JOURNAL OF COMPUTER SCIENCE AND INFORMATICS ENGINEERING
VOL. 05, NO. 1, JANUARY 2026: 98-106
Doi: https://doi.org/10.55537/cosie.v5i1.1401

e-ISSN 2827-8356
p-ISSN 2827-9999

Klasifikasi Gen rbcL pada Tumbuhan Berbasis Representasi K-mer Protein dan Random Forest

Classification of rbcL Gene in Plants Based on Protein K-mer Representation and Random Forest

Armansyah
[1]Program Studi Ilmu Teknik, Fakultas Teknik, Universitas Sriwijaya
03013682530007@student.unsri.ac.id

Received: November 23, 2025 | Revised: December 3, 2025 | Accepted: January 31, 2026

Abstrak

Gen *ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit* (rbcL) merupakan penanda molekuler yang umum digunakan dalam penelitian filogenetika, DNA barcoding, dan taksonomi tumbuhan. Identifikasi gen ini secara otomatis dari kumpulan sekuens protein skala besar merupakan tantangan bioinformatika yang menarik, terutama dengan meningkatnya jumlah data genom dan proteom yang tersedia secara publik. Penelitian ini bertujuan mengembangkan model klasifikasi berbasis pembelajaran mesin untuk membedakan gen rbcL dari gen plastid lain seperti ndhF, psbA, atpB, dan rbcS. Dataset diperoleh dari database NCBI Protein, kemudian diberi label otomatis berdasarkan metadata FASTA. Representasi numerik sekuens dilakukan menggunakan k-mer protein (k = 2) dengan total 400 fitur per sekuens. Model Random Forest dengan 300 estimator dilatih pada pembagian data 80% latih dan 20% uji. Evaluasi menggunakan *confusion matrix, precision, recall, f1-score*, dan ROC–AUC menunjukkan performa klasifikasi yang tinggi, dengan akurasi ≥ 0.95 dan nilai AUC yang mendekati 1.0. Hasil penelitian ini menunjukkan bahwa pendekatan berbasis k-mer tanpa alignment dapat digunakan secara efektif untuk identifikasi gen rbcL, dan berpotensi diterapkan dalam pipeline anotasi gen tumbuhan berbasis kecerdasan buatan.

Kata kunci: *rbcL, Viridiplantae, k-mer protein, Random Forest, klasifikasi gen, bioinformatika.*

Abstract

The ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit gene (rbcL) is one of the most widely used molecular markers in plant phylogenetics, DNA barcoding, and taxonomy. Automatic identification of this gene from large-scale protein sequence collections represents a compelling bioinformatics challenge, especially with the rapid growth of publicly available genomic and proteomic data. This study aims to develop a machine-learning–based classification model to distinguish rbcL from other plastid genes such as ndhF, psbA, atpB, and rbcS. Protein sequence data were obtained from the NCBI Protein database and automatically labeled using FASTA metadata. Numerical representation of sequences was performed using protein k-mers (k = 2), yielding 400 features per sequence. A Random Forest model with 300 estimators was trained using an 80% training and 20% testing split. Evaluation using confusion matrix, precision, recall, F1-score, and ROC–AUC demonstrated high classification performance, with accuracy ≥ 0.95 and an AUC value approaching 1.0. These findings indicate that an alignment-free k-mer approach can effectively identify rbcL sequences and has strong potential for application in AI-driven gene annotation pipelines for plant plastid genomes.

Keywords: *rbcL, Viridiplantae, protein k-mer, Random Forest, gene classification, bioinformatics.*

e-ISSN 2827-8356
p-ISSN 2827-9999

1. PENDAHULUAN

Gen rbcL (*ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit*) merupakan salah satu marker molekuler paling penting dalam studi evolusi dan identifikasi tumbuhan. Gen ini mengkode subunit besar enzim RuBisCO yang berperan dalam fiksasi karbon, dan telah menjadi gen kloroplas yang paling banyak digunakan dalam DNA barcoding, analisis filogenetika, dan studi biodiversitas tumbuhan[1], [2], [3], [4], [5]. Konsistensi evolusinya membuat rbcL sangat informatif sebagai penanda taksonomi, dibuktikan dengan lebih dari 300.000 sekuens rbcL yang kini tersedia di NCBI per tahun 2024[6].

