Универзитет у Београду

Математички факултет

Примена генетског алгоритма у откривању молекула са потенцијалним лековитим дејствима

Лазар Савић

Професор: Владимир Филиповић

Асистент: Стефан Капунац

Јануар 2025.

Садржај

1 Увод

Већ преко три деценије, методе откирвања лекова употребом рачунара (eng. Computational drug discovery) играју значајну улогу у развоју мањих молекула од фармаколошког значаја. Оне представљају ефективну стратегију за убрзавање и економизацију процеса развијања нових молекулских структура.

Традиционални приступи у откривању лекова често подразумевају вишегодишњи рад са великим бројем клиничких испитивања, што захтева значајне ресурсе и време. Иако употребом рачунарских метода овај проблем није у потпуности отклоњен, он је значајно убрзан. Како је простор потенцијалних молекулских структура од интереса изузетно велики, различите методе вештачке интелигенције наметнуле су се као природна техника за претрагу таквог простора.

Једну од њих представљају и генетски алгоритми (ГА). Генетски алгоритми припадају широј групи метахеуристичких алгоритама глобалне оптимизације или претраге који користе технике инспирисане биологијом. Генетски алгоритми користе појмове као што су селекција, укрштање, наслеђивање, мутација, итд. У природи, еволуција је процес у којем јединке које су најбоље прилагођене околини преживљавају и остварују потомство, које је најчешће исто тако или боље прилагођено околини. У контексту нашег проблема, јединке, као саставни елементи популације, представљају молекуле. Њихову прилагођеност, начине селекције, укрштања и мутације, описујемо у посебном поглављу.

Кључни аспект примене генетских алгоритама у генерисању молекула са лековитим својствима је њихова способност да се "науче" из постојећих података. Коришћењем хемијских и биолошких база података, ГА могу створити нови скуп молекула који нису били претходно идентификовани, а који могу имати потенцијал за лечење болести као што су рак, неуродегенеративне болести, инфекције изазване резистентним бактеријама и многе друге.

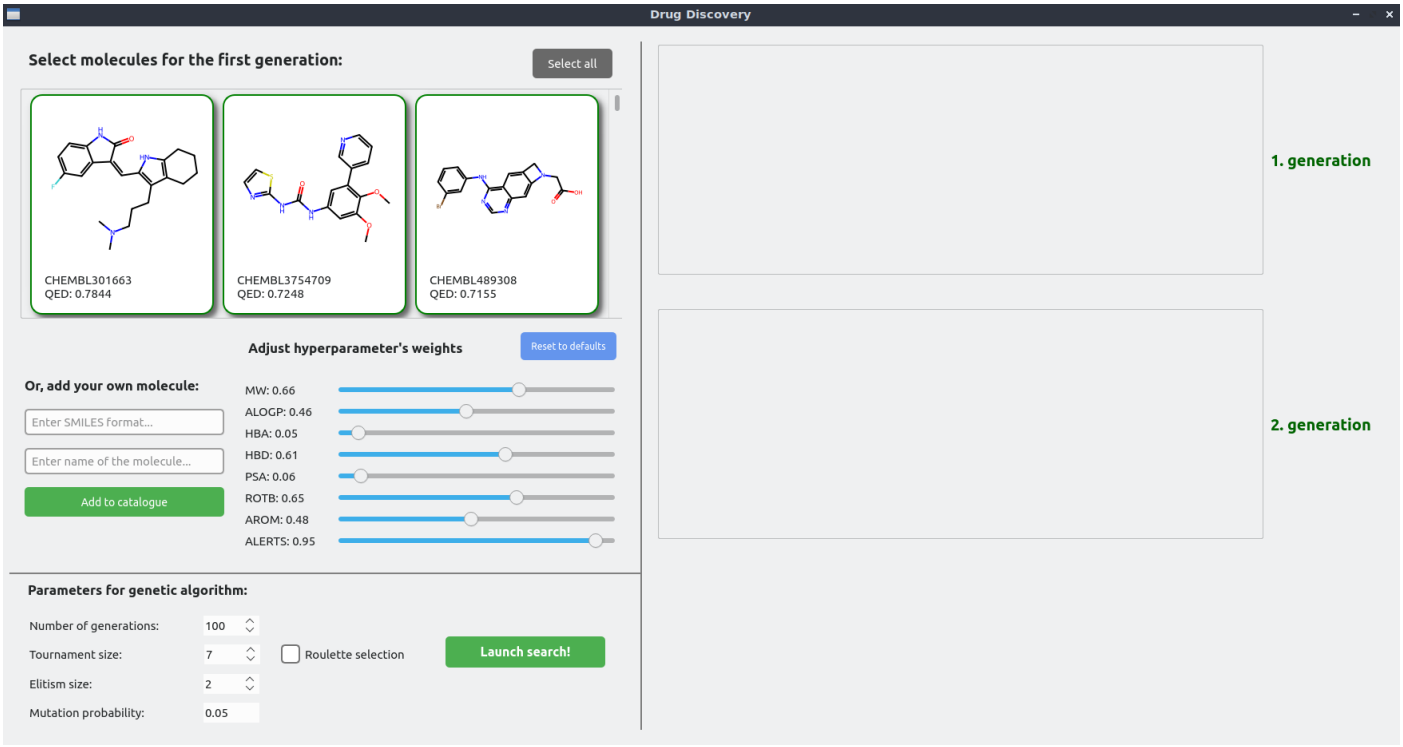
У овом раду, приказани су и резултати, односно молекулске структуре од потенцијалног значаја, добијене употребом ГА. Такође, разматрамо како конфигурисање параметара генетског алгоритма утиче на квалитет добијених јединки. Потребно је нагласити да приказани молекули представљају само идеје које тек треба да прођу све фазе будућих клиничких испитивања, а не лекове за комерцијалну употребу.

2 Генетски алгоритам

У овом поглављу приказујемо развијени софтвер за молекуларни дизајн. Графички кориснички интерфејс апликације направљен је употребом PyQt5 библиотеке у програмском језику Python. PyQt5 је популаран алат за развој десктоп апликација, будући да дозвољава једноставну изградњу интерфејса и интеграцију са Python кодом. Визуализација структура молекула је у разматраном проблему кључна, те се овакав софтвер намеће као природно решење.

Од велике помоћи су и постојеће хемоинформатичке библиотеке, чије се функционалности ослањају на методе машинског учења. У овом случају, коришћена је библиотека RDKit, која има API развијен за Python и C++, а чије ћемо конкретне примене у развоју ове апликације разматрати у наредним деловима рада.

На слици 1, приказан је изглед апликације приликом њеног покретања.



Слика 1: Почетни прозор.

