Pemodelan QSAR untuk Prediksi pIC50 Inhibitor COX-2 dengan Memanfaatkan Avalon Deskriptor

QSAR Modeling for Prediction of pIC50 COX-2 Inhibitors by Utilizing Avalon Descriptors

Teguh Arif Mulyana[1)], Dhea Sukma Agustiana[2)], Hanna Septiani[3)], Aldy Samuel[4)], Muhammad Dhoni Apriyadi[5)].

Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera

Email: teguh.120450075@student.itera.ac.id [1)], dhea.120450035@student.itera.ac.id [2)], hanna.120450064@student.itera.ac.id [3)], aldy.120450049@student.itera.ac.id [4)], muhammad.120450111@student.itera.ac.id [5)].

Abstrak

Penelitian ini memfokuskan pada pemanfaatan Avalon Deskriptor dalam memodelkan Quantitative Structure-Activity Relationship (QSAR) untuk memprediksi pIC50 inhibitor COX-2, enzim yang terlibat dalam peradangan. Menggunakan dataset ChEMBL dan Random Forest, penelitian ini mengidentifikasi fitur-fitur penting, seperti Col_A_1026, Col_A_8, Col_A_3991, Col_A_4052, Col_A_3553, Col_A_2462, Col_A_2721, Col_A_1946, dan Col_A_3014, melalui analisis Tanimoto coefficient. Hasil prediksi model QSAR menunjukkan kinerja yang memadai, dengan metode Shuffle Split dan Train-Test Split memberikan performa serupa (r2=0.58-0.59, mae=0.62, rmse=0.86-0.88). Kesimpulannya, penelitian ini memberikan wawasan mendalam tentang hubungan struktur-molekul dan aktivitas biologis inhibitor COX-2, membuka peluang untuk pengembangan obat antiinflamasi yang lebih efektif dan selektif.

Kata kunci: Model *QSAR, Random Forest, pIC50, Avalon*

Abstract

This study focuses on utilizing Avalon Descriptors in modeling Quantitative Structure-Activity Relationship (QSAR) to predict the pIC50 of COX-2 inhibitors, an enzyme involved in inflammation. Using ChEMBL and Random Forest datasets, this study identified important features, such as Col_A_1026, Col_A_8, Col_A_3991, Col_A_4052, Col_A_3553, Col_A_2462, Col_A_2721, Col_A_1946, and Col_A_3014, through Tanimoto coefficient analysis. QSAR model prediction results showed adequate performance, with the Shuffle Split and Train-Test Split methods providing similar performance (r2=0.58-0.59, mae=0.62, rmse=0.86-0.88). In conclusion, this study provides deep insight into the structure-molecule relationship and biological activity of COX-2 inhibitors, opening up opportunities for the development of more effective and selective anti-inflammatory drugs.

Keywords: QSAR Model, Random Forest Regression Model, pIC50

PENDAHULUAN

Pemodelan QSAR (Quantitative Structure-Activity Relationship) telah menjadi pendekatan yang kritis dalam pengembangan obat, memungkinkan pemahaman yang mendalam tentang hubungan antara struktur molekul dan aktivitas biologis suatu senyawa. Dalam konteks ini, fokus penelitian terarah pada prediksi pIC50 sebagai parameter kuantitatif untuk aktivitas inhibitor COX-2, enzim yang berperan dalam proses peradangan. Kajian literatur menyoroti pentingnya deskriptor

molekuler sebagai representasi struktur yang akurat dan informatif. Dalam upaya ini, Avalon Deskriptor dipilih sebagai dasar untuk pemodelan, mengingat keunggulannya dalam menggambarkan sifat fisikokimia, struktur, dan aktivitas biologis suatu senyawa. Keterlibatan Avalon Deskriptor dalam QSAR membuka peluang untuk mengidentifikasi pola dan hubungan yang mendasari aktivitas inhibitor COX-2.

Meskipun penelitian-penelitian sebelumnya telah mencoba pemodelan QSAR untuk berbagai senyawa, penelitian ini memberikan kebaruan dengan mengeksplorasi Avalon Deskriptor secara khusus untuk memprediksi pIC50 inhibitor COX-2. Pemahaman lebih lanjut tentang keterkaitan struktur-molekul dan aktivitas biologis ini diharapkan dapat memberikan kontribusi pada pengembangan obat antiinflamasi yang lebih selektif dan efektif. Oleh karena itu, tujuan penelitian ini adalah untuk memodelkan QSAR dengan menggunakan Avalon Deskriptor guna meningkatkan pemahaman tentang aktivitas inhibitor COX-2 dan membuka jalan bagi penemuan obat baru yang lebih terarah.

