УДК 574.587.51:574.791/792

ПОСТРОЕНИЕ КЛАССИФИКАЦИОННОЙ МОДЕЛИ ДЛЯ ВИРТУАЛЬНОГО СКРИНИНГА ИНГИБИТОРОВ ТИРОЗИНОВЫХ КИНАЗ

К.В.Балакин, Я.А.Иваненков, А.В.Скоренко,
С.Н.Коваленко*, И.А.Журавель*, В.П.Черных*

Институт химического разнообразия,
141700, Россия, Московская обл., г. Долгопрудный, ул. Первомайская, 1. Тел. (095) 408-80-51
* Национальный фармацевтический университет

Ключевые слова: виртуальный скрининг; тирозиновые киназы; ингибирование; метод опорных векторов; биомишень; МОВ-модель

Описана процедура разработки модели количественной связи "структура — активность" для классификации соединений по их способности ингибировать тирозиновые киназы. Построенная классификационная модель позволяет корректно предсказывать более 72% ингибиторов тирозиновых киназ и более 83% агентов, неактивных по отношению к киназам, в независимых тренирующих выборках.

CONSTRUCTION OF CLASSIFICATION MODEL FOR VIRTUAL SCREENING OF TYROSINE KINASES INHIBITORS

K.V.Balakin, Ya.A.Ivanenkov, A.V.Skorenko, S.N.Kovalenko, I.A.Zhuravel, V.P.Chernykh

In work procedure of model building of quantitative relation "structure — activity" for classification of connections by their ability to inhibit tyrosine kinases is described. The classification constructed model correctly allows predicting more than 72% of inhibitors of tyrosine kinases and more than 83% of agents, inactive in relation to kinases, in independent training samples.

ПОБУДОВА КЛАСИФІКАЦІЙНОЇ МОДЕЛІ ДЛЯ ВІРТУАЛЬНОГО СКРИНІНГУ ІНГІБІТОРІВ ТИРОЗИНОВИХ КІНАЗ

К.В.Балакін, Я.А.Іваненков, А.В.Скоренко, С.М.Коваленко, І.О.Журавель, В.П.Черних

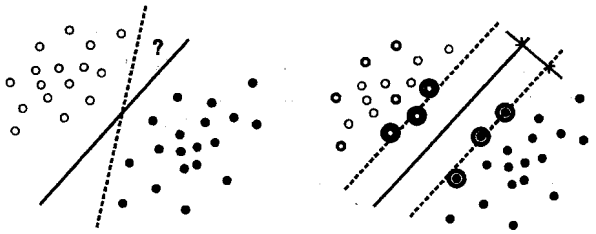
Описана процедура розробки моделі кількісного зв'язку "структура — активність" для класифікації сполук за їх здатністю інгібувати тирозинові кінази. Побудована класифікаційна модель дозволяє коректно виділити більш ніж 72% інгібіторів тирозинових кіназ і більш ніж 83% агентів, неактивних по відношенню до кіназ, у незалежних тренувальних вибірках.

Тирозиновые киназы как биомишень для лекарственных препаратов

В настоящей работе освещены некоторые теоретические и практические аспекты построения классификационной модели количественных связей "структура — активность" (КССА) для ингибиторов тирозиновых киназ (ТК). ТК играют важнейшую роль в передаче сигнальной информации в клетках посредством катализа переноса γ-фосфатной группы аденозин-5'-трифосфата (АТФ) на белковый субстрат [1]. Хотя ТК разных типов отличаются размерами, механизмами активации, организацией субъединиц и внутриклеточной локализацией, они все имеют структурно консервативный каталитический сайт связывания АТФ, одновременно являющийся сайтом связывания для большинства открытых к настоящему времени ингибиторов ТК. Консервативная природа этого сайта представляет собой серьезную проблему при открытии селективных ингибиторов. В то же время из нее следует то, что связывание различных ингибиторов происходит в сходном микроокружении, характеризующемся специфическими стерическими параметрами, липофильностью, числом центров образования водородных связей, величиной и знаком зарядов и т. д. Это сходство представляет собой базис, на котором возможно эффективное нахождение количественной функциональной зависимости, дискриминирующей между ингибиторами ТК и фармацевтическими соединениями, действующий по другим механизмам активности.

