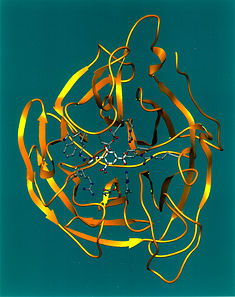
**Drug Discovery: nram — Neuraminidase**

Oleh: Andhika Prima Prasetyo

**1. Latar Belakang**



Enzim neuraminidase (nram) adalah enzim glikosida hidrolase (EC 3.2.1.18) yang membelah ikatan glikosidik asam neuraminik. Enzim neuraminidase adalah kelompok enzim yang dapat ditemui dalam berbagai organisme. Neuraminidase yang paling dikenal adalah neuraminidase viral yang disasar oleh obat-obatan untuk mencegah penyebaran infeksi influenza. Neuraminidase viral sering digunakan sebagai determinan antigenik yang dapat ditemukan di permukaan virus influenza. Beberapa varian neuraminidase influensa dianggap lebih virulen terhadap virus daripada yang lain. Homolog yang lain dapat ditemui dalam sel-sel mamalia yang memiliki berbagai fungsi. Setidaknya empat homolog sialidase mamalia terdapat di dalam genom manusia (Gambar 1).

Gambar 1. Neuraminidase diagram (Wikipedia, 2020)

Project ini memiliki beberapa tujuan, yakni (1) mencari metode machine learning terbaik untuk mensimulasi molecule yang berinteraksi dengan enzime neuraminidase; (2) optimasi metode terbaik dengan memodifikasi hyper paramter untuk memberikan hasil prediksi yang lebih baik; dan (3) screening molekul drug yang saat ini berada dipasaran untuk melihat hasil prediksi model.

**2. Metode**

Data molekul target neuraminidase diunduh dari <http://dude.docking.org/targets/nram> dan kemudian di ektrak molekul yang aktif dan non-aktif. Selanjutnya molekul molekul tersebut di visualisasi menggunakan RDKit (<https://www.rdkit.org/>). Selanjutnya molekul aktif and decoy masing-masing diambil 50 dan 100 molekul untuk mempermudah analisis. Selanjutnya molekul-molekul tersebut di ektrasi featurenya dengan menggunakan modul PyFingerprint (<https://github.com/hcji/PyFingerprint>). Dua metode ektrasi diujikan terhadap molekul tersebut, yakni pubchem dan Klekota-Roth. Output pubchem selanjutnya digunakan dengan pertimbangan beban analisis dan hasil analisis yang lebih variatif. Sebelum dianalisis dengan machine learning, terlebih dahulu data fiture dicek keseragaman nama kolom dan jumlah kolom.

