УДК 615.277:004.032.26

И. В. Семенюта, В. В. Ковалишин, И. Н. Коперник, А. Н. Василенко, В. В. Прокопенко, В. С. Броварец

Создание QSAR моделей для поиска ингибиторов тубулина

(Представлено академиком НАН Украины В. П. Кухарем)

Описаны новые QSAR модели для поиска ингибиторов тубулина. Точность прогноза для учебных и тестовых выборок составляет $\mathrm{Ac} = 0{,}96 \div 0{,}97$ *и* $\mathrm{Ac} = 0{,}95 \div 0{,}97$ *соответственно. Для построения моделей использованы ассоциативные нейронные сети. Оценка качества моделей проведена методами внутренней и внешней проверки. На выборке из 75 новых соединений правильно классифицированно 63% всех веществ, а также 69% активных соединений. С помощью индекса Дайса рассчитана область применения созданных QSAR моделей. Показано, что количество правильно спрогнозированных соединений с DI* $0{,}6-0{,}7$ *и* $\geqslant 0{,}7$ *составляет 74 и 85% соответственно.*

Заболевания онкологического характера являются одной из наиболее острых проблем современной медицины, а в связи с прогнозируемым увеличением числа онкологических заболеваний к 2020 г. возникает необходимость создания новых противоопухолевых препаратов [1].

Значительный прогресс, достигнутый за последние десятилетия в разработке и создании новых противоопухолевых препаратов, нивелируется тем, что подавляющее большинство лекарств, применяемых в медицинской практике, а также находящихся на стадии клинических тестов, обладает общими недостатками, к которым относится низкая селективность биологического действия, что приводит к высокой системной токсичности и снижению терапевтической активности ввиду множественной лекарственной устойчивости опухолевых клеток [2]. Одним из важных направлений в разработке противоопухолевых препаратов является получение новых антимитотических агентов, поскольку ключевая роль митотического веретена в клеточном делении сделала его перспективной мишенью для противоопухолевой химиотерапии [3]. Антимитотические агенты таксол и таксотер относятся к наиболее эффективным препаратам, используемым в современной клинической практике, а ряд антимитотических противоопухолевых препаратов в настоящее время проходят клинические испытания [4]. В связи с этим актуален поиск новых синтетических антимитотических агентов на основе ингибиторов тубулина.

Создание нового лекарственного препарата в настоящее время неразрывно связано с использованием различных математических методов анализа данных, реализованных в форме программного обеспечения, что позволяет создавать прогнозирующие компьютерные модели, которые устанавливают связь между химической структурой и биологической активностью исследуемых соединений (Quantitative Structure — Activity Relationship (QSAR)) [5]. QSAR является важным инструментом для автоматизированного предварительного виртуального скрининга баз данных, разработки комбинаторных библиотек молекулярных фраг-

ментов, позволяет проводить идентификацию и количественное выражение структурных параметров или физико-химических свойств физиологически активных веществ в виде дескрипторов с целью выявления факта влияния каждого из них на биологическую активность. Поэтому применение методов QSAR при создании новых соединений с заданными свойствами позволяет значительно сокращать время и ресурсы, а также осуществлять более целенаправленный синтез соединений, обладающих необходимым заданным комплексом свойств.

Материалы и методы. *Выборка данных.* Мы проанализировали выборку соединений ингибиторов тубулина, воздействующих на динамику клеточных микротрубочек, отобранных из литературных источников и систематизированных в PubChem базе данных [6]. Молекулы ингибиторов смоделировали с помощью программы ChemAxon Standardizer [7], 2D координаты атомов пересчитали заново, ионы и соли удалили из молекулярной структуры, молекулы привели к нейтральной форме и удалили дубликаты. 3D структуры соединений оптимизировали с помощью программы ChemAxon Standardizer и сохранили в SDF формате.

Расчет дескрипторов. Для расчета молекулярных дескрипторов использовали пакет DRAGON [8], который обеспечивает расчет более чем 3200 молекулярных дескрипторов. Каждый дескриптор имеет уникальный код, который позволяет провести его дальнейшую идентификацию. В результате для каждого соединения рассчитали QSAR дескрипторы, такие как гидрофобность, молекулярный объем, количество атомов, количество доноров и акцепторов электронов, количество подвижных связей и другие. Затем первоначальное количество рассчитанных дескрипторов было сокращено. Сначала удаляли дескрипторы, которые имели постоянные значения для всех молекул, затем взаимно коррелированные дескрипторы, т. е. если коэффициент корреляции дескриптора с другими дескрипторами был равен или превышал 0,95, то он удалялся из исходной выборки [8].

