«Вестник МИТХТ», 2008, т. 3, № 4

Математические методы и информационные технологии в химии и химической технологии

УДК 681.3

ПРОГНОЗИРОВАНИЕ АКТИВНОСТИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Е.В. Бурляева, профессор, П.А. Ушаков, аспирант
кафедра Информационных технологий МИТХТ им. М.В. Ломоносова

Выполнен анализ применимости нейронных сетей для прогнозирования активности производных дитиокарбаминовой кислоты и производных TIBO. Предложен алгоритм регуляризации нейронных сетей, позволяющий улучшить прогностические возможности сетей

Ключевые слова: *нейронные сети, компьютерная химия, производные дитиокарбаминовой кислоты, TIBO, структура и свойства органических соединений*

Одной из важнейших задач компьютерной химии является предсказание физических, химических и биологических свойств химических соединений. Такое прогнозирование позволяет проводить дорогостоящие экспериментальные исследования более прицельно и оценивать возможность использования соединения в качестве основы для создания лекарственного препарата на ранних стадиях его изучения [1]. В основе исследований лежит предположение о том, что структура соединения определяет свойства, проявляемые этим соединением. Построение гипотез о взаимосвязи между структурой и свойствами исследуемых соединений (зависимостей «структура – активность) сводится к выявлению общей закономерности на основе ряда примеров ее проявления. Эта задача относится к классу задач индуктивного вывода. Как правило, гипотеза описывается в виде формулы, в которую необходимо подставить параметры структуры молекулы для того, чтобы получить оценку ее активности. Основным результатом прогнозирования являются оценки значений свойств соединений тестовой выборки.

В последнее время для построения гипотез о зависимостях «структура – активность» стали активно применяться подходы, основанные на методах искусственного интеллекта, в том числе методы, основанные на использовании искусственных нейронных сетей (ИНС), позволяющих формировать гибкие нелинейные модели зависимостей «структура-свойство».

Целью исследований является анализ применимости искусственных нейронных сетей для прогнозирования активности органических соединений. Анализ выполнялся на 2-х групп соединений: производные дитиокарбаминовой кислоты, синтезированные и экспериментально изученные в отделе Медицинской химии Государственного научного центра по антибиотикам (ГНЦА) – 32 соединения, и производные тетрагидроимидазобензодиазипенона (TIBO), взятые из нескольких литературных источников [2–5] – 49 соединений. Соединения первой группы рассматриваются как перспективная основа для противотуберкулезных препаратов, для них в качестве активности определялась способность подавлять рост микобактерий туберкулеза на твердой среде. Общая структура соединений представлена на рис. 1, расшифровка заместителей и активность – в табл. 1.

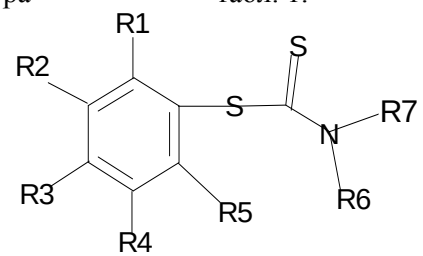

Рис.1. Структура производных дитиокарбаминовой кислоты.

Соединения второй группы используются в комплексных анти-ВИЧ препаратах, в качестве их активности рассматривались значения отрицательного логарифма 50% ингибирующей концентрации (IC_{50}). Общая структура соединений представлена на рис. 2, расшифровка заместителей и активность – в табл. 2. Для обеих групп соединений активность была задана количественно. Особенностью набора соединений, предоставленных ГНЦА,

является наличие большого количества полностью неактивных соединений (24 из 32), что существенно усложняет построение количественных зависимостей «структура-активность» [6].

Обучающую выборку для построения ИНС составляют векторы числовых значений параметров, описывающих структуры молекул (такие параметры называют молекулярными дескрипторами), и числовые значения изучаемых свойств. Для описания молекул были использованы квантовохимические параметры, рассчитанные с помощью программы MOPAC 7.0. После оптимизации всех построенных конформаций в качестве структурного представителя молекулы выбирался конформер с наименьшей теплотой образования (наиболее энергетически устойчивый).

Таблица 1. Структура и активность производных дитиокарбаминовой кислоты.

