Глава 8. ПРОГНОЗ СВОЙСТВ ГУМУСОВЫХ КИСЛОТ С ИСПОЛЬЗОВАНИЕМ ДЕСКРИПТОРОВ СОСТАВА

В результате выполнения предыдущего блока исследований были созданы все необходимые предпосылки для решения важной теоретической и практической задачи – разработки прогностических моделей “состав – свойство”. Создание таких моделей может служить теоретическим базисом для разработки системы скрининга гумусовых кислот, основанного на определении их состава. Это позволит осуществлять целевой поиск препаратов, обладающих максимальными связывающими и/или детоксицирующими свойствами по отношению к различным экотоксикантам, не проводя экспериментов по связыванию или детоксикации. Появление таких моделей способствовало бы расширению сферы применения природного гуминового сырья (уголь, торф, сапропель) в целях рекультивации загрязненных сред, а также явилось бы необходимой предпосылкой для разработки лекарственных препаратов на основе гуминовых веществ.

Спецификой полученного в работе массива данных является высокая размерность матрицы дескрипторов состава (число столбцов > 20) и сравнимое с ней число препаратов в выборке (число строк 16-26 в зависимости от количества найденных констант связывания). Создание прогностических моделей с использованием таких массивов данных требует применения методов, позволяющих снизить размерность матрицы независимых переменных, в нашем случае – дескрипторов состава.

Для этой цели использовали метод множественной регрессии (МР), вводя ограничение на количество членов в уравнении регрессии и предусматривая выбор оптимальных дескрипторов путем их автоматического перебора (автор алгоритма и программы – А. В. Кудрявцев), и методы многокомпонентного анализа – регрессии на главных компонентах (РГК) и дробного метода наименьших квадратов (ДМНК). Оба метода позволяют снизить размерность матрицы исходных данных, разлагая ее на компоненты, не снижая при этом количество исходных дескрипторов. В методе РГК разложение исходной матрицы проводят только с учетом ее внутренней структуры [Дубров и др., 2000]. В методе ДМНК осуществляется совместное разложение как матрицы дескрипторов, так и прогнозируемых свойств [Geladi and Kowalski, 1986, Gunst and Maston, 1980]. Это повышает надежность получаемой прогностической модели [Lindberg at al, 1983].

В дальнейшей работе использовали все три метода. На выходе все они дают прогностическую модель, которая представляет собой линейную (МР,

РГК и ДМНК) или нелинейную (МР) функциональную зависимость прогнозируемого свойства от дескрипторов.

8.1 Корреляционная взаимосвязь дескрипторов состава и прогнозируемых свойств

Расчет прогностических моделей предваряли оценкой наличия корреляционных взаимосвязей между дескрипторами состава, набор которых для всех выборок был стандартным и включал все три уровня дескрипторов (за исключением выборки для ПАУ, где набор дескрипторов был дополнен за счет ε*), и прогнозируемыми свойствами – константами связывания и детоксикации Hg(II), ПАУ (Py, Flt, An):и атразина гумусовыми кислотами.

Проведение корреляционного анализа показало, что тесные корреляционные взаимосвязи между константами устойчивости комплексов Hg(II) с гумусовыми кислотами и интегральными дескрипторами состава всех уровней отсутствуют.

Для констант связывания и детоксикации ПАУ наблюдалось наличие тесной корреляционной взаимосвязи с прямыми и косвенными показателями содержания ароматических фрагментов в составе гумусовых кислот (C_{Ar}, $\sum C_{Ar}$, H/C, ε*). Пример такой корреляционной зависимости для констант связывания ПАУ приведен на рис. 8.1.

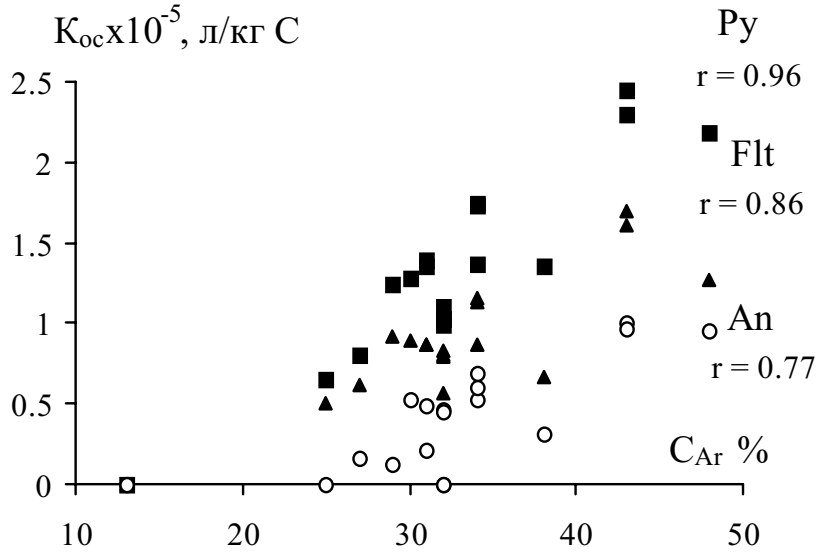


Рис. 8.1. Корреляционное поле для пары переменных "ΣC_{Ar} – K_{OC}" для выборки из 19 препаратов, использованной в экспериментах по связыванию ПАУ.

Особо следует отметить, что указанная зависимость наблюдалась для всех трех ПАУ, и для обоих типов констант – связывания и детоксикации. Наиболее тесной она была для наиболее гидрофобных Py и Flt. Полученные результаты хорошо согласуются с данными работ [Gauthier et al, 1987; Chin et

al, 1997], где авторами показано наличие тесной корреляции между содержанием ароматического углерода в гумусовых кислотах и их сродством к ПАУ.

Корреляционная взаимосвязь двух косвенных показателей ароматичности гумусовых кислот – H/C и ε* – с константами связывания Py, Flt и An была существенно слабее, чем для ^{13}C ЯМР-дескриптора. Значения коэффициентов корреляции для Py, Flt и An составили 0.85, 0.84, 0.76 и 0.66, 0.74, 0.55 для H/C и ε*, соответственно. Все зависимости являются значимыми при P = 0.95.

Корреляционное поле для K_{OC} всех исследованных ПАУ и M_w гумусовых кислот приведено на рис. 8.2. Как видно из представленных данных, явной тенденции увеличения степени сродства гумусовых кислот к ПАУ по мере возрастания ММ не наблюдалось ни для одного из трех ПАУ. Гумусовые кислоты с ММ ниже 10000 Да характеризовались весьма слабым сродством к ПАУ.

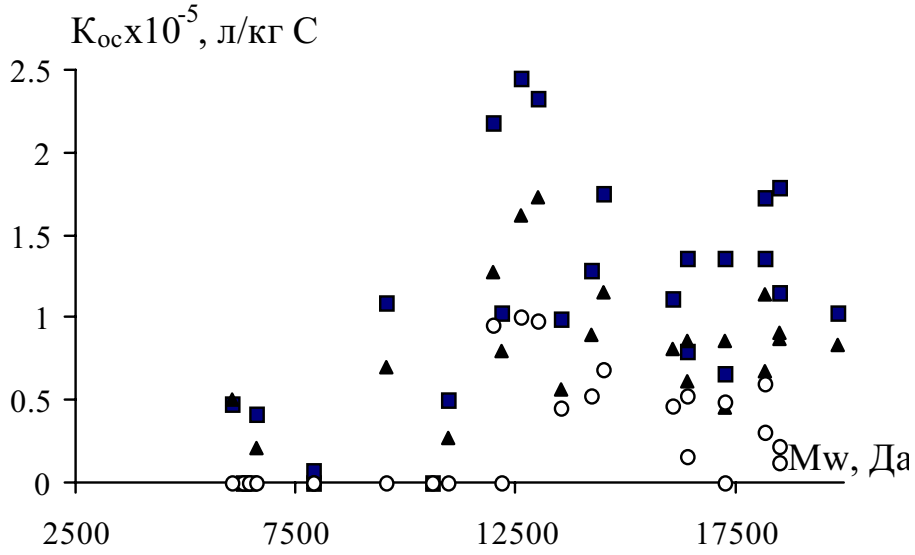


Рис. 8.2. Корреляционное поле для пары переменных "M_p-K_{OC}" для полной выборки из 26 препаратов, использованной в экспериментах по связыванию ПАУ (■ Py, ▲ Flt, O An).

