

на правах рукопису

ТЕТКО ІГОР ВОЛОДИМИРОВИЧ

ЗАСТОСУВАННЯ НЕЙРОННИХ МЕРЕЖ ДЛЯ ВИЯВЛЕННЯ
ЗВ'ЯЗКУ СТРУКТУРА-БІОЛОГІЧНА АКТИВНІСТЬ

02.00.10 - Біоорганічна хімія, хімія природних
та фізіологічно активних речовин

Автореферат
дисертації на здобуття наукового ступеня
кандидата хімічних наук

Київ - 1994

Дисертацією в рукописі.

Робота виконана у відділі медико-біологічних проблем Інституту біоорганічної хімії та нафтохімії АН України

Науковий керівник: доктор медичних наук,
професор О.І. Луйк

Офіційні опоненти: доктор хімічних наук,
професор А.Я. Ільченко,
кандидат технічних наук
О.М. Різник

Провідна установа: Фізико-хімічний інститут АН України
ім. О.В. Богатського,
м. Одеса

Захист відбудеться "25" марта 1994 р. на
засіданні спеціалізованої вченої ради Д 016.65.01 в Інституті
біоорганічної хімії та нафтохімії АН України
(253660, м. Київ, вул. Мурманська, 1).

З дисертацією можна ознайомитися у бібліотеці Інституту
біоорганічної хімії та нафтохімії АН України (253660, м. Київ, вул.
Мурманська, 1).

Автореферат розісланий "25" февраля 1994 р.

Вчений секретар
спеціалізованої вченої ради

Федорук

Д.М. Федорук

ЛННБ України ім.В.Стефаніка



00801813 (L)

Актуальність проблеми Передбачення біологічної активності хімічних речовин, виходячи з їх будови, та створення нових препаратів із заданими властивостями є однією з найголовніших задач сучасної медичної хімії (Fujita, 1984; Dunn III, 1989). Згідно даних статистики (Knight, 1990) в провідних фірмах світу тільки одна із 10-15 тисяч синтезованих та випробованих хімічних речовин знаходить використання в медичній практиці. Через це вартість створення нового лікарського препарату в середньому сягає 600-700 мільйонів доларів і на це витрачається 8-12 років. Методи комп'ютерного скринінгу дозволяють значно прискорити процедуру такого пошуку і заощадити значні матеріальні кошти, які були б марно витрачені на синтез та наступні біологічні випробовування малоефективних речовин.

Знаходження зв'язку структура-біологічна активність є однією з найскладніших математичних задач пошуку закономірності тому, що кількість доступних для аналізу даних, як правило, вельми обмежена, а сама залежність параметри-активність має складний нелінійний характер. Отримання нових даних пов'язано із значними затратами фінансів та часу, а крім цього існує ймовірність неправильного формування навчальної вибірки через немінучі експериментальні похибки. Тому дослідники, які працюють над виявленням зв'язку структура-активність (ЗСА) приділяють особливу увагу пошуку та розробці методів знаходження залежностей, котрі б дозволяли отримувати вірогідні прогнози для таких складних задач. Серед цих методів однією з найбільш перспективних є група методів, відомих під загальною назвою нейронні мережі. Ці методи з успіхом застосовувалися для вирішення задач розпізнавання рукописних знаків, мови, фільтрації зображення від шуму, аналізу енцефалограм (Kohonen 1991; Fukushima 1987; Lando e.a., 1993; Gasteiger e.a., 1991). Існує чимало типів нейронних мереж, що відрізняються своєю архітектурою, цілями функціонування та алгоритмами навчання. Найчастіше для пошуку ЗСА застосовується метод нейронних мереж із зворотним розповсюдженням помилки - back-propagation neural networks (НМ). Але безпосереднє застосування НМ до проблеми ЗСА призвело до суперечливих результатів: з одного боку, ряд дослідників повідомляють про значну перевагу НМ над стандартними методами аналізу ЗСА, такими як лінійні регресійні методи, методи лінійного дискримінантного аналізу, метод головних компонент та іншими методами (Aoyama e.a., 1990; Andrea e.a., 1991; Chastrette e.a., 1991 и др.), а деякі дослідники, навпаки, звертають увагу на випадки

досить поганого узагальнення даних цим методом (Liu et al., 1992).

Мета роботи полягала в пошуку алгоритмів поліпшення узагальнення даних НМ та розробці методу для розв'язання задач виявлення закономірностей структура-активність.

Задачі дослідження полягали у наступному:

1) виявити можливі причини погіршення узагальнення даних НМ та розробити алгоритми, які будуть спроможні подолати ці недоліки методу.

2) створити методи оцінки інформативності вхідних параметрів нейронних мереж;

3) проаналізувати результати класифікації молекул методом НМ як попередню обробку для подальшого дослідження методами, що аналізують їх просторову будову, наприклад, електрон-топологічним методом (ЕТМ). Передбачалося використати як приклади деякі класичні задачі пошуку ЗСА, добре досліджені багатьма методами, а також ряд нових задач цього класу: класифікація лікарських речовин по типу їх дії на основні сигнальні системи клітини (ОССК); аналіз інгібіторів зворотної транскриптази (ЗТ) вірусу імунодефіцита людини 1-го типу (ВІЛ-І) та інші задачі.