METODE

Dataset

ChEMBL (Chemical Biology Database of the European Bioinformatics Institute) adalah basis data yang dipilih dengan cermat dari senyawa aktif biologis untuk penggunaan obat[1]. ChEMBL menyediakan kumpulan data besar yang tidak berlebihan dari 7504 senyawa dengan nilai IC50 yang dipublikasikan terhadap Cox-2, yang digunakan dalam analisis hubungan aktivitas struktur kuantitatif (QSAR) untuk mempelajari lebih lanjut tentang asal-usul bioaktivitas mereka. Penghambat Cox-2 dijelaskan menggunakan satu set deskriptor sidik jari, dan model prediksi random forest[2].

Setelah itu, jumlah sidik jari substruktur diperiksa secara menyeluruh untuk mendapatkan informasi yang berguna tentang aktivitas penghambatan inhibitor AChE. Informasi ini dapat diterapkan dalam berbagai cara. Tujuan makalah kami adalah untuk memprediksi nilai plC50 menggunakan notasi SMILES dan ID ChEMBL sebagai input.

Penelitian ini menggunakan ChEMBL dapat membantu dalam identifikasi potensial obat baru, memahami mekanisme aksi senyawa, dan mendukung penelitian lebih lanjut di bidang farmakologi dan biologi molekuler. Ini menggabungkan data kimia, data aktivitas biologis, dan data genetik untuk membantu mengubah data genom menjadi obat baru yang efektif. Kumpulan penghambat Cox-2 manusia (ID subjek CHEMBL.230) dikumpulkan menggunakan basis data ChEMBL 230 yang berisi 4733 senyawa. Sebanyak 4726 senyum kanonik unik diperoleh setelah menghapus senyum kanonik yang memiliki nilai pIC50 *infinite*[3].

Pembersihan dan pembuatan model dilakukan dengan menggunakan Google Collaboratory.

Shuffle split + train test split ratio(70 30)

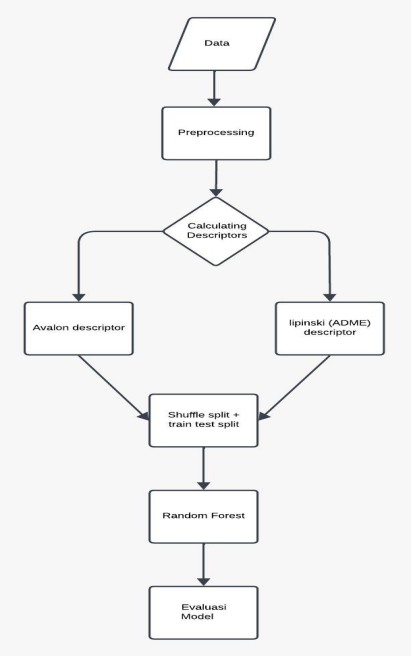
Shuffle split merujuk pada metode dalam pemrograman Python, khususnya library scikit-learn, yang digunakan untuk validasi silang acak. Metode ini menghasilkan indeks untuk membagi data menjadi set test dan train. Shuffle split melakukan iterasi pengacakan dan pemisahan. Pada dataset ini penulis membagi data test dan train menjadi 70:30 dengan menggunakan metode shuffle split[4].

Deskripsi Inhibitor

Inhibitor COX-2 merujuk kepada senyawa yang memiliki kemampuan menghambat aktivitas enzim COX-2[5]. Enzim COX-2 memainkan peran penting dalam proses sintesis prostaglandin, suatu substansi yang menjadi penyebab peradangan dan rasa nyeri dalam berbagai kondisi

penyakit, seperti rheumatoid arthritis, osteoarthritis, migrain, atau nyeri haid. Penggunaan inhibitor untuk COX-2 dapat diarahkan sebagai obat antiinflamasi nonsteroid (OAINS) yang lebih selektif dan aman dibandingkan dengan OAINS konvensional yang mempengaruhi baik enzim COX-1 maupun COX-2..

