

Методы нелинейного картирования для прогнозирования фармакологических свойств химических соединений

Я.А.Иваненков, Е.В.Бовина, К.В.Балакин

*Институт физиологически активных веществ Российской академии наук
142432 Черногловка Московской обл., Северный пр., 1, факс (496) 524–9508*

Рассмотрено современное состояние исследований в области разработки и применения специальной группы методов анализа данных — алгоритмов картирования, позволяющих снижать размерность многомерных пространств свойств анализируемых объектов (рядов химических соединений) до их двух- или трехмерных проекций (отображений) с сохранением топологии исходного пространства. Показано, как с помощью таких карт можно визуально анализировать сложные зависимости между свойствами соединений и их структурными особенностями, что используется для прогнозирования фармакологически значимых свойств химических соединений.

Библиография — 86 ссылок.

Оглавление

I. Введение	503
II. Современные технологии и значимость виртуального скрининга	504
III. Анализ химического пространства с использованием различных методов картирования	507
IV. Использование алгоритма SPE для анализа химического пространства	518
V. Специализированные компьютерные программы	521
VI. Заключение	522

I. Введение

Интенсивное развитие в XX веке методологии органического синтеза привело к существенному качественному и количественному росту новых химических соединений. Качественный рост выражается в том, что в настоящее время при наличии необходимых ресурсов исследователи могут синтезировать почти любое реально существующее соединение, включая сложные молекулы, характерные для живой материи. В частности, о колоссальных возможностях разработанных синтетических подходов свидетельствует получение в последние годы огромного числа природных соединений.¹ Количественный рост синтезированных соединений связан, в первую очередь, с достижениями комбинаторной химии.^{2,3} В этой сравнительно молодой дисциплине в последние годы достигнуты впечатляющие результаты.

Я.А.Иваненков. Инженер-исследователь лаборатории научно-информационных технологий в медицинской химии ИФАВ РАН. Телефон: (496)524–2563, e-mail: smart_people@inbox.ru

Е.В.Бовина. Кандидат химических наук, научный сотрудник той же лаборатории. Телефон: (496)524-2563, e-mail: bovin_a_e@ipac.ac.ru

К.В.Балакин. Доктор химических наук, заведующий той же лабораторией. Телефон: (496)524-2563, e-mail: balakin@ipac.ac.ru
Область научных интересов авторов: органическая и медицинская химия, химическая информатика.

Дата поступления 29 января 2009 г.

Комбинаторная химия ориентирована прежде всего на поиск новых лекарственных средств. С ее помощью удается получать большие ряды («библиотеки») сложных структур при минимальных затратах времени и средств. Методы комбинаторной химии постоянно совершенствуются, аккумулируя лучшие достижения традиционных методов органического синтеза. Это приводит к тому, что многие соединения со сложной структурой (например, молекулы с хиральными центрами, антибиотики, пептиды, олигонуклеотиды), синтез которых еще недавно осуществлялся только с использованием классических методов органической химии, сегодня могут быть успешно получены с применением высокопроизводительных схем комбинаторного синтеза.

В настоящее время комбинаторный синтез стал одним из ключевых методов органической химии и смежных дисциплин. В сочетании с методами высокопроизводительного экспериментального тестирования (скрининга), он составляет основу многих исследовательских программ в химико-фармацевтической и агрохимической промышленности. Однако в последнее время стало очевидным, что неограниченный рост размеров синтезируемых и тестируемых «библиотек» соединений сам по себе, без предварительного рационального отбора соединений-кандидатов для синтеза и последующих испытаний, не приводит к интенсификации исследовательских программ. Очевидно, что исчерпывающий синтез огромного количества соединений и их последующее тестирование являются чрезвычайно трудоемкими и экономически нецелесообразными процессами.^{4,5} Эти процессы еще более усложняются в случае проведения многостадийных синтезов, при использовании многокомпонентных

реакций и параллельных комбинаторных схем. Следовательно, даже в относительно простых случаях исследователю необходимо прибегать к методам рационального досинтетического отбора соединений-кандидатов. Эти методы позволяют существенно уменьшить размеры выборок соединений, обладающих требуемыми свойствами. Такой подход является необходимым условием успешной реализации многих научно-исследовательских программ.

Потребность в досинтетическом анализе еще не синтезированных виртуальных «библиотек» соединений с целью рационального отбора структур, наиболее перспективных для решения конкретных задач, привела к возникновению новой области науки — компьютерного, или виртуального, скрининга. Эта сравнительно молодая дисциплина приобретает все большую практическую значимость.^{4–6} Хотя начало систематических работ в области компьютерного моделирования свойств органических соединений относится еще к 50-м годам прошлого столетия,⁷ именно в наше время этому подходу отводится роль важнейшего инструмента планирования синтеза.

Виртуальный скрининг библиотек соединений включает целый ряд компьютерных методов и алгоритмов, позволяющих предсказывать многие принципиальные характеристики молекул, включая различные фармакокинетические и физико-химические параметры, а также исследовать различные взаимодействия в системе лиганд–биологическая мишень.

II. Современные технологии и значимость виртуального скрининга

Практическая ценность методов, составляющих основу виртуального скрининга (применительно к разработке новых лекарственных препаратов), заключается в возможности оценки потенциального воздействия низкомолекулярных органических соединений (потенциальных лекарственных препаратов) на различные белковые мишени, вовлеченные в процесс возникновения и развития заболевания. Именно это воздействие приводит к желаемому терапевтическому эффекту. Описанные в настоящем обзоре компьютерные модели в полной мере подпадают под понятие методов «виртуального скрининга», в том смысле, как оно трактуется в современной химико-фармацевтической промышленности.^{5, 8}

Наиболее значимые современные подходы к виртуальному скринингу представлены на рис. 1. Все они с некоторой долей условности могут быть разделены на две большие группы.⁹ В первую группу входят методы, позволяющие конструировать новые молекулы, исходя из структур уже известных активных соединений (Structure-based design). Вторая группа методов исходит из структуры биологической мишени (Target-based design).

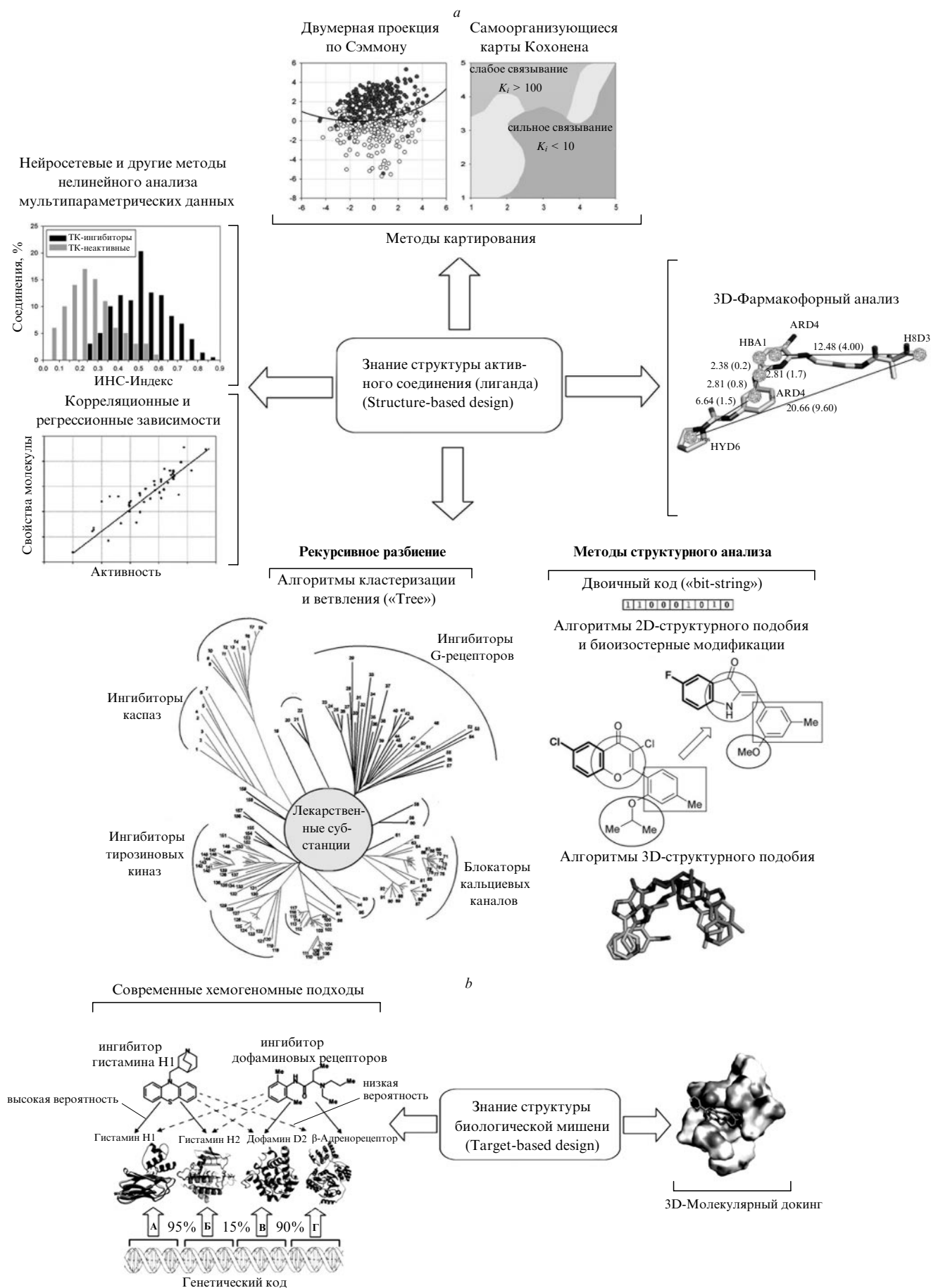
Различные алгоритмы отбора соединений по принципу их структурного подобия известным лекарственным препаратам включают поиск по фрагменту структуры,¹⁰ по коэффициенту подобия малых структурных фрагментов молекулы, так называемых «фингерпринтов» (от англ. fingerprints — отпечатки пальцев),^{11, 12} а также поиск биоизостерных^{13–16} аналогов или соединений, обладающих сходным профилем «многоуровневого атомного окружения».^{17–19} К принципиальным недостаткам методов, основанных на структурном подобии, можно отнести тяготение к уже известным химическим типам соединений.

К современным, более мощным, методам первой группы относятся различные алгоритмы анализа мультипараметрических данных, такие как самоорганизующиеся карты Кохонена,²⁰ нелинейные карты Сэммона,²¹ искусственные нейронные сети,^{22, 23} метод опорных векторов (МОВ),²⁴ алгоритмы иерархической кластеризации и ветвления, генетические алгоритмы, а также различные регрессионные и корреляционные зависимости. Перечисленные выше алгоритмы, в отличие от прямых методов структурного анализа, основаны на поиске неявных зависимостей в системе молекулярные дескрипторы–свойство(активность). Такие подходы характеризуются достаточной точностью, эффективностью и производительностью.

В современных исследовательских проектах широко используются методы поиска новых лекарственных препаратов, основанные на оценке их трехмерного (3D) структурного подобия уже известным физиологически активным молекулам.²⁵ Широкое распространение получило 3D-фармакофорное моделирование.^{26, 27} В этом методе в основу поиска структур со специфическим типом активности положена 3D-фармакофорная гипотеза, согласно которой отбор трехмерных моделей ведется по принципу их пространственного соответствия найденным ранее фармакофорным элементам, таким как доноры или акцепторы водородной связи, области с суммарным положительным или отрицательным зарядом, гидрофобные поверхности и т.п. Эти методы позволяют предлагать новые структурные хемотипы на базе уже известных активных веществ, однако они требуют больших вычислительных ресурсов.

Во вторую группу входит гораздо меньшее количество методов, в основном ввиду того что для их реализации требуется знание структуры исследуемой биологической мишени — особенностей ее строения, формы и состава активного сайта связывания, — а также особенностей взаимодействия известного лиганда с конкретной биомишенью. Наиболее популярным в этой группе методов является 3D-молекулярный докинг.^{28–30} Не отличаясь особой производительностью, этот метод во многих случаях позволяет более точно охарактеризовать взаимодействие лиганд–мишень и на основании полученных данных построить трехмерную компьютерную модель, которая в дальнейшем может быть использована для прогнозирования физиологической активности органических соединений.

Теоретически молекулярный докинг является наиболее адекватным способом оценки потенциальной мишень-специфической активности соединений. Однако его применение в реальных задачах виртуального скрининга в настоящее время сопряжено с целым рядом теоретических и практических проблем, резко ограничивающих возможности его прикладного использования. Построение трехмерной модели комплекса лиганда с биомишенью является чрезвычайно ресурсоемкой задачей даже при наличии доступа к высокопроизводительным вычислительным системам, а учитывая недостаточную изученность многих аспектов взаимодействия лиганда с мишенью (например, динамики взаимодействия, конформационной подвижности лиганда и мишени, влияния молекул растворителя и других веществ, находящихся в растворе, на связывание лиганда с мишенью и т.д.), можно предположить, что качество предсказания мишень-специфической активности лиганда с использованием молекулярного докинга будет ненамного лучше, чем предсказание, основанное на менее ресурсоемких методах двумерного структурного подобия.³⁰ Не стоит забывать и о том, что на сегодняшний день в подавляющем большинстве



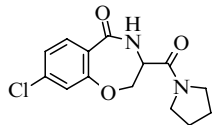
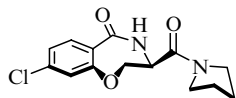
компьютерных программ вообще не учитывается конформационная подвижность белков (мишеней), т.е. белки рассматриваются как стационарные системы. Учитываются лишь незначительные структурно-энергетические изменения в области непосредственного связывания лиганда с белком, что, очевидно, не может в полной мере отражать все аспекты взаимодействия в таких системах. В значительной мере многое из сказанного справедливо и для других методов анализа трехмерных структур, связанных с предварительным нахождением трехмерной конформации молекулы с минимальной потенциальной энергией.

К рассматриваемой категории компьютерных методов можно отнести и сравнительно молодую, но крайне перспективную научную дисциплину — хемогеномику.³¹ В отличие от всех остальных методов виртуального скрининга, хемогеномика оперирует в первую очередь геномной информацией. В основу методов хемогеномики положена концепция, в рамках которой делается предположение о том, что подобные гены кодируют подобные белки, а подобные белки связываются с одним и тем же лигандом или группой лигандов (в упрощенной формулировке). Таким образом, выбрав какое-либо известное лекарственное вещество и зная биологическую мишень (или мишени), на которую это вещество действует, можно отобрать другие клеточные биомишени, на которые, вероятно, рассматриваемое соединение тоже будет действовать. В результате удастся выяснить функциональный профиль действия лекарственного вещества, а также определить и уточнить внутриклеточные сигнальные каскады и сопутствующие клеточные ответы. Все сказанное в полной мере относится и к новым активным соединениям, которые находятся на стадии исследований.

