УДК 615.214

Прогнозування біодоступності лікарських засобів методом класифікаційних моделей

М. Я. Головенко, В. Є. Кузьмін, А. Г. Артеменко, М. А. Кулінський
П. Г. Поліщук, І. Ю. Борисюк
Фізико-хімічний інститут ім. О. В. Богатського НАН України, Одеса

Резюме

Використовувались методи класифікаційних дерев і «випадкового лісу» (Random Forest). Обсяг вибірки — 628 сполуки. Були побудовані класифікаційні моделі, що прогнозували біодоступность за двома (низька та допустима) або за трьома класами (низька, середня і висока). Класифікаційні QSPR моделі будувались методом симплексного представлення молекулярної структури. По кожній виборці окремо були побудовані QSPR моделі методом «випадкового лісу». Міжкласифікаційна помилка для моделей з трьома класами висока, звідси прогнозуюча спроможність даної моделі є низькою. Регресійна модель також має низьку прогнозуючу здатність (R^2_{oob} = 0,294). Спрощення класифікації за двома класами має кращу прогнозуючу здатність. Враховуючи перетинання моделей та варіювання біодоступності в деякому діапазоні значень в отримані моделі введено довірчий інтервал у межах 10% від кордону. Були отримані класифікаційні моделі з врахуванням довірчого інтервалу, які значно збільшують прогнозуючу здатність. Метод класифікаційних дерев є достатньо перспективним інструментом для попереднього аналізу біодоступності потенційних ЛЗ. Має місце значний вплив фізіологічних факторів, що зменшують біодоступність ліків до їх попадання в системний кровообіг, які важко встановити моделюванням, це потребує додаткових досліджень. Метод добре прогнозує біодоступність низькомолекулярних сполук, всмоктування яких відбувається шляхом простої дифузії.

Ключові слова: біодоступність, метод класифікаційних дерев, метод «випадкового лісу».

Клин. информат. и Телемед. 2011. Т.7. Вып.8. с.88–92

Вступ

Біодоступність – універсальна властивість препаратів, що охоплює швидкість та ступінь надходження активного фармацевтичного інгредієнта (АФІ) з відповідної лікарської форми і місця введення в системний кровообіг, внаслідок чого вони стають доступними в біофазі дії. Цей показник є дуже важливим при створенні та впровадженні в медичну практику інноваційних препаратів [1], а також у випадку генеричних препаратів, які є взаємозамінними інноваційному. Останнє досягається шляхом вивчення біоеквівалентності або порівняльними дослідженнями *in vitro* [2]. Отже, дослідження біодоступності є відповідним «фільтром» надходження неякісних лікарських засобів (ЛЗ) в медичну практику.

Визначення відповідних показників біодоступності (баланс мас, $AUC_{тест/референт}$, тощо) та розчинення ЛЗ все ще залишається трудомістким процесом, тому створення спрощених методів, що дають можливість їх прогнозувати є актуальним і необхідним у біофармацевтичних дослідженнях. Варте уваги те, що моделювання завжди надає системі, що описується структурний відтінок. Тому структурне подання отримує широке розповсюдження в біофармації. Ці методи засновані на описі структури хімічної речовини за допомогою числових характеристик-дескрипторів і побудові статистичних залежностей між значеннями дескрипторів і проникненням через біомембрани, швидкістю всмоктування або біодоступністю. Вирішальне значення при цьому має обраний набір дескрипторів, які відображають особливості молекулярної структури, від яких може залежати біодоступність [3].

Умовно, моделі, що описують таку залежність, можуть бути якісними (SPR) або кількісними (Quantitative structure-activity relationship, QSAR / Quantitative structure-property relationship, QSPR). Перевагою перших служило те, що для їх реалізації не було необхідним отримувати дані біодоступності експериментальним шляхом. Автори розробок використовували відповідні показники з бази даних WDI і ACD. У 1997 р. Ліпінські і співавтори [4] провели аналіз біодоступності 2245 відомих комерційних препаратів і показали, що тільки 10% з них мають logP > 5, а 8% мають > 5 донорів водневого зв'язку (ДВЗ) OH і NH, 10% > 10 акцепторів водневого зв'язку (АВЗ) і 11% мали молекулярну масу (ММ) більше 500. Отримані дані дозволили авторам встановити «Правило 5», згідно з яким низька проникність мембран або біодоступність має місце в тому випадку, коли ЛЗ відповідають наступним показникам: ММ > 500, logP > 5, ДВЗ > 5 та АВЗ > 10.

