БИООРГАНИЧЕСКАЯ ХИМИЯ, 2001, том 27, № 4, с. 303–313

УДК 541.6:519.688

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ, ИСПОЛЬЗУЮЩИХ АЛГОРИТМ ПРОСТРАНСТВЕННОГО ОБУЧЕНИЯ ДЛЯ ИССЛЕДОВАНИЯ СВЯЗЕЙ СТРУКТУРА–АКТИВНОСТЬ ХИМИЧЕСКИХ СОЕДИНЕНИЙ

© 2001 г. В. В. Ковалишин#, И. В. Тетко, А. И. Луйк, Ж. Р. Кретиен*, Д. Ливингстоун**

Институт биоорганической химии и нефтехимии, 02094, Киев, Мурманская, 1, Украина;

**Лаборатория биоинформатики, Университет Орлеана, Орлеан, Франция;*

*** Центр молекулярного моделирования, Университет Портсмута, Нант, Великобритания*

Поступила в редакцию 12.04.2000 г. Принята к печати 12.01.2001 г.

На примере производных *N*-бензилпиперидинов показано, что для успешного решения задач описания пространственных количественных соотношений структура–активность (3D-ККСА) может быть использован алгоритм пространственного обучения (АПО) искусственных нейронных сетей (ИНС). Новый алгоритм является комбинацией двух типов нейронных сетей – сети Кохонена и ИНС с прямым распространением сигналов, объединенных для решения общих проблем. Применение методов отбора наиболее информативных признаков позволило выявить наиболее важные регионы вокруг молекулы, влияющие на ее активность. Кластерные зоны, определенные с помощью нового алгоритма, показали хороший прогноз активности молекул из контрольной выборки.

Ключевые слова: структура–активность; искусственные нейронные сети; алгоритм пространственного обучения; кластеры; N-бензилпиперидины.

ВВЕДЕНИЕ

При решении практических задач в различных областях науки исследователь все чаще вынужден привлекать моделирование. В частности, в области химии одной из таких задач является выявление связей между структурой вновь синтезированных органических соединений и их биологической активностью с помощью построения математических моделей. Известно, что биологическая активность любого вещества определяется совокупностью факторов, обеспечивающих (1) ее способность проникать внутрь организма и переноситься к мишени взаимодействия, избегая связывания и накапливания в инертных компонентах и тканях, (2) противодействовать разлагающим ферментам и (3) взаимодействовать с определенным фармакологическим рецептором [1]. Мера, в которой вещество соответствует этим требованиям, есть функция его структуры и, соответственно, физико-химических свойств. Для выявления связи структура–активность применяются различные методы, использование которых основано на определенных допущениях. Наиболее важные среди них: существование объективной связи между строением вещества и его биологическим действием; возможность описать структуру с помощью определенного языка; возможность экстраполяции найденных закономерностей на свойства новых соединений. Используемые подходы условно классифицируются в зависимости от метода выявления закономерностей и способа описания структуры [2].

По способу описания структуры можно выделить две группы методов. В первую группу входят методы, использующие дискретные (качественные) признаки, во вторую – методы, использующие непрерывные (количественные) признаки. Последние в основном применяются в регрессионных методах анализа, а качественные – в статистических методах и методах распознавания образов. Однако не исключено использование как в одних и тех же методах признаков разных типов, так и при одинаковых типах признаков – разных методов выявления зависимости структура–активность.

Одним из современных методов, использующих непрерывные признаки, является CoMFA (Comparative Molecular Field Analysis) [3, 4]. Авторы метода исходили из принципов зависимости биологической активности от конформации молекулы и неизменности ковалентных связей при лиганд-рецепторном взаимодействии [4]. Основ-

Сокращения: АПО – алгоритм пространственного обучения; ИНС – искусственные нейронные сети; ККСА – количественное соотношение структура–активность; ЧНК – метод частичных наименьших квадратов; AChE – ацетилхолинэстераза.

#Автор для переписки (e-mail: vkov@bioorganic.kiev.ua; факс: 380-044-573-25-52; тел.: 380-044-573-25-95).

ное допущение метода состоит, как правило, в том, что молекулярная активность чувствительна к локализованным в пространстве отличиям в интенсивностях электростатического (кулоновское) и стерического (потенциал Леннарда–Джонса) полей [5]. В некоторых случаях используются также гидрофобные поля [6]. В методе CoMFA интенсивность полей измеряют в узлах трехмерной решетки как энергию взаимодействия молекулы с пробным атомом, обладающим заданным зарядом и стерическими свойствами. При этом анализ количественных признаков проводится с помощью метода частичных наименьших квадратов (ЧНК) [7].

Метод ЧНК представляет собой итерационную процедуру, на каждой стадии которой образуется новая латентная переменная по принципу максимизации подобия к зависимой переменной (активностям молекул). Таким образом, набор данных моделируется набором новых латентных переменных, которые в большей степени скоррелированы с зависимой переменной, нежели исходный набор [4]. Модификация метода ЧНК, используемая в CoMFA, позволяет учесть вклады анализируемых признаков и определить участки пространства вокруг молекул, наиболее важные для функционирования анализируемых соединений [5]. Прогнозирующая способность выявленной зависимости между активностью молекул и латентными переменными проверяется с помощью перекрестной оценки [7]. Результаты, полученные с помощью данного подхода, имеют наглядную геометрическую интерпретацию. Это возможно благодаря тому, что каждой переменной в CoMFA отвечает точка в окружающем молекулу пространстве. В результате строятся контурные карты, описывающие области пространства, наиболее важные для реализации определенного вида активности.

Следует отметить, что метод ЧНК дает хорошие результаты, когда взаимосвязь между активностью и признаками линейна. Известны работы, описывающие нелинейные варианты ЧНК (например, [8]), однако найденный в них тип нелинейной зависимости, зачастую искусственен и поэтому не всегда адекватно отражает взаимосвязь между структурой и активностью.

В качестве альтернативного подхода следует отметить группу алгоритмов, известных под общим названием искусственные нейронные сети (ИНС). Эти методы позволяют смоделировать взаимосвязи между структурой и активностью, учитывая их нелинейный характер. Поэтому в данной работе мы описываем применение метода ИНС вместо метода ЧНК для анализа CoMFA-модели.