Метод опорных векторов

Существует целый ряд математических алгоритмов, позволяющих строить классификационные количественные связи "структура — актив-



а б

Рис. 1. (а) Две возможных линейных разделяющих гиперплоскости; (б) наилучшая разделяющая гиперплоскость максимизирует отступ.

ность" (КССА). Среди них следует выделить различные варианты нелинейного картирования, такие как самоорганизующиеся карты Кохонена [2] и карты Саммона [3], а также Искусственные нейронные сети (ИНС) [4]. ИНС за последние годы стали стандартным инструментом в области разработки КССА моделей в области виртуального скрининга компьютерных библиотек соединений. Они относительно просты в использовании, являясь при этом мощным и гибким алгоритмом. В то же время им присущи некоторые типичные недостатки. Например, ИНС можно сравнить с "черным ящиком", поскольку они не позволяют выразить связи между входными и выходными переменными в явном виде; это может приводить к трудностям в интерпретации и анализе построенных моделей, а также в их оптимальной настройке. Далее, для алгоритма ИНС существует проблема перетренированности. Этот эффект выражается в том, что построенная модель с высокой точностью описывает зависимости между входными и выходными переменными, включая фоновые шумовые сигналы, в то время как реальная зависимость остается не выявленной. Наконец, многие ИНС модели требуют значительного времени на тренировку.

В последнее время все более популярным становится альтернативный алгоритм, который обеспечивает как минимум не меньшую точность и гибкость моделирования и при этом лишен многих недостатков, присущих нейронно-сетевому подходу. Метод опорных векторов (МОВ) был разработан в 80-е годы прошлого века в Советском Союзе [5]. Сообщалось об его эффективном использовании в различных областях — от анализа геномных последовательностей до распознавания лиц. Появились также публикации об использовании МОВ для разработки лекарственных препаратов [6]. Недавно нами разработана МОВ-модель, позволяющая предсказывать активность соединений по отношению к карбоновой ангидразе II [7]. Показательно, что во всех случаях МОВ превосходил ИНС по классификационной способности и скорости тренировки.

Существует несколько публикаций, посвященных теории МОВ [5, 7-10]. В настоящей работе мы кратко изложим суть алгоритма опорных векторов в применении к классификационным задачам.

Ключевой концепцией МОВ является принцип *минимизации структурного риска* (МСР) [11]. Предположим, мы имеем тренирующий набор объектов, состоящий из m точек, описываемых как $\{(x_1,y_1)...(x_m,y_m)\}$, где x является совокупностью свойств (дескрипторов; X называется вводным пространством), а y_m является меткой класса, например, -1 и 1 для задач бинарной классификации. Предположим также, что существует неизвестная функция распределения вероятностей $P(x,y)$, которая описывает отнесение свойств к классам. Попытаемся соотнести дескрипторы с классами путем введения функции классификации $f(x,a)$, значения которой лежат в диапазоне от -1 до 1 в зависимости от класса. Параметр a этой функции может быть найден путем минимизации функциональной или ожидаемой ошибки:

$$I(a) = \int Q(x,a,y)P(x,y)dxdy,$$

где $Q(x,a,y)$ является функцией потерь. Например, $Q = (y - f(x,a))^2$ соответствует обычной функции наименьших квадратов. Очевидной проблемой является тот факт, что указанный интеграл зависит от неизвестного истинного распределения P, определенного на всем пространстве вводных дескрипторов, тогда как мы имеем лишь некоторую выборку из этого распределения, а именно, тренирующую выборку. Следовательно, для практических задач этот интеграл должен быть заменен другой функцией, описывающей интегральную сумму только для точек тренирующей выборки и называемой *структурный риск*. Однако, может существовать целый ряд функций, обеспечивающих хорошую классификацию на тренирующей выборке. Проблема заключается в том, чтобы выбрать из них такую, которая бы обеспечила наилучшую классификацию также на других, пока неизвестных объектах, то есть обеспечила бы наилучшую общую предсказывающую способность. В соответствии с принципом МСР эта задача может быть решена путем минимизации структурного риска, а также *доверительного интервала*. Последняя величина пропорциональна отношению размерности модели (измеряемой в рамках так называемой размерности Вапника-Червоненкиса) и числа объектов в тренирующей выборке. Опуская избыточные математические выражения, укажем, что МСР постулирует следующее: оптимальный классификатор обеспечивает баланс между ошибкой тренировки и снижением размерности модели, при этом ограничивая возможности перетренировки.

Рассмотрим пример классификации в двумерном пространстве вводных дескрипторов (рис. 1а). Для указанной тренирующей выборки как сплошная, так и прерывистая линия обеспечивают одинаково хорошую классификацию. Какая же из линий будет более оптимальной? Интуитивно нам понятно, что сплошная линия будет менее чувствительна к небольшим изменениям в поло-

жении тренирующих объектов. Другими словами, разделяющая линия должна максимально отстоять от тренирующих объектов, принадлежащих разным классам. Именно последняя мысль является стержневой в принципе МСР, составляющем суть метода опорных векторов: оптимальный классификатор должен обеспечивать наибольший отступ, разделяющий классы (отступ определяется как сумма наикратчайших расстояний от разделяющей линии до ближайших объектов, принадлежащих обоим классам (рис. 1б). С геометрической точки зрения, оптимальная линия делит пополам наикратчайшее расстояние между выпуклыми оболочками двух классов.

Примечательным образом оказывается, что относительно небольшое число объектов, расположенных в непосредственной близости от разделительной линии (указаны кружками на рис. 1б), полностью определяет положение *оптимальной разделительной линии* (или так называемой *оптимальной разделительной гиперплоскости* (ОРГ) в случае многомерного пространства свойств). Эти объекты называются *опорными векторами* (ОВ).

Как ОРГ, так и ОВ могут быть найдены путем решения соответствующих квадратных уравнений.

К сожалению, возможности линейного классификатора ограничены. Реальные задачи зачастую линейно неразделимы в пространстве R^n. Возможности линейного классификатора можно существенно расширить путем нелинейного отображения исходного пространства в пространство потенциально намного более высокой размерности $\Phi: Rn \rightarrow F$ и применения линейного классификатора в пространстве F. Классифицирующую функцию $f(x)$ можно преобразовать таким образом, что она будет представлять собой линейную комбинацию скалярных произведений вектора X с векторами тренировочного набора. А линейный классификатор, использующий только скалярные произведения, может неявно оперировать в пространстве F, используя аппарат *ядерных функций* [12], не работая с векторами пространства F и даже не зная отображения Φ. Таким образом, можно использовать линейный классификатор (опорные вектора) для корректной работы с линейно неразделяемыми классами без существенного усложнения вычислительных операций.

В качестве краткого резюме можно отметить, что использование метода опорных векторов позволяет: а) получить функцию классификации с минимальной верхней оценкой ожидаемого риска (ошибкой классификации); б) использовать линейный классификатор для работы с нелинейно разделяемыми данными, сочетая простоту с эффективностью.

