Penerapan Optimasi Portfolio Untuk Pemilihan Kandidat Molekul Dalam Menghambat Ptp1b Pada Penyakit Diabetes Melitus Menggunakan Non-dominated Sorting Genetic Algorithm

Rafanzhani Elfarizy[1], Deni Saepudin[2], Isman Kurniawan[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]erafanzhani@students.telkomuniversity.ac.id, [2] denisaepudin@telkomuniversity.ac.id,
[3]ismankrn@telkomuniversity.ac.id

Abstrak

PTP1B atau Protein Tyrosine Phosphatase 1B merupakan protein yang menjadi salah satu penyebab dari penyakit Diabetes Melitus. Salah satu cara untuk menanganinya adalah dengan menghambat PTP1B. Terdapat banyak kandidat molekul yang berpotensi untuk menghambat pertumbuhan protein ini. Untuk membantu meringankan pemilihan, molekul dalam jumlah yang besar ini dipilih berdasarkan tingkat probabilitas kesuksesan yang tinggi. Masalah pemilihan ini menyerupai masalah pemilihan saham untuk portfolio yang optimal pada keuangan. Permasalahan ini dapat diselesaikan dengan menggunakan NSGA-ii berdasarkan prinsip Multi-Objective Optimization. Pada algoritma ini dalam setiap generasi dilakukan evaluasi berdasarkan non-dominted sorting terhadap individu pada populasi untuk mendapatkan individu terbaik yang akan menjadi parent pada generasi selanjutnya. Parent kemudian akan digunakan untuk menghasilkan himpunan turunan (*offsprings*). Pada akhir generasi akan didapatkan himpunan individu terbaik yang digambarkan dengan grafik *efficient frontier*. Sebanyak 3715 data yang digunakan dalam penelitian ini diambil dari www.ebi.ac.uk . Setelah dilakukan preprocessing terdapat sebanyak 1452 data yang memenuhi klasifikasi untuk dapat digunakan. Pengujian yang dilakukan terhadap dataset sebanyak 10 kali pengujian untuk 5 molekul dalam 1 portfolio. Untuk setiap peningkatan jumlah generasi didapatkan grafik dengan tingkat *confidence* terhadap konvergen yang semakin tinggi juga. Kenaikkan jumlah molekul dalam 1 portfolio berpengaruh terhadap kenaikkan *expected return* dan *diversity.*

Kata kunci : multi-objective optimization, NSGA-II, expected return, diversity, efficient frontier

Abstract

PTP1B or Protein Tyrosine Phosphatase 1B is a protein that is one of the causes of Diabetes Mellitus. One way to handle it is by inhibiting PTP1B. There are many candidate molecules that have the potential to inhibit the growth of this protein. For make selection easier, these large numbers of molecules are chosen based on a high probability of success. This selection problem is similar to the issue of stock selection for an optimal portfolio in finance. This problem can be solved by using NSGA-ii based on the principle of Multi-Objective Optimization. In this algorithm, each generation is evaluated based on non-dominted sorting of individuals in the population to get the best individual that will be the parent of the next generation. The parent will then be used to produce offsprings. At the end of the generation, it will be obtained the best set of individuals that are depicted with efficient frontier graphs. A total of 3715 data used in this study were taken from www.ebi.ac.uk. After preprocessing there are 1452 data that meet the classification to be used. Tests carried out on a dataset of 10 times testing for 5 molecules in 1 portfolio. For each increase in the number of generations a graph with a higher level of confidence in convergence is also obtained. An increase in the number of molecules in a portfolio influences the expected return and diversity.

Keywords: multi-objective optimization, NSGA-II, expected return, diversity, efficient frontier

1. Pendahuluan

1.1 Latar Belakang

Setiap organisme mempunyai peranannya sendiri dalam menyerang sistem tubuh manusia. Salah satu contohnya adalah PTP1B. PTP1B atau *Protein tyrosine phosphatase 1B* merupakan sebuah jenis protein yang menjadi salah satu penyebab dari penyakit Diabetes Melitus. Langkah yang dapat diambil untuk mengobati penyakit Diabetes Melitus adalah dengan membuat obat yang di dalamnya terkandung molekul yang bisa menghambat pertumbuhan PTP1B [1] .

Pada awalnya pemilihan molekul penghambat ini hanya didasari oleh probabilitas kesuksesan dari masing-masing molekul tersebut. Namun, hal ini dinilai tidak cukup, bahkan berisiko. Untuk mengatasi permasalahan tersebut pendekatan alternatif yang dilakukan adalah dengan melihat nilai *diversity* dari masing – masing molekul. Hal ini dimaksudkan agar untuk setiap kandidat molekul terpilih mengalami kegagalan, molekul lain yang memiliki keberagaman akan memberi peluang lain [2] .

Masalah pemilihan molekul tersebut mirip dengan masalah pemilihan saham pada portfolio keuangan. Pada portfolio keuangan, komposisi saham ditentukan berdasarkan nilai *expected return* tertinggi untuk

suatu tingkat risiko tertentu pada suatu portfolio. Pemilihan portfolio dengan mempertimbangkan 2 fungsi objektif ini dapat diselesaikan dengan metode *multi-objective optimization* [2].

Pada tugas akhir ini akan diimplementasikan portfolio pada keuangan untuk menyelesaikan masalah pemilihan kandidat molekul yang menghambat pertumbuhan PTP1B. Dalam menyelesaikan masalah ini ada dua fungsi objektif yang harus dimaksimumkan yaitu, *expected return* yang direpresentasikan sebagai probabilitas kesuksesan molekul dan *diversity* atau keberagaman molekul. Dengan menggunakan konsep *multi-objective* akan dihasilkan kemungkinan solusi yang dapat dipilih dengan mempertimbangkan kedua fungsi tersebut. salah satu algoritma untuk yang dapat menyelesaikan permasalahan *multi-objective optimization* adalah NSGA-II. Algoritma ini akan membentuk kumpulan solusi dengan melihat tingkat dominansi antar solusi. Sehingga tidak ada solusi yang saling mendominasi satu dengan yang lainnya.

