Jurnal Ilmiah Sinus (JIS) Vol : 21, No. 2, Juli 2023
ISSN (Print) : 1693-1173 , ISSN (Online): 2548-4028

Ekstraksi Fitur Rantai Markov untuk Klasifikasi Famili Protein

Toto Haryanto[1)], Rizky Kurniawan[2)], Sony Muhammad[3)], Aziz Kustiyo[4*)], Endang Purnama Giri[5)]
[1)] Departemen Ilmu Komputer, FMIPA, Institut Pertanian Bogor
[1)] totoharyanto@apps.ipb.ac.id, [2)]rizkyasagikurniawan@gmail.com. [3)]sonywr10@gmail.com, [4)]azizku@apps.ipb.ac.id, [5)]endang_pg@apps.ipb.ac.id

ABSTRACT

Proteins, as complex molecules, have a variety of tasks in living organisms. Proteins are organic molecules made up of twenty amino acid combinations that perform diverse tasks such as transportation, chemical reaction catalysts for metabolism, and food stores. This research aims to classify proteins family based on sequences of amino acids as the primary structure. There are 300 amino acid fragments obtained from Pfam database. The proteins family database subset with three sub sample classes, including 1-cysPrx_C, 4HBT, and ABC_Tran, were obtained. In this research, we apply first and second order of the Markov Chain for extracting features. In comparison to the joint probability strategy, we then use a Probabilistic Neural Network (PNN) as a classifier. The evaluation process was conducted by comparing sensitivity and specificity of both classification techniques. In general, the results show that PNN has slightly better performance than the joint probability technique for classifying protein family.

***Keywords**: amino acid, joint probability, markov chain, probabilistic neural network, protein family*

I. PENDAHULUAN

Protein merupakan unsur penting bagi makhluk hidup. Secara struktural, protein terdiri dari struktur primer, struktur sekunder dan struktur tersier. Sifat-sifat fisikokimia dari sekuen asam amino berisi informasi mengenai fungsi dan evolusi protein (Hatton & Warr, 2015). Klasifikasi famili protein memiliki peran penting seperti pada perbaikan identifikasi protein, untuk membantu memelihara basisdata famili protein, untuk mengambil informasi biologis dengan data yang sangat besar secara efektif, serta untuk mewakili ekspresi gen dari famili protein untuk keperluan analisa filogenetik dan untuk profil sekuen biologis (Naveenkumar *et al.*, 2018). Protein memiliki beberapa basisdata dengan tujuan yang spesifik, seperti basisdata famili hierarkis (PIR-PSD dan Peta Proto), basisdata domain protein (Pfam) (Finn et al., 2014), basisdata sekuen motif (PROSITE dan PRINT), basisdata struktural (SCOP dan CATH) (Andreeva et al., 2020), dan basisdata famili terintegrasi InterPro (Mitchell et al., 2015). Klasifikasi protein juga penting untuk pencarian informasi seperti struktur, aktivitas dan anotasi, dan metabolisme sistem (Wu et al., 2003) .

Di sisi lain, bidang proteomik yang memanfaatkan pendekatan machine learning memiliki potensi besar yang dibuktikan dengan jumlah publikasi yang mencapai 5000 artikel setiap tahunnya (Desaire et al., 2022). Klasifikasi famili protein dan anotasi protein dilakukan menggunakan metode rule base (Wu et al., 2003). Selain itu, klasifikasi protein juga dilakukan menggunakan metode *support vector machine* (SVM) (Gupta et al., 2019). Penelitian sebelumnya adalah berdasarkan *Sparse Markov Transducer* (SMT) (Eskin et al., 200), yang berfokus pada dua teknik dari SMT yaitu model prediksi SMT dan model pengklasifikasi SMT. Klasifikasi menggunakan model prediksi SMT memperoleh akurasi tertinggi hingga 100% untuk kelas-kelas FGF, MCP sinyal dan MHC_I. Untuk model pengklasifikasi SMT, akurasi tertinggi diperoleh dari kelas-kelas TIM, S12, MCP Kelas MHC_I, FGF, Cys_knot, ATP-synt-ab dan 7tm_3 dengan akurasi sebesar 100%.

Penelitian terkait famili protein telah dilakukan oleh beberapa peneliti. Penggunaan arsitektur *deep learning* seperti DenseNet diterapkan dalam identifikasi famili protein (Imrie et al., 2018). Beberapa metode baru berbasis *deep learning* juga telah diusulkan, seperti DeepFam (Seo et al., 2018) dan ProtCNN (Bileschi et al., 2019) untuk mengidentifikasi dan menganotasi protein. Selain itu (Sandaruwan & Wannige, 2012) mengusulkan perbaikan arsitektur model untuk klasifikasi famili protein.

Fitur rantai Markov merupakan salah satu model probabilistik yang sering digunakan dalam bidang penelitian bioinformatika. Rantai Markov dapat diterapkan untuk menganalisis sekuens DNA, RNA, dan sekuens asam amino (Usotskaya & Ryabko, 2009). Sekuen asam amino terbentuk melalui dua proses yaitu transkripsi DNA dan translasi RNA. Asam amino dapat dihasilkan oleh kodon RNA atau triplet. Sekuen-sekuen asam amino akan merepresentasikan protein. Pada rantai Markov, sekuen baru dapat diidentifikasi menggunakan peluang gabungan (*joint probability*) dengan asumsi Markov (Teugels, 2008). Sebagai model generatif, rantai Markov banyak digunakan dalam bidang penelitian bioinformatika seperti untuk aplikasi ontologi (Robert & Alexa, 2012), ekspresi gen (Robert & Alexa, 2012), analisis data segmentasi citra (Robinson et al., 2015), dan analisis interaksi jaringan (Robinson et al., 2017).

