Tools Bioinformatika untuk Pemrosesan Data dan Prediksi Fungsi Protein

Bioinformatics Tools for Data Processing and Prediction of Protein Function

Green Arther Sandag[1], Semmy Wellem Taju[2]
[1]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Klabat
e-mail: [1]greensandag@unklab.ac.id

Abstrak

Bioinformatika semakin populer karena kemampuannya untuk menganalisis dan memproses data biologis dengan cepat dan efektif. Bagian penting dari bioinformatika adalah untuk mengidentifikasi fungsi dan karakteristik protein dengan membangun metode prediksi menggunakan algoritma pembelajaran mesin. Ini termasuk bagaimana pembelajaran mesin dapat digunakan untuk menganalisis dan mengklasifikasikan fungsi protein yang cocok untuk digunakan sebagai deteksi penyakit, merancang perawatan medis yang tepat untuk pasien, dan mengembangkan obat untuk beberapa penyakit. Permintaan untuk pembuatan predictive tools dalam menentukan model protein-ligand dan fungsi protein meningkat untuk mempromosikan penelitian biologi dalam lingkungan desain obat yang inovatif. Namun, dibutuhkan banyak waktu dan upaya untuk mengembangkan alat prediksi yang dapat diterapkan pada protein. Dalam penelitian ini kami mengembangkan tools bioinformatika yang dapat secara otomatis mengembalikan data protein dalam bentuk komposisi asam amino (AAC), komposisi pasangan dipeptida (DPC), dan matriks penentuan spesifikasi posisi (PSSM). Data protein, telah kita ambil dari database uniprot yang berisi file fasta. Penelitian ini, kami membuat alat untuk memfasilitasi ilmuwan dalam memproses atau menganalisis data protein dan juga dapat memprediksi fungsi protein menggunakan algoritma pembelajaran mesin seperti Neural Network dan Random Forest.

Kata Kunci—Bionformatika, AAC, DPC, PSSM

Abstract

Bioinformatics increasingly popular due to its ability to analyze and process biological data quickly and effectively. The essential part of bioinformatics is to identify protein function and characteristics by building a prediction method using machine learning algorithms. This include how machine learning can be used to analyze and classification function of proteins that are suitable for use as detection of diseases, designing the right medical treatment for patient, and develop drugs for several diseases. Predictive tools of model protein–ligand binding on demand and protein function are needed to promote biology research in an innovative drug-design environment. However, it takes considerable time and effort to develop predictive tools that can be applied to protein. In this research we developed bioinformatics tools that can automatically restore the protein data in the form of amino acid composition (AAC), dipeptide pair composition (DPC), and position specification scoring matrix (PSSM). The protein data, we have taken from uniprot database that contains fasta file. This work we build tools to facilitate

data scientist or researches in processing or analyze data of protein and also can be predict protein function by machine learning algorithms such as Neural Network and Random Forest.

Keywords—Bioinformatics, AAC, DPC, PSSM

1. PENDAHULUAN

Bioinformatika adalah bidang baru yang menggabungkan teknik komputer, matematika, dan statistik dalam biologi untuk mengambil dan menganalisis data biologis. Hari ini bioinformatika menjadi salah satu bidang yang paling cepat berkembang dalam biologi dan ilmu komputer. Selama ini, bioinformatika masih hanya terfokus pada beberapa tujuan, seperti: untuk menganalisis tiga jenis dataset, yaitu: (1) Urutan genom (urutan baik pengkodean protein, peptida, RNA dan encode beberapa makromolekul lainnya), (2) Struktur makromolekul 3D meliputi protein, DNA, RNA, dan lain-lain, dan (3) fungsi genom, termasuk: pola ekspresi gen tertentu, atau pola interaksi antara protein [1-3], dan lain-lain. Namun, selain berfokus pada tiga jenis dataset itu, analisis bioinformatika juga sering digunakan untuk menganalisis beberapa jenis data lainnya, seperti: membuat taksonomi pohon, memprediksi hubungan protein atau jalur makromolekul lainnya, mengambil dan menganalisis urutan gen, dan mengembangkan basis data. Sampai saat ini, berbagai jenis teknik telah dikembangkan untuk analisis bioinformatika, antara lain: (1) untuk menyelaraskan urutan primer (urutan asam nukleat dalam gen atau urutan asam amino dalam protein) [4], (2) menyelaraskan struktur 3D protein [5], (3) menyelaraskan struktur sekunder dan tersier dari protein dengan protein lain [6], (4) membuat pohon filogenetik [7], (5) untuk memprediksi dan mengklasifikasikan struktur protein [8], (6) untuk memprediksi struktur RNA dan DNA [9], (7) untuk memprediksi fungsi protein [10], dan (8) mengelompokkan gen berdasarkan pola ekspresi mereka [11]. Sehubungan dengan beberapa teknik di atas, ketersediaan algoritma memainkan peran yang sangat penting dalam bioinformatika untuk melakukan validasi dan akurasi. Semua data bank seperti yang disebutkan di atas gratis sehingga kita dapat mengakses data secara langsung melalui web dan di sisi lain, setiap peneliti juga diizinkan untuk menambahkan data mentah yang baru diperoleh dari penelitian mereka. Salah satu bank data yang umum digunakan untuk research di bidang bioinformatika adalah Uniprot, yang berisi sekumpulan rangkaian asam amino penyusun protein yang sebenarnya adalah produk terjemahan dari semua urutan DNA yang disimpan di bank data Uniprot [12] [13]. Sejauh ini, data biologis seperti sekuens DNA atau sekuens protein relatif dominan dibandingkan dengan jenis data biologis lainnya. Data yang telah disimpan dalam database harus ditampilkan secara informatif dan sebisa mungkin harus mudah diakses dan saling berhubungan (melalui tautan) dengan data lain yang disimpan dalam database yang sama atau basis data lainnya. Namun, karena dataset besar yang dihasilkan sehingga memerlukan pendekatan komputasi seperti teknik pembelajaran mesin untuk menganalisis dan menginterpretasikan data yang relevan. Data biologis ini dapat mencakup berbagai informasi dalam kode genetik dan protein.