Seiring berkembangnya teknologi sekuensing, data protein dan genom tumbuhan meningkat secara eksponensial. NCBI mencatat lebih dari 18 juta sekuens protein terarsip, dengan pertumbuhan sekitar 15% setiap tahun, sejalan dengan ekspansi data omics dan proyek genomik global . Peningkatan ini memunculkan tantangan baru: bagaimana mengidentifikasi gen tertentu, seperti rbcL, dalam skala besar secara cepat, mudah, dan akurat. Metode berbasis pencarian sekuens seperti BLAST masih menjadi standar, tetapi BLAST memiliki biaya komputasi tinggi dan tidak efisien untuk data berskala besar [7], [8], [9], [10].

Untuk mengatasi keterbatasan ini, pendekan *alignment-free* mulai mendapat perhatian, terutama metode berbasis k-mer, yang mengubah sekuens biologis menjadi representasi numerik. Pendekatan ini terbukti 50–100 kali lebih cepat dibanding BLAST dan kompatibel dengan algoritma pembelajaran mesin seperti Random Forest[10], [11], [12], [13]. Namun, sebagian besar penelitian berfokus pada mikroorganisme dan genom virus, bukan gen plastid tumbuhan, sehingga penerapannya pada rbcL masih terbatas.

Berdasarkan kajian literatur, terdapat tiga gap penelitian utama, yaitu (1) mayoritas studi masih mengutamakan aplikasi filogenetik, bukan klasifikasi gen plastid; (2) pendekatan k-mer lebih sering diterapkan pada DNA atau RNA, bukan protein; dan, (3) belum banyak studi yang secara eksplisit membedakan rbcL dari gen plastid lain seperti ndhF, psbA, atpB, dan rbcS menggunakan pendekatan *machine learning*.

Untuk menjawab gap tersebut, penelitian ini bertujuan untuk: (1) mengembangkan model klasifikasi berbasis *machine learning* untuk mendeteksi sekuens rbcL dari protein tumbuhan; (2) menggunakan representasi protein k-mer (k = 2) sebagai fitur numerik yang bersifat alignment-free; (3) menerapkan algoritma Random Forest sebagai model klasifikasi utama; dan, (4) mengevaluasi performa model menggunakan metrik akurasi, presisi, recall, F1-score, dan ROC-AUC.

Pemilihan Random Forest didukung oleh berbagai studi yang menunjukkan bahwa algoritma ini stabil, cepat, dan sangat efektif dalam memodelkan data biologis berbasis sekuens, termasuk representasi numerik seperti k-mer. Penelitian Wang et al.[14] memperlihatkan bahwa Random Forest mampu menghasilkan performa klasifikasi yang lebih baik dibandingkan SVM dan Naive Bayes pada prediksi kelas struktur protein, bahkan ketika fitur berdimensi sangat tinggi. Studi Yin et al.[15] juga menunjukkan bahwa Random Forest memberikan akurasi, presisi, dan F1-score tertinggi dibandingkan KNN, SVM, Naive Bayes, maupun model deep learning (CNN dan DNN) dalam prediksi situs ubikuitilasi protein. Selain itu, penelitian Simón et al.[16] menegaskan bahwa Random Forest merupakan pendekatan yang sangat kuat, sensitif, dan robust untuk klasifikasi protein berbasis fitur komposisi dipeptida/tripeptida — suatu bentuk representasi *alignment-free* yang sangat dekat dengan pola k-mer. Konsistensi temuan dari tiga studi ini memperkuat alasan pemilihan Random Forest sebagai model utama dalam klasifikasi sekuens rbcL.

Hasil penelitian ini menunjukkan bahwa model yang dikembangkan mampu mencapai akurasi di atas 0.95, sehingga membuktikan bahwa pendekatan gabungan k-mer dan Random Forest merupakan solusi efisien, cepat, dan terukur dalam proses anotasi gen rbcL di era big data biologis.

e-ISSN 2827-8356
p-ISSN 2827-9999

2. METODE PENELITIAN

Penelitian ini merupakan studi bioinformatika dengan pendekatan *supervised machine learning*, bertujuan mengembangkan model klasifikasi untuk membedakan gen rbcL dari gen plastid lain berdasarkan sekuens protein. Metode yang diterapkan bersifat *alignment-free*, yaitu analisis sekuens tanpa proses penyelarasan (alignment) seperti BLAST, sehingga lebih efisien dan dapat diterapkan pada dataset berskala besar[17], [18], [19], [20]. Representasi sekuens dilakukan menggunakan model protein k-mer, sedangkan algoritma klasifikasi utama yang digunakan adalah Random Forest, karena keakuratannya, stabil, dan mampu menangani data dimensi tinggi.

