мах с участием супероксид–аниона, и, как следствие, — разработке антиоксидантов с необходимым уровнем антиоксидантной активности.

Список использованной литературы

1. Афанасьев И.Б. Анион-радикал кислорода O_2^- в химических и биохимических процессах / И.Б. Афанасьев // Успехи химии. — 1986. — Т. 48. — № 6. — С. 977.

2. Помещенко А.И. Особенности влияния некоторых фенольных соединений на процесс окисления кумола в присутствии O_2^- / А.И. Помещенко, И.В. Ефимова, И.А. Опейда // Наукові праці ДонНТУ. Серія: Хімія і хімічна технологія. — 2011. — Вип. 16(184). — С. 101–105.

3. Frimer A.A. The Superoxide Mediated Base Catalyzed Autoxidation of Tetraphenylcyclopentadiene / A.A. Frimer, G. Strul and H. E. Gottlieb // J. Org. Chem. — 1995. — № 60. — P. 4521–4524.

4. Bradshaw M.P. Defining the Ascorbic Acid Crossover from Anti-Oxidant to Pro-Oxidant in A Model Wine Matrix Containing (+)-Catechin / M.P. Bradshaw, V. Cheyner, G.R. Scollary, P.D. Prenzler // J. Agric. Food Chem. — 2003. — Vol. 51. — P. 4126–4132.

5. Peyrat-Maillard M.N. Antioxidant activity of phenolic compounds in 2,2′-azobis (2-amidinopropane) dihydrochloride (AAPH)-induced oxidation: Synergistic and antagonistic effects / M.N. Peyrat-Maillard, M.E. Cuvelier, C. Berset // J. Am. Oil Chem. Soc. — 2003.— Vol. 80, N. 10. — P. 1007–1012.

6. Эмануэль Н.М. Цепные реакции окисления в жидкой фазе / Н.М. Эмануэль, Е.Т. Денисов, З.К. Майзус. — М.: Наука, 1965. — 374 с.

7. Ефимова И.В. Реакции O_2^--содержащих супрамолекул с алкилгалогенидами / И.В. Ефимова, А.И. Помещенко, И.А. Опейда // Журнал общ. Химии. — 2004.— Т. 74.— Вып. 7.— С. 1100–1104.

8. Опейда И.А. Радикально-цепное окисление в присутствии супрамолекулярных систем, содержащих супероксид-анион / И.А. Опейда , А.И. Помещенко, И.В. Ефимова // Журнал физ. химии. — 2004.— Т. 78, № 11.— С. 1976–1979.

9. Wilfred L.F. Armarego, Christina L.L. Chai. Purification of Laboratory Chemicals. Elsevier Science, 2003. — 608 p.

10. Ефимова И.В. Кинетика реакции супероксиданиона с пропилбромидом / И.В. Ефимова, А.И. Помещенко, И.О. Опейда // Журнал орг. Химии. — 2002.— Т. 38, Вып. 11. — С. 1746–1747.

11. Эмануэль Н.М. Роль среды в радикально-цепных реакциях окисления органических соединений / Н.М. Эмануэль, Г.Е. Заиков, З.К. Майзус. — М.: Наука, 1973. — 297 с.

12. Ковтун Г.А. Реакционная способность взаимодействия фенольних антиоксидантов с пероксильными радикалами / Г.А. Ковтун // Катализ и нефтехимия. — 2000. — № 4.— С. 1–9.

Надійшла до редколегії 05.03.2012

УДК 544.35.032.72+54.04+547.892

С.Ю. Суйков, А.В. Яковлева, О.І. Луцик (ИнФОУ ім. Л.М. Літвіненка НАН України)

ФОРМУВАННЯ ТЕСТОВОГО НАБОРУ ДАНИХ ТА ОЦІНКИ ЕФЕКТИВНОСТІ МОДЕЛЕЙ РОЗПОДІЛУ Н-ОКТАНОЛ/ВОДА

Проведені дослідження методів аналізу прогнозних даних емпіричних моделей розподілу н-октанол / вода. Запропоновано новий контрольний набір значень Pow та методи аналізу ефективності прогнозу. Виділені найбільш адекватні емпіричні моделі. Показано, що більш достовірним є прогноз зміни величини Pow по відношенню до спорідненої сполуки.