Пізніше, Вебер і Джонсон [5] з метою прогнозу біодоступності використовували тільки два дескриптора: число зв'язків, що обертаються, повинно складати < 12 і площа молекулярної поверхні 140 A^2. При цьому автори не заперечували значення такого показника, як кількість можливих водневих зв'язків та їх стійкість.

Кількісна модель (QSAR/QSPR) це математичні співвідношення, за допомогою яких можна описати властивість (біодоступність). Для її побудови використовують різні методи, як наприклад, множинна лінійна регресії (MLR) [3], або її модифікація – регресія латентних змінних (метод часткових найменших квадратів – PLS) [6, 7], а також нелінійн-

ний статистичний метод – штучні нейронні мережі (ANN) [8].

Оскільки різноманітні біологічні механізми вносять відповідний внесок у процес біодоступності (наприклад, ефект первинного проходження або вплив P-gp) може виникнути дуже складний нелінійний зв'язок між розглянутими параметрами. Звідси складність її кількісного опису і прогнозування. У цих випадках використовують методи класифікації, в яких визначають, до якого з двох або більше класів належить молекула. Для досягнення такої можливості використовують кілька методів:

1) лінійний дискримінантний аналіз (LDA), де для QSAR/QSPR аналізу (класифікації ЛЗ) використовуються не регресійні рівняння, а дискримінантні функції [9];

2) метод класифікаційних дерев рішень представляє собою гіллясту (ієрархічну) структуру, в якій кожен вузол відповідає конкретній умові для певного дескриптора [10, 11].

Класифікаційні дерева – це непараметричний статистичний метод аналізу, що дозволяє аналізувати вибірки будь-якого розміру незалежно від кількості об'єктів (сполук) і характеризують їх атрибути (молекулярниі дескриптори). Метод класифікаційних дерев має ряд переваг перед широко поширеними регресійнии методами QSAR/QSPR аналізу, таких як: швидкий процес навчання, інтуїтивно зрозуміла класифікаційна модель на природній мові, нелінійність одержуваних моделей, можливість побудови моделей для випадків, коли використовується не числова шкала, а класифікаційна (що дозволяє аналізувати вибірки з різнорідними значеннями активності).

Тому як альтернативу широко поширеним регресійним методам для побудови QSAR моделей (прогнозування класу біодоступності для сполук) ми пропонуємо використовувати метод класифікаційних дерев, що і стало метою нашої роботи.

Матеріали та методи

В роботі використані методи класифікаційних дерев і «випадкового лісу» (Random Forest). Побудова класифікаційного дерева відбувається наступним чином. На вхід (в корінь) подається деяка навчальна множина (вибірка), яка містить об'єкти (сполуки), що характеризуються атрибутами (дескрипторами), один з яких визначає приналежність кожного об'єкта до певного класу. Далі алгоритм виробляє загальні критерії для об'єктів одного класу. Для цього в кожній вершині, починаючи з кореня, визначається правило, на підставі якого і відбувається подальший розподіл навчальної вибірки. Під правилом розуміється логічна конструкція, представлена у вигляді «якщо ... – то ... ». Наприклад, «якщо ліпофільність > 3, то сполука активна». Таке зростання дерева відбувається до того часу, доки в вершині не залишаться об'єкти, що належать до одного класу, тобто в результаті ми отримуємо термінальну вершину. Інакше кажучи, класифікаційні дерева – це спосіб подання правил в ієрархічній, послідовній структурі, де кожному об'єкту відповідає єдина вершина, що дає рішення.

Для оцінки надійності одержаних моделей класифікаційних дерев традиційно використовується стандартна процедура ковзаючого (перехресного) контролю. Одним з важливих етапів визначення прогнозуючої здатності QSAR моделей є наявність зовнішньої тестової вибірки. Таким чином, на попередньому етапі необхідно виключити з вихідної вибірки частину сполук, які після побудови моделі будуть використані для оцінки її надійності. Слід зазначити, що для більш достовірного результату тестова вибірка повинна включати різнорідні сполуки, які охоплюють всі класи активності.