Важным достоинством большинства методов виртуального скрининга, основанных на знании структуры биологической мишени, является их применимость в условиях, когда известно лишь малое число лигандов, активных по отношению к рассматриваемой мишени, или даже когда такие лиганды вообще неизвестны. Однако в последнем случае для рационального отбора лигандов необходимо знать структуру активного центра биомишени.

Задачи, решаемые с применением различных методов виртуального скрининга, весьма разнообразны. Выбор того или иного алгоритма компьютерного анализа зависит в равной степени от наличия необходимой информации, которую в случае использования первой группы методов аккумулируют из различных источников, а также от специфических особенностей самого метода. Например, для использования нейронных сетей и методов картирования необходимо наличие, по крайней мере, двух обязательных факторов — обучающей выборки и ряда рассчитанных или

Таблица 1. Пример классификации молекулярных дескрипторов по их «размерности».

Категория	Исходная структура для расчета	Примеры дескрипторов
Одномерные (1D)	Брутто-формула: $C_{14}H_{15}ClN_2O_3$	Молекулярный вес, число атомов углерода, число атомов кислорода и т.д.
Двумерные (2D)	Двумерная структурная формула 	Топологические индексы, $\lg P$, структурные фрагменты и т.д.
Трехмерные (3D)	Трехмерная структура в конформации с минимальной расчетной энергией 	Радиус Ван-дер-Ваальса, площадь полярной поверхности, дипольный момент и т.д.

экспериментально определенных параметров. В роли рассчитанных параметров выступают молекулярные дескрипторы (табл. 1), такие как логарифм коэффициента распределения вещества в системе октанол – вода ($\lg P$), количество доноров и акцепторов водородной связи, количество свободно вращающихся связей, топологические и электротопологические индексы, ван-дер-ваальсовы радиусы атомов и т.д. Подобного рода параметры рассчитываются исходя из структуры химического соединения при помощи специальных компьютерных программ, о которых будет сказано ниже. Молекулярные дескрипторы используются для построения компьютерных моделей. Необходимая информация для формирования обучающей выборки может быть извлечена из соответствующих коммерческих баз данных (табл. 2), в которых, как правило, помимо структурной формулы вещества приводится ряд его принципиальных характеристик, например основная биологическая мишень, стадия биологического тестирования, некоторые молекулярные дескрипторы и экспериментальные параметры.

Среди разнообразных направлений компьютерного скрининга следует особо отметить направление, связанное с построением моделей, которые описывают количественную связь между структурой и активностью (QSAR) или структу-

Таблица 2. Специализированные базы данных, использованные в настоящей работе для создания обучающих выборок.

Название базы данных	Вид информации	Разработчик
Prous Ensemble	Структуры и фармакологически значимые свойства соединений	Prous Science, Испания (URL: http://www.prous.com)
Beilstein Cross-Fire	Структуры и физико-химические свойства соединений	Beilstein, США (URL: http://www.beilstein.com)
MetaDrug	Структуры и метаболические свойства соединений	GeneGo Inc., США (URL: http://www.genego.com)
ToxSys	Структуры и токсические свойства соединений	SciVision Inc. & ToxSys, США (URL: http://www.toxsys.com)

рой и свойствами (QSPR) химических соединений. Именно при помощи QSAR/QSPR-моделирования построены все прогностические системы, описанные в настоящей работе. В общем случае QSAR/QSPR-моделирование состоит из следующих принципиальных стадий: 1) подготовка модельного эксперимента; 2) построение QSAR/QSPR-модели; 3) тестирование и подтверждение построенной модели. Более подробно все эти стадии будут рассмотрены далее при обсуждении методов картирования.

III. Анализ химического пространства с использованием различных методов картирования

Для анализа мультипараметрических данных помимо искусственных нейронных сетей применяются также различные методы картирования, позволяющие получать двумерные или трехмерные отображения многомерных пространств свойств исследуемых выборок соединений. Эти методы, как правило, относят к более общей категории алгоритмов понижения размерности многомерного пространства признаков. В настоящее время из методов картирования наиболее часто используются самоорганизующиеся карты Кохонена,^{21, 22} нелинейные карты Сэммона^{32, 33} и SPE-метод (SPE — Stochastic Proximity Embedding).³⁴ Существуют и другие методы, при помощи которых удается получать двумерные отображения исходного пространства признаков, например алгоритмы кластеризации,³⁵ диффузионные карты,³⁶ алгоритм ИЗОМЭП.³⁷ Ниже будет рассмотрен ряд примеров эффективного применения основных компьютерных алгоритмов картирования в сфере анализа химического пространства.

1. Самоорганизующиеся карты Кохонена

Самоорганизующиеся карты Кохонена — современный мощный инструмент для анализа и визуализации многопараметрических данных высокой размерности и объема, позволяющий создавать двумерные или трехмерные отображения многомерных пространств свойств.²¹ Метод основан на известном принципе обучения «победитель забирает все» и специальном правиле обучения нейронов решетки по Кохонену. Самоорганизация достигается за счет пошагового (итерационного) обучения каждого нейрона сети с учетом его активационной функции. В качестве активационной функции

часто используется нормальное распределение Гаусса. В результате такого обучения нейроны решетки постепенно упорядочиваются в кластерно-связанную структуру, отражающую в пределах допустимой ошибки на двумерной плоскости или в трехмерном объеме закономерности и топологию, обнаруженные в исходном многомерном пространстве признаков. Этот метод находит все более широкое применение в различных областях органической и медицинской химии, включая предсказание физиологических и физико-химических свойств соединений (см., например,^{22, 38–41}).

а. Компьютерный дизайн мишень-специфических библиотек органических соединений

Одной из наиболее важных задач современной медицинской компьютерной химии является исследование нековалентных взаимодействий в системе лиганд–рецептор (мишень). В частности, актуальной задачей является выяснение возможного механизма действия лекарственного препарата на биологическую мишень, взаимодействие с которой приводит к терапевтически значимому физиологическому ответу. Благодаря высокой производительности компьютерные модели (*in silico*), позволяющие предсказывать мишень-специфическую активность органических соединений, являются незаменимыми средствами для анализа больших массивов соединений. Обычно еще до стадии непосредственного биологического тестирования виртуальные библиотеки соединений «фильтруют» при помощи специальных QSAR/QSPR-моделей, предсказывающих мишень-специфическую активность соединений. Примеры некоторых типичных классификационных задач, которые решаются на этапе рационального отбора соединений с заданной мишень-специфической активностью, представлены в табл. 3.

Однако следует отметить, что для успешного использования классификационных статистических QSAR/QSPR-моделей необходимо существование достаточно большой обучающей выборки соединений, проявивших заданную активность или свойство. Это требование ограничивает применимость такого подхода в виртуальном скрининге. Отличительная особенность алгоритма Кохонена состоит в том, что он позволяет в рамках одной QSAR/QSPR-модели предсказывать эффективность связывания лиганда не с одной, а с целым рядом терапевтически значимых биологических

Таблица 3. Типичные задачи в области виртуального скрининга.

Задача	Обучающая база данных	Ссылки
Классификация соединений по их способности ингибировать фосфодиэстеразу V (ФДЭ-V)	61 ингибитор ФДЭ-V и 1739 соединений, неактивных по отношению к ФДЭ-V	42
Классификация соединений по их способности ингибировать ацил-CoA-холестерин-О-ацилтрансферазу (АКАТ)	446 ингибиторов АКАТ и 1771 соединение, неактивное по отношению к АКАТ	42
Классификация соединений по их активности по отношению к эстрогеновым рецепторам (ЭР)	463 аналога эстрогена с различными уровнями антиэстрогенной активности	43
Классификация соединений по их активности по отношению к девяти различным G-белок-связанным рецепторам (GPCR)	3365 соединений, активных по отношению к девяти различным GPCR	44
Классификация соединений по их способности ингибировать карбоновую ангидразу II (КА-II)	337 соединений с различными уровнями ингибирующей активности по отношению к КА-II	45
Классификация соединений по способности проявлять антибактериальную активность	661 соединение, обладающее или не обладающее антибактериальной активностью	46

мишеней. Теоретическим и практическим аспектам использования самоорганизующихся карт Кохонена посвящены многочисленные публикации (см., например, работы^{47–51} и ссылки в них).

В работе⁵² описана уникальная компьютерная модель на базе алгоритма самоорганизующихся карт Кохонена, при помощи которой удалось эффективно разделить группы лекарственных веществ, обладающих различным профилем биологической активности. В качестве тренирующей выборки была использована база данных Prous Ensemble, состоящая из 18 703 соединений с экспериментально установленной биологической активностью. В нее входят как известные лекарственные средства, так и соединения, находящиеся на стадии клинических или доклинических испытаний. Для каждого соединения из этой базы данных известна его биологическая мишень, принадлежащая одному из пяти терапевтически значимых семейств белков: 1) тирозинкиназам; 2) рецепторам ионных каналов; 3) рецепторам, связанным с G-белками (GPCR-рецепторам); 4) фосфодиэстеразам; 5) протеазным ферментам. Для каждого соединения были рассчитаны пять молекулярных дескрипторов: молекулярная масса (MW), логарифм коэффициента распределения вещества в системе октанол–вода ($\lg P$), число свободно вращающихся связей (RB), число доноров (HBD) и акцепторов (HBA) водородной связи. Эти параметры относятся к классическим параметрам QSAR/QSPR-анализа; они напрямую связаны с такими важными свойствами молекулы, как размер, липофильность, конформационная подвижность, способность к образованию водородных связей.

После стадии предподготовки и фильтрации обучающей выборки для нее была построена карта Кохонена (10 × 10). Обучение сети проводилось независимо от класса соединений (неконтролируемая тренировка), что составляет главное отличие данного алгоритма от стандартных методов «обучения с учителем». После построения модели все соединения были распределены по группам, соответствующим перечисленным выше пяти мишень-специфическим активностям. Все вычисления были произведены с использованием программы Informagenesis (URL: <http://www.Informagenesis.com>), о которой будет сказано ниже. В результате была получена двумерная проекция пятимерного пространства свойств, причем объекты, расположенные в непосредственной близости друг от друга на двумерной карте Кохонена, были близкими соседями и в пятимерном пространстве. После проведенной процедуры тренировки алгоритма Кохонена была получена общая карта распределения свойств соединений из полной исходной обучающей выборки (рис. 2).

Из рисунка видно, что соединения распределены по всей карте. На карте отчетливо видны области неправильной формы, в которых сконцентрировано большое число лекарственных веществ, и, наоборот, области, в которые попало лишь незначительное число соединений. Различия в заселенности отдельных областей может превышать величину двух порядков.

После построения общей карты на ней были выявлены области распределения соединений, связывающихся с каждой из пяти исследуемых мишеней (рис. 3, а–е). Видно, что соединения из разных групп располагаются в разных областях карты. Например, области расположения соединений, активных по отношению к GPCR-рецепторам (рис. 3, с), почти не пересекаются с областями расположения ингибиторов протеазных ферментов (рис. 3, е). Для всех пяти групп соединений можно выделить области повышенной заселенности, а также незаселенные области. Средний процент раз-

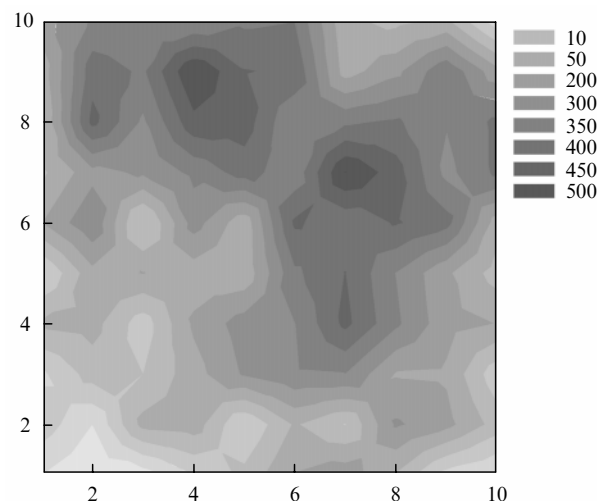


Рис. 2. Карта Кохонена (10 × 10) полной обучающей выборки лекарственных веществ (18 703 соединения).

деления мишень-специфических групп на построенной карте составил 80%, что вполне приемлемо с точки зрения общей концепции рационального отбора соединений и математической статистики.

Предложенный в работе⁵² алгоритм во многих случаях позволяет выявлять тонкие различия в свойствах разных групп объектов (в частности, наличие кластеров), а его совместимость с высокопроизводительными методами виртуального скрининга дает возможность тестировать с помощью построенной карты Кохонена большие виртуальные библиотеки соединений в реальном времени. Разработанная модель может оказаться весьма полезной при планировании комбинаторного синтеза, так как исследованные пять групп биомишеней являются наиболее типичными в современной фармакологии: на них действует не менее 60% известных коммерческих лекарственных препаратов. При помощи построенной карты Кохонена можно отбирать наиболее эффективные соединения (находящиеся в высокозаселенных областях) с целью их дальнейшего скрининга на аналогичных белковых системах.