Ключові слова: *коефіцент розподілу, октанол, вода, гетероциклічні сполуки, прогнозні моделі.*

Проведены исследования методов анализа прогнозных данных эмпирических моделей распределения н-октанол/вода. Предложен новый контрольный набор значений Pow и методы анализа эффективности прогноза. Выделены наиболее адекватные модели. Показано,

что более достоверным является прогноз изменения величины Pow по отношению к родственному соединению.

Ключевые слова: *коэффициент распределения, октанол, вода, гетероциклические соединения, прогнозные модели.*

Вступ

Останні десятиліття розробка QSAR моделей значною мірою стимульована потребами фармакології та екології [1–3]. Емпіричні моделі вигідно виділяються відносною простотою і легкістю реалізації. Окремим напрямком є оцінка їх ефективності. Відсутність прийнятих методів та складні види розподілу величин дають широке поле для досліджень — від використання дуже великих контрольних наборів даних [2] до специфічних методів теорії інформації [1, 4]. Одним з важливих фізико-хімічних параметрів, є коефіцієнт розподілу н-октанол / вода (P_{ow}). Ця суто ефективна величина [5] широко використовується у процесі створення ліків, прогнозі біологічної активності, екологічних ефектів тощо [3]. Її експериментальне вимірювання є дуже витратним і проведено для обмеженого кола речовин. Тому значні зусилля зосереджено на створенні прогнозних моделей [1–2], в першу чергу емпіричних, постійно удосконалюються і програмні засоби для прогнозу P_{ow} [1].

Формулювання «ефективності» моделі у відношенні до прогнозу P_{ow} не є неочевидним. Мала кількість доступних експериментальних значень призводить до певної циклічності – введення нового набору «навчає» існуючі моделі, і він втрачає ефективність. Прогнозні моделі, побудовані з перших принципів (наприклад, на основі РСМ), також містять значну кількість підгонних параметрів, і цим не принципово відрізняються від емпіричних [6].

Дослідження статистичних характеристик моделей та вибору відповідних оптимальних функціоналів і методів для представлення розбіжностей в літературі практично відсутні. В роботі [1] для цієї мети використана сума арифметичних середніх нев'язок, в [7] сума середніх квадратичних похибок. Остання міра досить популярна [7] але жодною чином не доведено, що є ефективною. В дослідженні [7] запропоновані рівні модуля різниці експериментального та розрахункового значень $\log P_{ow}$: 0–0,5 «прийнятне», 0,5–1,0 «дискусійне» більше 1,0 – «неприйнятне» – зазвичай вважається, що рівень між лабораторної збіжності значень $\log P_{ow}$ становить до 0,35 одиниці [3, 7]. Обговорювалася ідея використовувати розбіжності в прогнозі різними моделями як міру надійності прогнозу. В [7] для неї навіть запропоновано окремий термін консенсусне (узгоджене) значення $\log P_{ow}$. На жаль, ефективність такого підходу a priori неочевидна і вимагає перевірки.

Питання створення ефективних вибірок для параметризації та оцінки якості прогнозних моделей зрідка розглядається в літературі, проте до теперішнього часу не має достатньо надійного рішення. Існуючі набори даних не є випадковими [4] і закономірності, з яких речовини потрапляють в них здатні забезпечувати локальну ефективність в цілому неприйнятних моделей [5].

Завданням роботи було формування контрольного набору даних помірного розміру та порівняльний аналіз на ньому ефективності ряду відомих прогнозних моделей.

Експериментальна частина

Статистичну обробку результатів проводили за допомогою програми GNU R методами візуального аналізу даних за [9]. Дані для аналізу та дослідження моделей в цій роботі було

отримано фільтруванням відомої відкритої бази даних Санкстера [3] по ключу «azepin». На даний момент в базі міститься більше 28 тисяч сполук. Прогнозні значення отримані за допомогою сервісу vcclab та для ряду програм безпосередньо через їх web-інтерфейси. Далі посилання на моделі дається у формі: ALOGPs — ALOGPs, AC logP — AclogP, AB/logP — ABlogP, а milogP, ALOGP, MLOGP, KOWWIN, XLOGP2, XLOGP3 як є. Всього було виділено 262 сполуки. Обсяг вибірки є порівнянним, наприклад, з використаним в [7] списком «Star» з 223 сполук, який, у свою чергу, є скоригованими набором даних з роботи [8] і є не принципово меншим від тестового набору «Nycomed» з 882 сполук.

На рис. 1 наведено гістограму значень $logP_{ow}$ для всього використаного набору. Легко бачити, що за наявними даними вибірка є досить представницькою.

