УДК 541.6

ПРИМЕНЕНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ В ХИМИЧЕСКИХ И БИОХИМИЧЕСКИХ ИССЛЕДОВАНИЯХ

И. И. Баскин, В. А. Палюлин, Н. С. Зефиров

(кафедра органической химии)

В статье рассмотрены основные направления применения искусственных нейронных сетей в химических и биохимических исследованиях. Наибольшее внимание уделено работам по корреляциям между строением химических соединений и проявляемыми ими физико-химическими свойствами и биологической активностью, которые открывают возможность использования нейросети для дизайна материалов с заранее заданными свойствами и новых фармакологических препаратов. Сформулированы основные направления работ по применению искусственных нейронных сетей в органической, аналитической, физической и биологической областях химии.

В настоящее время в вычислительной математике и практически во всех связанных с ней научных и технических дисциплинах происходят коренные изменения в понимании принципов организации вычислительного процесса и подходов к решению прикладных задач, что вызвано значительным ростом интереса к теории и практике использования искусственных нейронных сетей (ИНС). Это уже привело к появлению новых научных дисциплин, таких как нейроинформатика, нейрокомпьютинг и нейроматематика [1, 2]. Появившись относительно недавно как одно из направлений в области искусственного интеллекта, призванное моделировать на компьютере процессы обработки информации, происходящие в человеческом мозгу, ИНС утвердили себя в качестве ведущего направления развития вычислительной математики, превзойдя по числу публикаций все остальные направления, вместе взятые.

Причина такого беспрецедентного роста интереса к ИНС кроется в изначально присущей им способности подходить к обработке информации как к процессу распознавания и классификации образов совершенно произвольной и сколь угодно сложной структуры при помощи неформализуемых алгоритмов, которые сами же нейронные сети и находят. Это выгодно отличает их от традиционных способов компьютерной обработки информации при помощи операций с числами и символами, заранее специфицированных по какому-либо жесткому алгоритму. Элементарной единицей обрабатываемой информации в ИНС является образ произвольной длины, обычно представляемый в виде либо одномерного вектора, либо двумерной матрицы чисел, а элементарной операцией – срабатывание искусственного нейрона (рис. 1), в процессе которого происходит сравнение внешнего образа с хранящимся в связанных с нейроном синапсах эталоном путем вычисления скалярного произведения вектора образа и вектора синаптических весов, вслед за чем нейрон реагирует на такое сравнение путем взятия специальной функции активации от этого скалярного произведения. При этом формируется сигнал, входящий в состав новых образов, подающихся на вход уже другим нейронам сети. Поскольку вследствие «комбинаторного взрыва» для произвольных образов принципиально невозможно полностью описать произвольный алгоритм их обработки, то ИНС решают эту задачу другим способом: они пытаются восстановить алгоритм по его неполному описанию, заданному в виде набора примеров. Иными словами, в основе работы ИНС лежит использование алгоритмов, получаемых при помощи обучения на примерах. Эти алгоритмы кодируются в ИНС в виде значений весов синапсов. Обученная таким образом ИНС способна решать задачу аппроксимации нелинейной функции произвольного вида от многих переменных и, как частный случай этого, задачу отнесения образов к одному из классов, а также задачу категоризации данных и определения их внутренней структуры. ИНС разного строения по-разному ориентированы на решение этих задач.

Для решения первой задачи обычно применяют сети прямых связей (*feed-forward*) (рис. 2), обучаемые при помощи алгоритма обратного распространения ошибок (*backpropagation*), но используются также сети встречного распространения (*counterpropagation*), а также сети радиальных гауссовых функций RBF, каскадные сети корреляций (*cascade correlation networks*) и сети функциональных связей (*functional links networks*). Для решения второй задачи чаще применяют сети Хопфилда (*Hopfield*) и сети адаптивного резонанса ART, тогда как для решения третьей задачи обычно используют самоорганизующиеся

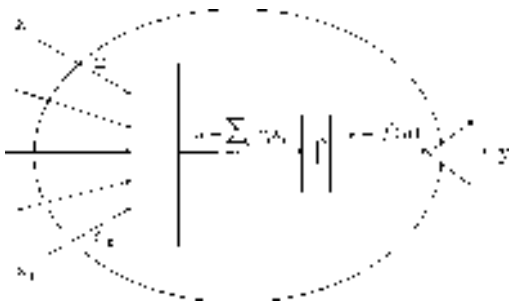

Рис. 1. Схема работы нейрона

карты Кохонена (*Kohonen*). Кроме того, ИНС могут быть использованы для извлечения явных знаний из набора данных (*data mining*). Для этих целей служат специальные процедуры прореживания, упрощения и «вербализации» нейронных сетей [1, 2]). ИНС могут быть реализованы с помощью как нейрокомпьютеров, так и эмуляторов нейронных сетей на компьютерах обычной архитектуры.