Teknik pembelajaran mesin telah digunakan secara luas untuk menganalisis data dari berbagai bidang biologi, khususnya, metode pembelajaran mesin telah diterapkan untuk menghasilkan teknik analisis transkripsi dan protein terjemahan dan gen identifikasi yang membawa penyakit [14]. Menurut Yang [15] pembelajaran mesin dapat diterapkan ketika masing-masing sampel telah dijelaskan oleh label kuantitatif. Pembelajaran mesin yang diawasi melibatkan model pelatihan berdasarkan data sampel yang telah diberi label sebagai kelas. Ini berbeda dengan klasifikasi atau pengelompokan tanpa pengawasan. Machine Learning (ML) teknik telah berhasil diterapkan ke berbagai klasifikasi protein yang terkait dengan fungsi protein. Secara khusus, pembelajaran mesin telah terbukti sangat berguna di bidang prediksi struktur protein dan mengarah pada pengembangan sejumlah alat dan aplikasi. Ini termasuk protein prediksi struktur sekunder, protein diskriminasi, komposisi protein, gangguan protein, lipatan protein dan protein model [14].

Secara umum, metode pembelajaran mesin bekerja pada fitur dalam dataset. Pembelajaran mesin mempelajari beberapa hubungan yang bermakna antara elemen-elemen dalam fitur. Fitur utama dari pekerjaan ini adalah urutan protein. Data urutan protein yang telah disimpan dalam database belum diproses, dalam penelitian ini kami mengembangkan alat bioinformatika yang dapat secara otomatis mengembalikan data protein dalam bentuk komposisi asam amino (AAC), komposisi pasangan dipeptida (DPC), dan spesifikasi posisi scoring matrix (PSSM). Urutan protein yang telah kita ambil dari database Uniprot mengandung informasi protein. File ini harus dikonversi menjadi jenis nomor yang berarti bagi para peneliti. Pekerjaan ini kami membangun alat untuk memudahkan ilmuwan data atau penelitian dalam mengolah atau menganalisis data protein dan juga dapat memprediksi fungsi protein