Selain itu, *Probabilistic Neural Network* (PNN), sebagai pengklasifikasi, memiliki peran penting dalam penelitian bioinformatika. Dalam bidang proteomik, PNN digunakan sebagai pendekatan untuk mengklasifikasikan superfamili protein dengan 497 fitur, tidak hanya dari informasi asam amino (Rao et al., 2005). PNN juga digunakan untuk prediksi struktur sekunder dari protein folding (Ibrahim & Yasseen, 2017). PNN juga digunakan dalam bidang penelitian genomik untuk pola sekuens analisis DNA (Wu et al., 2005), selain itu PNN digunakan sebagai dasar untuk menganalisis data genom berdimensi tinggi (Baliarsingh et al., 2020). Penelitian ini bertujuan untuk mengklasifikasikan famili protein yang terutama diperoleh dari sekuen asam amino sebagai struktur primer. Teknik klasifikasi yang digunakan dalam penelitian ini adalah peluang gabungan dengan asumsi Markov dan PNN. Evaluasi terhadap kedua teknik klasifikasi tersebut dilakukan dengan membandingkan kinerja keduanya.

II. TINJAUAN PUSTAKA

3.1 Famili Protein

Protein dengan kelas yang sama memiliki ancestor yang sama dan umumnya serupa dalam sekuen, struktur tiga dimensi, dan fungsionalitas. Pada kondisi tertentu, sulit untuk mengevaluasi signifikansi kemiripan di antara famili protein. Namun, protein dengan *ancestor* yang berbeda memiliki sekuen asam amino yang berbeda. Inilah alasan mengapa banyak teknik dikembangkan untuk mengklasifikasikan famili protein berdasarkan sekuen asam amino. Sebelumnya, terdapat lebih dari 60.000 kelas famili protein. Basisdata famili protein dapat diakses di Uniform Resource Locator (URL) pusat penelitian Sanger Genome Institute yang dapat diakses di ftp://ftp.sanger.ac.uk/pub/databases/Pfam.

Setelah lebih dari dua dekade, pengembangan dalam basisdata famili protein semakin banyak dilakukan. Saat ini basisdata tersebut secara resmi dikenal sebagai InterPro (Mitchell et al., 2015). Basisdata ini disumbangkan oleh lebih dari tiga puluh peneliti (Blum et al., 2021). Versi terbaru dari basisdata protein dapat diakses di https://www.ebi.ac.uk/interpro/.

3.2 Model Rantai Markov dan Hidden Markov

Metode rantai Markov diperkenalkan oleh matematikawan Rusia, AA Markov pada awal abad ke-20. Menggunakan proses Markov, fenomena stokastik di dunia nyata dapat dimodelkan. Masalah yang mendasari rantai Markov adalah bagaimana menentukan

kondisi transisi (*state transition*) dengan tepat sehingga proses stokastik tersebut memenuhi sifat Markov. Hal ini berarti bahwa informasi terkait kondisi (*state*) tersebut sudah cukup tersedia untuk memprediksi perilaku stokastik berikutnya (Teugels, 2008). Selanjutnya, Hidden Markov Model (HMM) juga banyak digunakan dalam bioinformatika seperti prediksi struktur protein (Lasfar & Bouden, 2018). Konsep rantai markov dan perluasannya juga telah diterapkan dalam banyak bidang penelitian seperti *hidden markov* berbasis *kernel* (De Gooijer et al., 2022) dan *kernel tree* (Chang et al., 2022).

Rantai Markov dapat disebut sebagai Markov waktu diskrit jika ruang proses Markov adalah himpunan terbatas atau dapat dihitung. Jika nilai kondisi dalam periode tertentu hanya bergantung pada satu kondisi sebelumnya maka disebut sebagai rantai Markov orde pertama. Persamaan (1) menjelaskan orde pertama dari rantai Markov.

$$P\{X_{n+1} = j | X_n = i\} \ldots \quad (1)$$

Jika nilai kondisi di dalam periode kondisi tertentu bergantung pada m keadaan sebelumnya, maka ini disebut sebagai rantai Markov orde-m seperti pada Persamaan (2).

$$P\{X_{n+1} = j | X_{(n+1)-m} = i_1, X_{(n+1)-m+1} = i_2, \ldots X_n = i_n\} \ldots (2)$$

Himpunan kondisi dalam penelitian ini direpresentasikan sebagai sekuen asam amino. Peluang X_{n+1} pada kondisi j diberikan X_n pada state i disebut sebagai peluang orde pertama mengacu pada Persamaan (3).

$$P_{ij}^{n,n+1} = P\{X_{n+1} = j | X_n = i\} \ldots (3)$$

Peluang kondisi transisi direpresentasikan sebagai matriks transisi yang dikenal sebagai matriks transisi peluang P. P_{ij} didefinisikan sebagai P{X_(n+1)=j|X_n=i} mengacu pada Persamaan (4). Untuk setiap kelas dalam penelitian ini memiliki matriks sebagai model. Untuk mengidentifikasi sekuen asam amino baru, diterapkankan peluang gabungan dengan asumsi Markov mengacu pada Persamaan (5).

$$P = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \ldots (4)$$

$$P(X_1, \ldots X_n) = \prod_{i=1}^{n} P(X_i | X_{i-1}) \ldots (5)$$

Dari Persamaan (5) akan dihasilkan peluang sekuen baru berdasarkan matriks transisi yang diberikan. Nilai peluang maksimum dari matriks tersebut akan menentukan hasil klasifikasi dari suatu sekuen baru.

