УДК 541.6:547.789:615.076

ВИВЧЕННЯ ВЗАЄМОЗВ'ЯЗКУ «СТРУКТУРА – ПРОТИПУХЛИННА АКТИВНІСТЬ» ПОХІДНИХ 4-ТІАЗОЛІДИНОНІВ МЕТОДАМИ РЕГРЕСІЙНОГО АНАЛІЗУ ТА КЛАСИФІКАЦІЙНОГО МОДЕЛЮВАННЯ

Б.С.Зіменковський, О.Т.Девіняк*, Р.Б.Лесик

Львівський національний медичний університет ім. Данила Галицького
79010, м. Львів, вул. Пекарська, 69. E-mail: dr_r_lesyk@org.lviv.net
* Державний вищий навчальний заклад «Ужгородський національний університет»

Ключові слова: 4-тіазолідинони; протипухлинна активність; QSAR; множинна лінійна регресія, покрокова регресія; алгоритм; Random Forest

Проведено багатофакторний лінійний регресійний аналіз та класифікаційне моделювання (алгоритм Random Forest) протипухлинної активності похідних 4-тіазолідинонів з метою спрямованого синтезу сполук з протиракovою дією. Прогностична здатність підтверджена статистичними показниками, отриманими в процесі створення моделі, та при прогнозі тестової вибірки. Здійснено аналіз вкладів дескрипторів у прийняття рішення, визначено найбільш важливі з них для проведення класифікації, розкрито їх фізико-хімічний зміст та проілюстровано зв'язок виділених молекулярних дескрипторів зі здатністю до інгібування росту онкоклітин.

THE STUDY OF THE «STRUCTURE – ANTICANCER ACTIVITY» RELATIONSHIP OF 4-THIAZOLIDINONES USING REGRESSION ANALYSIS AND CLASSIFICATION MODELING
B.S.Zimenkovsky, O.T.Devinyak, R.B.Lesyk
In order to perform directed synthesis of antineoplastic compounds the multiple linear regression analysis and classification modeling (by Random Forest algorithm) of 4-thiazolidinones as anticancer agents have been carried out. The predictive ability is confirmed by the statistic data obtained during model developing and test sample predicting. Descriptors contribution in the decision making has been estimated, the most important descriptors for the classification have been identified and interpreted. The relationship between the molecular descriptors selected and cancer cells growth inhibition is given.

ИЗУЧЕНИЕ ВЗАИМОСВЯЗИ «СТРУКТУРА – ПРОТИВООПУХОЛЕВАЯ АКТИВНОСТЬ» ПРОИЗВОДНЫХ 4-ТИАЗОЛИДИНОНОВ МЕТОДАМИ РЕГРЕССИОННОГО АНАЛИЗА И КЛАССИФИКАЦИОННОГО МОДЕЛИРОВАНИЯ
Б.С.Зименковский, О.Т.Девиняк, Р.Б.Лесык
Проведен многофакторный линейный регрессионный анализ и классификационное моделирование (алгоритм Random Forest) противоопухолевой активности производных 4-тиазолидинонов с целью целенаправленного синтеза соединений с противораковым действием. Прогностическая способность подтверждена статистическими показателями, полученными в процессе создания модели и при прогнозе тестовой выборки. Осуществлен анализ вкладов дескрипторов в принятие решения, определены наиболее важные из них для проведения классификации, раскрыто их физико-химическое значение и проиллюстрирована связь выделенных молекулярных дескрипторов со способностью к ингибированию роста онкоклеток.

Пошук протипухлинних засобів серед похідних 4-тіазолідинонів є одним з перспективних напрямків наукових досліджень у хімії лікарських засобів [1-6]. Використання з цією метою засобів математичного моделювання та *in silico* скринінгу успішно себе зарекомендувало отриманими результатами [7-9]. Об'єктом нашого дослідження є бібліотека похідних 4-азолідонів та споріднених гетероциклічних сполук, синтезованих на кафедрі фармацевтичної, органічної і біоорганічної хімії ЛНМУ ім. Данила Галицького, яка серед 5000 сполук містить 1076 похідних (402 з протипухлинною активністю різного рівня), протестованих в Національному інституті раку США (NCI) [1, 2]. Попередньо нами було здійснено класифікацію 4-тіазолідинонів за механізмами протипухлинної активності та виявлено 3 класи (А, Б і В) активних сполук [10]. Також сформовано три масиви речовин для дослідження та встановлено, що сполуки, при дії яких середній відсоток росту ракових клітин (GP) складає >86%, слід визначити як неактивні. У запропо-