1.2 Topik dan Batasannya

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, pada Tugas Akhir ini akan dibahas tentang bagaimana cara menentukan komposisi kandidat molekul penghambat PTP1B dengan menggunakan metode optimasi portfolio pada keuangan dengan menerapkan algoritma NSGA-II dalam menyelesaikan permasalahan *multi-objective optimization*. Kemudian Bagaimana pengaruh jumlah molekul dalam satu portfolio terhadap probabilitas kesuksesan dana diversity dari portfolio tersebut. Adapun batasan masalah yang digunakan pada tugas akhir ini adalah:

1. Dataset yang digunakan diambil dari www.ebi.ac.uk
2. Jumlah molekul dalam satu portofolio telah ditentukan.
3. Hanya data molekul dengan nilai *activity* lebih dari 10000 nM yang dapat digunakan.
4. Penggunaan *Molecular Fingerprint* berdasarkan *library* rdkit pada *python* yang merupakan hasil transformasi dari SMILES yang disediakan pada dataset.
5. Harga untuk setiap molekul diasumsikan sama

1.3 Tujuan

Penelitian ini bertujuan untuk menghasilkan komposisi kandidat molekul penghambat pertumbuhan PTP1B dengan nilai *expected return* tertinggi untuk suatu tingkat *diversity* tertentu dengan menerapkan metode *Multi-Objective Optimization* pada portfolio keuangan menggunakan algoritma NSGA-II. Selanjutnya dilakukan analisis untuk membuktikan pengaruh jumlah molekul dalam satu portfolio terhadap probabilitas kesuksesan dan *diversity* dari portfolio tersebut.

1.4 Organisasi Tulisan

Urutan penulisan dalam tugas akhir ini adalah sebagai berikut : bagian 2 menunjukkan studi literatur yang terkait dengan penelitian ini. Bagian 3 menjelaskan sistem yang akan dibangun mulai dari *preprocessing* data sampai dengan mengolah data menggunakan NSGA-II. Bagian 4 menjelaskan Analisa mengenai penelitian yang dilakukan. Bagian 5 menjelaskan kesimpulan yang didapat dan penelitian yang mungkin bisa dikerjakan untuk selanjutnya.

2. Studi Terkait

2.1 Drug Discovery

PTP1B atau *Protein tyrosine phosphatase 1B* merupakan jenis protein yang terbukti sebagai *negative regulator* dalam penyaluran sinyal insulin yang dapat menyebabkan penyakit *metabolic* seperti obesitas dan diabetes melitus . Protein ini telah dibuktikan berpotensi tinggi untuk menginduksi resistensi insulin dalam tubuh manusia. Hal ini yang menyebabkan PTP1B menjadi target yang tepat untuk menemukan pengobatan yang tepat untuk penyakit diabetes melitus [1].

2.2 Portfolio Keuangan

2.2.1 Portfolio

Portfolio sering dikaitkan dengan manajemen investasi. Para *investor* dalam melakukan investasi pasti selalu mengharapkan keuntungan semaksimal mungkin dengan risiko yang seminimal mungkin atas biaya yang dikeluarkan [3] . Markowitz membuktikan dalam teorinya bahwa untuk membentuk suatu portfolio yang optimal, diperlukan pembagian modal untuk beberapa sekuritas investasi. [4].

2.2.2 *Return* dan Risiko

Nilai pengembalian dalam portfolio keuangan diartikan sebagai jumlah keuntungan yang didapat dari setiap sekuritas yang dipilih sebagai portfolio. *Expected return* merupakan rata – rata keuntungan yang didapat untuk suatu portfolio yang dibangun [5]. Risiko dapat digambarkan sebagai ukuran seberapa jauh hasil keuntungan yang didapat dengan perkiraan awal. Semakin besar risiko, maka semakin besar pula peluang untuk tidak mendapatkan

keuntungan seperti yang diharapkan. Namun, dengan risiko yang tinggi pula investor mendapat peluang mendapat keuntungan yang lebih besar [5] [6] [7] .

2.2.3 Bobot Aset

Bobot aset merupakan komposisi untuk suatu aset sekuritas. Pada tugas akhir ini, bobot aset direpresentasikan sebagai komposisi pemilihan molekul yang akan dipilih. Pemilihan diputuskan dengan memilih molekul (bobot = 1) atau tidak (bobot = 0) [2].

2.3 Portfolio pada *Drug Discovery*

2.3.1 Probabilitas Kesuksesan

Tingkat kesuksesan molekul merepresentasikan nilai *return* pada saham. Probabilitas Kesuksesan merupakan suatu parameter yang dijadikan acuan untuk melihat peluang keberhasilan molekul untuk menjadi kandidat obat. Nilai probabilitas kesuksesan p_i untuk suatu molekul pada nTotal dataset dapat diperoleh dengan persamaan sebagai berikut [2]:

$$p_i = \frac{a_i}{\sum_{i=1}^{nTotal} a_i} \quad (1)$$

Dimana nilai a_i merupakan nilai logaritma dari aktivitas molekul i, salah satu ukuran parameter untuk aktifitas molekul yang dapat digunakan adalah IC50. Nilai IC50 menggambarkan banyaknya molekul yang diperlukan untuk menghambat 50% dari penyebaran target. Semakin kecil nilai IC50 maka semakin baik molekul tersebut. nilai a_i dapat dituliskan sebagai berikut :

$$a_i = e^{-IC_i} \quad (2)$$

Keterangan :