Наукова новизна Розроблено новий метод пошуку ЗСА на основі застосування алгоритму НМ із зворотним розповсюдженням помилок. Запропоновано для уникнення випадкових кореляцій НМ використовувати статистичний аналіз, який також дозволяє виявити достовірність класифікації молекул із контрольної вибірки. Запропоновано новий оригінальний алгоритм зменшення НМ та оцінки інформативності аналізованих параметрів під час навчання мережі. Цей алгоритм дозволяє збільшити швидкість навчання НМ, знайти найменшу архітектуру мережі, а використання тільки найбільш ефективного набору параметрів допомагає значно поліпшити прогностичні властивості НМ. Результати, отримані за допомогою НМ, мають самостійну наукову цінність, проте вони можуть використовуватись ще й для попередньої обробки даних для більш складних методів пошуку ЗСА.

Теоретичне значення роботи Запропоновано нові ефективні алгоритми поліпшення прогнозу НМ. Розроблені методи можуть використовуватися у всіх галузях науки та техніки, де потрібен розподіл об'єктів на класи.

Практичне значення роботи Запропонований метод пошуку ЗСА був з успіхом застосований для аналізу та прогнозування ЗСА інгібіторів ЗТ ВІЛ-І. Для цього було створено банк найбільш типових інгібіторів ЗТ ВІЛ-І, який включав такі високоактивні інгібітори, як азидотимідин (АЗТ), ddI, ddI, TIBO (Pauwels, 1993) та інші. Деякі з цих речовин застосовуються для лікування синдрому набутого імунодефіциту людини (СНІД). Нейронні мережі навчалися на основі цього банку і використову-

валися для комп'ютерного скринінгу речовин, що були вперше синтезовані в Інституті біоорганічної хімії та нафтохімії АН України. Результати прогнозів ІМ повністю підтвержені наступними біологічними випробовуваннями, проведеними в референс-лабораторії Національного комітету по боротьбі із захворюваннями СНІДом при Президенті України.

Застосування запропонованого методу для задачі класифікації лікарських препаратів по типу їх дії на основі сигнальні системи клітини дозволило з високою ймовірністю (74%) правильно передбачити активність 90 препаратів із контрольної вибірки. На основі послідовного аналізу результатів прогнозу ІМ була створена навчальна вибірка найбільш типових препаратів. Аналіз цієї вибірки електрон-топологічним методом дозволив виявити фармакофорний елемент, який визначає загальний профіль дії лікарських препаратів на ОСК.

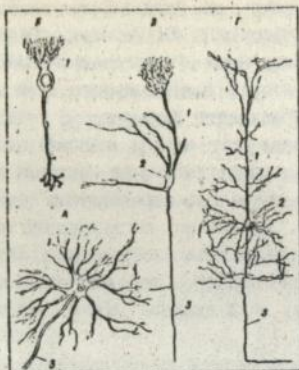
Апробація роботи Основні результати роботи доповідалися на I Всеукраїнській конференції по теоретичній органічній хімії (Волгоград, 1991); Всеросійській науковій конференції "Створення лікарських препаратів" (Москва, 1992); Міжнародному конгресі по комп'ютерним системам та прикладній математиці "CSAM'93" (Санкт-Петербург, 1993); 12^й щорічній конференції Товариства по молекулярній графіці (Швейцарія, 1993); на третій щорічній конференції по еволюційному програмуванню "EP'94" (Сан-Дієго, 1994); на міжнародному семінарі по альдозо редуктазі, (US-Japan Aldose Reductase Workshop, Hawaii, 1994); на науковому семінарі лабораторії нейронних мереж медичного факультету Католицького університету Левена (Бельгія, 1993); на наукових семінарах Інституту проблем математичних машин та систем АН України (Київ, 1993 - 1994); на наукових семінарах відділу медико-біологічних досліджень та конференціях молодих вчених Інституту біоорганічної хімії та нафтохімії АН України (Київ, 1989-1993).

Розмір та структура дисертації Дисертація складається із вступу, п'яти розділів, висновків, списку літератури та додатку. Робота викладена на 154 аркушах машинописного тексту, ілюстрована 11 малюнками та 22 таблицями. Список літератури нараховує 118 джерел.

1. ЗАГАЛЬНИЙ ОПИС МЕТОДУ НЕЙРОННИХ МЕРЕЖ

Штучні нейронні мережі створювалися по аналогії з реальними нейронними мережами головного мозку людини. Спрощена схема реального нейрона показана на Мал. 1. Нейрон складається з тіла, багатьох входів (дендритів) та одного виходу - аксону. Тіло нейрона здійснює просторово-часову сумачію та перетворення сигналів, які надходять до цього нейрону.

Аксон розгалужується на відростки, які передають сигнали на входи

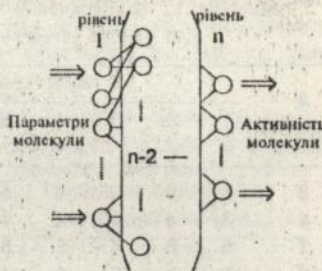


Мал. 1. Схема нейронів з головного мозку тварин та їх штучного аналога. 1-дендрити, 2-тіло клітини, 3-аксон.