3.3 *Probabilistic Neural Networks* (PNN)

Probabilistic Neural Network (PNN) adalah salah satu teknik klasifikasi dengan struktur empat lapis (*layer*), yaitu lapisan input, lapisan pola, lapisan penjumlahan dan lapisan keputusan. Berbeda dengan *Neural Network* (NN), PNN tidak memiliki proses mundur (*backward*). Oleh karena itu, PNN lebih efisien dibandingkan *back-propagation neural networks* (Specht, 1990; Mohebali et al., 2020).

Lapisan input adalah lapisan pertama di PNN. Pada penelitian ini lapisan input berupa input vektor fitur dengan panjang 400. Angka ini diperoleh dari ekstraksi fitur

menggunakan rantai Markov dari nilai matriks transisi antara 20 asam amino. Oleh karena itu terdapat nilai peluang 20x20 yang akan menjadi vektor input di PNN.

Lapisan pola sebagai lapisan kedua mewakili pola kelas. Pada lapisan ini, data latih akan dikelompokkan menjadi tiga kelas famili protein yaitu 1-cysPrx_C, 4HBT dan ABC_Tran. Terdapat 75 fragmen sebagai data latih untuk setiap kelas. Pada prinsipnya, pada lapisan pola, terdapat proses perhitungan kemiripan antara sekuen baru dengan sekuen yang ada pada data latih menggunakan *kernel Gaussian* dengan parameter *smoothing* mengacu pada Persamaan (6).

$$f(x) = exp\left(-\frac{(x-x_{A_i})^T(x-x_{A_i})}{2\sigma^2}\right)\text{.....(6)}$$

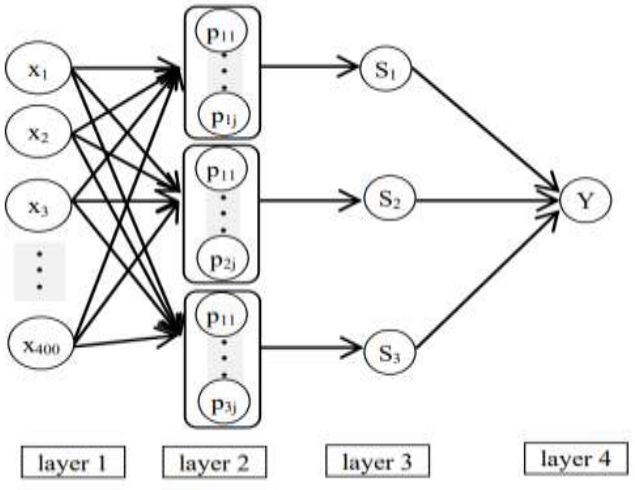
Lapisan penjumlahan adalah lapisan ketiga di PNN. Di lapisan ini, akan dilakukan penjumlahan peluang untuk data pengujian untuk setiap kelas di lapisan pola. Persamaan (7) adalah perhitungan pada lapisan penjumlahan (Mohebali et al., 2020).

$$p(\omega_A)p(x|\omega_A) = \frac{1}{2\pi^{\frac{d}{2}}\sigma^d N}\sum_{i=1}^{N_A} exp\left(-\frac{(x-x_{A_i})^T(x-x_{A_i})}{2\sigma^2}\right)\text{.....(7)}$$

Keterangan :

$p(\omega_A)$: peluang kelas A
$p(x|\omega_A)$: peluang bersyarat x pada kelas A
x_{A_i} : vektor data latih ke-i pada kelas A
d : dimensi vektor input
N : jumlah pola latih seluruh kelas
N_A : jumlah pola latihan di kelas A
σ : parameter pemulusan (smoothing parameter)

Lapisan keputusan sebagai lapisan keempat merupakan penentuan asam amino baru untuk dapat diklasifikasikan ke dalam salah satu dari tiga kelas dalam lapisan pola. Lapisan ini menghitung nilai maksimum yang dihasilkan oleh lapisan penjumlahan. Struktur PNN dapat dilihat pada Gambar 1, yang menunjukkan 400 input hasil ekstraksi fitur, tiga lapisan pola mewakili 3 kelas famili protein, dan tiga lapisan penjumlahan untuk setiap kelas Y sebagai lapisan keputusan untuk mendapatkan hasil klasifikasi.



Gambar 1. Struktur PNN dengan 400 input dan 3 kelas

 DOI : https://doi.org/10.30646/sinus.v21i2.748

3.4 Metode Evaluasi

Kinerja metode klasifikasi diukur menggunakan sensitivitas dan spesifisitas untuk setiap famili protein berdasarkan confusion matrix (Tabel 1).