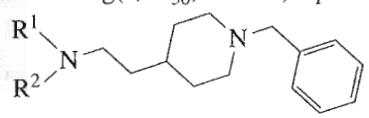
Нейронные сети, как и другие методы распознавания образов, различаются по типу использования обучающей информации: обучение с учителем и обучение без учителя. ИНС, использующие обучения без учителя, как правило, представлены большими сетями (сотни нейронов) и применяются в моделях ассоциативной памяти и восстановления изображения из шума. Успехи ИНС в этой области заслужено связаны с именем японского ученого Фукушимы и его моделями когнитрона и неокогнитрона [9, 10], а также с работами Хопфилда [11], Кохонена [12, 13] и др. Среди сетей, которые используют обучение с учителем, наибольшее распространение получили нейронные сети с прямым распространением сигналов. Среди алгоритмов обучения наиболее известен алгоритм с обратным распространением ошибки [14, 15]. Хотя ИНС очень часто используются для предсказания разного рода биологической активности молекул, сфера их применения часто ограничена анализом качественных признаков [16–18]. Это связано в первую очередь с тем, что скорость обучения нейросетей, имеющих большое число нейронов на входном слое, сравнительно невысокая, поэтому прямое использование ИНС для анализа молекул с большим числом признаков требует длительных вычислений. Во-вторых, конвертирование трехмерной матрицы в линейный вектор признаков приводит к потере информации о структуре данных. Наконец, поскольку каждый признак рассматривается как независимый, возможно появление корреляционных эффектов между признаками, что отрицательно сказывается на обучении ИНС.

Мы старались избежать вышеприведенных проблем, используя для анализа данных CoMFA-алгоритм, который, по сути, является комбинацией метода ИНС с обратным распространением ошибки и сетей Кохонена [12]. Основная идея метода заключается в том, чтобы вначале определить кластерные зоны в пространстве вокруг молекул, а затем использовать для обучения ИНС величины признаков, обобщенных для каждого из кластеров.

АНАЛИЗИРУЕМЫЕ ДАННЫЕ

Эффективность разработанного подхода была исследована на основе ряда молекул производных *N*-бензилпиперидинов, являющихся ингибиторами ацетилхолинэстеразы (AChE) [19]. Некоторые из AChE-ингибиторов положительно влияют на процесс восстановления функций памяти у людей, страдающих болезнью Альцгеймера [20]. По сравнению с другими классами AChE-ингибиторов производные *N*-бензилпиперидинов обладают более высокой активностью и меньшей токсичностью. Эти качества позволяют рассматривать их как наиболее вероятных кандидатов в препараты, перспективные для лечения болезни Альцгеймера.

Таблица 1. Структуры и величины ингибирующей активности $lg(1/IC_{50}$, $мкМ^{-1})$ производных *N*-бензилпиперидинов, использованные для построения CoMFA-модели



Номер	R^1	R^2	$lg(1/IC_{50})$	Номер	R^1	R^2	$lg(1/IC_{50})$
(1)	PhCO	H	0.25	(21)	PhCO	$PhCH_2$	0.03
(2)	*o*-$CH_3C_6H_4CO$	H	0.00	(22)	PhCO	Ph	1.46
(3)	*m*-$CH_3C_6H_4CO$	H	0.33	(23)	*p*-$(PhCH_2SO_2)$-C_6H_4CO	CH_3	3.22
(4)	*p*-$CH_3C_6H_4CO$	H	0.74	(24)	*p*-$(PhCH_2SO_2)$-C_6H_4CO	C_2H_5	3.52
(5)*	*o*-$NO_2C_6H_4CO$	H	0.06	(25)*	*p*-$(PhCH_2SO_2)$-C_6H_4CO	Ph	3.22
(6)	*m*-$NO_2C_6H_4CO$	H	0.64	(26)	*p*-$CH_3OC_6H_4CO$	Ph	0.23
(7)	*p*-$NO_2C_6H_4CO$	H	1.26	(27)	*p*-FC_6H_4CO	Ph	1.74
(8)	*p*-$CH_3OC_6H_4CO$	H	1.06	(28)	*p*-$NO_2C_6H_4CO$	Ph	2.27
(9)	*p*-$OHCC_6H_4CO$	H	0.92	(29)	*p*-PyCO	Ph	1.19
(10)*	*p*-ClC_6H_4CO	H	0.74	(30)*	$C_6H_{11}CO$	Ph	–0.97
(11)	*p*-FC_6H_4CO	H	1.07	(31)	CH_3CO	Ph	1.28
(12)	*p*-$CH_3COC_6H_4CO$	H	1.29	(32)	CH_3CH_2CO	Ph	0.08
(13)	*p*-$(PhCH_2SO_2)C_6H_4CO$	H	1.54	(33)	CH_3CO	*m*-$CH_3OC_6H_4$	1.34
(14)	*o*-PyCO	H	0.10	(34)	CH_3CO	*p*-$CH_3OC_6H_4$	0.15
(15)*	*m*-PyCO	H	1.16	(35)*	CH_3CO	*m*-FC_6H_4	1.19
(16)	*p*-PyCO	H	1.14	(36)	CH_3CO	*p*-FC_6H_4	0.69
(17)	$C_6H_{11}CO$	H	–0.20	(37)	CH_3CH_2	Ph	–1.08
(18)	$PhCH_2$	H	–1.66	(38)	CH_3CO	*p*-Py	0.97
(19)	PhCO	CH_3	0.77	(39)	CH_3CO	CH_3	0.18
(20)*	PhCO	C_2H_5	0.89				

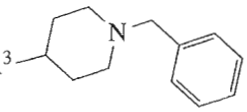
* Соединения, вошедшие в тестовый набор данных. Py – остаток пиридина C_5H_4N.

Молекулярные структуры *N*-бензилпиперидинов представлены в табл. 1, 2. 13 соединений (каждое пятое) не включали в первоначальный набор обучения – из них был сформирован тестовый набор, использованный в дальнейшем для оценки качества полученной модели [21]. Ингибирующее действие соединений оценивалось величиной $lg(1/IC_{50})$.