В последние годы ИНС находят все более широкое применение при разработке новых химических соединений и материалов с заранее заданными свойствами (библиографию по применению ИНС в изучении связи химическая структура – свойство и химическая структура – биологическая активность можно найти в компьютерной сети Интернет по адресу http://org.chem.msu.su/~baskin/neurchem.html), а также при создании новых лекарственных препаратов [3–7]. Подобные подходы обычно основаны на свойствах ИНС распознавать сложные образы и аппроксимировать непрерывные функции произвольного вида. Для этого наиболее часто используют методики, включающие предварительный перевод информации о связности молекулярного графа, однозначно описывающего строение химического соединения, в вектор инвариантов графа, называемых молекулярными дескрипторами. В дальнейшем векторы молекулярных дескрипторов, вычисленные для структур химических соединений с известными свойствами, вместе с присоединенными значениями свойств используют в качестве множества примеров для обучения ИНС корректно воспроизводить значение прогнозируемого свойства исходя из подаваемого на вход ИНС вектора молекулярных дескрипторов (рис. 3). Для этой цели используют, как правило, многослойные ИНС с прямыми связями (*feed-forward*), обучаемые по методу обратного распространения ошибок (*backpropagation*), хотя в ряде работ успешно использовали ИНС других типов, например сети встречного распространения (*counterpropagation*), гауссовы сети, сети каскадных корреляций (*cascade correlation*) и сети функциональных связей. В процессе обучения осуществляется контроль за прогнозирующей способностью ИНС при помощи заранее выбранной контрольной выборки

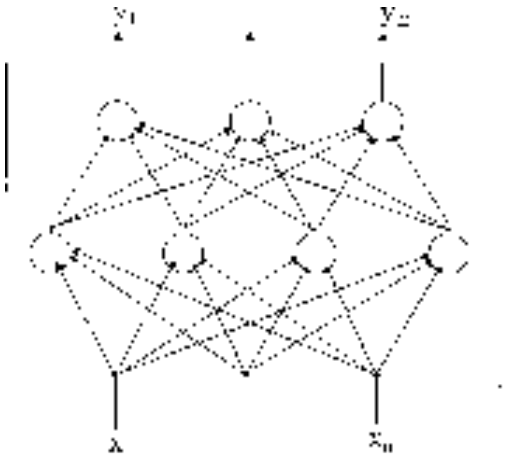

Рис. 2. Строение многослойной искусственной нейронной сети с прямыми связями

Рис. 3. Схема использования искусственной нейронной сети для прогнозирования свойств химических соединений

примеров, что позволяет избежать переобучения (нами было показано, что наиболее объективные оценки прогнозирующей способности ИНС при преждевременном прерывании обучения дает использование двух контрольных выборок [8]). Кроме того, во многих случаях при обучении применяются различные виды регуляризации сети и прунинга (частичного удаления ненужных нейронов и разрежения связей между ними) для упрощения структуры ИНС, уменьшения числа необходимых молекулярных дескрипторов, следствием чего является повышение прогнозирующей способности сети [8, 9]. В дальнейшем обученная таким образом ИНС может быть использована для прогнозирования свойств химических соединений с целью поиска тех из них, которые удовлетворяют заранее заданным параметрам.

В рамках изложенной выше методологии мы одни из первых провели при помощи ИНС работы по прогнозированию температуры кипения, октанового числа, молярного объема, молярной рефракции, теплоты испарения, критического давления и поверхностного натяжения алканов [10], мутагенности гетероциклических аналогов ароматических углеводородов [11], а также значений констант заместителей [12].

В качестве альтернативы рассмотренной выше методологии, подразумевающей использование молекулярных дескрипторов и стандартных ИНС, были разработаны и успешно опробованы на практике сети особого строения, позволяющего обучаться зависимости свойств химических соединений непосредственно от структур соответствующих им помеченных графов, минуя стадию произвольного выбора молекулярных дескрипторов [13]. ИНС такого строения была успешно применена нами для прогнозирования температуры кипения алканов, вязкости, плотности и теплоты испарения углеводородов, теплоты сольватации и поляризуемости произвольных органических соединений, а также давления различных газов, достаточного для проявления анестезирующего эффекта [14].