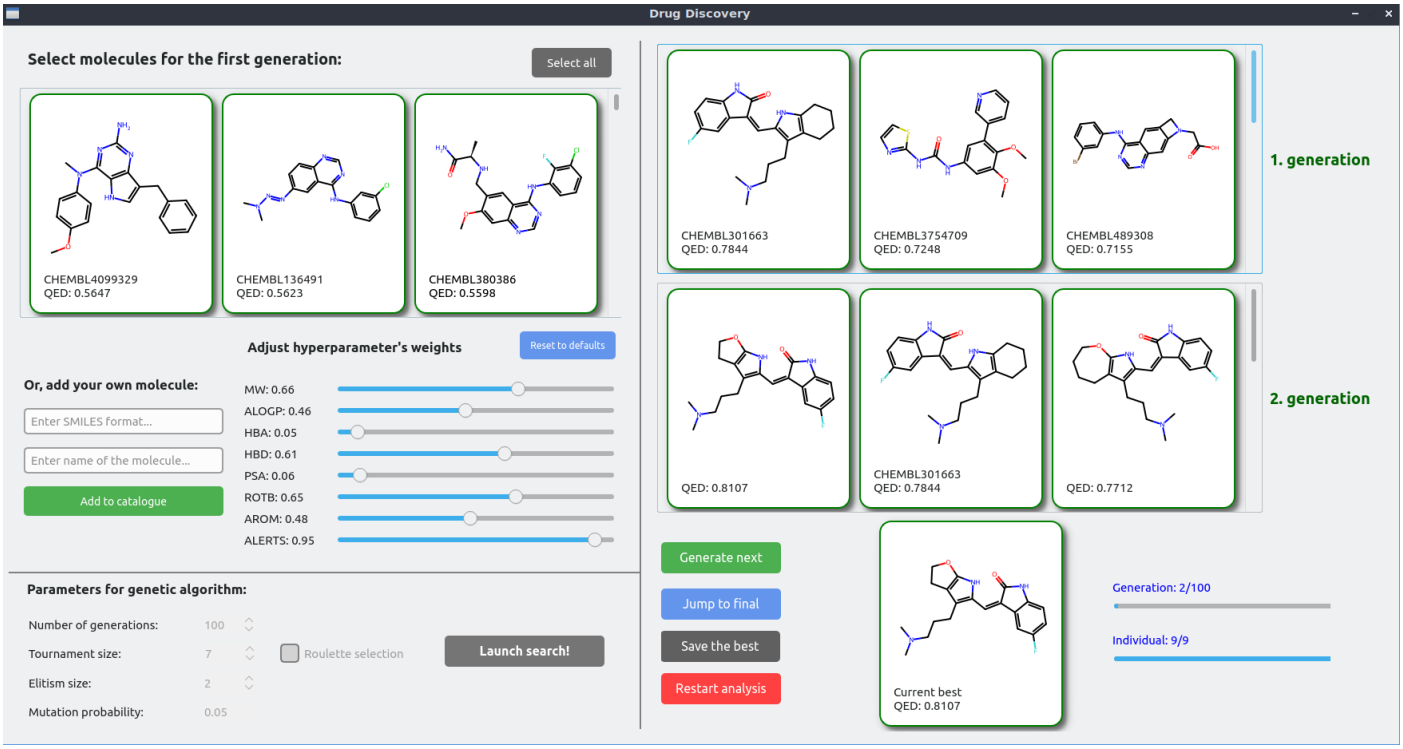
У горњем левом делу, приказан је каталог са доступним молекулима. Сваки од њих представљен је структурном формулом, идентификатором, и QED коефицијентом, о ком ће више речи бити у поглављу 2.2. Избор молекула који ће ући у прву генерацију врши се кликом на картицу одговарајућег молекула. Том приликом, та картица се пребацује у горњи одељак на десној половини екрана. Омогућен је и обрнут процес, тј. кликом на картицу у десној половини, она се враћа у почетни каталог и тиме избацује из прве генерације одабраних молекула. Дугме Select all убрзава процес селекције иницијалних молекула, у случајевима када желимо да одаберемо већину доступних молекула.

Могуће је проширивање почетне базе молекулских структура, попуњавањем формулара у средишњем делу леве половине прозора. Да би се убацио нови молекул, потребно је навести његов SMILES формат и опис (најчешће комерцијални назив, ако је молекул комерцијално доступан, IUPAC назив, или неки други тип идентификатора). SMILES формат молекула је ниска карактера која кодира атоме од којих је молекул сачињен, као и начин на који су они повезани (вишеструкост хемијских веза, циклуси, ароматични прстени, стереохемија, и слично). Библиотека RDKit на основу примљеног SMILES формата, у случају да је он исправан, формира сликовити приказ хемијске структуре учитаног молекула, конвертујући SMILES формат у погодну структуру графа. Израчунава се QED коефицијент молекула, и он се убацује у опадајуће сортирану листу молекула из каталога (на основу свог QED коефицијента).

QED коефицијент се рачуна на основу осам хиперпараметара, чија је вредност између 0 и 1. Њихове тежине, односно важности, постављене су у складу са резултатима (1), као и у складу са иницијалним вредностима које користи библиотека RDKit. Оне се могу контролисати на једноставан начин приказан у средишњем делу леве стране прозора, а њиховом променом се аутоматски ажурирају QED вредности свих молекула из каталога. Тако је могуће "испробавати" и увидети на који начин свака од ових тежина утиче на вредности QED коефицијената.

У доњем левом делу, могуће је конфигурисати параметре генетског алгоритма. Након подешавања свих параметара, могуће је покренути алгоритам, кликом на дугме Launch search! Тада се на основу одабране прве генерације молекула и подешених параметара формира друга генерација молекула, чије се јединке приказују у средишњем делу десне половине прозора. Пример оваквог извршавања приказан је на слици 2.

Сада је могуће посматрати развој молекулских јединки из генерације у генерацију (кликом на дугме Generate next), или директно "скочити" на последњу жељену генерацију (кликом на дугме Jump to final). Брзина одвијања



Слика 2: Покретање генетског алгоритма и формирање друге генерације молекула

ових процеса може се пратити кроз траке напретка (енг. progress bar) у доњем десном углу прозора. Молекул из последње формиране генерације са најбољим QED коефицијентом се такође приказује у издвојеној картици у подножју десног дела екрана. Уколико је то молекул од интереса, могуће је сачувати информације о њему, кликом на дугме Save the best. Том приликом се SMILES репрезентација тог молекула, његов QED коефицијент и тренутно време бележе у посебан фајл, који је од интереса за касније анализе. Као индикатор успешности, појављује се и лабела Saved!. Том приликом се појављују и два прозора, која кориснику нуде могућност рачунања неких додатних статистика. Дугме Restart analysis омогућава повратак апликације на стање приказано сликом 1.