Первый шаг CoMFA-процедуры заключается в получении активной конформации молекулы лиганда и ее пространственном ориентировании в полости рецептора или фермента ("докинг") в соответствии с фармакофорной моделью или кристаллографическими данными. В данной работе, поскольку пространственная структура фермента была известна, применялось ориентирование лиганда на основе "докинга".

Наборы координат атомов двух белков AChE доступны в Брукхевенском банке данных белковых структур: первый из ската *Torpedo californica* (TAChE, PDB-код 1ACE) и второй из мыши (MAChE, PDB-код 1MAH). Так как ингибиторная активность используемых в настоящей работе соединений измерялась для MAChE, то "докинг" проводился с этим ферментом. К сожалению, даже наличие сведений о кристаллической структуре фермента не устраняет полностью неоднозначность при выборе конформации связывающихся с ним лигандов. Недостающая информация может быть получена, например, при исследовании кристаллической структуры лиганд-ферментных комплексов. В литературе не были описаны комплексы лиганд–MAChE, но имелись структурные данные для комплексов лиганд–TAChE. В результате исследования кристаллической структуры комплексов TAChE с ингибиторами этого фермента, а также сравнения обоих ферментов, было показано [19, 21], что четвертичный пиперидиновый атом азота *N*-бензилпиперидинового кольца может быть использован как якорь для процедуры автоматического "докинга" лигандов с ферментами.

Каждый из 66 лигандов был смоделирован с помощью программы Sybyl 6.3 (Tripos Corp., США) на рабочей станции Silicon Graphics INDY R5000. Начальные конформации были оптимизированы с помощью алгоритма молекулярной механики с

Таблица 2. Структуры и величины ингибирующей активности $lg(1/IC_{50}$, мкМ$^{-1})$ для производных *N*-бензилпиперидинов, использованные для построения CoMFA-модели 

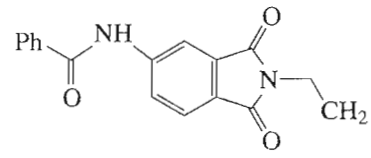
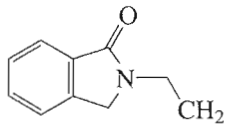
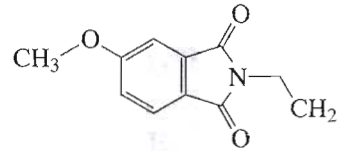
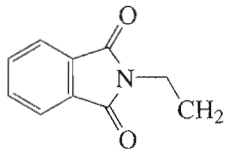
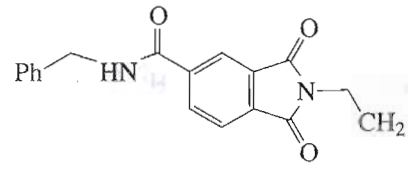
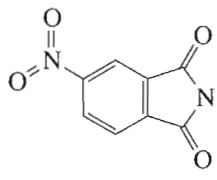
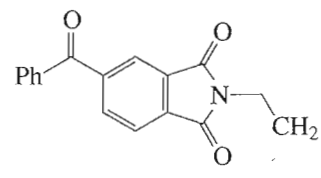
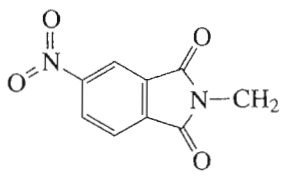
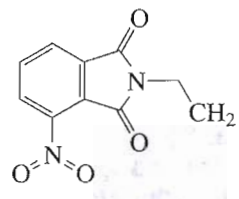
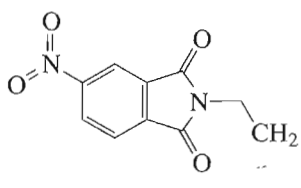
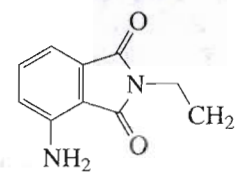
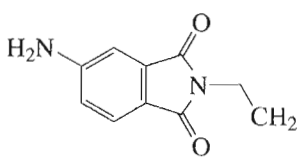
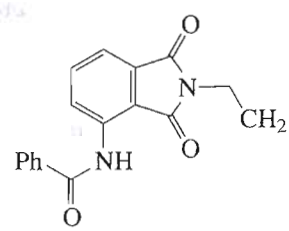
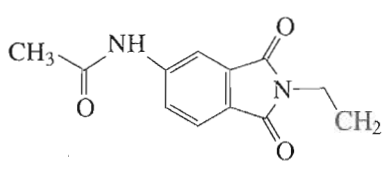
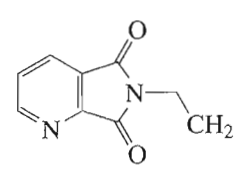
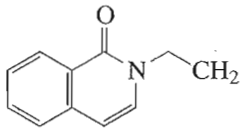
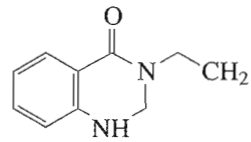
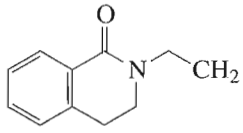
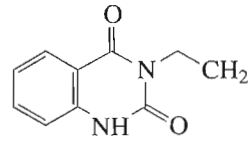
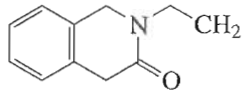
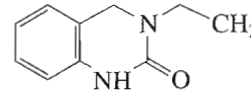
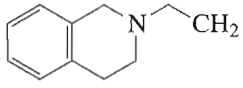
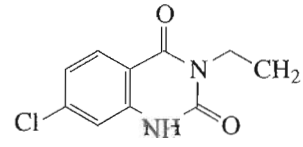
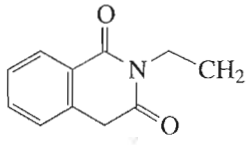
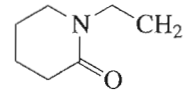
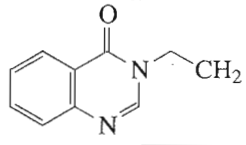
Соединение	R^3	$lg(1/IC_{50})$	Соединение	R^3	$lg(1/IC_{50})$
(40)*	$PhCO(CH_2)_3$	0.28	(48)		2.92
(41)		1.01	(49)		2.10
(42)		1.52	(50)*		2.66
(43)		−1.43	(51)		2.62
(44)		−0.48	(52)		2.05
(45)*		1.90	(53)		1.96
(46)		2.06	(54)		0.47
(47)		2.55	(55)*		1.89

Таблица 2. Окончание

Соединение	R^3	$lg(1/IC_{50})$	Соединение	R^3	$lg(1/IC_{50})$
(56)		−0.04	(62)		0.10
(57)		0.00	(63)		2.38
(58)		1.77	(64)		1.89
(59)		−0.20	(65)*		2.35
(60)*		1.64	(66)		0.57
(61)		−0.08			

* Соединения, вошедшие в тестовый набор данных.