Математический аппарат QSAR. Математическим аппаратом QSAR являются методы многомерного статистического анализа данных: линейный и нелинейный регрессионный анализ, дисперсионный анализ, различные методы классификации и распознавания образов, такие как искусственные нейронные сети (ИНС), генетические алгоритмы и др. [5]. Для построения прогнозирующих моделей мы использовали метод ассоциативных нейронных сетей (Associative Neural Networks (ASNN)) [9]. Для выбора наиболее информативных дескрипторов применяли специальные методы анализа информативности дескрипторов [10].

Статистические коэффициенты. Для оценки качества и прогнозирующей способности классификационных моделей [11] использовали такие параметры, как чувствительность (Sn), специфичность (Sp) и общая точность модели (Ac):

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{2}$$

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \tag{3}$$

где TP — количество активных соединений, предсказанных правильно, т. е. как активные; FP — количество активных соединений, предсказанных неправильно, т. е. как неактивные;

TN — количество неактивных соединений, предсказанных правильно, т. е. как неактивные; FN — количество неактивных соединений, предсказанных неправильно, т. е. как активные.

Методика внешней оценки качества QSAR моделей. Методика внешней оценки качества QSAR моделей состоит в использовании тестовых наборов соединений, которые не участвуют в построении модели. Точность всех индивидуальных моделей оценивали с помощью метода пятиразовой перекрестной проверки [5], который заключается в использовании 20% соединений, случайным образом отобранных в тестовый набор, тогда как оставшиеся 80% соединений из общего набора данных использовались для построения QSAR моделей. Эту процедуру последовательно повторили пять раз, при этом получили пять различных тестовых наборов данных и, соответственно, пять наборов для обучения. Таким образом, для каждого набора данных было создано пять моделей и обобщенный прогноз на основе тестовых наборов данных.

Результаты и обсуждение. На первом этапе для лучшего понимания механизмов, лежащих в основе ингибирующей активности тубулина, был исследован набор данных из более чем 190 000 соединений. Для построения QSAR моделей использовали все активные соединения (1621) — ингибиторы тубулина и такое же количество неактивных соединений. Вначале из всех неактивных соединений с помощью программы Instant Jchem [12] было отобрано 10 000 соединений с индексом Дайса (Dice Index (DI)) [12] 0,5–0,6, 1621 относительно активных соединений, т. е. соединений отличных по своей структуре от активных. Далее из этих 10 000 соединений с помощью метода Kennard–Stone Design (KSD) [13] к имеющимся 1621 активному соединению было отобрано 1623 неактивных соединения.

Для каждого соединения с помощью пакета DRAGON получили 1314 дескрипторов. Для выбора наиболее информативных дескрипторов применяли специальные методы анализа информативности дескрипторов, известные в литературе как “pruning methods” [10]. В результате анализа из 1314 было отобрано 134 наиболее информативных дескриптора. Точность прогноза для наборов обучения была в пределах $\mathrm{Ac} = 0{,}96 \div 0{,}97$ и $\mathrm{Ac} = 0{,}95 \div 0{,}97$ для тестовых наборов данных, что свидетельствует о хорошей прогнозирующей способности созданных моделей. Результаты статистического анализа приведенны в табл. 1.

Из литературных источников известно, что многие производные оксазола и тиазола проявляют антибактериальную, фунгицидную, а также противораковую активность [14]. Поэтому на втором этапе анализа созданные модели были использованы для предсказания противоопухолевой активности и выяснения механизма действия 75 соединений, состоящих из гетероциклических соединений, синтезированных в Институте биоорганической химии и нефтехимии НАН Украины, среди которых большинство — производные тиазола и окса-

Таблица 1. Статистический анализ результатов

Название набора данных	Набор обучения				Тестовый набор			
	Количество молекул	Sn	Sp	Ac	Количество молекул	Sn	Sp	Ac
Набор 1	2595	0,97	0,97	0,97	649	0,95	0,96	0,95
Набор 2	2595	0,96	0,96	0,96	649	0,95	0,96	0,96
Набор 3	2595	0,97	0,96	0,96	649	0,99	0,95	0,97
Набор 4	2595	0,97	0,96	0,97	649	0,95	0,96	0,95
Набор 5	2596	0,97	0,96	0,96	648	0,95	0,95	0,95
Общий набор	—	—	—	—	3244	0,97	0,96	0,96