IC_i = Nilai aktivitas IC50 molekul i

2.3.2 *Expected Return*

Expected return merupakan nilai pengembalian yang diharapkan atas berhasilnya molekul menjadi elemen sebuah obat. Nilai *expected return* pada drug discovery dapat didefinisikan sebagai nilai *Gain* dikurangin nilai *Losses,* seperti berikut [2]:

$$E(X) = G\boldsymbol{p}\,.\,x - B = \left(\sum_{i=1}^{nTotal} G\,p_i x_i\right) - B \quad (3)$$

Keterangan :

E(X) = Nilai *expected return* untuk 1 protfolio

G = Nilai *gain* untuk keberhasilan molekul

nTotal = Jumlah molekul dalam dataset

x_i = Bit molekul

2.3.3 *Molecular Fingerprint*

Molecular fingerprint merupakan *bit string* yang merepresentasikan struktur molekul untuk menentukan tingkat *similarity* antar molekul. Ada beberapa tipe dari *molecular fingerprint* tergantung metode apa yang digunakan uintuk merepresentasikan struktur molekul menjadi sebuah *bit string*. *Topological fingerprint* merupakan salah satu tipe yang merepresentasikan struktur dengan melihat path pada molekul sampai dengan jumlah dari ikatan yang ada. *Bit string* berisi nilai 1 (untuk struktur yang ada pada molekul) dan 0 (untuk struktur yang tidak ada pada molekul [8] [9].

2.3.4 *Diversity*

Diversity merupakan nilai yang merepresentasikan keberagaman suatu populasi. Semakin besar nilai *diversity*, maka semakin besar pula tingkat keragaman suatu populasi. Nilai diversity yang tinggi diperlukan untuk memperkecil risiko kegagalan pemilihan molekul dalam satu portfolio. Dengan besarnya *diversity*, kumpulan molekul dalam portfolio akan semakin beragam, sehingga memberikan peluang kepada molekul lain (dalam satu portfolio) untuk berhasil jika yang lainnya mengalami kegagalan. Berdasarkan persamaan Solow-Polasky, *diversity* dapat dihitung dengan menjumlahkan semua elemen pada matrix invers dari matriks korelasi. Matriks korelasi dapat dimulai dengan mencari nilai *similarity* antar 2 molekul. [2]. *Tanimoto similarity* dapat digunakan untuk menghitung nilai *similarity* Sim_T dengan melihat *molecular fingerprint* antar molekul, seperti berikut [8] :

$$Sim_T(A,B) = \frac{N_{AB}}{N_A + N_B + N_{AB}} \quad (4)$$

Keterangan :

N_A = Banyaknya bit 1 pada molekul A saja

N_B = Banyaknya bit 1 pada molekul A saja

N_{AB} = Banyaknya bit 1 pada molekul A dan B

Nilai *distance* $d(x_i, x_j)$ dapat digambarkan sebagai nilai *disimilarity/diversity* antar molekul dengan melihat struktur antar molekul tersebut, seperti berikut :

$$d(x_i, x_j) = 1 - Sim_T(x_i, x_j) \quad (5)$$

Dimana, nilai elemen pada matriks F(X) dapat direpresentasikan sebagai :

$$f_{ij} = e^{-\theta d(x_i, x_j)} \quad (6)$$

Keterangan :

θ = Konstanta dengan nilai 0.5

$d(x_i, x_j)$ = Nilai *distance* antara molekul i dan molekul j

Kemudian dapat dihitung nilai *diversity* D(X) untuk suatu portfolio, seperti berikut :

$$D(X) = \sum_{i=1}^{nPortfolio} \sum_{j=1}^{nPortfolio} F(X)_{ij}^{-1} \quad (7)$$

Dimana $F(X)^{-1}$ adalah *invers* dari matriks F(X) yang didapat pada persamaan (6)

2.4 Non-Dominated Sorting Genetic Algortihm (NSGA-II)

2.4.1 Algoritma Genetika

Pada algoritma NSGA-II pemilihan parent ditentukan dengan melakukan pengurutan individu dengan mempertimbangkan nilai dominansi antar individunya (*Non-Dominated Sorting*) pada populasi. Individu yang mempunyai nilai *rank* tertinggi menggambarkan individu dengan nilai *fitness* terbaik. Selain melihat nilai *fitness*, algoritma ini juga memerhatikan nilai kedekatan antar indvidu dengan individu disampingnya (*Crowding Distance*). Individu dengan nilai *crowding distance* terbesarlah yang akan dipilih sebagi *parent* untuk menghasilkan individu baru dengan menggunakan *crossover* dan *mutation* [10].

2.4.2 *Non-Dominated Sorting*

Non-Dominated Sorting merupakan suatu metode untuk mengurutkan individu pada suatu populasi dengan memerhatikan dominansi antar individu. Suatu individu p dikatakan

mendominasi individu q jika salah satu atau kedua nilai fungsi objektif individu p lebih baik dari individu q [6] [7].

2.4.3 *Crowding Distance*

Crowding distance merepresentasikan jarak kerapatan antara *front* satu dengan *front* yang berada di dekatnya. Perhitungan ini dilakukan hanya pada tingkat *front* yang sama [11] .