использованием силового поля Tripos [22]. С помощью опции SYBYL/SEARCH были найдены конформации с наименьшей энергией, использованные затем как начальные в процедуре "докинга". Так как данные лиганды взаимодействуют с ферментом при pH 8, то все соединения рассматривали в протонированной форме. При "докинге" каждый из лигандов изначально ориентировали в месте связывания фермента так, что четвертичный пиперидиновый атом азота использовался как якорь, т.е. фиксировался. Затем молекулу вращали вокруг трех осей координат с шагом 30°. В каждом положении проводили конформационный поиск с помощью опции SYBYL/SYSTEMATIC SEARCH, который заключался в том, что для каждой из конформаций, полученных вращением вокруг всех подвижных связей с шагом 30°, рассчитывали энергию комплекса лиганд–фермент. Конформации с наименьшей энергией анализировали более детально путем вращения вокруг всех подвижных связей с шагом 5° в интервале ±30°. Конформации молекул с минимальными величинами энергии комплекса использовались в CoMFA-анализе.

Затем каждую молекулу лиганда с фиксированной пространственной структурой помещали в пространственную решетку, составленную из кубических ячеек величиной 0.125 Å^3. Частичные атомные заряды для вычисления электростатического поля были получены на основе метода MOPAC AM1 [23]. CoMFA-область выбиралась так, чтобы в ней помещалась любая молекула из набора с дополнительным запасом в 2 Å. Все вычисления проводили в вакууме, константу диэлектрической проницаемости брали равной 1. Между каждой молекулой и атомом углерода ("щупом") с зарядом +1 вычисляли электростатическую и стерическую энергию взаимодействия в каждом узле решетки [19]. Если энергия стерического или электростатического взаимодействия превышала +/– 30 ккал/моль, она была ограничена этим пороговым значением. Электростатические взаимодействия между двумя атомами вычисляли по формуле:

$$E_{el} = q_i q_j / 4\pi\varepsilon_0\varepsilon_r r_{ij},$$

где q_i и q_j – заряды атомов, r_{ij} – расстояние между ними, ε_0 – диэлектрическая проницаемость вакуума, ε_r – диэлектрическая константа окружающей

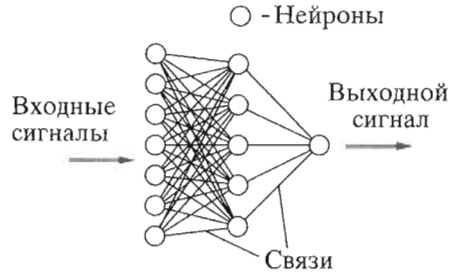


Рис. 1. Топология искусственной нейронной сети с прямым распространением сигналов и одним скрытым слоем.

среды. Для расчета стерических взаимодействий использовалось уравнение Леннарда–Джонса:

$$E_{vdw} = \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{1}{a_{ij}^{12}} - \frac{2}{a_{ij}^{6}}\right),$$

где $E_{ij} = (E_i E_j)^{1/2}$ константа Ван-дер-Ваальса, $a_{ij} = r_{ij}/(R_i + R_j)$, R_i – ван-дер-ваальсовый радиус i-го атома молекулы [22], N – число атомов. В результате расчетов были сформированы два набора данных, состоящих из 14112 признаков. Каждый признак соответствовал энергии стерического или электростатического взаимодействия в данном узле решетки, измеренной в ккал/моль.

Использованный подход, обоснование процедуры "докинга", а также детали расчетов более полно описаны в работе [19]. Следует отметить, что цель данной работы – исследование возможности замены метода ЧНК, который применялся в CoMFA-процедуре в предыдущих исследованиях [19, 21], методом ИНС с использованием той же модели и тех же наборов признаков. Поэтому никаких дополнительных по сравнению с работами [19, 21] исследований с учетом диэлектрических свойств системы молекула–растворитель, величины атомных зарядов, конформации молекул и т.п. мы не проводили. Такой анализ, безусловно, представляет большой интерес, но заслуживает отдельного исследования, которое выходит за рамки данной работы, впервые показывающей возможность применения ИНС вместо метода ЧНК для решения задач 3D-QSAR.

НЕЙРОННЫЕ СЕТИ

Простейшая составляющая каждой ИНС – искусственный нейрон, отражающий свойства биологического аналога: суммирование и нелинейное преобразование сигналов, а также возможность адаптации своих связей (рис. 1). На вход нейрона поступает вектор входных воздействий X, каждая компонента которого является выходом другого нейрона. Входное возбуждение нейрона определяется как взвешенная сумма его входов, где W – вектор весов связей. Выходной сигнал нейрона определяется через активационную функцию от входного возбуждения. Структура связей и вид активационной функции определяются в зависимости от используемого метода нейронных сетей. Входная информация поступает на первый слой, а выходную информацию (целевой вектор) получают на слое N. Для задач поиска связей структура–активность входная информация является вектором параметров молекулы (каждый нейрон во входном слое соответствует одному параметру молекулы), а выходная – есть вектор активностей молекулы (каждый нейрон в выходном слое соответствует одному типу активности молекулы). Входной и целевой вектор составляют обучающую пару. При прохождении через скрытые слои (как правило, используется один такой слой) входной поток информации нелинейно трансформируется. Обучение сети состоит в таком настраивании связей сети, когда каждое входное возбуждение (вектор параметров молекулы) приводит к появлению на выходе нейронной сети желаемого целевого вектора (вектора активностей молекул).