Data sekuens protein diperoleh dari NCBI Protein Database dialamat (https://www.ncbi.nlm.nih.gov/protein), yang merupakan basis data publik terbesar yang memuat informasi sekuens lintas organisme. Pengambilan data dilakukan menggunakan kueri spesifik seperti rbcL[Gene Name] AND Viridiplantae[Organism], ndhF[Gene Name] AND Viridiplantae[Organism], psbA[Gene Name] AND Viridiplantae[Organism], atpB[Gene Name] AND Viridiplantae[Organism], dan rbcS[Gene Name] AND Viridiplantae[Organism].

Seluruh sekuens diperoleh dalam format FASTA[21] dan terdiri atas ID, deskripsi, dan rangkaian asam amino. Kelima gen plastid tersebut diambil dari kingdom Viridiplantae, dengan rbcL sebagai kelas positif (satu/1), dan empat gen lainnya sebagai kelas negative(nol/0). Rangkuman dataset ditampilkan pada Tabel 1:

Tabel 1. Representasi Dataset Protein Gen Plastid

Kelas Gen	Jumlah Sekuens	Format
rbcL	± 250	Protein FASTA
ndhF	± 200	Protein FASTA
psbA	± 150	Protein FASTA
atpB	± 130	Protein FASTA
rbcS	± 130	Protein FASTA

Ekstraksi fitur dilakukan menggunakan pendekatan k-mer protein, yaitu memecah sekuens menjadi potongan *substring* sepanjang k. Pada penelitian ini, digunakan k = 2 (dipeptida), sehingga jumlah fitur dihitung dengan rumus:

$$F = n^k \quad (1)$$

Dimana:

- F = jumlah total kemungkinan k-mer
- n = jumlah jenis simbol dalam alfabet (untuk protein, n = 20 karena ada 20 asam amino standar)
- k = panjang k-mer (jumlah residu asam amino dalam satu unit k-m

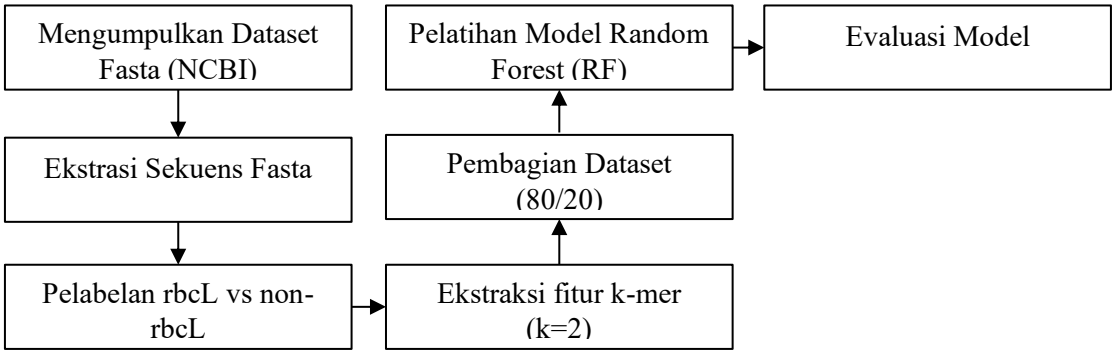
Frekuensi setiap k-mer dihitung dan dinormalisasi menggunakan:

$$f(k) = \frac{jumlah\ kemunculan\ k-mer}{total\ k-mer\ dalam\ sequence} \quad (2)$$

Setelah fitur diekstraksi, dataset dibagi menggunakan *stratified sampling* untuk menjaga keseimbangan proporsi kelas, dengan komposisi 80% data latih dan 20% data uji. Proses klasifikasi dilakukan menggunakan algoritma Random Forest, yang dikenal efektif dalam menangani dataset berdimensi tinggi dan berbasis frekuensi biologis, serta memiliki toleransi yang kuat terhadap *overfitting*. Kinerja model dievaluasi menggunakan metrik standar dalam

e-ISSN 2827-8356
p-ISSN 2827-9999

pembelajaran mesin, termasuk *confusion matrix, accuracy, precision, recall, F1-score*, dan *ROC-AUC*, sehingga memberikan gambaran lengkap mengenai performa prediktif model. Seluruh rangkaian analisis dilakukan menggunakan Python 3.10 di lingkungan Jupyter Notebook, dengan dukungan pustaka Biopython, NumPy, Pandas, scikit-learn versi 1.2 ke atas, dan Matplotlib, yang merupakan perangkat lunak umum dalam analisis bioinformatika berbasis pembelajaran mesin. Diagram alur penelitian ditunjukkan pada Gambar 1:



Gambar 1. Diagram Alur Penelitian

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil analisis data serta performa model klasifikasi yang dibangun menggunakan representasi protein k-mer. Tahap awal analisis dimulai dengan pemeriksaan struktur dataset yang diperoleh dari berkas FASTA dan dikonversi ke format CSV. Dataset terdiri dari empat atribut utama, yaitu *id, sequence, gene*, dan *label*, yang mencerminkan identitas sekuens, rangkaian asam amino, asal gen plastid, dan kategori kelas. Cuplikan dataset ditampilkan pada Tabel 2, yang menunjukkan contoh sekuens dari gen atpB hingga rbcL yang digunakan dalam pemodelan.

Tabel 2. Cuplikan Dataset Hasil Konversi FASTA ke CSV

Index	id	sequence	gene	label
0	sp|P83505.1|ATPB BRACM	KDALVYGQMNEPPGARMRVGLKDG SITSIQAV	atpB	0
1	sp|P81663.1|ATPB M_PINPS	ESIASFQGVLDGKDYD	atpB	0
2	sp|P80083.1|ATPB M_SPIOL	ASSAAAAAAXAPSTPLA	atpB	0
3	XGD01388.1	MRINPTTSRP	atpB	0
4	XGD01383.1	MRINPTTSRP	atpB	0
…	…	…	…	…
43503	CAA54016.1	MSPQTETKASVQFGKAGVKEYKLTY YTPEDYEKTKDTILAARFVRPQ…	rbcL	1
43504	CAA53995.1	MSPQTETKAGVQFGKAGVKEYKLTY YTPEDYEKTKDTILAARFVRPQ…	rbcL	1
43505	CAA54007.1	MSPQTETKAGVQFGKAGVKEYKLTY YTPEDYEKTKDTILAARFVRPQ…	rbcL	1
43506	CAB55342.1	MS	rbcL	1
43507	CAA54002.1	MSPQTETKAGVQFGKAGVKEYKLTY YTPEDYEKTKDTILAARFVRPQ…	rbcL	1

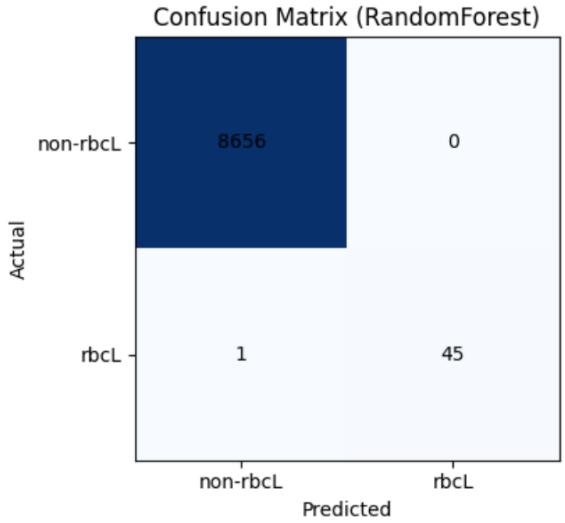
e-ISSN 2827-8356
p-ISSN 2827-9999

Secara keseluruhan, dataset yang digunakan berjumlah 43.508 sekuens protein, dengan 230 sekuens rbcL sebagai kelas (satu/1) dan 43.278 sekuens berasal dari empat gen plastid lainnya sebagai kelas (nol/0). Distribusi ini menjadi dasar dalam pembagian data ke dalam subset pelatihan dan pengujian menggunakan teknik *stratified sampling,* yaitu metode pembagian sampel yang mempertahankan proporsi setiap kelas pada seluruh subset data. Teknik ini secara luas direkomendasikan dalam pembelajaran mesin, terutama ketika jumlah sampel antar-kelas tidak seimbang, karena memastikan bahwa baik data latih maupun data uji tetap mencerminkan distribusi label sebenarnya[22], [23].

Model dilatih menggunakan 80% dari total data dan diuji menggunakan 20% sisanya. Evaluasi performa model dilakukan menggunakan *confusion matrix, classification report*, serta metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil evaluasi ditampilkan pada Tabel 3 dan divisualisasikan pada Gambar 2 dan Gambar 3. *Confusion matrix* menunjukkan bahwa dari 8.702 sekuens uji, model memprediksi seluruh sekuens non-rbcL dengan benar dan mengidentifikasi 45 dari 46 sekuens rbcL dengan tepat.