Для оценки устойчивости полученных корреляционных взаимосвязей между константами связывания ПАУ и различными дескрипторами ароматичности ($\sum C_{Ar}/\sum C_{Alk}$, $\sum C_{Ar}$, C_{Ar}, H/C, ε*) гумусовых кислот и их M_w, была осуществлена процедура кросс-валидации. Эта часть работ подробно изложена в соответствующей публикации [Perminova et al, 1999]. Ее суть заключалась в следующем: исходную выборку препаратов (полная – 26 и охарактеризованная методом спектроскопии ЯМР ^{13}C – 19 препаратов) разбивали на непересекающиеся подмножества, сгруппированные по

сходству источника происхождения и/или фракционного состава. Указанные подмножества включали: 8 препаратов ГФ торфа (ГК+ФК), 8 ГК и 5 ФК почв. Одна из выборок была сформирована из 5 препаратов, не вошедших ни в одно из подмножеств (ГФК почв, 2 ГК угля, 2 ГФК вод) и обозначена как ГФК*. Полученные зависимости между $К_{ос}$ Ру и указанными дескрипторами гумусовых кислот приведены на рис. 8.3. Аналогичные результаты были получены для Flt и An.

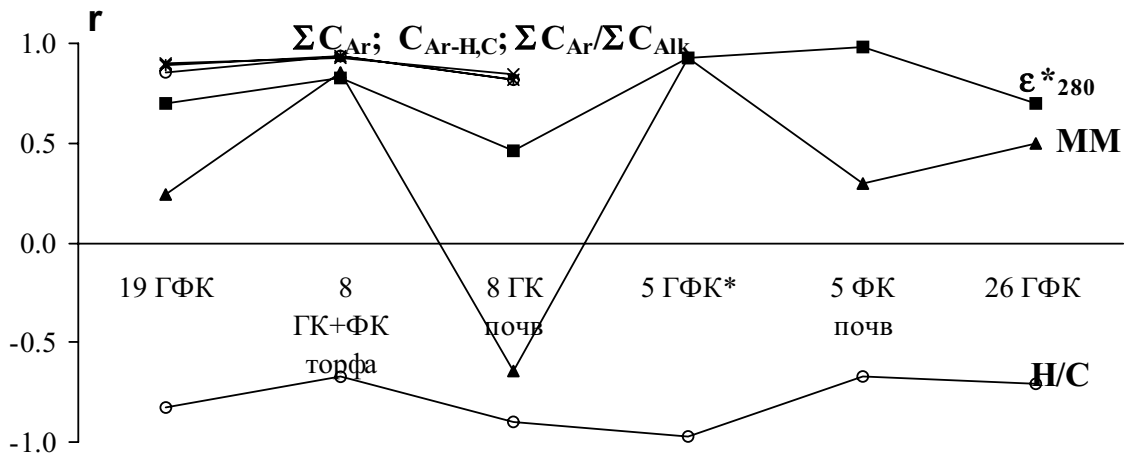


Рис. 8.3. Зависимость коэффициента корреляции, характеризующего взаимосвязь между $К_{ос}$ пирена и исследуемыми дескрипторами, от состава выборки препаратов гумусовых кислот.

Как видно из представленных данных, наибольшей устойчивостью характеризовалась корреляционная взаимосвязь между $К_{ОС}$ и $^{13}С$ ЯМР дескрипторами ($С_{Ar}$, $\Sigma С_{Ar}$, $\Sigma С_{Ar}/\Sigma С_{Alk}$) гумусовых кислот. Довольно высокая устойчивость наблюдалась и для соотношения Н/С и ε*. Наиболее чувствительным дескриптором к изменению состава выборки препаратов гумусовых кислот оказалась $М_р$. Как видно из рисунка, корреляция между $К_{ОС}$ и $М_р$ является значимой только для выборок из 8 препаратов торфяных ГФК, 5 ГФК* и 8 ГК почв. Причем в двух первых случаях корреляция прямая, а для ГК почв – обратная. Следовательно, $М_р$ можно использовать для прогноза сродства к ПАУ только для гумусовых кислот аналогичных по происхождению и/или фракционному составу.

Полученные зависимости имеют очевидный физический смысл, свидетельствуя о ведущей роли гидрофобных взаимодействий в процессах связывания и детоксикации ПАУ гумусовыми кислотами: чем больше содержание ароматических фрагментов, тем выше гидрофобность макромолекул гумусовых кислот и выше их сродство к гидрофобным молекулам ПАУ.

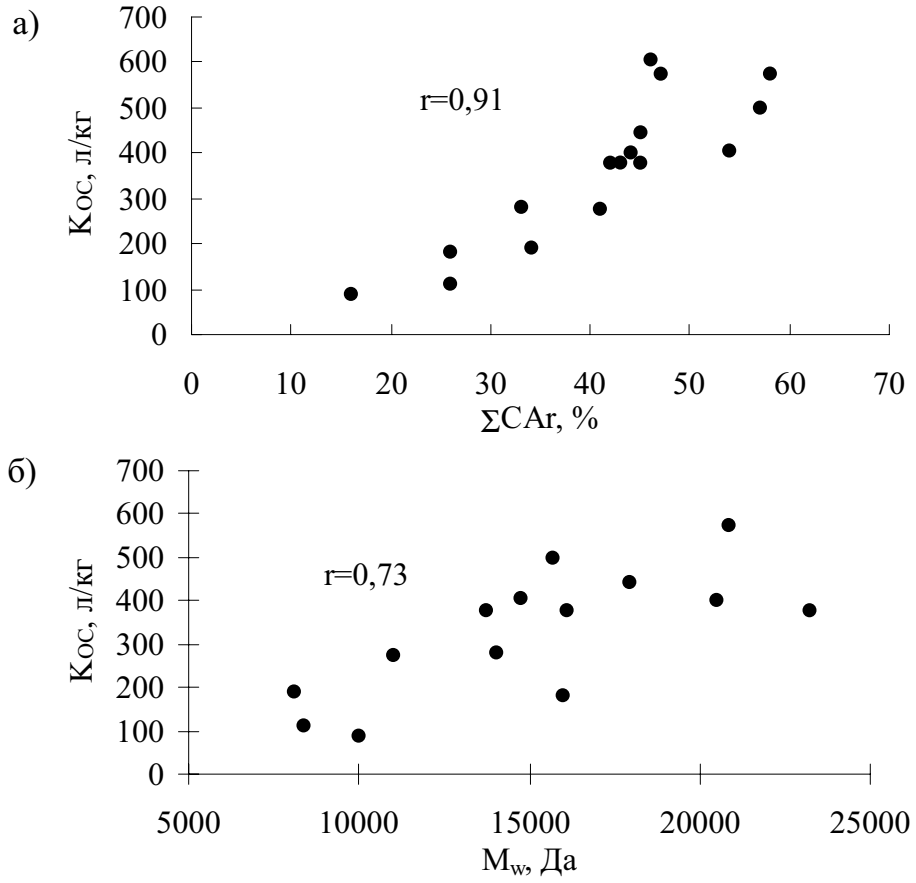


Рис. 8.4. Корреляционное поле для пары переменных "ΣC_{Ar} – K_{OC}" (а) и "M_w – K_{OC}".(б) для выборки из 16 препаратов, использованной в экспериментах по связыванию атразина.