Математически обучение нейронной сети заключается в минимизации функции ошибок:

$$F = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - O_i)^2}{N \times M}},$$

где O_i – расчетный и Y_i – целевой векторы активностей молекулы, N – число соединений, M – число выходных нейронов. Процедура обучения проводится для всех входных данных до тех пор, пока значение F не станет достаточно малым (как правило, <0.01).

Для избежания проблем переобучения ИНС [18] был применен метод "раннего останова" ансамбля нейронных сетей [18, 24, 25], который позволяет существенно улучшить прогнозирующую способность сетей. Кратко остановимся на его работе. Каждый ансамбль ИНС формировали из 200 сетей. Из усредненного по всем сетям значения активности вычисляли статистические коэффициенты. Качество полученных моделей оценивали с помощью метода скользящего контроля: каждую молекулу последовательно удаляли из обучающей выборки и затем использовали для оценки качества модели. Прогнозирующую способность сетей определяли с помощью коэффициента перекрестной оценки q^2, предложенного Крамером и др. [5]. Исходные данные разделяли на два набора – первый использовали для обучения ИНС, а второй – для контроля процесса обучения. В процессе обучения сети определяли точку "раннего останова" [18], которая соответствовала минимуму функции ошибки сети F для контрольного набора. Процесс обучения ограничивался 10000 итерациями.

СЕТЬ КОХОНЕНА

Основное назначение сети Кохонена – создание нелинейной проекции данных большой размерности на область малой размерности. Мы использовали сеть, состоящую из двух слоев нейронов, образующих прямоугольную решетку на плоскости (рис. 2). Входные сигналы, X_s

$$X_s = (x_{s1}, x_{s2}, \ldots, x_{sm}) \quad s = 1, 2, \ldots, p,$$

где p – число параметров, m – число молекул исходной выборки данных, представляют собой вектора действительных чисел, которые последовательно предъявляют сети. В качестве X_s использовали энергии взаимодействия в узлах CoMFA-решетки, а также веса обученных ИНС.

Каждому нейрону j соответствует вектор весовых элементов

$$W_j = (w_{j1}, w_{j2}, \ldots, w_{jm}) \quad j = 1, 2, \ldots, n,$$

где n соответствует числу нейронов в сети Кохонена. Для необученной сети, веса нейронов инициализируют случайным образом.

На каждом шаге обучения входной вектор воздействий (X_s) сравнивают со всеми нейронами сети путем расчета евклидового расстояния между X_s и всеми n весовыми векторами W_j:

$$d_{si} = \sqrt{\sum_i^m (x_{si} - w_{ji})} \quad j = 1, 2, \ldots, n.$$

Нейрон, чей вектор весов W имеет наименьшее расстояние d_{sj}, считается "выигравшим" нейроном. В процессе обучения веса "выигравшего" нейрона изменяют таким образом, чтобы они стали близки входному воздействию X. Такой процесс называется самоорганизацией, и, следовательно, сеть Кохонена называется самоорганизующейся сетью. При каждой итерации процесса обучения сети все входные воздействия по очереди представляют каждому нейрону сети, используя их для настройки весовых коэффициентов сети. Процесс адаптации весов производят согласно уравнению:

$$W_j^* = W_j + \alpha(t)h_{cj}(t)(X - W_j),$$

где W_j^* – новое значение весов для нейрона j, $\alpha(t)$ – функция, отображающая скорость обучения сети, $h_{cj}(t)$ – функция, характеризующая меру близости весовых параметров нейронов. Обычно $h_{cj}(t) = h(\|r_c - r_j\|, t)$, где r_c и r_j координаты нейронов c и j. Как $\alpha(t)$, так и $h_{cj}(t)$ представляют собой функции, монотонно убывающие во времени.

После окончания процесса обучения весь набор данных вновь представляется по очереди каждому нейрону и для каждой молекулы определяется "выигравший" нейрон. Эти нейроны соответствуют проекции исходных данных на двумерную карту Кохонена. Таким образом, в результате процесса самоорганизации сети формируется карта, на которой вектора входных воздействий располагаются согласно их близости в многомерном пространстве. Качество обучения нейронной сети оценивается как сумма расстояний между "выигравшими" нейронами и исходными данными.

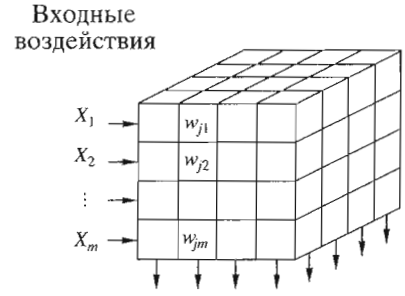


Рис. 2. Архитектура сети Кохонена, состоящая из 16 нейронов. Входные воздействия – вектор X, состоящий из m входов, поступает на вход сети. Каждый нейрон сети Кохонена представлен столбцом с m весовыми коэффициентами w_{ij}.

АЛГОРИТМ ПРОСТРАНСТВЕННОГО ОБУЧЕНИЯ ИНС

Алгоритм пространственного обучения (АПО) реализует циклическую повторяющуюся итерационную процедуру, комбинирующую поочередное применение сети Кохонена и искусственных нейронных сетей. Анализ данных проводили согласно следующему АПО, блок-схема которого представлена на рис. 3:

1) формирование исходной выборки данных (блок 1);

2) инициализация параметров сети Кохонена (блок 2);

3) подача исходных данных на вход сети Кохонена. Кластеризация пространства входных признаков (блок 3);

4) сжатие данных, замена исходной выборки значениями кластерных центров;

5) подача сжатой выборки данных на вход мини-ансамбля ИНС (блок 6). Расчет средней ошибки прогноза F для всего мини-ансамбля сетей;

6) сравнение значения ошибки на текущем этапе E_{cur} с ошибкой предыдущего этапа обучения F_{prev}. Если $F_{cur} < F_{prev}$, то текущее кластерное распределение сохраняется. Весовая матрица нейросети сохраняется только при уменьшении текущей ошибки F_{cur} в точке "раннего останова" [18];

7) формирование таблицы весовых коэффициентов (блок 5);

8) изменение размера карты Кохонена (блок 4);

9) повторение шагов 2–8 до тех пор, пока размер карты для сети Кохонена не уменьшится до

Рис. 3. Общая схема анализа данных с помощью алгоритма пространственного обучения. 1 – выборка исходных экспериментальных данных; 2, 3 – сеть Кохонена; 4 – блок контроля и изменения карт Кохонена; 5 – таблица весовых коэффициентов; 6 – блок тестирования модели (кластерного разбиения); 7 – блок выбора наиболее оптимальной модели; 8 – блок расчета на основе ИНС; 9 – блок прогноза новых соединений.