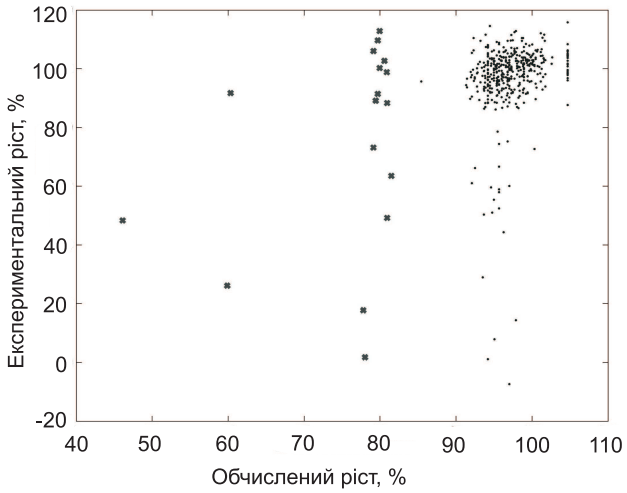


Рис. 1. Проекція експериментальних результатів для масиву А на вісь обчислених за допомогою QSAR-моделі.

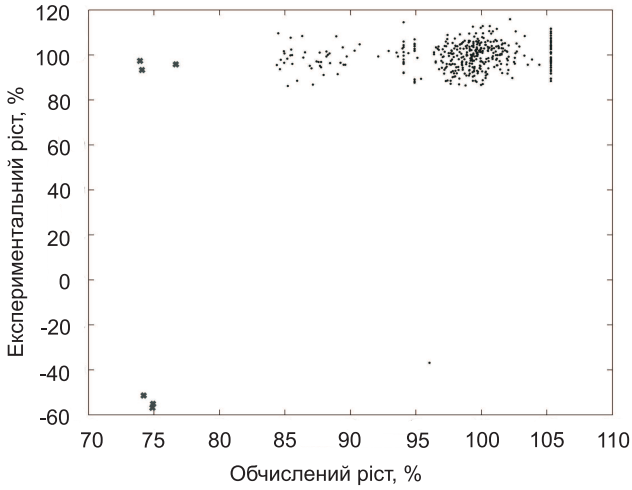


Рис. 2. Проекція експериментальних результатів для масиву Б на вісь обчислених за допомогою QSAR-моделі.

нованій статті продовжено дослідження зв'язку між структурою та біологічною активністю наведених вище класів за допомогою багатофакторного лінійного регресійного аналізу та класифікації з використанням методу *Random Forests*.

1. Багатофакторний лінійний регресійний аналіз

Багатофакторний лінійний регресійний аналіз полягає у створенні математичної моделі зв'язку між залежною змінною (активністю) та лінійною комбінацією незалежних змінних (дескрипторів). Метою такого аналізу є визначення ступеня детермінованості зв'язку та внесків окремих незалежних змінних у варіацію залежної, а також передбачення значення активності сполуки за допомогою її молекулярних дескрипторів. Дескриптори були обчислені за допомогою програми E-Dragon (>1600 дескрипторів) [12] після попередньої оптимізації геометрії молекул у силовому полі MMFF94 [13]. В одержаному первинному масиві здійснювався пошук помилок обчислення з наступним видаленням відповідного рядка чи стовпця, керуючись принципом мінімізації втраченої інформації. Крім того, видалено константні та попарно корельовані ($|r|>0,95$) дескриптори. Побудова регресійної функції здійснювалась за допомогою методу покрокової регресії [11].

Таким чином одержано функцію для масиву А:

$$GP = (3.27\pm25.07)-(21,38\pm3.03)*'ARR'+ +(33.84\pm8.31)*'H0u'-(19.25\pm3.68)*'reactive' \quad (1)$$
$$r^2=12.9,\ RMSE=14.20,\ F=22.13,\ p=2.2*10^{-13}$$

Низька точність моделі означає слабку лінійну залежність активності аналізованих сполук від будь-якої комбінації із обчислених дескрипторів. При проектуванні експериментальних результатів на вісь розрахованих (рис. 1) отримана QSAR-модель виділяє групу, що містить 7 активних та 10 неактивних сполук. Тобто застосування цієї моделі дає шанс дещо менший, ніж 7:10 (похибка передбачення повинна бути більшою за похибку регресії) здійснити синтез активної сполуки класу А, що приблизно в 10 разів краще за здійснення випадкового синтезу (28:426).