Наличие наиболее тесной корреляции между K_{OC} и содержанием углерода в составе ароматических фрагментов было установлено и для атразина (рис. 8.4), что может трактоваться как проявление действия сходного с ПАУ механизма связывания – гидрофобного. Для K_{OC} атразина также наблюдалась и значимая (при $P = 0.95$) корреляционная взаимосвязь с M_w гумусовых кислот. Однако в отличие от ПАУ, для K_{OC}^D атразина корреляция с содержанием ароматических фрагментов отсутствовала. Эти константы не имели значимой корреляции ни с одним из интегральных дескрипторов состава. Однако тесно коррелировали с содержанием низкомолекулярных фракций (<5000 Да – предела проницаемости клеточных мембран [Del Agnola et al., 1986] в образце гумусовых кислот ($r = 0.93$). Полученные данные могут свидетельствовать о различии механизмов связывания и детоксикации атразина гумусовыми кислотами.

Как было показано в Главе 7, в случае атразина детоксикация обусловлена стимулирующим действием гумусовых кислот на тест-объект, приводящим к усилению его резистентности к химическим стрессорам. Данный эффект неоднократно описывался в литературе для высших растений [Христева, 1973] Однако механизм его до сих пор не ясен. Существенный

прогресс в этой области может быть достигнут с помощью установления корреляционных соотношений “строение – детоксицирующая способность” и “строение – физиологическая активность” гумусовых кислот.

Для описания конкретных типов взаимосвязей между дескрипторами состава и связывающими/детоксицирующими свойствами гумусовых кислот использовали методы регрессионного анализа.

8.2 Прогностические модели, полученные методом множественной регрессии

Одно из основных требований, которое предъявляет метод МР к калибровочной выборке, используемой для расчета прогностической модели, – это существенное превышение количества прогнозируемых признаков (известных констант связывания/детоксикации) над числом независимых переменных (дескрипторов состава) [Geladi and Kowalski, 1986]. Принимая во внимание размерность матрицы интегральных дескрипторов состава, в виде которой записывается информация о строении каждого препарата гумусовых кислот ($n > 20$), оптимальный объем выборки должен был бы составлять 50-60 препаратов. Однако в связи с трудоемкостью формирования обширных выборок препаратов это требование весьма сложно реализовать на практике. Каждый из полученных нами массивов прогнозируемых свойств содержал 15-25 значений, что не позволяло удовлетворить сформулированное выше условие.

Для решения указанной дилеммы использовали вариант метода МР с ограничением на количество членов в регрессии (не больше четырех), но при этом предусматривали поиск оптимальных наборов дескрипторов путем перебора. Для этого нами был реализован алгоритм МР (автор – А.В. Кудрявцев), позволяющий проводить автоматический перебор всех возможных сочетаний дескрипторов (в том числе комбинированных – произведения и отношения исходных дескрипторов) с ограничением на число членов в полиномах (до 4-х исходных и 2 комбинированных дескриптора). Для расчетов моделей задавали полиномиальный вид уравнения регрессии:

$$y = b_0 + b_1x_1 + \mathrm{b2x2} + \ldots + b_\mathrm{n}x_\mathrm{n} \qquad (8.1)$$

где y – признак свойства, x_i – i-й дескриптор, в качестве которого могут выступать как исходные величины, так и их комбинации (x_1x_2, x_1/x_2, $x_1x_2^2$ и т.д.). Значения коэффициентов b_i рассчитывали по методу МНК.

На выходе модели отбирали 10 лучших полиномов. Качество моделей оценивали с помощью параметров R^2 и Q^2, характеризующих, соответственно, описательную и прогностическую способность. R^2 представляет собой долю объясненной дисперсии и рассчитывался традиционно:

$$R^2 = 1 - s_r^2/s_t^2 \qquad (8.2)$$

где s_r^2 – остаточная (необъясненная в рамках данной модели) дисперсия, а s_t^2 – общая дисперсия описываемого свойства.

Чем ближе r^2 к 1, тем лучше модель описывает имеющиеся данные, но может при этом плохо предсказывать данные, которые не использовались при ее построении. Если разделить имеющуюся выборку препаратов (признаки их свойств и дескрипторы) на обучающую и контрольную выборки, то для характеристики предсказывающей способности можно рассчитать R^2 для контрольной выборки, что будет более адекватной оценкой предсказывающей способности модели.

Однако далеко не всегда имеется возможность выделить контрольную выборку, особенно при работе с выборками малых объемов. В то же время, чем больше препаратов использовано при построении модели, тем лучше ее качество. Поэтому альтернативным методом оценки прогностической способности моделей является расчет Q^2 в результате кросс-валидации (перекрестного оценивания) [Krzanowski, 1987; Geladi and Kowalsi, 1986a]. Для этой цели рассчитывают модель, используя весь набор препаратов. Затем исключают один или несколько препаратов, пересчитывают модель, предсказывают значения свойств для исключенных препаратов и считают дисперсию между предсказанным и известным свойством. Данную процедуру повторяют, исключая из исходного набора следующий препарат. Полученную таким образом среднюю непредсказанную дисперсию, используют для расчета Q^2:

$$Q^2 = 1 - s_p^2/s_t^2 \qquad (8.3)$$

где s_p^2 – средняя непредсказанная дисперсия; s_t^2 – общая дисперсия описываемого свойства. Чем ближе Q^2 к 1, тем лучше модель предсказывает имеющиеся данные. Таким образом, Q^2 может служить оценкой прогностической способности моделей, использоваться для их оптимизации и сравнения.

Для расчета моделей использовали весь набор дескрипторов. Результаты расчета лучших полиномов по методу МР приведены в табл. 8.1. Полиномы с 3-4 членами (исходные дескрипторы), как правило, предсказывают лучше, чем с двумя комбинированными дескрипторами (произведения и отношения исходных дескрипторов). Увеличение числа дескрипторов в наборе приводит к улучшению качества модели, при этом максимальные Q^2 наблюдаются для смешанного набора дескрипторов всех трех уровней.

Таблица 8.1

Наилучшие прогностические модели МР с тремя исходными параметрами

Токсикант	Полином	Q^2	R^2
Hg(II)	$lgK_{РСЦ} = 13.5 - 6010/Q_{50} + 5.40\times C_{Ar}O + 0.106\times CHO$	0.79	0.89
	$lgK_{РСЦ}^{D} = 10.7 + 146/C_{Ar} - 26.3\times C + 0.000140\times Q_{25}$	0.59	0.75
	$lgK_{РСЦ}^{B} = 14.1 - 0.552/E - 20.6\times C + 2.77\times10^{-5}\times M_z$	0.65	0.80
Py	$K_{OC} = 170000 - 10240/CHO - 48100\times H + 836000\times C_{Ar}$	0.85	0.89
	$K_{OC}^{D} = 0.172 + 1900/Q_{25} - 2100/M_n + 0.345/(H/C)$	0.83	0.85
Атразин	$K_{OC} = -34600 + 579000/(H/C) - 242\times Q_{25} + 71.3\times Q_{75}$	0.57	0.76
	$K_{OC}^{D} = 513 - 1620\times C_{Ar}O + 0.0196\times M_z - 134\times M_w/M_n$	0.90	0.94

* E – эксцесс, Q – квантили ММР.