2.4.4 *Crossover dan Mutation*

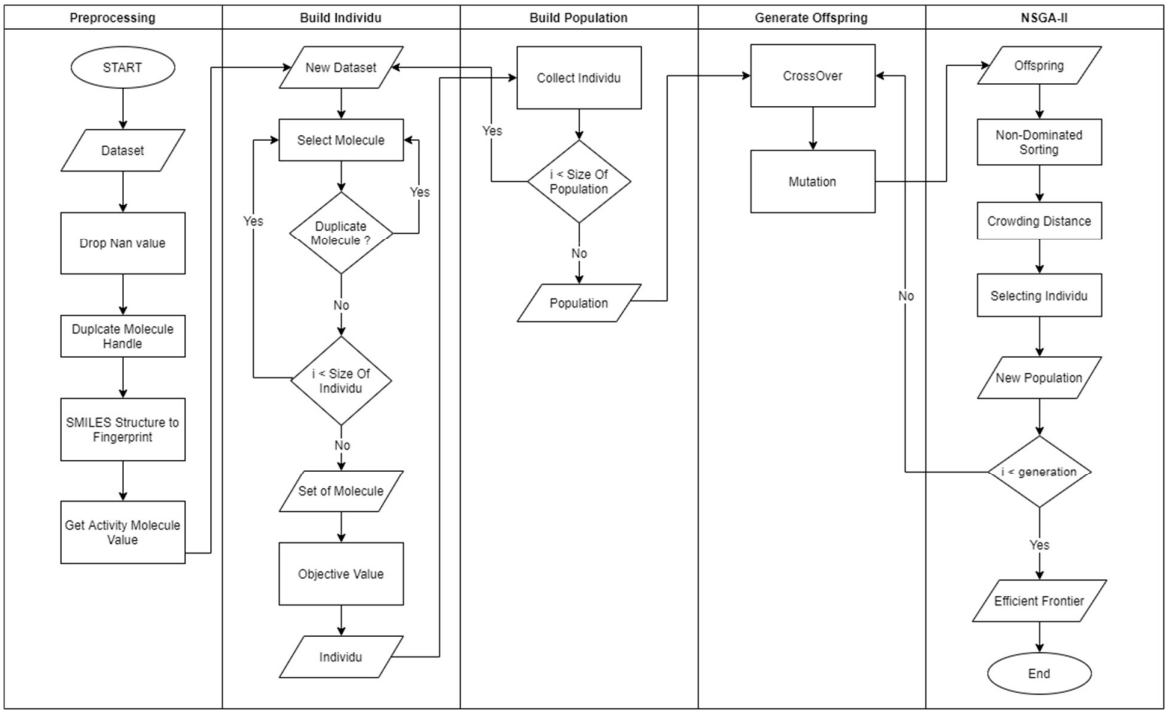
Crossover merupakan proses persilangan antara 2 individu (*parent*) yang berbeda dan telah terpilih pada proses sebelumnya. Kedua individu ini nantinya akan disilangkan untuk menghasilkan indivdu baru (*offspring*). Setelah dilakukan *crossover*, langkah selanjutnya adalah melakukan *mutation*. *Mutation* dilakukan dengan cara mengubah beberapa gen secara acak pada individu yang dihasilkan dari crossover [6].

2.5 Penelitian Terkait

Menurut penelitian yang dilakukan oleh Anagnostopoulos dalam membandingkan algoritma PESA, NSGA-II dan SPEA2 mendapatkan hasil bahwa PESA mempunyai perfomansi lebih baik secara pemilihan kedekatan individu pada *efficient frontier.* Sedangkan algoritma NSGA-II dan SPEA2 memiliki rata rata performansi yang baik untuk *hypervolume indicator.* Kelebihan lainnya dari kedua algoritma ini adalah bersifat fleksibel untuk menentukan *constraint* dan objektif untuk risiko [12].

3. Sistem yang Dibangun

Sistem yang akan dibangun pada penyelaesaian Tugas Akhir ini digambarkan dalam bentuk diagram alir pada Gambar 1 di bawah ini.



Gambar 1 Perancangan Sistem

3.1 Data Praprocessing

Dataset yang digunakan untuk pembangunan sistem ini berasal dari www.ebi.ac.uk. Sebelum dataset tersebut digunakan, praproses data perlu dilakukan. Beberapa preproses data yang dilakukan adalah sebagai berikut :

3.1.1 *Drop Nan Value*

Beberapa molekul memiliki *Nan value* pada salah satu fiturnya. Dengan adanya *Nan value* ini mengakibatkan beberapa proses yang akan ada tidak dapat dilanjutkan. Untuk mengatasi hal ini dilakukan *drop* pada molekul yang memiliki *Nan value* ini. Berikut merupakan beberapa molekul yang mempunyai *Nan value*.

Tabel 1 Data dengan *Nan Value* pada fiturnya

Molecule Name	SMILES	Activity		
		Standard Relation	Standard Value	Standard Units
CHEMBL179166	Nan	Nan	Nan	Nan
CHEMBL179166	Nan	'='	152000	nM
CHEMBL86473	Nan	'='	308000	nM
CHEMBL86473	Nan	'>'	10000000	nM
CHEMBL179166	Nan	Nan	Nan	Nan
CHEMBL179166	Nan	'='	78300	nM
CHEMBL2000194	Nan	'='	79300	nM

3.1.2 *Duplicate Molecule Handle*

Pada dataset yang tersedia memiliki beberapa nama molekul yang sama namun memiliki nilai *activity* dan SMILES yang berbeda. Untuk mengatasi hal ini data dengan nama molekul yang sama akan di-*drop* dari dataset.

3.1.3 *SMILES to Fingerprint Structure*

Nilai *Similarity* antar molekul didapatkan dengan menghitung jarak antar molekul dengan menggunakan *bit String Molecular Fingerprint*. Pada dataset hanya disediakan struktur SMILES, sehingga perlu diubah dalam *bit string Fingerprint*. Berikut merupakan contoh perubahan struktur yang dilakukan pada dataset :