2. METODE PENELITIAN

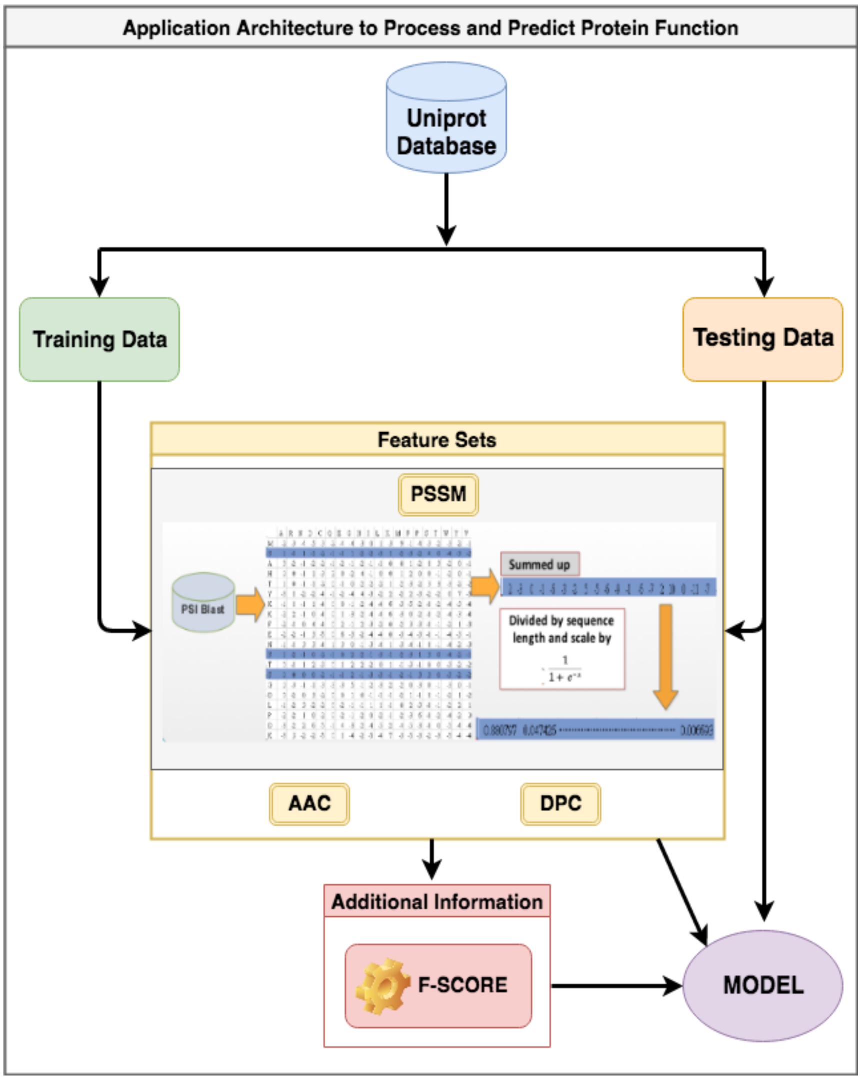
2.1 Classification methods

Supervised machine learning dapat digunakan untuk klasifikasi, model yang dibangun dari data pelatihan, yang mencakup label kelas untuk setiap sampel, dengan menilai nilai atribut. Model ini kemudian digunakan untuk menentukan kelas dari setiap sampel dalam dataset [16]. Kelas dapat memiliki kelompok yang berbeda, seperti kelompok penyakit atau jenis protein. Atribut dapat berupa 20 protein yang diidentifikasi atau fitur baru seperti komposisi asam amino. Klasifikasi dapat digunakan dalam mendiagnosis penyakit, dan untuk klasifikasi protein. Dimungkinkan juga untuk mempertimbangkan atribut khusus sebagai biomarker untuk kelas yang telah ditentukan [17].

Metode klasifikasi terdiri dari dua langkah, yaitu [18]:

a) Langkah pertama, bangun model untuk mendeskripsikan set kelas data yang telah ditentukan. Model dibangun dengan menganalisis tuple yang menggambarkan atribut. Setiap tupel diasumsikan memiliki kelas yang telah ditentukan sebelumnya, sebagaimana ditentukan oleh satu atribut, yang disebut atribut label kelas. Data dianalisis untuk menetapkan model berdasarkan set data pelatihan. Beberapa tupel yang membentuk perangkat pelatihan disebut pelatihan dan sampel kumpulan data yang dipilih secara acak. Karena label kelas setiap sampel pelatihan disediakan, langkah ini dikenal sebagai pembelajaran yang diawasi. Model pembelajaran ini diwakili dalam aturan klasifikasi, pohon keputusan, atau rumus matematika
b) Langkah kedua adalah klasifikasi. Pada langkah ini, classifier yang telah dibangun akan digunakan untuk mengklasifikasikan data. Pertama, keakuratan prediksi pengklasifikasi yang diprediksi. Jika menggunakan satu set pelatihan untuk mengukur keakuratan classifier, maka perkiraan akan optimis karena data yang digunakan untuk membentuk classifier adalah pelatihan yang ditetapkan juga. Oleh karena itu, gunakan set tes, yang merupakan kumpulan tupel bersama dengan label kelas dipilih secara acak dari kumpulan data. Set tes independen dari set pelatihan karena set tes tidak digunakan untuk membangun penggolong

Keakuratan classifier ditentukan oleh oleh testing data yang merupakan persentase set tes yang diklasifikasikan dengan benar oleh classifier. Label kelas dari setiap set tes dibandingkan dengan prediksi dari classifier. Jika keakuratan classifier dapat diterima maka classifier dapat digunakan untuk mengklasifikasikan data baru. Arsitektur aplikasi seperti yang ditunjukkan pada Gambar 1.