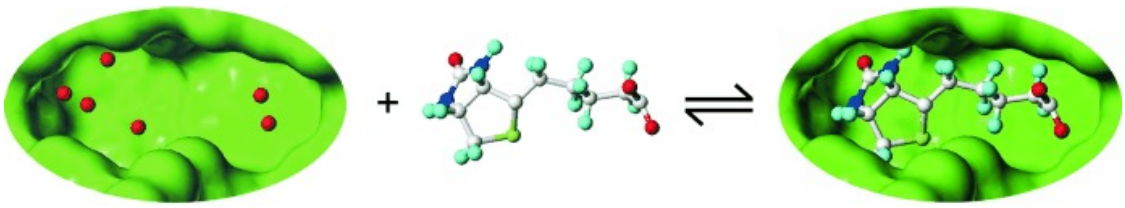
2.1 Јединке

Јединке у једној генерацији су молекулске структуре. Све молекулске структуре су представљене својим SMILES форматом, који на компактан и једноставан начин, у облику ниске карактера, чува све неопходне информације о хемијској композицији молекула. За иницијалне молекуле, приказан је и њихов идентификатор.

2.2 QED коефицијент

Фитнес функција, односно функција прилагођености, говори колико је јединка адаптивна у околини у којој се налази, односно, даје оцену квалитета те јединке. У контексту нашег проблема, фитнес функција би требала да показује колико одређени молекул има потенцијала да поседује лековита својства.

У пракси, критеријуми којима се одређује да ли молекул има лековита својства или не, постављају се у виду правила. Најпознатији скуп таквих правила је Липинскова петорка (енг. *Lipinski's rule of five* (*Ro*5)) (2). То је петорка правила које би сваки молекул требало да испуни да би могао да се сматра фармаколошки интересантним. Проблем са оваквим приступом је што је сувише "бинаран". На основу њега, не постоји начин да од два молекула која задовољавају свих пет правила одредимо бољи. Стога, овај приступ треба на неки начин фазификовати, односно, предиспозицију молекула да има лековита својства је неопходно *квантификовати*. Постоје разни приступи који се разликују по сложености и ефикасности. У случају да истражујемо одређену групу молекула, са специфичним фармаколошким својствима, приступ који се може користити јесте *афинитет везивања за циљни протеин*. Молекули од интереса се везују за одређени протеин, као што је приказано на слици 3. (3)



Слика 3: Везивање молекула за циљни протеин

Процес везивања лиганда за протеин је егзотерман. Већа количина ослобођене енергије приликом везивања указује на јачу кохезију лиганда и протеина. Развијени су разни алати који су у стању да израчунају овај афинитет. Већи афинитет везивања чини молекул бољим лигандом, те се за фитнес функцију управо може узети та метрика. Израчунавање овог афинитета је изузетно рачунарски захтевно, будући да подразумева испитивање великог броја могућности просторног уклапања лиганда у зависности од геометрије протеина. Ефикасност израчунавања фитнес функције је пресудна приликом примене генетског алгоритма, јер се извршава над огромним бројем јединки, кроз генерације. Због тога, искоришћен је знатно једноставнији приступ. QED (енг. Quantitative estimation of drug − likeness) развили су Ричард Бикертон и његови сарадници (1). Основу ове метрике чине физиохемијске карактеристике молекула (тј. *молекулски дескриптори*), и то:

- Моларна маса (енг. molecular weight, MW)
- Октанол-вода партициони коефицијент (енг. octanol − water partition coefficient, ALOGP)
- Број донора водоничних веза (енг. Number of hydrogen bond donors, HBD)
- Број акцептора водоничних веза (енг. Number of hydrogen bond acceptors, HBA)
- Поларна молекулска површина (енг. Molecular polar surface area, PSA)
- Број ротабилних веза (енг. Number of rotatable bonds, ROTB)
- Број ароматичних прстенова (енг. Number of aromatic rings, AROM)
- Број структурних упозорења (енг. Number of structural alerts, ALERTS)

Различита испитивања показала су да наведене карактеристике могу да сугеришу да ли молекул потенцијално поседује лековита својства. Такође, њихово израчунавање је углавном једноставно и веома брзо, што је битно својство сваке фитнес функције, и своди се на израчунавање вредности следеће функције:

$$d(x) = a + \frac{b}{\left[1 + \exp\left(-\frac{x-c+\frac{d}{2}}{e}\right)\right]} \cdot \left[1 - \frac{1}{\left[1 + \exp\left(-\frac{x-c-\frac{d}{2}}{f}\right)\right]}\right]$$

Параметри a, b, c, d, e и f за сваку од функција d_{MW}, d_{ALOGP}, d_{HBD}, d_{HBA}, d_{PSA}, d_{ROTB}, d_{AROM} и d_{ALERTS}, могу се пронаћи у (4).

Описани молекуларни дескриптори могу бити различите важности приликом одређивања QED коефицијента, те се он може израчунати као *тежинска геометријска средина вредности функција молекуларних дескриптора*, што је приказано наредном једначином:

$$QED = \exp\left(\frac{\sum_{i=1}^{n} w_i \ln(d_i)}{\sum_{i=1}^{n} w_i}\right)$$

Уколико се не нагласи другачије у наставку овог рада, коришћене су подразумеване вредности тежина молекуларних дескриптора. У библиотеци RDKit постоји модул rdkit.Chem.QED, у ком је имплементирана функција која рачуна овај коефицијент на описани начин.

2.3 Селекција

Имплементиране су две најпопуларније стратегије селекције молекула једне генерације који ће учествовати у укрштању, рулетска и турнирска.

Као и што је очекивано, у случају рулетске селекције, вероватноћа да i-та јединка буде одабрана једнака је:

$$p_i = \frac{QED_i}{\sum_{j=1}^{n} QED_j}$$

Турнирска селекција заснована је на случајном узорковању јединки из популације и одређивању оне која има најбољи QED коефицијент.

2.4 Укрштање

Након завршене фазе селекције, потребно је извршити укрштање изабраних јединки. У контексту овог проблема, фаза укрштања представља рачунарски најзахтевнији део генетског алгоритма.

Јединке су представљане својим SMILES нискама. Рекомбинација две ниске може се извршити на велики број начина, али главни проблем код сваког од њих представља *валидност* новодобијених ниски. Библиотека RDKit омогућава овакву проверу.

Наиван метод своди се на покушавање једнопозиционог укрштања док се не добију две валидне ниске. Иако прилично наиван, овај метод је погодан због своје једноставности и брзине извршавања, те је због тога управо он коришћен у имплементацији алгоритма. Огромна мана овог приступа је потенцијалан велики број неуспешних укрштања пре достизања валидног решења, па се због тога укупан број итерација ограничава, након чега се одустаје од даљих покушаја једнопозиционог укрштања.

Уколико једнопозиционо укрштање не успе, на сличан начин покушава се двопозиционо укрштање, које је такође ограничено максималним бројем итерација. Ако пак ни ово укрштање не успе, молекули се неизмењени пропуштају у наредну генерацију. Иако ово звучи проблематично, показује се да су овакве ситуације ретке уколико се максималан број дозвољених итерација за претходна два описана начина укрштања постави на довољно високу вредност.

