УДК 541.6

ОЦЕНКА СПОСОБНОСТИ К ПРОТОНИРОВАНИЮ НЕКОТОРЫХ ОРГАНИЧЕСКИХ ОСНОВАНИЙ ПО ИХ СТРУКТУРНЫМ ФОРМУЛАМ

Р.А. Осипов, аспирант, *Н.С. Рукк, доцент, М.И. Скворцова, профессор, *В.В. Замалютин, студент, *А.Ю. Скрябина, аспирант

кафедра Высшей и прикладной математики

**кафедра Неорганической химии им. А.Н. Реформатского*

МИТХТ им. М.В. Ломоносова Москва, 119571 Россия

e-mail: skvorivan@mail.ru

В рамках статистического подхода к моделированию связи «структура–свойство» построено около 980 тысяч линейных уравнений для параметра $\lg K_a$ (K_a *– константа кислотности*) *ряда органических оснований. Для построения этих уравнений использовались различные обучающие выборки соединений и различные наборы молекулярных топологических дескрипторов. Из полученного множества моделей по определенному критерию отобрано около 90 наилучших моделей, используемых в дальнейшем для*: *а*) *прогнозирования величин* pK_a *некоторых соединений, не входящих в обучающие выборки*; *б*) *выявления структурных особенностей, наиболее существенно влияющих на* pK_a.

Ключевые слова: *способность к протонированию, константы кислотности, органические основания, модели связи «структура–свойство», молекулярные графы.*

Введение

Проблема моделирования связи между структурой и свойствами органических соединений является одной из важнейших математических задач современной теоретической химии. Найденные закономерности позволяют прогнозировать свойства химических соединений непосредственно по их структуре, минуя эксперимент, и могут быть использованы для планирования целенаправленного поиска соединений с заданным набором свойств. Согласно Программе фундаментальных научных исследований государственных академий наук на 2013-2020 гг., одним из направлений научных исследований в области химических наук и наук о материалах является «…обнаружение и изучение зависимостей «структура–свойство» в целях получения новых фундаментальных знаний о химической структуре и свойствах веществ» [1].

Наиболее распространенным подходом к моделированию связи «структура-свойство» является так называемый статистический подход, суть которого заключается в следующем. Имеется выборка соединений с известными численными значениями некоторого свойства y этих соединений. Структура соединений описывается при помощи набора молекулярных параметров $x_1, x_2, ..., x_n$, в качестве которых могут быть использованы топологические, электронные, геометрические характеристики молекул или значения каких-либо их физико-химических свойств. Как правило, математическая модель связи «структура-свойство» в рамках этого подхода имеет вид уравнения $y = f(x_1, x_2, ..., x_n)$, связывающего свойство y и параметры $x_1, x_2, ..., x_n$ при помощи некоторой функции f. Тип функции f предполагается известным, однако f зависит от ряда подгоночных параметров. Эти параметры подбираются по известным данным для заданной выборки соединений так, чтобы вышеуказанное соотношение выполнялось бы как можно более точно на этой выборке. В качестве количественной характеристики точности такой аппроксимации может служить, например, средняя (или максимальная) относительная ошибка расчета свойств соединений исходной выборки. Для этой цели могут быть также использованы коэффициент корреляции и среднеквадратичное отклонение для линейной регрессии, построенной по наборам экспериментальных и расчетных значений свойств рассматриваемых соединений выборки. Критерии, согласно которым достигнутая точность аппроксимации считается удовлетворительной, зависит от класса соединений, рассматриваемого свойства и поставленной химической задачи. В ряде случаев оценивается также и прогнозирующая способность построенной модели. Для этого составляется некоторая «тестовая» выборка соединений с известными значениями рассматриваемого свойства, не использованная для построения модели. Для соединений тестовой выборки рассчитываются значения свойств, строится корреляция между экспериментальными и расчетными значениями, и затем оцениваются ее статистические характеристики. Если модель признана удовлетворительной (по каким-либо критериям), то она используется в дальнейшем для прогноза свойств соединений, для которых отсутствуют экспериментальные данные.

Важное место в исследованиях связи «структура–свойство» занимают способы количественного описания структуры молекул, т. е. выбор параметров $x_1, x_2, ..., x_n$. Для вычисления

топологических параметров $x_1, x_2, ..., x_n$ структурные формулы органических соединений представляют в виде меченых или взвешенных графов, вершины и ребра которых соответствуют атомам и связям молекулы, а метки (веса) вершин и ребер кодируют атомы и связи различной химической природы. В качестве $x_1, x_2, ..., x_n$ используются инварианты этих графов.