Соед.	R^1	R^2	R^3	R^4	R^5	R^6	R^7	Акт.
I	NO_2	H	H	H	H	CH_2CH_3	CH_2CH_3	18
II	NO_2	H	NO_2	H	H	CH_2CH_3	CH_2CH_3	25
III	NO_2	H	F_3	H	H	CH_2CH_3	CH_2CH_3	19
IV	NO_2	H	F_3	Cl	H	CH_3	CH_3	0
V	NO_2	H	F_3	H	NO_2	CH_3	CH_3	0
VI	NO_2	H	NO_2	Cl	H	CH_3	H	0
VII	N	H	NO_2	H	H	CH_3	H	0
VIII	NO_2	H	Cl	NO_2	H	CH_3	H	0
IX	NO_2	H	COOH		H	CH_3	H	0
X	NO_2	H	COOH	NO_2	H	CH_3	H	0
XI	NO_2	H	H	NO_2	H	CH_3	H	0
XII	NO_2	NO_2	Cl	H	H	CH_3	H	0
XIII	NO_2	Cl	H	H	H	CH_3	H	0
XIV	NO_2	H	CNH_2O	H	H	CH_3	H	0
XV	NO_2	H	CNH_2O	H	H	CH_2CH_3	CH_2CH_3	0
XVI	NO_2	H	COOMe	H	H	CH_3	CH_3	0
XVII	NO_2	H	CN	H	H			0
XVIII	NO_2	H	CN	H	H	CH_2CH_3	CH_2CH_3	0
XIX	NO_2	H	NO_2	OCH_3	NO_2	CH_3	CH_3	0
XX	NO_2	H	COOMe	H	NO_2	CH_3	CH_3	0
XXI	NO_2	H	H	CCH_3O	NO_2	CH_3	CH_3	0
XXII	NO_2	H	NO_2	CH_3	NO_2	CH_3	CH_3	0
XXIII	NO_2	H	NO_2	OCH_3	NO_2	H	H	0
XXIV	NO_2	F_3	H	H	H	H	H	0
XXV	NO_2	F_3	H	H	H	$CH_2CH_2CH_2CH_2$		0
XXVI	NO_2	H	CNH_2O	H	H	$CH_2CH_2CH_2CH_2$		0
XXVII	NO_2	H	CN	H	H	$CH_2CH_2CH_2CH_2$		0
XXVIII	NO_2	H	NO_2	H	CN	$CH_2CH_2CH_2CH_2$		22
XXIX	NO_2	H	NO_2	H	CNH_2O	$CH_2CH_2CH_2CH_2$		18
XXX	NO_2	H	NO_2	H	CN	CH_2CH_3	CH_3	25
XXXI	NO_2	H	NO_2	H	$CNCH_3CH_3$	$CH_2CH_2CH_2CH_2$		14
XXXII	NO_2	H	NO_2	H	$CNCH_3CH_3$	$CH_2CH_2CH_2CH_2$		21

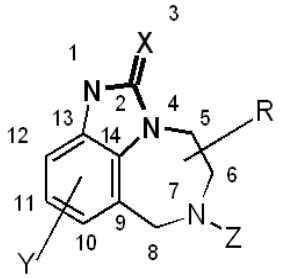

Рис.2. Структура производных TIBO.

Обучение нейронной сети для этой задачи эквивалентно установлению зависимостей между параметрами, характеризующими структуру молекулы, и исследуемыми свойствами соответствующих соединений [7]. Нами использовались многослойные гомогенные нейронные сети, в которых каждый нейрон последующего слоя связан со всеми нейронами предыдущего слоя. В такой сети выделяют входной слой, осуществляющий распределение значений входных параметров, скрытый слой и выходной слой, в котором определяется значение выходного параметра сети.

Таблица 2. Структура и активность производных TIBO.