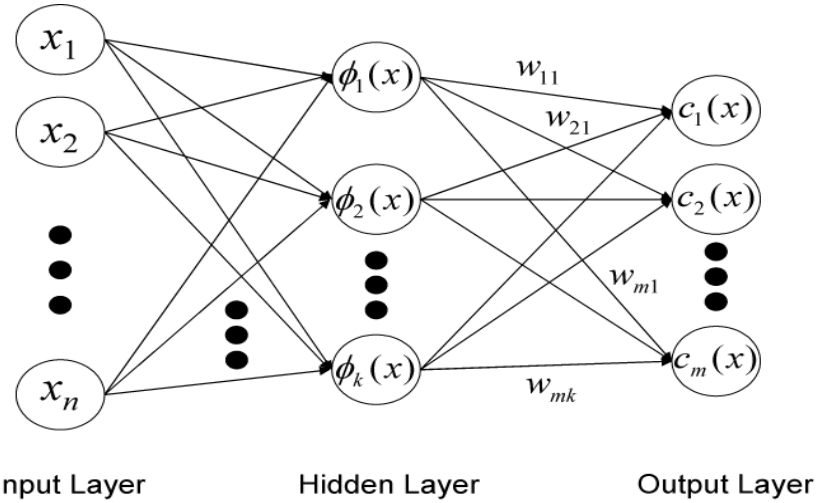
Gambar 1 Arsitektur aplikasi

2.2 Design of the radial basis function netoworks

Dalam penelitian ini, kami menggunakan paket QuickRBF untuk membangun model klasifikasi dengan pengaturan default [19]. Dalam pekerjaan ini, kami menggunakan semua data pelatihan sebagai hidden neuron untuk mendapatkan hasil terbaik dalam mengidentifikasi fungsi protein transport. Klasifikasi berdasarkan *Radial Basis Function (RBF)* memiliki beberapa aplikasi dalam bioinformatika. Telah banyak digunakan untuk memprediksi cleavage sites in proteins [20], inter-residue contacts [21], protein disorder [22], discrimination of b-barrel proteins [23], diskriminasi protein transpor [24] dan sebagainya. Arsitektur RBFN ditunjukkan pada Gambar 1. Dalam penelitian ini, kami menggunakan semua data pelatihan sebagai hidden neuron untuk mendapatkan hasil terbaik.

Bentuk matematis umum dari node output dalam RBFN dinyatakan sebagai berikut:

$$g_j(x) = \sum_{i=1}^{k} w_{ji}\phi(\|x-\mu_i\|;\sigma_i) \quad \cdots\cdots (1)$$



Gambar 2 Arsitektur RBF Network

2.3 Random Forrest

Random Forrest dibangun dari pohon keputusan, di mana beberapa pohon dibangun dari data pelatihan. Setiap pohon hanya memiliki akses ke subkumpulan sampel acak dari atribut. Ketika memprediksi kelas dari set tes, setiap pohon individu memprediksi kelas [24].

2.4. Features Selection Methods

Metode pemilihan fitur adalah peran utama dalam memilih atribut yang signifikan, melalui penghapusan atribut yang redundan atau tidak relevan, dan karena itu dapat juga digunakan untuk identifikasi fitur-fitur penting. Metode ini diterapkan untuk mengidentifikasi protein yang berbeda secara signifikan di antara protein lain yang menentukan protein mana yang dapat menyebabkan penyakit tertentu, baik secara individu atau dalam kombinasi dengan yang lain, dan karena itu bisa menjadi biomarker potensial untuk mengidentifikasi penyakit.

2.4.1 Amino Acid Composition (AAC)

Protein mengandung 20 jenis asam amino yang tidak identik sifatnya, setiap asam amino memiliki kode, yaitu A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V. Gambar 10 menunjukkan metode AAC untuk menghitung proporsi 20 jenis asam amino dalam masing-masing komponen di seluruh rangkaian protein. Rumus untuk menghitung AAC didefinisikan sebagai

$$AAC_{aa,i} = \frac{\Sigma n_{aa,i}}{nres_i} \quad (2)$$

Dimana $\Sigma n_{aa,i}$ adalah jumlah residu asam amino (aa) yang ditemukan dalam protein (i) dan $nres_i$ adalah total jumlah sequence atau residu protein (i)

2.4.2 Dipeptide Pair Composition(DPC)

Dipeptida adalah molekul yang terdiri dari dua asam amino yang disatukan oleh ikatan peptida tunggal atau satu asam amino. Metode ini dapat memperoleh 400 kombinasi dipeptida yang berbeda dari 20 asam amino yang berbeda. Komposisi pasangan dipeptida adalah metode yang penting, dapat memberikan informasi tentang preferensi pasangan residu serta distribusi asam amino [3]. Rumus untuk menghitung DPC didefinisikan sebagai:

$$DPC_{aa,i,j} = \frac{\Sigma n_{aa,i,j}}{N} \quad \text{..... (3)}$$

Dimana, $\Sigma n_{aa,i,j}$ adalah jumlah type residu asam amino (aa) yang bertetangga dengan type residu *j* dan *N* adalah total jumlah residu, *i* dan *j* merujuk kepada asam amino dari 1 sampai 20 dan total kombinasinya akan berjumlah 400.