Тренирующая база

Использованная нами концепция КССА моделирования основана на статистическом анализе свойств большой тренирующей выборки соединений, для которых было экспериментально показано существование ингибирующей активности по отношению хотя бы к одному члену семейства тирозиновых киназ. В качестве такой выборки мы использовали 1249 соединений, описанных в специальной коммерческой базе данных Ensemble в качестве ингибиторов ТК. В качестве выборки сравнения (отрицательная тренирующая выборка) мы использовали примерно 9000 соединений, для которых в базе Ensemble была указана активность иная, чем киназная. Поскольку не все эти соединения были испытаны на киназную активность, то эта база сравнения не является абсолютно корректной отрицательной выборкой. Однако это допущение выглядит оправданным, так как обычная доля соединений с ТК-активностью при испытаниях рандомических выборок соединений не превышает 0,5-1%. Именно этой цифрой можно оценить потенциальную погрешность нашего метода, связанную с отрицательной тренирующей выборкой. Более существенным фактором, вносящим дополнительную неточность в эксперимент, является различный уровень активности ингибиторов ТК, собранных из разных источников, включая исследовательские статьи, патенты, сообщения на научных конгрессах. Это приводит к тому, что в позитивной тренирующей выборке могут одновременно присутствовать ингибиторы, отличающиеся по своей активности в 10000 раз (например, ингибиторы с IC_{50}=1 нМ и IC_{50}=10 мкМ). Как следствие, оценка факторов активности может быть несколько искажена. Следует заметить, что отмеченные проблемы являются характерными для настоящего этапа развития методов классификационного КССА моделирования. Накопление экспериментального материала позволит в будущем свести эти проблемы к минимуму.

При создании указанных тренирующих выборок были выполнены все фильтрующие процедуры, описанные в нашей предыдущей работе [13]. Для обеспечения единого масштаба измерений для обеих выборок, а также соответствия анализируемых молекул приблизительному диапазону молекулярных масс, характерных для лекарственных соединений, в полную тренирующую выборку были включены соединения с молекулярной массой от 200 до 700 г/моль. В целях обеспечения корректного расчета молекулярных дескрипторов все соединения тренирующей выборки содержали только следующие атомы: C, H, N, O, S, P, F, Cl, Br, I.

Описанные тренирующие базы данных были использованы во всех описанных далее вычислительных модельных экспериментах.

Молекулярные дескрипторы

В данной работе для отбора дескрипторов мы применили анализ главных компонентов. Теоретические и методологические аспекты этого подхода изложены во многих работах, например, [14] и поэтому в настоящей статье не описаны. Полученный набор был оптимизирован в серии модельных экспериментов. В результате были отобраны 8 дескрипторов.

Таблица

Параметры МОВ-модели

Параметр	Величина
Тренирующая база	Положительная (ТК-активные): 1249 соединений; негативная (ТК-неактивные): 8592 соединений.
Дескрипторы	LogP - логарифм коэффициента распределения в системе 1-октанол/вода; HBA - число акцепторов водородной связи; HBD - число доноров водородной связи; BRot - число вращающихся связей; Dens - плотность; Zagreb - загребский индекс; MR - молекулярная рефракция; RPSA - относительная площадь положительно заряженной поверхности
Параметры моделирования	Метод опорных векторов (Support Vector Machine); радиальная базисная функция (RBF), γ=0,4; отношение размеров тренирующей/кроссвалидирующей/тестируемой выборок: 3-1-1
Процент корректно классифицированных соединений	Для тестируемой выборки: ТК-активные соединения: 72%; ТК-неактивные соединения: 83% (при пороговом индексе 0,4)

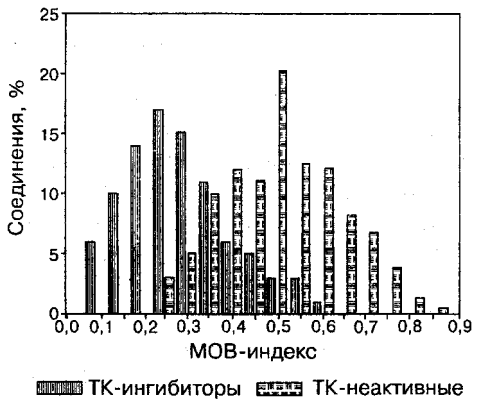


Рис. 2. Распределение соединений тестируемой выборки по МОВ-индексам.