В последние годы для прогнозирования биологической активности химических соединений и создания новых лекарственных препаратов все более широко применяются

подходы, основанные на свойствах самоорганизующейся сети Кохонена. В одном из используемых при этом подходов используют сеть Кохонена низкого разрешения для отображения создаваемого вокруг молекулы электростатического потенциала на плоскость. При этом оказывается, что молекулы совершенно разного строения, но действующие на одну и ту же биологическую мишень (например, рецептор или фермент), дают сходное отображение, что позволяет предсказывать биологическую активность молекул на качественном уровне [15]. Другой не менее перспективный подход включает задание критерия «близости» химических структур и использование этого критерия для отображения либо больших баз потенциально возможных биологически активных соединений, либо специально сконструированных комбинаторных библиотек химических соединений («первая база») на узлы двумерной решетки при помощи сетей Кохонена высокого разрешения [16]. После этого обученная таким образом ИНС используется для отображения множества химических соединений, обладающих заданной биологической (например, фармакологической) активностью («вторая база»), на эту же решетку нейронов. Вслед за этим соединения из «первой базы», отобразившиеся на те же узлы, что и соединения «второй базы», выбираются для дальнейшего исследования как потенциально активные соединения. Наконец, третий подход к использованию сетей Кохонена [17] (а также сетей ART-2 [18]) в этой области включает их использование для кластеризации баз данных с целью их разбивки на обучающую и контрольную выборки.

Кроме прогнозирования физико-химических свойств органических соединений и их биологической активности ИНС применяют также и для предсказания спектров. Возможность предсказать спектр позволяет, во-первых, осуществлять идентификацию химических соединений, что крайне важно, например, для экологического контроля, и получения новых химических соединений с заданными спектральными свойствами (например, с целью создания новых лазерных красителей). При прогнозировании масс-спектров [19, 20] и спектров поглощения света в инфракрасной области [20–22] в качестве входного набора дескрипторов используют обычно признаки наличия определенной подструктуры в составе химической структуры, при этом в ряде случаев хорошо себя зарекомендовало использование комбинированной сети, состоящей из множества простых ИНС. При прогнозировании спектров ядерного магнитного и электронного парамагнитного резонанса в качестве векторов дескрипторов используют либо, как и в предыдущем случае, формально-структурные параметры, либо результаты квантовомеханических расчетов [23]. Следует, однако, упомянуть работы, в которых эти спектры прогнозировались без использования входного вектора дескрипторов за счет введения информации о химической структуре непосредственно в ИНС [24]. Наконец, при прогнозировании спектров поглощения в ультрафиолетовой и видимой областях при помощи ИНС хорошо себя зарекомендовало, как мы показали в одной из работ на примере длинноволновых полос поглощения цианиновых красителей, сочетание квантовомеханических и формально-структурных дескрипторов [25].

Кроме прогнозирования свойств органических соединений и их спектров ИНС используют также для поиска методов их синтеза и предсказания химической реакционной способности [26, 27]. В этих работах в качестве входной информации для ИНС используются либо формально-структурные параметры, либо непосредственно матрица смежности молекулярного графа, соответствующего химическому соединению, а в качестве выхода – направление либо синтетической, либо ретросинтетической реакции. Во всех этих работах использована стандартная ИНС с прямой связью и с обратным распространением ошибки.

Еще одним направлением является использование ИНС в аналитической химии для определения химического состава анализируемой смеси, что крайне важно для контроля химического производства и экологического мониторинга. В этом случае работа может протекать в двух вариантах: динамическом и статическом.

В первом случае используют ИНС, обученную по значению поступающих на ее вход сигналов, снятых с датчиков (в качестве таких сигналов могут, например, выступать потенциалы, снятые с ион-селективных электродов или с газовых сенсоров), воспроизводить процентный состав анализируемой смеси [28, 29].

Во втором случае ИНС может быть обучена, например, разлагать спектр смеси на спектры индивидуальных компонентов [30].

Из работ по применению ИНС в физической химии наиболее интересны исследования, связанные с моделированием кинетики химических процессов с использованием рекуррентных сетей Хопфилда. Было показано, что моделирование возможно в двух направлениях. В одном случае при помощи ИНС можно моделировать протекание химических реакций [31], а в другом – с помощью специально подобранного набора химических реакций моделировать динамику работы рекуррентных нейронных сетей, что, в сущности, эквивалентно проведению вычислений при помощи «химического компьютера» [32]. Большой интерес вызывает также моделирование гиперповерхности потенциальной энергии молекул при помощи ИНС с прямыми связями. В этом случае после подачи на вход обученной ИНС сигналов, описывающих геометрию молекулы, на выходном нейроне формируется сигнал, соответствующий потенциальной энергии молекулы. Обученные таким образом ИНС могут быть в дальнейшем использованы для моделирования динамических свойств молекул и молекулярных систем [33–35].