Deskriptor Avalon, pada konteks ini, merujuk pada parameter numerik yang mencerminkan sifat fisikokimia, struktur, dan aktivitas biologis suatu senyawa. Avalon Deskriptor, dihasilkan melalui perangkat lunak Avalon yang berbasis pada representasi grafis struktur molekul, dapat diterapkan dalam pembentukan model QSAR (Quantitative Structure-Activity Relationship). Model ini berfungsi untuk memprediksi nilai aktivitas/pIC50 suatu senyawa terhadap enzim COX-2. Aktivitas/pIC50 disini adalah ukuran kekuatan penghambatan suatu senyawa terhadap enzim COX-2, dengan nilai yang semakin tinggi menandakan tingkat kekuatan yang lebih besar sebagai inhibitor. Model QSAR yang menggunakan Avalon Deskriptor memberikan informasi tentang keterkaitan struktur molekul dengan aktivitas biologisnya, dan dapat digunakan untuk melakukan penapisan virtual terhadap senyawa-senyawa baru yang memiliki potensi sebagai inhibitor untuk COX-2.



Gambar 1. Diagram Alir

Tanimoto Koefisien

Tanimoto koefisien, juga dikenal sebagai Tanimoto similarity atau Jaccard coefficient, adalah ukuran kemiripan antara dua himpunan, sering digunakan dalam konteks struktur molekuler atau fingerprint. Koefisien ini mengukur seberapa besar elemen-elemen yang sama atau serupa antara dua himpunan terhadap total elemen yang unik dalam himpunan tersebut. Dalam konteks kimia dan bioinformatika, Tanimoto coefficient sering digunakan untuk membandingkan fingerprint molekuler, yang merepresentasikan kehadiran atau ketidakhadiran fitur-fitur spesifik dalam struktur molekuler. Nilai Tanimoto coefficient berkisar antara 0 (tidak ada kemiripan) hingga 1 (kemiripan sempurna). Penggunaan Tanimoto coefficient memungkinkan evaluasi tingkat kemiripan struktural antara senyawa-senyawa, menjadi alat penting dalam pengelompokan, screening virtual, dan pemilihan senyawa berdasarkan kemiripan struktural[6].

Analisis Multivariat

Supervised training adalah proses melatih model pada data pelatihan berlabel,

yang dapat digunakan untuk memprediksi data yang tidak diketahui atau data yang akan datang. Penelitian ini membangun model regresi yang memprediksi (pIC5O) sebagai variabel respons kontinu (yaitu, fungsi dari variabel prediktor (deskriptor sidik jari).

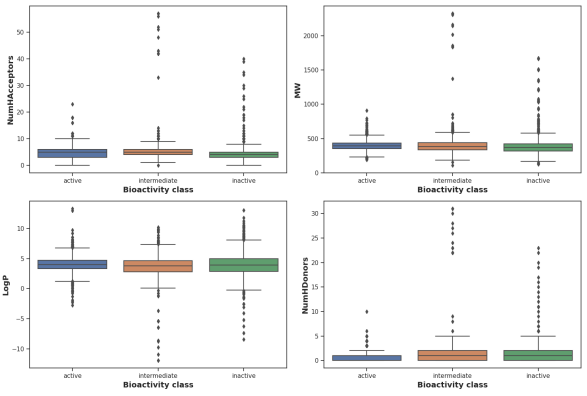
Penelitian ini membangun model regresi yang memprediksi variabel respons kontinu (plC5O) sebagai fungsi dari variabel prediktor (deskriptor sidik jari).

Pengklasifikasi random forest (RF) adalah pengklasifikasi ensembel yang terdiri dari beberapa pohon keputusan. Sederhananya, ide utama dari RF adalah bahwa alih-alih membangun pohon keputusan yang dalam dengan jumlah node yang terus meningkat yang mungkin rentan terhadap overfitting data dan overtraining, RF menciptakan beberapa pohon untuk meminimalkan varians daripada memaksimalkan akurasi[7].