Регресійна функція для масиву Б має наступний вигляд:

$$GP = (105.31\pm1.45)-(10.41\pm2.33)*'nR= =Ct'-(11.26\pm2.36)*'nC= =N-N<'-(0.026\pm0.006)*'ASA+' \quad (2)$$
$$r^2=12.6,\ RMSE=14.51,\ F=20.47,\ p=2.0*10^{-12}$$

Дана функція виокремлює групу, в якій 3 активні та 3 неактивні сполуки (рис. 2). Шанс коректного передбачення є дещо меншим за 1:1.

Регресійна функція для масиву В є наступною:

$$GP = (102.86\pm1.00)+(0.045\pm \pm0.009)*\ 'T(O..S)'-(13.56\pm3.09)*\ 'nArCOSR'- -(0.091\pm0.015)*'PEOE\_VSA+1' \quad (3)$$
$$r^2=14.0,\ RMSE=8.11,\ F=23.22,\ p=6.0*10^{-14}$$

Наведена функція виокремлює групу, в якій 2 активні та 7 неактивних сполук (рис. 3). Шанс коректного передбачення дещо менше за 2:7.

Жодна зі створених QSAR-моделей не задовольняє загальноприйнятих критеріїв точності, хоча значення критерію Фішера (*F*) перебувають на задовільному рівні. Це зумовлено наявністю неактивних сполук із значеннями дескрипторів, дуже близькими до дескрипторів активних молекул. Внаслідок цього спостерігається ефект «важеля», тобто зменшення похибки моделі для активних сполук одночасно підвищує похибку наближених до них неактивних. З огляду на те, що одержані моделі проявляють тенденцію до про-

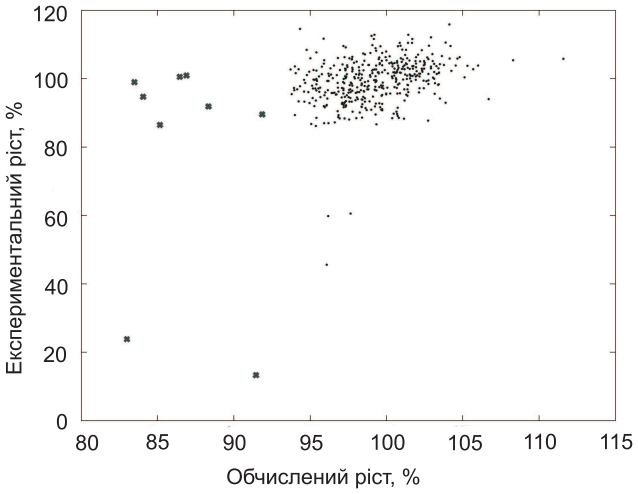


Рис. 3. Проекція експериментальних результатів для масиву В на вісь обчислених за допомогою QSAR-моделі.

стого розділення сполук (на активні та неактивні) замість точного встановлення відсотків інгібування росту, можна зробити висновок, що адекватний опис зв'язку між структурою та активністю можливий в рамках класифікаційних моделей.

2. Класифікаційне моделювання протипухлинної активності похідних тіазолідинону

Призначенням класифікаційних моделей є розпізнавання самого факту прояву біологічної дії сполуками, тобто розділення активних та неактивних молекул. Для їх створення було використано метод *Random Forest*, за допомогою якого здійснюється побудова ансамблю дерева рішень [14]. Згідно з технічною документацією методу необхідності у проведенні перехресної валідації чи виділення окремої тестової вибірки для встановлення прогностичної здатності моделі немає, оскільки кожне дерево будується із використанням випадкової вибірки із генеральної сукупності, отриманої за допомогою техніки бутстрепінгу (*bootstraping*). Одержані при цьому вимірники похибки для сполук, що не ввійшли в навчальну вибірку (*out-of-bag compounds*), є адекватними і неупередженими [15]. Модель *Random Forest* містить значну кількість дерев рішень, побудованих на різних вибірках із використанням випадкових дескрипторних підпросторів для формування кожного вузла дерева, тому будь-яке дерево потенційно спроможне охопити усі наявні механізми активності за поділом у різних вузлах. Враховуючи наведене, було прийнято рішення щодо створення єдиної моделі на основі всіх відомих та доступних результатів тестування протипухлинної дії 4-тіазолідинонів. Таким чином, навчальна вибірка містила 674 сполуки. Геометрія цих сполук була оптимізована в силовому полі MMFF94 [13], а >1600 дескрипторів було обчислено за допомогою E-Dragon [12]. Помилки обчислення дескрипторів видалялись таким чином, щоб мінімізувати втрачену інформацію. Крім того, константні та попарно корельовані ($|r|>0,95$) дескриптори також були видалені з масиву даних. Таким чином, залишилось 739 дескрипторів для здійснення моделювання. На основі середнього відсотка росту онкоклітин GP = 86% як межі між класами «активних» і «неактивних» сполук та параметрів алгоритму по замовчуванню (500 дерев рішень та число дескрипторів для здійснення розгалуження $mtry = round(\sqrt{N_{desc}})$ = 27, де N_{desc} – розмірність простору дескрипторів) сформовано модель *Random Forest* із статистичними показниками: частота помилок (*Error Rate*) = 16,62%, чутливість моделі = 22,8%, специфічність = 98,15%, кореляційний коефіцієнт *MCC* (*Matthew's correlation coefficient*) = 0,273.