В процессе построения моделей в рамках статистического подхода возникают проблемы выбора: 1) способа представления химических структур в виде меченых (взвешенных) графов; 2) инвариантов графов $x_1, x_2, ..., x_n$; 3) функции f. Эти проблемы связаны с тем, что в задачах такого типа заранее неизвестно, от каких именно структурных особенностей и каким образом зависит данное свойство.

Постановка задачи

Способность органических соединений к протонированию (количественно характеризуемая величиной $pK_p = -pK_a = \lg K_a$, где K_p – константа протонирования, K_a – константа кислотности), является их важным физико-химическим свойством. Однако экспериментальное определение этих констант связано со значительными техническими трудностями. Поэтому разработка любых методов теоретической оценки параметра pK_a непосредственно по структурной формуле молекулы весьма актуальна.

В настоящей работе поставлены следующие задачи:

1) в рамках описанного выше статистического подхода к моделированию связи «структура–свойство», используя литературные данные, получить ряд уравнений, описывающих зависимость параметра pK_a от различных структурных характеристик молекул;

2) на основе анализа всей совокупности этих моделей выявить наиболее существенные структурные характеристики, влияющие на значение pK_a;

3) оценить величины pK_a для двух конкретных соединений, являющихся производными пиразолона – (а) 5-метил-2-фенил-4Н-пиразолона-3 и (б) 2-(4-хлорофенил)-5-метил-4Н-пиразолона-3, для которых они не известны, для объяснения их относительно низкой склонности к комплексообразованию (по сравнению с 1-фенил-2,3-диметилпиразолоном-5 (антипирином), способным входить в состав различных комплексов *d*- и *f*-элементов).

Отметим, что производные пиразолона обладают противораковой и противовирусной активностью, а также являются атиоксидантами. Кроме того, известно, что в ряде случаев биологическая активность соединений может быть усилена за счет синергетического эффекта при образовании их комплексов с *d*- и *f*-элементами. Поэтому исследование различных физико-химических свойств соединений этого класса, связанных определенным образом с их биологической активностью, представляет несомненный интерес.

В качестве исходных данных для решения вышеуказанных задач использовалась выборка из 26 соединений, заданных своими структурными формулами, с известными значениями величин pK_a (рис. 1, 2, табл. 1,).

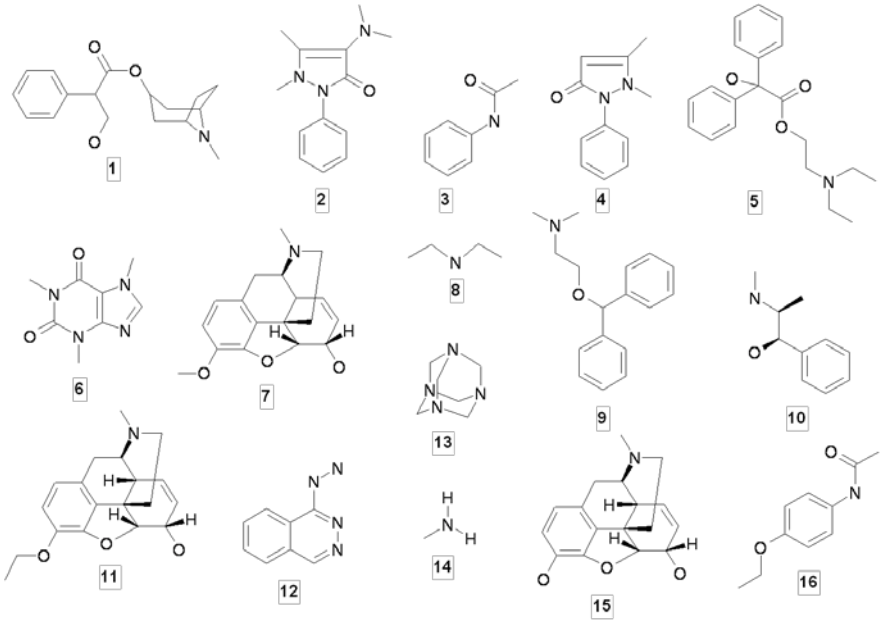

Рис. 1. Строение молекул **1–16** из таблицы 1.