На основании полученных результатов можно сделать вывод о том, что в исследуемом пятимерном пространстве свойств имеются компактные регионы, соответствующие определенным мишень-специфическим активностям. Построенная карта Кохонена, являющаяся двумерным отображением этого пятимерного пространства, позволяет уловить различия в свойствах разных групп активных соединений. Авторы работы⁵² высказали предположение относительно причин подобного статистического различия свойств. Как правило, у биомишеней одного семейства активные места связывания низкомолекулярных лигандов имеют более или менее сходное строение, что вытекает из гомологичности аминокислотных последовательностей. Сходство строения активных мест определяется близостью их пространственных характеристик, сходной липофильностью, наличием одних и тех же положительно или отрицательно заряженных групп и т.п. Как следствие, низкомолекулярные лиганды, способные эффективно связываться с такими активными местами, также должны обладать определенной степенью сходства. Однако уловить это сходство, носящее вероятностный характер и зависящее от специфических особенностей

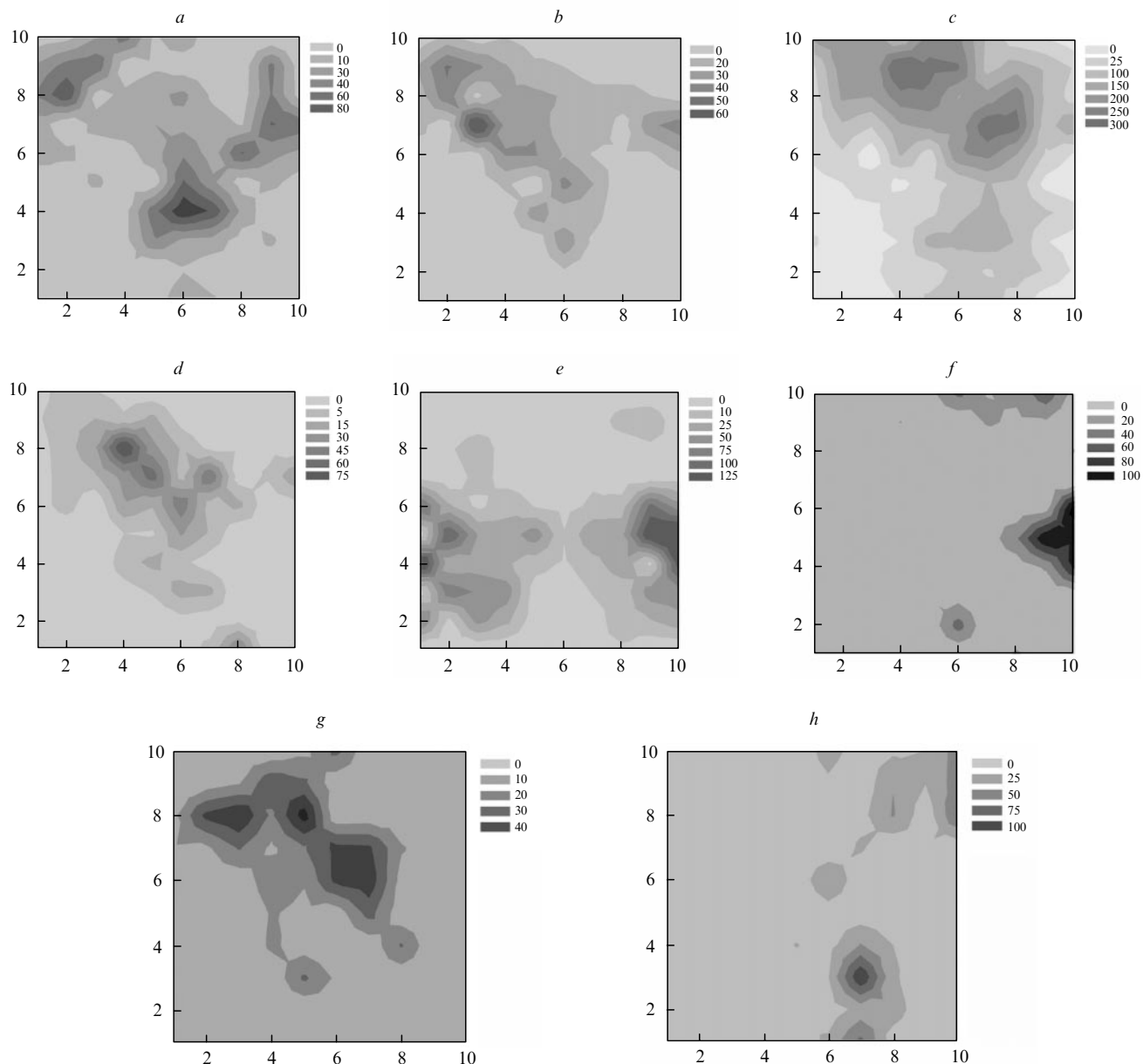


Рис. 3. Области распределения восьми мишень-специфических групп активных соединений на карте Кохонена (10 × 10). *a* — ингибиторы тирозинкиназ (1400 соединений); *b* — модуляторы рецепторов ионных каналов (1080 соединений); *c* — модуляторы активности GPCR-рецепторов (12 660 соединений); *d* — ингибиторы фосфодиэстераз (730 соединений); *e* — ингибиторы протеазных ферментов (2833 соединения); *f* — антагонисты/агонисты ядерных рецепторов (1021 соединение); *g* — активаторы калиевых каналов (1060 соединений); *h* — антагонисты кальциевых каналов (1230 соединений). Примеры *a–e* взяты из работы ⁵², *f–h* — из работы ⁵⁰.

конкретного члена данного семейства, можно лишь при наличии достаточно больших обучающих выборок.

Альтернативный вариант самоорганизующейся карты Кохонена был предложен в работе ⁴¹. В этом случае полная обучающая выборка была составлена из известных лекарственных препаратов (2105 агентов орального применения), отобранных из базы данных World Drug Index, ⁵³ и из более чем 52 500 нелекарственных соединений, взятых из базы данных Maybridge. ⁵⁴ Для каждого соединения были рассчитаны 26 молекулярных дескрипторов, после чего с использованием алгоритма анализа главных компонент (PCA) их число было уменьшено до трех: QsumN (суммарный заряд атомов водорода), Espmin (минимальный электростатический потенциал) и covNBac (кислотность ковалентных водо-

родных связей). На основании этих трех дескрипторов была построена общая карта Кохонена (100 × 100), на которой поочередно были выявлены области, соответствующие лекарственным и нелекарственным соединениям (рис. 4, *a*). Обучение сети проводилось в два этапа, что позволило получить более точное распределение объектов на карте (второй этап обучения часто называют этапом тонкой настройки нейронов Кохонена). Для обучения использовали следующие параметры: † количество итераций (время обуче-

† Радиус обучения определяет, сколько нейронов, кроме нейрона-победителя, участвует в обучении (в коррекции весов) на данной итерации; скорость обучения — величина коррекции весов нейронов на этой итерации.

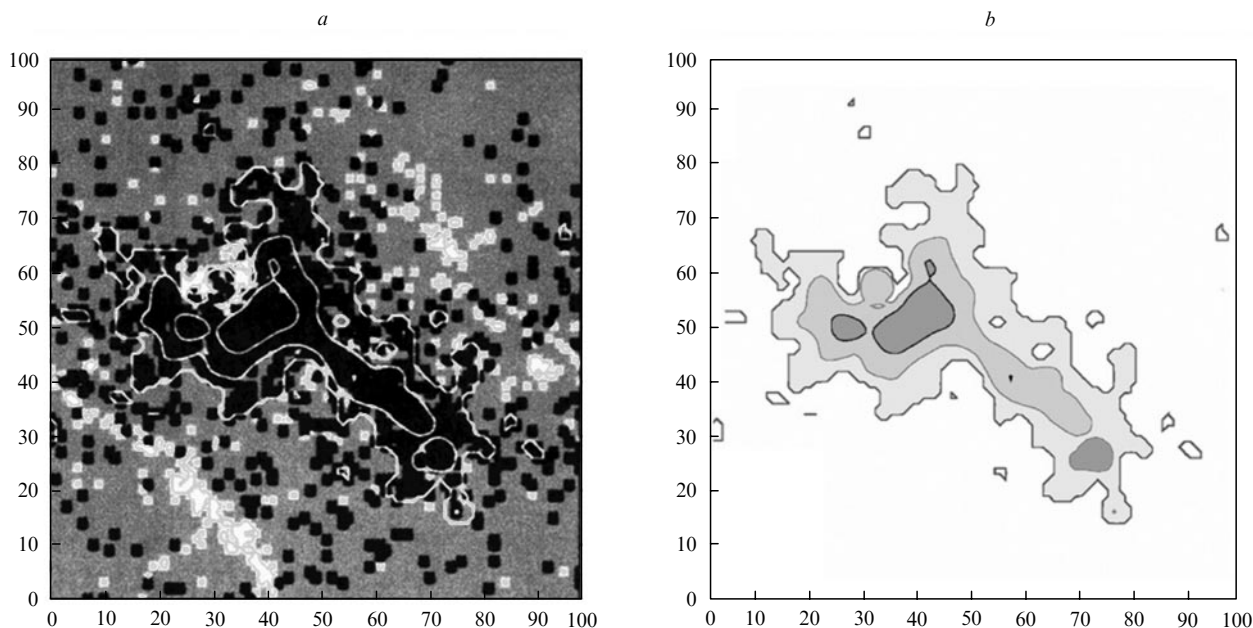


Рис. 4. Карты Кохонена (100×100), построенные для более чем 54 000 органических соединений. Черные области на карте (a) соответствуют лекарственным веществам, серые — нелекарственным; контурными линиями на карте (b) показано количественное распределение лекарственных соединений (%): градиент серого цвета отражает количество соединений, приходящихся на один нейрон Кохонена (от 0.2 — светло-серый до 0.8 — темно-серый). За внешней границей контурных линий концентрация лекарственных веществ не превышает значений случайного распределения.

ния) на первой стадии обучения — 10 000, начальное значение радиуса обучения — 200, начальное значение скорости обучения — 0.03; количество итераций на второй стадии обучения — 100 000, радиус обучения — 40, начальная скорость обучения — 0.01.

На приведенной на рис. 4,a карте Кохонена наибольшая часть лекарственных соединений сосредоточена в центральной области. Эта область имеет форму неправильного эллипса и занимает около одной четверти всей площади карты (рис. 4,b). На основании полученных данных можно говорить о том, что в той системе молекулярных дескрипторов, которая была использована для моделирования, все отобранные соединения обладают довольно схожими харак-

теристиками, что неудивительно в рамках фундаментальной концепции лекарственного подобия.

Построенную карту Кохонена авторы работы⁴¹ применили для изучения распределения каждого из трех дескрипторов, используемых при ее построении. В результате они получили три отдельные градиентные карты (рис. 5), из рассмотрения которых следует, что разные молекулярные дескрипторы имеют различные направленности вектора распределения значений, однако все они вносят свой вклад в формирование эллиптической области, в которой располагаются лекарственные вещества.

Таким образом, авторам работы⁴¹ на основании всего лишь трех молекулярных дескрипторов удалось построить

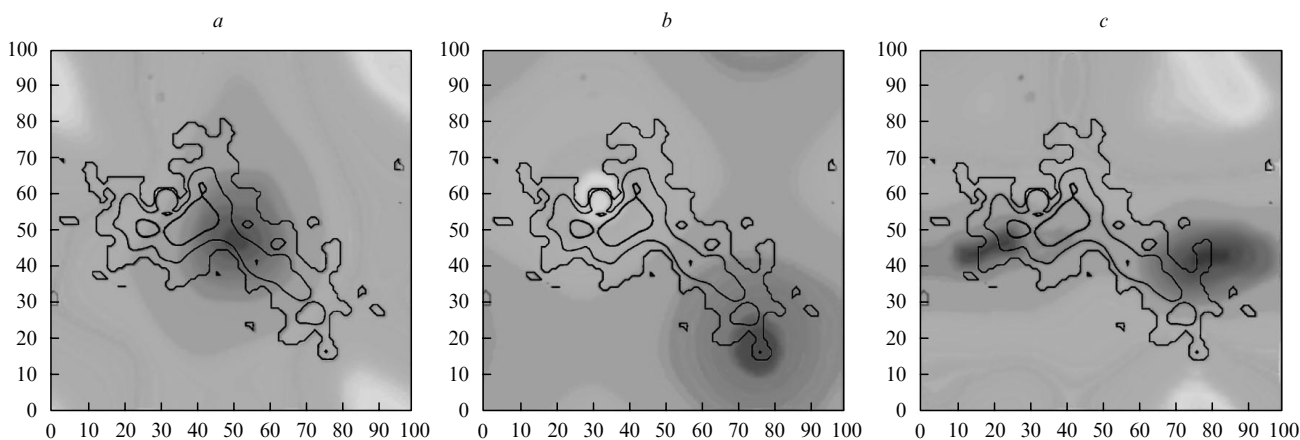


Рис. 5. Градиентные карты Кохонена (100×100), показывающие (на фоне распределения лекарственных веществ) распределение по значениям трех дескрипторов — QsumH (a), Espmin (b), covHBac (c), — использованных для построения карты Кохонена (рис. 4,a). Светло-серый цвет соответствует минимальным значениям дескрипторов, черный — максимальным.

компьютерную модель, при помощи которой можно прогнозировать степень подобия исследуемых соединений известным лекарственным веществам.

б. Компьютерная оценка возможных токсических эффектов

Известно, что нежелательные токсические эффекты являются одним из наиболее распространенных факторов, негативным образом влияющих на внедрение новых лекарственных препаратов и агрохимикатов. Как следствие, создание и внедрение в практику надежных методов раннего предсказания токсичности может принести существенный экономический эффект. Вместе с тем предсказание потенциальной токсичности химических соединений является одной из наиболее трудно решаемых задач в современных системах анализа данных. В литературе описано множество подходов к решению этой задачи (см., например, недавние обзоры^{55–58}), но до сих пор их применение было ограничено либо локальным характером рассматриваемых эффектов, либо неполнотой структурных химических типов, представленных в обучающей выборке.

Отдельные алгоритмы предсказания токсичности основаны на использовании больших баз данных, аккумулирующих сведения о структурах токсичных соединений и их токсическом действии. Статистический анализ корреляций между токсичностью соединения и наличием в его молекуле определенных структурных элементов позволяет в ряде случаев делать точные прогнозы относительно характера и интенсивности токсического воздействия. По этому принципу построены наиболее известные в токсикологии прогностические системы, такие как DEREK, HazardExpert, ТОРКАТ.^{56, 57} У подобного подхода есть ряд принципиальных недостатков. Так, предсказание основано на информации, содержащейся в конкретной базе данных, каждая из которых в той или иной степени страдает от неполноты и неточности сведений, их исторически обусловленной направленности, разнородности источников данных и экспериментальных условий определения токсичности, отсутствия сведений о новых хемотипах соединений и т.п. Кроме того, подобные экспертные системы, как правило, не учитывают влияния физико-химических факторов на токсикокинетический профиль соединений.

Главным объективным фактором, сдерживающим создание надежных алгоритмов для предсказания токсичности, является множество видов и механизмов токсического действия. За редким исключением, при QSAR/QSPR-моделировании токсичности не удается определить ни конкретной биомишени, ни конкретного принципа токсического действия. Это обстоятельство затрудняет проведение исследований, поскольку с необходимостью обуславливает неоднородность пространства свойств химических соединений — потенциальных токсифоров. Для решения реальных практических задач, связанных, например, с ранним определением потенциальной токсичности низкомолекулярных органических соединений, необходимо, чтобы предсказательная модель могла работать с разнообразными структурными хемотипами, а не с одним определенным классом соединений. Отсутствие таких моделей осложняет предсказание токсичности.

