УДК 577.152.3:004.032.26

И. В. Семенюта, В. В. Ковалишин, В. В. Прокопенко

Создание QSAR моделей для поиска ингибиторов трипсина

(Представлено академиком НАН Украины В. П. Кухарем)

В исследовании представлены новые QSAR модели для поиска ингибиторов трипсина. Для построения моделей использовали ассоциативные нейронные сети. Оценку качества моделей проводили методами внутренней и внешней проверки. На основании анализа трех выборок веществ (с известными значениями IC_{50} и K_i) был получен ряд регрессионных моделей с точностью прогноза $q^2 > 0{,}7$ и классификационные модели с прогнозирующей способностью 69–80%.

Ключевое положение трипсина в системе пищеварительных ферментов объясняется тем, что он не только участвует в расщеплении пищевых белков, но и активирует все проферменты, образующиеся в поджелудочной железе [1]. При этом активность трипсина угнетается большим числом природных ингибиторов, которые предохраняют ткани от разрушения трипсином. Дефицит (или дефект) некоторых ингибиторов трипсина служит причиной развития патологических состояний [2]. Ингибиторы трипсина применяются также при остром панкреатите и остром некрозе поджелудочной железы, хроническом панкреатите, для профилактики послеоперационного некроза поджелудочной железы, сепсисе, заболеваниях кроветворных органов, карциномах, а также при остром неспецифическом послеоперационном раннем паротите [3, 4]. Принимая во внимание изученность трипсина и механизмов его регуляции, актуальным является создание новых ингибиторов фермента с повышенной специфичностью и улучшенной эффективностью.

В настоящее время создание нового лекарственного препарата неразрывно связано с использованием вычислительной техники, в частности различных математических методов анализа данных, реализованных в форме программного обеспечения. Данный подход позволяет создавать прогнозирующие компьютерные QSAR[1] модели, которые устанавливают связь между химической структурой и биологической активностью исследуемых соединений [5]. QSAR является важным инструментом для автоматизированного предварительного виртуального скрининга баз данных, разработки и комбинаторных библиотек молекулярных фрагментов, дает возможность 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 [6]. Поэтому применение методов QSAR при создании новых соединений с заданными свойствами позволяет значительно сократить время и ресурсы, а также осуществлять более целенаправленный синтез соединений, обладающих необходимым заданным комплексом свойств.

Материалы и методы исследования. *Выборка данных.* В ходе исследования мы проанализировали три выборки соединений с известными значениями IC_{50} и K_i, которые отобраны на основе литературных данных и систематизированных в PubChem [7] и ChEMBL [8]

[1]Quantitative Structure-Activity Relationship.

базах данных соединений. Для всех наборов данных использовали стандартную процедуру. Каждая из молекул была смоделирована с помощью программы Chemaxon standardizer [9]. 2D координаты атомов были пересчитаны заново, ионы и соли удалены из молекулярной структуры, молекулы приведены к нейтральной форме, дубликаты удалены. 3D структуры смоделированы с помощью программы Chemaxon standardizer [9] и сохранены в SDF-формате.

Расчет дескрипторов. Для расчета молекулярных дескрипторов использовали пакет DRAGON [10], обеспечивающий расчет более чем 3200 молекулярных дескрипторов. Каждый дескриптор имеет уникальный код, позволяющий его дальнейшую идентификацию. В результате для каждого соединения были рассчитаны 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.

Математический аппарат QSAR — методы многомерного статистического анализа данных: линейный и нелинейный регрессионный анализ, дисперсионный анализ, различные методы классификации и распознавания образов, такие, как ИНС (искусственные нейронные сети), ЧНК (частные наименьшие квадраты), генетические алгоритмы, метод k-БС (k-ближайших соседей) и другие [6]. В нашей работе для построения прогнозирующих моделей использовался метод ИНС [5]. Для выбора наиболее информативных дескрипторов применяли специальные методы анализа информативности дескрипторов, известные как "pruning methods" [11]. Удаление наименее информативных дескрипторов повышает надежность результатов и увеличивает скорость обучения ИНС [11].