Ця модель стала відправним пунктом для подальшої поетапної оптимізації параметрів алгоритму, зокрема, з метою зниження частоти помилок та підвищення чутливості моделі. Спочатку досліджувались зміни ефективності моделі при переміщенні межі між класами із кроком 4% при фіксованому числі дескрипторів для здійснення розгалуження = 27 та зменшеній (з метою підвищення швидкодії) кількості дерев у моделі до 100. Одержані результати (рис. 4) не дають змоги чітко визначити межу, при якій модель характеризується найкращими статистичними показниками, однак дозволили визначити інтервали оптимальності. Наступний етап оптимізації полягав у дослідженні поведінки моделі при зміні кількості молекулярних дескрипторів (крок 10 дескрипторів) для здійснення розгалуження, приймаючи як межу між класами послідовно усі цілочисельні значення із знайденого інтервалу оптимальності. Отримані результати показали, що кращі статистичні параметри досягаються в тому випадку, коли активними вважаються сполуки із GP<20%. Залежність ефективності моделі із 100 дерев рішень від кількості дескрипторів для здійснення розгалуження при фіксованій межі між класами = 20% (рис. 5). У знайденому інтервалі оптимальності [170, 190] проводилось додаткове дослідження ефективності моделі зі зміною кількості дескрипторів на одиницю. Таким чином, було встановлено, що модель досягає найкращих показників при використанні випадкового 180-вимірного підпростору дескрипторів для прийняття рішення в кожному розгалуженні дерева. На останньому етапі здійснювалась оптимізація кількості дерев у моделі *Random Forest*. При цьому знайдено декілька рішень даної задачі, одне з яких дорівнює 100. В результаті одержано кінцеву оптимальну модель при наступних параметрах: ме-

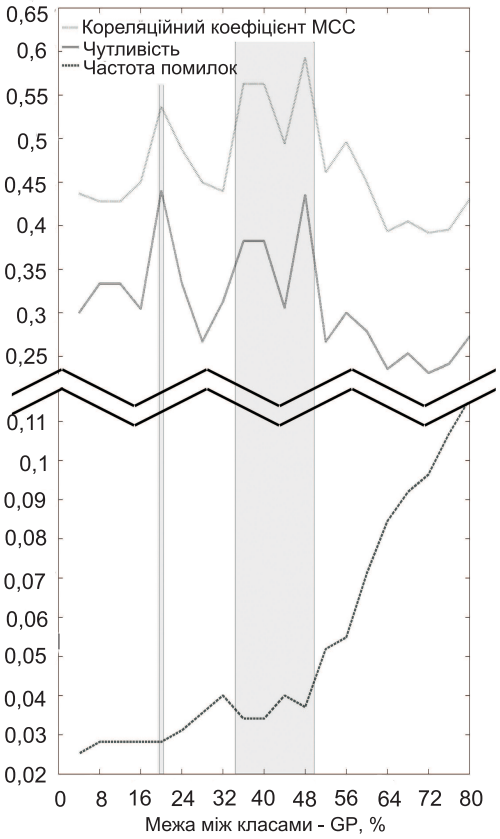


Рис. 4. Зміна статистичних параметрів моделей Random Forest при переміщенні межі між класами. Інтервали оптимальності виділені сірим кольором.

жа між класами = 20%, кількість дескрипторів для здійснення розгалуження = 180, кількість дерев рішень = 100. Статистичні показники цієї моделі наступні: частота помилок (*Error Rate*) = 2,077%, чутливість моделі = 56,00%, специфічність = 99,64%, кореляційний коефіцієнт *MCC* (*Matthew's correlation coefficient*) = 0,684. Варто зазначити, що завдяки проведеній оптимізації у 8 разів зменшено частоту помилок, у 2,5 рази підвищено чутливість та кореляційний коефіцієнт *MCC*.