В этих условиях для предсказания токсичности соединений приходится ограничиваться локальными моделями,

разработанными для конкретных мишеней. Можно предположить, что эволюция моделей со временем приведет к накоплению необходимых знаний и послужит толчком к созданию интегральных высокоэффективных систем предсказания токсичности.

Одна из таких локальных моделей была разработана для предсказания токсичности соединений в водных растворах.⁵⁹ Эта модель позволяет прогнозировать токсические эффекты широкого круга низкомолекулярных органических соединений по отношению к толстоголовым гольянам (*Pimephales promelas*). Она является чрезвычайно полезным прикладным инструментом исследований водной экологии, в частности, она позволяет предсказывать биоаккумуляцию токсичных веществ в организмах этих морских животных. При разработке такой QSAR/QSPR-модели использовалась уникальная экспериментальная база данных Duluth,⁶⁰ в которой, в частности, содержатся сведения об органических соединениях, обладающих определенной степенью токсичности по отношению к изучаемому виду рыб.

Обучающая выборка состояла из 562 органических соединений, специальным образом отобранных из базы данных Duluth. Для всех соединений из обучающей выборки были рассчитаны 169 индивидуальных молекулярных дескрипторов. Для рациональной селекции наиболее значимых дескрипторов авторы использовали алгоритм Кохонена. Отбор производили на основании простого правила, согласно которому из каждого кластера на результирующей карте Кохонена отбирали всего один дескриптор, распределение значений которого наилучшим образом характеризовало распределение соединений из обучающей выборки. В результате было показано, что для описания токсичности в условиях проводимого эксперимента достаточно использовать три дескриптора: логарифм коэффициента распределения вещества в системе октанол–вода, молекулярный вес, а также топологический индекс Кира и Холла, характеризующий строение молекулы.

На рис. 6,а показано распределение значений $\lg P$ на карте Кохонена (30 × 30). Для построения этой карты использовали 800 обучающих циклов (итераций). Как видно из рисунка, распределение значений $\lg P$ практически полностью соответствует распределению токсичности органических соединений по отношению к толстоголовым гольянам, т.е. чем выше значение молекулярного дескриптора $\lg P$, тем более токсично соединение. Этот результат подтвердил найденную ранее в ряде экспериментов закономерность о взаимосвязи липофильности некоторых органических соединений с их токсичностью по отношению к *Pimephales promelas*.

Существующие на сегодняшний день линейно-корреляционные зависимости $\lg P$ –токсичность не позволяют с достаточной точностью предсказывать токсичность соединений, в то время как построенная карта Кохонена обладает более высокой предсказательной способностью и может быть использована для классификации органических структур по степени их токсичности по отношению к толстоголовым гольянам, а также другим морским животным. Предложенная QSAR/QSPR-модель позволяет в автоматическом режиме тестировать большие виртуальные библиотеки соединений (вплоть до десятков миллионов молекулярных объектов), отбирая те из них, которые обладают наименьшей потенциальной токсичностью.

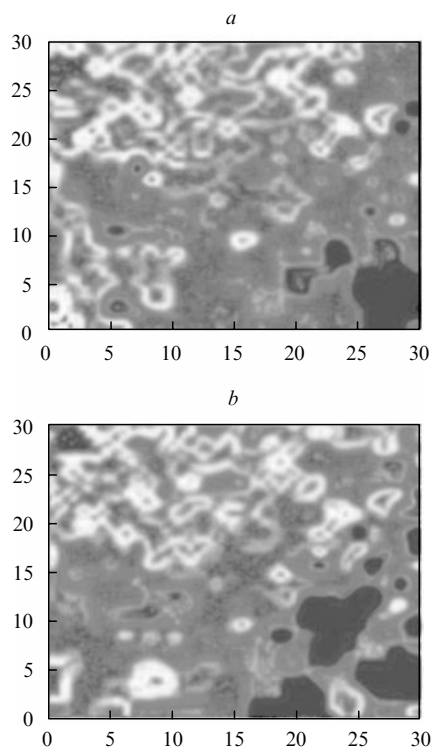


Рис. 6. Распределение значений $\lg P$ органических соединений из базы данных Duluth на карте Кохонена (30×30) (a), и токсичность этих соединений в водном растворе по отношению к толстоголовым гольянам (*Pimephales promelas*) (b).

Белые области соответствуют низким значениям $\lg P$ и токсичности, темно-серые — высоким.

в. Анализ противоопухолевой активности лекарственных веществ

Разработка лекарственных препаратов, обладающих противоопухолевой активностью, — одно из наиболее важных и востребованных направлений медицинской химии. На сегодняшний день множество различных органических соединений, включая природные белковые комплексы, а также

разнообразные низкомолекулярные органические соединения, находятся на различных стадиях биологического тестирования на противоопухолевую активность по отношению к различным злокачественным новообразованиям. Ряд препаратов уже выпущен на мировой фармацевтический рынок. В соответствии с классической схемой разработки лекарственных веществ, на начальной стадии дизайна новых противоопухолевых препаратов довольно часто применяют различные методы компьютерного моделирования, в частности алгоритм самоорганизующихся карт Кохонена.

Так, в работе⁴² описана компьютерная модель, позволяющая предсказывать наличие той или иной противоопухолевой активности у органических соединений. В качестве обучающей выборки использовались вещества, отобранные из базы данных Национального института рака США (National Cancer Institute, NCI). Эта выборка включала в себя 362 лекарственных соединения с экспериментально установленной степенью активности по отношению к различным типам опухолевых клеток. После стадии расчета и рационального отбора наиболее значимых молекулярных дескрипторов была построена прямоугольная карта Кохонена (17×9). На этой карте были выявлены области, в которые попадают соединения с различным типом противоопухолевой активности (рис. 7).

Видно, что различные классы противоопухолевых препаратов локализованы в разных областях карты. Пунктирной линией выделены области, в которые попадают соединения, действующие на различные стадии митоза (область А), и соединения, основной терапевтической мишенью которых является тубулин (область В). Степень разделения модельных соединений превысила 77%. Таким образом, авторам работы⁴² удалось построить компьютерную модель, которая в дальнейшем может быть использована для предсказания противоопухолевой активности органических соединений.

2. Нелинейные карты Сэмона

Помимо самоорганизующихся карт Кохонена существуют и другие алгоритмы снижения размерности многомерных пространств признаков, результатом действия которых является создание двух- и трехмерных карт, удобных для визуального анализа. Одним из таких алгоритмов является метод дву-

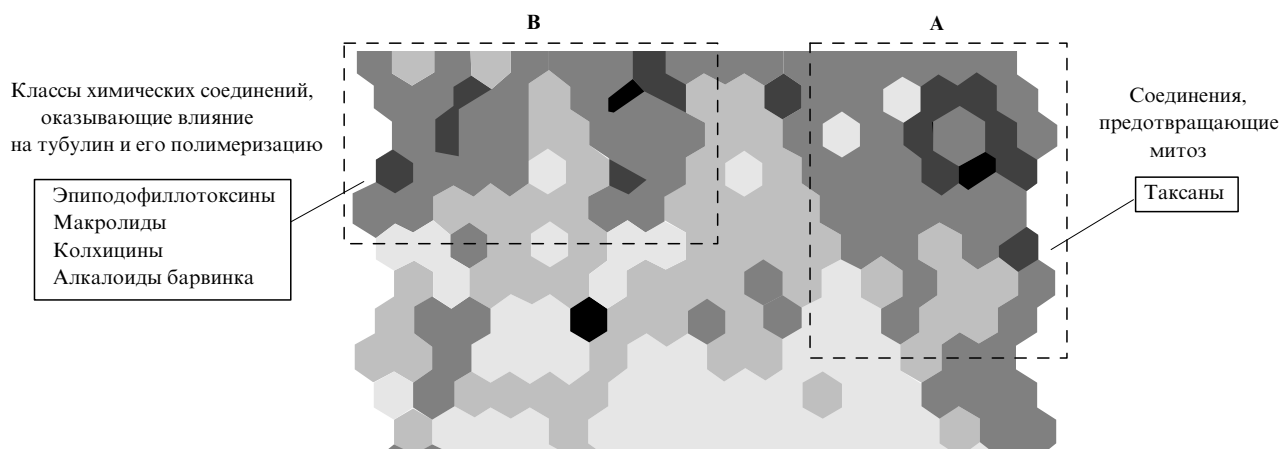


Рис. 7. Карта Кохонена (17×9 , гексагональная ячейка), построенная для 362 лекарственных препаратов с противоопухолевой активностью.

мерного отображения исходного пространства признаков по Сэммону.^{32, 33} Применительно к компьютерной медицинской химии алгоритм Сэммона позволяет создавать двумерное отображение многомерного пространства свойств биологически активных соединений, взятых из экспериментальных баз данных. В результате исследователь получает удобное средство визуального анализа пространства свойств, кодируемых молекулярными дескрипторами. Например, он может визуальнo выявлять зоны преимущественной локализации соединений, принадлежащих к интересующим его классам. (В некоторых случаях для этой цели используют метод опорных векторов или алгоритм, предложенный в работе⁶¹.)

Необходимо отметить, что в отличие от карт Кохонена, корректно отображающих в пространстве более низкой размерности лишь кластерную топологию исходного пространства, метод Сэммона позволяет сохранить общую топологию многомерного пространства, где расстояния между объектами (векторами, отображающими эти объекты) в проекции более низкой размерности соответствуют (в пределах допустимой ошибки аппроксимации) исходным расстояниям (например, евклидовым). Алгоритм Сэммона находит применение при решении различных задач компьютерной медицинской химии.

а. Моделирование процесса проникновения органических соединений через липидные мембраны

Рассмотрим применение метода картирования по Сэммону для предсказания степени проникновения органических соединений через липидные мембраны — гематоэнцефалический барьер (ГЭБ) и стенки желудочно-кишечного тракта (ЖКТ).[‡] В недавно опубликованной работе⁶² описан способ построения карты Сэммона для предсказания указанного выше свойства.

В ходе работы авторам удалось построить общую карту Сэммона, включающую в себя две обучающие выборки — проникновение соединений через ГЭБ и стенки ЖКТ. В качестве входного сигнала они использовали набор молекулярных дескрипторов — логарифм коэффициента распределения вещества в системе октанол–вода, количество доноров и акцепторов водородной связи и количество свободно вращающихся связей С–С. С помощью построенной карты Сэммона была проведена оценка степени разделения двух исследуемых категорий соединений — хорошо проникающих (обозначены «+») и плохо проникающих («–») через ГЭБ и стенки ЖКТ — путем поочередной визуализации этих категорий и проведения между ними разделяющей линии по методу опорных векторов.²⁵ Эта линия определяет оптимальную функцию разделения объектов в двумерных координатах. На построенной карте Сэммона (рис. 8) соединения, хорошо и плохо проникающие через липидные мембраны, попадают, как правило, в разные зоны.

Для обеих мембран степень разделения приближается к 90% (табл. 4); средняя точность классификации в обоих случаях практически одинакова. Достигнутый уровень разделения очень примечателен, учитывая тот факт, что построение карты Сэммона является неконтролируемой

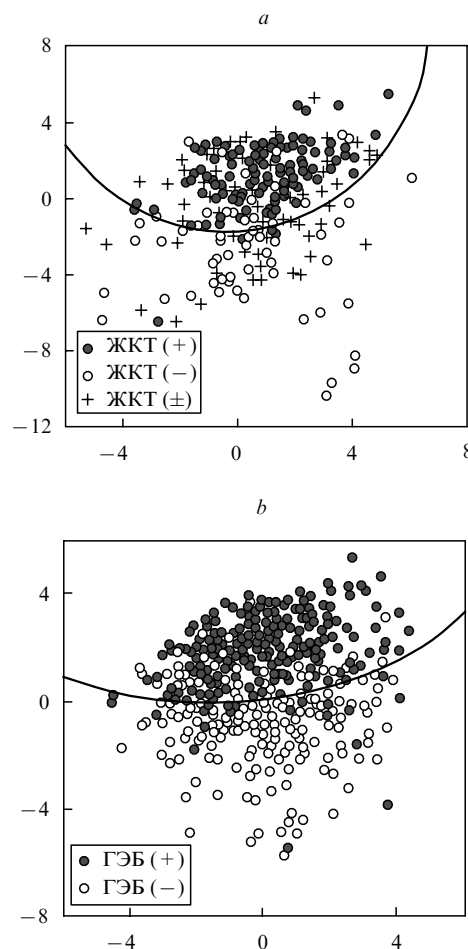


Рис. 8. Нелинейные карты Сэммона, демонстрирующие различия свойств соединений, обладающих высокой (темные кружки) и низкой (светлые кружки) способностью проникновения через стенки ЖКТ (a) и ГЭБ (b).

Здесь и далее оси на картах Сэммона имеют относительную градацию. Линия, разделяющая две категории соединений (хорошо и плохо проникающих через липидные мембраны), проведена при помощи алгоритма МОВ, обеспечивающего оптимальное (математически обоснованное) разделение.

Таблица 4. Показатели качества классификации при помощи карты Сэммона.

Свойство	Категория соединений	Число соединений		
		общее число испытанных соединений	корректно классифицированные	некорректно классифицированные
Проникновение через ГЭБ	ГЭБ (+)	267	237 (88.8%)	30 (11.2%)
	ГЭБ (–)	189	163 (86.2%)	26 (13.8%)
	Среднее:		87.5%	12.5%
Проникновение через стенки ЖКТ	ЖКТ (+)	114	110 (96.5%)	4 (3.5%)
	ЖКТ (–)	61	48 (78.7%)	13 (21.3%)
	Среднее:		87.6%	12.4%

‡ Для этих же целей был использован метод Кохонена.

объективной процедурой, которая не учитывает экспериментальной классификации молекулярных объектов.

Используя разработанные карты Сэммона, авторы провели дополнительные исследования, в частности изучили корреляции между степенью проникновения лекарственных веществ через эти липидные мембраны и значениями отобранных молекулярных дескрипторов. В результате проведенных исследований были выявлены различия в свойствах соединений с высокой степенью проникновения через ГЭБ и стенки ЖКТ. Таким образом, построенные карты Сэммона продемонстрировали возможность разделения соединений при помощи отобранных дескрипторов. Однако наряду с очевидными преимуществами применение алгоритма Сэммона сопряжено и с некоторыми трудностями.

б. Оценка других фармакокинетических параметров с использованием карт Сэммона

Как уже говорилось выше, фармакокинетические параметры (абсорбция, распределение, метаболизм и выведение — ADME-параметры) являются главнейшими характеристиками всех лекарственных веществ. В частности, в соответствии с директивой о лекарственных препаратах, каждое лекарственное соединение должно быть охарактеризовано по целому ряду признаков, в том числе по фармакокинетике и возможной цитотоксичности. Перспективные лекарственные вещества, не обладающие должными фармакокинетическими параметрами, как правило, снимаются со стадии биологического тестирования и подвергаются химической модификации. Таким образом, разработка компьютерных методов предсказания различных фармакокинетических параметров представляет чрезвычайно важную и актуальную задачу современной компьютерной медицинской химии.