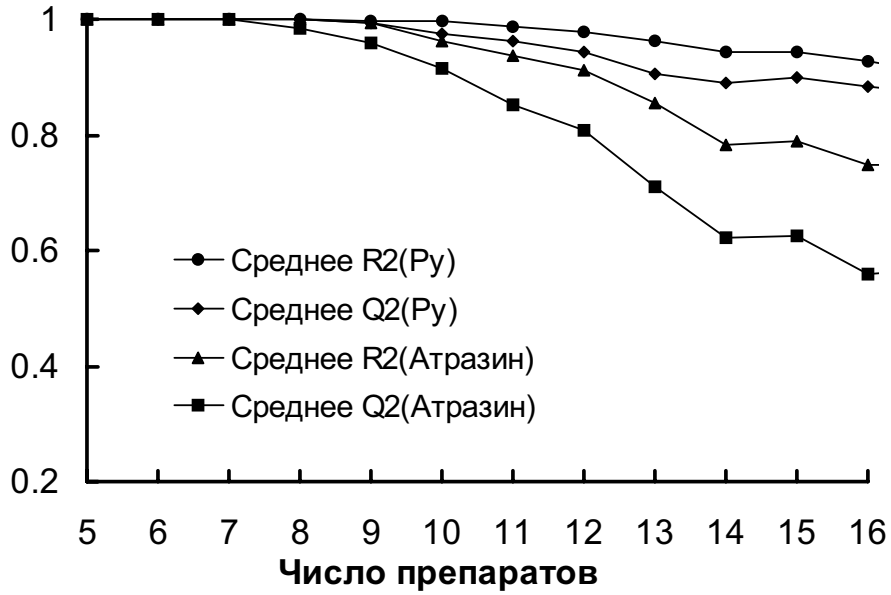


Рис. 8.5. Зависимость средних R^2 и Q^2 (усреднение по пяти моделям со случайным образом удаленными препаратами) от размера выборки, использованной для моделирования K_{OC}^{D} атразина и K_{oc} пирена.

Из прогнозируемых свойств наиболее удовлетворительные результаты получены для K_{OC} и K_{OC}^{D} всех трех ПАУ и для K_{OC}^{D} атразина. Использование метода МР позволило выявить форму зависимости между K_{OC}^{D} атразина и ММ характеристиками гумусовых кислот. Удалось также описать зависимости $K_{РСЦ}$ от дескрипторов состава.

Анализ характера изменения величин R^2 и Q^2 в зависимости от размера исходной выборки препаратов позволил обнаружить их завышение для прогностических моделей, построенных для малых выборок, что может

указывать на случайность подбора дескрипторов в таких полиномах. Данное предположение было проверено с помощью численного эксперимента путем последовательного исключения препаратов из выборки, которое проводилось случайным образом. Результаты в виде зависимости R^2 и Q^2 от числа препаратов (на примере K_{OC}^{D} атразина и K_{OC} Ру) представлены на рис. 8.5.

Из характера полученных зависимостей видно, что при $n < 15$ значения R^2 и Q^2 начинают резко возрастать. Это позволяет сделать вывод о том, что для построения адекватных прогностической модели на основании данных аналогичного качества с использованием алгоритма МР размер выборки должен быть не менее 20 препаратов.

8.3 Прогностические модели, полученные методами многокомпонентного анализа

С целью снижения размерности исходной матрицы дескрипторов состава без опасности потери информативных признаков использовали методы многокомпонентного анализа – РГК и ДМНК. Применение указанных методов к массивам разнородных данных (например, как в нашем случае – при использовании дескрипторов состава, определяемых разными методами) включает обязательную процедуру их центрирования и масштабирования [Стьюпер и др., 1982], что позволяет заменить исходную матрицу данных на матрицу их дисперсий.

Согласно методу РГК, полученную матрицу разлагают на сумму компонент по принципу максимального объяснения дисперсии исходных данных [Дубров и др., 2000; Geladi and Kowalski, 1986]. Выбор компонент обусловлен спецификой внутренней структуры матрицы дескрипторов. К преимуществам данного метода относится простота реализации, высокая скорость расчета, использование информации о специфике матрицы дескрипторов, возможность получения дополнительных сведений об изучаемой зависимости строение – свойство за счет анализа состава главных компонент и степени их влияния на прогнозируемое свойство. К недостаткам метода относится то, что не используется информация о структуре матрицы свойств. Как следствие, выделяемые главные компоненты могут описывать факторы, непосредственно не связанные с прогнозируемым свойством.

Указанного недостатка лишен метод ДМНК. В отличие от РГК, он предусматривает выбор главных компонент путем минимизации необъясненной дисперсии не только матрицы независимых, но и зависимых переменных [Geladi and Kowalski, 1986; Clementi et al., 1986; Sjöström et al., 1983]. Это повышает надежность получаемой прогностической модели, снижает ее чувствительность к появлению в калибровочной выборке образца сравнения с несколько отличными свойствами [Lindberg et al, 1983]. Кроме

того данный метод хорошо работает в случае сильной взаимокорреляции дескрипторов [Новиков и Шпигун, 1993]. Метод также относительно прост в реализации, характеризуется высокой скоростью расчета, состав выделяемых главных компонент и характер их влияния на прогнозируемое свойство может дать дополнительную информацию об исследуемой зависимости строение – свойство. К недостаткам метода, как впрочем и РГК, относится отсутствие возможности выбора оптимального набора параметров и необходимость достаточно линейного характера зависимости между свойствами и дескрипторами для получения удовлетворительного качества прогноза [Geladi and Kowalski, 1986]. С учетом указанной специфики, в дальнейшей работе для получения прогностических моделей использовали оба метода.

Для расчета использовали полный набор дескрипторов, включающий в себя комбинации исходных дескрипторов по первой и второй степени. Оптимальное число компонент определяли методом кросс-валидации.

Таблица 8.2.

Характеристики прогностических моделей (R^2 и Q^2), рассчитанных методами РГК и ДМНК