Соед.	R	X	Y	Z	IC_{50}(μм)	Акт
s-I	H	S	8-Cl	DMA	0.046	1.337
s-II	H	S	9-Cl	DMA	0.16	0.796
s-III	7-Me	O		DMA	12.1	-1.083
s-IV	7-Me	O	8-Cl	DMA	0.145	0.839
s-V	7-Me	O	9-Cl	DMA	0.16	0.796
s-VI	7-Me	S		DMA	0.078	1.108
s-VII	7-Me	S	8-Cl	DMA	0.012	1.921
s-VIII	7-Me	S	9-Cl	DMA	0.023	1.638
s-IX	4,5-ди-Me(*цис*)	O		DMA	56.7	-1.754
s-X	4,5-ди-Me(*цис*)	S		DMA	2.22	-0.346
s-XI	5,7-ди-Me(*транс*)	S		DMA	0.042	1.377
s-XII	5,7-ди-Me(*цис*)	S		DMA	1.15	-0.061
s-XIII	5,7-ди-Me(*R,R;транс*)	O	9-Cl	DMA	0.023	1.638
s-XIV	5,7-ди-Me(*R,R;транс*)	S	9-Cl	DMA	0.48	0.319
s-XV	4,7-ди-Me(*транс*)	S		DMA	25.8	-1.412
s-XVI	5-Me(*S*)	S	8-Cl	DMA	0.005	2.301
s-XVII	5-Me(*S*)	O	9-Cl	DMA	0.18	0.745
s-XVIII	5-Me(*S*)	S	9-Cl	DMA	0.043	1.367
s-XIX	5-Me	O		DMA	0.097	1.013
s-XX	5-Me(*S)*	S		DMA	3.32	-0.521
p-I	4-Me(*S*)	O		DMA	32.1	-1.507
p-II	4-Me(*S*)	S	9-Cl	DMA	0.67	0.174
p-III	7-Me(*S*)	S		$CH_2CH_2CH_3$	2.43	-0.386
p-IV	5-Me(*S*)	O		$CH_2CH_2CH_3$	59.7	-1.776
p-V	5-Me(*S*)	S		$CH_2CH_2CH_3$	1.67	-0.223
p-VI	5-Me(*S*)	O		DMA	34.7	-1.540
p-VII	5-Me(*S*)	S		DMA	0.026	1.585
i-I	5-Me(S)	S	H	DMA	0.044	1.357
i-II	5-Me(S)	S	8-F	DMA	0.006	2.222
i-III	5-Me(S)	S	8-Br	DMA	0.003	2.523
i-IV	5-Me(S)	S	8-CH_3	DMA	0.014	1.854
i-V	5-Me(S)	S	8-O-CH_3	DMA	0.034	1.469
i-VI	5-Me(S)	S	CHO	DMA	0.186	0.730
i-VII	5-Me(S)	S	8-I	DMA	0.048	1.319
i-VIII	5-Me(S)	S	8-CN	DMA	0.056	1.252
i-IX	5-Me(S)	S	8-C=CH	DMA	0.03	1.523
i-X	5-Me(S)	S	10-Br	DMA	1.072	-0.030
i-XI	5-Me(S)	S	10-O-CH_3	DMA	4.677	-0.670
i-XII	5-Me(S)	S	9,10-diCl	DMA	0.026	1.585
h-I	4-Me	S		DMA	0.028	1.553
h-II	4-Me	S		$CH_2CH_2CH_3$	0.06	1.222
h-III	4-Me	S		2-MA	0.026	1.585
h-IV	4-Me	S	9-Cl	2-MA	0.026	1.585
h-V	4-Me	S	9-Cl	DMA	0.0015	2.824
h-VI	4-Me	S	9-Cl	2-MA	0.42	0.377
h-VII	4-Me	S	9-Cl	DEtA	0.012	1.921

Примечания: DMA – диметилаллил-; 2-MA – 2- метилаллил-; DetA – диэтилаллил-

Данные о соединениях, имеющих в названии префикс s-, взяты из [2],
данные о соединениях, имеющих в названии префикс p- – из [3],

данные о соединениях, имеющих в названии префикс i-, - из [4],
данные о соединениях, имеющих в названии префикс h- – из [5].

Показано, что наилучшие результаты достигаются при использовании трехслойных сетей. Для определения оптимального количества нейронов в скрытом слое проводились специальные исследования, в которых было доказано, что наилучшие результаты достигались при использовании 3 нейронов [8].