HASIL DAN PEMBAHASAN

Analisis Ruang Kimia Inhibitor Cox-2

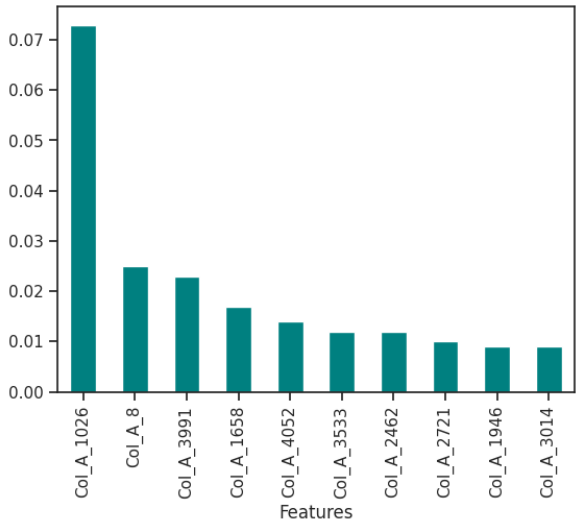
Analisis ruang kimia inhibitor Cox-2 dilakukan untuk mendapatkan pemahaman mengenai korelasi struktur-aktivitas dengan memeriksa deskriptor aturan lima Lipinski. Pendekatan ini memungkinkan penilaian karakteristik umum senyawa yang mengatur sifat penghambatan senyawa. Data eksplorasi dianalisis dengan menggunakan deskriptor aturan lima Lipinski, termasuk MW (ukuran molekul), LogP (logaritma partisi oktanol-air), NumHDonors (jumlah donor ikatan hidrogen), dan NumHAcceptors (jumlah akseptor ikatan hidrogen). MW mencerminkan dimensi molekul yang sering digunakan karena dapat dihitung dan diinterpretasikan dengan mudah, sementara LogP mengukur lipofilisitas senyawa dan berguna dalam memahami penetrasi membran. NumHDonors dan NumHAcceptors digunakan untuk mengukur kemampuan ikatan hidrogen. Visualisasi ruang kimia LogP sebagai fungsi MW menunjukkan distribusi inhibitor dalam rentang MW (molecular weight) sekitar 300-600 Da dan LogP sekitar -2,5 hingga 2,5. Senyawa dengan nilai LogP mendekati nol kemungkinan besar adalah inhibitor tidak aktif, sementara inhibitor aktif cenderung memiliki nilai LogP yang lebih rendah. Plot kotak deskriptor Lipinski menunjukkan perbedaan signifikan dalam nilai LogP dan MW antara kelas bioaktivitas, dengan senyawa aktif memiliki nilai LogP terendah dan MW terbesar, sementara senyawa tidak aktif memiliki nilai yang lebih kecil. Analisis lebih lanjut menunjukkan bahwa perbedaan terbesar terletak pada nilai LogP untuk senyawa aktif, sementara perbedaan lainnya dapat diabaikan. Berikut adalah visualisasi box-plot Lipinski Deskriptor.



Gambar 2. Cox-2 Lipinski Deskriptor

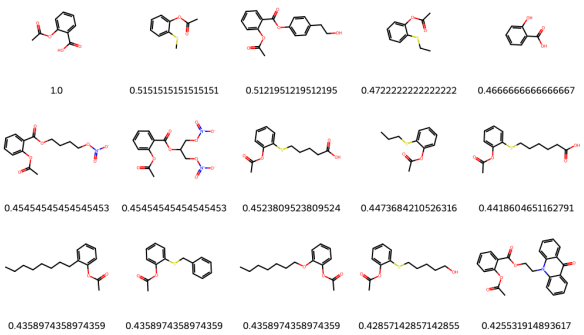
Fitur Penting dan Nilai Tanimoto

Penelitian ini akan menelusuri elemen-elemen kunci yang memiliki dampak signifikan dalam proses prediksi aktivitas inhibitor COX-2. Analisis mendalam terhadap fitur-fitur yang dianggap berperan penting akan memberikan wawasan yang lebih baik terkait hubungan antara struktur molekul dan aktivitas biologis, dengan tujuan mengoptimalkan model prediksi pIC50 pada enzim COX-2.



Gambar 3. Top 10 Fitur Penting

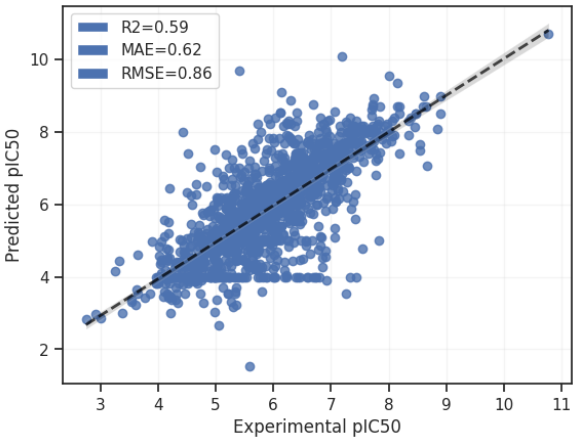
Penelusuran nilai Tanimoto sebagai alat untuk mengukur kemiripan antar-senyawa dalam dataset akan memberikan pemahaman mendalam tentang hubungan struktural yang dapat memengaruhi prediksi aktivitas inhibitor COX-2. Analisis nilai Tanimoto diharapkan memberikan perspektif yang lebih luas terkait kesamaan dan variasi struktural dalam dataset, memberikan dukungan yang diperlukan untuk interpretasi hasil prediksi. Senyawa paling mirip dengan senyawa Aspirin adalah sebagai berikut.