2.5 Мутација

Примена мутација омогућава очување диверзитета, односно разноврсности молекула једне генерације. Углавном је вероватноћа доласка до мутације изузетно мала, и она не би требала да направи значајне промене на молекулу. Како ове структуре могу бити изузетно сложене, потребно је опрезно изабрати скуп дозвољених мутација, чије извршавање треба реализовати врло пажљиво, како се не би нарушила валидност молекулске структуре. Ово захтева добро познавање хемије, и донекле искуство пожељних и типичних фрагмената молекула од биолошког значаја. За потребе ове апликације, изабран је (мали) скуп могућих мутација. Када се утврди да је потребно извршити мутацију, насумично се бира један од четири имплементирана типа, након чега се извршава конкретна, насумично одабрана, мутација изабраног типа. Мутације се чувају у посебним фајловима, и подељене су у четири категорије:

- *Замена атома* - насумично се бира један хетероатом, и са одређеним вероватноћама се конвертује у други хетероатом. На пример, уколико је изабрани атом кисеоник, он се замењује атомом азота са вероватноћом 50%, атомом сумпора вероватноћом 37.5%, а атомом фосфора са вероватноћом 12.5%.

- *Замена група* - испитује се присуство одређених функционалних група у молекулу. Функционалне групе су специфични делови молекула који диктирају његово хемијско понашање, тј. типове реакција у које ступа, али и физичка својства. Због тога је смислено мутирати овакве делове молекула. Конкретно, омогућена је замена карбонилне групе карбоксилном, карбоксилне групе амидном, карбонилне групе амино групом, као и цијано групе карбонилном. Такође, омогућена су формирања етарске или карбонилне функционалне групе између два суседна атома угљеника, али и продужетак угљеничног ланца. Постојањем последње наведене мутације, гарантује се извршење барем једне мутације група (остале мутације можда неће успети, јер се примењују једино уколико постоји специфична функционална група у структури молекула).

- *Делеција* - уколико је присутно гранање у молекулу, покушава се брисање произвољне гране. Уколико гране не постоје, или их је немогуће обрисати, извршава се брисање произвољног атома, уколико се тиме не нарушава валидност хемијске структуре молекула.

- *Инсерција* - из унапред одређеног скупа функционалних група, бира се произвољна, након чега се покушава њено уметање на произвољном месту у молекулу.

У случају да изабрани тип мутације не успе да произведе исправан трансформисан молекул након одређеног броја итерација, покушава се са другим типом мутације. Како се мутације замене група гарантовано извршавају успешно, обезбеђено је да ће у случајевима када се захтева извршење некакве мутације, заиста бити формиран модификовани молекул.

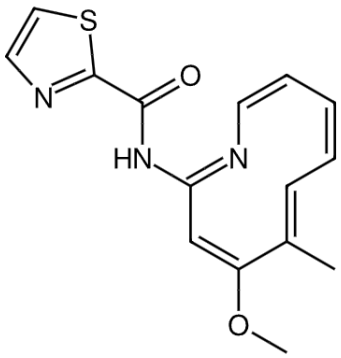
3 Експериментални резултати

3.1 Претпроцесирање и иницијални резултати

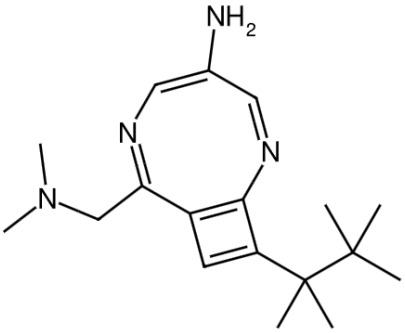
Резултате добијене применом описаног алгоритма приказујемо на подацима преузетим са (5), упоређујући их са резултатима рада (6). Скуп података садржи огроман број малих једињења преузетих из ChEMBL базе података, који нападају и везују се за протеинске тирозинске киназе.

Подаци су најпре претпроцесирани. Избачени су сви молекули који не показују биолошку активност (у колони 'class' немају вредност 'active'). Од добијених молекула, на случајан начин одабрано је 100, над којима је покренут алгоритам. Сви ови молекули су изабрани да учествују у првој генерацији, а могу се пронаћи у датотеци molecules.json у оквиру data директоријума. Као и у раду (6), број генерација постављен је на 50, број јединки које учествују у једном турниру једнак је 5, а вероватноћа мутације постављена је на 0.01. Иако се концепт елитизма не адресира у наведеном раду, за потребе нашег алгоритма, број елитних јединки постављан је на 0, 1, 2 или 3.

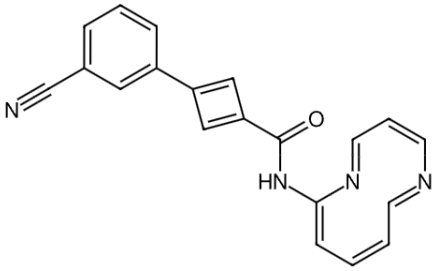
Структуре и SMILES формати пет најбољих јединки добијених покретањем овог алгоритма приказане су на сликама 4а, 4б, 4в, 4г и 4д.



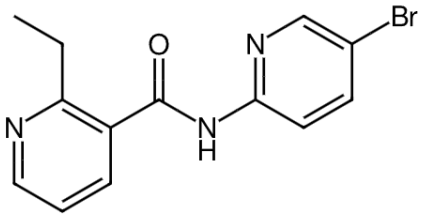
Слика 4а: Молекул 1
SMILES : COc1cc(NC(= O)c2nccs2)nccccc1C



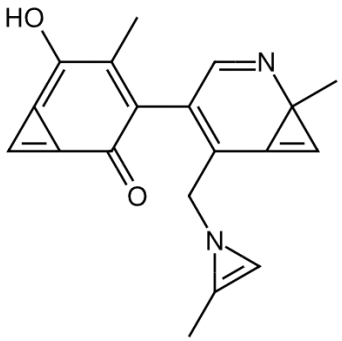
Слика 4б: Молекул 2
SMILES :
CN(C)Cc1ncc(N)cnc2c1cc(C(C)(C)(C(C)(C)C))2



Слика 4в: Молекул 3
SMILES : N#Cc1cccc(−c2cc(C(=
O)Nc3ccccncccn3)c2)c1



Слика 4г: Молекул 4
SMILES : O = C(Nc1ncc(Br)cc1)c1c(CC)nccc1



Слика 4д: Молекул 5
SMILES : Cc1c(O)c2c(c2)c(= O)c1 − c1cnc2(C)cc2c1C − n1cc(C)1

3.2 Својства молекула и упоредна анализа

У табели 1 приказана су својства генерисаних молекула. QED коефицијент је добијен самим алгоритмом, а својство Drug score добијено је применом јавно доступног алата Osiris property explorer. SA коефицијент (енг. Synthetic accessibility) објашњава колико је једињење компликовано синтетисати (већи број означава тежу синтезу), и израчунат је употребом SwissADME алата. Молекули 6, 7, 8, 9 и 10 ће се од сада па на даље односити се на резултате приказне у раду (6). Њихове структуре приказане су на сликама 5а, 5б, 5в, 5г и 5д.