Tabel 3. Confusion Matrix Model Random Forest

	Prediksi non-rbcL	Prediksi rbcL
Aktual non-rbcL	8656	0
Aktual rbcL	1	45



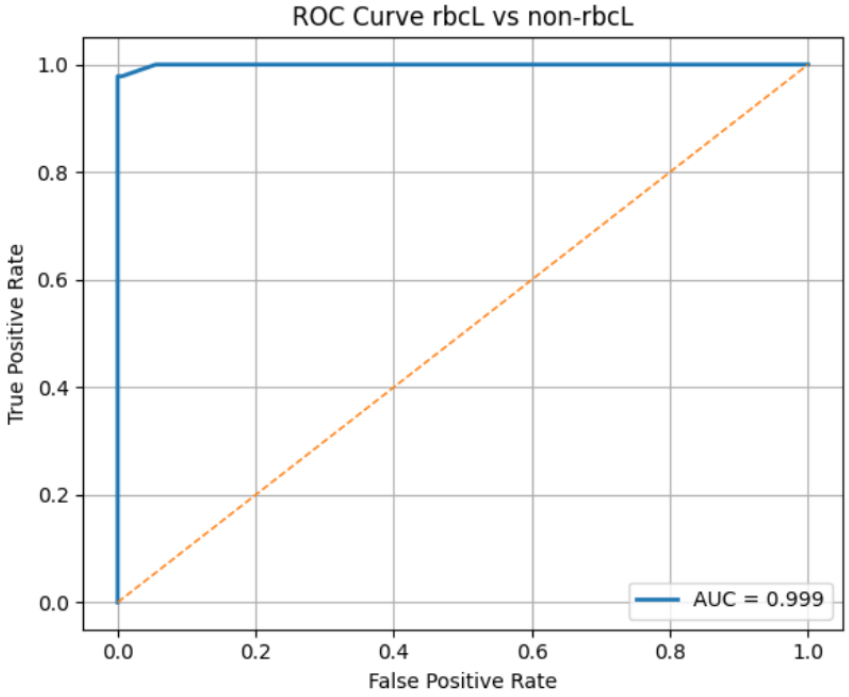
Gambar 2. Klasifikasi Gen Plastid

```
Classification report:
              precision    recall  f1-score   support

    non-rbcL       1.00      1.00      1.00      8656
        rbcL       1.00      0.98      0.99        46

    accuracy                           1.00      8702
   macro avg       1.00      0.99      0.99      8702
weighted avg       1.00      1.00      1.00      8702
```

Gambar 3. Rekapitulasi Hasil Klasifikasi (*classification report*)

e-ISSN 2827-8356
p-ISSN 2827-9999

Kurva ROC hasil evaluasi ditampilkan pada Gambar 4, dengan nilai AUC sebesar 0.999. Kurva menunjukkan pemisahan yang sangat jelas antara kelas positif dan negatif berdasarkan probabilitas prediksi model.



Gambar 4. Kurva ROC dan Nilai AUC

Hasil evaluasi model yang telah disajikan pada Tabel 3 dan Gambar 2 menunjukkan bahwa algoritma Random Forest mampu mempelajari pola k-mer protein dengan sangat baik untuk membedakan sekuens gen rbcL dari empat gen plastid lainnya. Nilai precision dan recall untuk kelas non-rbcL masing-masing sebesar 1.00 menandakan bahwa model tidak menghasilkan false positive, sebuah indikator penting bahwa pola dipeptida pada gen ndhF, psbA, atpB, dan rbcS sangat konsisten dan mudah dipisahkan oleh model. Hal ini juga konsisten dengan struktur dataset pada Tabel 2, yang menunjukkan variasi panjang sekuens dan karakteristik komposisi asam amino yang berbeda antargen.

Untuk kelas rbcL, model mencatat precision sempurna (1.00), menunjukkan bahwa seluruh sekuens yang diprediksi sebagai rbcL tepat sasaran. Namun, recall sebesar 0.98 mengindikasikan bahwa terdapat satu sekuens rbcL yang tidak terklasifikasi dengan benar dan masuk ke kategori non-rbcL. Kasus ini dapat dihubungkan dengan temuan pada dataset awal yang memperlihatkan adanya beberapa sekuens rbcL berukuran sangat pendek atau dengan variasi asam amino yang lebih ekstrem dibanding mayoritas sekuens lainnya, sehingga representasi k-mer miliknya cenderung tumpang tindih dengan pola gen lain.

