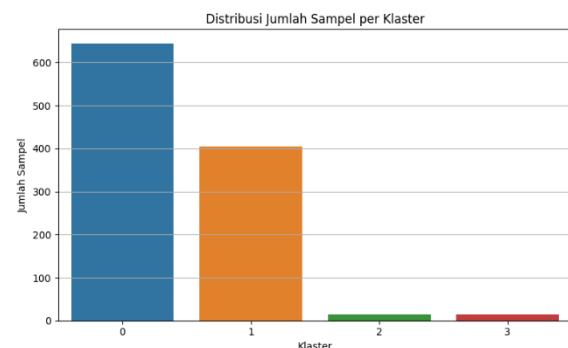
Replikasi eksperimental tidak dilakukan karena dataset telah tersedia secara lengkap dan tidak memerlukan eksperimen tambahan. Semua proses analisis dilakukan menggunakan Python 3.10 dengan dukungan pustaka pandas, numpy, scikit-learn, matplotlib, dan seaborn.

3. Hasil dan Pembahasan

3.1. Visualisasi Distribusi Kluster

Berdasarkan hasil pemrosesan menggunakan algoritma Spectral Clustering, diperoleh empat kluster utama. Visualisasi distribusi jumlah sampel pada setiap kluster ditampilkan dalam bentuk diagram batang. Kluster 0 merupakan kelompok dengan jumlah sampel terbanyak, diikuti oleh kluster 1, sementara kluster 2 dan 3 memiliki jumlah sampel yang jauh lebih sedikit. Ketimpangan distribusi ini dapat mengindikasikan adanya variasi signifikan dalam ekspresi protein antar kelompok, serta potensi adanya outlier atau subgrup khusus dalam data.

Distribusi ini penting untuk diperhatikan dalam interpretasi hasil klasterisasi, karena ketidakseimbangan dapat memengaruhi validitas kesimpulan biologis yang diambil, khususnya jika akan dilakukan analisis lanjutan seperti identifikasi gen target atau eksplorasi biomarker.

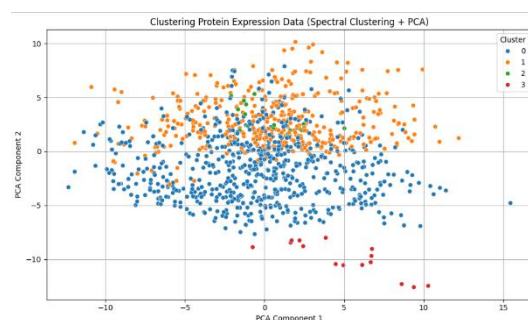


Gambar 1 Grafik distribusi jumlah sampel

3.2. Visualisasi Klaster dalam Ruang PCA

Hasil klasterisasi dalam ruang dua dimensi menggunakan PCA (Principal Component Analysis) untuk reduksi dimensi. Setiap titik mewakili sampel dan diwarnai berdasarkan label klaster hasil dari Spectral Clustering. Terlihat bahwa meskipun terdapat beberapa area tumpang tindih, masing-masing klaster menunjukkan pola distribusi yang relatif unik.

Pemisahan ini memberikan gambaran awal bahwa struktur internal data ekspresi protein memang memiliki kecenderungan kelompok-kelompok tertentu. Hal ini dapat diasumsikan berhubungan dengan perlakuan berbeda yang diterapkan terhadap masing-masing kelompok tikus percobaan, seperti perbedaan kondisi biologis atau lingkungan.



Gambar 2 Klasterisasi ekspresi protein dengan PCA.

3.3. Analisis dan Interpretasi Hasil Klasterisasi

Pendekatan Spectral Clustering terbukti efektif dalam menangkap struktur non-linier dalam data. Tidak seperti K-Means yang cenderung membentuk klaster berbentuk bulat (isotropik), Spectral Clustering lebih fleksibel dalam mengidentifikasi pola kompleks yang tersembunyi dalam distribusi data biologis. Ini sangat relevan untuk data ekspresi protein, yang umumnya memiliki variasi tinggi dan noise [8].