Молекул	QED (Фитнес)	Drug score	SA
1	0.945	0.88	3.04
2	0.9457	0.45	4.27
3	0.9369	0.44	3.49
4	0.948	0.51	2.17
5	0.9484	0.91	5.13
6	0.93	0.76	3.78
7	0.92	0.93	3.97
8	0.86	0.87	3.06
9	0.91	0.89	4.04
10	0.88	0.87	3.81

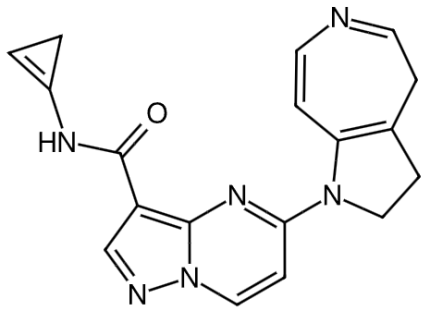
Табела 1: Својства и потенцијална лековитост генерисаних молекула

Табела 2 приказује вредности важнијих молекулских дескриптора, добијене употребом поменутог алата SwissADME. Важно је напоменути да сви добијени молекули пролазе тестове Липинског (енг. Lipinski), Госеа (енг. Ghose), Вебера (енг. Veber), Егана (енг. Egan) и Муегеа (енг. Muegge). Детаљи ових тестова такође су доступни у коришћеном софтверу.

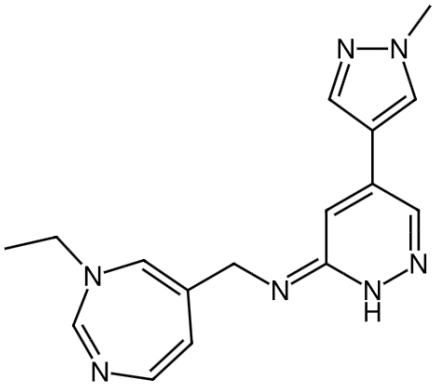
Молекул	MW	ALOGP	HBD	HBA	PSA	ROTB
1	301.36	2.48	1	4	92.35	4
2	310.52	1.88	3	4	53.32	4
3	330.38	2.65	1	4	78.67	4
4	306.16	2.46	1	3	54.88	4
5	310.39	1.86	2	3	52.57	3
6	332.37	2.20	1	4	74.89	4
7	307.36	1.80	1	4	74.46	4
8	267.29	1.57	2	2	75.66	3
9	316.36	2.96	2	2	75.01	4
10	335.42	2.00	2	2	74.72	4

Табела 2: Молекулски дескриптори генерисаних молекула. Моларна маса изражена је у $\frac{g}{mol}$, а поларна молекулска површина у $Å^2$.

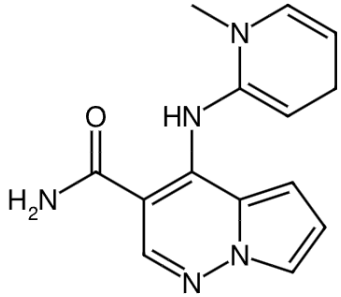
На основу изложеног, закључујемо да молекул 5 има највећи потенцијал да поседује лековита својства, мада је његова синтеза тежа од свих осталих. Уколико нам је то значајан фактор, онда се молекул 1 чини као приближно добра, али знатно приступачнија опција. У поређењу са резултатима рада (6), молекул број 5 има већи drugscore од скоро свих приказаних молекула, али је, нажалост, оцењен као релативно тежак за синтезу (присуство 1H - азиринског прстена је тешко изоловати, јер он показује тенденцију да се таутомеризује у 2H - азиридински прстен, а поред тога, потребно је формирати и два циклопропенска фрагмента, који су изузетно нестабилни због великог угаоног напона).



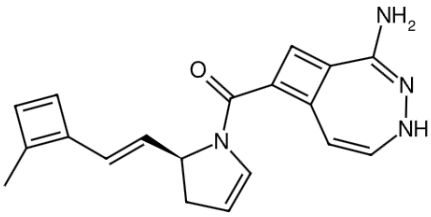
Слика 5а: Молекул 6
SMILES : C1C = NC = CC2 = C1CCN2C = 3C = CN4N = CC(C(= O)NC = 5CC = 5) = C4N = 3



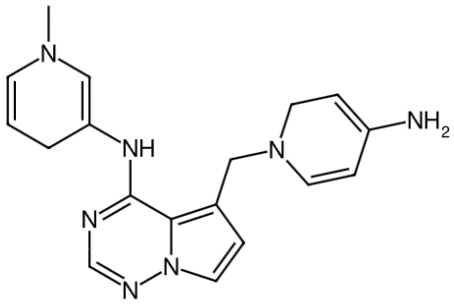
Слика 5б: Молекул 7
SMILES : C[C@@H1]N1C = NC = CC([C@@H1]N = C2C = C(C = 3C = NN(C)C = 3)C = NN2) = C1



Слика 5в: Молекул 8
SMILES : [C@@H1]N1C = CCC = C1NC2 = C(C(N) = O)C = NN3C = CC = C23



Слика 5г: Молекул 9
SMILES : NC1 = N[NH1]C = CC2 = C(C(= O)N3C = CC[C@H1]3C = CC4 = C([C@H1])C = C4)C = C12



Слика 5д: Молекул 10
SMILES : CN1C = CCC(NC2 = NC = NN3C = CC(CN4C = CC(N) = CC4) = C23) = C1

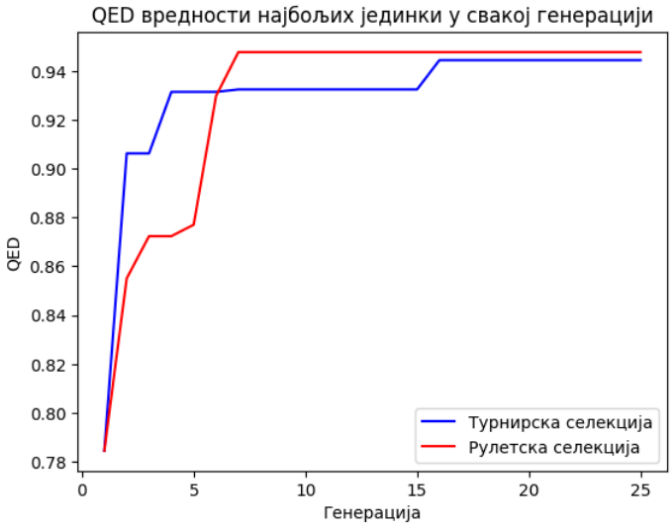
Интересантно је упоредити структуре молекула 1-5 са структурама 6-10. Иако суштински доста различите, ипак постоје неке заједничке карактеристике, што и није сувише необично, будући да су алгоритми извршавани при сличним условима, а могуће је и поклапање одређених јединки које су одабране за полазну генерацију. Азот је, поред кисеоника, најзаступљенији хетероатом у органским једињењима, а поготово оним са лековитим својствима. Због тога не чуди овакав атомски састав добијених молекула (велики број амидних и амино група, као и хетероцикала са азотом). Оно што смањује квалитет добијених молекула јесте наиван приступ укрштања који је имплементиран у овим радовима. Једнопозиционо (или двопозиционо) укрштање често доводи до формирања прстенова димензије 3 или 4. Иако су засићени четворочлани прстенови неретка појава у хемијским структурама лекова, постојање таквих подструктура смањује стабилност молекула и отежава његову синтезу, због огромног угаоног напона. Угаони напон је утолико већи уколико је у малом прстену присутно и неко додатно незасићење, у виду двоструких веза. Имајући то у виду, није зачуђујућа чињеница да се молекули који садрже циклопропенски или циклобутадиенски прстен (као што су молекули 2, 5 и 9) значајно теже синтетишу, што се јасно уочава кроз

повишене вредности њихових SA коефицијената.