Tabel 2 Perubahan dari struktur SMILES menjadi *bit string fingerprint*

Nama Molekul	SMILES	Fingerprint
CHEMBL140954	O=P(O)(O)C(F)(F)c1ccc(COc2ccc(OCc3ccc(C(F)(F)P(=O)(O)O)cc3)cc2)cc1	[1 0 0 ... 0 0 0]
CHEMBL440955	O=C1C[C@@H](c2ccc(C[C@H](Nc3nc4ccccc4s3)c3nc4ccccc4[nH]3)cc2)S(=O)(=O)N1	[1 0 1 ... 1 0 1]
CHEMBL246869	O=C1CC(c2ccc(C[C@H](Nc3nc4ccc(Br)cc4s3)c3nc4ccccc4[nH]3)cc2)S(=O)(=O)N1	[1 0 1 ... 1 0 1]
CHEMBL210836	O=C1C[C@@H](c2ccc(C[C@H](NS(=O)(=O)c3ccccc3)c3nc4ccccc4[nH]3)cc2)S(=O)(=O)[N-]1.[Na+]	[1 0 1 ... 1 0 1]
CHEMBL382311	CC(=O)N[C@@H](Cc1ccccc1)C(=O)N[C@@H](Cc1ccc([C@@H]2CC(=O)NS2(=O)=O)cc1)C(N)=O	[0 1 1 ... 1 0 1]
CHEMBL1935500	C/C=C/C=C/C(=O)N1Cc2cc(OCCc3nc(/C=C/C(C)CCC)oc3C)ccc2C[C@H]1C(=O)O.CC(C)(C)N	[1 1 1 ... 1 0 1]
CHEMBL1935493	C/C=C/C=C/C(=O)N1Cc2cc(OCCc3nc(/C=C/C(C)C)oc3C)ccc2C[C@H]1C(=O)O.CC(C)(C)N	[1 1 1 ... 1 0 1]
CHEMBL253500	O=C1CC(c2ccc(CC(NS(=O)(=O)c3cccc(F)c3)c3ncc(CCCc4ccccc4)[nH]3)cc2)S(=O)(=O)N1	[1 0 1 ... 1 0 1]

3.1.4 *Get Activity Molecule Value*

Dalam pembangunan sistem ini molekul yang digunakan hanya yang memenuhi beberapa syarat, seperti :

1. Relasi antara molekul dan nilai *Activity* yang digunakan adalah '='
2. Satuan unit yang digunakan adalah nM
3. Nilai aktivitas molekul kurang dari 10000 nM

Setelah beberapa molekul terpilih, dilakukan konversi dari satuan nM ke mM. Berikut merupakan beberapa contoh molekul yang terpilih :

Tabel 3 Seleksi molekul berdasarkan nilai *Activity*

Molekul dari Dataset

Molecule Name	Standard Relation	Standard Value	Standard Units
CHEMBL440955	'='	2700	nM
CHEMBL246869	'='	5000	nM
CHEMBL381633	'>'	1000000	nM
CHEMBL1946252	'>'	100000000	nM
CHEMBL210836	'='	320	nM
CHEMBL382311	'='	2400	nM
CHEMBL2414203	'>='	200	ug.mL-1
CHEMBL370470	'='	3801893963	nM

Molekul terpilih

Molecule Name	Standard Relation	Standard Value	Standard Units
CHEMBL440955	'='	2.7	mM
CHEMBL246869	'='	5	mM
CHEMBL210836	'='	0.32	mM
CHEMBL382311	'='	2.4	mM

3.2 Build Individu

Suatu individu terdiri dari beberapa molekul unik yang memiliki nilai *activity* dan SMILES nya masing-masing. Setelah proses pemilihan molekul dilakukan, nilai *Expected Retrun* dan *Diversity* dapat dicari. Kedua nilai ini yang nantinya akan digunakan untuk melakukan evaluasi untuk membentuk setiap individu dalam populasi setiap generasinya berdasarkan algoritma NSGA-II. Pemilihan molekul untuk membentuk individu pada generasi pertama dilakukan secara *random*.

3.3 *Build Population*

Beberapa individu yang telah dibentuk akan dikumpulkan menjadi satu populasi. Pada sistem ini, tidak dimungkinkan terdapat beberapa individu yang memiliki himpunan molekul yang sama persis antar molekul tersebut. Proses pembentukan individu akan terus dilakukan sampai batas jumlah individu dalam populasi terpenuhi.

3.4 *Generate Offspring*

Pada proses ini jumlah populasi ditambahkan menjadi 2 kali lipat dari jumlah populasi awal. Populasi baru ini terdiri dari parent atau himpunan individu dari populasi sebelumnya dan *offspring* yang akan dibentuk dari *parent* dengan beberapa proses sebagai berikut :

3.4.1 *Crossover*

Crossover merupakan proses pertukaran silang yang dilakukan antara 2 individu (*parent*) dan menghasilkan 2 individu baru (*offspring*). Individu dipilih berdasarkan Probabilitas *Crossover* (PCo) yang telah ditentukan. Setiap iterasi dalam pemilihan individu dilakukan *generate* bilangan antara 0 sampai 1 secara *random*. Untuk setiap individu yang memiliki nilai probabilitas kurang dari nilai (PCo) akan dilakukan proses *crossover*.

3.4.2 *Mutation*

Mutation merupakan proses pergantian beberapa molekul terpilih yang ada pada individu dengan molekul yang ada di dataset. Pemilihan molekul dilakukan dengan men-*generate* bilangan random antara 0 sampai 1 untuk setiap iterasi molekul. Nilai yang memiliki nilai kurang dari nilai Probabilitas *Mutation* (PMut) yang telah diberikan.

3.5 NSGA-II

NSGA-II merupakan salah satu algoritma yang dapat digunakan untuk menyelesaikan masalah *Multi Objective Optimization* (MOO). Algoritma ini salah satu jenis *Evolutionary Algorithm* (EA). Algoritma ini akan menghasilkan beberapa individu terbaik. Offspring yang telah dibentuk pada proses *Generate Offspring* akan dilakukan beberapa proses pemilihan individu terbaik untuk setiap generasinya. Iterasi akan terus berlangsung sampai jumlah generasi terpenuhi. Berikut merupakan beberapa proses yang dilakukan pada algoritma ini :

3.5.1 *Non-Dominated Sorting*

Seperti yang sudah dijelaskan pada bagian .. Berikut merupakan langkah yang harus dilakukan pada *Non-Dominated Sorting* , yaitu : [6] [7].