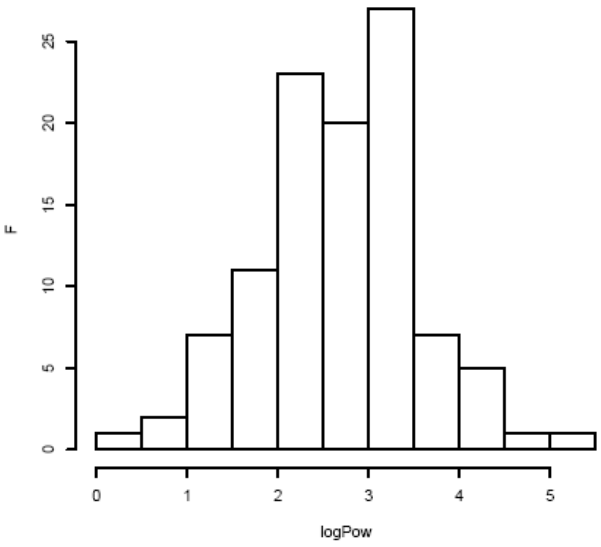

Рис. 1. Гістограма значень $logP_{ow}$ для обраного масиву сполук

Загальне порівняння результатів прогнозу з експериментом на всьому обраному наборі наведено на рис. 2. Для оцінювання ефективності моделі більш важливим є не чисельний збіг з експериментальними даними, а наявність лінійної (у загальному випадку — функціональної) залежності між прогнозом і експериментом. При наявності такого зв'язку додаткова «підгонка» результатів прогнозу не потребує переробки моделі й може бути легко проведена кінцевим користувачем за кількома експериментальним значенням. Зв'язок прогнозу і реального значення $logP_{ow}$ знаходиться в першому рядку матриці (рис 2). В ідеальному випадку високоефективної моделі в ньому повинна бути лінійна залежність.

Видно, що найкращі результати демонструє модель XLOGP3, дещо гірше ALOGPs і miLogP. Видається дуже цікавою мала кореляція між прогнозами за спорідненими XLOGP2 і XLOGP3. XLOGP у теоретичному сенсі – найпростіша за структурою модель атомних вкладів з корегуючими факторами. XLOGP3 найкраще ілюструє розвиток ідеології всього напрямку. Біля витоків її стояло використання розрахункових схем для оцінки відхилення $logP_{ow}$ від відомого значення для структурно близької сполуки. Тривалого часу розвивалися моделі прогнозу не відхилення, а безпосередньо значення коефіцієнту розподілу. Наразі сімейство ALOGP і XLOGP3 найкращі показники мають при розрахунку «споріднених» з відомими структур, тобто для відхилення, а не величини.

Поруч з повним набором даних було проаналізовано рандомізовану вибірку, створену за наявністю експериментального значення у виводі vcclab і меншу за загальну приблизно у 2 рази. Результати прогнозу у цьому наборі (який є підмножиною загального) наведено на рис 3. Порівняння рис. 2 та 3 наочно демонструють низьку стабільність висновків щодо ефективності прогнозу та їх високу залежність від вибірки. І на рис. 3 найкращі результати у XLOGP3 та ALOGP(-s), але значно ліпші і для інших моделей. За значеннями коефіцієнтів кореляції як адекватна виділяється лише XLOGP3.

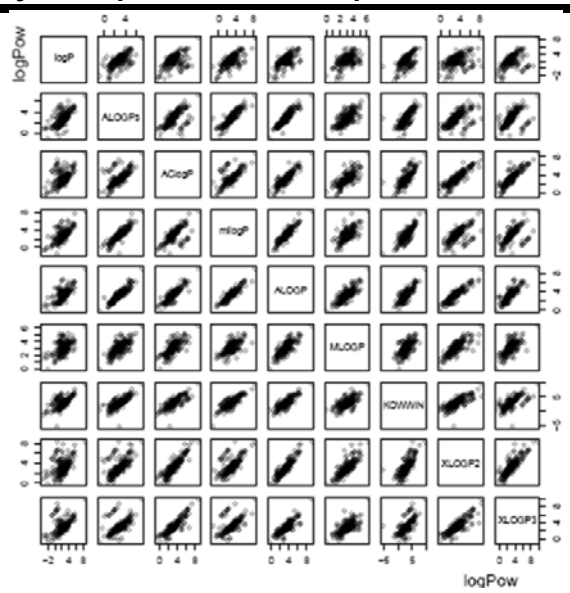

Рис. 2. Порівняння ефективності прогнозних моделей ($logP_{ow}$ calc vis $logP_{ow}$ exp) на всьому наборі даних