С друге стране, уочава се присуство релативно дугачких конјугованих система у структурама ових молекула, који повољно утичу на његову стабилност. Конјуговани системи представљају непрекидан низ наизменичних једноструких и двоструких веза. То на пример узрокује повишену вредност Drug score и снижену вредност SA коефицијента молекула 1.

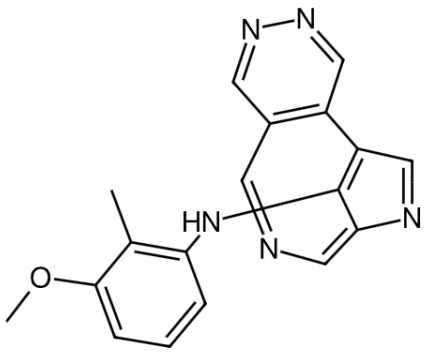
3.3 Турнирска против рулетске селекције

Размотримо сада разлику између добијених резултата када се примењују турнирска и рулетска селекција. Почетни скуп података је остао непромењен, број генерација је 25, број јединки које учествују у једном турниру једнак је 5 (у случају турнирске селекције), број елитних јединки у генерацији је 2, а вероватноћа доласка до мутације износи 0.05. График на слици 6 приказује тренд промене QED вредности најбољих јединки у свакој од 25 генерација, за оба поменута типа селекције.

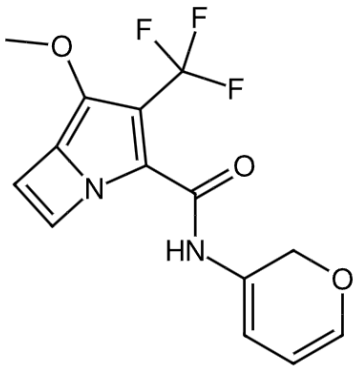


Слика 6: QED вредности најбољих јединки кроз генерације, при турнирској, односно рулетској селекцији

Показује се да се рулетском селекцијом постигао QED резултат од 0.9477, што је за нијансу боље од 0.9444, колико износи QED резултат добијен турнирском селекцијом. На сликама 7а и 7б су приказане структуре најбољих јединки које су добијене претходно описаном анализом.



Слика 7а: Молекул добијен турнирском селекцијом
SMILES : N#Cc1cccc(−c2cc(C(=O)Nc3ccccncccn3)c2)c1
QED : 0.9444
Drugscore : 0.76
SA : 3.67

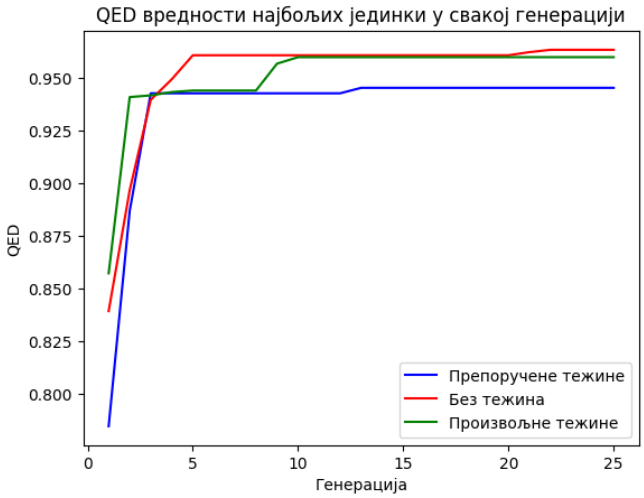


Слика 7б: Молекул добијен рулетском селецкијом
SMILES : COc1c(C(F)(F)F)c(C(=O)Nc2cccOC2)n2ccc12
QED : 0.9477
Drugscore : 0.36
SA : 4.08

3.4 Утицај тежина молекуларних дескриптора

У одељку 2.2 објаснили смо да се QED коефицијент рачуна као тежинска геометријска средина вредности функција молекуларних дескриптора. Развијени апликативни интерфејс пружа могућност конфигурације тежина ових параметара. Испитајмо разлике у QED вредностима најбољих јединки кроз генерације у случајевима када се

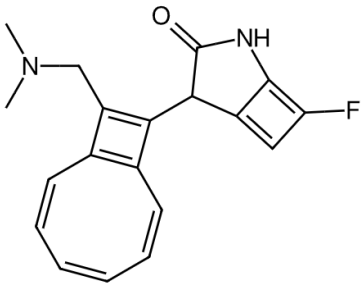
користе препоручене (подразумеване) вредности тежина, када се све тежине поставе на 1 (односно када QED коефицијент рачунамо као нетежинску геометријску средину вредности функција молекуларних дескриптора), и када се за тежине узму произвољне вредности. Добијени резултати приказани су на графику на слици 8. Почетни скуп података је остао непромењен, број генерација је 25, број јединки које учествују у једном турниру једнак је 10 (одабрана је турнирска селекција), број елитних јединки у генерацији је 2, а вероватноћа доласка до мутације износи 0.05.



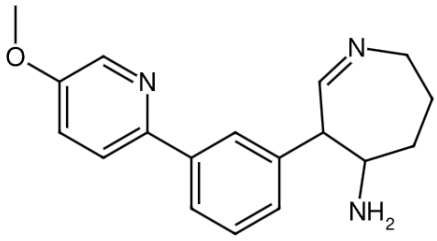
Слика 8: QED вредности најбољих јединки кроз генерације, при различитим вредностима тежина функција молекуларних дескриптора. За произвољне вредности тежина (зелена линија) изабране су следеће вредности: MW : 0.66, ALOGP : 0.9, HBA : 0.38, HBD : 0.38, PSA : 0.9, ROTB : 0.62, AROM : 1.0, ALERTS : 0.5

Важно је напоменути да приказане вредности QED коефицјената на претходном графику служе само за уочавање и упоређивање интервала у којима се ти коеифицијенти крећу у зависности од узетих тежина. Другим речима, несигуран би био закључак да се након 25 генерација најбоља јединка добија уколико све тежине имају вредност један (црвена линија на претходном графику).