1. Untuk setiap individu p yang berada pada populasi A , dilakukan :
   a. Inisialisasi Sp = {}, dimana Sp merupakan himpunan individu q yang didominasi oleh p.
   b. Inisialisasi np = 0, dimana np menyimpan jumlah invidu q yang mendominasi p.
   c. Untuk setiap individu q dalam populasi A akan diperiksa :
      i. Jika p mendominasi q, maka tambahkan q kedalam himpunan Sp atau dapat ditulis sebagai Sp = Sp U {q}
      ii. Jika q mendominasi p, maka nilai np ditambahkan 1 atau dapat ditulis sebagai np = np + 1
   d. Jika tidak ada satupun individu q yang mendominasi p (np = 0), maka lakukan :
      i. Individu p menjadi *front* pertama.
      ii. Individu p diberi *rank* 1 atau dapat ditulis sebagai pRank = 1
      iii. *Update front* pertama dengan menambahkan individu p (F1 = F1 U {p})
2. Inisialisasi *front* = 1 dan i = 1
3. Jika *front* ke i mempunyai anggota didalamnya (Fi ≠ {}), maka lakukan :
   a. Inisialisasi Q = {}, dimana Q merupakan himpunan individu untuk *front* setelahnya (Fi+1)
   b. Untuk setiap inividu p di dalam Fi, dilakukan :
      i. Untuk setiap individu q yang merupakan anggota Sp, dilakukan :
      1) Lakukan pengurangan jumlah untuk himpunan individu yang mendominasi q atau dapat ditulis sebagai nq = nq – 1, dimana nq merupakan jumlah individu yang mendominasi q.

2) Jika tidak ada satupun individu yang mendominasi q atau dapat ditulis sebagai nq = 0, maka lakukan :
   a. Individu q diberi rank i+1 (qrank = i + 1)
   b. *Update* Q dengan menambahkan individu q (Q = Q U q)
   c. Tetapkan Q yang sudah terbentuk sebagai *front* berikutnya (Fi = Q)

3.5.2 *Crowding Distance*

Adapun cara mendapatkan nilai *crowding distance* yaitu [11]:

1. Inisialisasi *distance* untuk semua individu ke- k sampai n dengan nilai 0, dengan n merupakan jumlah individu yang ada pada satu front.
2. Untuk setiap fungsi objektif m dilakukan :

a. Pengurutan setiap individu dengan melihat nilai fungsi objektif m, sehingga I = *sort*(I,m)
b. Inisialisasi distance[1] = distance[n] = ∞
c. Untuk setiap individu ke k+1 sampai n-1 dilakukan :
   i. Proses perhitungan jarak untuk masing – masing individu dengan tetangga terdekatnya menggunakan rumus :

$$\boldsymbol{distance}[\mathbf{k}] = \mathbf{distance}[\mathbf{k}] + \left(\frac{\mathbf{I[k+1].m - I[k-1].m}}{\boldsymbol{f_m^{max} - f_m^{min}}}\right) \quad (8)$$

Keterangan :
I[k].m = Nilai fungsi objektif m dari individu k di I
f_m^{max} = Nilai maksimum fungsi objektif m
f_m^{min} = Nilai minimum fungsi objektif m

3.5.3 *Selecting Individu*

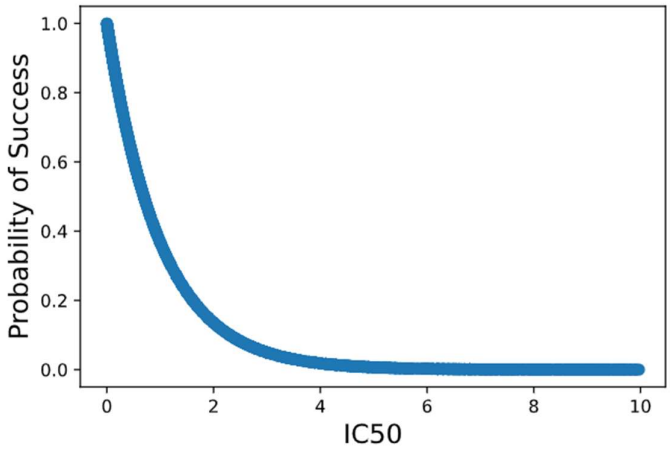
Nilai yang didapat pada Proses *Crowding Distance* digunakan pada proses *Selecting Individu*. Nilai dari setiap individu yang terdapat pada *offspring* tersebut dilakukan *Ascending Order*. Berdasarkan kumpulan molekul yang telah diurutkan, dipilih beberapa molekul pertama sebanyak besarnya populasi yang telah ditentukan. Kumpulan molekul terpilih ini yang akan dijadikan populasi baru untuk generasi selanjutnya.

4. Evaluasi

Dataset yang digunakan pada penelitian kali ini diambil dari www.ebi,ac,uk. Dataset molekul yang digunakan merupakan molekul dengan target protein PTP1B. Pada kategori ini, data yang tersedia sebanyak 3715 molekul. kemudian dilakukan praprocessing untuk mendapatkan data yang dapat diolah. Setelah dilakukan preprocessing data, molekul yang dapat digunakan adalah sebanyak 1452 molekul. Data inilah yang kemudian digunakan untuk melakukan pencarian individu terbaik.