2.4.3 Position Specific Scoring Matrix (PSSM)

Berdasarkan struktur protein, beberapa residu asam amino dapat bermutasi tanpa mengubah struktur protein, dan dapat membuat dua protein memiliki struktur yang mirip dengan komposisi asam amino yang berbeda [25]. Profil PSSM diperoleh dengan menggunakan database protein PSI-BLAST dan non-redundan (NR) database. Dalam mengidentifikasi fungsi protein dalam transporter, profil PSSM digunakan untuk menghasilkan vektor input 400D sebagai fitur input dengan menjumlahkan setiap baris asam amino dalam profil PSSM. PSSM diperoleh dengan kemungkinan mutasi 20 residu asam amino. Setiap residu akan memiliki 19 kemungkinan ada perubahan atau tetap sama oleh karenanya probabilitas totalnya adalah 400, yang memiliki dimensi 400D. Gambar 1 menunjukkan rincian arsitektur untuk mengidentifikasi fungsi protein transporter dan menghasilkan 400 PSSM fitur dari profil PSSM asli. Setiap elemen dari vektor masukan 400D dibagi oleh panjang urutan dan kemudian discalae atau dinormalisasi menggunakan persamaan berikut:

$$\frac{1}{1+e^{-x}} \quad \text{..... (4)}$$

2.5 F-Score

F-score untuk *i* atribut didefinisikan dengan persamaan berikut [26]:

$$F(i) = \frac{\left(\bar{X}_i^{(+)} - \bar{X}_i\right)^2 + \left(\bar{X}_i^{(-)} - \bar{X}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{X}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{X}_i^{(-)})^2} \quad \text{..... (5)}$$

Dimana $\bar{X}_i$, $\bar{X}_i^{(+)}$, dan $\bar{X}_i^{(-)}$ adalah rata-rata dari *i* atribut untuk positive dan negative dataset pada semua dataset, n_+ adalah jumlah positive dataset dan n_- adalah jumlah negative dataset, $x_{k,i}^{(+)}$ adalah *i* atribut *k* yang merupakan positive dataset, dan $x_{k,i}^{(-)}$ adalah *i* atribut dari *k* yang merupakan negative dataset. Pada penelitian ini, setiap kelas dan family adalah positif dan yang lain negative dataset. F-score adalah salah satu metode penting untuk metode seleksi fitur yang diawasi. Ide utama dari f-score adalah memilih fitur penting yang jarak antara titik data tidak sama.

2.6 Performance Evaluation

Cross-validation adalah prosedur untuk mengevaluasi kinerja model dengan membagi sampel asli ke dalam satu set pelatihan dan menggunakan subset untuk melatih model, dan satu set tes untuk mengevaluasinya [27]. Performance evaluation juga akan menghitung recall, precision, specificity, accuracy, and F-measure.

$$Recall = \frac{TP}{TP + FN} \quad \text{..... (6)}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{.................................... (7)}$$

$$Specificity = \frac{TN}{TN + FP} \quad \text{.................................... (8)}$$

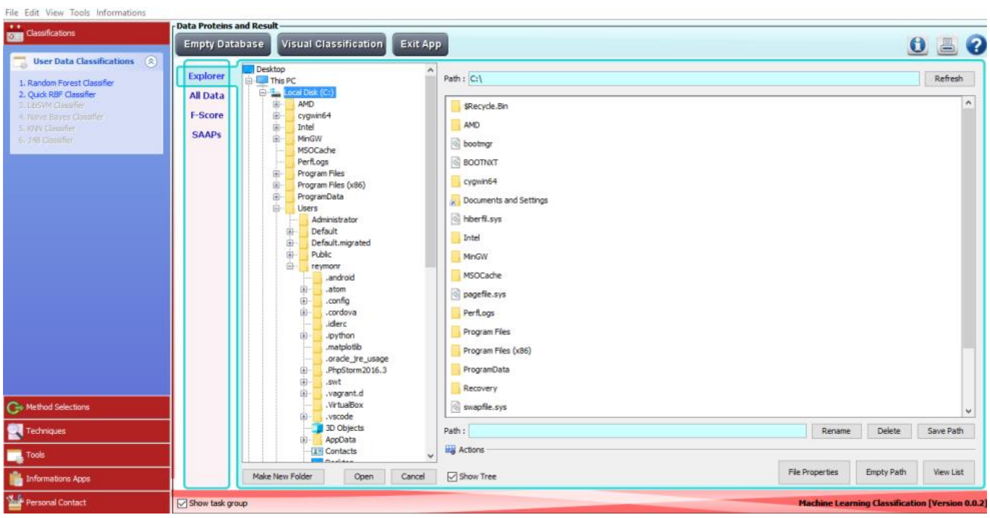
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{.............................. (9)}$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precison + Recall} \quad \text{............................ (10)}$$