Skor F1 untuk kelas rbcL yang mencapai 0.99 menunjukkan keseimbangan yang sangat baik antara kemampuan model mengidentifikasi kelas positif dan menekan kesalahan prediksi. Stabilitas performa model juga terlihat dari nilai macro average dan weighted average yang konsisten tinggi (≥0.99), meskipun dataset sangat tidak seimbang, dengan rasio kelas yang jauh berpihak pada kelompok non-rbcL. Kemampuan Random Forest menangani ketidakseimbangan ini menunjukkan ketangguhan algoritma dalam mempelajari distribusi fitur frekuensi k-mer, terlepas dari variasi panjang sekuens atau disparitas jumlah sampel.

Kinerja model semakin diperkuat melalui kurva ROC yang ditampilkan pada Gambar 4, dengan nilai AUC sebesar 0.999. Bentuk kurva yang hampir tegak lurus pada sumbu Y

e-ISSN 2827-8356
p-ISSN 2827-9999

menegaskan bahwa model memiliki kapasitas diskriminatif yang sangat tinggi dalam memisahkan dua kelas. Nilai AUC mendekati 1 bukan hanya mengindikasikan prediksi yang akurat, tetapi juga kestabilan performa saat model dihadapkan pada berbagai threshold probabilitas. Dalam konteks bioinformatika, pencapaian ini memberikan nilai tambah signifikan karena mengurangi kebutuhan akan validasi manual berbasis alignment seperti BLAST yang memerlukan waktu dan sumber daya komputasi lebih besar.

Untuk memberikan interpretasi biologis terhadap model, dilakukan analisis feature importance pada Random Forest. Hasilnya menunjukkan bahwa hanya sebagian kecil dipeptida memberikan kontribusi dominan terhadap keputusan klasifikasi. Dipeptida seperti MS, SP, PQ, QT, dan KA muncul sebagai fitur dengan bobot tertinggi. Pola ini konsisten dengan struktur awal sekuens rbcL yang umumnya diawali oleh motif konservatif MSPQTET…, sebagaimana terlihat pada contoh sekuens di Tabel 2. Temuan ini mengindikasikan bahwa model tidak sekadar menghafal panjang atau komposisi global sekuens, tetapi benar-benar menangkap motif biologis khas rbcL pada tingkat lokal (dipeptida). Dengan demikian, pendekatan k-mer yang digunakan tidak hanya efektif secara komputasional, tetapi juga bermakna secara biologis, karena mampu mengekstraksi ciri sekuens yang relevan dengan identitas gen rbcL.

Secara keseluruhan, performa hampir sempurna yang dicapai model menunjukkan bahwa kombinasi representasi protein k-mer dan algoritma Random Forest sangat sesuai untuk tugas klasifikasi gen plastid. Implementasi ini tidak hanya efektif dalam mengolah jumlah data besar, tetapi juga adaptif terhadap variasi internal antar-gen yang tercermin dalam dataset. Keberhasilan ini mendukung potensi pendekatan *alignment-free* sebagai alternatif cepat, efisien, dan akurat untuk anotasi gen rbcL dalam skala besar, terutama ketika diterapkan pada dataset yang berasal dari sumber beragam dalam kingdom Viridiplantae.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model klasifikasi berbasis representasi protein k-mer dan algoritma Random Forest untuk membedakan gen rbcL dari empat gen plastid lainnya pada tumbuhan. Proses ekstraksi fitur menggunakan k-mer berukuran dua terbukti mampu menangkap pola dipeptida yang khas dari sekuens rbcL, sehingga menghasilkan representasi numerik yang stabil, efisien, dan alignment-free. Berdasarkan hasil evaluasi, model menunjukkan performa yang sangat tinggi dengan akurasi mencapai 0.99, precision 1.00, recall 0.978, serta F1-score 0.989 untuk kelas rbcL, yang menunjukkan kemampuan model dalam mengidentifikasi sekuens target secara akurat meskipun terjadi sedikit kesalahan klasifikasi pada satu sekuens.Nilai AUC sebesar 0.999 pada kurva ROC mempertegas bahwa model memiliki kapasitas diskriminatif yang hampir sempurna dalam memisahkan kelas rbcL dan non-rbcL. Temuan ini membuktikan bahwa pendekatan k-mer + Random Forest merupakan solusi komputasi yang efektif dan skalabel dibandingkan metode penyelarasan tradisional seperti BLAST, terutama dalam analisis dataset berukuran besar. Dengan demikian, model yang dikembangkan pada penelitian ini layak digunakan sebagai alat anotasi otomatis untuk identifikasi gen rbcL pada sekuens protein tumbuhan, serta berpotensi diperluas ke aplikasi bioinformatika lain, termasuk klasifikasi gen plastid dan karakterisasi sekuens pada berbagai taxa.