Статистические коэффициенты. Прогнозирующая способность регрессионных моделей оценивалась с помощью коэффициента перекрестной оценки q^2, предложенного Крамером с соавторами [12]:

$$q^2 = 1 - \frac{\sum\limits_{i=1}^{N}(O_i - Y_i)^2}{\sum\limits_{i=1}^{N}(Y_i - Y_{\text{mean}})^2}, \tag{1}$$

где O_i — расчетный вектор активности молекулы i; Y_i — целевой вектор активности молекулы i; Y_{mean} — среднее значение Y_i; N — количество соединений. Кроме этого, для каждой модели рассчитывалась среднеквадратическая ошибка прогноза (root mean squared error, RMSE).

Другие параметры, такие, как чувствительность (Sn), специфичность (Sp) и общая точность модели (Ac), использовались для оценки качества классификационных моделей [13]:

$$Sn = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$Sp = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{3}$$

$$Ac = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}. \tag{4}$$

Здесь 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 состоит в использовании тестовых наборов соединений, которые не берут участия в построении модели. В нашей работе выполнялось внешнее тестирование с использованием альтернативного подхода [14], состоящего в применении 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 и при этом получили пять тестовых наборов данных, пять наборов для обучения. Таким образом, для каждого набора данных было создано пять моделей и обобщенный прогноз на основе тестовых наборов данных.

Результаты и их обсуждение. Для лучшего понимания факторов, лежащих в основе ингибиторной активности трипсина, были исследованы три различных набора данных. Первый — состоял из 1240 соединений из ChEMBL базы данных. Степень активности соединений оценивалась величиной константы ингибирования (K_i). Для каждого соединения с помощью пакета DRAGON получили 1753 дескриптора. Применение методов отбора дескрипторов позволило сократить их количество до 53–66, при этом значительно улучшив точность и качество полученных моделей. Результаты данного анализа приведены в табл. 1. Точность прогноза, полученная методом ИНС для учебных выборок, была приблизительно равна и составила $q^2 = 0{,}81$–$0{,}83$, тогда как для тестовых наборов значения незначительно варьировались от 0,71 до 0,79. Коэффициент перекрестной оценки q^2 для всей выборки равнялся 0,76, что свидетельствует о хорошей прогнозирующей способности предложенных моделей (см. табл. 1).

Соотношения экспериментальных значений $-\log(K_i)$ к их предсказанным значениям для всего набора данных демонстрирует рис. 1. Большинство предсказанных значений не отличается от экспериментальных более чем на один порядок, что также подтверждает приведенные выше выводы.

Второй набор данных состоял из 620 соединений, полученных с PubChem базы данных. Степень активности соединений оценивалась величиной IC_{50}. Наилучшая модель была построена на основе 106–280 дескрипторов (см. табл. 1). Точность прогноза составила $q^2 = 0{,}74$–$0{,}77$ для учебных выборок и $q^2 = 0{,}61$–$0{,}66$ для тестовых наборов. Качество мо-

Таблица 1. Статистический анализ результатов для наборов данных №1 и №2

№ п/п	Название	Число дескрипторов	Набор данных № 1 (K_i)	Число дескрипторов	Набор данных № 2 (IC_{50})
			Набор обучения (тестовый набор), q^2		Набор обучения (тестовый набор), q^2
1	Набор данных № 1	65	0,81 (0,79)	106	0,75 (0,61)
2	Набор данных № 2	60	0,83 (0,77)	280	0,76 (0,61)
3	Набор данных № 3	66	0,83 (0,71)	138	0,77 (0,63)
4	Набор данных № 4	56	0,82 (0,77)	270	0,74 (0,66)
5	Набор данных № 5	53	0,81 (0,78)	216	0,77 (0,66)
6	Общий набор	—	0,76	—	0,63