Основными проблемами, возникающими при использовании ИНС для прогнозирования количественной зависимости «структура-свойство» являются большая корреляция между значениями молекулярных дескрипторов и малый объем выборок (как правило, в таких задачах количество дескрипторов на порядок превышает объем выборки), что приводит к переобучению сети. При переобучении сети ошибка, полученная по обучающей выборке, мала, но при подстановке в сеть новых данных, т.е. при прогнозировании, резко возрастает. Для предотвращения переучивания имеющуюся выборку разделяют на 3 группы: обучающую, по которой ведется минимизация функции ошибки сети, контрольную, по которой проверяется качество прогноза, и тестовую, которая используется для окончательного контроля качества полученной ИНС. Для повышения достоверности результатов использовался метод скользящего контроля, при котором перебирается большое количество комбинаций разбивки полной выборки соединений на обучающую, тестовую и контрольную и анализируется среднее значение ошибки по тестовой выборке.

Для реализации ИНС использовалась программа STATISTICA Neural Networks [9]. Для обучения сети был выбран метод Левенберга-Маркардта, который позволяет уменьшить возможность переобучения сети [10]. Основным показателем качества модели является отношение стандартного отклонения ошибки к стандартному отклонению данных (это коэффициент в программе статистика обозначается S.D. Ratio).

Для производных дитиокарбаминовой кислоты проводился скользящий контроль 2-х видов: полный перебор всех вариантов разделения исходной выборки на группы (при этом контрольная и тестовая выборка включали по 1 соединению) и случайный выбор 4 соединений, из которых далее по величине S.D. Ratio выбирались 3 контрольных и 1 тестовое. При полном переборе среднее значение S.D. Ratio составляет 0.1. Наилучшие результаты достигаются при выборе XI соединения в качестве контрольного, в этом случае среднее значение S.D. Ratio, рассчитанное по всем тестовым соединениям, составляет 0.04. При случайном выборе среднее значение S.D. Ratio, рассчитанное по 28 попыткам, составило 0.05. Таким образом, ИНС демонстрирует прогностические способности, приемлемые для отбора наиболее перспективных соединений. Значимость полученных результатов особенно возрастает, если учесть, что выборка содержит большое количество неактивных соединений, то есть нулевых значений выходного параметра. Анализ таких выборок классическими регрессионными методами существенно затруднен.

Для производных TIBO полный перебор вариантов разбиения на обучающую, тестовую и контрольную выборки затруднен, поэтому использовался случайный выбор. По 10 попыткам среднее значение S.D. Ratio составило 0.2.

Для сокращения входных параметров ИНС нами предложен алгоритм регуляризации, основанный на анализе включенных в ИНС параметров. Анализ выполняется в группе ИНС, обученных на том контрольном веществе, которое дало наименьшую величину S.D. Ratio. Выбираются параметры, входящие в наибольшее число построенных ИНС, обучаются новые ИНС, включающие в качестве входных только отобранные параметры, и выполняется анализ полученных значений S.D. Ratio (рис. 3). В том случае, если среднее значение S.D. Ratio не возрастает, можно считать регуляризацию выполненной и в дальнейшем использовать только отобранные параметры. В противном случае количество отобранных параметров следует увеличить.

При выполнении регуляризации для первой группы соединений количество отобранных параметров уменьшилось с 61 до 24, величина S.D. Ratio уменьшилась с 0.04 до 0.0006. Для второй группы количество параметров уменьшилось с 58 до 29, а величина S.D. Ratio – с 0.2 до 0.08. Из этих результатов видно, что регуляризация позволяет существенно улучшить качество ИНС и их прогностические возможности.

Таким образом, показано, что гомогенные однонаправленные трехслойные нейронные

сети могут быть использованы для построения моделей зависимости «структура-свойство»; применение нейронных сетей эффективно для смешанных выборок, включающих как неактивные соединения, так и соединения, активность которых измерена количественно; предложенный алгоритм регуляризации нейронных сетей позволяет улучшить прогностические возможности сетей.