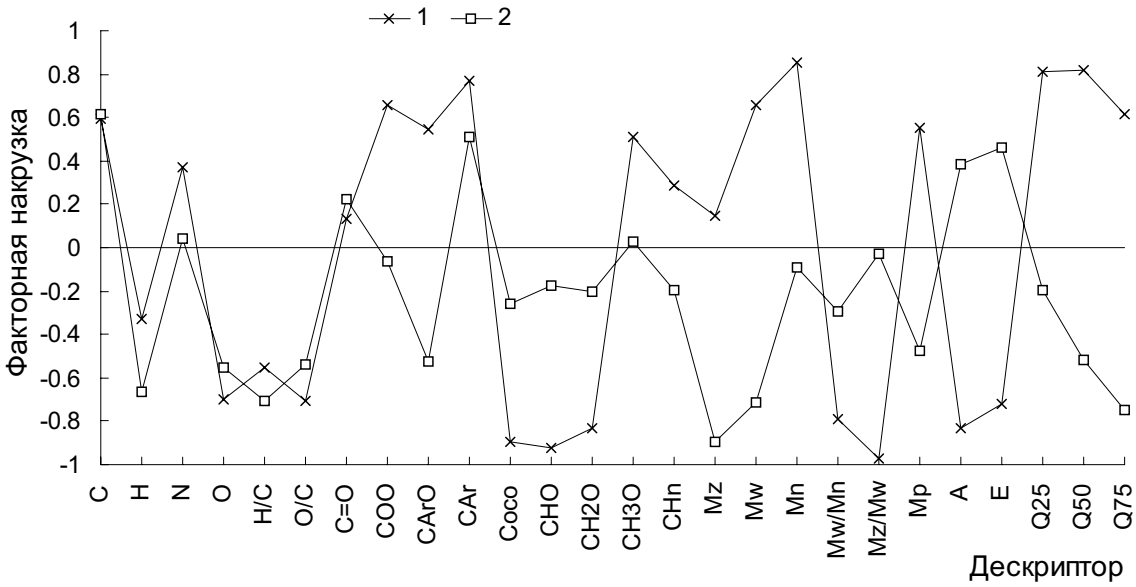
Константы связывания					
	Hg	Py	Flt	An	Атразин
	$lgK_{РСЦ}$	$K_{ОС}$			
ДМНК					
R^2	0.69	0.9	0.9	0.41	0.78
Q^2	0.27	0.52	0.55	0.09	0.47
РГК					
R^2	0.62	0.9	0.91	0.4	0.87
Q^2	0.47	0.67	0.78	0.24	0.73

Константы детоксикации						
	Hg		Py	Flt	An	Атразин
	$lgK_{РСЦ}^{D}$	$lgK_{РСЦ}^{B}$	$K_{ОС}^{D}$			
ДМНК						
R^2	0.37	0.36	0.78	0.48	0.79	0.72
Q^2	–0.49	–0.79	0.57	–0.02	0.63	0.31
РГК						
R^2	0.12	0.15	0.77	0.60	0.79	0.73
Q^2	0.10	0.02	0.63	0.28	0.73	0.45

Как видно, прогностическая способность полученных моделей невысока – значительно хуже, чем моделей МР. Для объяснения этого факта проводили анализ главных компонент, которые выделяются с помощью указанных методов из исходного пространства дескрипторов. Число компонент, используемых для прогноза по методу РГК, составляло от 2 до 9. Для метода ДМНК оно, как правило, не превышало 2.

Первые пять компонент, на которые метод РСЦ разлагает матрицу исходных дескрипторов для выборки из 19 препаратов, использованных для изучения их взаимодействия с ПАУ, приведены на рис. 8.6а (первые две компоненты) и рис. 8.6б (третья, четвертая и пятая компоненты).

а)



б)

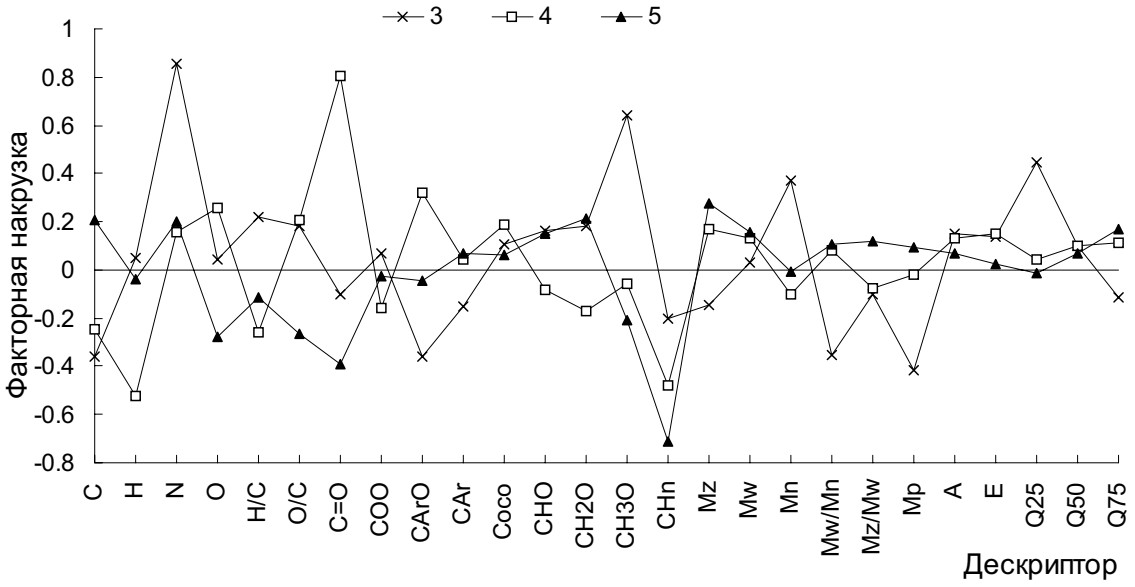


Рис. 8.6. Первая и вторая (а) и третья, четвертая, пятая (б) РГК-компоненты для выборки из 19 препаратов, использованных в экспериментах по связыванию ПАУ.

Как видно из рис. 8.6а, первая компонента преимущественно описывает совместный вклад ароматических и углеводных фрагментов в состав макромолекул гумусовых кислот, а также их полидисперсность (наиболее высокие факторные нагрузки наблюдаются для C_{Ar}, CHO, CH_2O, M_w/M_n, M_z/M_w). Вторая компонента определяется, в основном, ММ характеристиками и частично – параметрами состава ароматической части гумусовых кислот ($C_{Ar}O$ и C_{Ar}). Максимальный вклад в третью компоненту вносит содержание N и O/C, четвертая компонента определяется содержанием карбонилов, пятая – алифатического незамещенного углерода.