інших нейронів. Точки з'єднання називаються синапсами і діють по типу електрохімічної передачі сигналу. Нейрони в мозку об'єднані в складні ієрархічні системи, які обробляють поступаючу інформацію. Штучний нейрон побудовано аналогічно реальному. Він складається із суматора вхідних сигналів та нелінійного перетворювача, який обробляє ці сигнали згідно з нелінійною функцією, такою як гіперболічний тангенс, логістична функція та інші. Доля сигналу, котрий передається від одного нейрону до іншого, пропорційна величині ваги зв'язку між i -м нейроном на рівні v та j -им нейроном на рівні $v+1$ - w_{ij} . Існує чимало типів нейронних мереж, котрі поділяються на дві основні категорії - мережі, які функціонують по типу асоціативної пам'яті та мережі з прямим розповсюдженням сигналів. Ми використовували тільки один із підтипів нейронних мереж з прямим розповсюдженням, а саме нейронні мережі із зворотним розповсюдженням помилки (НМ), котрі, як було вказано у вступі, найбільш часто застосовуються для вирішення проблем пошуку ЗСА (Gasteiger e.a., 1991). НМ, що побудована з n послідовних рівнів, показана на Мал. 2. Нейрони позначені кульками. На перший (вхідний) рівень поступає потік інформації, наприклад, вектор аналізованих параметрів молекул. При проходженні через приховані рівні, а звичайно використовується всього один прихований рівень, цей потік нелінійно трансформується. Для кожного вхідного вектора існує відповідний йому цільовий (вихідний) вектор. Ці вектори складають навчальну пару. Навчання мережі полягає в такому настроюванні зв'язків мережі, коли кожне вхідне збудження призводить до появи на виході нейронної мережі бажаного вектору (спочатку ваги зв'язків іні-

ціються випадково). Навчання НМ відноситься до методу навчання з вчителем і проводиться згідно наступному алгоритму:

1. Вибрати із сукупності всіх навчальних векторів слідуочу навчальну пару. Подати на вхід мережі вхідний вектор.
2. Вирахувати вихід мережі.
3. Вирахувати різницю між отриманими та бажаними виходами мережі.
4. Зкоригувати вагу зв'язків так, щоб зменшити помилку мережі.
5. Повторювати етапи 1-5 до тих пір, поки помилка на всій сукупності не досягне заданого рівня.



Мал. 2. Нейронна мережа з прямим розповсюдженням сигналів, яка складається із n -рівнів.

Існує чимало алгоритмів, котрі можуть використовуватись на четвертому етапі. Ми використовували один з найбільш простих алгоритмів, а саме дельта-правило (Rumelhart e.a., 1986). У всіх своїх дослідженнях ми використовували НМ з одним прихованим рівнем.

1.2. Застосування нейронних ансамблів для усунення неоднозначності прогнозу НМ

Безпосереднє застосування НМ до пошуку ЗСА може призвести до поганого передбачення активності нових молекул (Liu e.a., 1992). Саме такою була наша перша спроба застосувати метод нейронних мереж для пошуку ЗСА похідних карбохінонів - потенційних протипухлинних засобів (Aoyama e.a., 1990). Ми спробували повторити результати цих авторів. 11 молекул використовувались як навчальна вибірка, та 5 молекул використовувались для контролю якості передбачення активності. Нейронні мережі мали ту ж саму архітектуру та параметри навчання, що використовувались японськими вченими. Нейронна мережа чудово навчалася, але передбачення деяких молекул із контрольної вибірки було неоднозначне (Табл. 1).

Для усунення неоднозначності передбачення активності нових молекул ми запропонували використовувати статистичний аналіз ряду незалежних прогнозів НМ або нейронні ансамблі (НА). Можливість поліпшення передбачення при використуванні НА відмічалась в роботі (Hansen, Solomon, 1990). Нами вперше було запропоновано використати закони біноміального розподілу, для визначення вірогідності отриманого прогнозу. Нехай нам потрібно визначити, до якого з двох класів А чи В належить молекула, що аналізується. Якщо ця молекула в m випадках із n була передбачена НМ з

Таблиця 1. Передбачення активності молекул, похідних мітоміцину, для різних початкових ваг w_{ij} матриці зв'язків^a

№ сполуки	№ початкової матриці w_{ij}										еспе- римент- дані
	1	2	3	4	5	6	7	8	9	10	
2	5	5	5	4	5	5	5	5	5	5	5
4	4	4	4	4	4	4	4	4	4	4	4
7	5	5	4	5	5	5	5	5	5	5	4
10	5	4	5	5	4	5	5	5	5	4	2
16	2	3	2	1	3	2	3	2	2	1	1

^a - всі молекули були віднесені до 5 класів активності - 1, ..., 5 (Аоуата, 1990), нейронна мережа мала 4 нейрони на прихованому рівні

типом активності A ($m > (n-m)$), то з достовірністю

$$Pr(p_A > p > 0.5) = \sum_{i=m}^n \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=m}^n \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \cdot 2^{-n} < \alpha/2 \quad (1)$$

молекула має тип активності A (Джонсон та ін., 1987). Формула (1) відповідає перевірці гіпотези $p_A > 1/2$ ($p_A = m/n$, $p_B = (n-m)/n$), тобто того, що в більше половини випадків НМ буде достовірно передбачати молекулу як ту, що має активність типу A . Формула (1) узагальнюється і для випадку класифікації молекул на більшу кількість типів активності. Використання цього критерію дозволяє уникнути неоднозначності передбачення молекул із контрольної вибірки, як показано у Табл. 2.