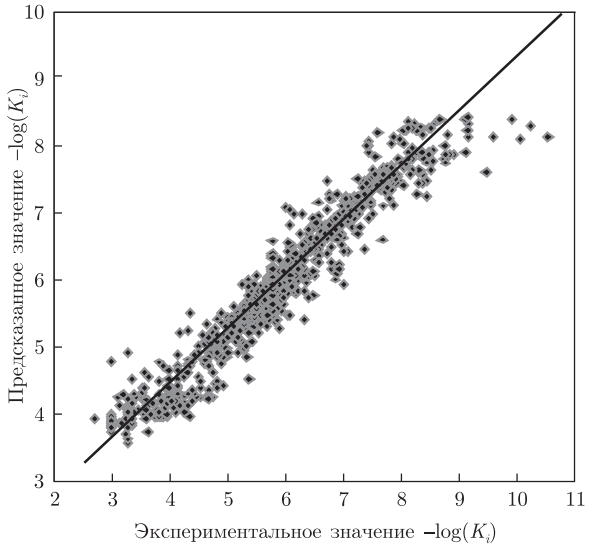

Рис. 1

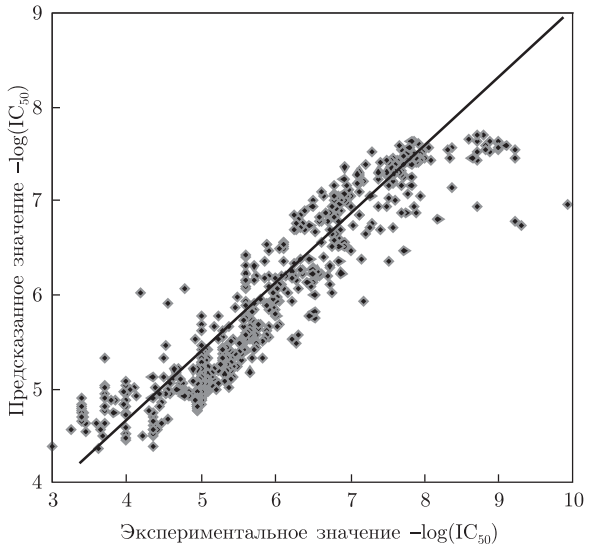

Рис. 2

дели для всей выборки — $q^2 = 0{,}63$, что обусловлено разнообразием данных, представленных в данном наборе (см. табл. 1; рис. 2). Соотношения экспериментальных значений $-\log(\mathrm{IC}_{50})$ к их предсказанным значениям для всего набора данных демонстрирует рис. 2. Большинство предсказанных значений отличается от экспериментальных в пределах одного значения $\log(\mathrm{IC}_{50})$. Следует отметить, что активность некоторых соединений предсказана как более низкая, т. е. отличается от экспериментальных значений активности более чем 2 $\log(\mathrm{IC}_{50})$, что несколько ухудшает значение q^2 для обобщенной модели (см. № 6 в табл. 1; рис. 2).

В качестве дополнительного критерия для поиска потенциальных ингибиторов трипсина был построен также ряд классификационных моделей. Для построения моделей использовалась выборка из 446 соединений, отобранных из PubChem базы данных [7]. Среди этих

Таблица 2. Статистический анализ результатов для набора данных №3

№ п/п	Название	Набор обучения				Тестовый набор			
		Количество молекул	*Sn*	*Sp*	*Ac*	Количество молекул	*Sn*	*Sp*	*Ac*
1	Набор данных № 1	356	0,88	0,90	0,89	90	0,71	0,70	0,70
2	Набор данных № 2	357	0,90	0,88	0,89	89	0,69	0,67	0,68
3	Набор данных № 3	357	0,88	0,89	0,88	89	0,77	0,77	0,76
4	Набор данных № 4	357	0,87	0,85	0,86	89	0,70	0,80	0,75
5	Набор данных № 5	357	0,89	0,90	0,90	89	0,65	0,72	0,69
6	Общий набор	—	—	—	—	446	0,74	0,73	0,74

соединений 50% являются ингибиторами трипсина, а другие 50% — активаторами трипсина, что является необходимым условием создания полноценных моделей. Соединения аналогично разделили на пять выборок и построили пять QSAR моделей. Точность прогноза для учебных выборок составила 85–90% и 69–80% для тестовых выборок (табл. 2).