минимально допустимого – 8 узлов (4 × 2, 4 узла в 2 рядах). После первого цикла сжатия в качестве входных данных используют сохраненные весовые коэффициенты нейронных сетей;

10) выбор наилучшего кластерного распределения на основе минимального среднего значения F (блок 7);

11) сжатие данных аналогично шагу 4;

12) обучение ансамбля ИНС (блок 8);

13) прогнозирование активности новых соединений (блок 9).

При формировании выборки данных вектор входных воздействий X для сети Кохонена имел размерность, равную числу молекул исходной выборки данных. Количество входных воздействий было равно числу исследуемых признаков. Оптимальный размер карты Кохонена для первой итерации был найден путем последовательного тестирования карт различных размеров от 2 × 2 до 100 × 100 для исходных данных. Дальнейший анализ показал, что для качественного анализа этих данных, достаточно использовать карты небольшого размера. Использование карт меньше 154 узлов ухудшало результаты, поэтому первоначальный размер карты Кохонена принимали равным 154 узлам (14 × 11). При каждом последующем сжатии размер карт уменьшался на две единицы как для числа узлов, так и для числа рядов. Обучение сети, как рекомендовано Кохоненом [12], было разделено на две фазы. Количество итераций составляло 20000 и 50000 соответственно для первой и второй фаз обучения. В первой фазе начальный радиус области обучения брали равным 5, а α – равным 0.7. Для второй фазы обучения стартовый радиус равнялся 2 и α – 0.02. Эти значения параметров были определены путем перебора с использованием в качестве граничных значения, рекомендованные Кохоненом [12]. При этом количество итераций изменяли от 100000 до 10000 с шагом в 10000 итераций, начальный радиус варьировали от 7 (радиус начальной карты Кохонена) до 1 с единичным шагом, а α – от 1 до 0.01 с шагом 0.01. Значения параметров, для которых с помощью сети Кохонена была получена наилучшая кластеризация входных данных, использованы в данной работе.

Для расчета более качественной матрицы входных весов на шаге 2–3 вместо сети Кохонена применяли алгоритм быстрой кластеризации, позволяющий незначительно сжимать исходный набор данных. Суть метода заключается в том, что если уровень корреляции между двумя признаками больше порогового, то их объединяют в один кластер. Использование данного подхода вместо сети Кохонена на первом цикле обучения существенно улучшало процедуру кластерного разбиения. Пороговый уровень корреляции принимали равным 0.99.

При формировании сжатой выборки данных (шаг 4), использовали четыре альтернативных варианта выбора значения данных в кластерных центрах:

1) усредненное значение всех признаков, входящих в кластер, рассчитанное по формуле:

$$C_n = \Sigma X_{ni}/m,$$

где X_{ni} – значение i-го признака для n-ой молекулы, m – количество признаков, входящих в кластер;

2) значение признака из данного кластера, максимально скоррелированного с усредненным значением всех признаков;

3) значение признака, евклидово расстояние между которым и усредненным значением всех признаков было минимально;

4) значение признака, ближайшего к геометрическому центру кластера.

Наилучший результат был получен на основе первого варианта, несколько хуже – на основе второго и намного хуже на основе третьего и четвертого вариантов. Поэтому все представленные результаты рассчитывали, применяя первый подход.

Для тестирования сжатых данных использовали нейронные сети, состоящие из трех слоев. Количество нейронов на входном слое соответство-

Таблица 3. Суммарный статистический анализ обучающего (53 соединения) и тестового (13 соединений) наборов *N*-бензилпиперидинов

Тип поля	Алгоритм пространственного обучения				Метод частичных наименьших квадратов		
	Число		Коэффициент q^2				Число латентных переменных
	признаков	кластеров	МСК*	Тестовый набор	МСК*	Тестовый набор	
Стерическое (С)	14112	10	0.58 ± 0.02	0.76 ± 0.02	0.42	0.66	9
Электростатическое (Э)	14112	35	0.65 ± 0.02	0.64 ± 0.02	0.53	0.59	9
С + Э	28224	45	0.58 ± 0.03	0.75 ± 0.02	0.59	0.69	9

* Метод скользящего контроля.

вало количеству кластеров, полученных на выходе сети Кохонена. На скрытом слое использовали пять нейронов. Третий слой состоял из одного нейрона, который соответствовал прогнозируемой активности молекул.

Для первоначального тестирования кластерного разбиения использовали мини-ансамбль, включающий 50 сетей. В процессе обучения мини-ансамбля ИНС формировали весовые матрицы входного слоя нейросети, отражающие информативность центров кластеров. На основе этих матриц определяли весовые матрицы входных признаков. При этом весовые коэффициенты входов, принадлежащие одному кластеру, принимали равными и определяли по формуле:

$$W_{ki} = W_k/m,$$

где k – номер кластера, i – номер признака, m – количество признаков, входящих в кластер. Эти матрицы формировали таблицу весовых коэффициентов, частично меняющуюся на каждом цикле обучения. Данную таблицу затем использовали в качестве входной информации для сети Кохонена вместо таблицы исходных данных. На заключительном этапе (блок 8) для анализа данных использовали 200 независимых ИНС.