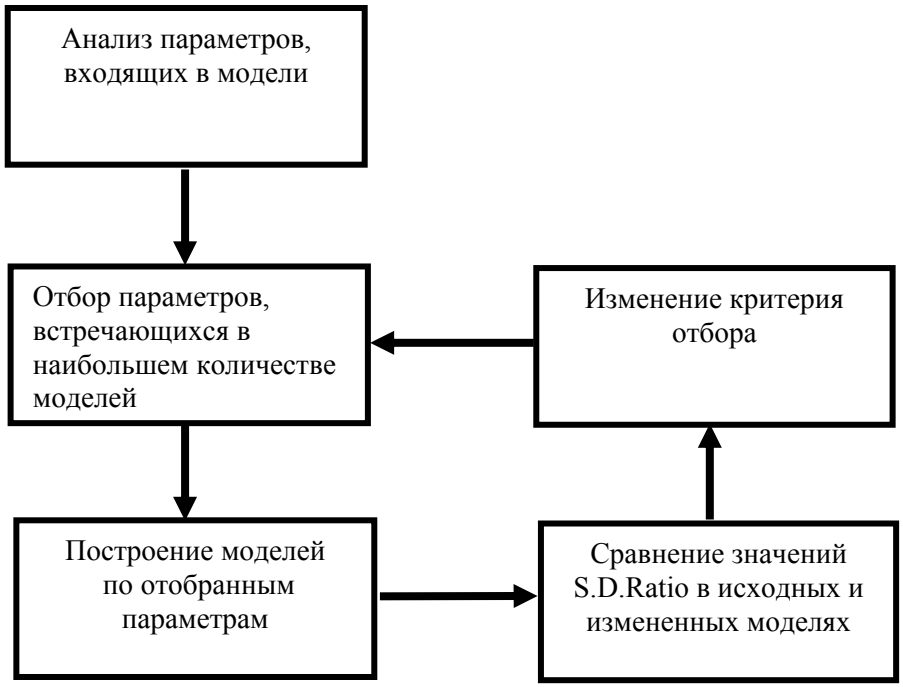

Рис.3 Алгоритм регуляризации.

ЛИТЕРАТУРА:

1. Поройков, В. В. Компьютерное предсказание биологической активности веществ: пределы возможного / В. В. Поройков // Химия в России. –1999. – № 2. – С. 8–12.

2. Crystal Structure at 3,5 Å Resolution of HIV-1 Reverse Transcriptase Complexed with an Inhibitor / L. A. Kohlstaedt [et al.] // Science. – 1992. –Vol. 256. – P.1783–1790.

3. The structure of HIV-1 reverse transcriptase complexed with 9-cloro-TIBO: lessons for inhibitor design / J. Ren [et al.] // Structure. – 1995. – Vol. 3. – P. 915–926.

4. Quantitative structure-activity relationships and comparative molecular field analysis of TIBO derivatised HIV-1 reverse transcriptase inhibitors / S. Hannongbua [et al.] // Journal of Computer-Aided Molecular Design. – 1999. – № 13. – P. 563–577.

5. Eriksson, Mats A. L. Prediction of the Binding Free Energies of New TIBO-like HIV-1 Reverse Transcriptase Inhibitors Using a Combination of PROFEC, PB/SA, CMC/MD and Free Energy Calculations / Mats A. L. Eriksson, J. Pitera, P. Kollman // Journal of Medicinal Chemistry. – 1999. – Vol. 42, № 5. – P. 868–881.

6. Компьютерное моделирование противотуберкулезной активности производных дитиокарбаминовой кислоты / А. М. Юркевич, В. В. Бурляев, В. С. Боридко, С. В. Разливинская // Ученые зап. МИТХТ. – 2000. – Вып.1. – С. 39–42.

7. Баскин, И. И. Применение искусственных нейронных сетей в химических и биохимических исследованиях / И. И. Баскин, В. А. Палюлин, Н. С. Зефиров // Вестник Московского Университета. Химия. – 1999. – Т. 40, № 5. – С. 323–325.

8. The Use of Artificial Neural Networks in QSAR / D. W. Salt [et al.] // Pesticide Science. – 2002. – Vol. 32, № 1. – P. 161–170.

9. www.statsoft.ru

10. Marquardt, D. W. An algorithm for least-squares estimation of non-linear parameters / D. W. Marquardt // Journal of the Society of Industrial and Applied Mathematics. – 1997. – Vol. 11 (2). – P. 431–441.