З метою зовнішньої валідації моделі на основі відкритої бази даних результатів NCI DTP було сформовано вибірку із 129 сполук, що містять структурний фрагмент 4-тіазолідинону. Частота похибки класифікації моделі при цьому склала 1,55%.

Наступним кроком було проведення аналізу вкладів дескрипторів у прийняття рішення. Важливість кожного дескриптора обчислювалась як середнє значення різниці між відсотком правильних відповідей *i*-го дерева при незмінному стані та при випадковому перемішуванні цього дескриптора. Результати не виявили дескрипторів із високими рівнями важливості, тобто в даному масиві дескрипторів не існує невеликої комбінації, здатної розділити активні та неактивні сполуки, а вірне розпізнавання здійснюється за участю багатьох факторів. Це, власне, також є од-

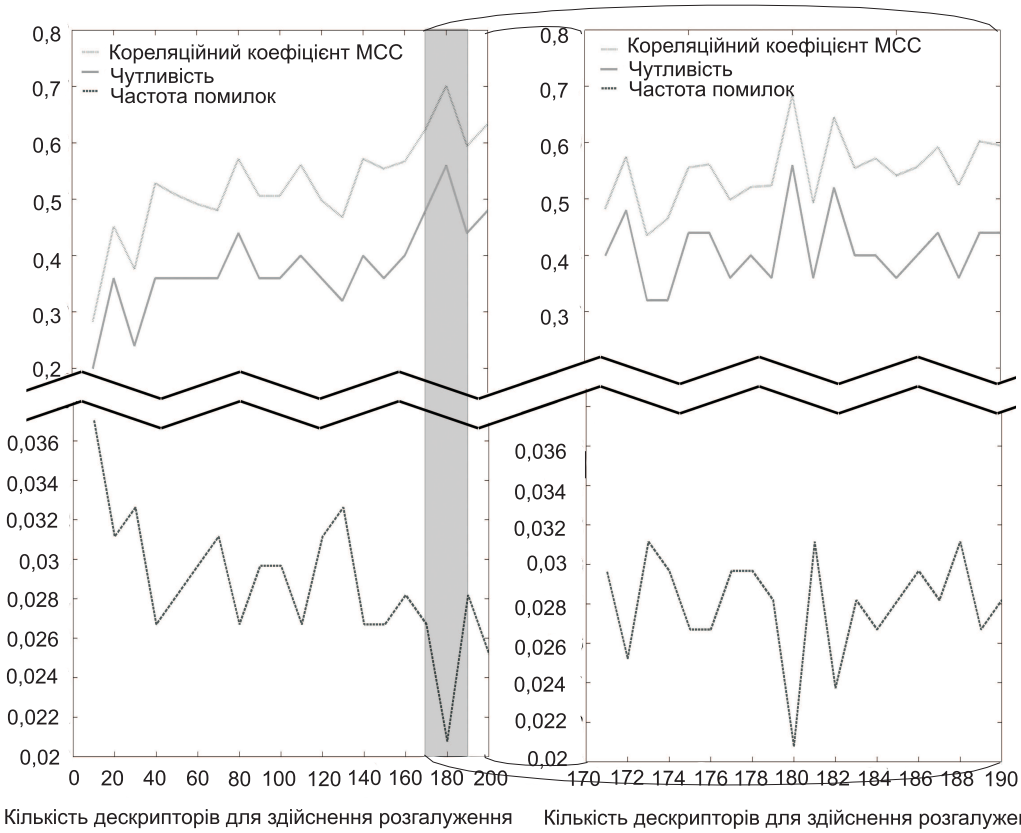


Рис. 5. Залежність статистичних параметрів моделей Random Forest від кількості дескрипторів для здійснення розгалуження. Інтервали оптимальності виділені сірим кольором.

Таблиця

Три найважливіші дескриптори моделі *Random Forests*

Дескриптор	Важливість	Частка важливості
D/Dr05 –індекс відстані/обходу циклів 5-го порядку	0,0069	12,65%
BEHv2 – друге найвище власне значення матриці Бурдена, зваженої атомними масами	0,0052	9,65%
VEA1 – сума коефіцієнтів власного вектора матриці суміжності, асоційованого з найменшим власним значенням	0,0040	7,30%

нією з причин низької точності лінійних регресійних функцій. Найважливіші три дескриптори наведені у таблиці.