В этом исследовании был представлен ряд новых QSAR моделей с точностью прогноза $q^2 > 0{,}7$ для поиска новых ингибиторов трипсина. Для построения этих моделей были использованы известные ингибиторы трипсина различных химических классов, что позволит получать достоверные результаты прогноза для химических веществ различных классов. Высокая прогнозирующая способность полученных классификационных моделей (69–80%) дает возможность с высокой степенью достоверности на начальном этапе исследований определять у веществ активирующую либо ингибирующую направленность действия на фермент. Следует добавить, что представленные методы позволяют быстро строить прогнозирующие QSAR модели не только для поиска новых ингибиторов трипсина, но и для других видов биологической активности среди различных классов химических соединений.

1. *Антонов В. К.* Химия протеолиза. – Москва: Наука, 1991. – 504 с.
2. *Веремеенко К. Н., Голобородько О. П., Кизим А. И.* Протеолиз в норме и при патологии. – Киев: Здоровье, 1988. – 199 с.
3. *Веремеенко К. Н.* Протеолитические ферменты и их ингибиторы. Новые области применения в клинике // Врач. дело. – 1994. – № 1. – С. 8–13.
4. *Stenman U.* Role of the tumor-associated trypsin inhibitor SPINK1 in cancer development // Asian J. Androl. – 2011. – **13**. – P. 628–629.
5. *Gasteiger J., Zupan J.* Neural networks in chemistry // Chem. Int. Ed. Engl. – 1993. – **32**. – P. 503–527.
6. *Tetko I., Sushko I., Pandey A. et al.* Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection // J. Chem. Inform. Model. – 2008. – **48 (9)**. – P. 1733–1746.
7. *http:*//pubchem.ncbi.nlm.nih.gov.
8. *https:*//www.ebi.ac.uk/chembl/.
9. *http:*//www.chemaxon.com/products/.
10. *http:*//www.talete.mi.it/products/dragon_description.htm.
11. *Tetko I., Villa A., Livingstone D.* Neural network studies. 2. Variable selection // J. Chem. Inform. Comput. Sci. – 1996. – **36**. – P. 794–803.
12. *Cramer R. D., Patterson D. E., Bunce J. D.* Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins // J. Am. Chem. Soc. – 1988. – **110**. – P. 5959–5967.
13. *Li Q., Lai L.* Prediction of potential drug targets based on simple sequence properties // BMC Bioinformatics. – 2007. – **8.** – P. 353–363.
14. *Tropsha A.* Best Practices for QSAR Model Development, Validation, and Exploitation // Mol. Inf. – 2010. – **29**. – P. 476–488.

Институт биоорганической химии и нефтехимии НАН Украины, Киев

Поступило в редакцию 21.03.2012

І. В. Семенюта, В. В. Ковалишин, В. В. Прокопенко

Створення QSAR моделей для пошуку інгібіторів трипсину

У дослідженні представлено нові QSAR моделі для пошуку інгібіторів трипсину. Для побудови моделей використовували асоціативні нейронні сітки. Оцінку якості моделей здійснювали методами внутрішньої і зовнішньої перевірки. На підставі аналізу трьох вибірок речовин (з відомими значеннями IC_{50} й K_i), був отриманий ряд регресійних моделей з точністю прогнозу $q^2 > 0{,}7$ та класифікаційні моделі з прогнозуючою здатністю 69–80%.

I. V. Semenyuta, V. V. Kovalishin, V. V. Prokopenko

Creation of QSAR models to search for inhibitors of trypsin

New QSAR models to search for inhibitors of trypsin are presented. The models are built with the use of associative neural networks. The quality of models has been evaluated using both internal and external validation methods. Based on the analysis of three samples of substances with the known values of IC_{50} and K_i, a number of regression models with a prediction accuracy of $q^2 > 0.7$ and the classification models with a predictive ability of 69–80% are obtained.