Параметры построенной модели представлены в табл. 1. Полная тренирующая база, состоящая из 9841 соединения, включая ТК-активные и ТК-неактивные, была рандомизирована и поделена на три части: 1) тренирующая выборка (60% от общего числа соединений), 2) кросс-валидирующая выборка (20%) и 3) тестируемая выборка (20%). МОВ классификатор был основан на нелинейной ядерной функции RBF.

Валидация модели

Для оценки классификационной способности построенной МОВ-модели мы рассчитали МОВ-индексы для независимой тестируемой выборки соединений, не использовавшихся в тренирующей или кросс-валидирующей выборках. Результаты тестирования представлены на рис. 2, демонстрирующем гистограмму распределения МОВ-индексов для ТК-ингибиторов и ТК-неактивных соединений. Видно, что модель обеспечивает хорошее, статистически значимое разделение между исследуемыми категориями соединений. Максимальное качество разделения обеспечивается при пороговом индексе 0,4: 72% ТК-ингибиторов и 83% ТК-неактивных соединений классифицируются при этом корректно. Аналогичное качество разделения было обеспечено при двух других независимых рандомизациях исходной тренирующей базы (данные здесь не представлены). Можно сделать вывод о том, что МОВ-модель, построенная при использовании данной тренирующей выборки соединений и описанного набора молекулярных дескрипторов, позволяет успешно дискриминировать соединения по их способности ингибировать тирозиновые киназы.

Примеры соединений с высоким и низким МОВ-индексом

На рис. 3 и 4 мы приводим соответственно структуры типичных ТК-активных и ТК-неактив-

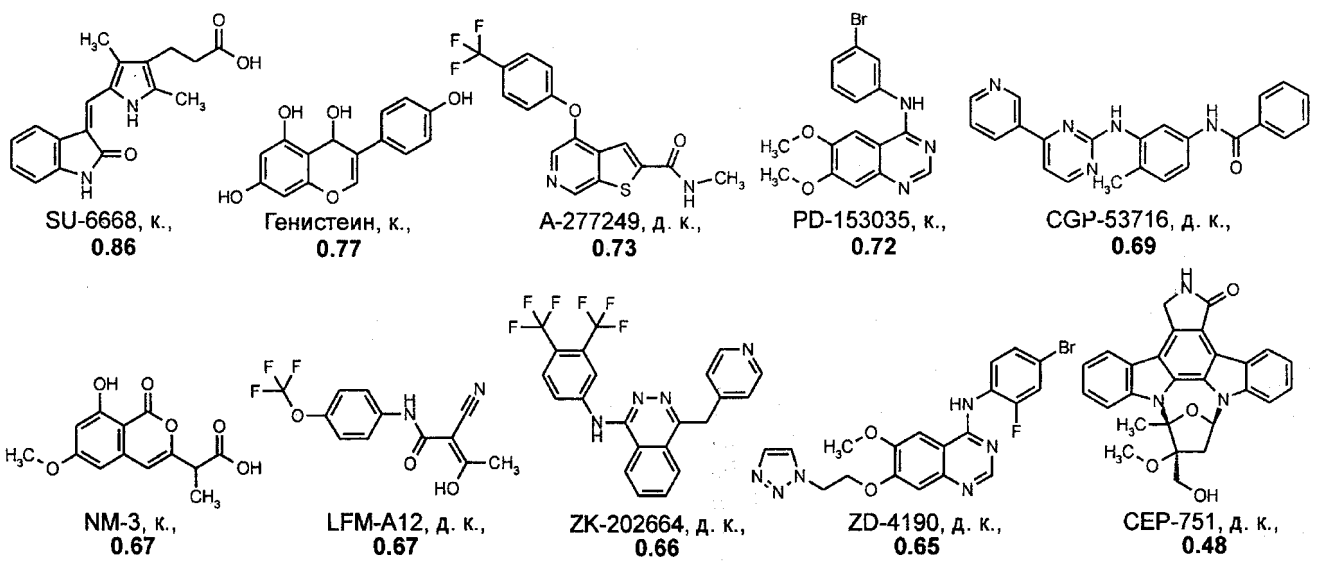


Рис. 3. Структуры и МОВ-индексы типичных ингибиторов тирозиновых киназ, вошедших в клиническую (к.) или доклиническую (д. к.) фазу испытаний.