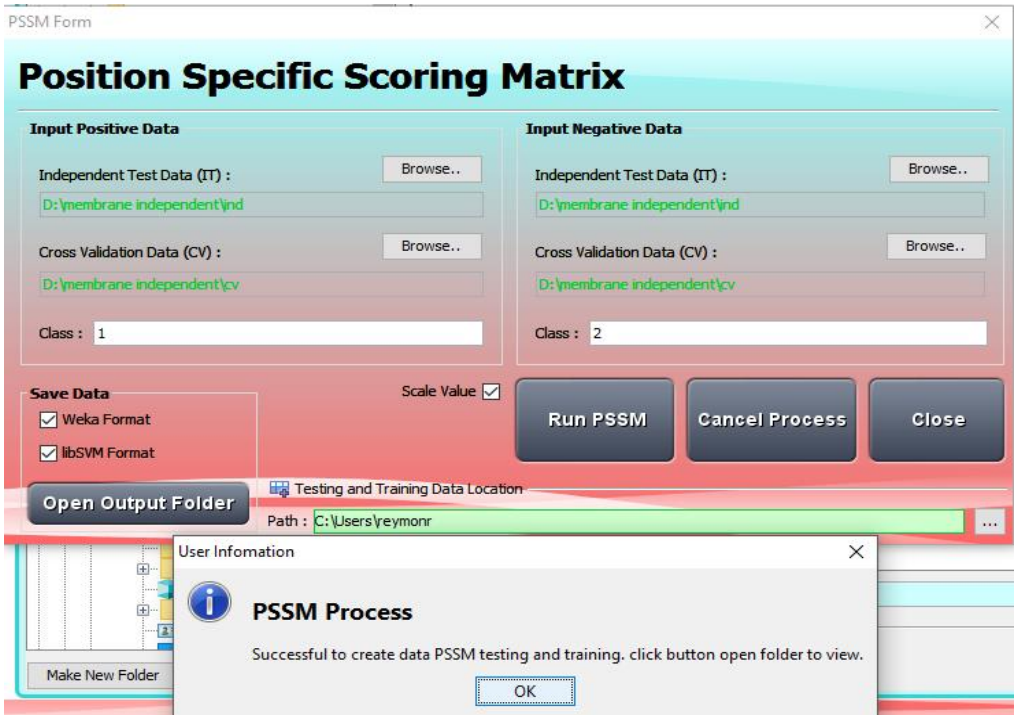
[TP – True Positive; FN – False Negative; TN – True Negative; FP – False Positive]

3. HASIL DAN PEMBAHASAN

Kami mengembangkan tools ini dengan bahasa java dan library weka. Pada tahap pertama harus menyediakan data protein yang sudah tersedia di berbagai database, salah satunya adalah uniprot database. Setelah pengguna mengunduh data protein, data harus diproses terlebih dahulu. Data biasanya terdiri dari file fasta. File fasta yang berisi nama-nama protein dan urutan protein. Data dipisahkan menjadi data pelatihan dan data pengujian dan kemudian pengguna dapat menentukan atribut kelas. Gambar 3 adalah antarmuka alat bioinformatika untuk memprediksi fungsi protein.

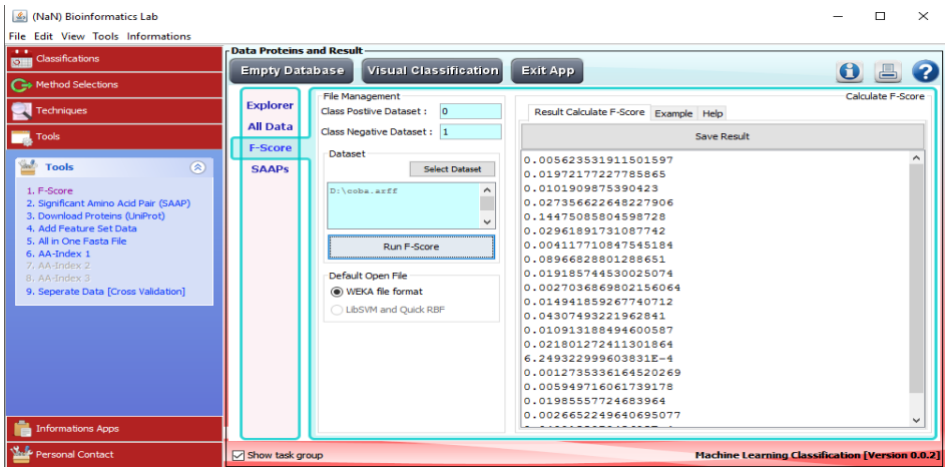


Gambar 3 Antarmuka tools bioinformatika untuk prediksi protein.