Sensitivitas mengukur proporsi positif dari famili protein baru yang dapat diklasifikasikan dengan benar mengacu pada Persamaan (8) sedangkan spesifisitas mengukur proporsi negatif dari famili protein baru yang dapat diklasifikasikan dengan benar mengacu pada Persamaan (9).

Tabel 1. *Confusion matrix* untuk setiap famili protein

		Diprediksi sebagai	
		Sekuen famili protein	Bukan sekuen famili protein
aktual	Sekuen famili protein	tp	fn
	Bukan sekuen famili protein	fp	tn

Keterangan

tp : *true positive* (jumlah sekuen famili protein yang diprediksi sebagai sekuen famili protein)

tn : *true negative* (jumlah bukan sekuen famili protein yang diprediksi sebagai bukan sekuen famili protein)

fp : *false positive* (jumlah bukan sekuen famili protein yang diprediksi sebagai sekuen famili protein)

fn : *false negative* (jumlah sekuen famili protein yang diprediksi sebagai bukan sekuen famili protein)

Sensitivitas mengukur proporsi positif dari famili protein baru yang dapat diklasifikasikan dengan benar mengacu pada Persamaan (8) sedangkan spesifisitas mengukur proporsi negatif dari famili protein baru yang dapat diklasifikasikan dengan benar mengacu pada Persamaan (9).

$$sensitivity = \frac{tp}{tp+fn} \quad \text{.....(8)}$$

$$spesificiy = \frac{tn}{tn+fp} \quad \text{.....(9)}$$

III. METODE PENELITIAN

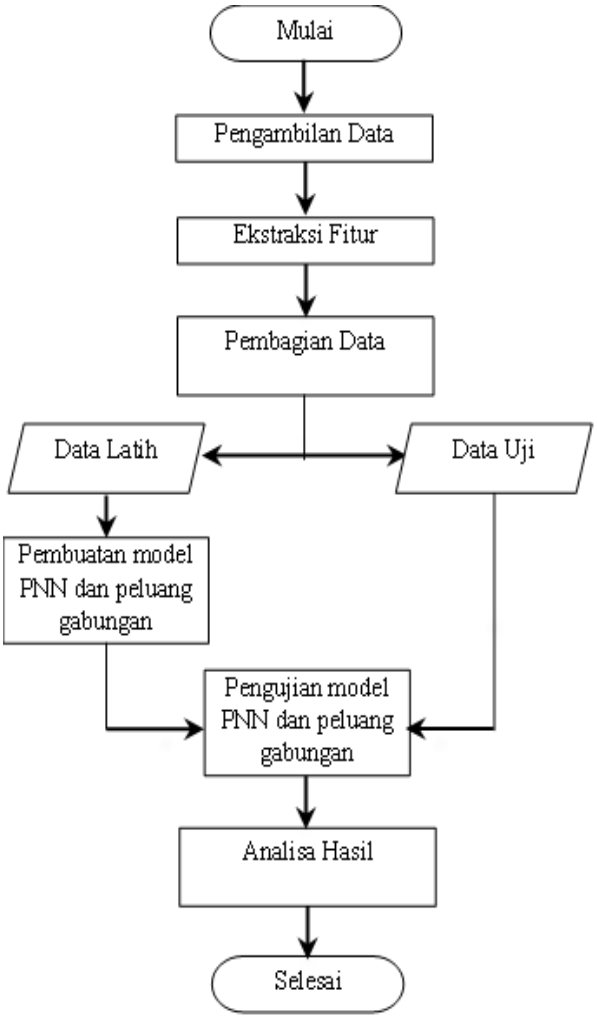
Penelitian ini dilakukan dengan beberapa tahapan untuk mengetahui kinerja PNN dan metode peluang gabungan. Tahapan-tahapan mulai dari pengambilan data hingga analisa hasil akan dijelaskan sesuai Gambar 2.

3.1 Pengambilan Data

Data protein family diperoleh dari Pfam yang merupakan pusat database komprehensif yang meliputi berbagai macam koleksi data. Pfam juga telah dipergunakan secara luas oleh berbagai macam komunitas yang bergerak di bidang *structural biology* untuk mengidentifikasi jenis protein baru. Data yang digunakan terdiri dari tiga kelas famili protein kelas 1-cysPrx_C, 4HBT dan ABC_Tran, yang masing-masing kelasnya mempunyai 100 data protein dengan format FASTA (sehingga total terdapat 300 data). Data tersebut diunduh dari http://www.pfam.sanger.ac.uk.

3.2 Ekstraksi Fitur

Pada penelitian ini, matriks transisi akan dibangun berukuran 20×20 sesuai dengan banyaknya monomer atau jenis asam amino yang ada. Terdapat dua jenis matriks transisi yang dibangun yaitu matriks transisi rantai Markov orde pertama dan orde kedua. Proses ekstraksi ciri dilakukan pada semua data, baik data latih maupun data uji. Jadi akan terdapat 300 matriks data hasil ektraksi ciri yang akan diproses ke tahap selanjutnya. Hasil ekstraksi ciri yang berupa matriks berukuran 20x20 lalu diubah menjadi sebuah vektor berukuran 1x400.