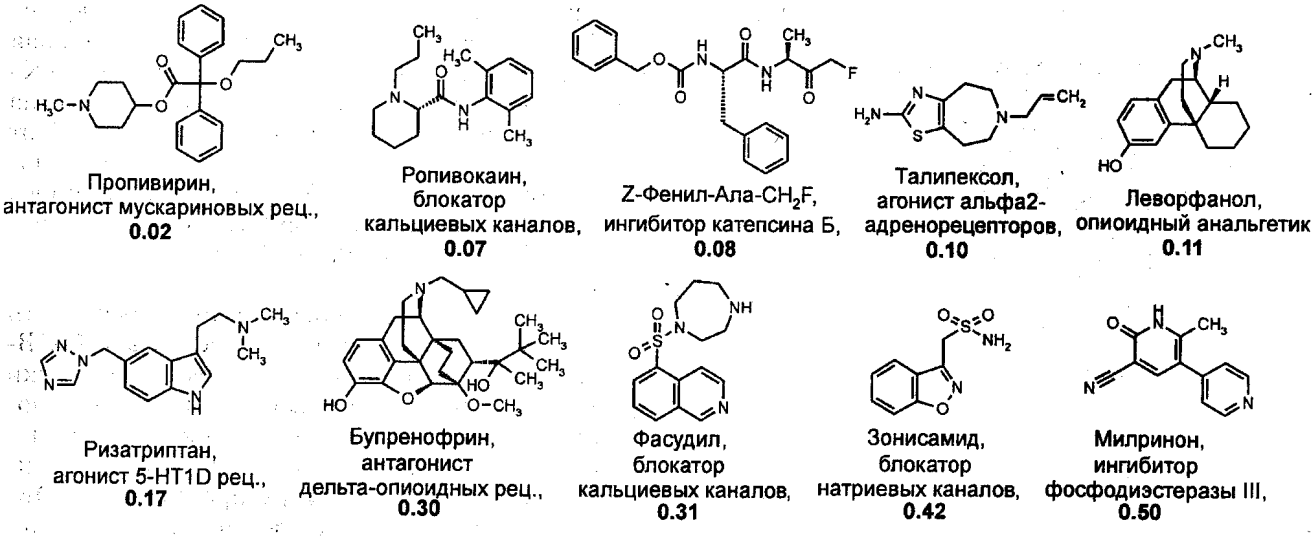


Рис. 4. Структуры и МОВ-индексы лекарственных препаратов, действующих на некиназные биомишени.

ных соединений, а также их рассчитанные МОВ-индексы. Визуальный анализ структур позволяет оценить типичные различия, которые были эффективно выявлены компьютерной моделью, основанной на расчетных физико-химических и топологических параметрах молекул. Так, ингибиторы тирозиновых киназ, как правило, обладают большим числом ароматических колец, преимущественно содержащих гетероатомы. Они менее конформационно подвижны и практически не содержат алифатических липофильных цепочек. Очевидно, что во многих случаях отнесение к потенциальным ингибиторам тирозиновых киназ может быть сделано по наличию определенных подструктурных фрагментов. Так, пиримидин-содержащие скаффолды, являющиеся миметиками АТФ (натуральный субстрат тирозиновых киназ), обладают повышенной вероятностью проявления ТК-ингибирующей активности. Однако отбор по подструктурному принципу, как правило, не позволяет находить новые структурные хемотипы активных соединений. По этой причине разработанный метод, основанный на физико-химических и топологических параметрах молекул, является предпочтительным для направленного отбора оригинальных малоисследованных структур.