Ми пропонуємо використовувати для цього такі дії. Будується допоміжна модель класифікаційних дерев на основі всіх наявних у вибірці сполук. Потім з кожного термінального вузла отриманої моделі, що містить більше 5% сполук (від їх загальної кількості), 15–20% сполук, які належать одному класу активності, поміщаються в тестову вибірку.

Важливо зазначити, що таким чином фактично оцінюють подібність сполук на основі структурних параметрів найбільш важливих для активності, що досліджується. Така процедура дозволяє сформувати представницьку тестову вибірку, яка включає сполуки всіх класів. При оцінці структурної подібності не слід керуватися випадковим набором структурних параметрів. Коректніше оцінювати подібність можна на основі тих параметрів, які дійсно впливають на властивість. Вибираючи сполуки тестової вибірки на основі оцінки їх подібності із сполуками, які залишаються в навчальній вибірці, вдається уникнути втрати унікальної структурної інформації, що може статися в разі формування випадкової тестової вибірки.

Вибір тільки термінальних вузлів, що містять більше 5% сполук, обумовлений тим, що в методі класифікаційних дерев немає процедури оцінки «відкидів» і подібні сполуки часто потрапляють в окремий вузол дерева [12]. Відповідно ймовірність того, що у вузлі з малим числом сполук присутня велика частка «відкидів» дуже висока.

Підхід, що застосовуємо тільки до моделей, в яких активність виражена класифікаційною шкалою є певним обмеженням методу. Однак для багатьох завдань, пов'язаних з дослідженням біологічної активності, така оцінка рівня активності сполук цілком прийнятна [13].

На основі методу класифікаційних дерев будується метод «випадкового лісу». «Випадковий ліс» – порівняно новий статистичний метод аналізу, який набуває все більшої популярності для побудови QSAR моделей, завдяки таким своїм перевагам як: відсутність проблеми перенавчання моделі; відсутність необхідності відбору (преселекції) змінних; наявність адекватної внутрішньої процедури оцінки якості та прогнозуючої здатності моделей; стійкість моделей до наявності шуму у вихідній вибірці; ефективна робота з великими базами даних; інтерпретованість одержуваних моделей; можливість коректно аналізувати вибірки, що включають сполуки з різним механізмом дії.

Модель «випадкового лісу» представляє собою ансамбль окремих дерев рішень. Таким чином, за допомогою цього методу можливо рішення як класифікаційних, так і регресійних завдань. Кожне з дерев в моделі «випадкового лісу» будується відповідно до таких правил:

1) з усього набору сполук навчальної вибірки з використанням процедури бутстреп формується нова вибірка, яка і є навчальною для даного конкретного дерева. Сполуки, що не ввійшли в навчальну вибірку поміщають в так звану out-of-bag вибірку, яка використовується при оцінці якості та прогнозуючої здатності моделі лісу;

2) при поділі даних у кожному вузлі дерева використовується алгоритм CART, однак розглядаються не всі змінні, а тільки невелика їх частина, яка в кожному вузлі вибирається випадково. Число цих змінних фіксоване і залишається постійним на всьому протязі побудови моделі. Це єдиний параметр, до якого моделі більш-менш чутливі;

3) кожне дерево будується до максимально можливої глибини, процедура відсікання гілок відсутня.

Для регресійних завдань прогноз здійснюється усередненням всіх прогнозів окремих дерев у лісі. Для класи-

фікаційних завдань прогноз проводять за найбільшою кількістю голосів поданих за будь-який клас. Аналогічним чином здійснюється прогноз і для out-of-bag вибірки, тільки кожне дерево видає прогноз тільки для тих сполук, які не увійшли в навчальну вибірку цього конкретного дерева. Величина помилки класифікації out-of-bag вибірки є визначальним параметром при виборі кінцевої моделі.

Результати та їх обговорення

Згідно з основними вимогами із 1000 ЛЗ, зареєстрованих на Україні, була створена первинна біофармацевтична база даних [14], в якій вказані як фізико-хімічні властивості сполук, так і біодоступність. Потім з урахуванням всіх обмежень («Правило 5») і наших вимог (встановлення пероральної біодоступності для таблетованої лікарської форми, тобто відкидали всі інші лікарські форми) залишилося 628 сполук, що і стало обсягом нашої вибірки. Окрім того, були відкинуті контрацептиви та ЛЗ, що впливають на моторику ШКТ.