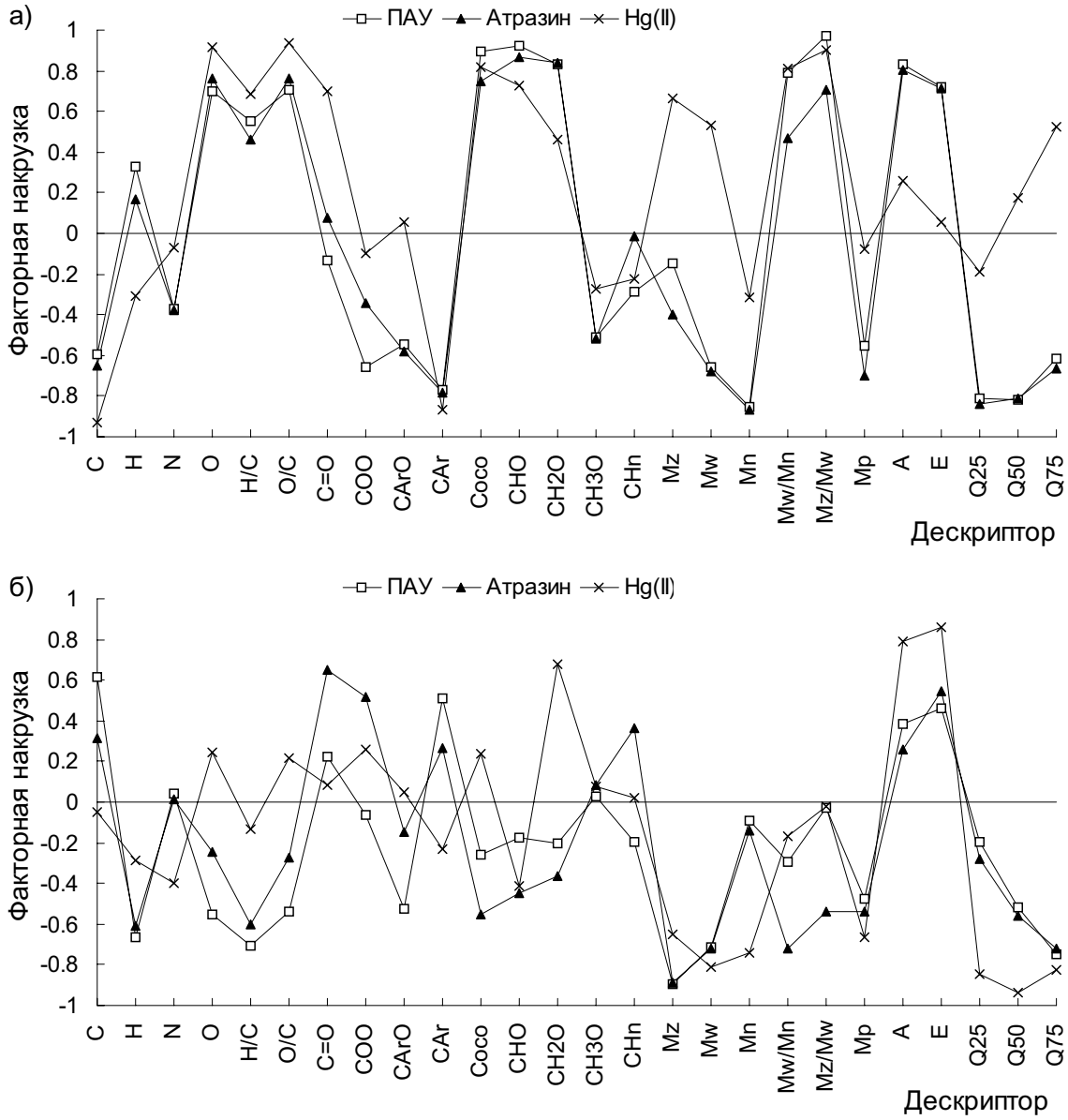


Рис. 8.7. Первые (а) и вторые (а) РГК-компоненты, полученные для различных выборок препаратов, использованных в экспериментах по связыванию и детоксикации Hg(II) (n = 16), ПАУ (n = 19) и атразина (n = 16) гумусовыми кислотами.

Рассмотрение первых двух РГК-компонент для выборок препаратов, использованных для изучения взаимодействия с Hg(II) и атразином (пересечение всех выборок только по 4 препаратам), показало их существенное сходство с описанными выше для выборки по ПАУ (рис. 8.7). Это позволяет сделать важный вывод об адекватности использованного в нашей работе принципа формирования выборок для установления зависимостей “строение – свойство”, который предусматривал задание максимального разнообразия строения и свойств гумусовых кислот по выборке путем включения в нее препаратов различного происхождения и фракционного состава.

В связи со спецификой выбора главных компонент по методу ДМНК, они должны нести в себе информацию не только о внутренней структуре матрицы дескрипторов, но и о ее взаимосвязи с матрицей прогнозируемых свойств. Поэтому проводили сопоставление первых компонент, полученных для одной и той же выборки препаратов, но в одном случае для констант связывания, а в другом – детоксикации. В случае ПАУ выделяемые компоненты оказались практически идентичны (рис. 8.8), что согласуется с наличием тесной корреляции между K_{OC} и K_{OC}^{D} для ПАУ.

В тоже время в случае атразина, для которого корреляция между K_{OC} и K_{OC}^{D} отсутствует, наблюдалось принципиальное различие между первыми ДМНК компонентами (рис. 8.8б). Так, если для K_{OC} она была весьма сходна с таковой для ПАУ и, соответственно, характеризовалась максимальным вкладом дескрипторов структурно-группового состава, то для K_{OC}^{D} явно превалировали дескрипторы ММ состава. В случае Hg(II), для всех трех наборов данных ($K_{РСЦ}$, $K_{РСЦ}^{D}$, $K_{РСЦ}^{B}$) наблюдалось хорошее совпадение первых ДМНК компонент. При этом они существенно различались от таковых для ПАУ и атразина по факторным нагрузкам дескрипторов структурно-группового состава, но практически совпадали по вкладу ММ-дескрипторов.

Проведенный анализ ДМНК-компонент показывает его полезность для выявления характера взаимосвязи между прогнозируемыми свойствами и дескрипторами, обеспечивая дополнительный источник информации о физической сущности исследуемых процессов. Так, максимальный вклад дескрипторов – показателей гидрофильно-гидрофобного баланса (углеводы/ароматика), в первые компоненты для данных по связыванию ПАУ и атразина может трактоваться как подтверждение гипотезы о гидрофобном взаимодействии как основном механизме связывания органических экотоксикантов гумусовыми кислотами. В тоже время большой вклад ММ-дескрипторов в описание дисперсии K_{OC}^{D} для атразина указывает на превалирующую роль факторов, ответственных за проникновение гумусовых кислот в клетки, в процесс детоксикации атразина. Это хорошо согласуется с