Экспериментальная часть

Расчет и отбор молекулярных дескрипторов, а также манипуляции со структурными базами соединений проведены при помощи программного продукта ChemoSoft компании "Chemical Diversity Labs" (Сан Диего, США, www.chemdiv.com). Построение МОВ-модели осуществлялось при помощи программы SVMlight, доступной для свободного некоммерческого использования [11]. Все расчеты выполнены в операционной системе Windows 2000 на стандартном IBM PC-совместимом персональном компьютере с ЦПУ AMD Athlon 1.4 ГГц и 512 Мб оперативной памяти.

Выводы

Ингибиторы тирозиновых киназ могут быть использованы в качестве онколитических препаратов, антидиабетиков и антипсориазных агентов. Данный вид мишень-специфической активности является чрезвычайно актуальным для современной фармацевтической индустрии. Большинство крупных фармацевтических концернов имеет исследовательские подразделения, специально занимающиеся киназной тематикой. Многие вендоры исследовательской химии для биоскрининга предлагают специальные библиотеки, фокусированные на тирозиновых киназах. Разработаны многочисленные варианты биологических in vitro тестов для эффективного обнаружения ингибиторов ТК. Регулярно проводятся тематические конгрессы, посвященные исследованиям в этой области. Однако, несмотря на интенсивные исследования последнего десятилетия, в настоящий момент лишь два соединения были выведены на рынок в качестве ингибиторов тирозиновых киназ — Gleevec и Iressa [15]. Основные проблемы при поиске новых активных агентов связаны с необходимостью обеспечения высокой ингибирующей активности, селективности действия относительно определенных тирозиновых киназ, а также благоприятных ADMET характеристик.

Очевидным ресурсом на пути поиска новых высокоэффективных ингибиторов ТК является разработка и использование методов виртуального скрининга. В настоящей работе мы построили классификационную КССА модель активности органических соединений по отношению к тирозиновым киназам. Модель, основанная на алгоритме опорных векторов, обеспечивает высокое качество разделения между ингибиторами тирозиновых киназ и соединениями, не обладающими киназной активностью. Эта модель может быть полезна на ранних стадиях разработки лекарственных препаратов, действующих по механизму ингибирования тирозиновых киназ.

Литература

1. *Cohen P. // Curr. Opin. Chem. Biol. — 1999. — Vol. 3. — P. 459-465.*
2. *Kohonen T. Self-organizing maps. — Springer-Verlag: Heidelberg, 1996.*
3. *Sammon J.W. // IEEE Trans. Comp. — 1969. — Vol. C-18. — P. 401-409.*
4. *Devillers J. Neural Networks in QSAR and Drug Design. — Academic Press: London, 1996.*
5. *Vapnik V. Statistical Learning Theory. — New York: Whiley, 1998.*
6. *Burbidge R., Trotter M., Buxton B., Holden S. // Comput. Chem. — 2001. — Vol. 26. — P. 5-14.*
7. *Zernov V.V., Balakin K.V., Ivashchenko A.A. et al. // J. Chem. Inf. Comput. Sci. — 2003. — Vol. 43. — P. 2048-2056.*
8. *Cristianini N., Shawe-Taylor J. An Introduction to Support Vector Machines. — Cambridge University Press, Cambridge, UK, 2000.*
9. *Joachims T. In: Advances in Kernel Methods — Support Vector Learning, Scholkopf, B.; Burges, C.; Smola, A.; Eds.; MIT Press, 1999.*
10. *URL: http: //www.kernel-machines.org*
11. *Vapnik V., Chervonenkis A. // Automation and Remote Control. — 1974. — №8, 9. — P. 29.*
12. *Mutller K., Mika G., Ratsch G. et al. // IEEE Neutral Networks. — 2001. — Vol. 12. — P. 181-201.*
13. *Балакин К.В., Иваненков Я.А., Шкуренко Н.Е. и др. // ЖОФХ. — 2004. — Т. 2, вип. 3 (7). — С. 47-53.*
14. *Jolliffe I.T. Principal Comp. Analysis. — New York: Springer-Verlag, 1986.*
15. *Levitzki A. // Acc. Chem. Res. — 2003. — Vol. 36. — P. 462-469.*

Надійшла до редакції 01.02.2004 р.