Таблиця 2. Розраховані активності похідних мітоміцину для 50 різних початкових матриць ваг зв'язків w_{ij} ^a

№ сполуки	Розрахований тип активності (кількість разів)					активність	
	5	4	3	2	1	розр.	екс.
2	48	2	0	0	0	5	5
4	0	50	0	0	0	4	4
7	43	6	0	0	2	5	4
10	38	12	0	0	0	4	3
16	0	0	21	3	26	3 чи 1	1

^a - нейронна мережа мала 4 нейрони в прихованому рівні

Порівняння прогностичної сили методу нейронних мереж із іншими методами пошуку ЗСА було проведено методом змінного контролю. НМ

показали такі ж результати передбачення активності молекул, як і метод адаптивних найменших квадратів (АНК, Moriguchi, 1986). Але, як показано у Табл. 3, ці результати були дещо гірші, ніж отримані методом розмитих АНК (РАНК, Moriguchi e.a., 1990). Слід зазначити, що АНК та РАНК вважаються одними з найкращих методів пошуку ЗСА.

Ми не проводили вибір найбільш ефективних параметрів спеціально для НМ, як це було зроблено для АНК та РАНК (Moriguchi e.a., 1990). Такий вибір параметрів зміг би значно поліпшити передбачення НМ. Проте й отримані результати слід розцінювати як цілком задовільні.

1.3. Розробка методу зменшення нейронних мереж

Знаходження оптимальної архітектури нейронної мережі для вирішення конкретної задачі є однією з найбільш важливих проблем при застосуванні НМ. Результати теоретичних та експериментальних досліджень показують, що чим менша нейронна мережа, тим краще вона має узагальнювати інформацію, що аналізується. Але, з іншого боку, існують теоретичні обмеження найменших розмірів нейронної мережі, за якими нейронна мережа не може взагалі навчитися конкретній задачі (Karnin, 1990). Слід також зазначити, що мережі з невеликою кількістю нейронів на прихованому рівні гірше навчаються, частіше потрапляють в локальні мінімуми. В літературі описані методи зменшення мереж, котрі дозволяють успішно знаходити найменшу архітектуру мережі під час навчання НМ (Omlin e.a., 1993; Karnin, 1990; Moser, Smolensky, 1987).

Проте, відомі методи не дозволяють виявити найбільш інформативні параметри молекул. Ця задача найтіснішим шляхом пов'язана з проблемою пошуку найменшої архітектури мережі. В обох випадках мають бути вилучені деякі нейрони, котрі не мають значного впливу на процес навчання нейронних мереж. Для вирішення цих проблем нами було запропоновано та розроблено оригінальний метод зменшення нейронних мереж. Цей метод реалізує на рівні зв'язків одного нейрону принцип посилення ("reinforcement"), котрий із успіхом використовувався в іншому типі нейронних мереж, що працюють по типу асоціативної пам'яті (Fukushima 1975; 1980). Для кожного зв'язку нейрона розраховується його чутливість згідно з

$$\frac{|w_{ij}^{s-1}|}{|w_j^{s-1}|} \quad (2)$$

де

$$w_{k,1}^{S-1} = w_{\max,1}^{S-1} = \max_k (|w_{k,1}^{S-1}|) \quad (3)$$

Метод полягає в додатковому коригуванні ваг зв'язків згідно:

$$\Delta w_{j1}^S = (\Delta w_{j1}^S)_{old} \cdot \begin{cases} \frac{w_{j1}^S}{(w_{j1}^S)^2}, & \text{якщо } w_{j1}^S < w_{j1}^S \\ -\sum_{k=1}^n \frac{(w_{k1}^S)^2}{(w_{j1}^S)^3}, & \text{якщо } w_{j1}^S = w_{j1}^S \end{cases} \quad (4)$$

де α - коефіцієнт нормування. При цьому зв'язки з найбільшою чутливістю отримують додаткове стимулювання для свого росту, а інші зв'язки пригнічуються.

Аналогічно чутливості зв'язку розраховується чутливість нейронів:

$$S_i = \sum_j \frac{(w_{ij}^S)^2}{(w_{ij}^S)^2} \quad (5)$$

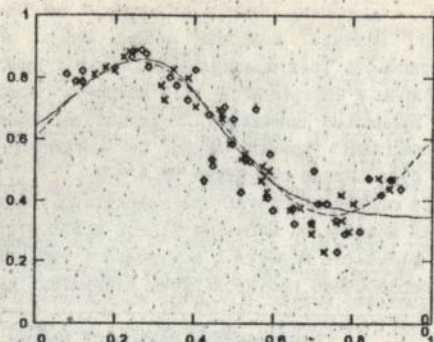
яка відображає величину зв'язку нейрона i на рівні v з усіма нейронами на наступному рівні. Вилючення найменш чутливих нейронів дозволяє знайти найменшу архітектуру нейронної мережі для вирішення поставленої задачі. Після завершення навчання НМ ми можемо проаналізувати, згідно формули (5), важливість кожного параметру для формування досліджуваної залежності та вилючити найменш важливі. Адекватність методу була перевірена на прикладі ряду теоретичних задач розпізнавання образів. Наприклад, для інтерполяції функції синуса (Мал. 3) запропонований метод дозволив вірогідно виявити з двох тестових наборів найменш зашумлений набір даних. Результати по знаходженню найменшої архітектури нейронної мережі для проблем декодування та моделювання бульової функції "виключаючого або" були аналогічні результатам, отриманим іншими відомими методами зменшення нейронних мереж, такими як розпад зв'язків, розпад чутливості нейронів та іншими (Kruschke e.a., 1991).

Таким чином, теоретичне моделювання дозволило нам переконалися в адекватності запропонованого алгоритму для вилучення зайвих нейронів для динамічного знаходження найменшої архітектури нейронних мереж. Оскільки запропонований нами алгоритм виявився ефективним для вирішення теоретичних задач проблем розпізнавання образів, він міг далі використовуватися для вирішення практичних задач пошуку ЗСА.