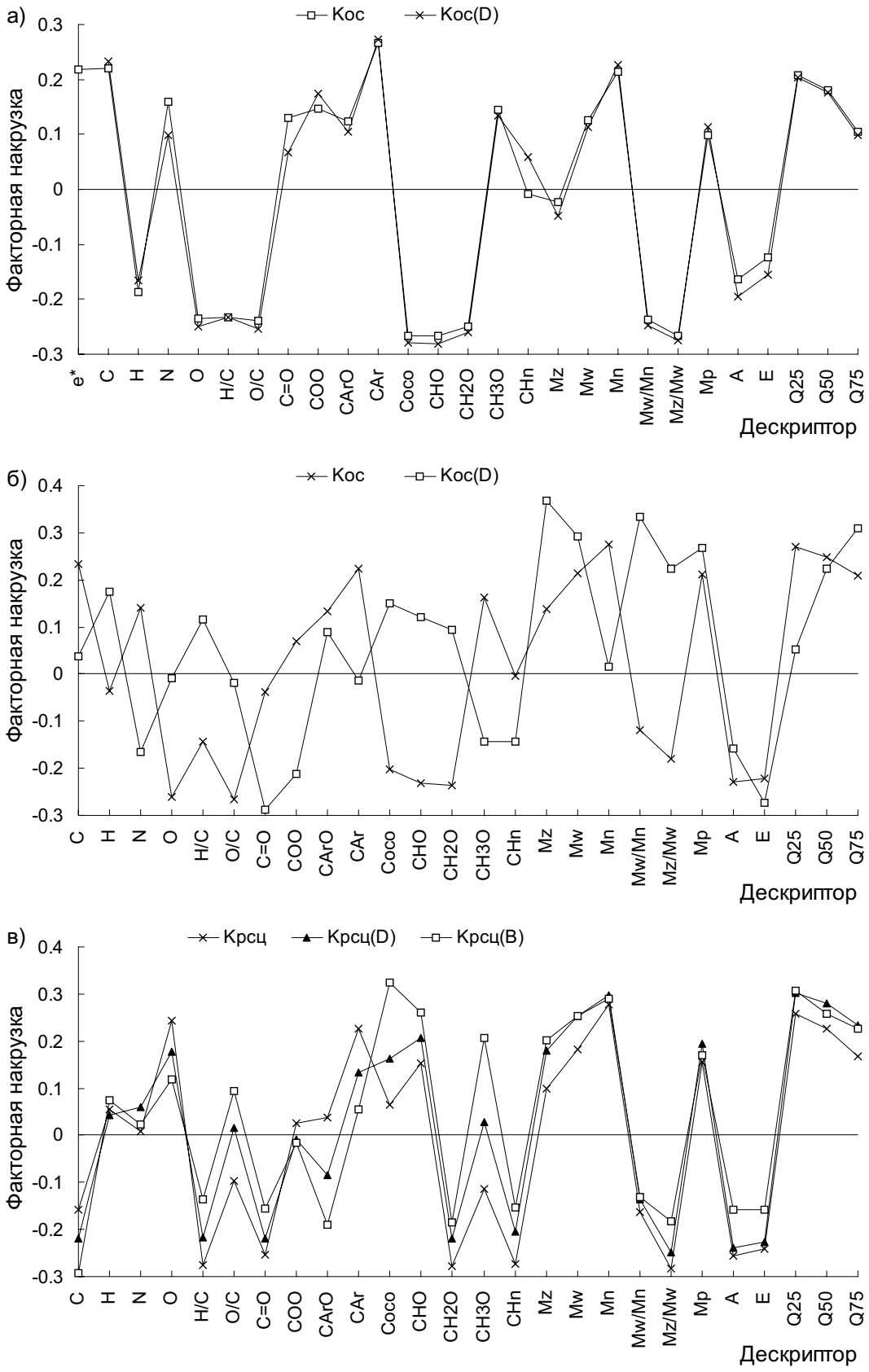


Рис. 8.8. Первые ДМНК компоненты для выборок препаратов, использованных для экспериментов по связыванию и детоксикации ПАУ (а), атразина (б) и Hg(II) гумусовыми кислотами.

установленным нами механизмом детоксикации атразина гумусовыми кислотами (Глава 7), который не является процессом истинной детоксикации, а заключается в повышении сопротивляемости организма к негативным факторам среды под воздействием гумусовых кислот. По-видимому, физиологическая активность гумусовых кислот в существенной степени определяется характером их ММР.

В результате анализа РГК- и ДМНК-компонент было установлено, что многие дескрипторы из использованного набора входят во все компоненты с очень малыми факторными нагрузками. Это может свидетельствовать об избыточности исходного набора независимых переменных, что приводит к информационному шуму при их использовании. Для наглядной демонстрации данного утверждения на рис. 8.9 графически представлены нормированные коэффициенты *b*, которые вошли в лучшие из полученных РГК-моделей для всех прогнозируемых свойств – K_{OC} и K_{OC}^{D} ПАУ (рис. 8.9а,б), K_{OC} и K_{OC}^{D} атразина (рис. 8.9в) и $K_{РСЦ}$, $K_{РСЦ}^{D}$ и $K_{РСЦ}^{B}$ комплексов Hg(II) с гумусовыми кислотами (рис. 8.9г).

Как следует из приведенных значений нормированных коэффициентов, прогностические модели РГК, полученные с использованием заданного набора дескрипторов состава, характеризуются очень высоким уровнем шума, который выражается в придании низких значений коэффициентов всем дескрипторам. Особенно наглядно этот эффект заметен для моделей, описывающих взаимосвязь дескрипторов состава и $K_{РСЦ}$. Для данных констант не было обнаружено значимой корреляционной зависимости ни с одним из интегральных дескрипторов состава. По-видимому, не существует и такой их линейной оптимальной комбинации, которая бы удовлетворительно описывала данное свойство. В итоге, как следует из рис. 8.9г, значения всех нормированных коэффициентов в РГК-модели для $K_{РСЦ}$ не превышают 0.08 и несущественно отличаются друг от друга.

В то же время для K_{OC} ПАУ, которые обнаружили тесную корреляционную взаимосвязь с дескрипторами состава (содержание ароматических фрагментов), характерен совсем иной вид распределения коэффициентов по их значимости в прогностической модели – отчетливо выявляется высокий вклад в прогнозируемое свойство таких дескрипторов, как $C_{Ar,}$ H, C=O, CH_n, M_w/M_n (K_{OC} Py и Flt). При этом для An, обладающего минимальным сродством к гумусовым кислотам из трех исследованных ПАУ, указанный характер распределения коэффициентов, как и в случае Hg(II), вырождается в их равномерное распределение по всем дескрипторам, что свидетельствует о низкой прогностической способности такой модели.

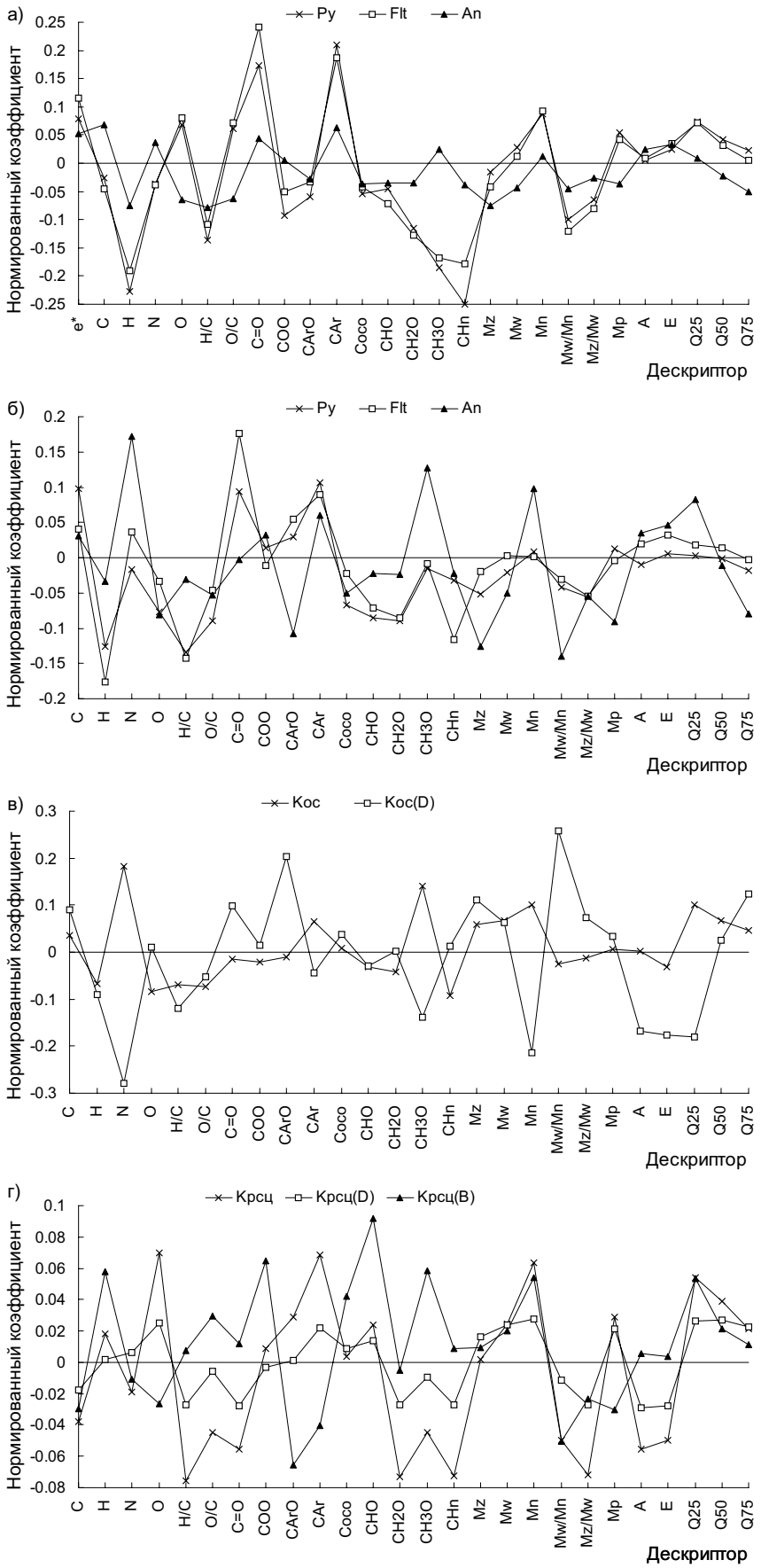


Рис. 8.9. Нормированные коэффициенты, полученные при расчете РГК моделей для K_{oc} ПАУ (а), K_{oc}^{D} ПАУ (б), атразина (в), Hg(II) (г).