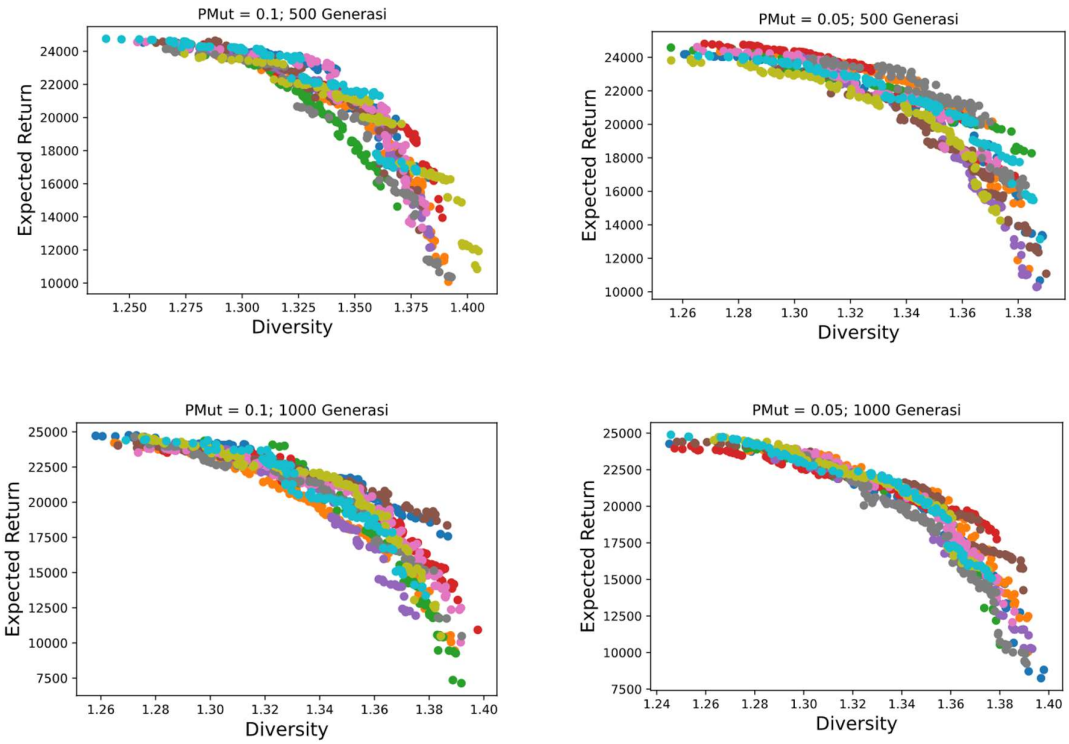
4.1 Hasil Pengujian

Berikut merupakan persebaran hubungan data aktivitas molekul dengan probabilitas kesuksesan suatu molekul, seperti berikut :



Gambar 2 Persebaran data molekul berdasarkan nilai *Activity*

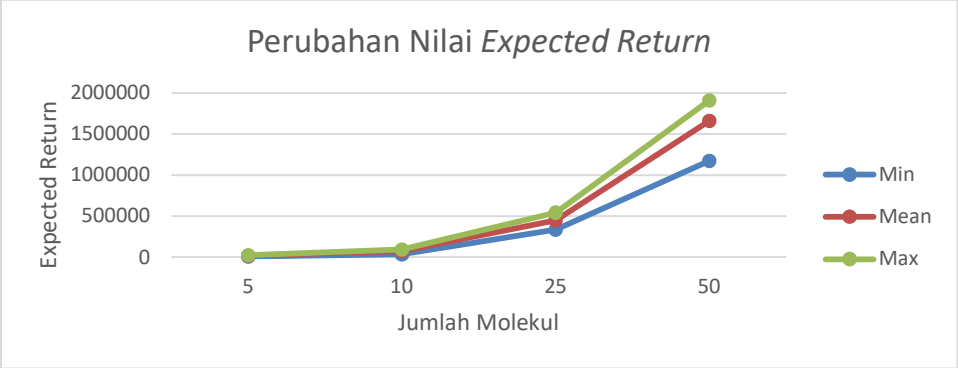
Berdasarkan gambar 2 dapat dilihat bahwa semakin besar nilai IC50 untuk setiap molekul maka tingkat probabilitas kesuksesan untuk molekul tersebut juga semakin kecil. Selanjutnya pengujian dilakukan dengan melakukan 30 kali *running* untuk suatu jumlah molekul tertentu dalam suatu portfolio. Gambar 3 menunjukkan hasil dari 10 kali percobaan dengan beberapa parameter yang digunakan seperti probabilitas *Crossover* (PCo) sebesar 0,8 , Probabilitas *Mutation (*PMut) sebesar 0.05 dan 0,1 untukjumlah molekul untuk suatu portfolio adalah 5 molekul. Adapun parameter nilai generasi maksimum untuk setiap pengujian yang diubah untuk dilihat perbandingan yaitu 500 dan 1000 generasi. Hasil yang didapatkan dalam setiap pengujian adalah grafik *efficient frontier* yang disatukan dalam satu grafik untuk setiap nilai maximum generasi dan nilai PMut yang sama.



Gambar 3 Grafik *efficient frontier* dari 10 kali *running* untuk nilai generasi maksimum 500 dan 1000 generasi

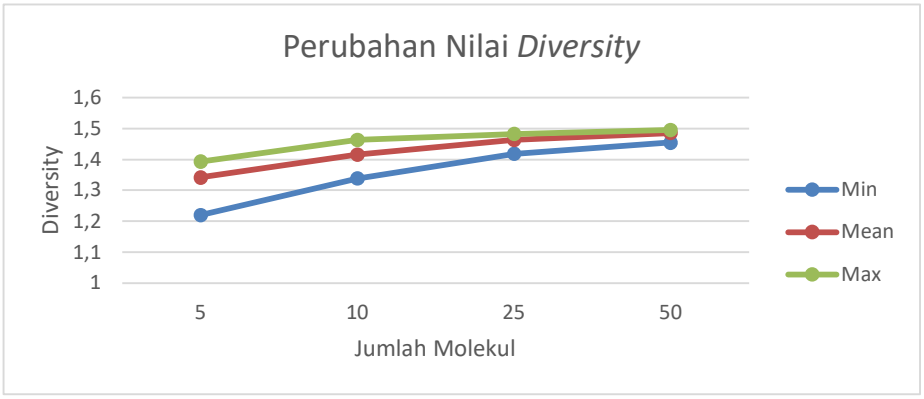
4.2 Analisis Hasil Pengujian

Berdasarkan hasil pengujian terlihat bahwa untuk semua hasil pengujian untuk msaing-masing nilai generasi maksimum yang berbeda didapatkan grafik dengan tingkat *confidence* terhadap konvergen yang relative mingkat sejalan dengan meningkatnya jumlah gnerasi. Selain jumlah generasi, nilai PMut ynang semakin kecil juga menghasilkan grafik dengan tingkat *confidence* yang semakin tinggi. Namun, semakin tinggi jumlah generasi yang harus dicapai, nilai *cost time* yang harus diambil juga semakin besar. Dari hasil pengujian untuk portfolio 5 molekul didapatkan bahwa *running time* untuk percobaan 500 generasi adalah 258,836 detik sedangkan untuk 1000 generasi adalah 605,404 detik.