2. ПЕРЕВІРКА ЗАПРОПОНОВАНИХ АЛГОРИТМІВ ПОЛІПШЕННЯ НМ ДЛЯ ВИРІШЕННЯ ПРАКТИЧНИХ ЗАДАЧ ЗСА

2.1 Передбачення активності інгібіторів зворотної транскриптази ВІЛ-І.

Для вирішення цієї проблеми нами було проведено аналіз сучасної літератури по інгібіторам ЗТ ВІЛ-І і створено комп'ютерний банк молекул,



Мал. 3. Інтерполяція зашумленої функції $y = a \sin(x) + b$. Для навчання НМ використовувалось два набори даних, що позначені на малюнку як \circ та \times . НМ виявила набір 1 (позначений \circ), згідно формулі (5), як більш чутливий для навчання, у порівнянні з набором 2. Це повністю відповідає дійсності, оскільки набір 1 менш зашумлений від набору 2.

який включав 44 структури. Всі речовини були поділені на два класи: активні та малоактивні речовини. Ті сполуки, для яких відношення $1x$ ED_{50} до ED_{50} АЗТ було більш, ніж 10^3 , вважали неактивними. Частина з цих молекул (30) використовувалась для навчання нейронної мережі, а інші речовини (14) використовувалися як контрольна вибірка.

Для кодування молекул розраховувалось 46 топологічних індексів. Ці індекси включали як добре відомі індекси Кіра, Балабана, Рандіча та ін., які вже раніш широко використовувалися для передбачення біологічної активності (Kier, 1976), так і ряд нових індексів, запропонованих нами вперше (Пода та ін., 1993). Навчання нейронної мережі одночасно з усіма індексами неможливе. В цьому випадку нейронна мережа давала б дуже погане узагальнення внаслідок перенавчання. З іншого боку, неможливо використовувати запропонований алгоритм зменшення, тому що це вимагало б забагато часу для зменшення такої великої мережі. Тому ми використали метод попередньої оцінки вхідних параметрів згідно ієрархічному кластерному аналізу (принцип середнього зв'язку). Всі параметри нормувалися на одиничну дисперсію. Кластерний аналіз дозволив поділити всі параметри на 6 груп. З кожного кластеру використовувалось для навчання нейронної мережі по одному параметру, що мав найбільший коефіцієнт кореляції з вектором активності молекул із навчальної вибірки.

Спочатку нейронна мережа мала структуру 6-10-2. Шість нейронів на

першому рівні відповідали кількості параметрів молекул, а 2 нейрони на вихідному рівні - кількості класів ("активні" та "неактивні" сполуки) для класифікації молекул. Нейронна мережа швидко навчалася і вірно передбачала активність усіх молекул з навчальної вибірки. Навіть після скорочення 7 нейронів внутрішнього та 2 нейронів першого рівня нейронна мережа все ще правильно навчалася. Подальше скорочення призводило до помилкової класифікації деяких сполук із навчальної вибірки. Тому як остаточна нами була обрана структура НМ 4-3-2.

У Табл. 4 приведені результати класифікації молекул із контрольної вибірки нейронними мережами різної архітектури.

Таблиця 4. Кількість молекул з контрольної вибірки, для яких було невірно передбачено противірусну активність щодо ЗТ ВІЛ-1^а

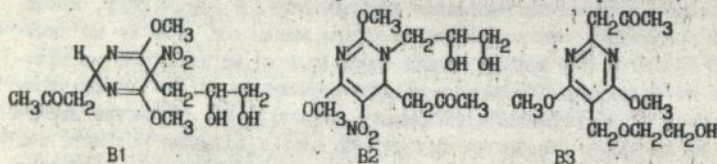
КІЛЬКІСТЬ МОЛЕКУЛ	Архітектура нейронної мережі			
	6-10-4	6-3-4	4-10-4	4-3-4
	4	5	3	2

^а - результати отримані на основі статистичного аналізу прогнозу 40 незалежних мереж згідно формули (1). Отримані результати статистично достовірні ($p < 0.01$).

Передбачення активності для молекул із контрольного набору значно поліпшилось при використанні НМ із меншою архітектурою.

НМ із будовою 4-3-2 використовувалися для прогнозу активності 20 нових сполук, які були вперше синтезовані в Інституті біоорганічної хімії та нафтохімії АН України і люб'язно надані нам С.Г. Вишневським та ін. Тільки одна з речовин В2 (Мал. 4) була передбачена як активна з рівнем достовірності $p < 0.01$. Всі інші молекули були передбачені з тією ж достовірністю як неактивні.

Біомедичні дослідження проводилися тільки для 3 речовин,



Мал. 4. Молекули, які використовувались в біомедичних дослідженнях. Молекула В2 була передбачена як високоактивний інгібітор ЗТ ВІЛ-1, що було підтверджено біологічними випробуваннями.

молекули В2, передбаченої, згідно прогнозу, як найбільш активна, та двох її найближчих структурних аналогів В1 та В2 (Мал. 4). Молекула В2 виявилася високотоксичною, і для неї подальші біологічні випробування на противірусну активність не проводилися. Активність сполук В1 та В2 оцінювалася чотирма методами. Було виявлено, що сполука В2 має таку ж активність проти ЗТ ВІЛ-І, як і АЗТ, але в 10 разів менш токсична для клітин. Сполука В1 виявилася неактивною. Таким чином, біологічні випробування повністю підтвердили комп'ютерний прогноз активності молекул.