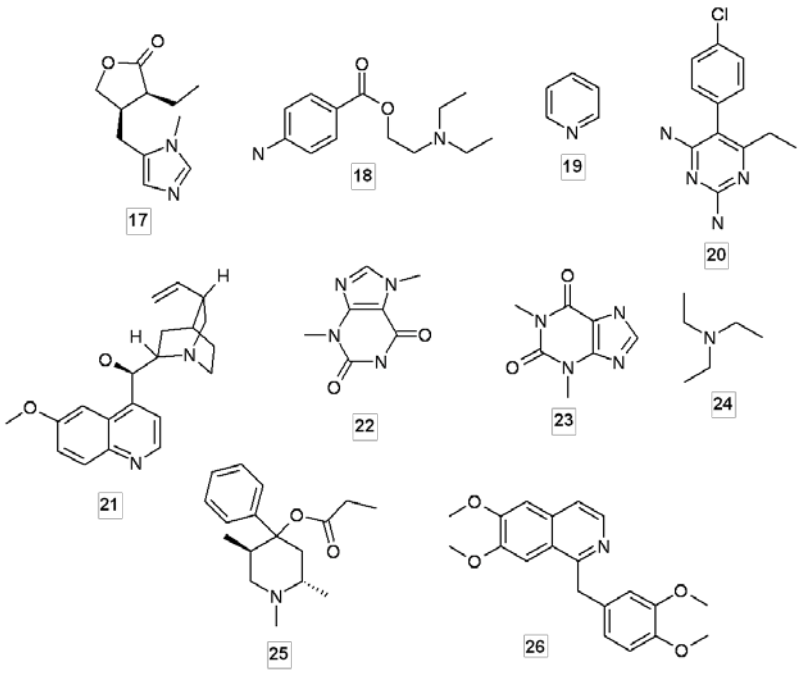

Рис. 2. Строение молекул **17–26** из таблицы 1.

Таблица 1. Соединения, использованные для построения моделей [2, 3]

№	Брутто-формула	Название	pK_a
1	$C_{17}H_{23}NO_3$	(8-метил-8-азабицикло[3.2.1]октил-3)3-гидрокси-2-фенилпропаноат (атропин)	9.65
2	$C_{13}H_{17}N_3O$	4-диметиламино-1,5-диметил-2-фенил-пиразолон-3 (пирамидон)	4.84
3	C_8H_9NO	N-фенилацетамид (антифебрин)	1.4
4	$C_{11}H_{12}N_2O$	1-фенил-2,3-диметилпиразолон-5 (антипирин)	2.2
5	$C_{20}H_{25}NO_3$	2-диэтиламиноэтил-2-гидрокси-2,2-дифенил ацетат (амизил)	8.5
6	$C_8H_{10}N_4O_2$	1,3,7-триметилпурин-2,6-дион (кофеин)	0.61
7	$C_{18}H_{21}NO_3$	(5-α,6-β)-3-метокси-17-метил-7,8-дидегидро-4,5-эпоксиморфинан-6-ол	7.95
8	$C_4H_{11}N$	диэтиламин	10.7
9	$C_{17}H_{21}NO$	2-[ди(фенил)метокси]-N,N-диметилэтанамин (димедрол)	8.2
10	$C_{10}H_{15}NO$	(1R,2S)-2-метиламино-1-фенилпропан-1-ол	9.66
11	$C_{19}H_{23}NO_3$	этилморфин	7.9
12	$C_8H_8N_4$	1-гидразинофталазин (апрессин)	7.1
13	$C_6H_{12}N_4$	гексаметилентетрамин (уротропин)	4.9
14	CH_3N	метиламин	10.6
15	$C_{17}H_{19}NO_3$	морфин	7.87
16	$C_{10}H_{13}NO_2$	1-этокси-4-ацетаминобензол (фенацетин)	2.2
17	$C_{11}H_{16}N_2O_2$	(3S, 4S)-3-этил-4-[(3-метилимидазол-4-ил) метил]оксолан-2-он (пилокарпин)	6.85
18	$C_{13}H_{20}N_2O_2$	2-диэтиламиноэтил-4-аминобензоат (новокаин)	8.85
19	C_5H_5N	пиридин	5.31
20	$C_{12}H_{13}ClN_4$	5-(4-хлорофенил)-6-этилпиримидин-2,4-диамин	5.6
21	$C_{20}H_{24}N_2O_2$	(R)-(6-метокси-4-хинолил)-[(2S,5R)-5-винил-2-хинуклидинил]-метанол (хинин)	8.0
22	$C_7H_8N_4O_2$	3,7-диметилпурин-2,6-дион (теобромин)	0.11
23	$C_7H_8N_4O_2$	1,3-диметил-7H-пурин-2,6-дион (теофиллин)	2.6
24	$C_6H_{15}N$	триэтиламин	10.7 (10.78)
25	$C_{17}H_{25}NO_2$	(2S,5R)-1,2,5- триметил-4-фенилпиперидин-4-ил пропионат (промедол)	8.4
26	$C_{20}H_{21}NO_4$	6,7-диметокси-1-(3,4-диметоксибензил) изохинолин (папаверин)	5.9