Наконец, следует отметить, что одним из наиболее важных направлений применения ИНС в областях, смежных с химией, является их использование в биохимии и структурной молекулярной биологии. Наиболее важным применением ИНС в этой области является прогнозирование вторичной структуры белка по его первичной аминокислотной последовательности [36, 37]). Актуальность этой задачи определяется тем, что знание пространственной структуры белка крайне важно для возможности проводить целенаправленное создание новых лекарственных препаратов, механизм действия которых заключается в связывании с этим

белком. В этом случае в качестве входных сигналов можно использовать свойства аминокислот, находящихся рядом с текущей аминокислотой. Выходной сигнал дает информацию о классификации конформации основной цепи аминокислоты на типы, соответствующие вторичной структуре белка (α-спираль, β-слой и т.д.). Для этой цели, как правило, используется стандартная ИНС с прямыми связями и обратным распространением ошибок при обучении. Еще больший интерес представляют работы по прогнозированию третичной структуры белка исходя из его первичной аминокислотной последовательности [38, 39], однако в этом случае достигнутые успехи не так значительны, как при прогнозировании вторичной структуры белка. Другим направлением применения ИНС в молекулярной биологии является анализ генетической последовательности и распознавание важных участков генов, например промоторов их экспрессии, участков кодирования белков, участков связывания с белками, а также выделение скрытых периодичностей (мотивов) [40–43]. ИНС также используются для прогнозирования мест связывания в белках с нуклеиновыми кислотами, другими белками или с низкомолекулярными лигандами [44, 45]. Результаты подобных прогнозов могут быть в дальнейшем использованы при разработке новых эффективных лекарственных средств.