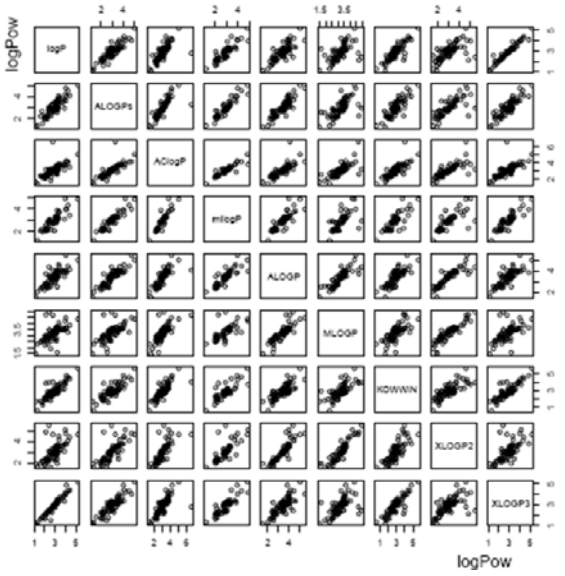

Рис. 3. Аналіз ефективності прогнозних моделей ($logP_{ow}$ calc vis $logP_{ow}$ exp) на рандомізованому наборі

Потенційно зв'язок ефективності прогнозу і розкиду прогнозних оцінок різних моделей міг би бути незалежною оцінкою надійності прогнозу. На рис. 4 порівняно нев'язки експеримент — медіана прогнозу (тобто по відношенню до консенсусного значення) від варіативності прогнозу — суми другого та третього квартилів.

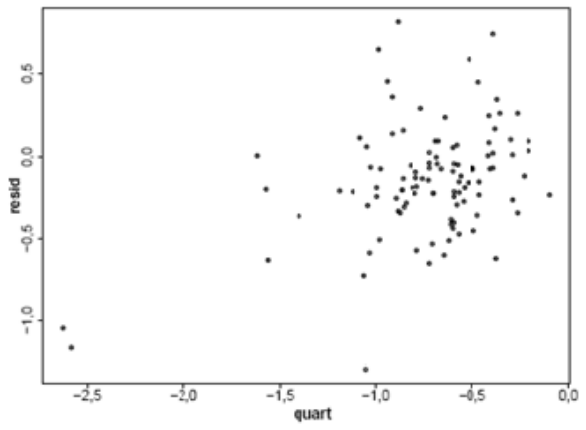

Рис. 4. Зв'язок величин нев'язок експеримент / медіана прогнозу P_{ow} та квартильної міри варіації значень прогнозу (границі другого та четвертого квартилів)

При наявності функціональної залежності, на графіку можна було б очікувати залежність. Але, на жаль, розбіжності у прогнозних значеннях не є ключем до оцінки надійності прогнозу.

Формування підвибірок для контролю надійності є досить поширеним прийомом, що лежить в основі ряду прикладних статистичних методів (бутстреп, джекнайф та інші [9.]) розроблених для використання у галузях з недостатньо розробленими аналітичними методами. Як не дивно, але цей прийом ніколи не використовується у аналізі адекватності моделей P_{ow}. Було сформовано два типи підвибірок: ряди R та T. Ряд R створювався виходячи з формальної оцінки хімічної активності сполуки лише з структурної формули.

Підвибірки ряду R — класифікація за хімічними особливостями:
Rall всі сполуки; R1 не здатні до сильних специфічних взаємодій; R2 сполуки, в яких присутні легко йонізуємі групи (карбонова, фенольна тощо); R3 з ізольованими гетероциклами і мінімальною кількістю замісників; R4 з неанельованими гетероциклами і без замісників, здатних до специфічних взаємодій з розчинником; R5 з подвійним зв'язком у одного з ендоциклічних атомів азоту (обмеження рухливості 7-членного циклу).

Ряд T (табл.1) класифікація сполук за структурними особливостями молекул.