Для оцінки біодоступності нами були побудовані класифікаційні моделі. Класифікаційні моделі прогнозували біодоступність за двома класами (низька та допустима) або за трьома (низька, середня і висока).

Для побудови класифікаційних QSPR моделей було використано метод симплексного представлення молекулярної структури [15], який дозволяє прогнозувати абсолютну біодоступність ЛЗ. З цією метою всі сполуки були розділені на три (висока, середня та низька біодоступність, табл. 1) або два (висока та допустима, табл. 2) класи. По кожній виборці окремо були побудовані QSPR моделі методом «випадкового лісу».

Для визначення чи впливає зсунення кордонів класів на якість моделей прогнозу і який цей вплив вибиралися різні границі класів, що показано в табл. 1.

Як видно з представлених результатів міжкласифікаційна помилка висока, а звідси і прогнозуюча спроможність моделей достатньо низька. Найкращий проноз у II моделі особливо, що стосується 1 и 2 класу. На основі данної виборки (628 сполуки) також була отримана регресійна модель, яка мала низьку прогнозуючу здатність (R^2_{oob} = 0,294).

Спрощення класифікації за двома класами (табл. 2) має кращу прогнозуючу здатність.

Табл. 1. Класифікаційні моделі по трьом класам біодоступності.

№ п/п моделі	Межа класів	Кількість молекул в класі	Кількість помилок по класам	Загальна помилка, %
I	1) 0–20 2) 20–80 3) 80–100	114 304 210	71 69 85	36
II	1) 0–10 2) 10–90 3) 90–100	68 432 128	45 28 98	27
III	1) 0–20 2) 20–70 3) 70–100	114 250 264	68 86 71	36

Табл. 2. Класифікаційні моделі по двум класам біодоступності.

№ п/п моделі	Межа класів	Кількість молекул в класі	Помилка класифікації, %
IV	1) 90–100 2) 0–90	156 472	20
V	1) 80–100 2) 0–80	222 406	24
VI	1) 70–100 2) 0–70	280 348	27

В табл. 2 показано, що зсув меж моделей знижують точність їх прогнозу. Це перш за все обумовлено збільшенням кількісті молекул із високою біодоступністю. Отже з метою виявлення причин даного феномену нами було проведено ретельний аналіз ЛЗ по всім відповідним моделям (IV–VI).

Результати аналізу показали перетин між всіма трьома моделями. Доведено, що практично у всіх моделях IV–VI для 1 класу (біодоступність 90–100, або 80–100, або 70–100) помилка в прогнозі пов'язана з активною участю специфічних транспортерів при всмоктуванні. Для 2 класу – відмічена незначна участь транспортерів в процесі всмоктування ЛЗ, але має місце інтенсивний метаболізм сполук та зв'язування з мішенню (причому більшість ЛЗ мають декілька мішеней, з якими відбувається міцний та тривалий зв'язок).

Враховуючи перетинання моделей та варіювання біодоступності в деякому діапазоні значень в отримані моделі ми ввели довірчий інтервал у межах 10 відсотків від кордону (рис. 1).

Таким чином, межа між класами має певну товщину. Похибки класифікації точно визначаються в тому випадку, коли молекули виходять за кордони межи в протилежний клас.

В результаті були отримані класифікаційні моделі з врахуванням довірчого інтервалу (табл. 3).

Представлені моделі (IV–1 – VI–1) значно збільшують прогнозуючу здатність. Для моделі IV–1 з найменшою помилкою класифікації приведені ЛЗ, для яких було проведено детальний аналіз, з врахуванням різних фізіологічних факторів, що зменшують біодоступність ліків до їх попадання в системний кровообіг. До числа таких факторів належать: фізичні властивості ЛЗ, зокрема, гідрофобність, ступінь дисоціації, розчинність; лікарські форми препарату (негайне, уповільнене, подовжене або тривале вивільнення, застосування допоміжних речовин, методи виробництва); введено ЛЗ натщесерце або після прийому їжі; розбіжності протягом доби; швидкість спорожнення шлунку; індукування / інгібування іншими ЛЗ або їжею: взаємодія з іншими ліками (антацидами, алкоголем, нікотином), взаємодія з окремими продуктами харчування (грейпфрутовий сік, помело, журавлинний сік); білки-переносники, субстрати для білка-переносника

(напр., Р-глікопротеїн); стан шлунково-кишкового тракту, його функція і морфологія.