Gambar 2. Tahapan-tahapan penelitian

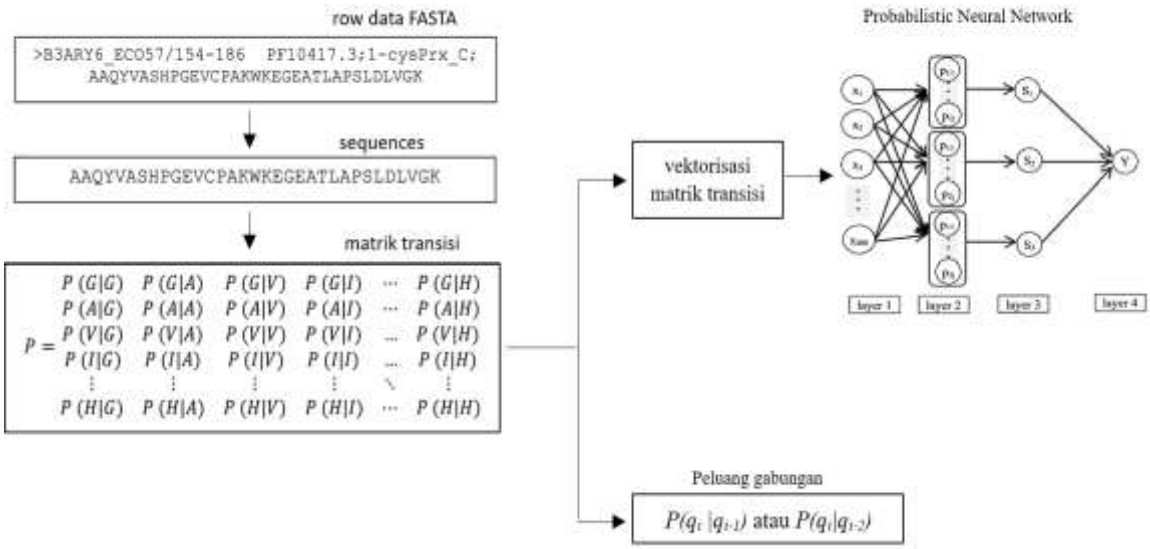
3.3 Pembagian Data

Setelah dilakukan ekstraksi ciri, data kemudian dibagi menjadi data latih dan data uji meenggunakan metode *k-fold cross validation* dengan k = 4. Semua data akan dibagi menjadi 4 *fold* yaitu S1, S2, S3, S4 yang masing-masing fold memiliki data yang sama. Proses identifikasi akan dilakukan 4 kali iterasi berdasarkan 4-*fold cross validation* dan data latih serta data uji memiliki *fold* yang berbeda pada setiap iterasinya sehingga terciptanya model klasifikasi yang terbaik. Dengan total 300 data, pembagian 4-*fold* ini akan menjadikan jumlah data latih sebanyak 225 dan data uji sebanyak 75 untuk setiap skenario *fold*.

3.4 Pembuatan dan Pengujian Model PNN dan Peluang Gabungan untuk klasifikasi

Pada tahap ini dilakukan pembuatan model PNN dan model peluang gabungan menggunakan data latih. Pengujian kedua model tersebut menggunakan data uji. Data latih

dan data uji telah ditentukan pada tahapan sebelumnya. Secara umum proses klasifikasi dengan PNN dan peluang gabungan disajikan pada Gambar 3.



Gambar 3. Proses Klasifikasi Protein Family dengan PNN dan Peluang Gabungan

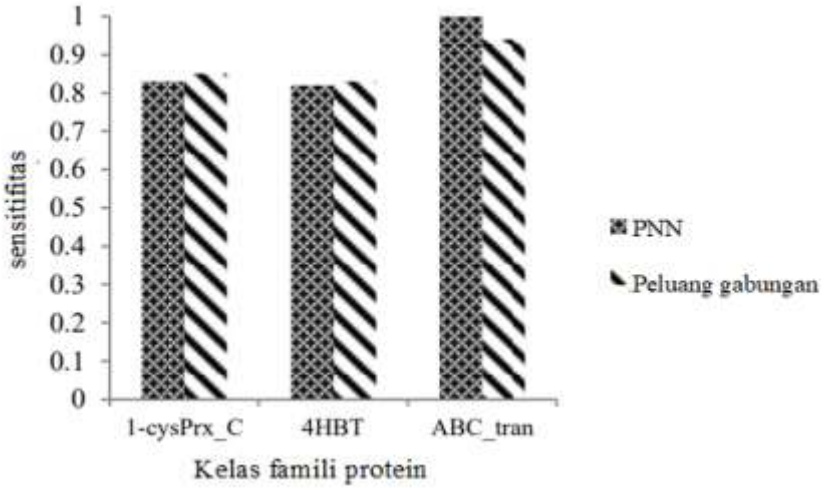
3.5 Analisa Hasil

Pada tahap ini dilakukan analisa terhadap hasil pengujian model PNN dan model peluang gabungan. Analisa hasil tersebut berdasarkan nilai sensitivitas dan spesifitas kedua model tersebut.