Способы представления молекул и выбор молекулярных параметров

Для представления молекул в виде графов использовались графы следующих двух видов:

1) простые графы, вершины которых соответствуют неводородным атомам молекулы, а ребра соответствуют связям между соответствующими атомами; веса вершин считаются равными нулю, а веса ребер считаются равными единице;

2) взвешенные графы, которые строятся так же, как и графы в п. 1), но веса вершин равны числам $Z-h$, где Z – число валентных электронов соответствующего атома, а h – число атомов водорода, связанных с этим атомом.

Параметры $x_1, x_2, ..., x_n$ отбирались из следующего множества 15 параметров, вычисляемых или по структурной формуле молекулы, или по соответствующим молекулярным графам (простому или взвешенному):

1) количество атомов водорода в молекуле;

2) отношение числа атомов углерода к общему числу неводородных атомов молекулы;

3) число атомов хлора;

4) число простых циклов в структурной формуле молекулы;

5) индекс Винера W молекулярного графа, построенного без учета атомов водорода;

6) отношение индекса W к общему числу неводородных атомов молекулы;

7) обобщенный взвешенный индекс связности графа молекулы, учитывающий наличие разнообразных подграфов;

8), 9) минимальное и максимальное собственные числа матрицы смежности простого молекулярного графа, построенного без учета атомов водорода;

10) разность между максимальным и минимальным собственными числами матрицы, указанной в пп. 8), 9);

11) сумма положительных собственных чисел матрицы, указанной в пп. 8), 9);

12)–15) – параметры, аналогичные параметрам из пп. 8)–11), но вычисляемые по взвешенной матрице смежности молекулярного графа.

Поясним выбор такого исходного набора параметров [4]. Параметры, равные числу фрагментов определенного вида (например, атомов), очень часто применяются при построении моделей связи «структура-свойство». Формулы для расчета свойств соединений, в которых величина свойства представляется в виде суммы вкладов отдельных фрагментов, так называемые аддитивные схемы, издавна используются в теоретической химии. Однако такие схемы дают удовлетворительные результаты лишь для структурно-родственных соединений и, как правило, содержат довольно много параметров. Поэтому при моделировании используются некоторые модификации этих параметров. Например, один из способов такой модификации заключается в следующем: находят общее количество n_1 фрагментов определенного типа (например, неводородных атомов), затем находят общее количество n_2 фрагментов некоторого подтипа (например, атомов углерода) и в качестве параметра используют величину n_2/n_1. Отметим, что для количественной характеристики способности молекулы к образованию водородных связей (или проявлению кислотных свойств) можно использовать параметр n_H, равный числу атомов водорода в молекуле или параметр n_H/n, где n – общее число атомов.

Очевидно, что наличие или отсутствие в молекуле циклов определенным образом влияет на ее свойства. Поэтому мы рассматривали и такой параметр, как число циклов в молекуле.

Индекс Винера W определяется по следующей формуле:

$$W = \sum_{i<j} d_{ij},$$

где d_{ij} – расстояние между i-ой и j-ой вершинами графа. Индекс W может служить количественной мерой компактности (или разветвленности) молекулы. Если рассмотреть все простые графы-деревья с одним и тем же числом вершин n, то W принимает свои экстремальные значения на наиболее и наименее компактных графах. Однако сравнение компактностей молекул при помощи W оправдано лишь для молекул с одним и тем же числом атомов n. В общем случае целесообразно использовать «нормированный» индекс Винера, т.е. величину W/n или W/n^p при некотором $p>1$.Отметим, что индекс Винера находит применение во многих корреляциях «структура-свойство» для разнообразных свойств и классов соединений.

Обобщенный взвешенный индекс связности графа молекулы (без атомов водорода), учитывающий наличие разнообразных подграфов, был построен следующим образом:

$$\chi^v = \Sigma (v_1 v_2 \ldots v_k)^{-0.5}.$$

Здесь суммирование распространено на все подграфы молекулярного графа, представляющие собой цепочки длины 1, 2, 3, а также граф-звезду с 3 ребрами; $v_1, v_2, \ldots, v_k$ – веса вершин конкретного подграфа, где $v_j = Z_j^v - h_j$, Z_j^v – число валентных электронов соответствующего неводородного атома, h_j – число атомов водорода, связанных с ним.