Для определения наиболее информативных кластеров, отвечающих за проявление биологической активности, были использованы методы отбора признаков, известные в литературе как "pruning methods" [26]. По этим методам чувствительность нейронов рассчитывают исходя из амплитуды их весов или исходя из изменения ошибки нейронных сетей без учета весов. Данные методы хорошо зарекомендовали себя при анализе количественных признаков [26, 27].

РЕЗУЛЬТАТЫ

Для лучшего понимания факторов, лежащих в основе активности *N*-бензилпиперидинов, были исследованы три различных набора данных. Первый набор состоял только из стерических параметров, второй – из электростатических, третий содержал оба типа данных. В процессе анализа исходное пространство признаков было разделено на 10 кластеров по стерическим признакам и на 35 кластеров по электростатическим признакам (табл. 3). Как видно из таблицы, результаты, полученные на основе АПО для первых двух наборов, лучше, чем результаты, достигнутые с помощью ЧНК. Результаты, полученные обоими методами для суммарного набора, практически совпадают. В табл. 4 приведены результаты распределения признаков по кластерам. Большинство признаков обоих наборов было сосредоточено в одном кластере, тогда как остальные кластеры были сравнительно небольшими.

Второй этап анализа состоял из анализа полученного кластерного разбиения данных с помощью методов оптимизации признаков. Полученные результаты приведены в табл. 5. В процессе отбора большинство кластеров были оценены как малоинформативные для прогноза активности исследуемых соединений и удалены из дальнейшего анализа. Для стерических данных было отобрано 5 кластеров из 10, для электростатических – 8 кластеров из 35, а для суммарного набора данных – всего 6 наиболее информативных кластеров. Сравнивая величину q^2 в табл. 3 и 5, следует отметить значительное улучшение q^2 для обучаемой выборки и практически те же результаты для тестовой выборки. Интересен тот факт, что в суммарный набор данных было отобрано всего по три кластера по каждому из типов признаков. Однако размеры кластеров для электростатических параметров значительно превышают размеры кластеров для стерических параметров, что свидетельствует о большем влиянии признаков данного типа на проявление активности. Данные результаты вполне согласуются и с величинами полученных статистических коэффициентов – величина q^2 для электростатических параметров несколько выше, чем для стерических параметров (см. табл. 3, 5).

Данный вывод также подтверждается и распределением кластеров в пространстве. На рис. 4 приведены стерические и электростатические

Таблица 4. Распределение признаков по кластерам

Тип поля	Число признаков в кластере									
	1	2	3	4	5	6	7	8	9	10
Стерическое	8*	13*,#	5	5	8	14048	6*,#	7*	6*,#	6
Электростатическое	69*,#	33*,#	6	9	8	5	5	8	11	8
	6	8	5	6*	9*	7	6	6	5	8
	5	6	9	7	5*	8	8	13	6	8*
	13*	12	5	13695	84*,#					

* Кластеры, отобранные с помощью методов оптимизации параметров отдельно по каждому типу признака; #кластеры, отобранные с помощью методов оптимизации параметров для суммарного набора признаков.

Таблица 5. Величина коэффициента перекрестной оценки q^2 для наборов кластеров, найденных с помощью методов оптимизации параметров

Тип поля	Исходное количество кластеров	Оптимальное число кластеров	МСК	Тестовый набор
Стерическое (С)	10	5	0.62 ± 0.02	0.75 ± 0.02
Электростатическое (Э)	35	8	0.77 ± 0.02	0.60 ± 0.02
С + Э	45	6	0.73 ± 0.02	0.77 ± 0.02

контурные карты молекул, отображающие участки, способствующие (светло-серый цвет) и препятствующие (черный) связыванию и увеличению (черный) и уменьшению (темно-серый) отрицательного заряда. Полученные результаты согласуются с результатами предыдущих исследований для производных *N*-бензилпиперидинов. Например, светло-серый цвет в районе аминобензильной группы фталимидного заместителя *N*-бензилпиперидина (48) свидетельствует об активности данного участка при связывании с рецептором. Наличие черной области возле атома азота фталимидного заместителя свидетельствует о стерически запрещенной области пространства (рис. 4*а*). Черная область, огибающая карбонильный кислород фталимидной части молекулы (48), подчеркивает наличие отрицательно заряженной ингибирующей группировки (рис. 4*б*). Локализация найденных в настоящей работе функционально значимых областей в большинстве случаев совпадает с описанной в работе [19], однако наблюдаются и значительные отличия в их размерах, форме и расположении вокруг молекулы. Например, в работе [19] стерически важные области пространства намного больше, чем найденные с помощью АПО, что свидетельствует о более точной локализации нашим методом стерических зон вокруг молекул, определяющих их способность связываться с рецептором.

В итоге проделанной работы можно сделать следующие выводы.

1) Разработан новый перспективный метод (АПО) для исследования количественной связи структура–активность физиологически активных веществ.

2) Использование предложенного метода улучшает статистические параметры модели по сравнению с результатами ЧНК для тех же наборов данных.

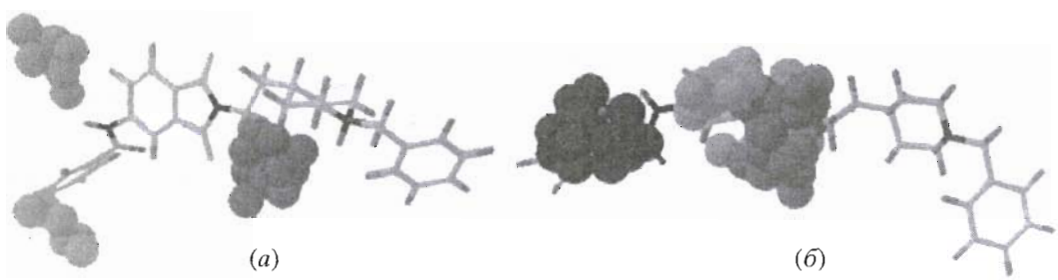


Рис. 4. Карта распределения кластеров вокруг молекулы соединения (48) (табл. 2) для суммарного набора данных: *а* – стерический тип признаков, *б* – электростатический тип признаков. Светло-серым цветом обозначены области, стерические или электростатические свойства которых способствуют проявлению активности, а черным и темно-серым цветом – области, стерические или электростатические свойства которых препятствуют проявлению активности. Интенсивность оттенка показывает степень активности кластера.