Dengan bantuan PCA, visualisasi hasil klasterisasi menjadi lebih intuitif dan dapat dievaluasi secara visual [9]. Kombinasi ini membuka peluang untuk digunakan sebagai langkah eksploratif awal sebelum melangkah ke analisis yang lebih mendalam, seperti identifikasi fitur penting (feature selection), korelasi antar protein, atau integrasi dengan data biologis lainnya [10].

4. Kesimpulan

Penelitian ini bertujuan untuk mengeksplorasi pola dalam data ekspresi protein menggunakan metode Spectral Clustering yang dikombinasikan dengan reduksi dimensi PCA. Proses analisis mencakup tahap pra-pemrosesan data, pembentukan klaster tanpa label (unsupervised clustering), serta visualisasi hasil dalam ruang dua dimensi. Hasil pemodelan menunjukkan terbentuknya empat kelompok utama dengan karakteristik distribusi yang berbeda-beda.

Visualisasi klaster mengindikasikan adanya struktur internal dalam data yang relevan secara biologis, meskipun terdapat area tumpang tindih antar klaster. Pendekatan ini menunjukkan potensi dalam membantu identifikasi hubungan tersembunyi antar fitur, khususnya dalam konteks data kompleks seperti ekspresi protein.

Daftar Rujukan

- [1] M. Mafruchati, *Eksresi Gen dan Teknik Biologi Molekuler untuk kajian veteriner*. Sidoarjo: Zifatama Jawara, 2023.
- [2] S. P. Nabila, N. Ulinuha, and A. Yusuf, "Model Prediksi Kelulusan Tepat Waktu Dengan Metode Fuzzy C-Means Dan K-Nearest Neighbors Menggunakan Data Registrasi Mahasiswa," *Netw. Eng. Res. Oper.*, vol. 6, no. 1, pp. 39–47, 2021, doi: 10.21107/nero.v6i1.199.
- [3] S. A. B. Telaumbanua, F. Setiadi, and S. Nurjanah, "Analisis Clustering Menggunakan Metode Enhanced Fuzzy C-Means Clustering Dengan Algoritma Rock Pada Student Performance Dataset," *Bin. Digit. -Technology*, vol. 7, no. 3, pp. 984–994, 2025, doi: 10.32877/bt.v7i3.2287.
- [4] R. R. Ananda, Arnita, R. Maharani, M. A. I. Lubis, S. P. Sutompul, and Hasian, "ANALISIS HARGA RUMAH DI DAERAH JAKARTA SELATAN DENGAN MENGGUNAKAN ALGORITMA SPECTRAL CLUSTERING," *J. Ilmu Komput. Revolusioner*, vol. 08, no. 10, pp. 65–72, 2024.
- [5] A. Syarif and A. Ramadhanu, "BERDASARKAN BENTUK WARNA MENGGUNAKAN METODE K-NEAREST NEIGHBO," vol. 7, no. 4, pp. 1578–1583, 2024.
- [6] D. Hedyati and I. M. Suartana, "Penerapan Principal Component Analysis(PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro," *J. Inf. Eng. Educ. Technol.*, vol. 05, no. 02, pp. 49–54, 2021.
- [7] R. H. Ngovangari, G. C. Rorimpandey, O. Kembuan, and H. M. Sumual, "Implementasi Algoritma Needleman-Wunsch dalam Pengujian Tingkat Kemiripan DNA Babi Sulawesi Utara," *JOINTER J. Informatics Eng.*, vol. 5, no. 01, pp. 20–26, 2024, doi: 10.53682/jointer.v5i01.32.
- [8] Y. Ariani, "LITERATURE REVIEW INFORMATION EXTRACTION PADA TEKS BIOMEDICAL DENGAN PENDEKATAN MACHINE LEARNING DAN NAMED ENTITY RECOGNITION," vol. 10, no. 1, pp. 1–12, 2023.
- [9] M. H. F. Dian Atia Ihsani, Achmad Arifin, "Klasifikasi DNA Microarray Menggunakan Principal Component Analysis (PCA) dan Artificial Neural Network (ANN)," *Tek. ITS*, 2020.
- [10] R. B. Ishlahiyah Nur Rizky, Rosa Prahasti, NatriaSelina, "Penggunaan Algoritma Komputasi untuk Analisis Sederhana Data DNA dalam Studi Bioinformatika," pp. 1–11, 2025.