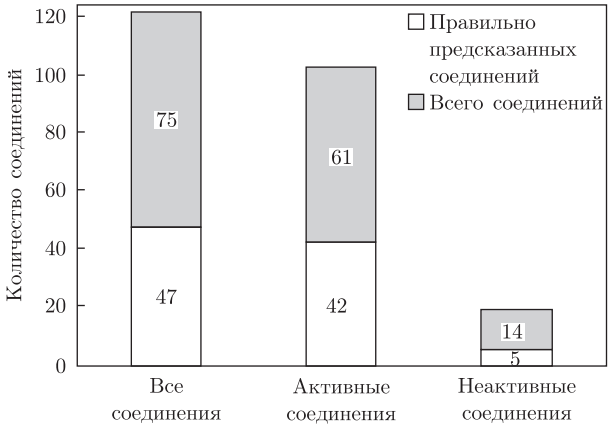

Рис. 1. Результаты прогноза противораковой активности для 75 тестовых соединений

зола. Противораковая активность данных соединений изучалась в рамках международной программы Национального института здоровья США — DTP (Developmental Therapeutic Programm) Национального института рака (NCI) [15]. Результаты прогноза противораковой активности представлены на рис. 1. По результатам предсказания противоопухолевой активности следует отметить 63% суммарного прогноза классификационными моделями. Полученные модели корректно предсказали противоопухолевую активность для 47 соединений из 75 в соответствии с результатами биологических испытаний NCI. При этом следует подчеркнуть, что количество правильно классифицированных активных соединений — 42 из 61, а неактивных — 5 из 14. Таким образом, можно предположить, что для большинства активных молекул данной выборки механизм подавления роста опухолей включает связывание соединений с мономерами тубулина и ингибирование веретена деления, что приводит к нарушению механизма клеточного деления раковых клеток.

Для того чтобы избежать некорректных прогнозов, для QSAR моделей определяют область применения (applicability domain (AD)) [5], которую оценивают с помощью различных мер так называемого расстояния до модели, таких как стандартное отклонение ансамбля моделей, корреляция в пространстве моделей или значения индекса Дайса (DI) и др. от тестируемой молекулы до всех молекул в наборе обучения на основе используемого набора дескрипторов. Мы для оценки области применения модели использовали индекс Дайса [5]. С помощью программы Instant Jchem для 75 тестовых соединений был рассчитан индекс Дайса по отношению к набору обучения. Согласно результатам рассчета (рис. 2), количество правильно спрогнозированных соединений с индексом Дайса 0,6–0,7 составляет 74%, а для веществ с $\mathrm{DI} \geqslant 0{,}7$ — 85%. При использовании соединений с $\mathrm{DI} \leqslant 0{,}6$ количество правильно спрогнозированных соединений снижается до 14%. Данные результаты показывают, что наиболее близкими к набору обучения являются соединения с $\mathrm{DI} > 0{,}6$, поэтому предсказание активности для этих соединений является наиболее точным и реалистичным.

Таким образом, описан ряд новых QSAR моделей с точностью прогноза для учебных выборок $\mathrm{Ac} = 0{,}96 \div 0{,}97$ и $\mathrm{Ac} = 0{,}95 \div 0{,}97$ для тестовых наборов. Для построения этих моделей были использованы известные ингибиторы тубулина различных химических классов, что позволяет получать достоверные результаты прогноза для химических веществ различных классов. Высокая прогнозирующая способность полученных классификационных мо-

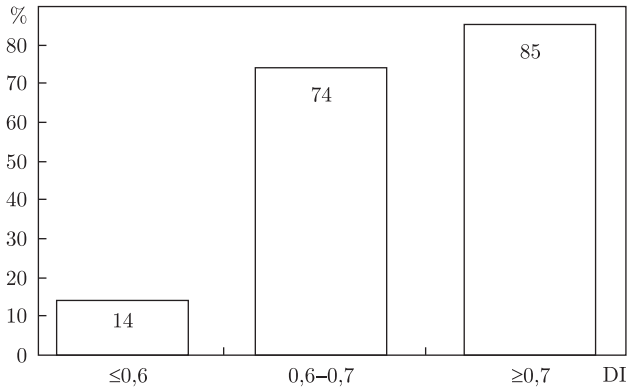

Рис. 2. Процент правильно спрогнозированных соединений в зависимости от индекса Дайса

делей, которая составляет 95–99%, дает возможность с высокой степенью достоверности на начальном этапе исследований определять у веществ активирующую либо ингибирующую направленность действия на белок. На выборке из 75 новых соединений было получено 63% суммарного прогноза, а также 69% правильно классифицированных активных соединений, что говорит о высокой чувствительности созданных моделей к гетероциклическим соединениям классов тиазолов и оксазолов. Также изучена область применения созданных QSAR моделей с помощью индекса Дайса и получены результаты, свидетельствующие о том, что процент правильно спрогнозированных соединений с DI 0,6–0,7 и $\geqslant$0,7 составляет 74 и 85% соответственно.