Таблиця 1. Підвибірки ряду Т

Азепіни			Діазепіни					
			1,2-діазепіни			1,4-діазепіни		
ізольоване ядро	анельований з		ізольоване ядро	анельований з		ізольоване ядро	анельований з	
	феніл	гетеро-цикл		феніл	гете-роцикл		феніл	гетеро-цикл
T1	T2	T3	T4	T5	T6	T7	T8	T9

На рис. 5 порівняні статистичні характеристики загального набору (1) та підвибірок R2 – 4.

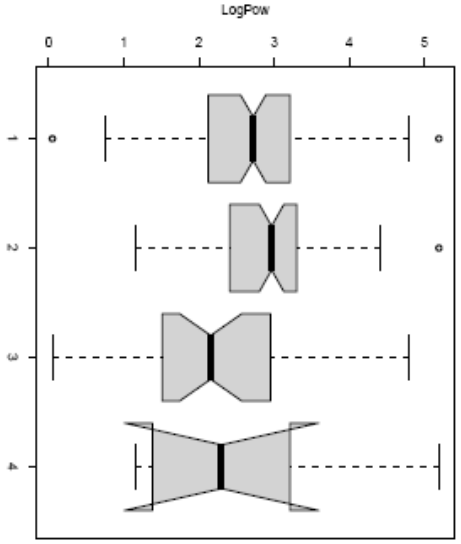

Рис. 5. Статистичні характеристики загального набору (1) та підвиборок (2–4)

Видно, що підвибірки мають близькі до основної вибірки характеристики, зокрема, збігається медіана і є близьким розмах (другий + третій квартилі). Тільки остання вибірка має трохи більшу варіативність (як наслідок меншого обсягу та більшого розкиду значень).

Показовим виявилося використання масиву (вибірки) R3. Рис. 6 демонструє, що для набору гетероциклічних сполук з мінімальною кількістю замісників дві з моделей – XLOGP3 та ALOGPs (гірше) – здатні забезпечити стабільно високу якість прогнозу, MilogP виявилася неспроможною відобразити запропоновані структури через принциповий недолік «словника».

Досить цікавою є значна різниця між XLOGP2 і XLOGP3 та дійсно кращі результати останньої.

Гетероциклічні сполуки складний об'єкт для атомарних моделей виходячи з відомих особливостей їх структури, зокрема з кооперативного характеру ароматичності тим не менше результат є задовільним. Деякою мірою він збігається з отриманим у [5] для похідних дібензодіоксинів, у масиві яких було представлено також сполуки без екзоциклічних замісників.

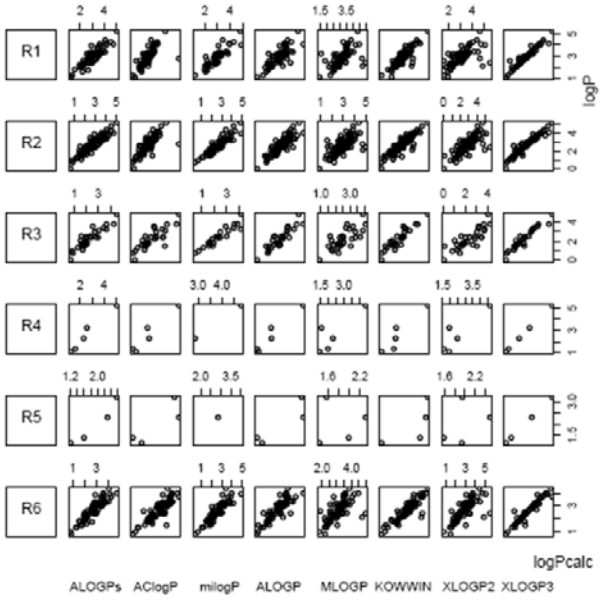

Рис 6. Порівняння ефективності прогнозних моделей ($logP_{ow}$ calc vis $logP_{ow}$ exp) для вибірки ряду R

Легко бачити у порівнянні з рис. 2 що вибірки ряду Т демонструють поведінку від майже тотожної до основної вибірки (відсутність лінійної залежності розрахунку та експерименту, Т8) до майже ідеальних прямих Т9, Т3 (рис. 7).