UCAPAN TERIMA KASIH

Hasil penelitian ini menunjukkan bahwa pendekatan klasifikasi berbasis representasi protein k-mer dan algoritma Random Forest mampu mengidentifikasi gen rbcL dengan akurasi sangat tinggi. Meski demikian, terdapat beberapa arah pengembangan yang dapat dilakukan pada penelitian selanjutnya. Model dapat diperluas untuk mengklasifikasikan lebih banyak gen plastid atau nuklir, sehingga berguna dalam pipeline anotasi genom yang lebih komprehensif. Selain itu, penggunaan nilai k yang lebih besar (misalnya k=3 atau k=4) ataupun representasi embedding

e-ISSN 2827-8356
p-ISSN 2827-9999

modern seperti ESM atau ProtT5 dapat dieksplorasi untuk meningkatkan sensitivitas pada sekuens yang sangat bervariasi. Studi lebih lanjut juga dapat memasukkan metode deep learning seperti CNN, LSTM, atau transformer untuk dibandingkan dengan performa model Random Forest, sehingga diperoleh gambaran menyeluruh terkait efektivitas berbagai pendekatan alignment-free dalam anotasi sekuens protein. Integrasi data filogenetik tambahan atau metadata geografis juga berpotensi meningkatkan akurasi klasifikasi pada kasus nyata.

Penelitian ini tidak terlepas dari dukungan berbagai sumber daya dan perangkat ilmiah. Penulis mengucapkan terima kasih kepada pengelola NCBI Protein Database atas akses terbuka terhadap data biologis yang menjadi fondasi penelitian ini. Ucapan terima kasih juga disampaikan kepada komunitas pengembang perangkat lunak Python, Biopython, scikit-learn, serta platform Jupyter Notebook, yang telah menyediakan ekosistem komputasi ilmiah yang stabil dan mudah digunakan. Dukungan teknis dan masukan dari berbagai pihak selama proses analisis dan penyusunan artikel ini sangat membantu dalam menghasilkan penelitian yang lebih sistematis dan komprehensif.

DAFTAR PUSTAKA

[1] W. J. Kress and D. L. Erickson, “A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region,” *PLoS ONE*, vol. 2, no. 6, p. e508, Jun. 2007, doi: 10.1371/journal.pone.0000508.

[2] D. Alzahrani, E. Albokhari, S. Yaradua, and A. Abba, “Complete chloroplast genome sequences of Dipterygium glaucum and Cleome chrysantha and other Cleomaceae Species, comparative analysis and phylogenetic relationships,” *Saudi J. Biol. Sci.*, vol. 28, no. 4, pp. 2476–2490, Apr. 2021, doi: 10.1016/j.sjbs.2021.01.049.

[3] M. K. Vasquez, E. K. Stock, K. J. Terrell, J. Ramirez, and J. A. Kyndt, “Unraveling Evolutionary Dynamics: Comparative Analysis of Chloroplast Genome of Cleomella serrulata from Leaf Extracts,” *Int. J. Plant Biol.*, vol. 15, no. 3, pp. 914–926, Sep. 2024, doi: 10.3390/ijpb15030065.

[4] A. Muwaffiq Faza *et al.*, “In Silico Evaluation of rbcL, matK, and psbA-trnH Regions on the Genus Spatholobus (Fabaceae),” *J. Ris. Biol. Dan Apl.*, vol. 6, no. 2, pp. 73–81, Sep. 2024, doi: 10.26740/jrba.v6n2.p73-81.

[5] S. Letsiou *et al.*, “DNA Barcoding as a Plant Identification Method,” *Appl. Sci.*, vol. 14, no. 4, p. 1415, Feb. 2024, doi: 10.3390/app14041415.