Так, в работе⁶¹ описаны уникальные компьютерные модели на базе карт Сэммона для оценки некоторых принципиальных ADME-параметров различных физиологически активных соединений, включая описанные выше степени проникновения через стенки ЖКТ (рис. 8,а), ГЭБ (рис. 8,б), связывание с белками крови (Plasma protein binding, PPB (рис. 9,а), с одним из важнейших цитохромов печени CYP3A4 (на карте, приведенной на рис. 9,б, выражено в константах Михаэлиса, K_m) и с П-гликопротеинами (П-ГП, рис. 9,с), а также объемное распределение соединений в плазме крови (V_d , рис. 9,д) и время их полужизни (период выведения из организма) в плазме крови ($T_{1/2}$, рис. 9,е).

В качестве общей обучающей выборки авторы использовали аннотированную библиотеку лекарственных соединений, составленную на основе опубликованных материалов^{63–67} и информации, полученной из частных источников (любезно предоставленной С.Экинсом и взятой из тематических баз данных Ивановского института химии растворов). Всего в аннотированную библиотеку вошли 1500 различных лекарственных веществ, которые были использованы для построения компьютерной модели.

Расчет молекулярных дескрипторов осуществлялся на базе специализированных компьютерных программ Dragon (URL: <http://www.taletе.mi.it/dragon.htm>) и Chemosoft (URL: <http://www.chemosoft.com>). С помощью этих программ для каждого соединения было рассчитано более 1000 уникальных молекулярных дескрипторов. Отбор наиболее значимых дескрипторов (сокращение числа входных признаков) проводили путем анализа главных компонент. В итоге удалось показать, что более 90% случаев распределения может быть описано с использованием 6–8 главных компонент.

Вклады этих компонентов оценивались на основании стандартного критерия Кайзера–Гутмана⁶⁸ и критерия Каттеля (Scree-Test).⁶⁹ На основании этих данных были отобраны те молекулярные дескрипторы, которые вносят наибольший вклад в значения главных компонент. Аббревиатуры и соответствующие определения наиболее известных дескрипторов можно найти в книге⁷⁰ и в технической инструкции к программе Dragon.

Карты Сэммона для полной обучающей выборки строились с использованием программы Informagenesis, оптимальную границу разделения между исследуемыми классами соединений находили по методу опорных векторов с использованием программы LibSVM-2.477 (URL: <http://www.csie.ntu.edu.tw>). Полученные результаты для всех рассматриваемых в работе⁶¹ ADME-параметров проиллюстрированы на рис. 9.

Из приведенных данных видно, что во всех случаях «крайние» группы соединений располагаются в разных областях карт с незначительным перекрытием. Для нахождения оптимальной кривой разделения двух противоположных категорий соединений (промежуточный класс не использовался) применялся метод опорных векторов, после чего для всех карт была рассчитана точность предсказания (табл. 5).

Из приведенных в табл. 5 данных видно, что разработанные модели обладают хорошей предсказательной способностью (80–90% соединений были корректно классифицированы). Авторы провели еще ряд интересных компьютерных экспериментов, в которых исследовались закономерности расположения на картах Сэммона соединений из различных фармакокинетических классов. Так, было построено несколько дополнительных карт, отражающих различия между разными группами соединений, но принадлежащих одинаковым категориям. Пример такого разделения представлен на рис. 9,ф. Видно, что соединения, обладающие высокой проникающей способностью через ГЭБ и стенки ЖКТ, располагаются в разных областях карты, что, в свою очередь, говорит о различной природе этих биологических мембран.

Кроме того, авторы проанализировали вклад каждого из дескрипторов, использованных для построения соответствующей модели, в разделение тестируемых соединений по категориям (см. табл. 5). Закономерности распространения градиента значений трех репрезентативных молекулярных дескрипторов с учетом соответствующих линий МОВ представлены на рис. 10. Видно, что значения этих дескрипторов характеризуются собственным направленным градиентом. Так, в случае дескриптора, характеризующего липофильность соединений ($\lg P$), область наименьших значений ($\lg P < 0$) располагается в правой части карты, а область высоких значений ($\lg P > 0$) — в левой. Переход от низких значений к более высоким сравнительно плавный и соответствует монотонному распределению.

Ранее предполагалось, что разделение соединений, обладающих различной способностью к проникновению через ГЭБ, зависит в первую очередь от их липофильности. Однако авторам работы⁴¹ при помощи карт Сэммона удалось показать, что общая площадь гидрофильной поверхности (TPSA)⁷⁰ (рис. 10,б) вносит не меньший вклад в разделение этих классов соединений, хотя линия МОВ в этом случае проходит и не по границе области резкого перехода явных низких значений TPSA к более высоким.

Выше уже отмечалось, что для нахождения оптимальной границы разделения двух противоположных категорий соединений применяют метод опорных векторов. Линии МОВ

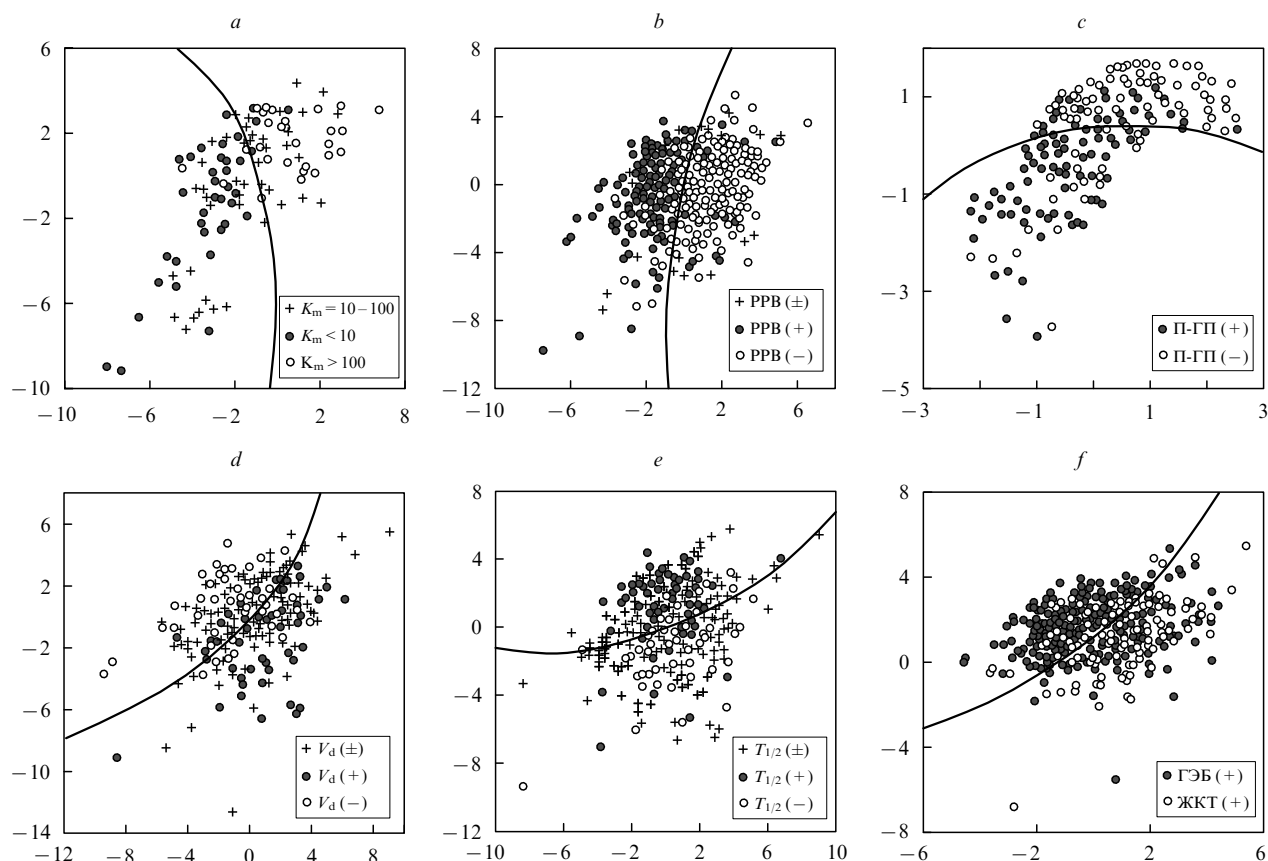


Рис. 9. Оценка некоторых наиболее важных фармакокинетических параметров различных биологически активных соединений с помощью карт Сэммона. Обозначения *a–f* см. в тексте.

Таблица 5. Оценка классификационной способности построенных карт Сэммона (оптимальная граница разделения противоположных категорий соединений найдена по методу опорных векторов).⁶¹

Фармакокинетический параметр	Категория соединений	Признак отнесения к данной категории	Количество тестируемых соединений	Корректно классифицированные соединения	Оценка предсказания, %
Проникновение через стенки ЖКТ	ЖКТ (+)	ЖКТ > 80%	114	110	96.5
	ЖКТ (±)	ЖКТ = 20–80%	145	—	—
	ЖКТ (–)	ЖКТ < 20%	61	48	78.7
Проникновение через ГЭБ	ГЭБ (+)	$\lg BB \geq 0^a$	267	237	88.8
	ГЭБ (–)	$\lg BB < -1$	189	163	86.2
Связывание с белками крови (РРВ) (главным образом с альбумином)	РРВ (+)	РРВ > 90%	200	169	84.5
	РРВ (±)	РРВ = 50–90%	172	—	—
	РРВ (–)	РРВ < 50%	177	143	80.8
Связывание с цитохромом СYP3A4	СYP (+)	$K_m \leq 10$	37	34	91.9
	СYP (–)	$K_m \geq 100$	28	26	92.9
Связывание с П-ГП	П-ГП (+)	Связывается	89	66	74.2
	П-ГП (–)	Не связывается	78	62	79.5
Объемное распределение вещества в плазме крови (V_d)	V_d (+)	$V_d < 0.4 \text{ л} \cdot \text{кг}^{-1}$	47	39	83.0
	V_d (±)	$V_d = 0.4–8 \text{ л} \cdot \text{кг}^{-1}$	162	—	—
	V_d (–)	$V_d > 8 \text{ л} \cdot \text{кг}^{-1}$	44	34	77.3
Время полужизни вещества в плазме крови ($T_{1/2}$)	$T_{1/2}$ (+)	$T_{1/2} > 20 \text{ ч}$	63	51	81.0
	$T_{1/2}$ (±)	$T_{1/2} = 1–20 \text{ ч}$	347	—	—
	$T_{1/2}$ (–)	$T_{1/2} < 1 \text{ ч}$	48	40	83.3

^a $\lg BB = \lg([c_{\text{кровь}}]/[c_{\text{мозг}}])$, где c — концентрация вещества.

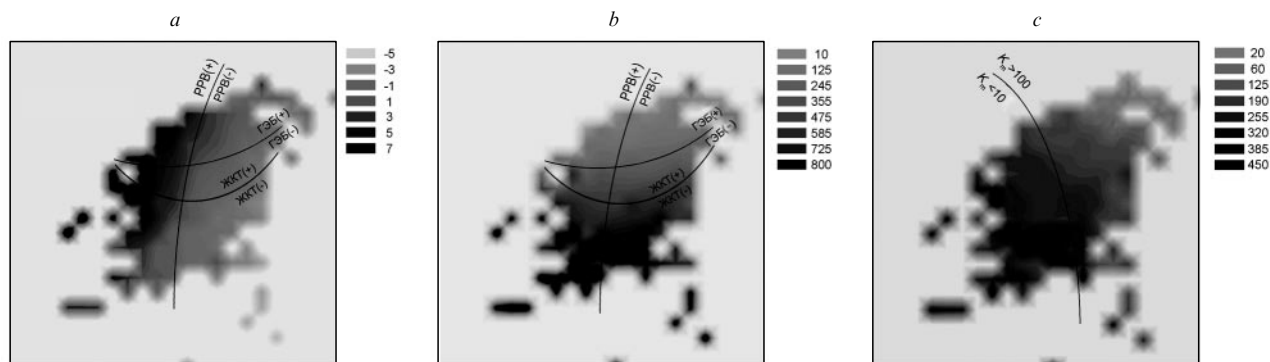


Рис. 10. Градиентные карты Сэмона, показывающие распределение по значениям трех репрезентативных дескрипторов, характеризующих липофильность соединений ($\lg P$, *a*), общую площадь гидрофильной поверхности (TPSA, *b*) и площадь отрицательно заряженной поверхности (PNSA, *c*), относительно МОВ-линий, разделяющих различные категории исследуемых соединений.

на рис. 10 были проведены с использованием компьютерной программы LibSVM-2.4. Эти линии соответствуют адекватной, с точки зрения математической статистики, границе разделения. Таким образом, используя алгоритм Сэмона и несколько простых молекулярных дескрипторов, можно анализировать и предсказывать многие важные фармакокинетические параметры.

в. Отбор наиболее значимых молекулярных дескрипторов с использованием карт Сэмона

Рациональный выбор наиболее значимых признаков, описывающих исходное многомерное пространство, является актуальной задачей современной компьютерной химии. Наряду с уже известными статистическими алгоритмами, используемыми в этой области исследований, такими как анализ главных компонент, факторный анализ, анализ чувствительности, методы прямой корреляции и *t*-статистика, часто прибегают к более точным средствам компьютерного моделирования, обладающим более высокой обобщающей и описательной способностью.

В работе⁷¹ было показано, что для этой цели можно применять и карты Сэмона. В качестве обучающей выборки авторы использовали цифровую базу данных медицинского центра Soroka (Индия), описывающую хромосомы человека. (Ранее эта база уже анализировалась с применением нейросетевых подходов.⁷²) Образцы хромосом, представленные в этой базе данных, были охарактеризованы специальным набором уникальных признаков, состоящим из четырех дескрипторов и описывающих структурно-геометрические особенности хромосом человека (центромерный индекс, длина, периметр и площадь). Обучающая выборка включала 100 различных хромосом, разделенных на три основных класса: хромосомы «13» и «19», а также X-хромосомы. Независимая контрольная выборка состояла из 10% образцов, случайным образом отобранных из обучающей выборки. Начальная скорость обучения составляла 0.9, а количество обучающих циклов — 30. Построенная карта Сэмона, представляющая собой двумерное отображение четырехмерного пространства признаков, представлена на рис. 11.