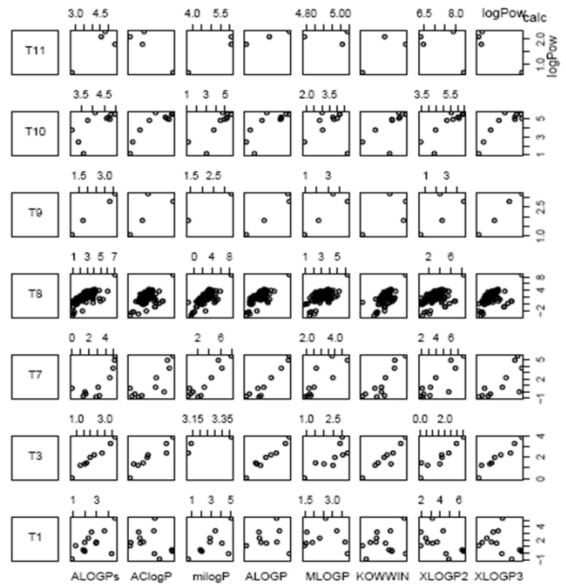

Рис. 7. Порівняння ефективності прогнозних моделей ($logP_{ow}$ calc vis $logP_{ow}$ exp) для вибірки ряду T

Слід зазначити, що найбільш вдалі прогнози демонструють лише ALOGPs та XLOGP3, і цей висновок можна вважати більш-менш надійним та незалежним від вибірки. Таким чином, обраний тестовий набір сполук не дивлячись на малий об'єм, дозволяє отримати всі види поведінки результатів прогнозу і наочно демонструє:

- нестійкість оцінювання моделей при використанні невипадкових наборів даних
- достатність обраного набору для тестування моделей.

Таким чином, прийнята методика формування вибірки повністю виправдалася.

Висновки

Запропоновано детермінований підхід до формування тестових наборів даних для QSAR моделей на прикладі коефіцієнту розподілу н-октанол/вода.

Створено тестовий набір та на ньому відпрацьован широкий ряд існуючих емпіричних прогнозних моделей.

Показано, що з тестованих моделей лише ALOGPs та XLOGP3 та деякою мірою ALOGP здатні при певних умовах до прийнятної якості прогнозу. Зазначено, що ці кращі моделі реалізують відмінну від інших стратегію прогнозу – розрахунок не значення, а відмінності від спорідненої сполуки.

Перелік використаної літератури

1. Calculation of molecular lipophilicity : State-of-the-art and comparison of logP methods on more than 96,000 compounds [Text] / R. Mannhold, G. I. Poda, C. Ostermann, I. V. Tetko // J. Pharm. Sci. — 2009. — Vol. 98, № 3. — P. 861–893.

2. Balakin, K. V. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds : trends, problems and solutions [Text] / K. V. Balakin, N. P. Savchuk, I. V. Tetko // Curr. Med. Chem. — 2006. — Vol. 13, № 2. — P. 223–241.

3. Sangster, J. Octanol-water partition coefficients : fundamentals and physical chemistry [Text] / J. Sangster. — Chichester : John Wiley & Sons Ltd, 1997. — 178 p.

4. Jeffry W. Godden Analysis of Chemical Information Content Using Shannon Entropy / Jeffry W. Godden, Jürgen Bajorath // Reviews in Computation Chemistry. — 2007. — Vol. 23 (5). — P. 263–298.

5. Дослідження впливу структури N-гетроциклічних сполук на їх ліпофільність [Текст] : Звіт про НДР (заключ.) / Ін-т фізико-органічної хімії і вуглехімії ; кер. Луцик О. І. ; викон. : Суйков С. Ю. [та ін.].— Донецк, ІнФОВ, 2007. — 142 с. — № ДР 0105U001017. — Інв.№ 0208U004220.

6. Solvation enthalpies of neutral solutes in water and octanol [Text] / A. Bidon-Chanal, O. Huertas, M. Orozco, F. J. Luque // Theor. Chem. Acc. — 2009. — Vol. 123, № 1–2. — P. 11–20.

7. Benchmarking of linear and nonlinear approaches for quantitative structure−property relationship studies of metal complexation with ionophores [Text] / I. V. Tetko [et al.] // J. Chem. Inf. Model. — 2006. — Vol. 46 (2). — P. 808–819.

8. Тьюки, Дж.У. Анализ результатов наблюдений : разведочный анализ [Текст] / Дж.У.Тьюки. — М. : Мир, 1981. — 696 с.

9. Орлов А.И. Эконометрика Учебник. / А.И. Орлов. — М.: Издательство «Экзамен», 2002. — 576 с.

10. Avdeef, A. Absorption and drug development : solubility, permeability, and charge state [Text] / A. Avdeef. N. Y. : Wiley-Interscience, 2003. — 294 p.

Надійшла до редколегії 01.02.2012