Так, ЛЗ, що дають помилку в прогнозуванні, можно розподілити на слідуючі групи:

а) при їх всмоктуванні активну роль приймають транспортери та вони мають певні мішені (1 клас – буметанид, цефрадин, дифлунизал, фолієва кислота, леветирацетам, локарбеф, праміпексол);

б) інтенсивно піддаються метаболізму та мають відповідні мішені (1 клас – амобарбітал, анастрозол, клофибрат, клоназепам, ципротерон, дапсон, дофетилид, доласетрон, донепезил, етосуксимид, фелбамат, галантамин, глімепирид, гексобарбітал, лінезолід, метимазол, мідодрін, невірапін, оксапрозин, пентобарбітал, фенпрокумон, фенілпропаноламін, прімаквін, псевдоефедрин, тамсулозин, тіагабін гідрохлорид, токаінід, зонізамід; 2 клас – бусульфан, етофілін);

в) мають мішені (1 клас – диазоксид, індапамід, міноксіділ, пенбутолол, фенірамін, практолол, римантадин, роксатідин, соталол, тригексифенидил; 2 клас – бутабарбітал);

г) всі чинники мають місце (1 клас – цетазоламід, безафібрат, клонидин, кортикостерон, циклопентиазид, доксициклин, етодолак, гемфиброзил, гидрокортизон, іматініб, ламотригін, ліотиронін, фенілбутазон, пробенецид, ребокседин, розіглітазон, сертралін);

д) мають інші чинники, що впливають на біодоступність (1 клас – аципімокс, амосулалол, антипірин, бетаксолол, цикапрост, циклопролол, хлороквин, фенспирид, флупиртин, гестринон, гитоксин, ізосорбід-2-мононітрат, летрозол, лоразепам, нікоранділ, пірпрофен, рілменідін, сульфадімезин, тіанептин, трапіділ, треосульфан; 2 клас – реніцин, різатріптан, сульфадимідин).

Таким чином, метод класифікаційних дерев є достатньо перспективним інструментом для попереднього аналізу активності (біодоступності) потенційних ЛЗ. Однак, має місце значний вплив різних фізіологічних факторів, що зменшують біодоступність ліків до їх попадання в системний кровообіг, які важко встановити моделюванням, так як це потребує додаткових експериментальних досліджень. На нашу думку, цей метод добре прогнозує біодоступність низькомолекулярних сполук, всмоктування яких відбувається шляхом простої дифузії.

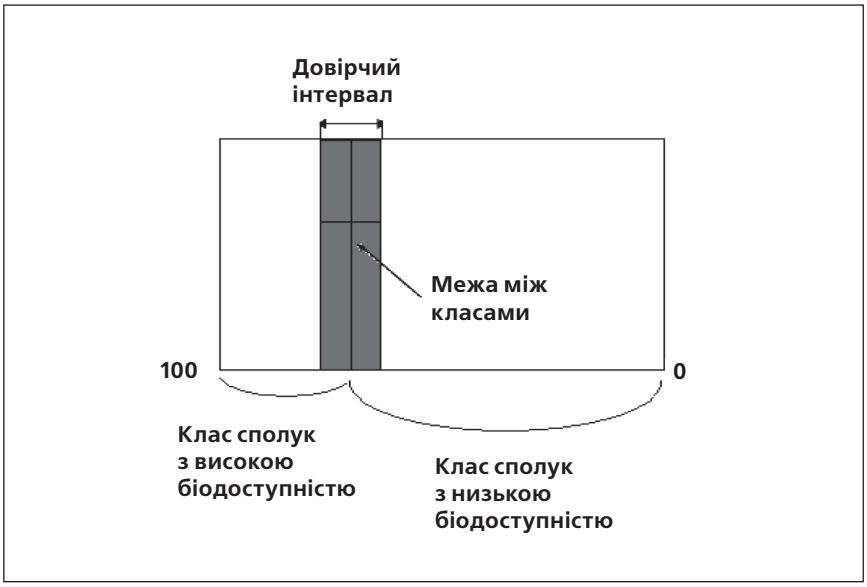

Рис. 1. Довірчий інтервал класифікації.