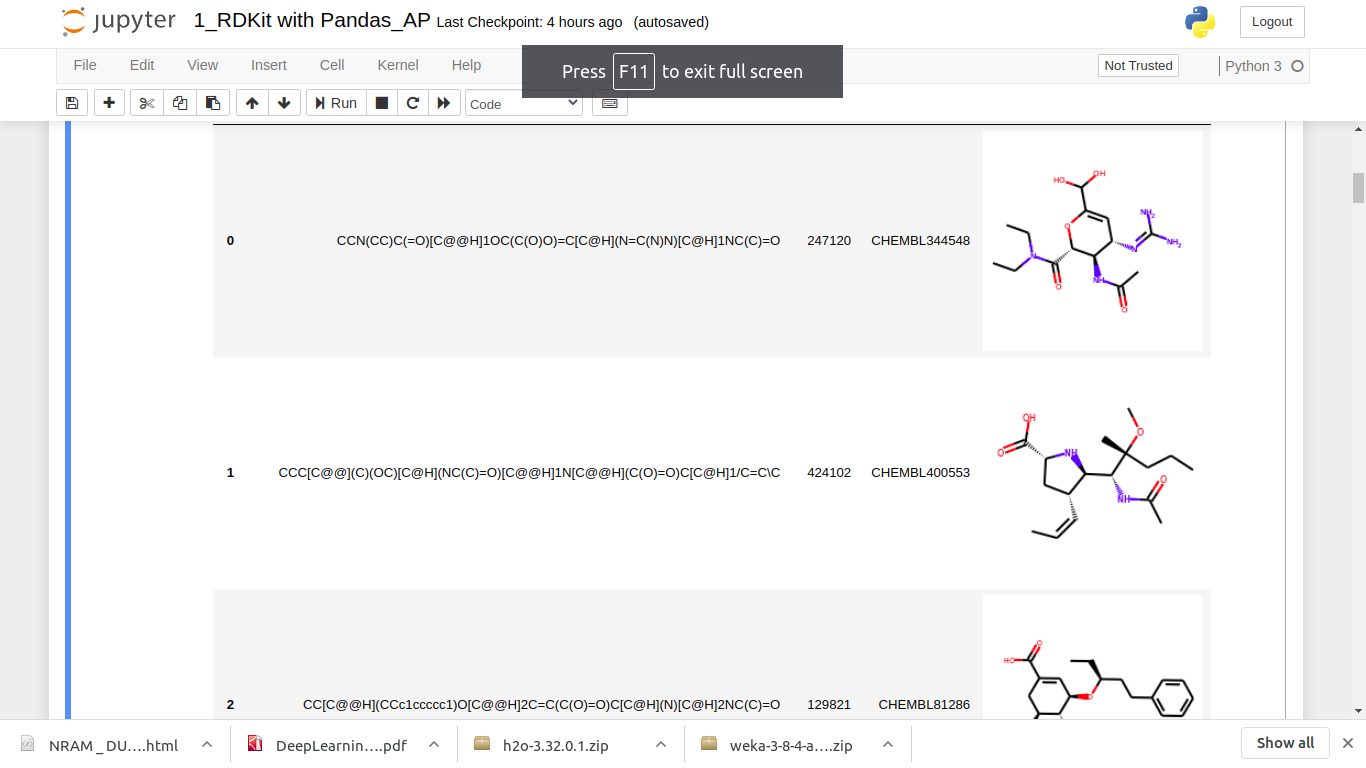
Tiga metode machine learning diujicobakan, yakni logistic regression, support vector machine (SVM) dan random forest dengan modul sklearn (<https://scikit-learn.org/stable/index.html>). Pemilihan metode terbaik didasarkan pada nilai akurasi, sensitifitas, specitifity dan presisi serta dibandingkan nilai AUC dari ROC diagram. Jumlah training dan test data, masing-masing diambil 60% dan 40%.

Model selnajutnya dioptimasi dengan menggunakan metode kFold cross validation dengan jumlah split 5 dan range 25. Model tersebut akan diiterasi nilai C dan tolerasinya kemudian dipilih paramater terbaik untuk digunakan sebagai model prediksi. Hasil akurasi Kfold dan spliting manual akan dibandingkan. Kedua model standar dan optimasi di simpan untuk digunakan kembali.

Screening dilakukan untuk dua molekul yang berpotensi menjadi inhibitor enzim neuraminidase yakni oseltamivir (<https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:7798>) dan zanamivir (<https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:50663>) serta du molekul lain dari database DUD.E yakni chembl293591 dan chembl344548. Selanjutnya kedua molekul tersebut di extrak fiturnya dengan menggunakan PyFingerprint dan disimpan untuk menjadi data test. Model standar dan optimasi dipanggil kembali untuk melakukan prediksi.

**3. Hasil dan Pembahasan**

**3.1. Visualisasi**

****

**3.2. Perbandingan metode**

Tabel 1. Akurasi model

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Model** | **Akurasi**  **pubchem** | **Akurasi**  **klekota-roth** |
| 1 | Logistic Regression | 0.983 | 1 |
| 2 | SVM | 0.917 | 0.633 |
| 3 | Random Forest | 0.967 | 1 |