Вказані дескриптори належать до топологічних і характеризують взаємне розміщення атомів у молекулі. Дескриптор D/Dr05 обчислюється за допомогою матриці частки відстань/обхід (D/Δ), що є симетричною матрицею A×A (де A – кількість атомів у молекулі), недіагональними елементами якої є відношення довжин найкоротшого до найдовшого шляху між будь-якою парою вершин. Вона визначається як

$$[D/\Delta]_{ij} = \begin{cases} \dfrac{d_{ij}}{\Delta_{ij}} & якщо \;\; i \neq j \\ 0 & якщо \;\; i = j \end{cases}, \qquad (4)$$

де d_{ij} та Δ_{ij} – топологічна та обхідна відстані між вершинами ν_i та ν_j відповідно. D/Dr05 являє собою суму рядків матриці D/Δ, що відповідають вершинам 5-членного циклу, і відображає локальне геометричне оточення 5-членних циклів. Дескриптор BEHv2 є другим найвищим власним значенням матриці Бурдена, зваженої атомними масами. Матриця Бурдена B, в свою чергу, визначається наступним чином: діагональні елементи B_{ii} є атомними номерами відповідних атомів, недіагональні елементи B_{ij}, що характеризують два зв'язані атоми *i* та *j*, є рівними умовному порядку зв'язку, що становить 0,1, 0,2, 0,3 та 0,15 для одинарного, подвійного, потрійного та ароматичного зв'язку відповідно; елементи матриці, що відповідають кінцевим зв'язкам, збільшуються на 0,01; усі інші елементи матриці, що відповідають незв'язаним атомам, рівні 0,001. VEA1 обчислюється як сума коефіцієнтів власного вектора матриці суміжності, асоційованого з найбільш від'ємним власним значенням. Матриця суміжності являє собою квадратну матрицю, недіагональні елементи якої рівні одиниці, якщо між відповідною парою атомів існує зв'язок, в іншому ж випадку елементи рівні нулю. На відміну від D/Dr05 ці обидва дескриптори описують геометрію молекули в цілому і застосовуються як міри близькості топологічних структур різних молекул [16].

Для ілюстрації зв'язку даних дескрипторів із задачею класифікації обчислено їх прототипи за наступним алгоритмом. На основі матриці попарних близькостей між сполуками у моделі (обчисленої за алгоритмом, наведеним у технічній документації методу [15]) для кожного з класів знайдено таку сполуку, що має найбільше число представників свого класу серед *k* найближчих сусідів. Враховуючи сильну асиметрію у розмірах класів, *k* встановлювався для кожного з класів окремо і дорівнював половинній кількості сполук у класі. Серед цих *k* сполук для кожного дескриптора обчислювалась медіана, верхній та нижній квартилі. Таким чином, отримана медіана є прототипом класу, а квартилі дають оцінку його стабільності. Ці характеристики були стандартизовані шляхом віднімання мінімуму значень дескрипторів та наступного ділення на різницю між максимальними та мінімальними значеннями. Порівняльний аналіз прототипів (рис. 6) дає змогу охарактеризувати вклад наведених дескрипторів у прийняття рішення щодо віднесення сполуки до того чи іншого класу.

Відсутність перетину границь інтерквартильних проміжків різних класів свідчить про достатню роздільну здатність вибраних молекулярних дескрипторів. Крім того, високі значення D/Dr05 активних сполук свідчать про деякі особливості

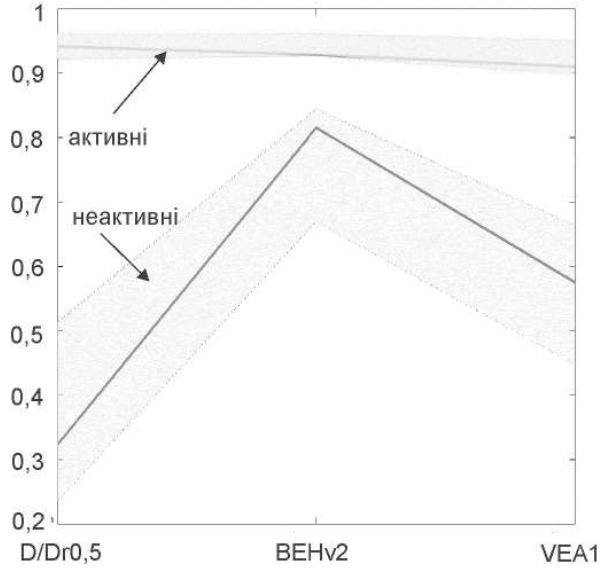
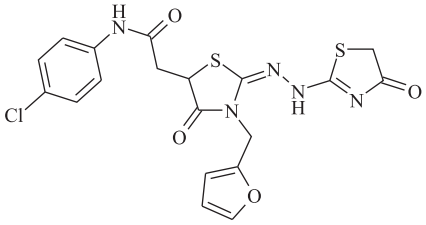
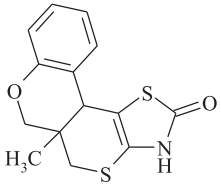


Рис. 6. Прототипи класів активних та неактивних сполук. Інтервали між відповідними квартилями позначено сірим кольором.