Gambar 4. Top 15 Senyawa Paling Mirip

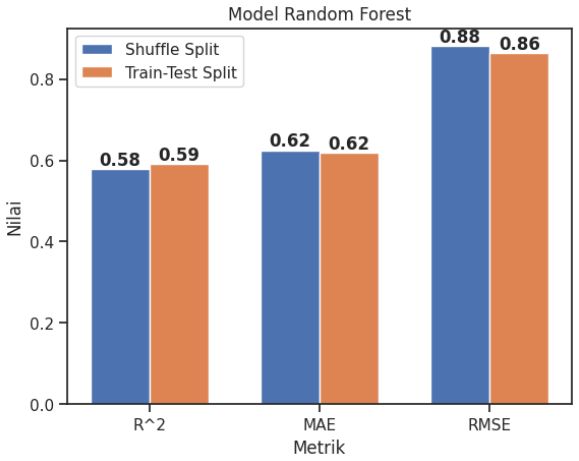
Prediksi pIC50 Model QSAR

Dataset terdiri dari 4726 senyawa digunakan untuk membangun model QSAR. Deskriptor sidik jari Avalon digunakan untuk menerjemahkan struktur molekuler ke fitur numerik yang mencerminkan sifat fisiko-kimia penting molekul. Kemudian model dibangun menggunakan rasio pemisahan data 70/30, dimana 70% kumpulan data digunakan sebagai data latih dan 20% data sebagai data uji, Dimana 4096 data independen digunakan untuk konstruksi model. Hasil Prediksi pIC50 ditunjukkan pada gambar 5.



Gambar 5. Eksperimental vs Prediksi pIC50

Adapun perbedaan nilai metrik antara metode pemisahan data menggunakan shuffle split dan train_test_split ditunjukkan pada gambar 6.



Gambar 6. Perbandingan Metrik

KESIMPULAN

Dalam mengevaluasi model QSAR untuk prediksi pIC50 inhibitor COX-2 dengan memanfaatkan Avalon Deskriptor, hasil analisis menyoroti beberapa fitur yang terbukti paling penting dalam mempengaruhi aktivitas biologis. Fitur-fitur tersebut

melibatkan Col_A_1026, Col_A_8, Col_A_3991, Col_A_4052, Col_A_3553, Col_A_2462, Col_A_2721, Col_A_1946, dan Col_A_3014, yang diidentifikasi sebagai variabel yang paling berpengaruh dalam pembentukan model. Metrik evaluasi model menggunakan metode Shuffle Split dan Train-Test Split secara berturut-turut menunjukkan hasil yang memadai, dengan nilai r2 mencapai 0.58 dan 0.59, mae sebesar 0.62, dan rmse sekitar 0.88 dan 0.86. Hasil ini menunjukkan bahwa model memiliki kemampuan yang relatif baik dalam meramalkan aktivitas inhibitor COX-2, dengan metode Shuffle Split dan Train-Test Split memberikan performa yang serupa.

Kesimpulannya, model QSAR yang dikembangkan dengan memanfaatkan Avalon Deskriptor dan fitur-fitur penting tersebut memberikan pemahaman yang signifikan tentang hubungan struktur-molekul dan aktivitas biologis senyawa, membuka peluang untuk pengembangan obat antiinflamasi yang lebih efektif dan selektif.

DAFTAR RUJUKAN

[1] "Exploring ChEMBL Data with the new ChEMBL Interface", doi: 10.6019/tol.chembl-w.2020.00001.1.

[2] R. E. Harris, *Inflammation in the Pathogenesis of Chronic Diseases: The COX-2 Controversy*. Springer Science & Business Media, 2007.

[3] *Progress in Medicinal Chemistry*. Elsevier, 2009.

[4] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. "O'Reilly Media, Inc.," 2016.

[5] R. A. Copeland, *Evaluation of Enzyme Inhibitors in Drug Discovery: A Guide for Medicinal Chemists and Pharmacologists*. John Wiley & Sons, 2013.

[6] A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*. Springer, 2007.

[7] Yu. L. Pavlov, *Random Forests*. Walter de Gruyter GmbH & Co KG, 2019.