2.2 Застосування НМ для класифікації лікарських препаратів по типу їх дії на основі сигнальної системи клітини (ОССК).

В ІВОНХ під керівництвом проф. О.І. Луйка інтенсивно проводяться дослідження по класифікації лікарських препаратів. На першому щаблі ієрархічної класифікації пропонується розподіл речовин на два основні класи - "+1/-2" - ксенобіотики, які посилюють сигнали, що проходять через аденілатіциклазну систему (АДЦ) і пригнічують передачу сигналів через фосфоліпідну сигнальну систему (ФЛСС) та "-1/+2" - речовини, які мають протилежний тип дії. Нами було проаналізовано більш як 200 препаратів, які відносяться до цих груп. Частина цих молекул вже аналізувалася раніше (Пода, 1992). Оскільки стандартні методи лінійного дискримінантного аналізу (ЛДА Фішера, ЛДА методом найменших квадратів) показали низькі результати при навчанні (60-70%) та прогнозі нових молекул (50-60%), Пода застосовував для аналізу оригінальний ймовірнісно-статистичний метод (ИСМ). Неefективність методів ЛДА вказує на складність вирішення цієї задачі.

Ми спробували вирішити цю задачу методом НМ. Контрольна вибірка була значно розширена за рахунок включення 50 нових молекул, які були віднесені до одного з двох класів активності згідно принципам, що викладені в роботі Луйка, Могілевича (1992). Слід відмітити, що внаслідок неповноти інформації про дію сполук для деяких молекул як із навчальної, так і контрольної вибірки міг бути неправильно визначений тип їх активності щодо ОССК.

Навчання НМ проводилося, як описано вище при аналізі інгібіторів ЗТ ВІЛ-І. Застосування НМ з меншою архітектурою дозволило значно поліпшити передбачення активності молекул із контрольної вибірки, як показано на Табл. 5.

НМ показали кращі результати при навчанні та прогнозі активності молекул із контрольної вибірки, в порівнянні з ІСМ та методом k-найближчих сусідів. Проте, аналіз молекул, котрі кодується вектором їх параметрів, не дозволяє виявити ті структурні елементи чи фармакофор,

Таблиця 5. Класифікація сполук по типу їх дії на ОССК

Метод	% вірних класифікацій для вибірок ^а	
	навчальної	контрольної
k найближчих сусідів	75	68
МСМ	87	70
HM ^b (15-15-2)	100	70
HM (15-6-2)	100	68
HM (10-15-2)	100	72
HM (10-6-2)	95	74

^а - навчальна вибірка складалась із 111, а контрольна-із 90 молекул

^б - в дужках показана архітектура нейронної мережі.

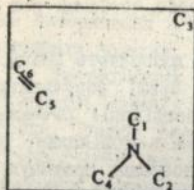
які відповідають за прояв типу активності, що спостерігається. Такий аналіз можна зробити, наприклад, ЕТМ.

Для аналізу ЕТМ було взято по 20 молекул із кожної групи. Це були, головним чином, жорсткі структури з обмеженим числом ступенів свободи, для яких була правильно передбачена активність методом НМ. Застосування ЕТМ дозволило виявити фармакофор (Мал. 5), який зустрічався у 17 із 20 молекул з групи "+1/-2". Цей фармакофор взагалі не зустрічався у структурі речовин групи "-1/+2" і може бути екзамплярною ознакою для віднесення молекул до групи "+1/-2". Спроби знайти який-небудь загальний фрагмент у будові речовин із групи "-1/+2" не призвели до статистично значимих результатів. Цей факт можна пояснити наявністю декількох альтернативних механізмів взаємодії молекул групи "-1/+2" з компонентами ОССК, для кожного з яких характерна своя топологія активного центра.

Обговорення результатів досліджень

Як було вказано у вступі, великий потік нових синтезованих хімічних сполук, необхідність їх біологічних випробувань вимагають значних матеріальних та людських ресурсів. Ці проблеми постійно стимулюють дослідників на створення теоретичних розрахункових методів передбачення біологічної активності речовин на основі їх структурної формули. Проблема пошуку ЗСА є складною задачею. НМ, які мають чудові якості щодо нелінійної екстраполяції, можуть виявитися найбільш адекватними для вирішення цієї надзвичайно важливої задачі.

Проте, в деяких випадках безпосереднє застосування НМ для аналізу ЗСА може привести до виявлення випадкових зв'язків, що в результаті погіршує результати передбачення активності нових молекул. Для випадку



	C ₁ (0.12)	N (-0.34)	C ₂ (0.13)	C ₃ (-0.06)	C ₄ (0.12)	C ₅ (-0.06)	C ₆ (-0.04)
C ₁		1.46	2.40	4.86	2.42	5.01	6.12
N			1.46	5.70	1.46	6.38	7.37
C ₂				5.65	2.39	7.28	8.28
C ₃					6.86	5.62	6.50
C ₄						7.08	8.25
C ₅							1.40

Мал. 5. Матриця відстаней між атомами фармакофора, що був знайдений ЕТМ. У дужках вказані заряди на атомах.