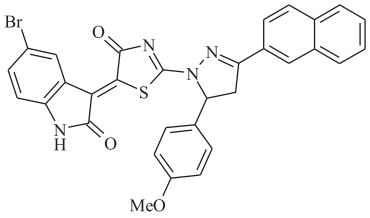


Les-2627, D/Dr05 = 289.8, GP = 14,4% – три п'ятичленні цикли, між всіма циклами є містки різного розміру

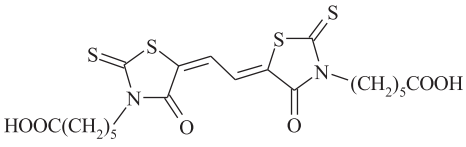


Les-651, D/Dr05 = 22.9, GP = 96,0% – найменше значення дескриптора – п'ятичленний цикл входить до поліциклічної конденсованої системи

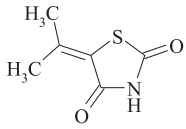
Рис. 7. Приклади сполук з високим та низьким індексами відстані/обходу циклів 5-го порядку.



Les-3833, BEHv2 = 3.93, GP = -60% – наявний галоген, значна кількість подвійних та ароматичних зв'язків



Les-2573, BEHv2 = 3.58, GP = 105,8% – молекулярна маса близька до Les-3833, але відсутній галоген та ароматичні зв'язки, значна кількість одинарних зв'язків

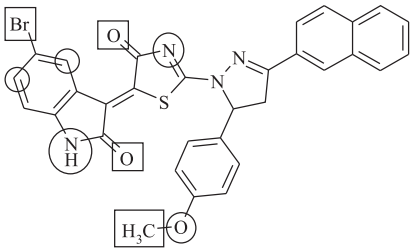


Les-2609, BEHv2 = 3,07, GP = 99,27% – найменше значення дескриптора – відсутній галоген та ароматичні зв'язки, низька молекулярна маса

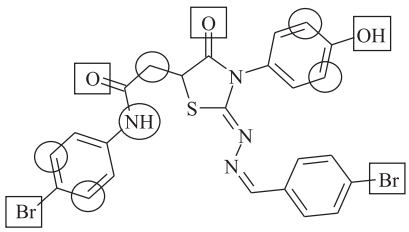
Рис. 8. Приклади сполук з високим та низьким другим найвищим власним значенням матриці Бурдена, зваженої атомними масами.

геометрії сполук, що мають позитивний вплив інгібування росту ракових клітин. Зокрема, наявність декількох п'ятичленних циклів та збільшення відстані між циклами у молекулі підвищують імовірність онкоцитотоксичної дії, і навпаки – наявність одного п'ятичленного циклу, розташованого в безпосередній близькості від інших циклів, знижує шанси щодо активності. На рис. 7 отриманий висновок проілюстровано на представниках різних класів.

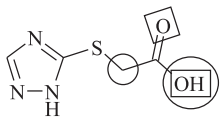
Вищі значення BEHv2 для активних сполук свідчать, що збільшення кількості гетероатомів, подвійних та ароматичних зв'язків підвищують імовірність прояву протипухлинної активності. У той же час факт, що для класифікації обрано саме **друге** найвище власне значення матриці Бурдена, зваженої атомними масами, свідчить про бажану присутність «важкого» гетероатома (найчастіше галогену). Необхідно відзначити, що природа «важкого» гетероатома (атомна



Les-3833, VEA1 = 5,230, GP = -60% – 4 термінальні групи (обведені квадратом), поряд 5 атомів (обведені колом), що зв'язані з двома сусідами



Les-1003, VEA1 = 4.117, GP= 94,3% – молекулярна маса близька до Les-3833, 5 термінальних груп (обведені квадратом), поряд 8 атомів (обведені колом), що зв'язані з двома сусідами



Les-2787, VEA1 = 2.898, GP = 97,2% – найменше значення дескриптора – 2 термінальні групи (обведені квадратом), розташовані поруч біля краю молекули, неподалік від яких 1 атом з двома сусідами та 1 атом, зв'язаний всього з одним сусідом (обведені колом)

Рис. 9. Приклади сполук з високою та низькою сумою коефіцієнтів власного вектора матриці суміжності, асоційованого з найменшим власним значенням.