СПИСОК ЛИТЕРАТУРЫ

1. *Горбань А.Н., Россиев Д.А.* Нейронные сети на персональном компьютере. Новосибирск, 1996.
2. *Ежов А.А., Шумский С.А.* Нейрокомпьютинг и его применение в экономике и бизнесе. М., 1998.
3. *Zupan J., Gasteiger J.* Neural Networks for Chemists - An Introduction. Weinhem, 1993.
4. *Gasteiger J, Zupan J.* // Neural Networks in Chemistry, Angew. Chem. Int. Ed. Engl. 1993. **105.** № 4. P. 503.
5. Neural Networks in QSAR and Drug Design / Ed. J. Devillers. L., 1996.
6. *Баскин И.И., Гальберштам Н.М., Палюлин В.А., Зефиров Н.С.* // Информационные технологии. 1997. № 9. С. 27.
7. *Баскин И.И., Палюлин В.А., Зефиров Н.С.* // Нейрокомпьютер. 1997. №3/4. С. 17.
8. *Baskin I.I., Skvortsova M.I., Palyulin V.A., Zefirov N.S.* // Foundations of Computing and Decision Sciences. 1997. **22.** № 2. P. 107.
9. *Tetko I.V., Villa A.E.P., Livingstone D.J.* // J. Chem. Inf. Comput. Sci. 1996. **36.** № 4. P. 794.
10. *Баскин И.И., Палюлин В.А., Зефиров Н.С.* // ДАН. 1993. **332.** № 6. С. 713.
11. *Abilev S.K., Lyubimova I.K., Baskin I.I., Halberstam N.M. Palyulin V.A.* // Karadeniz Journal of Medical Sciences 1995. **8.** № 4. P. 227.
12. *Baskin I.I., Palyulin V.A., Zefirov N.S.* 12th European Symposium on Quantitative Structure-Activity Relationships «Molecular Modelling and Prediction of Bioactivity», August 23–28, 1998. Copenhagen, Denmark, P. 140.
13. *Баскин И.И., Палюлин В.А., Зефиров Н.С.* // ДАН. 1993. **333.** № 2. С. 176.
14. *Baskin I.I., Palyulin V.A., Zefirov N.S.* // J. Chem. Inf. Comput. Sci. 1997. **37**. № 4. P. 715.
15. *Holzgrabe U., Wagener M., Gasteiger J.* // J. Mol. Graphics 1996. **14.** № 4. P. 185.
16. *Kireev D.B., Ros F., Bernard P., Chretien J.R., Rozhkova N.I.* Computer-Assisted Lead Finding and Optimization. Current Tools for Medicinal Chemistry. Wiley-VCH, 1997. P. 255.
17. *Domine D., Devillers J., Wienke D., Buydens L.* // Quant. Struct.-Act. Relat. 1996. 15. № 5. P. 395.
18. *Domine D., Devillers J., Wienke D., Buydens L.* // J. Chem. Inf. Comput. Sci. 1997. **37.** № 1. P. 10.
19. *Curry B., Rumelhart D.E.* // Tetrahedron Comput. Methodol. 1990. **3.** P. 213.
20. *Gasteiger J., Li X., Simon V., Novic M., Zupan J.* // J. Mol. Struct. 1993. **292.** P. 141.
21. *Robb E.W., Munk M.E.* // Mikrochim. Acta [Wien], 1990. P. 131.
22. *Munk M.E., Madison M.S., Robb E.W.* // Microchim. Acta [Wien] 1991. II. P. 505.
23. *Thomsen J.U., Mayer B.* // J. Magn. Res. 1989. **84.** P. 212.
24. *Kvasnicka V.* // J. Math. Chem. 1991. **6.** P. 63.
25. *Баскин И.И., Айт А.О., Гальберштам Н.М., Палюлин В.А., Алфимов М.В., Зефиров Н.С.* // ДАН. 1997. **357.** № 1. С. 57.
26. *Elrod D.W., Maggiora G.M., Trenary R.G.* // J. Chem. Inf. Comput. Sci. 1990. **30**. P. 477.
27. *Elrod D.W., Maggiora G.M., Trenary R.G.* // Tetrahedron Comput. Methodol. 1990. **3.** P. 163.
28. *Bos M., Bas A., Van-der-Linden W.E.* // Anal. Chim. Acta. 1990. **233.** № 1. P. 31.
29. *Sundgren H., Winquist F., Lukkarl I., Lundstroem I.* // Meas. Sci. Technol. 1991. **2.** P. 464.
30. *Wythoff B.J., Levine S.P., Tomellini S.A.* // Anal. Chem. 1990. **62.** P. 2702.
31. *Lebender D., Schneider F.W.* // J. Phys. Chem. 1993. **97.** № 34. P. 8764.
32. *Hjelmfelt A. Ross J.* // Proc. Natl. Acad. Sci. U.S.A. 1992. **89.** P. 398.
33. *Blank T.B., Brown S.D., Calhoun A.W., Doren D.J.* // J. Chem. Phys. 1995. **103.** P. 4129.
34. *Tafeit E., Estelberger W., Horejsi R., Moeller R., Oettl K., Vrecko K., Reibnegger G.* // J. Mol. Graphics. 1996. **14**. P. 12.
35. *No K.T., Chang B.H., Kim S.Y., Jhon M.S., Scheraga H.A.* /Chem. Phys. Lett. 1997. **271.** P. 152.
36. *Qian N., Sejnowski T.J.* // J. Mol. Biol. 1988. **202.** P. 865.
37. *Bohr H., Bohr J., Brunak S., Cotterill R.M.J., Lautrup B., Norskov L., Olsen O.H., Petersen S.B.* // FEBS Lett. **1988**. P. 223.
38. *Friedrichs M.S., Wolynes P.G.* // Science. 1989. **246.** P. 371.
39. *Bohr H., Bohr J., Brunak S., Cotterill R.M.J., Fredholm F., Lautrup B., Olsen O.H., Petersen S.B.* // FEBS Lett. 1990. **261.** P. 43.
40. *Brunak S., Engelbrecht J., Knudsen S.* // J. Mol. Biol. 1991. 220. № 1. P. 49.
41. *Ezhov A.A., Kalambet Yu.A., Cherny D.I.* // Stud. Biophys. 1989. **129.** P. 183.
42. *Ежов А.А., Токаев А.Г., Чечеткин В.Р.* // Научная сессия МИФИ – 99. Всероссийская научно-техническая конференция «Нейроинформатика-99»: Сборник научных трудов. Ч. 3. М., 1999. С. 182.
43. *Прохоров Р.В.* // Научная сессия МИФИ – 99. Всероссийская научно-техническая конференция «Нейроинформатика-99». Сборник научных трудов. Ч. 3. М., 1999. С. 204.
44. *Hirst J.D., Sternberg M.J.* // Protein Eng. **1991**. 4. № 6. P. 615.
45. *Hirst J.D., Sternberg, M.J.E.* // Biochemistry. 1992. **31.** № 32. P. 7211.

Поступила в редакцию 23.04.99