Gambar 4 Perubahan nilai *diversity* terhadap jumlah molekul dalam suatu portfolio

Berdasarkan grafik pada gambar 4 didapatkan bahwa semakin banyak molekul dalam suatu portfolio maka semakin tinggi pula kemungkinan nilai *expected return* yang diberikan. Selain itu rentang nilai yang dihasilkan semakin melebar dengan bertambahnya jumlah molekul.



Gambar 5 Perubahan nilai *expected return* terhadap jumlah molekul dalam suatu portfolio

Berdasarkan grafik pada gambar 5 sama halnya dengan niilai *expected return* semakin banyak jumlah molekul maka semakin tinggi pula kemungkinan nilai *diversity* yang dihasilkan. Namun, grafik menunjukkan rentang yang semakin kecil dan nilai diversity menuju nilai asimtotik tertentu.

5. Kesimpulan

Dari pembahasan teori dan beberapa pengujian yang telah dilakukan, maka dapat diambil beberapa kesimpulan. Penerapan Algoritma NSGA-II dalam menghasilkan beberapa individu terbaik dengan nilai *Expected Return* maksimum untuk tingkat *diversity* tertentu menghasilkan grafik dengan tingkat *confidence* yang relative meningkat sejalan dengan meningkatnya jumlah generasi. Selain tingkat generasi, nilai probabilitas *mutation* (PMut) yang semakin kecil juga sangat berpengaruh terhadap meningkatnya tingkat *confidence*. Jika dilihat pada gambar 3, perpaduan antara nilai PMut 0.05 dengan 1000 generasi menghasilkan grafik dengan tingkat c*confidence* yang tinggi. Namun, nilai *cost time* yang semakin meningkat juga patut menjadi bahan pertimbangan. Kemudian semakin banyak jumlah molekul dalam suatu portfolio juga berpengaruh terhadap meningkatnya nilai rentang untuk *expected return* dan *diversity.* Nilai rentang untuk *expected return* semakin melebar dengan bertambahnya jumlahh molekul. Hal yang berbeda terjadi untuk nilai *diversity.* Grafik dari gambar 5 menunjukkan semakin menyempitnya rentang nilai untuk diversity dan meunuju nilai asimtotik tertentu. Berbeda dengan *expected return* yang akan terus meningkat dengan semakin bertambahnya jumlah molekul, untuk nilai *diversity* tidak akan melewati batas nilai asimtotik.

Pada penelitian berikutnya, penulis menyarankan beberapa jenis peningkatan. Pertama, dikarenakan data yang sangat terbatas, pada penelitian ini harga untuk setiap molekul belum diperhitungkan sehingga diasumsikan sama untuk setiap molekul. Penulis menyarankan untuk mencari data yang memuat nilai harga untuk setiap molekul.

Daftar Pustaka

[1] J. Sun, C. Qu, Y. Wang, H. Huang, M. Zhang, H. Li, Y. Zhang, Y. Wang and W. Zou, "PTP1B, A Potential Target of Type 2 Diabetes Mellitus," *Molecular Biology,* vol. 5, 2016.

[2] I. Yevseyeva, E. B. Lenselink, A. d. Vries, A. P. I. A. H. Deutz and M. T. Emmerich, " Application of portfolio optimization to drug discovery," *Information Sciences,* 2018.

[3] I. Yunita, "Markowitz Model dalam Pembentukkan Portfolio Optimal (Studi Kasus pada Jakarta Islamic Index)," *Jurnal Manajemen Indonesia,* 2018.

[4] B. Pardosi and A. Wijayanto, "Aanalisis Perbedaan Return dan Risiko Saham Portfolio Optimal dengan Bukan Portfolio Optimal," *Mangement Analysis Journal,* 2015.

[5] J. Andreas, "Introduction of Portfolio Risk," *Pinnacle Investment Research,* 2016.

[6] S. Isnaeni, D. Saepudin and R. F. Umbara, "Penerapan Algoritma Genetika Multi-objective NSGA-II Pada Optimasi Portofolio," *e-Proceeding of Engineering,* vol. 2, p. 6841, 2015.

[7] A. Pratiwi, D. Saepudin and R. F. Umbara, "Optimasi Portofolio Mean-semivariance dengan Algoritma Genetika Multiobjective Evolutionary NSGA II," *e-Proceeding of Enginering,* vol. 5, pp. 8269-8281, 2015.

[8] I. Muegge and P. Mukherjee, "An overview of molecular fingerprint similarity search in virtual screening," *Expert Opinion on Drug Discovery ,* pp. 137-148, 2015.

[9] A. Ceret-Massague, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallve and G. Pujadas, "Molecular Fingerprint Similarity Search in Virtual Screening," *Methods,* pp. 58-63, 2015.

[10] P. L. L. Belluano, "Optimalisasi Solusi Terbaik dengan Penerapan Non-Dominated Sorting II Algorithm," *Jurnal Ilmiah ILKOM,* 2016.

[11] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II," *IEEE Transaction on Evolutionary Computation,* vol. 6, pp. 182-197, 2002.

[12] K. P. Anagnostopoulos and G. Mamanis, "Multiobjective evolutionary algorithms for complex portfolio optimization problems," *Springer,* pp. 259-279, 2019.