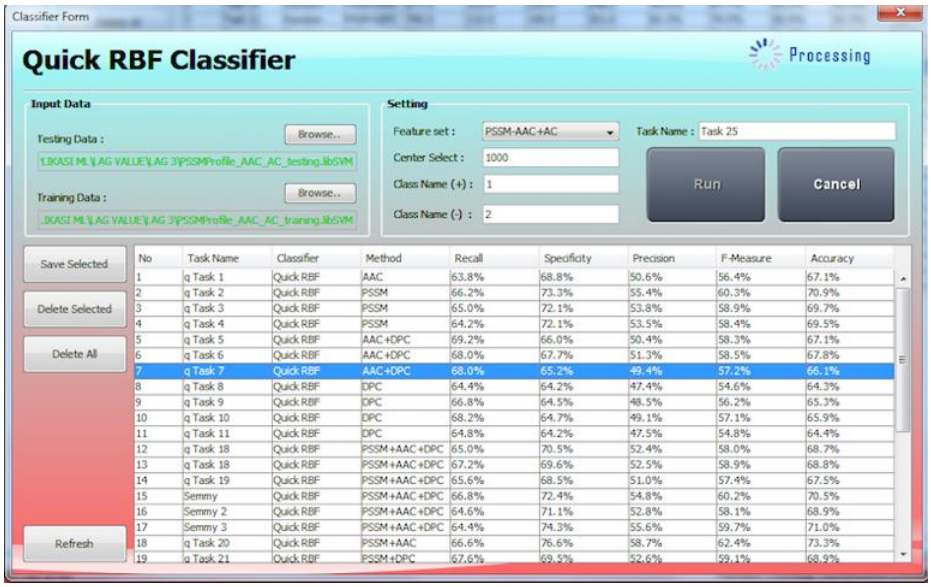
Gambar 4 Antarmuka untuk PSSM method

Dalam studi ini, kami menyediakan tiga metode untuk memproses data protein. Metodenya adalah komposisi asam amino (AAC), komposisi pasangan asam amino (DPC) dan position specification scoring matrix (PSSM). Pengguna dapat memilih metode yang akan digunakan, dan data dimasukkan ke dalam aplikasi. Proses input data adalah dua bagian, input pertama dataset positif dan masukan dataset negatif. Pengguna dapat memasukkan dataset uji independen (data pengujian) dan data validasi silang (data pelatihan). Kemudian masukkan kelas sesuai dengan pengguna. Untuk menyimpan hasil pengguna dapat menyimpan data ke dalam file weka atau libsvm. Antarmuka metode PSSM dalam aplikasi ini seperti pada Gambar 4. Aplikasi ini juga menyediakan fitur untuk menghitung F-Score dari data yang telah dibuat. Antarmuka seperti pada gambar 5. Setelah pengguna memproses data maka aplikasi akan mengembalikan ke dalam dua format file, weka (ARFF) dan libsvm. Untuk menghitung f-score, kami menggunakan file arff.



Gambar 5 Antarmuka untuk menghitung F-Score

Gambar 6 menunjukkan antarmuka untuk memprediksi dan mengklasifikasikan fungsi protein menggunakan algoritma pembelajaran mesin. Dalam studi ini, kami menyediakan dua metode yang dapat digunakan, Random forest and Neural network (quickrbf). Ketika aplikasi dijalankan, aplikasi akan menampilkan evaluasi kinerja sebagai recall, presisi, spesifisitas, akurasi, dan F-measure. Hasil ini dapat digunakan oleh peneliti untuk menganalisis data protein.



Gambar 6 Antarmuka menu prediksi dan klasifikasi protein menggunakan Quick RBF classifier

4. KESIMPULAN

Tools bioinformatika ini untuk pemrosesan data dan prediksi fungsi protein membantu para peneliti dalam memproses data mentah protein serta menentukan fungsi protein yang dapat diterapkan dalam metode pengobatan penyakit, deteksi penyakit, dan obat yang dikembangkan untuk beberapa penyakit. Aplikasi ini secara otomatis mengubah file fasta menjadi beberapa format seperti weka dan file libsvm. Ini membantu peneliti atau pengguna untuk menguji berbagai algoritme pembelajaran mesin yang tersedia di Weka dan disediakan di aplikasi ini. Hasil prediksi ini dapat digunakan sebagai referensi dalam penelitian selanjutnya.