Первоначально индекс связности (индекс Рандича χ) был определен для простых молекулярных графов, соответствующих углерод-

ным остовам алканов. При этом в качестве подграфов рассматривались все цепочки длины 2 (т.е. ребра графа), а в качестве весов вершин подграфов v_j – их степени в исходном графе. Индекс χ может быть использован в качестве количественной меры разветвленности ациклических молекул, так как он принимает свои наибольшее и наименьшее значения на наименее и наиболее разветвленных (с интуитивной точки зрения) молекулах с фиксированным числом атомов углерода. В дальнейшем были предложены различные модификации индекса χ. В частности, в качестве весов вершин графа стали использоваться другие атомные параметры, отражающие наличие гетероатомов и кратных связей. Примером таких весов служат указанные выше величины $v_j=Z_j^v-h_j$. Другое обобщение индекса χ связано с использованием в формуле для χ каких-либо других подграфов, отличных от цепочек длины 2, например, цепочек произвольной длины $h>2$. Отметим, что при построении индекса связности по вышеприведенной формуле обычно рассматривается только один вид подграфов, и для разных видов подграфов получают разные индексы. Индексы связности очень часто используются при моделировании связи «структура-свойство» для различных свойств и классов соединений [4].

Параметры спектрального типа 8–15 аналогичны спектральным параметрам, используемым в квантовой химии для описания электронного строения молекул (см., например, [4]).

Метод построения моделей и оценка их качества

Одним из традиционных способов построения линейной модели связи «структура-свойство» на основе заданной выборки соединений и некоторого заранее выбранного набора молекулярных парметров является метод пошаговой линейной регрессии. При этом выявляется относительно небольшой набор параметров, являющихся существенными для данного свойства и позволющих получить модель заданной точности.

Однако опыт показывает, что могут быть построены модели на каких-либо иных наборах параметров (из числа первоначально выбранных), которые на исходной выборке уступают по точности оптимальной модели, но дают более точный прогноз на соединениях, не включенных в исходную выборку. Кроме того, изменение обучающей выборки даже с сохранением прежнего набора параметров также приводит, вообще говоря, к изменению модели и, следовательно, к изменению прогнозов свойств новых соединений.

Так как основная цель моделирования в данной области – оценка свойств соединений, для которых отсутствуют экспериментальные данные, то представляется целесообразным строить много различных моделей, а не одну, варьируя как обучающую выборку, так и набор параметров, а для прогноза свойств усреднять результаты, полученные по разным моделям. При этом можно использовать какие-либо критерии для отбрасывания части моделей как неудовлетворительных, сокращая, таким образом, множество рассматриваемых вариантов. Кроме того, проанализировав полученные модели, можно выявить структурные параметры, наиболее существенные для данного свойства. При этом «существенными» мы называем те параметры, у которых наибольшая частота встречаемости в построенных удовлетворительных моделях.

Опишем метод построения достаточно большого множества моделей и оценки их качества, используемый в настоящей работе. Искомые модели строились следующим образом. Первоначально исходная совокупность соединений 30 раз разбивалась случайным образом на две равные части, одна из которых далее использовалась как обучающая выборка, а другая – как тестовая. Кроме того, из исходного множества 15 параметров формировались всевозможные наборы, содержащие от 1 до 15 параметров. Таких наборов $2^{15}-1=32767$, и они были занумерованы числами от 1 до 32767. Затем по каждой обучающей выборке и каждому набору параметров была построена линейная модель связи «структура-свойство» (т.е. модель вида $pK_a=b_0+b_1x_1+b_2x_2+...+b_nx_n$, где $x_1,x_2,...,x_n$ – выбранные параметры, а $b_0,b_1,b_2,...,b_n$ – некоторые константы, подбираемые по обучающей выборке соединений, $1\le n\le 15$). Очевидно, что общее количество таких моделей равно $30\cdot 32767=983010$.