дослідження кількісного зв'язку структура-активність (Andrea e.a., 1990) показана можливість погіршення передбачення активності нових молекул при зменшенні параметру ρ , який дорівнює відношенню кількості аналізованих даних до кількості ступенів свободи мережі. Наші дослідження задач ЗСА вказують на можливість виникнення за цих же причин неоднозначного передбачення активності молекул із навчальної вибірки НМ. Проблема уникнення нестабільності передбачення НМ є актуальною при застосуванні НМ не тільки для задач пошуку ЗСА, але і для великої кількості задач практичного використання НМ в інших галузях науки та техніки. Нами було показано, що подолання цього недоліку НМ можливе при використанні статистичного аналізу ряду незалежних передбачень методу (нейронних ансамблів), отриманих при навчанні нейронних мереж із випадковими початковими значеннями матриці зв'язків. Застосування НА, що дуже важливо, дозволяє оцінити рівень достовірності отриманого прогнозу.

Згідно з теоретичними дослідженнями (Sietsma e.a., 1991; Hinton, 1987; Hirose e.a., 1991), узагальнення нових даних НМ має поліпшуватися при використанні НМ з меншою архітектурою. Ці висновки були підтверджені нами при передбаченні активності інгібіторів 5T ВІЛ-І та класифікації речовин по типу їх дії на ОССК. Застосування НМ із меншою архітектурою збільшувало процент вірних передбачень. Особливо великий вплив на поліпшення передбачення активності речовин чинило зменшення числа неінформативних параметрів. Для вирішення проблеми вилучення таких параметрів та динамічного знаходження найменшої архітектури нейронної мережі нами було запропоновано новий оригінальний метод. Ідея запропонованого методу аналогічна ідеї конкуренції та посилення ("reinforcement") нейронів, яка була застосована в моделях когнітрона та неокогнітрона (Fukushima, 1980), але в нашому випадку ця ідея реалізована на рівні зв'язків одного нейрону. Ефективність запропонованого

методу було перевірено на прикладі як теоретичних задач проблеми розпізнавання образів, так і практичних задач пошуку ЗСА.

Розроблені методи поліпшення узагальнення даних дозволяють значно поліпшити передбачення НМ. Наприклад, НМ показали кращі результати передбачення активності нових молекул для задачі класифікації речовин по типу їх дії на основні сигнальні системи клітини в порівнянні з методом k-найближчих сусідів та ИСМ. Для задачі передбачення протипухлинної активності молекул, похідних мітоміцину, результати передбачення НМ співпали із результатами АНК (Moriguchi, 1986), але були дещо гірші, ніж отримані методом РАНК (Moriguchi e. a., 1990). Ми не проводили детального порівняння результатів передбачення НМ із результатами, отриманими іншими методами на великій кількості інших задач. Але й ці отримані результати є багатообіцяючими, зважаючи на велику складність розв'язуваних задач (Пода, 1992; Moriguchi e. a., 1990). На жаль, висновки про перевагу НМ, у порівнянні з традиційними методами виявлення залежності, неможливо зробити на основі аналізу інших літературних даних (Chastrette e. a., 1991; Li e. a., 1993; Sitting e. a., 1992), хоча в згаданих працях автори роблять як раз такі висновки. Дослідження, що описані в цій дисертації, вказують на можливість спостереження випадкових кореляцій при прогнозі активності нових молекул. Як раз такі кореляції мали місце в роботі (Aoyama e. a., 1990), результати якої ми використали в своїх дослідженнях. З іншого боку, критичні зауваження про випадки поганого передбачення НМ для нових молекул (Liu e. a., 1992) також можуть бути наслідком випадкових кореляцій НМ, з тих же причин.

Хоча наш аналіз торкався випадків застосування НМ для вирішення задач пошуку ЗСА, деякі висновки можуть бути зроблені також і для проблеми пошуку кількісного зв'язку структура- активність. Для отримання сталих прогнозів в цих випадках також доцільно проводити статистичний аналіз низки незалежних передбачень та розраховувати середнє значення та дисперсії активності кожної молекули, що аналізується.

На жаль, методи пошуку ЗСА, які використовують кодування молекули вектором її параметрів (1D методи), не дозволяють виявити конкретні фрагменти будови молекул, що відповідають за той чи інший вид біологічної активності молекул. Саме пошук такого фармакофору є найбільш важливим для практичного цілеспрямованого створення нових речовин. На прикладі задачі класифікації молекул по типу їх дії на ОССК ми проаналізували можливість використання 1D методів для попередньої обробки масиву молекул та вибору невеликої кількості найбільш типових структур для подальшого аналізу методами, що враховують просторову будову речовин.

Ми використали ЕТМ, який дозволяє найбільш чітко виявити фармакофорні елементи молекул.

Застосування ЕТМ дозволило виявити фармакофор, який існує у більшості молекул із групи "+1/-2" і відсутній у молекул із протилежної групи. Спроби виявити який-небудь загальний фрагмент для молекул із групи "-1/+2" не призвели до статистично значимих результатів. Це можна пояснити наявністю більш чіткого розподілу електронної густини оподук першої групи та передбачити, що механізми взаємодії цих речовин з регулюючим сайтом біополімеру визначаються наявністю знайденого фрагменту. Стосовно речовин групи "-1/+2" можна передбачити наявність декількох альтернативних або взаємодоповнювальних механізмів, для кожного з яких характерна своя топологія активного центру.

ВИСНОВКИ

1. Показана можливість виникнення нестабільності прогнозів, випадкових кореляцій ІМ при його застосуванні для задач пошуку ЗСА.