Можно видеть, что различные классы хромосом сосредоточены в разных областях карты, которые лишь незначительно перекрываются между собой. Качество предсказания для контрольной выборки при использовании всех четырех

дескрипторов составило 95%. Исключение из исходного набора дескрипторов наименее значимого приводит к потере предсказательной способности модели на 6%, что соответствует 89%-ной точности классификации. Заметим, что и 89%-ное качество предсказания вполне приемлемо для решения подобного рода задач.

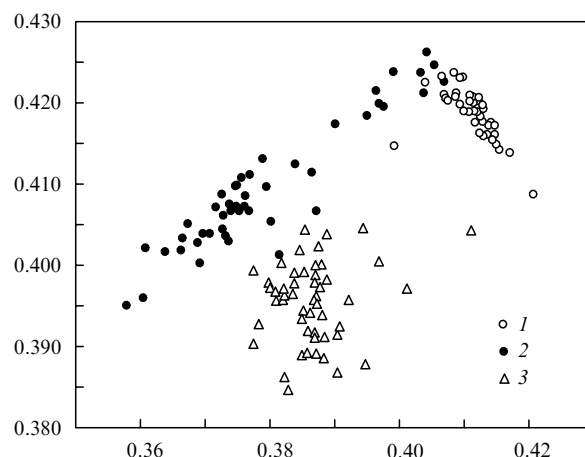


Рис. 11. Карта Сэмона для анализа хромосом «13» (1), «19» (2) и «X» (3).

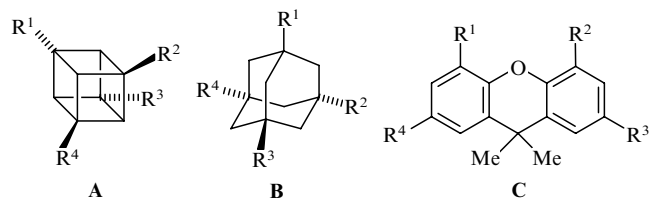
Следует отметить, что использование слишком большого количества дескрипторов в компьютерном моделировании зачастую ведет к переобучению модели. Такая модель не может быть использована для предсказания.

г. Исследование различных классов органических соединений

Помимо перечисленных задач компьютерной медицинской химии, непосредственно связанных с исследованием фармацевтических свойств органических соединений, карты Сэмона могут применяться и для систематизации органических соединений по структурным признакам, а также для оптимизации и планирования органического синтеза, в особенности комбинаторного. Необходимо отметить, что в этом случае молекулярные дескрипторы часто включают в себя признаки, которые описывают качественный и количественный структурный состав молекулы, например количество того

или иного периферийного структурного фрагмента. Применительно к решению таких задач карты Сэммона обладают рядом преимуществ по сравнению с самоорганизующимися картами Кохонена. Одним из преимуществ, о котором говорилось выше, является сохранение общей структуризации и топологии исходного пространства.

Авторами работы⁷³ была построена карта Сэммона для трех классов органических соединений — кубанов (A), адамантанов (B) и ксантенов (C) (рис. 12).



Обучающая выборка включала в себя несколько тысяч индивидуальных структур. Для построения карты Сэммона были использованы 12 автокорреляционных дескрипторов, полностью описывающих исходное пространство признаков. Для оценки разнообразия/сходства исследуемых соединений использовали метрику Евклида.

Построенная карта Сэммона позволяла улавливать различия между этими классами органических структур. Разделение соединений с использованием такой карты было четким и статистически значимым, что делает разработанную компьютерную модель весьма перспективным средством для классификации разнообразных комбинаторных библиотек соединений, содержащих рассмотренные циклические структуры.

Карты Сэммона по сравнению с картами Кохонена способны более точно передавать общую топологию входного пространства. Однако такое увеличение точности происходит за счет увеличения времени расчетов: время, необходимое для построения карт Сэммона, растет пропорционально второй степени от числа анализируемых объектов. В результате, даже используя мощные персональные компьютеры,

невозможно строить карты Сэммона с числом соединений, превышающим 5000–7000. Это обстоятельство, как правило, не мешает анализу обучающих баз данных, но является непреодолимым препятствием для тестирования больших виртуальных библиотек, зачастую включающих в себя сотни тысяч и даже миллионы соединений. Другим недостатком карт Сэммона является невозможность прямого тестирования при помощи уже построенной карты новых объектов. Для этого требуется полное переучивание карты, включающее в себя обучающую и контрольную выборки, причем в этом случае качество разделения, достигнутое при анализе только обучающей выборки, уже не гарантируется.

3. Алгоритм GalaxyTM

В заключение рассмотрим еще один интересный метод нелинейного картирования, который по своей сути является модифицированным алгоритмом Сэммона. Разработанный компанией BioWisdom (Великобритания, URL: <http://www.omniviz.com>) компьютерный продукт под названием OmniViz включает алгоритм GalaxyTM, с помощью которого можно генерировать двумерные проекции многомерного пространства признаков. Этот алгоритм находит применение при анализе разных типов химических данных. Так, в работе⁷⁴ была показана высокая эффективность алгоритма GalaxyTM для анализа химических структур в рамках комбинаторного синтеза и сопряженного с ним биологического тестирования. Обучающая выборка включала 1107 индивидуальных соединений, полученных методом высокоэффективного комбинаторного синтеза, которые были испытаны на биологическую активность в условиях высокопроизводительного биологического скрининга в 19 различных биологических тестах. Для всех протестированных соединений были рассчитаны структурные дескрипторы на базе различных подструктурных фрагментов. После чего с использованием алгоритма GalaxyTM была построена двумерная проекция многомерного пространства структурных признаков (рис. 13, a).

Как и в случае снижения размерности входного пространства по Сэммону, алгоритм позволяет получать проекции, в которых расстояния между всеми объектами в пространстве меньшей размерности соответствуют расстоя-

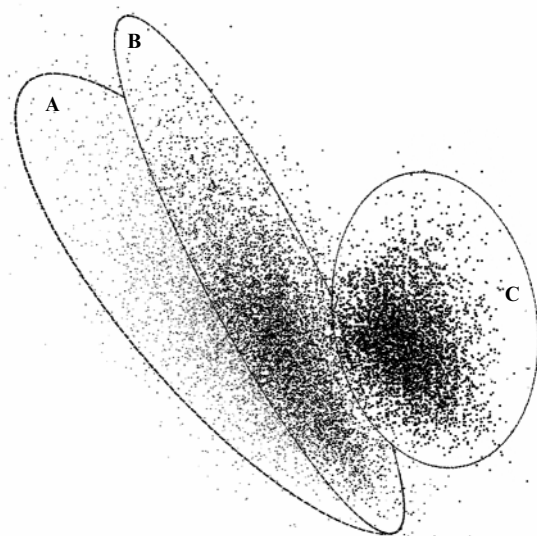


Рис. 12. Нелинейная проекция Сэммона, показывающая распределение соединений из трех библиотек органических соединений — кубанов (A), адамантанов (B) и ксантенов (C).

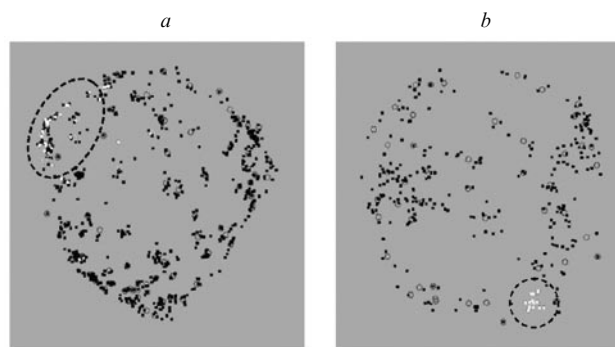


Рис. 13. Анализ различных химических структур с применением алгоритма GalaxyTM.

a — двумерная карта для обучающей выборки органических соединений (1107 структур), построенная на базе различных структурных фрагментов; b — двумерная карта, отражающая биологическую активность протестированных соединений. Каждая точка на карте соответствует одному соединению.

ниям в многомерном пространстве признаков. Таким образом, соединения, близко расположенные на карте, являются соседями и в исходном многомерном массиве.

На рис. 13,*b* приведена карта биологической активности протестированных соединений, построенная по аналогичному принципу, т.е. соединения, обладающие похожей биологической активностью, занимают близкие области на карте. Интересно, что при произвольном выборе группы близко расположенных на карте биологической активности соединений (выделены пунктиром) программа в автоматическом режиме идентифицирует эти соединения и на первой карте (рис. 13,*a*), на которой отражены структурные особенности соединений, выраженные в количестве тех или иных подструктурных фрагментов. Из анализа расположения этой группы соединений на карте (см. рис. 13,*a*) следует, что они обладают высокой степенью родства, однако есть ряд соединений с отличным структурным окружением. Последнее означает, что различные в структурном отношении соединения могут обладать одинаковой биологической активностью и наоборот. Таким образом, алгоритм GalaxyTM может рассматриваться в качестве перспективной альтернативы классическим картам Сэммона и может с успехом применяться для анализа многопараметрических данных в области компьютерной медицинской химии.

4. Исследование экспрессии генов с использованием комбинированного алгоритма картирования

В некоторых случаях для решения комплексных и сложных задач, таких как анализ профиля экспрессии генов, прибегают к комбинированному алгоритму картирования на базе самоорганизующихся карт Кохонена и Сэммона. Возможность использования такого гибридного метода для анализа данных, которые были получены в результате проведения ДНК-анализа вызванной диауксической реакцией экспрессии генов, была показана в работе⁷⁵. Обучающая выборка включала данные, описывающие экспрессию 6400 генов дрожжей. В качестве дескрипторов выступали значения объема экспрессии генов, зарегистрированные на различных временных интервалах; всего использовали семь контрольных точек.

На первой стадии данные анализировались с использованием карт Кохонена (16 × 16), каждый нейрон которых в зависимости от контрольной точки содержал от 10 до 64 генов (рис. 14,*a*) (в рассматриваемой работе использовалась иная визуализация результатов выполнения алгоритма картирования по Кохонену). Как показал анализ построенной карты, гены с одинаковым характером экспрессии попали в один и тот же нейрон Кохонена, гены с похожим характером экспрессии — в нейроны, локализованные в непосредственной близости друг от друга, в то время как нейроны, занимающие принципиально различные места на карте, соответствовали генам, обладающим различным профилем экспрессии. В качестве примера на рис. 14 отдельно показаны четыре различных нейрона Кохонена (C1–C4), выделенные на общей карте. Гистограмма, показанная в каждом из четырех нейронов, характеризует количественный характер экспрессии генов (среднее значение) в контрольных точках (семь временных интервалов (по два часа каждый) на протяжении всей диауксической реакции).

Далее полученные интегральные данные анализировались с применением алгоритма картирования по Сэммону (рис. 14,*b*). Главные оси карты Сэммона отражают количественную степень экспрессии, так что гены, характеризующиеся наибольшим интегральным значением активности,

располагаются в левой области карты, а гены с низкой активностью — в правой. Гены, экспрессия которых не регулировалась в условиях диауксической реакции, располагаются в центральной части карты. Таким образом, пространственная удаленность объектов на картах Сэммона соответствует различиям в среднем объеме экспрессии генов, а размер точек — их количеству в нейроне Кохонена.

Предложенная гибридная компьютерная модель на базе самоорганизующихся карт Кохонена и нелинейных карт Сэммона позволила авторам математически описать экспрессию 6400 генов дрожжей в условиях диауксической реакции.

IV. Использование алгоритма SPE для анализа химического пространства

К настоящему времени самоорганизующиеся карты Кохонена и метод нелинейного картирования по Сэммону уже успели зарекомендовать себя в качестве эффективных и точных методов компьютерной медицинской химии. Они наиболее часто используются для решения многих практических задач виртуального скрининга. Однако востребованность еще более мощных и точных алгоритмов, которые позволяют обойти ряд трудностей, связанных с применением перечисленных выше методов, таких как низкая производительность алгоритма Сэммона или кластерное приближение алгоритма Кохонена, сохраняется. Одним из таких методов является алгоритм нелинейного картирования SPE (отображение вероятностного приближения), сравнительно недавно разработанный Аграфитосом и сотр.⁷⁶

В основе метода SPE, как и метода нелинейного картирования по Сэммону, лежит фундаментальный принцип самоорганизации — построение наиболее адекватной и полной проекции структуры исходного пространства свойств в пространстве наименьшей размерности. Принципиальное отличие метода SPE от метода Сэммона состоит в том, что этот алгоритм не требует расчета абсолютно всех геодезических расстояний между объектами в новом пространстве низкой размерности. Для этого он использует модифицированную стратегию попарного отнесения, которая не требует расчета полносвязной матрицы подобия. В отличие от карт Сэммона производительность алгоритма SPE линейно зависит от объема входного пространства. Другими словами, SPE учитывает тот факт, что геодезическое расстояние всегда превосходит или равно исходной метрике подобия. Метод SPE в большей степени нацелен на сохранение принципиальных локальных расстояний, нежели общей структуры исходного пространства признаков. Используемая в алгоритме SPE модифицированная методика оценки расстояний между объектами позволяет получать полносвязные проекции с привлечением гораздо меньших ресурсов и в итоге генерировать двумерные карты, на которых в полной мере отражаются как общая топология исходного пространства, так и локальные закономерности. Эффективность применения алгоритма SPE в области виртуального скрининга органических соединений показана в целом ряде научных публикаций, некоторые из них рассмотрены ниже.