[6] NCBI, *NCBI Protein Database — Search results for “rbcL”. National Center for Biotechnology Information.*, Jan. 01, 2024. Accessed: Sep. 22, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/protein

[7] J. Shaw and Y. W. Yu, “Fast and robust metagenomic sequence comparison through sparse chaining with skani,” vol. Volume 20, Sep. 2023, doi: https://doi.org/10.1038/s41592-023-02018-3.

[8] L. He, S. Sun, Q. Zhang, X. Bao, and P. K. Li, “Alignment-free sequence comparison for virus genomes based on location correlation coefficient,” *Infect. Genet. Evol.*, vol. 96, p. 105106, Dec. 2021, doi: 10.1016/j.meegid.2021.105106.

[9] L. He *et al.*, “A new alignment-free method: K-mer Subsequence Natural Vector (K-mer SNV) for classification of fungi,” *BMC Bioinformatics*, vol. 26, no. 1, p. 170, Jul. 2025, doi: 10.1186/s12859-025-06152-x.

[10] M. T. Swain and M. Vickers, “Interpreting alignment-free sequence comparison: what makes a score a good score?,” *NAR Genomics Bioinforma.*, vol. 4, no. 3, p. lqac062, Jul. 2022, doi: 10.1093/nargab/lqac062.

[11] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, “Alignment-free sequence comparison: benefits, applications, and tools,” *Genome Biol.*, vol. 18, no. 1, p. 186, Dec. 2017, doi: 10.1186/s13059-017-1319-7.

[12] D. Tang *et al.*, “KCOSS: an ultra-fast k-mer counter for assembled genome analysis,” *Bioinformatics*, vol. 38, no. 4, pp. 933–940, **Jan. 2022,** doi: 10.1093/bioinformatics/btab797.

[13] J. T. Lee, X. Li, C. Hyde, P. A. Liberator, and L. Hao, “PfaSTer: a machine learning-powered serotype caller for Streptococcus pneumoniae genomes,” *Microb. Genomics*, vol. 9, no. 6, **Jun. 2023,** doi: 10.1099/mgen.0.001033.

e-ISSN 2827-8356
p-ISSN 2827-9999

[14] J. Wang *et al.*, "Scaffolding protein functional sites using deep learning," *Science*, vol. 377, no. 6604, pp. 387–394, Jul. 2022, doi: 10.1126/science.abn2100.
[15] S. Yin, "UPFPSR: a ubiquitylation predictor for plant through combining sequence information and random forest," vol. 19, no. 1, Nov. 2022, doi: 10.3934/mbe.2022035.
[16] D. Simón, O. Borsani, and C. V. Filippi, "RFPDR: a random forest approach for plant disease resistance protein prediction," *PeerJ*, vol. 10, p. e11683, Apr. 2022, doi: 10.7717/peerj.11683.
[17] S. Seo, M. Oh, Y. Park, and S. Kim, "DeepFam: deep learning based alignment-free method for protein family modeling and prediction," *Bioinformatics*, vol. 34, no. 13, pp. i254–i262, Jul. 2018, doi: 10.1093/bioinformatics/bty275.
[18] N. A. Saputra, L. S. Riza, A. Setiawan, and I. Hamidah, "A Systematic Review for Classification and Selection of Deep Learning Methods," *Decis. Anal. J.*, vol. 12, p. 100489, Sep. 2024, doi: 10.1016/j.dajour.2024.100489.
[19] D. J. Van Zyl *et al.*, "Alignment-Free Viral Sequence Classification at Scale," Dec. 11, 2024, *Genomics*. doi: 10.1101/2024.12.10.627186.
[20] A. L. Delcher, "Fast algorithms for large-scale genome alignment and comparison," *Nucleic Acids Res.*, vol. 30, no. 11, pp. 2478–2483, Jun. 2002, doi: 10.1093/nar/30.11.2478.
[21] Suraj Varma, *Analysis of a Protein sequence(fasta dataset)*, Dec. 30, 2024. Accessed: Sep. 23, 2025. [Online]. Available: https://github.com/suraj5424/Protein-sequence-analysis
[22] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, "Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced Dataset," *Electronics*, vol. 12, no. 21, p. 4423, Oct. 2023, doi: 10.3390/electronics12214423.
[23] E. E. Ojeda Avilés, D. Olmos Liceaga, and J.-H. Jung, "Stratified Sampling Algorithms for Machine Learning Methods in Solving Two-scale Partial Differential Equations," *J. Sci. Comput.*, vol. 104, no. 3, p. 110, Sep. 2025, doi: 10.1007/s10915-025-03024-7.