Для преодоления указанной проблемы информационного шума была введена процедура пошагового исключения дескрипторов. Удалялись те из них, исключение которых приводило к большему Q^2 рассчитываемой модели. Лучшими считались модели с наибольшим Q^2. Их характеристики и входящие в них дескрипторы приведены в табл. 8.3.

Таблица 8.3.

Характеристики описательной и прогностической способности моделей (R^2 и Q^2), рассчитанных с пошаговым исключением дескрипторов методами ДМНК и РГК

Токси-кант	св-во	ДМНК				РГК			
		Q^2	R^2	Число дескр	Число комп.	Q^2	R^2	Число дескр.	Число комп.
Hg (II)	$lgK_{РСЦ}$	0.98	1.00	14	11	0.80	0.89	5	4
	$lgK_{РСЦ}^{D}$	0.38	0.52	8	1	0.39	0.44	4	2
	$lgK_{РСЦ}^{B}$	0.24	0.60	7	2	0.30	0.41	7	4
Py	K_{OC}	0.86	0.94	9	4	0.86	0.93	8	6
	K_{OC}^{D}	0.78	0.85	4	2	0.80	0.82	12	3
Flt	K_{OC}	0.85	0.92	8	3	0.86	0.92	7	6
	K_{OC}^{D}	0.50	0.61	2	1	0.53	0.61	11	3
An	K_{OC}	0.58	0.66	2	2	0.58	0.65	8	3
	K_{OC}^{D}	0.84	0.89	12	2	0.82	0.85	14	4
Атразин	K_{OC}	0.57	0.73	9	2	0.60	0.71	15	5
	K_{OC}^{D}	0.83	0.90	4	1	0.86	0.92	5	4

Полученные данные показывают эффективность процедуры снижения информационного шума за счет исключения из исходной матрицы дескрипторов, имеющих наименьшее влияние на прогнозируемое свойство. Так, качество рассчитываемых РГК- и ДМНК-моделей существенно улучшилось, хотя в среднем оно оставалось хуже, чем для МР-моделей (табл. 8.1)

С целью выявления значимости дескрипторов для прогнозируемых связывающих и детоксицирующих свойств гумусовых кислот, был проведен анализ частоты встречаемости дескрипторов в лучших моделях (по каждому свойству), полученных методами РГК и ДМНК (характеристики приведены в табл. 8.3). Полученные результаты приведены на рис. 8.10.

Анализ частоты встречаемости дескрипторов в наилучших моделях позволил выявить наибольшую значимость структурно-групповых дескрипторов, описывающих распределение углерода в углеводной части гумусовых кислот, полидисперсности, содержания H и N. Следует отметить,

что полученный ряд дескрипторов, имеющих высокую значимость для прогнозирования связывающих и детоксицирующих свойств гумусовых кислот, оказался весьма близок найденному ранее набору дескрипторов, обладающих высокой дискриминирующей способностью с точки зрения происхождения и фракционного состава гумусовых кислот.

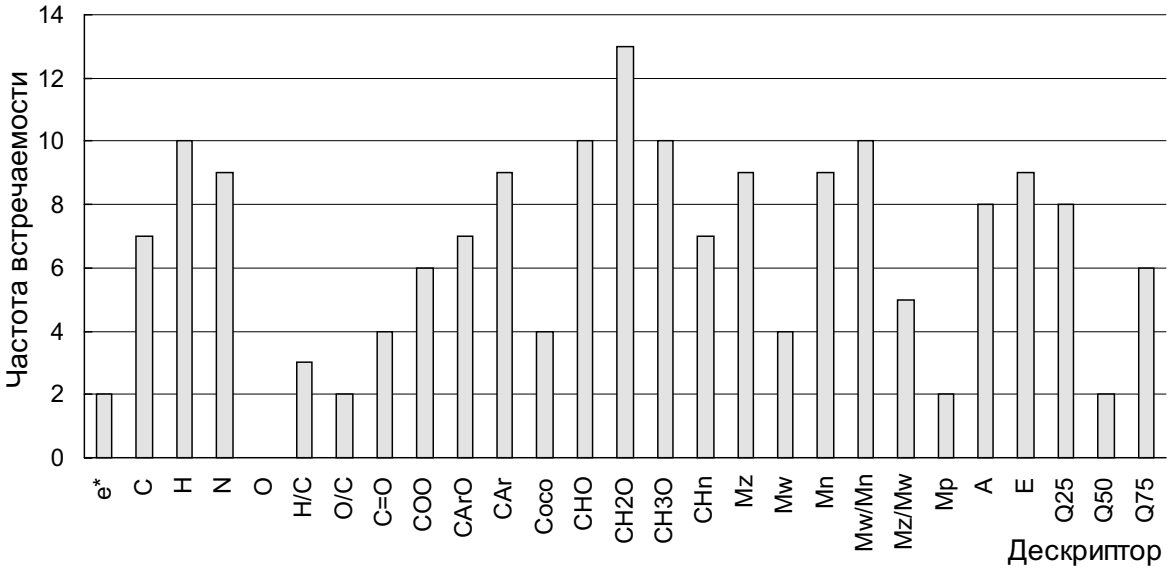


Рис. 8.10. Частота встречаемости дескрипторов в наилучших РГК- и ДМНК-моделях, прогнозирующих связывающие и детоксицирующие свойства гумусовых кислот по отношению к ПАУ, атразину и Hg(II) (характеристики всех 22 моделей приведены в табл. 8.3).

***

Таким образом, проведенный комплекс систематических исследований позволил разработать методические подходы к построению прогностических моделей “строение – свойство” с использованием набора интегральных дескрипторов состава для численного описания строения гумусовых кислот. Так, установлен минимальный размер выборки препаратов для определения признаков свойств, показаны преимущества реализованного в работе алгоритма метода МР по сравнению с методами многокомпонентного анализа, предложен способ повышения прогностической способности моделей, рассчитываемых методами РГК и ДМНК, путем оптимизации исходного набора дескрипторов.

На основании указанных методических подходов разработаны МР-, РГК- и ДМНК-модели, которые могут быть использованы для прогноза связывающих и детоксицирующих свойств гумусовых кислот различного происхождения и фракционного состава по отношению к ПАУ, атразину и Hg(II). Оценка прогностической способности полученных моделей

свидетельствует о возможности построения моделей удовлетворительного качества с использованием интегральных дескрипторов состава. Тем самым подтверждена состоятельность предложенного в работе подхода к численному описанию строения гумусовых кислот в терминах состава. Успешное решение задач классификации и прогноза свойств гумусовых кислот с использованием комплекса интегральных дескрипторов состава, отвечающих разным уровням структурной организации органических объектов, позволяет предположить справедливость данного подхода к численному описанию строения и других объектов стохастического характера.

Опыт применения рассчитанных моделей на практике позволит более полно выявить достоинства и недостатки заложенных в них алгоритмов и определить оптимальные сферы их использования.