5. SARAN

Aplikasi ini dapat berjalan di berbagai platform yang mendukung Java. Oleh karena itu penelitian ini dapat dikembangkan lagi untuk platform pemrograman yang lain yang lebih cepat. Penelitian ini dapat dikembangkan lebih lanjut dengan menambahkan algoritma pembelajaran mesin lainnya seperti naïve Bayes, decision tree, svm, dll., Serta menambahkan metode pemilihan fitur lainnya seperti, asam amino signifikan dan indeks pasangan asam amino.

DAFTAR PUSTAKA

[1] M. Pop, and S.L. Salzberg, *Bioinformatics challenges of new sequencing technology*. Trends in Genetics, 2008. **24**(3): p. 142-149.

[2] Martí-Renom, M.A., et al., "Comparative protein structure modeling of genes and genomes," *Annual review of biophysics and biomolecular structure*, 2000. **29**(1): p. 291-325.

[3] M. B. Eisen , et al., "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, 1998. **95**(25): p. 14863-14868.

[4] R. Wernersson, and A.G. Pedersen, "RevTrans: multiple alignment of coding DNA from aligned amino acid sequences," *Nucleic acids research*, 2003. **31**(13): p. 3537-3539.

[5] L. Holm, and C. Sander, "Protein structure comparison by alignment of distance matrices" *Journal of molecular biology*, 1993. **233**(1): p. 123-138.

[6] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of molecular biology*, 1999. **292**(2): p. 195-202.

[7] S. Li, D.K. Pearl, and H. Doss, "Phylogenetic tree construction using Markov chain Monte Carlo," *Journal of the American Statistical Association*, 2000. **95**(450): p. 493-508.

[8] D. T. -H. Chang, et al., "Prediction of protein secondary structures with a novel kernel density estimation based classifier," *BMC research notes*, 2008. **1**(1): p. 51.

[9] M. Zuker, D.H. Mathews, and D.H. Turner, "*Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide*, in *RNA biochemistry and biotechnology,"* 1999, Springer. p. 11-43.

[10] H. N. Chua, W.-K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions," *Bioinformatics*, 2006. **22**(13): p. 1623-1630.

[11] M. P. Brown, et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences*, 2000. **97**(1): p. 262-267.

[12] H. S. Bilofsky, and B. Christian, *The GenBank® genetic sequence data bank*. Nucleic acids research, 1988. **16**(5): p. 1861-1863.

[13] U. Consortium, *Reorganizing the protein space at the Universal Protein Resource (UniProt)*. Nucleic acids research, 2011: p. gkr981.

[14] J. H. Moore, F.W. Asselbergs, and S.M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, 2010. **26**(4): p. 445-455.

[15] S. Kotsiantis, "Feature selection for machine learning classification problems: a recent overview," *Artificial Intelligence Review*, 2011: p. 1-20.

[16] A. L. Swan, et al., "Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology," *Omics: a journal of integrative biology*, 2013. **17**(12): p. 595-610.

[17] J. Yang, et al., Image super-resolution via sparse representation, *IEEE transactions on image processing*, 2010. **19**(11): p. 2861-2873.

[18] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. 2011: John Wiley & Sons.

[19] Y.-Y. Ou, *QuickRBF: an efficient RBFN package*. software available at : http://csie/.org/~ yien/quickrbf/quickstart. php, 2005.

[20] Z. R. Yang, and R. Thomson, "Bio-basis function neural network for prediction of protease cleavage sites in proteins," *IEEE Transactions on Neural Networks*, 2005. **16**(1): p. 263-274.

[21] G.-Z. Zhang, and D.-S. Huang, "Prediction of inter-residue contacts map based on genetic algorithm optimized radial basis function neural network and binary input encoding scheme," *Journal of computer-aided molecular design*, 2004. **18**(12): p. 797-810.

[22] C.-T. Su, C.-Y. Chen, and Y.-Y. Ou, "Protein disorder prediction by condensed PSSM considering propensity for order or disorder," *Bmc Bioinformatics*, 2006. **7**(1): p. 319.

[23] Y.-Y. Ou, et al., "TMBETADISC-RBF: discrimination of-barrel membrane proteins using RBF networks and PSSM profiles," *Computational biology and chemistry*, 2008. **32**(3): p. 227-231.

[24] N.Q.K. Le, G. A. Sandag, and Y.-Y. Ou. "Incorporating post translational modification information for enhancing the predictive performance of membrane transport proteins," *Computational biology and chemistry* 77 (2018): 251-260.

[25] L. Breiman, *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

[26] S.-A. Chen, et al., "Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties," *Bioinformatics*, 2011. **27**(15): p. 2062-2067.

[27] Y.-W. Chen, and C.-J. Lin, *Combining SVMs with various feature selection strategies*, in *Feature extraction*. 2006, Springer. p. 315-324.

[28] G. Zhang, et al., "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," *European journal of operational research*, 1999. **116**(1): p. 16-32.