2. Вперше для аналізу передбачень методом ІМ застосовано статистичний підхід, основою якого є закони біноміального розподілу. Цей підхід дозволяє запобігти випадкових кореляцій та врахувати достовірність прогнозу активності нових молекул.

3. Розроблено оригінальний метод спрощення ІМ, що дозволяє виявити під час навчання найбільш важливі параметри та динамічно знаходити найменшу архітектуру ІМ. Ефективність алгоритму продемонстрована як на прикладі класичних задач теорії розпізнавання образів, так і практичних задачах пошуку ЗСА.

4. На прикладі прогнозування інгібіторів ЗТ ВІІ-І та класифікації речовин по типу їх дії на ОССК показано підвищення вірних передбачень активності молекул ІМ при використанні вказаних вище алгоритмів. Отримані результати підтверджують теоретичні висновки про поліпшення узагальнення даних нейронними мережами з меншою архітектурою.

5. Показано, що аналіз даних нейронними мережами може бути використаний як попередня класифікація молекул для наступних досліджень методами, що враховують просторову будову молекул. Застосування ЕТМ для задачі класифікації сполук по типу їх дії на ОССК дозволило виявити фармакофорні елементи, котрі можуть визначати збудження АІД системи та гальмування проведення сигналів ФЛСС.

6. Проведені дослідження підтверджують справедливність використаного підходу до раціональної класифікації фізіологічно активних речовин та відкривають нові перспективи їх розрахункового конструювання.

Основні матеріали дисертації викладені в публікаціях.

1. Луйк А.И., Кухарь В.П., Радченко И.В., Ильяшова Л.М., Найденова И.Ю., Пода Г.И., Кондрашова Л.Н., Тетко И.В. Стереотипный компонент механизма действия ксенобиотиков // ДАН УССР, Сер. Б. - 1990. - № 8. - С. 67-70.
2. Пода Г.И., Тетко И.В., Луйк А.И., Могилевич С.Е., Кухарь В.П. Расчет структуры универсального фармакофора, определяющего элементы стереотипы в действии лекарств // ДАН СССР. - 1991. - т. 317, № 2. - С. 488-491.
3. Луйк А.И., Тетко И.В. Анализ ряда флавоноидов - ингибиторов ЦАМФ фосфодиэстеразы модифицированным методом Хопфингера. // Тезисы I Всесоюзной конференции по теоретической органической химии. Волгоград. - 1991. - С. 520.
4. Пода Г.И., Тетко И.В., Могилевич С.Е., Луйк А.И. Выявление элементов структурной общности веществ, сходно действующих на клеточные сигнальные системы // Хим.-Фарм. журнал - 1992. - № 2. - С. 15-18.
5. Тетко И.В., Танчук В.Ю., Луйк А.И. Применение нейронных сетей для анализа взаимосвязи "структура-активность" при малом числе молекул // Тезисы Российской научной конференции "Конструирование лекарственных средств". Москва. - 1992. - С. 43-44.
6. Тетко И.В., Пода Г.И., Танчук В.Ю., Луйк А.И. Анализ модифицированным методом Хопфингера фосфодиэстеразной ингибирующей активности флавоноидов. // Теор. и эксп. химия. - 1993. - т. 29, № 1. - С. 63-69.
7. Пода Г.И., Танчук В.Ю., Тетко И.В., Косель М.И., Луйк А.И. Применение топологических индексов для прогнозирования 5-липоксигеназной ингибирующей активности в ряду гидроксаматов. // Теор. и эксп. химия. - 1993. - т. 29, № 2. - С. 122-125.
8. Tetko I.V., Luik A.I., Poda G.I. Applications of Neural Networks in Structure-Activity Relationships of a Small Number of Molecules. // J. Med. Chem. - 1993. - v. 36, № 7. - pp. 811-815.
9. Tetko I.V., Luik A.I. A Node Pruning Algorithm for Feed-Forward Neural Networks. // J. Intelligent Control, Neurocomputing and Fuzzy Logic. - 1993. - v.1, №2. - pp.
10. Tetko I.V., Koshel M.I., Luik A.I. The Simple Pruning Algorithm for Feed-Forward Neural Networks. // Abstracts of International Congress on Computer Systems and Applied Mathematics CSAM'93. St. Petersburg. - 1993. - pp. 219-220.
11. Tetko I.V., Tanchuk V.Yu., Luik A.I. Applications of Neural Networks to a Small Dataset Structure-Activity Relationship. // Poster

Presentations of 12th Annual Conference of the Molecular Graphics Society. Interlaken, Switzerland. - 1993. - p. 86.

12. Tetko I.V., Tanchuk V.Yu., Luik A.I. Application of an Evolutionary Algorithm to the Structure-Activity Relationship. // Proceeding of EP'94 (The Third Annual Conference on Evolutionary Programming). San Diego. - 1994. pp. 207-222.

13. А.С. Госкомизобретений СССР от 27.06.91, № 4951446/04 (055216): 3-(2-Бензофурил)-7-гидрокси-8-метилхромон, обладающий противовоспалительной активностью (В.П. Хилия, В.Н. Ковалев, И.А. Зупнец, Г.И. Пода, И.В. Тетко, С.А. Васильев).

И.Тетко

Підп. до друку. 18.02.93. Формат 60 × 84¹/₄ Папір офс

Друк. офс. Умовн. друк. арк. 0,93 Обл.-вид. арк. 0,66 тир. 100

Зам. 4-2006

Київська книжкова друкарня наукової книги. Київ, Репін, 4.

AB 29.863

AB 29.863