IV. HASIL DAN PEMBAHASAN

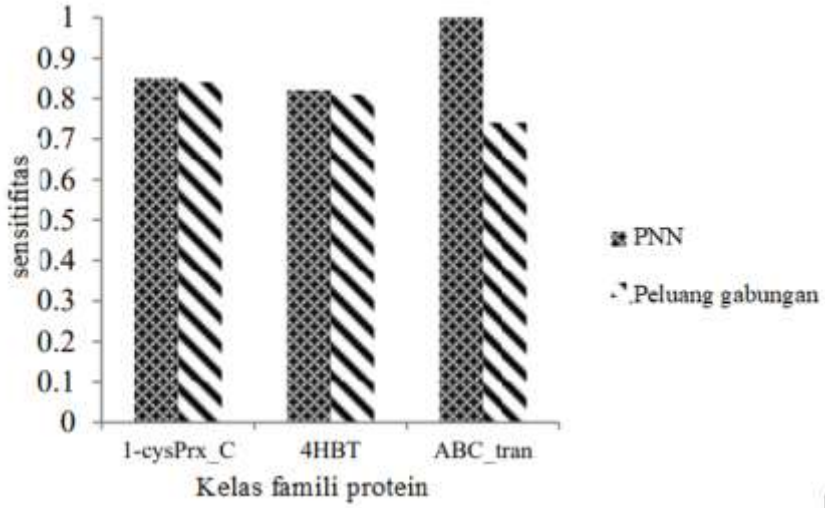
4.1 Perbandingan Nilai Sensitivitas Model PNN dan Peluang Gabungan

Rata-rata sensitivitas dan spesifisitas untuk semua fold dihitung untuk mengevaluasi kedua teknik klasifikasi, PNN dan peluang gabungan. Baik nilai sensitivitas maupun spesifisitas dibagi menjadi dua metode dari orde rantai Markov, yaitu orde pertama dan orde kedua. Gambar 4 menunjukkan perbandingan sensitivitas antara PNN dan peluang gabungan dengan asumsi Markov. Angka tersebut menunjukkan bahwa sensitivitas PNN dan peluang gabungan di kelas 1-cysPrx dan 4HBT sebanding. Namun demikian, pada kelas ABC_tran, sensitivitas PNN lebih tinggi 1,0 dibandingkan sensitivitas peluang gabungan (0,94). Jika sensitivitasnya 1,0, berarti semua ABC_tran dalam data aktual telah diklasifikasikan dengan benar sebagai ABC_tran.



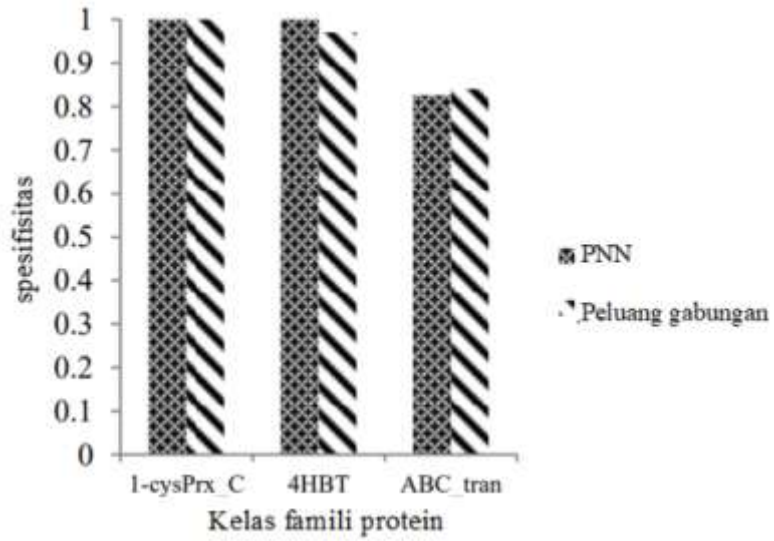
Gambar 4. Sensitivitas model PNN dan peluang gabungan dengan ekstraksi fitur rantai Markov orde pertama

Pada Gambar 5 disajikan nilai sensitivitas model PNN dan peluang gabungan menggunakan asumsi Markov orde kedua. Gambar 5 menunjukkan bahwa semua kelas ABC_tran telah diklasifikasikan dengan benar dengan nilai sensitivitas 1,0. Selain itu, untuk dua kelas lainnya, nilai sensitivitasnya sebanding. Dari Gambar 4 dan 5 dapat disimpulkan bahwa PNN mampu menangkap pola untuk ABC_tran lebih baik dibandingkan model peluang gabungan. Selain itu, ABC_tran memiliki pola yang berbeda dibandingkan kedua kelas lainnya.



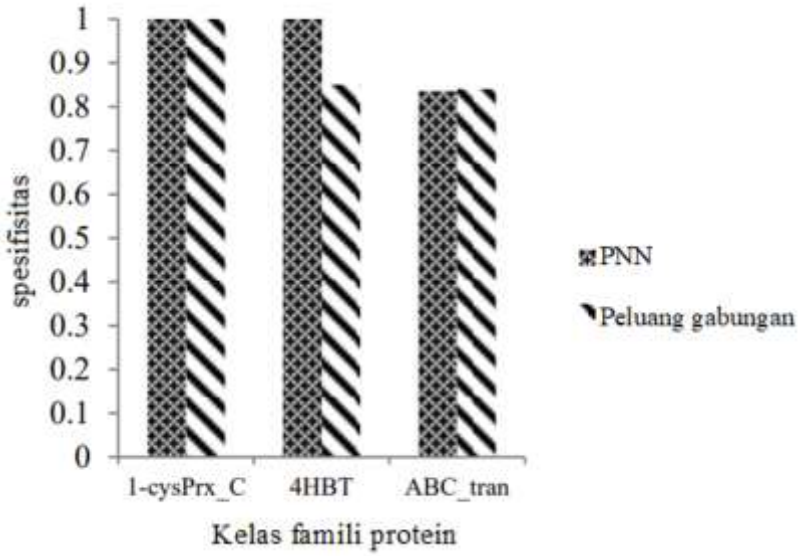
Gambar 5. Sensitivitas model PNN dan peluang gabungan dengan ekstraksi fitur rantai Markov orde kedua