Затем при помощи каждой построенной модели были проведены расчеты значений свойств соединений соответствующей тестовой выборки. Точность аппроксимации (для обучающей выборки) и точность прогноза (для тестовой выборки) оценивались по среднеквадратичным отклонениям s_1 и s_2 расчетных значений свойств от их экспериментальных значений для обучающей и тестовой выборки соответственно. Для количественной оценки s качества модели использовалась величина $0.5(s_1+s_2)$. Таким образом, каждому набору молекулярных параметров, использованному для построения 30 моделей на различных обучающих выборках, соответствует вектор с 30 компонентами, равными значениям величин s для 30 соответствующих моделей. Затем этот вектор нормировался на единицу, и вычислялось среднее арифметическое значение его новых компонент. Полученное число δ, названное нами «средняя нормализованная погрешность», характеризует

эффективность использования определенного набора молекулярных параметров при построении моделей. На рис. 3 представлена диаграмма распределений величин δ для всех 32767 рассматриваемых наборов (по вертикальной оси откладывается значение δ, а по горизонтальной оси – номер набора параметров; под горизонтальной осью указаны также числа параметров в наборах и номер первого по порядку набора с данным числом параметров).

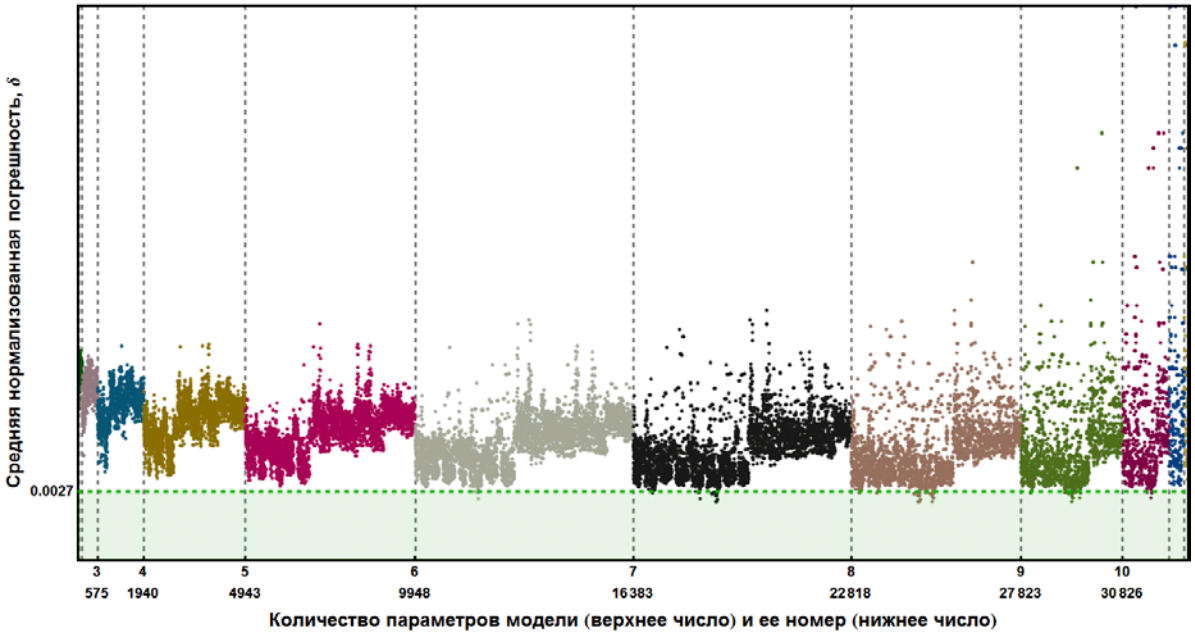

Рис. 3. Диаграмма распределения средней нормализованной погрешности для 32767 наборов молекулярных параметров.

Таким образом, в описанных выше исследованиях нами использовался один из возможных способов количественной оценки качества модели, позволяющий отобрать наилучшие модели из многих – так называемый «скользящий контроль». При этом рассматривалась та его разновидность, когда исходная выборка многократно случайным образом разбивается на две части – обучающую и тестовую выборки, и тестовая выборка содержит 50% от общего числа соединений [5]. Согласно [5], на практике в тестовую выборку включают, как правило, от 25 до 50% объектов исходного множества. Отметим, что величины s_1 и s_2, характеризующие модель на обучающей и тестовой выборке, вычисляются по одной и той же формуле, и в формулу для s они входят равноценным образом. Поэтому представляется естественным выбирать обучающую и тестовую выборку равноценными и по числу объектов в них, т.е. включать в них по 50% от общего числа соединений.

Выбор наилучших моделей и методика расчета pK_a

Визуальный анализ диаграммы на рис. 3 показывает, что имеется относительно небольшое количество наборов с малыми значениями δ ($\delta < 0.0027$). На рис. 3 соответствующие точки расположены ниже пунктирной горизонтальной линии δ=0.0027. Таких наборов оказалось 88, и они были отобраны для дальнейших исследований.