3) Число исходных признаков может быть уменьшено заменой их центрами кластеров, которые используются в качестве входных признаков для ИНС, а также для визуализации и интерпретации полученных результатов.

Работа выполнена в рамках проектов INTAS-Украина № 95-0060 и № 96-1115.

СПИСОК ЛИТЕРАТУРЫ

1. *Hanch C., Fugita T.* // J. Amer. Chem. Soc. 1964. V. 86. P. 1616–1626.
2. *Тюрина Л.А., Кадыров Ч.Ш., Симонов В.Д.* // Итоги науки и техники. ВИНИТИ. Сер. Органическая химия. 1989. Т. 18. С. 1–156.
3. *Cramer R.D., DePriest S.A., Patterson D.E., Hecht P.* // 3D QSAR in Drug Design / Ed. H. Kubinyi. Leiden: ESCOM, 1993. P. 443–485.
4. *Cramer R.D., Bunce J.D.* // QSAR in Drug Design and Toxicology / Eds D. Handzi, B. Jerman-Blazic. Amsterdam: Elsevier, 1987. P. 3–12.
5. *Cramer R.D., Patterson D.E., Bunce J.D.* // J. Am. Chem. Soc. 1988. V. 110. P. 5959–5967.
6. *Barreca M.L., Carotti A., Carrieri A., Chimirri A., Monforte A.M., Calace M.P., Rao A.* // Bioorg. Med. Chem. 1999. V. 7. P. 2283–2292.
7. *Wold S., Johansson E., Cocci M.* // 3D QSAR in Drug Design / Ed. H. Kubinyi. Leiden: ESCOM, 1993. P. 523–563.
8. *Wold S.* // Chemometr. Intell. Lab. Syst. 1992. V. 14. P. 71–84.
9. *Fukushima K.* // Biol. Cybernetics. 1975. V. 20. P. 121–136.
10. *Fukuchima K.* // Biol. Cybernetic. 1980. V. 36. P. 193–202.
11. *Hopfield J.J.* // Proc. Natl. Acad. Sci. USA. 1982. V. 79. P. 2554–2558.
12. *Kohonen T.* Self-organisation Maps. Berlin: Springer-Verlag, 1995.
13. *Simon V., Gasteiger J., Zupan J.A.* // J. Am. Chem. Soc. 1993. V. 115. P. 9148–9159.
14. *Rumelhart D.E., McClelland J.L.* Parallel Distributed Processing Extrapolation in Microstructure of Cognition. Cambridge: the MIT Press, 1986.
15. *Tetko I.V., Luik A.I., Poda G.I.* // J. Med. Chem. 1993. V. 36. P. 811–814.
16. *Maddalena D.* // Exp. Opin. Ther. Patents. 1996. V. 6. P. 239–251.
17. *Devillers J.* Neural Networks in QSAR and Drug Design. London: Academic Press, 1996.
18. *Tetko I.V., Livingstone D.J., Luik A.I.* // J. Chem. Inf. Comput. Sci. 1995. V. 35. P. 826–833.
19. *Bernard P., Kireev D.B., Chretien J.R., Fortier P.-L., Coppet L.* // J. Comp. Aid. Mol. Design. 1999. V. 13. P. 355–371.
20. *Perry E.K.* // Br. Med. Bull. 1986. V. 42. P. 408.
21. *Golbraikh A., Bernard P., Chretien J.R.* // Eur. J. Med. Chem. 2000. V. 35. P. 1–14.
22. *Clarc M., Cramer R.D. III, Opdenbosch N.V.* // J. Comput. Chem. 1989. V. 10. P. 982–1012.
23. *Dewar M.J.S., Zoebisch E.G., Healy E.F., Stewart J.J.P.* // J. Amer. Chem. Soc. 1985. V. 107. P. 3902–3909.
24. *Tetko I.C., Villa A.E.P.* // Neural Processing Letters. 1997. V. 6. P. 43–50.
25. *Tetko I.V., Villa A.E.P.* // Neural Networks. 1997. V. 10. P. 1361–1374.
26. *Tetko I.V., Villa A.E.P., Livingstone D.J.* // J. Chem. Inf. Comput. Sci. 1996. V. 36. P. 794–803.
27. *Kovalishyn V.V., Tetko I.V., Luik A.I., Kholodovych V.V., Villa A.E.P., Livingstone D.J.* // J. Chem. Inf. Comput. Sci. 1998. V. 38. P. 651–659.

Application of Neural Networks Using the Volume Learning Algorithm for Quantitative Study of the Three-dimensional Structure–Activity Relationships of Chemical Compounds

V. V. Kovalishim*, I. V. Tetko*, A. I. Luik*, J. R. Chretien, and D. J. Livingstone*****

e-mail: vkov@bioorganic.kiev.ua

**Institute of Bioorganic and Petroleum Chemistry, National Academy of Sciences of Ukraine, ul. Murmanskaya 1, Kiev, 02094 Ukraine*

***Laboratory of Chemometrics and Bioinformatics, University of Orleans, BP 6759, Orleans Cedex 2, 45067 France*

****ChemQuest, Delamere House 1, Royal Crescent, Sandown, Isle of Wight, PO36 8LZ and Centre for Molecular Design, University of Portsmouth, Hants, PO1 2EG UK*

A volume learning algorithm for artificial neural networks was developed to quantitatively describe the three-dimensional structure–activity relationships using as an example *N*-benzylpiperidine derivatives. The new algorithm combines two types of neural networks, the Kohonen and the feed-forward artificial neural networks, which are used to analyze the input grid data generated by the comparative molecular field approach. Selection of the most informative parameters using the algorithm helped to reveal the most important spatial properties of the molecules, which affect their biological activities. Cluster regions determined using the new algorithm adequately predicted the activity of molecules from a control data set. The English version of the paper: *Russian Journal of Bioorganic Chemistry*, 2001, vol. 27, no. 4; see also http://www.maik.ru.

Key words: artificial neural networks, N-benzylpiperidines, clusters, structure–activity relationships, volume learning algorithm