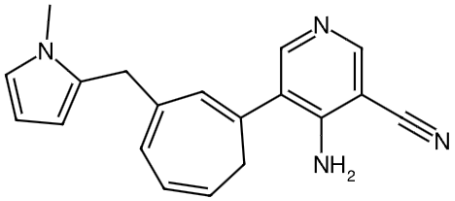
На сликама 8а, 8б и 8в су приказане структуре најбољих јединки у последњој (25.) генерацији добијених претходно описаном анализом.



Слика 8а: Случај препоручених тежина
SMILES : CN(C)Cc1c(C2(=O)Nc3c(F)cc32)c2c1cccccc2



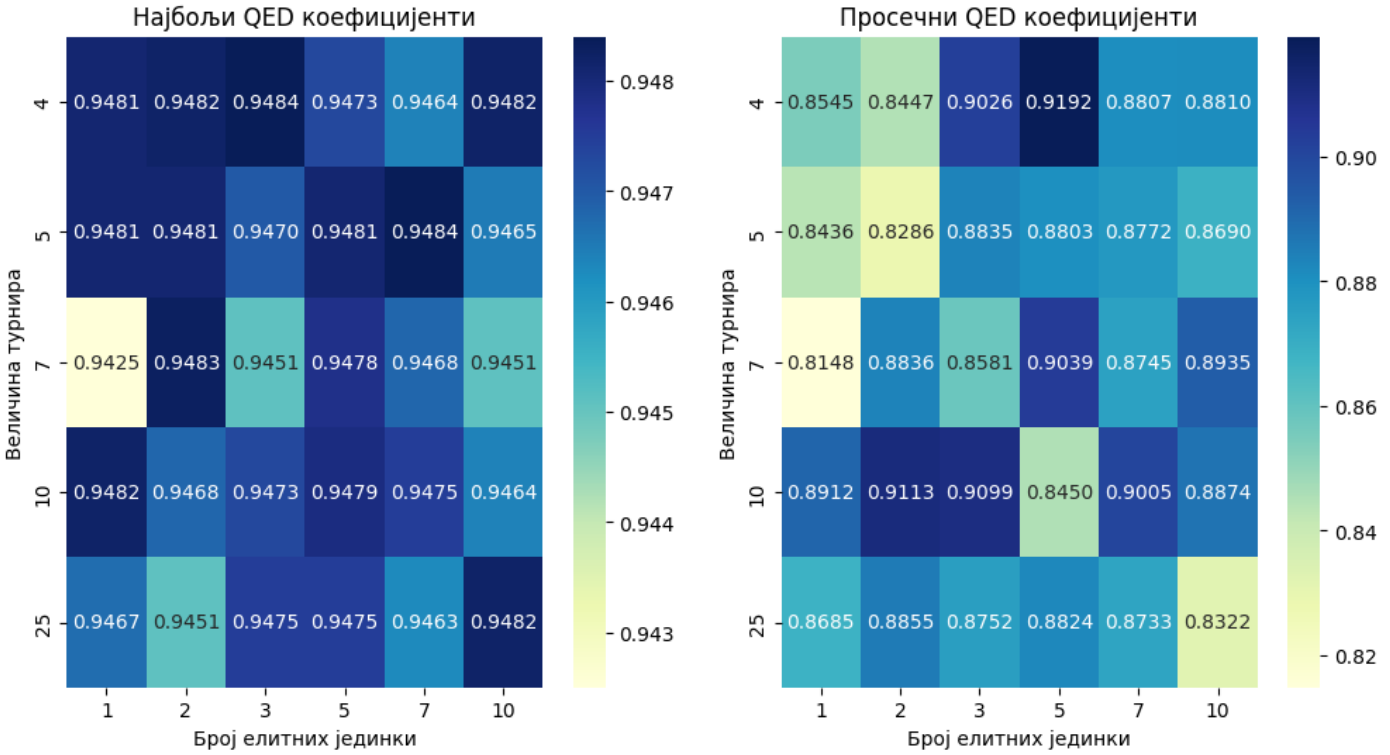
Слика 8б: Случај без тежина
SMILES : Nc1C(c2cc(−c4ccc(OC)cn4)ccc2)cnCcC1



Слика 8в: Случај произвољних тежина
SMILES : NCc1cncc(−c2cc([C@@H]c3cccn3C)ccc[C@@H]2)c(N)1

3.5 Општи утисак

Како бисмо стекли општи утисак о квалитету и стабилности имплементираног алгоритма, извршили смо га укупно 30 пута са различитим вредностима полазних параметара, над посебним скупом молекула, који се може пронаћи у датотеци molecules_set2.json у оквиру data директоријума. Број генерација је увек постављан на 50, а за вероватноћу доласка до мутације узиман је произвољан број између 0.02 и 0.15. За величину турнира узете су вредности 4, 5, 7, 10 и 25, а за број елитних јединки 1, 2, 3, 5, 7, 10, и свих 30 комбинација ових вредности представљале су улазе за 30 покретања алгоритма. Резултати који су овом приликом добијани приказани су на две топлотне мапе на слици 9. Лева топлотна мапа приказује QED вредности најбоље јединке у 50. генерацији, док друга представља просечну QED вредност јединки 50. генерације.



Слика 9: Најбоље и просечне QED вредности јединки из последње генерације

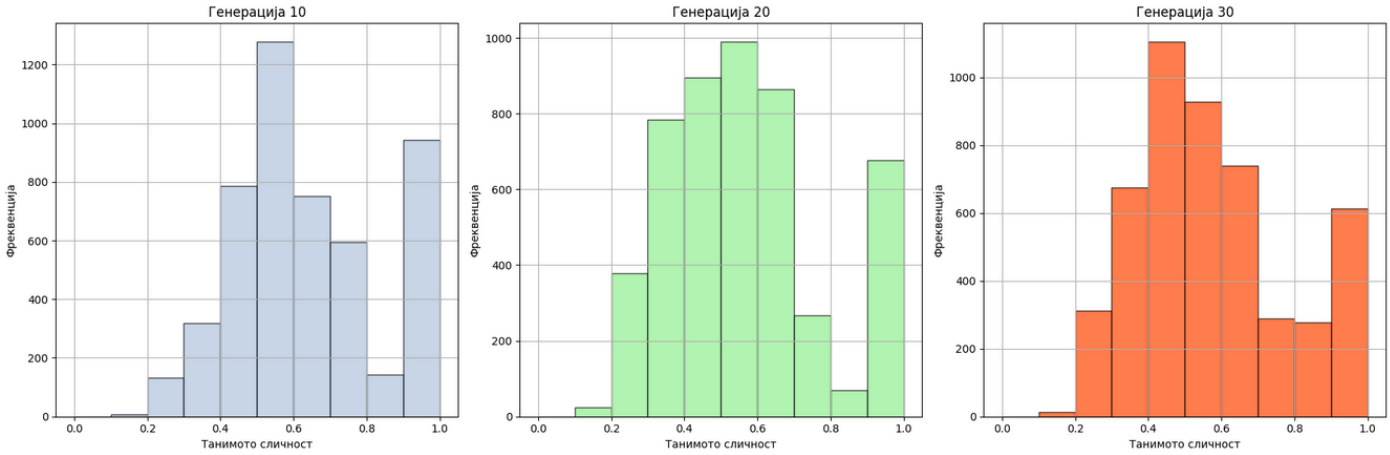
Просечно измерено време за једно извршавање претходно описаног поступка износи 4 минута и 33 секунде.