Для определения величины pK_a некоторого соединения использовалась следующая методика: значения pK_a вычислялись для него по всем моделям, построенным на основе отобранных 88 наборов молекулярных параметров, а в качестве прогнозируемой величины рассматривалось их среднее значение. Для оценки точности предлагаемого метода такие расчеты были проведены для всех соединений исходной выборки. Коэффициент корреляции между экспериментальными и расчетными значениями pK_a для этих соединений равен 0.9, что можно считать удовлетворительным для рассматриваемого свойства и исходных данных, полученных разными авторами и различными методами.

При помощи предложенного метода были рассчитаны величины pK_a для соединений (а) и (б), используемых нами в лабораторных исследованиях, для которых отсутствуют экспериментальные данные, но которые, по прогнозу [6] могут обладать биологической активностью.

Анализ моделей

Следует отметить также, что анализ полученных 88 моделей позволяет выявить основные структурные факторы, влияющие на величину рассматриваемого свойства. Для этой цели для каждого из 15 структурных параметров можно определить процент моделей, содержащих этот параметр, и затем упорядочить параметры по этим величинам. На рис. 4 представлена соот-

ветствующая диаграмма, показывающая частоту встречаемости (в процентах) каждого параметра в отобранных моделях. Согласно этой диаграмме, наилучшие восемь параметров – это параметры 1), 3), 6) – 11); при этом частота встречаемости параметров 1), 3), 6), 7) равна 100%. На диаграмме столбец для x_{15} отсутствует, так как x_{15} в этих моделях не встречается.

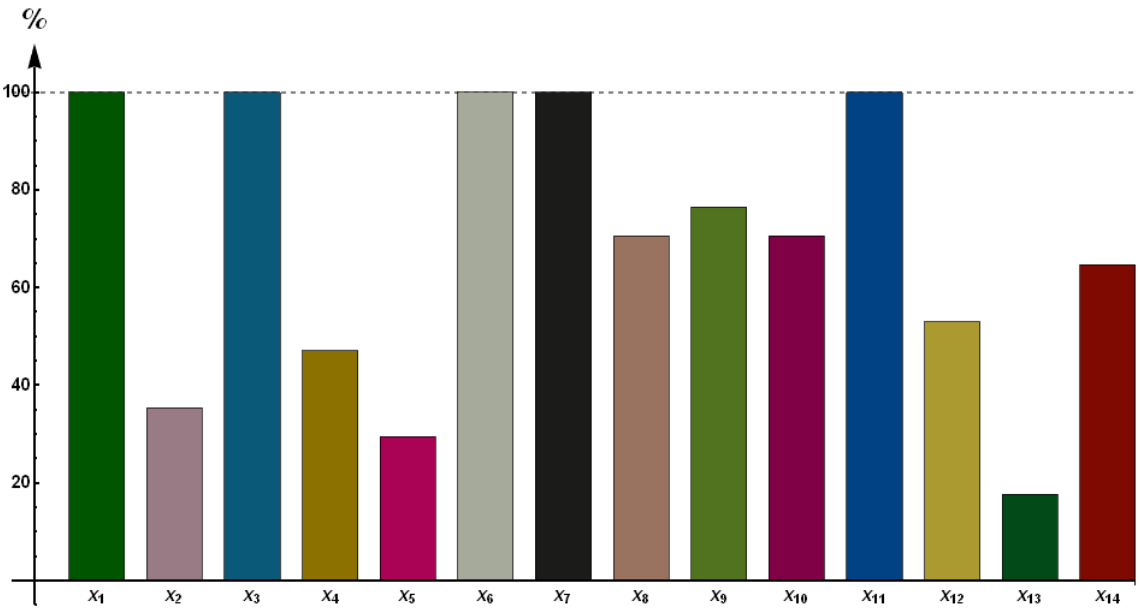

Рис. 4. Частоты встречаемости параметров $x_i\,(i=1,2,...,14)$ в наилучших 88 моделях.

Рассчитанные значения pK_a для указанных выше соединений (а) и (б) (3.35 и –3.39, соответственно) не позволяют в достаточной степени объяснить понижение способности к комплексообразованию указанных лигандов, однако квантово-химические расчеты энергии протонирования позволили [7, 8] классифицировать лиганды по их способности протонироваться и, следовательно, координироваться центральными атомами.