Авторы выражают благодарность за поддержку программе НАТО "Наука ради мира" (NATO Science for Peace) — грант № EAP.SFPP 984401.

1. *Mistry M., Parkin D., Ahmad A., Sasieni P.* Cancer incidence in the United Kingdom: projections to the year 2030 // Brit. J. Cancer. – 2011. – **105**. – P. 1795–1803.
2. *Nooter K., Stoter G.* Molecular mechanisms of multidrug resistance in cancer chemotherapy // Pathol. Res. Pract. – 1996. – **192**. – P. 768–780.
3. *Jordan M., Wilson L.* Microtubules as a target for anticancer drugs // Nat. Rev. Cancer. – 2004. – **4**. – P. 253–265.
4. *Wani M. C., Taylor H. L., Wall M. E. et al.* Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus brevifolia // J. Am. Chem. Soc. – 1971. – **93**. – P. 2325–2327.
5. *Sushko I., Novotarskyi S., Körner R., Pandey A., Kovalishyn V., Prokopenko, V., Tetko I.* Applicability domain for in silico models to achieve accuracy of experimental measurements // J. Chemometrics. – 2010. – **24**. – P. 202–208.
6. *http*://pubchem.ncbi.nlm.nih.gov.
7. *http*://www.chemaxon.com/products/.
8. *http*://www.talete.mi.it/products/dragon_description.htm.
9. *Tetko I. V.* Neural network studies. 4. Introduction to associative neural network // J. Chem. Inf. Comput. Sci. – 2002. – **42**. – P. 717–728.
10. *Tetko I., Villa A., Livingstone D.* Neural network studies. 2. Variable selection // Ibid. – 1996. – **36**. – P. 794–803.
11. *Li Q., Lai L.* Prediction of potential drug targets based on simple sequence properties // BMC Bioinformatics. – 2007. – **8**. – P. 353–363.
12. *http*://www.chemaxon.com/products/instant-jchem/.
13. *Kennard R. W., Stone L. A.* Computer aided design of experiment // Technometrics. – 1969. – **11**. – P. 137–148.

14. *Savariz F., Foglio M., de Carvalho J. et al.* Synthesis and Evaluation of New β-Carboline-3-(4-benzylidene)-4H-oxazol-5-one Derivatives as Antitumor Agents // Molecules. – 2012. – **17**. – P. 6100–6113.
15. *http*://dtp.nci.nih.gov/branches/btb/ivclsp.html.

Институт биоорганической химии и нефтехимии НАН Украины, Киев

Поступило в редакцию 23.04.2013

І. В. Семенюта, В. В. Ковалішин, І. М. Коперник, О. М. Василенко, В. В. Прокопенко, В. С. Броварець

Створення QSAR моделей для пошуку інгібіторів тубуліну

Описано нові QSAR моделі для пошуку інгібіторів тубуліну. Точність прогнозу для навчальних та тестових вибірок становить Ac $= 0{,}96 \div 0{,}97$ *та* Ac $= 0{,}95 \div 0{,}97$ *відповідно. Для побудови моделей використано асоціативні нейронні мережі. Оцінку якості моделей проведено методами внутрішньої і зовнішньої перевірки. На вибірці з 75 нових сполук правильно класифіковано 63% усіх речовин, а також 69% активних сполук. За допомогою індексу Дайса розраховано область застосування створених QSAR моделей. Показано, що кількість правильно спрогнозованих сполук з DI* $0{,}6{-}0{,}7$ *і* $\geqslant 0{,}7$ *становить 74 та 85% відповідно.*

I. V. Semenyuta, V. V. Kovalishin, I. N. Kopernik, A. N. Vasilenko, V. V. Prokopenko, V. S. Brovarets

Creation of QSAR models to search for inhibitors of tubulin

The study presents new QSAR models to search for tubulin inhibitors. The prediction accuracies for the training and test sets are Ac $= 0.95{-}0.97$ *and* Ac $= 0.95{-}0.97$*, accordingly. QSAR methodologies used Associative Neural Networks. The quality of models have been evaluated using both internal and external validation methods. In a sample of 75 new compounds, we correctly classified 63% of all compounds and 69% of active molecules. The applicability domain of QSAR models was evaluated by the Dice index. It is shown that the percentages of correctly predicted compounds with DI equal to* $0.6{-}0.7$ *and* $\geqslant 0.7$ *are 74 and 85%, respectively.*