маса) є неважливою, оскільки роль дескриптора BEHv1 (перше найвище власне значення), на який цей гетероатом чинить вплив, при класифікації незначна. Отриманий висновок проілюстровано прикладами на рис. 8.

Дескриптор VEA1 певним чином залежить від наявності та геометричного оточення у молекулі термінальних груп. Так, збільшення кількості атомних груп, що зв'язані лише з одним атомом-сусідом, збільшення відстані цих груп до центру молекули та до вершин високого ступеня (зв'язок з трьома-чотирма сусідами) призводять до зменшення цього дескриптора, що, в свою чергу, негативно впливає на шанс прояву активності (рис. 9).

Висновки

1. Проведено багатофакторний лінійний регресійний аналіз, який показав неможливість точного опису протипухлинної активності лінійною функцією від обчислених дескрипторів, однак передбачив ефективність класифікаційного моделювання.

2. За допомогою алгоритму *Random Forest* побудовано класифікаційну модель, здатну передбачати наявність чи відсутність протипухлинної активності для нових сполук з метою наступного спрямованого синтезу речовин, що володіють виразною протираковою дією. Прогностична здатність моделі підтверджена статистичними критеріями, отриманими в процесі її створення та при прогнозі тестової вибірки.

3. Здійснено аналіз вкладів дескрипторів у прийняття рішення, визначено найбільш важливі з них для здійснення класифікації, розкрито їх фізико-хімічний зміст та проілюстровано зв'язок вибраних дескрипторів зі здатністю до інгібування росту онкоклітин.

Література

1. *Зіменковський Б.С., Лесик Р.Б. 4-Тіазолідони. Хімія, фізіологічна дія, перспективи. – Вінниця: Нова книга, 2004. – 106 с.*
2. *Lesyk R.B., Zimenkovsky B.S., Kaminsky D.V. et al. // Biopolymers and Cell. – 2011. – Vol. 27, №2. – P. 107-117.*
3. *Гаврилюк Д.Я., Лесик Р.Б., Матійчук В.С., Обушак М.Д. // ЖОФХ. – 2006. – Т. 4, вип. 1 (13). – С. 42-47.*
4. *Havrylyuk D., Mosula L., Zimenkovsky B. et al. // Eur. J. Med. Chem. – 2010. – Vol. 45. – P. 5012-21.*
5. *Havrylyuk D., Zimenkovsky B., Vasylenko O. et al. // Eur. J. Med. Chem. – 2009. – Vol. 44, №4. – P. 1396-1404.*
6. *Kaminsky D., Zimenkovsky B., Lesyk R. // Eur. J. Med. Chem. – 2009. – Vol. 44. – P. 3627-3636.*
7. *Камінський Д.В., Роман О.М., Атаманюк Д.В., Лесик Р.Б. // ЖОФХ. – 2006. – Т. 4, вип. 1 (13). – С. 41-48.*
8. *Гаврилюк Д.Я., Зіменковський Б.С., Драпак І.В. та ін. // Фармац. журн. – 2009. – №6. – С. 69-75.*
9. *Мосула Л.М., Зіменковський Б.С., Огурцов В.В. та ін. // Фармац. журн. – 2010. – №2. – С. 77-83.*
10. *Зіменковський Б.С., Девіняк О.Т., Гаврилюк Д.Я., Лесик Р.Б. // ЖОФХ. – 2011. – Т. 9, вип. 3 (35). – С. 64-71.*
11. *Draper N.R., Smith H. Applied Regression Analysis. – Hoboken, NJ: Wiley-Intersci., 1998. – P. 307-312.*
12. *Tetko I.V., Gasteiger J., Todeschini R. et al. // J. Comput. Aid. Mol. Des. – 2005. – Vol. 19. – P. 453-463.*
13. *Halgren T.A. // J. Comp. Chem. – 1996. – Vol. 17 (5-6). – P. 490-519.*
14. *Breiman L. // Machine Learning. – 2001. – Vol. 45 (1). – P. 5-32.*
15. *Офіційна сторінка автора алгоритму «Random Forests». http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm*
16. *Todeschini R., Consonni V. Molecular Descriptors for Chemoinformatics. Second, Revised and Enlarged Ed. Vol. 1. – Weinheim (Germany): Wiley-VCH, 2009. – 1265 p.*

Надійшла до редакції 11.11.2011 р.