Программные средства для построения и анализа моделей связи «структура–свойство»

Для проведения исследований нами использовалась вычислительная программа *Wolfram Mathematica* 8. Эта программа позволяет создавать базы данных по химическим структурам и свойствам соединений, модифицировать эти базы и разбивать их на любые части, создавать программы вычисления различных структурных молекулярных параметров для соединений базы, строить модели (линейные и нелинейные) связи «структура-свойство» с любыми параметрами из числа запрограммированных и вычислять статистические характеристики этих моделей, получать различные графические представления результатов исследований (графики, диаграммы и т.д.), осуществлять прогноз свойств соединений по любой построенной модели [9].

ЛИТЕРАТУРА:

1. Программа фундаментальных научных исследований государственных академий наук на 2013-2020 гг. (утверждена распоряжением Правительства РФ от 3 декабря 2012 г. № 2237-р). http://government.ru/gov/results/21805/

2. Евстратова К.И., Гончарова Н.А., Соломко В.Я. Константы диссоциации слабых органических оснований // Фармация. 1968. Т. 17. С. 33–36.

3. Государственная фармакопея СССР / Глав. ред. М.Д. Машковский. – М.: Медицина, 1968. 1079 с.

4. Станкевич М. И., Станкевич И. В., Зефиров Н. С. Топологические индексы в органической химии // Успехи химии. 1988. Т. 57. С. 337–366.

5. Воронцов К.В. Лекции по методам оценивания и выбора моделей. 2007. http://ccas.ru/voron/download/Modeling.pdf.

6. Рукк Н.С., Скрябина А.Ю., Апрышко Г.Н. Поиск потенциальных противоопухолевых комплексных соединений редкоземельных металлов / Тез. докл. на VIII Всерос. научно-практ. конф. «Отечественные противоопухолевые препараты» // Рос. биотерапевт. журн. 2009. Т. 8. № 2. С. 14–15.

7. Rukk N., Shamsiev R., Skvortsova M., Osipov R., Zamalyutin V., Obukhova A. Basic properties of a number of organic ligands. Quantum-chemical calculatons and modelling the structure – protonation constant relationships / Book of Abstracts of the XIX Int. Conf. «Mathematics Education». Dubna, January 30 – February 3, 2012. – Dubna, 2012. P. 156.

8. Осипов Р.А., Рукк Н.С., Скворцова М.И., Михайлова Н.А. Исследование связи между структурой органических соединений и значениями их констант кислотности // Сб. трудов XXV Междунар. научной конф. «Математические методы в технике и технологиях – ММТТ-25» – Волгоград, 29–21 мая 2012 / под ред. А.А. Большакова. – Волгоград: Волгогр. гос. техн. ун-т, 2012. Т. 7. Секция 11. С. 5–8.

9. http://www.wolfram.com/mathematica/; http://vk.com/wolframmathematica.

ESTIMATION OF PROTONATION ABILITY OF SOME ORGANIC BASES BY THEIR STRUCTURAL FORMULAE

R.A. Osipov, N.S. Rukk, M. I. Skvortsova[@], V.V. Zamalyutin, A.Yu. Skryabina

M.V. Lomonosov Moscow University of Fine Chemical Technologies, Moscow, 119571 Russia

[@]*Corresponding author e-mail: skvorivan@mail.ru*

Searching the quantitative structure–property relationships (QSPR) is one of the most important tasks of the contemporary theoretical chemistry. The models obtained may be used for prognosis of the chemical substances properties on the basis of their structure and for searching compounds with predetermined properties. About 980,000 structure–property linear models have been constructed for parameter lgK_a (K_a – acidity constant) of a number of organic compounds. Different training and test sets obtained by means of multiple random halving the initial set of compounds have been used to desigh these models. Molecular descriptors have been selected from some set of topological molecular parameters reflecting different particularities of molecular structures. The Wiener index, the generalized weighted connectivity index, the number of hydrogen atoms, some spectral characteristics of graph, representing molecule, etc. are among these parameters. About 90 the best models have been selected on the basis of some quantitative criterium, characterizing the model precision both on training and test sets. These models have been used for evaluation of lgK_a for other compounds not included into the initial set of compounds, by calculation of lgK_a for every model and averaging the results obtained. Besides, the structural particularities possessing the most significant influence on the given property have been derived on the basis of analysis of these models. The computer program Wolfram Mathematica 8 has been used in this work.

***Key words**: protonation ability, acidity constants; organic bases; structure–property relations; molecular graphs.*