Nilai spesifitas model PNN dan peluang gabungan disajikan pada Gambar 6. Dapat dilihat dari Gambar 6 bahwa model PNN memperoleh nilai spesifisitas 1,0 untuk famili protein 1-cysPrx dan 4HBT, lebih tinggi dibandingkan model peluang gabungan. Hasil tersebut menunjukkan bahwa semua data yang bukan milik 1-cysPrx atau 4HBT diklasifikasikan ke dalam kelas selain 1-cysPrx atau 4HBT. Untuk kelas ABC_tran, model peluang gabungan memiliki nilai spesifitas lebih tinggi dibandingkan model PNN.



Gambar 6. Spesifitas model PNN dan peluang gabungan dengan ekstraksi fitur rantai Markov orde pertama

Nilai spesifisitas juga dihitung untuk rantai Markov orde kedua seperti halnya pada nilai sensitivitas. Gambar 7 menunjukkan bahwa nilai spesifisitas PNN masing-masing adalah 1.0, 1.0, dan 0.83 untuk 1-cycPrx, 4HBT dan ABC_tran, sedangkan metode peluang gabungan menghasilkan nilai spesifisitas adalah 1,0, 0,95, dan 0,84. Dibandingkan dengan Gambar 6, spesifisitas penggunaan rantai Markov orde kedua tidak lebih baik dibandingkan dengan penggunaan rantai Markov orde pertama dalam proses ekstraksi ciri.



Gambar 7. Spesifitas model PNN dan peluang gabungan dengan ekstraksi fitur rantai Markov orde kedua

Berdasarkan hasil sensitivitas dan spesifisitas, rantai markov mampu mengekstraksi informasi untuk famili sekuens protein. Hal ini dapat dilihat pada nilai sensitivitas dan spesifisitas baik menggunakan PNN maupun nilai peluang gabungan. Dengan demikian, rantai Markov masih menjadi metode yang menjanjikan untuk mengekstraksi informasi, terutama pada data sekuensial, seperti data asam amino.

V. KESIMPULAN DAN SARAN

5.1 KESIMPULAN

Sekuen asam amino sebagai struktur primer memiliki informasi yang dapat diekstrak oleh rantai Markov. Klasifikasi famili protein dapat dilakukan dengan menggunakan PNN atau peluang gabungan. Secara umum, kinerja klasifikasi PNN lebih baik dibandingkan dengan model peluang gabungan.

Nilai-nilai sensitivitas dan spesifisitas pada penelitian ini sebagian besar bernilai lebih dari 0,8. Hal ini menunjukkan bahwa rantai markov memiliki kinerja yang baik sebagai metode ekstraksi fitur. Di samping itu, rantai Markov orde pertama memberikan lebih banyak informasi dalam ekstraksi fitur dibandingkan orde kedua.

5.2 SARAN

Klasifikasi famili protein juga dapat memanfaatkan informasi kimia, fisika dan dikombinasikan dengan informasi statistik untuk metode ekstraksi fitur berikutnya.

DAFTAR PUSTAKA

Andreeva, A., Kulesha, E., Gough, J., & Murzin, A. G. (2020). The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, No. 48, Vol. D1, D376–D382. https://doi.org/10.1093/nar/gkz1064.

Baliarsingh, S.K., Vipsita, S., Gandomi, A.H. (2020). Analysis of high-dimensional genomic data using MapReduce based probabilistic neural network. *Computer Methods and Programs in Biomedicine*, No. 195, Vol. 105625. https: 10.1016/j.cmpb.2020.105625.

Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M. A., & Colwell, L. J. (2019). Using deep learning to annotate the protein universe. *Nature Biotechnology*, No. 40, Vol. 6, 932–937. https://doi.org/10.1038/s41587-021-01179-w.

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, No. 49, Vol. D1, D344–D354. https://doi.org/10.1093/nar/gkaa977.

Chang, D., Ding, L., Malmberg, R., Robinson, D., Wicker, M., Yan, H., Martinez, A., & Cai, L. (2022). Optimal learning of Markov k-tree topology. *Journal of Computational Mathematics and Data Science*, No. 4, Issue May, 100046. https://doi.org/10.1016/j.jcmds.2022.100046.

De Gooijer, J. G., Henter, G. E., & Yuan, A. (2022). Kernel-based hidden Markov conditional densities. *Computational Statistics and Data Analysis*, No. 169, Vol. 107431. https://doi.org/10.1016/j.csda.2022.107431.

Desaire, H., Go, E. P., & Hua, D. (2022). Advances, obstacles, and opportunities for machine learning in proteomics. *Cell Reports Physical Science*, No. 3, Vol. 10, 101069. https://doi.org/10.1016/j.xcrp.2022.101069.