Tabel 2. Confusion matrix

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** |  |  | **pubchem** | |  | **klekota-roth** | |
|  |  | **Prediksi** | |  | **Prediksi** | |
| 1. Linear regression |  |  | 1 | 0 |  | 1 | 0 |
|  | True value | 1 | 38 | 0 |  | 38 | 0 |
|  | 0 | 1 | 21 |  | 0 | 22 |
|  |  |  |  |  |  |  |  |
| 2. SVM |  |  | 1 | 0 |  | 1 | 0 |
|  | True value | 1 | 38 | 0 |  | 38 | 0 |
|  | 0 | 5 | 17 |  | 22 | 0 |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 3. Random forest |  |  | 1 | 0 |  | 1 | 0 |
|  | True value | 1 | 37 | 1 |  | 38 | 0 |
|  | 0 | 1 | 21 |  | 0 | 22 |

Tabel 3. Nilai sensitifitas dan presisi

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **pubchem** | | **klekota-roth** | |
| **No.** | **Model** | **Sensitifity** | **Precision** | **Sensitifity** | **Precision** |
| 1 | Logistic Regression | 0.955 | 1 | 1 | 1 |
| 2 | SVM | 0.773 | 1 | 0 | 0 |
| 3 | Random Forest | 0.955 | 0.955 | 1 | 1 |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| AUC Log. Regression:  1 | AUC SVM:  0.999 | AUC Random Forest:  0.978 |

Gambar 2. AUC masing-masing model dengan fingerprint pubchem

|  |  |  |
| --- | --- | --- |
|  |  |  |
| AUC Log. Regression:  1 | AUC SVM:  1 | AUC Random Forest:  1 |

Gambar 3. AUC masing-masing model dengan fingerprint klekota-roth

**3.3. Optimasi metode**

Optimasi model menemukan hal serupa, yakni logistic regression. Hasil cross valiation juga memukan perubahan nilai C dan tol konsisten. Selamjutnya, hasil iterasi dengan nilai akurasi terbaik dengan metode grid search menemukan kombinasi terbaik adalah dengan parameter sebagai berikut:

LogisticRegression(C=0.1, class\_weight=None, dual=False, fit\_intercept=True,

intercept\_scaling=1, max\_iter=100, multi\_class='warn',

n\_jobs=None, penalty='l2', random\_state=None, solver='warn',

tol=0.1, verbose=0, warm\_start=False)

Gambar 4. Cross validasi untuk parameter C

Tabel 4. Akurasi model standar dan cross validasi

|  |  |
| --- | --- |
| **Model** | **Nilai akurasi** |
| Logistic Regression standard | 0.983 |
| Logistic Regression optimasi | 0.983 |

**3.4. Screening**

**Tabel 5. Hasil prediksi molekul**

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Drug** | **Prediksi Model Standar** | **Prediksi Model Optimasi** |
| 1 | oseltamivir | Aktif (1) | Aktif (1) |
| 2 | zanamivir | Aktif (1) | Aktif (1) |
| 3 | chembl293591 | Aktif (1) | Aktif (1) |
| 4 | chembl344548 | Aktif (1) | Aktif (1) |

Kedua drug diprediksi aktif terhadap enzim neuraminidase. Meskipun begitu beberapa studi menunjukkan enzim neuraminidase memiliki resistensi terhadap kedua drug tersebut (<https://doi.org/10.1128/JVI.01200-09> dan <https://doi.org/10.2165/11531450-000000000-00000>). Selain itu hasil uji coba dengan molekul akftif lain juga menunjukkan prediksi aktif.

**4. Kesimpulan**

Hasil analisis machine learning untuk enzim nram diperoleh:

1. Metode logictic regression adalah metode terbaik untuk klasifikasi molekul aktif terhadap enzim nram;
2. Hasil optimasi model diperoleh akurasi yang serupa yakni 0.987;
3. Model logistic regression:

lr\_model = LogisticRegression(C=0.1, class\_weight=None, dual=False, fit\_intercept=True,

intercept\_scaling=1, max\_iter=100, multi\_class='warn',

n\_jobs=None, penalty='l2', random\_state=None, solver='warn',

tol=0.1, verbose=0, warm\_start=False)

1. Hasil prediksi menjukkan oseltamivir dan zanamivir aktif terhadap enzim neuraminidase dan berpotensi digunakan sebagai terapi;