Как и алгоритмы картирования по Сэммону, алгоритм SPE может с успехом применяться для исследования химического пространства сложной геометрии с неочевидными, скрытыми взаимосвязями.³⁴ На рис. 15 представлены две двумерные карты SPE для описания закономерностей внутри виртуальных библиотек органических соединений — продуктов двух виртуальных реакций: двухкомпонентной реакции

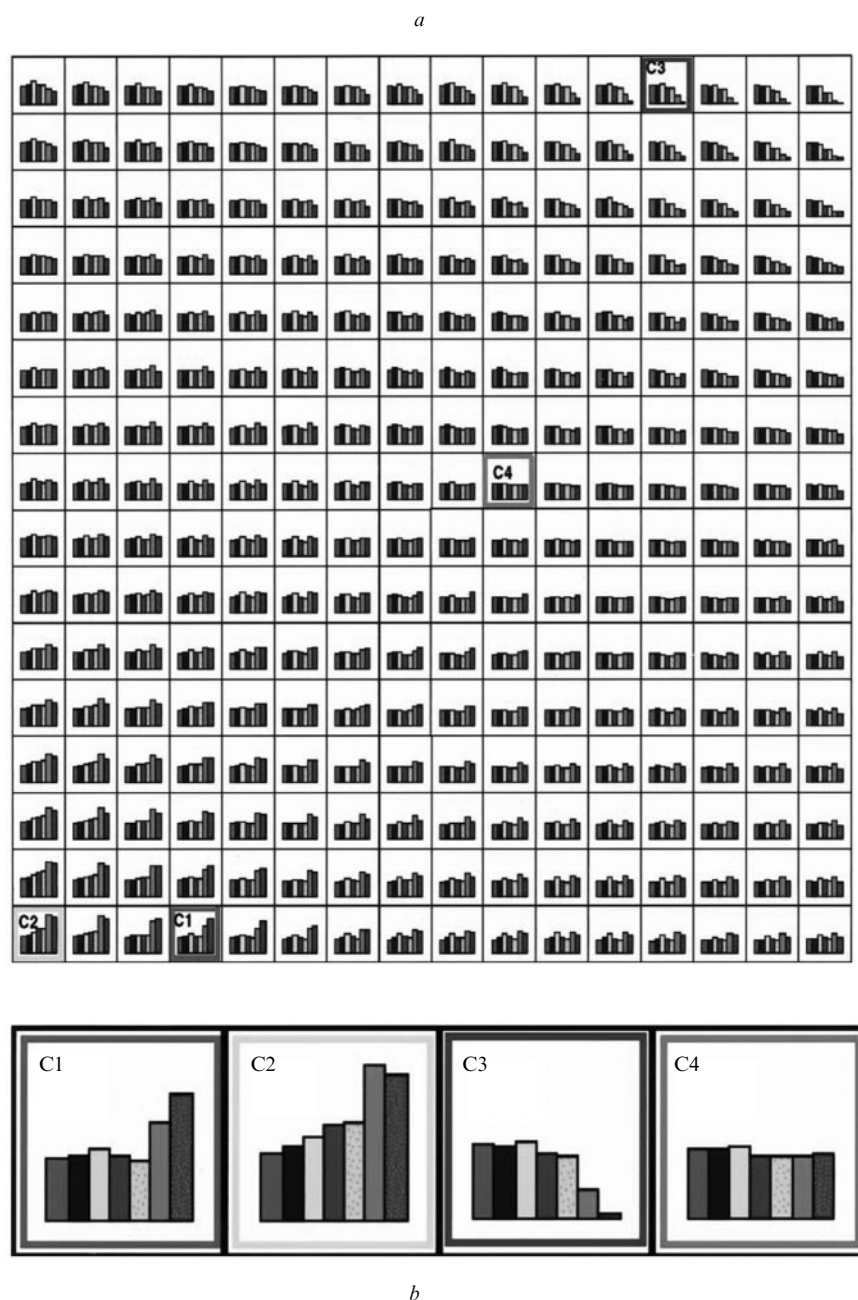


Рис. 14. Карта Кохонена (*a*), характеризующая профиль экспрессии 6400 генов дрожжей в зависимости от времени регистрации (различными оттенками серого обозначены семь различных контрольных точек), и построенная на ее основе интегральная карта Сэммона (*b*) (различными оттенками серого показаны позиции, соответствующие выделенным нейронам Кохонена).

аминирования (рис. 15,*a*) и многокомпонентной конденсации Уги (рис. 15,*b*).

Так, на рис. 15,*a* показано двумерное отображение SPE виртуальной библиотеки, состоящей из 10 000 органических

соединений, которые были получены из 100 аминов и 100 альдегидов различного строения. Каждый продукт виртуальной реакции аминирования был охарактеризован 117 различными топологическими дескрипторами, включая

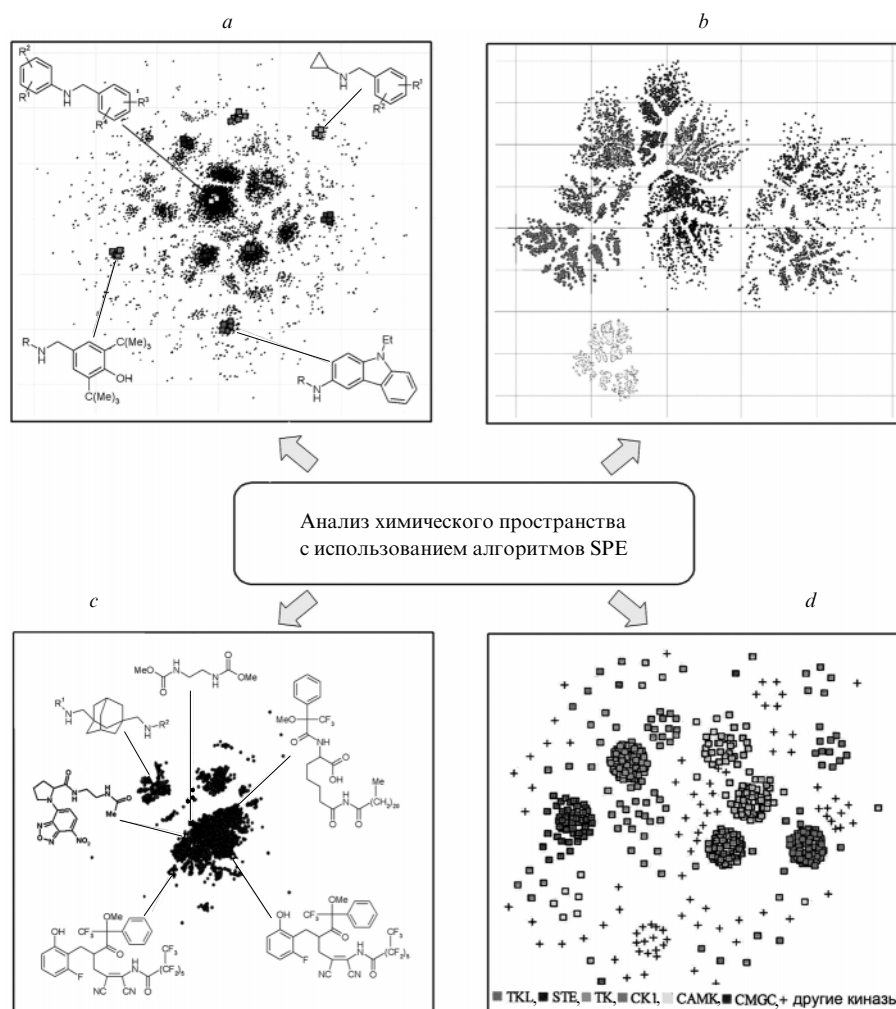


Рис. 15. Использование алгоритма SPE для решения различных задач компьютерной медицинской химии: распределение различных соединений из комбинаторной библиотеки, полученной реакцией аминирования (начальный радиус обучения — 0.4) (a); распределение соединений, полученных по классической реакции Уги (b); распределение различных диаминов (c); локализация различных типов киназ (начальный радиус обучения — 0.89) (d).

индексы молекулярной связности, k -индексы формы молекулы, информационно-теоретические индексы, индексы Бончева–Тринайстича и т.д.³⁴ Для снижения размерности входного пространства использовался метод анализа главных компонент. Было найдено, что наибольший вклад в изменение исходного пространства признаков (~99%) вносят 23 главных компонента, на базе которых и была построена карта SPE. Как видно из рис. 15,а, различные структурные классы соединений, полученные в результате реакции аминирования, занимают обособленные положения на карте, что говорит о высокой способности алгоритма SPE решать подобные задачи классификации.

Приведенная на рис. 15,б двумерная карта отражает закономерности, найденные внутри пространства виртуальной библиотеки из 10 000 структур. Эти соединения — продукты виртуальной четырехкомпонентной конденсации Уги между карбоновыми кислотами, первичными аминами алифатического и ароматического рядов, альдегидами и изонитрилами (по 10 различных соединений в каждой группе исходных реагентов). Каждый продукт виртуальной реакции был охарактеризован 166 различными структурными дескрипторами, кодирующими количество того или иного

фрагмента в целевом соединении. В данном случае при отборе дескрипторов авторы использовали метрику структурного подобия/различия, основанную на классическом уравнении и коэффициентах Танимото. Видно, что построенные карты содержат несколько очевидных компактных кластеров, которые соответствуют различным классам соединений. (Построение карт подробно описывается в работе⁷⁷.)

Метод SPE был успешно применен и для классификации органических соединений, существенно различающихся по строению и имеющих во многих случаях сложную структуру (см., например, работу⁷⁸). Как и в предыдущем случае, компьютерное моделирование осуществлялось на базе специальным образом отобранного универсального набора структурных дескрипторов, отражающих количество того или иного фрагмента в молекуле. Построенная двумерная карта приведена на рис. 15,с.

Алгоритм SPE показал свою состоятельность и высокую эффективность и при решении других, более сложных и комплексных задач, таких как конформационный анализ органических молекул,⁷⁹ определение структуры соединения методом спектроскопии ЯМР и предсказание структуры

белковых макромолекул.⁸⁰ Так, алгоритм SPE в сочетании с алгоритмом самоорганизующегося наложения (Self-organizing superimposition, SOS) рассматривается в качестве перспективного компьютерного метода анализа конформационных состояний органических молекул и осуществления поиска соединений по их конформации.^{81–83} Помимо этого, метод SPE был успешно применен для классификации и визуализации белковых аминокислотных последовательностей, где аминокислоты использовались в качестве дескрипторов. Так, на рис. 15,d показано двумерное отображение 21-мерного пространства признаков (аминокислот), характеризующих пространство киназных ферментов. Алгоритм SPE может также с успехом использоваться для снижения размерности данных, полученных в результате проведения различных исследований в области геномики и протеомики.⁸⁴ Высокая эффективность алгоритма SPE в решении задач, связанных с анализом мишень-специфической активности органических структур, может быть наглядно проиллюстрирована на примере построения компьютерных моделей для предсказания активности соединений по отношению к киназам и ядерным рецепторам.⁸⁵

Однако, как и все существующие на сегодняшний день методы компьютерного анализа, метод SPE не лишен некоторых недостатков и ограничений. К их числу можно отнести сравнительно сложную процедуру настройки алгоритма, во время которой приходится устанавливать множество внутренних параметров обучения и метрики подобия. Так же, как в случае метода IsoMap, результативность алгоритма SPE сильно зависит от выбора радиуса обучения. В случае слишком большого значения близко расположенные объекты разных классов будут существенно перекрываться между собой, захватывая также более удаленные объекты. В то же время использование малых значений радиуса обучения может привести к нарушению/искажению общей топологии и структурной организации пространства дескрипторов меньшей размерности, которое уже не будет соответствовать исходному многомерному пространству признаков.

Несмотря на возможные трудности, связанные с применением алгоритма SPE, он, как и описанные выше методы картирования Кохонена и Сэммона, рассматривается в качестве одного из наиболее эффективных и точных методов анализа химического пространства. В заключение следует сказать, что помимо методов Кохонена, Сэммона и SPE были разработаны и другие методы нелинейного картирования. Так, похожий на алгоритм SPE метод нелинейного

картирования, названный криволинейным компонентным анализом (CCA), был предложен авторами работы⁸⁶. Этот метод широко используется для решения различных задач математической статистики, в частности для анализа многомерных данных путем снижения размерности исходного пространства признаков.

V. Специализированные компьютерные программы

В сфере компьютерного моделирования существует большое число компьютерных программ, позволяющих строить самоорганизующиеся карты Кохонена и нелинейные карты Сэммона. Некоторые из них перечислены в табл. 6. Однако не все из этих компьютерных программ обладают достаточной эффективностью и удобны в применении. Многие программы работают под устаревшей операционной системой DOS и являются полностью консольными версиями без пользовательского интерфейса. Такие, чаще всего простые, программы предназначены для решения несложных задач из различных областей компьютерного моделирования. На сегодняшний день существует только одна компьютерная программа, которая была специально разработана для нужд компьютерной медицинской химии и виртуального скрининга на базе алгоритмов картирования — SmartMining (URL: <http://www.Informagenesis.com>).

Программа SmartMining является частью (специальным модулем) более мощной в математическом аспекте программы Informagenesis. Программа SmartMining предназначена для работы с химическими данными. С ее помощью можно осуществлять различные исследования в области виртуального скрининга и компьютерного моделирования. В модуль SmartMining встроены многие прогностические модели, которые в автоматическом режиме способны предсказывать целый спектр принципиальных лекарственных характеристик органических соединений, включая профиль мишень-специфической активности, степень проникновения через биологические мембраны, множество фармако- и токсикокинетических параметров. Например, программа SmartMining позволяет проводить исследования проникновения соединений через гематоэнцефалический барьер (на базе самоорганизующихся карт Кохонена) и через стенки желудочно-кишечного тракта (на базе карт Сэммона).⁶⁴

Таблица 6. Коммерчески доступные программы для построения QSAR/QSPR-моделей на базе карт Кохонена и Сэммона (с разделением по методу опорных векторов).

Программа	Алгоритм	Разработчик
Informagenesis	Искусственные нейронные сети, карты Кохонена и Сэммона	Компания Informagenesis (URL: http://www.informagenesis.com)
NeuroSolution	Искусственные нейронные сети и карты Кохонена	NeuroDimension Inc., США (URL: http://www.nd.com)
SVMlight	Метод опорных векторов	Дортмундский университет, Германия (URL: http://svmlight.joachims.org)
ChemoSoft	Искусственные нейронные сети, карты Сэммона	Chemical Diversity Labs Inc., США; ИИХР, Россия (URL: http://www.ChemDiv.com)
OmniViz (Galaxy™)	Модифицированный алгоритм Сэммона	Компания BioWisdom, Великобритания (URL: http://www.omniviz.com)

Разработанные с использованием программного комплекса SmartMining методы QSAR/QSPR-моделирования пригодны для автоматической оценки свойств больших баз соединений, содержащих до нескольких миллионов структур. При этом можно использовать стандартные персональные компьютеры под управлением операционной системы Windows. Опыт использования QSAR/QSPR-моделей убеждает, что компьютерный прогноз в большинстве случаев гораздо точнее, чем взгляд опытного эксперта, работающего в определенной области органической химии. Производительности компьютерных методов и человеческой экспертизы абсолютно несоизмеримы, и этот разрыв непрерывно растет по мере увеличения производительности доступных компьютерных систем.