3.6 Диверзитет

Испитајмо диверзитет јединки једне популације. Да бисмо квантификовали диверзитет, неопходно је да за почетак дефинишемо меру којом можемо упоредити колико су две јединке, односно два молекула, представљена SMILES форматом, слична. Ово је типичан задатак који се јавља у огромном броју хемоинформатичких проблема, те су алгоритми којима се ти задаци решавају имплементирани у оквиру библиотека као што је RDKit. На основу SMILES ниске, којом је пренета комплетна информација о атомима у оквиру молекула и начинима на које су они повезани, генеришу се *отисци молекула*, по аналогији са отисцима прстију. Претражују се и лоцирају типичне градивне подструктуре у оквиру молекула. Што је већи број заједничких подструктура два молекула, већи је и њихов коефицијент сличности, који називамо *Танимото коефицијент сличности.*(7) Другим речима, Танимото коефицијент сличности два молекула се може израчунати као:

$$T_{\text{Танимото}}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

где је A скуп подструктура првог, а B скуп подструктура другог молекула. Вредност овог коефицијента је увек између 0 и 1, где 0 означава да су молекули у потпуности различити, док 1 означава идентичне молекуле. Када имамо читаву популацију јединки, можемо израчунати сличност сваког пара јединки. Управо је на тај начин имплементирано испитивање сличности јединки једне популације у нашем алгоритму. Добијени резултати приказани су хистограмима на слици 10. Укупан број генерација је 30, број јединки које учествују у једном турниру је 10, број елитних јединки износи 2, док је вероватноћа доласка до мутације постављена на 0.05. Диверзитет је испитиван у 10, 20, и 30. генерацији.



Слика 10: Диверзитет јединки кроз генерације

У табели 3 приказане су дескриптивне статистике података представљених хистограмима.

Генерација	Математичко очекивање	Стандардна девијација
10	0.6446	0.2114
20	0.5685	0.2171
30	0.5783	0.2149

Табела 3: Дескриптивне статистике диверзитета популације

Показује се да је Танимото коефицијент благо повишен, односно да јединке показују тенденцију ка конвергенцији. Но, овакво понашање треба узети условно, и не треба га генерализовати, будући да диверзитет популације зависи од почетних параметара генетског алгоритма. На пример, вредности Танимото коефицијента се могу смањити повећањем вероватноће доласка до мутације, или смањењем броја елитних јединки. Добијене вредности диверзитета (које се групишу око 0.6), сматрају се горњом границом популације са добрим степеном диверзитета.

4 Закључак

Генетски алгоритми представљају моћан алат за откривање и оптимизацију молекулских структура са потенцијалним лековитим својствима, те су због тога честа техника која се користи у овом делу хемоинформатике. Простор претраге је бесконачан, те се употребом раличитих хеуристика, алгоритам усмерава ка молекулским структурама које су више обећавајуће. Ове хеуристике омогућавају алгоритму да ефикасније претражује простор и идентификује молекуле које имају веће шансе да задовоље специфичне критеријуме, као што су стабилност, биолошка активност и биокомпатибилност.

Развијена апликација пружа једноставан интерфејс којим се спровођење овог алгоритма и конфигурисање одговарајућих параметара знатно олакшава истраживачима. Визуализација молекулских структура кроз генерације је такође изузетно корисна, јер омогућава праћење структурних трансформација које се дешавају током еволуције популације. На овај начин, корисници могу лакше уочити промене у структури молекула и пратити како се оне развијају према оптималнијим решењима. Јасно истакнута структура најбоље јединке у оквиру сваке генерације омогућава прецизно препознавање најуспешнијих молекула, чиме се олакшава даља анализа.

Приступ који смо користили за фитнес функцију је прилично наиван, али и даље, пружа корисне резултате. Могућа је употреба другачијих, компутационо сложенијих функција. На пример, уколико се посматрају молекули који би се понашали као лиганди који се везују за одређени протеин, може се испитивати афинитет таквог везивања и то користити као функција прилагођености јединке. Овај приступ, који се заснива на молекуларној динамици и симулацијама молекуларног моделирања, могао би додатно побољшати резултате, омогућавајући прецизнију селекцију молекула са вишим потенцијалом за биолошку активност. Такође, важан аспект који би се могао узети у обзир јесте SA коефицијент, будући да он описује тежину извршавања самог процеса од идеје до реализације. Пожељно је и њега укључити у израз којим се дефинише функција прилагођености.

Уколико располажемо моћнијим рачунарским ресурсима, алгоритам би могао бити извршаван кроз већи број генерација, чиме би се добили још квалитетнији и прецизнији резултати. Такође, извршавањем алгоритма већи број пута, могућа је претрага ширег простора могућих решења, што може довести до открића нових молекула са потенцијалним лековитим својствима која нису била евидентна у претходним итерацијама.

С обзиром на брз развој рачунарске хемије и примене машинског учења у области хемоинформатике, генетски алгоритми представљају само један од алата који може бити употпуњен другим напредним техникама, као што су дубоко учење и симулације молекуларне динамике. Комбиновањем ових метода, могуће је додатно побољшати резултате и омогућити бржу и ефикаснију идентификацију нових кандидата за лекове, што је од кључне важности за убрзање процеса који тек предстоји, а то су осмишљавање синтезе и извршавање свих неопходних клиничких испитивања.

5 Литература

[1] Bickerton, G., Paolini, G., Besnard, J. et al. Quantifying the chemical beauty of drugs. Nature Chem 4, 90–98 (2012). https://doi.org/10.1038/nchem.1243

[2] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Del. Revs. 1997; 23 : 3–25. doi : 10.1016/s0169 − 409x(00)00129 − 0.

[3] Gohlke, H. and Klebe, G. (2002), Approaches to the Description and Prediction of the Binding Affinity of Small − Molecule Ligands to Macromolecular Receptors. Angewandte Chemie International Edition, 41 : 2644 − 2676.

[4] Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2012 Jan 24; 4(2) : 90 − 8. doi : 10.1038/nchem.1243. PMID : 22270643; PMCID : PMC3524573.

[5] Romero R. Tyrosine Kinases ligands with bioactivity data. Kaggle; 2024. https : //www.kaggle.com/dsv/7465979.

[6] Romero, Ricardo. "Integration of Genetic Algorithms and Deep Learning for the Generation and Bioactivity Prediction of Novel Tyrosine Kinase Inhibitors." arXiv preprint arXiv : 2408.07155 (2024).

[7] Bajusz, David, AnitaRacz, and Karoly Heberger. "Why is Tanimoto index an appropriate choice for fingerprint − based similarity calculations?." Journal of cheminformatics 7 (2015) : 1 − 13.