Eskin, E., Grundy, W. N., & Singer, Y. (2000). Protein Family Classification using Sparse Markov Transducers. *Proc Int. Conf Intell Syst Mol Biol*, No. 8, 134-45. PMID: 10977074.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research*, No. 42, Vol. D1, 222–230. https://doi.org/10.1093/nar/gkt1223.

Gupta, C.L.P., Bihari, A., Tripathi, S. (2019). Protein Classification using Machine Learning and Statistical Techniques: A Comparative Analysis. *https://arxiv.org/pdf/1901.06152*.

Hatton, L., & Warr, G. (2015). Protein structure and evolution: Are they constrained globally by a principle derived from information theory? *PLoS ONE*, No. 10, Vol. 5, 1–23. https://doi.org/10.1371/journal.pone.0125663.

Ibrahim, A.A., Yasseen, I.S. (2017). Using Neural Networks to Predict Secondary Structure for Protein Folding. *Journal of Comp and Comm*, Vol. 5, No. 1, 1-8.

Imrie, F., Bradley, A. R., Van Der Schaar, M., & Deane, C. M. (2018). Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. *Journal of Chemical Information and Modeling*, Vol. 58, No. 11, 2319–2330. https://doi.org/10.1021/acs.jcim.8b00350.

Lasfar, M., & Bouden, H. (2018). A method of data mining using Hidden Markov Models (HMMs) for protein secondary structure prediction. *Procedia Computer Science*, No. 127, 42–51. https://doi.org/10.1016/j.procs.2018.01.096.

Mitchell, A., Chang, H. Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S. Y., Bateman, A., Punta, M., Attwood, T. K., Sigrist, C. J. A., Redaschi, N., … Finn, R. D. (2015). The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Research*, Vol. 43, No. D1, D213–D221. https://doi.org/10.1093/nar/gku1243.

Mohebali, B., Tahmassebi, A., Meyer-Baese, A., & Gandomi, A. H. (2020). *Probabilistic neural networks: A brief overview of theory, implementation, and application*. In *Handbook of Probabilistic Models. Elsevier Inc*. https://doi.org/10.1016/B978-0-12-816514-0.00014-X.

Naveenkumar, K.S., Babu, R.M.H., Vinayakumar, R., Soman, K.P. (2018). Protein Family Classification using Deep Learning. BioRxiv, No. 414128. *https://www.biorxiv.org/content/early/2018/09/11/414128.full.pdf+html*

Rao, P. N., Edu, N., Devi, T. U., Kaladhar, D., Sridhar, G. R., & Rao, A. A. (2005). A Probabilistic Neural Network Approach for Protein Superfamily Classification. *JATIT*, Vol. 6, No. 1, 101-105,

Robert, F., & Alexa, M. (2012). Markov Chain Ontology Analysis (MCOA). *BMC Bioinformatics*, No. 13, Issue February, 23–23.

Robinson, S., Guyon, L., Nevalainen, J., Toriseva, M., Åkerfelt, M., & Nees, M. (2015). Segmentation of image data from complex organotypic 3D models of cancer tissues with markov random fields. *PLoS ONE*, Vol. 10, No. 12, 1–26. https://doi.org/10.1371/journal.pone.0143798.

Robinson, S., Nevalainen, J., Pinna, G., Campalans, A., Radicella, J. P., & Guyon, L. (2017). Incorporating interaction networks into the determination of functionally related hit genes in genomic experiments with Markov random fields. *Bioinformatics*, Vol. 33, No. 14, i170–i179. https://doi.org/10.1093/bioinformatics/btx244.

Sandaruwan, P. D., & Wannige, C. T. (2021). An improved deep learning model for hierarchical classification of protein families. *PLoS ONE*, No. 16, Issue October, 1–15. https://doi.org/10.1371/journal.pone.0258625.

Seo, S., Oh, M., Park, Y., & Kim, S. (2018). DeepFam: Deep learning based alignment-free method for protein family modeling and prediction. *Bioinformatics*, Vol. 34, No. 13, i254–i262. https://doi.org/10.1093/bioinformatics/bty275.

Specht, D. F. (1990). Probabilistic Neural Networks. Neural Network, No. 3, Vol. 1, 109–118. https://doi.org/10.1007/s11069-015-1595-z.

Teugels, J. L. (2008). Markov Chains: Models, Algorithms and Applications. *Journal of*

the American Statistical Association, Vol. 103, Issue 483. https://doi.org/10.1198/jasa.2008.s254.

Usotskaya, N., & Ryabko, B. (2009). Application of information-theoretic tests for the analysis of DNA sequences based on Markov chain models. *Computational Statistics and Data Analysis*, No. 53, Vol. 5, 1861–1872. https://doi.org/10.1016/j.csda.2008.07.002.

Wu, C. H., Huang, H., Yeh, L. S. L., & Barker, W. C. (2003). Protein family classification and functional annotation. *Computational Biology and Chemistry*, No. 27, Vol. 1, 37–47. https://doi.org/10.1016/S1476-9271(02)00098-1.

Wu, X., Lü, F., Wang, B., & Cheng, J. (2005). Analysis of DNA sequence pattern using probabilistic neural network model. *Journal of Research and Practice in Information Technology*, No. 37, Vol. 4, 353–362.