VI. Заключение

Сегодня уже ни у кого не вызывает сомнения, что виртуальный скрининг является неотъемлемой частью процесса разработки новых лекарственных веществ. Несмотря на сравнительно недолгую историю существования специализированных методов компьютерной медицинской химии, существует целый ряд эффективных *in silico* технологий, позволяющих производить рациональный отбор химических структур для непосредственного биологического тестирования, а также осуществлять различные исследования в этой области. Среди разнообразных методов QSAR/QSPR-моделирования методы картирования, позволяющие получать двумерные и трехмерные проекции исходного многомерного пространства молекулярных дескрипторов, занимают особое место. Помимо предсказательной способности такие карты весьма удобны для визуального восприятия; они позволяют исследователю проводить комплексный и, что немаловажно, разносторонний анализ найденных закономерностей. На основании полученной нелинейной проекции исследуются причины и основные факторы, обуславливающие разделение различных категорий лекарственных веществ. Основной задачей методов картирования является построение адекватной компьютерной модели, при помощи которой можно в дальнейшем прогнозировать различные терапевтические и физико-химические свойства органических молекул, такие как профиль мишень-специфической активности, различные фармакокинетические параметры, растворимость, температурная стабильность и т.д. Таким образом, исследователь получает в свои руки эффективный и точный компьютерный инструмент, без которого в условиях современного развития органического синтеза и биологического скрининга обойтись просто невозможно.

Литература

1. K.C.Nicolaou, E.J.Sorensen. *Classic in Total Synthesis*. Wiley-VCH, Weinheim, 1996. P. 798
2. N.K.Terrett, M.Gardner, D.W.Gordon, R.J.Kobylecki, J.Steele. *Tetrahedron*, **51**, 8135 (1995)
3. W.Bannwarth, E.Felder. *Combinatorial Chemistry: a Practical Approach*. Wiley-VCH, Weinheim, 2000. P. 430
4. H.-J.Böhm, G.Schneider. *Virtual Screening for Bioactive Molecules*. Wiley-VCH, Weinheim, 2000
5. W.P.Walters, M.T.Stahl, M.A.Murcko. *Drug Discov. Today*, **3**, 160 (1998)
6. F.L.Stahura, J.Bajorath. *Comb. Chem. High Throughput Screen.*, **7**, 259 (2004)
7. H.Kubinyi. In *Methods and Principles in Medicinal Chemistry*. Vol. 1. Wiley-VCH, Weinheim, 1993. P. 4
8. J.Bajorath. *Nature Rev. Drug Discov.*, **1**, 882 (2002)
9. К.В.Балакин, Я.А.Иваненков, А.В.Скоренко, С.Н.Коваленко, И.А.Журавель, В.П.Черных. *Укр. журн. орг. фарм. химии*, **2**, 47 (2004)
10. C.Merlot, D.Domine, C.Cleva, D.J.Church. *Drug Discov. Today*, **8**, 594 (2003)
11. P.Willett. *J. Chem. Inf. Comput. Sci.*, **38**, 983 (1998)
12. R.P.Sheridan, S.K.Kearsley. *Drug Discov. Today*, **7**, 903 (2002)
13. G.A.Patani, E.J.LaVoie. *Chem. Rev.*, **96**, 3147 (1996)
14. J.Krumrine, F.Raubacher, N.Brooijmans, I.Kuntz. *Meth. Biochem. Anal.*, **44**, 443 (2003)
15. K.V.Balakin, S.E.Tkachenko, I.Okun, A.V.Skorenko, Y.A.Ivanenkov, N.P.Savchuk, A.A.Ivashchenko, Y.Nikolsky. *Chem. Today*, **22**, 15 (2004)
16. N.P.Savchuk, K.V.Balakin. *Curr. Drug Discov.*, **5**, 23 (2004)
17. S.Anzali, G.Barnickel, B.Cezanne, M.Krug, D.Filimonov, V.Poroikov. *J. Med. Chem.*, **44**, 2432 (2001)
18. V.V.Poroikov, D.A.Filimonov, W.-D.Ihlenfeldt, T.A.Gloriozova, A.A.Lagunin, Y.V.Borodina, A.V.Stepanchikova, M.C.Nicklaus. *J. Chem. Inf. Comput. Sci.*, **43**, 228 (2003)
19. Y.Borodina, A.Sadym, D.Filimonov, V.Blinova, A.Dmitriev, V.Poroikov. *J. Chem. Inf. Comput. Sci.*, **43**, 1636 (2003)
20. T.Kohonen. *Self-organizing Maps*. 3rd Ed. Springer-Verlag, New-York, 2001. P. 528
21. J.W.Sammon Jr. *IEEE Trans. Comput.*, **C18**, 969 (1969)
22. J.P.Bigus. *Data Mining with Neural Networks*. McGraw-Hill, New York, 1996. P. 220
23. J.Devillers. *Neural Networks in QSAR and Drug Design*. Academic Press, London, 1996. P. 284
24. V.Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998. P. 768
25. R.D.Cramer, M.A.Poss, M.A.Hermsmeier, M.A.Hermsmeier, T.J.Caulfield, M.C.Kowala, M.T.Valentine. *J. Med. Chem.*, **42**, 3919 (1999)
26. A.D.van Dijk, R.Boelens, A.M.Bonvin. *FEBS J.*, **272**, 293 (2005)
27. D.B.Kitchen, H.Decornez, J.R.Furr, J.Bajorath. *Nat. Rev. Drug Discov.*, **3**, 935 (2004)
28. R.D.Taylor, P.J.Jewsbury, J.W.Essex. *J. Comput. Aid. Mol. Des.*, **16**, 151 (2002)
29. P.G.Gane, P.M.Dean. *Curr. Opin. Drug Discov. Dev.*, **10**, 401 (2000)
30. R.D.Brown, Y.C.Martin. *J. Chem. Inf. Comput. Sci.*, **36**, 572 (1996)
31. M.Bredel, E.Jacoby. *Nature Rev. Gen.*, **5**, 262 (2004).
32. S.Frickenhaus, S.Kannan, M.Zacharias. *J. Comput. Chem.*, **30**, 479 (2009)
33. D.K.Agrafiotis, V.S.Lobanov. *J. Chem. Inf. Comput. Sci.*, **40**, 1356 (1997)
34. D.K.Agrafiotis. *J. Comput. Chem.*, **24**, 1215 (2003)
35. M.A.Johnson, G.M.Maggiora. *Concepts and Applications of Molecular Similarity*. Wiley-VCH, Weinheim, 1990
36. S.Lafon, A.B.Lee. *IEEE Trans. Pattern Anal. Machine Intelligence*, **28**, 1393 (2006)
37. J.B.Tenenbaum. In *Advances in Neural Information Processing Systems. Vol. 10*. MIT Press, Cambridge, 1998. P. 682
38. H.Bauknecht, A.Zell, H.Bayer, P.Levi, M.Wagener, J.Sadowski, J.Gasteiger. *J. Chem. Inf. Comput. Sci.*, **36**, 1205 (1996)
39. S.Anzali, G.Barnickel, M.Krug, J.Sadowski, M.Wagener, J.Gasteiger, J.Polanski. *J. Comput. Aid. Mol. Des.*, **10**, 521 (1996)
40. M.Brüistle, B.Beck, S.T.Chindler, W.King, T.Mitchell, T.Clark. *J. Med. Chem.*, **45**, 3345 (2002)
41. A.A.Rabow, R.H.Shoemaker, E.A.Sausville, D.G.Covell. *J. Med. Chem.*, **45**, 818 (2002)
42. P.Labute, S.Nilar, C.Williams. *Comb. Chem. High Throughput Screen.*, **5**, 135 (2002)
43. H.Gao, C.Williams, P.Labute, J.Bajorath. *J. Chem. Inf. Comput. Sci.*, **39**, 164 (1999)

44. K.V.Balakin, S.A.Lang, A.V.Skorenko, S.E.Tkachenko, A.A.Ivashchenko, N.P.Savchuk. *J. Chem. Inf. Comput. Sci.*, **43**, 1553 (2003)
45. V.V.Zernov, K.V.Balakin, A.A.Ivashchenko, N.P.Savchuk, I.V.Pletnev. *J. Chem. Inf. Comput. Sci.*, **43**, 2048 (2003)
46. M.T.Cronin, A.O.Aptula, J.C.Dearden, J.C.Duffy, T.I.Netzeva, H.Patel, P.H.Rowe, T.W.Schultz, A.P.Worth, K.Voutzoulidis, G.Schüürmann. *J. Chem. Inf. Comput. Sci.*, **42**, 869 (2002)
47. Y.Nikolsky, K.V.Balakin, Y.A.Ivanenkov, A.A.Ivashchenko, N.P.Savchuk. *PharmaChem.*, **4**, 68 (2003)
48. K.V.Balakin. *Modern Drug Discov.*, **9**, 45 (2003)
49. Y.A.Ivanenkov, K.V.Balakin, A.V.Skorenko, S.E.Tkachenko, N.P.Savchuk, A.A.Ivashchenko, Y.Nikolsky. *Chem. Today*, **21**, 72 (2003)
50. N.P.Savchuk, K.V.Balakin. In *Virtual Screening in Drug Discovery*. CRC Press, New York, 2005. P. 121
51. N.P.Savchuk, S.E.Tkachenko, K.V.Balakin. In *Cheminformatics in Drug Discovery*. Wiley-VCH, Weinheim, 2004. P. 287
52. К.В.Балакин. Дис. д-ра хим. наук. Ивановский гос. ун-т, Иваново, 2005
53. *Derwent World Drug Index (Derwent WDI)*. Derwent Information, London, 1997
54. Maybridge Chemicals Company Ltd., England
55. A.G.Wilson, A.C.White, R.A.Mueller. *Curr. Opin. Drug Discov. Devel.*, **6**, 123 (2003)
56. M.T.Cronin. *Exp. Suppl.*, **93**, 259 (2003)
57. M.T.Cronin. *Altern. Lab. Anim. Suppl.*, **2**, 81 (2002)
58. J.D.McKinney, A.Richard, C.Waller, M.C.Newman, F.Gerberick. *Toxic. Sci.*, **56**, 8 (2000)
59. P.Barbieri, N.Piclin, A.Szymoszek, M.Novic, M.Vracko, E.Benfenati. In *Meeting of Slovenian Chemical Society 2001*. FKKT Maribor, 2001. P. 74
60. C.L.Russom, S.P.Brandbury, S.J.Broderius, D.E.Hammermeister, D.A.Drummond. *Environ. Toxicol. Chem.*, **16**, 948 (1997)
61. M.Jianchang, A.K.Jain. *IEEE Trans. Neural Netw.*, **6**, 296 (1995)
62. K.V.Balakin, Y.A.Ivanenkov, N.P.Savchuk, A.A.Ivashchenko, S.Ekins. *Curr. Drug Discov. Technol.*, **2**, 99 (2005)
63. K.V.Balakin, S.Ekins, A.Bugrim, Y.A.Ivanenkov, D.Korolev, Y.V.Nikolsky, A.V.Skorenko, A.A.Ivashchenko. *Drug Metab. Dispos.*, **32**, 1183 (2004)
64. U.Norinder, M.Haeberlein. *Adv. Drug Deliv. Rev.*, **54**, 291 (2002)
65. K.Thummel, D.D.Shen. In *Goodman and Gilman's the Pharmaceutical Basis of Therapeutics*. McGraw-Hill, New York, 2001, P. 1924
66. J.E.Penzotti, M.L.Lamb, E.Evenson, P.D.J.Grootenhuys. *J. Med. Chem.*, **45**, 1737 (2002)
67. S.Ekins, R.B.Kim, B.F.Leake, A.H.Dantzig, E.Schuetz, L.B.Lan, K.Yasuda, R.L.Shepard. *Mol. Pharmacol.*, **61**, 964 (2002)
68. L.Guttman. *Psychometrika*, **19**, 149 (1954)
69. R.B.Catell. *Multi Behav. Res.*, **1**, 245 (1966)
70. R.Todeschini, V.Consonni. *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, 2002
71. B.Lerner, H.Guterman, M.Aladjem, I.Dinstein. *Pattern Recognit. Lett.*, **20**, 7 (1999)
72. B.Lerner, H.Guterman, I.Dinstein, Y.Romem. *Pattern Recognit.*, **28**, 1673 (1995)
73. J.Sadowski, M.Wagener, J.Gasteiger. *Angew. Chem., Int. Ed. Engl.*, **34**, 2674 (1995)
74. P.Gedeck, P.Willett. *Curr. Opin. Chem. Biol.*, **5**, 389 (2001)
75. P.Toronen, M.Kolehmainen, G.Wong, E.Castren. *FEBS Lett.*, **451**, 142 (1999)
76. D.K.Agrafiotis, H.Xu. *Proc. Natl. Acad. Sci. USA*, **99**, 15869 (2002)
77. D.K.Agrafiotis, H.Xu. *J. Chem. Inf. Comput. Sci.*, **43**, 475 (2003)
78. D.N.Rassokhin, D.K.Agrafiotis. *J. Mol. Graph. Model.*, **22**, 133 (2003)
79. D.C.Spellmeyer, A.K.Wong, M.J.Bower, J.M.Blaney. *J. Mol. Graph. Model.*, **15**, 18 (1997)
80. T.F.Havel, K.Wuthrich. *J. Mol. Biol.*, **182**, 281 (1985)
81. F.Zhu, D.K.Agrafiotis. *J. Comput. Chem.*, **28**, 1234 (2007)
82. H.Xu, S.Izrailev, D.K.Agrafiotis. *J. Chem. Inf. Comput. Sci.*, **43**, 1186 (2003)
83. D.K.Agrafiotis, D.Bandyopadhyay, G.Carta, A.J.S.Knox, D.G.Lloyd. *Chem. Biol. Drug Des.*, **70**, 123 (2007)
84. M.A.Farnum, H.Xu, D.K.Agrafiotis. *Protein Sci.*, **12**, 1604 (2003)
85. D.K.Agrafiotis. *Protein Sci.*, **6**, 287 (1997)
86. P.Demartines, J.Herault. *IEEE Trans. Neural Netw.*, **8**, 148 (1997)

NONLINEAR MAPPING TECHNIQUES FOR PREDICTING THE PHARMACOLOGICAL PROPERTIES OF CHEMICAL COMPOUNDS

Ya.A.Ivanenkov, E.V.Bovina, K.V.Balakin

Institute of Physiologically Active Compounds, Russian Academy of Sciences

1, Pr. Severnyi, 142432 Chernogolovka Moscow Region, Russian Federation, Fax +7(496)524-9508

The state of the art of studies in the development and application of a special group of data analysis methods, namely, mapping algorithms, which allow decreasing the dimensionality of multidimensional spaces of the properties of analyzed objects (series of chemical compounds) to their two- or three-dimensional projections (images) with retention of the initial space topology is considered. It is shown how one can use such maps to visually analyze complex relationships between the properties and structural features of compounds, which is used for predicting the pharmacologically significant properties of chemical compounds.

Bibliography — 86 references.

Received 29th January 2009