Табл. 3. Класифікаційні моделі по двум класам з довірчим інтервалом.

№ п/п моделі	Межа класів	Кількість молекул в класі	Помилка по класам	Помилка класифікації, %
IV–1	1) 90–100 2) 0–90	156 222	83 6	14
V–1	1) 80–100 2) 0–80	222 406	81 20	16
VI–1	1) 70–100 2) 0–70	280 348	74 36	18

Література

1. Головенко Н. Я. Физико-химическая фармакология. – Одесса, Астропринт. – 2004. – 720 с.
2. Головенко М. Я., Баула О. П., Борисюк І. Ю. Біофармацевтична класифікаційна система. – Київ. – 2010. – 299 с.
3. Головенко Н. Я., Борисюк И. Ю. Теоретические основы биофармацевтической классификационной системы // Фармаком. – 2007. – №3. – С. 27–37.
4. Lipinski C. A., Lombardo E., Doming B., Feeney P. J. Experimental and computation approaches to estimate solubility and permeability in drug discovery and development settings // Adv. Drug Deliv. Rev. – 1997. – Vol. 23. – P. 3–25.
5. Veber D., Johnson S. Molecular properties that influence the drug bioavailability of drug candidates // J. Med. Chem. – 2002. – Vol. 45, №12. – P. 2615–2623.
6. Wold S., Johansson E., Cocchi M. PLS — Partial least-squares projections to latent structures, in: 3D QSAR in Drug Design // Kubinyi. H. – ESCOM, Leiden. – 1993. – P. 523–550.
7. Cramer R. D., Patterson D. E., Bunce J. D. Comparative Molecular Field Analysis (CoMFA) / 1. Effect of shape on binding of steroids to carrier proteins // J. Am. Chem. Soc. – 1988. – Vol. 110. – P. 5959–5967.
8. Баскин И., Палсолин В., Зефиров Н. Многослойные персептроны в исследовании зависимости «структура–свойство» для органических соединений // Рос. хим. ж. – 2006. – Т. 50. – Вып. 2. – С. 86–96.
9. Anderson T. An introduction to multivariate statistical and analysis (2nd). – New York, Wiley. – 1984. – 353 p.
10. Артеменко А. Г., Кузьмин В. Е., Муратов Е. Н., Полищук П. Г., Головенко Н. Я., Борисюк И. Ю. Анализ влияния структуры замещенных бензодиазепинов на их фармакокинетические свойства // Химико-фармацевтический журнал. – 2009. – т. 43, №8. – С. 27–35.
11. Артеменко А. Г., Поліщук П. Г., Муратов Є. М., Кузьмін В. Є., Головенко М. Я., Борисюк І. Ю. Прогнозування періо-

ду напіввиведення препаратів похідних 1,4-бенздіазепіну на основі комбінації симплексів // Медична хімія. – 2007. – т. 9, №3. – С. 10–17.
12. Fielding A. H. Cluster and classification techniques for the biosciences / Fielding AH – Cambridge University Press, 2007 – 240 p.
13. Kuz'min V. E., Artemenko A. G., Muratov E. N., Polischuk P. G., Ognichenko L. N., Liahovsky A. V., Hkomov A. I., Varlamova E. V. Virtual screening and molecular design based on hierarchical QSAR technology // Recent advances in QSAR studies methods and application / Puzyn T., Leszczynski J., Cronin M.T.D. – Springer. – 2010. – Vol. 8. – P. 127–176.
14. Головенко М. Я., Кузьмін В. Є., Ларіонов В. Б., Муратов Є. В., Борисюк І. Ю. Біофармацевтична інформатика: генерація нових знань і розроблення лікарських засобів // Вісник Національної Академії Наук України. – 2009. – №8. – С. 3– 10.
15. Kovdienko N. A., Polischuk P. G., Muratov E. N., Artemenko A. G., Kuz'min V. E., Gorb L., Hill F., Leszczynski J. Application of Random Forest and multiple linear regression technologues to QSPR prediction of an aqueous solubility for military compaunds // Molecular informatics. – 2010. – vol. 29. – P. 394–406.

Prediction of bioavailability of drugs by the method of classification models

N. Ya. Golovenko, V. E. Kuz'min
A. G. Artemenko, M. A. Kulinsky
P. G. Polishchuk, I. Yu. Borisyuk
A.V. Bogatsky Physico-Chemical Institute NAS Ukraine, Odessa

Abstract

The study was conducted using the methods of classification trees and «random forest» (Random Forest). Sample size was 628 compounds. To assess the bioavailability of classification models were constructed, predicted this option for two (low and acceptable), or three classes (low, medium, high). To construct the QSPR classification models used the method of simplex representation of molecular structure. For each sample separately QSPR models were constructed using «random forest».

Interclassification mistakes for models with three classes of high, hence the predictive ability of this model is low. Regression model also has a low predictive ability (R^2_{oob} = 0.294). Simplifying the classification of two classes has a better predictive ability.

Given the intersection of models and varying bioavailability in a certain range of values obtained in the model we have introduced a confidence interval within 10% percent of the border. Thus were obtained classification models taking into account the confidence interval, which significantly increase the predictive ability.

Thus, the method of classification trees is a promising tool for preliminary analysis of the bioavailability of potential drugs. However, there is significant influence of various physiological factors that reduce the bioavailability of drugs before they enter the systemic circulation, which are difficult to establish the simulation, since it requires additional experimental research. In our opinion, this method well predicts the bioavailability of low molecular weight compounds, absorption of which occurs by simple diffusion.

Key words: bioavailability, method of classification trees, method of «random forest».

Прогноз биодоступности лекарственных средств методом классификационных моделей

Н. Я. Головенко, В. Е. Кузьмин
А. Г. Артеменко, М. А. Кулинский
П. Г. Полищук, И. Ю. Борисюк
Физико-химический институт им. А. В. Богатского НАН Украины Одесса

Резюме

Исследование проведено с использованием методов классификационных деревьев и «случайного леса» (Random Forest). Объем выборки составил 628 соединений. Для оценки биодоступности были построены классификационные модели, прогнозировали этот параметр по двум (низкая и допустимая) или по трем классам (низкая, средняя и высокая). Для построения классификационных QSPR моделей был использован метод симплексного представления молекулярной структуры. По каждой выборке отдельно были построены QSPR модели методом «случайного леса».

Межклассификационная ошибка для моделей с тремя классами высокая, отсюда прогнозирующая способность данной модели является низкой. Регрессионная модель также имеет низкую прогнозирующая способность (R^2_{oob} = 0,294). Упрощение классификации по двум классам имеет лучшую прогнозирующая способность.

Учитывая пересечения моделей и варьирования биодоступности в некотором диапазоне значений в полученные модели мы ввели доверительный интервал в пределах 10% процентов от границы. Так были получены классификационные модели с учетом доверительного интервала, которые значительно увеличивают прогнозирующая способность.

Таким образом, метод классификационных деревьев является перспективным инструментом для предварительного анализа биодоступности потенциальных ЛС. Однако, имеет место значительное влияние различных физиологических факторов, уменьшающих биодоступность лекарств до их попадания в системный кровоток, которые трудно установить моделированием, так как это требует дополнительных экспериментальных исследований. По нашему мнению, этот метод хорошо прогнозирует биодоступность низкомолекулярных соединений, всасывание которых происходит путем простой диффузии.

Ключевые слова: биодоступность, метод классификационных деревьев, метод «случайного леса».

Листування

академік НАМН України
д.біол.н. **М. Я. Головенко**
фізико-хімічний інститут
ім. О.В. Богатського НАН України
тел. +380 (48) 766 23 93
ел. пошта: n.golovenko@gmail.com

к.біол.н. **І. Ю. Борисюк**
фізико-хімічний інститут
ім. О.В. Богатського НАН України
Люстдорфська дорога 86
Одеса, 65080, Україна
тел. +380 (48) 765 94 02
ел. пошта:borisyuk_kaynova@mail.ru