Глава 6. КЛАССИФИКАЦИЯ ГУМУСОВЫХ КИСЛОТ ПО ПРОИСХОЖДЕНИЮ И ФРАКЦИОННОМУ СОСТАВУ

Основные проблемы классификации гумусовых кислот обусловлены стохастическим характером данного органического объекта, а именно, переменным составом, нерегулярностью структуры и полидисперсностью. Указанные факторы создают большие трудности для поиска классификационных признаков гумусовых кислот. Здесь необходимо оговориться, что в связи с еще не устоявшейся русскоязычной терминологией классификационного анализа, зачастую под классификационным признаком понимают как количественную характеристику (в нашем случае – дескриптор состава или свойств), с использованием которой осуществляют в дальнейшем отнесение препарата к тому или иному классу, так и ту категорию, по которой проводят априорное разделение на классы (в нашем случае – “источник происхождения”, “фракционный состав”). Поэтому при использовании понятия “классификационный признак” в последнем смысле добавляли определение “априорный” или указывали конкретные категории, по которым проводили классификацию (“источник происхождения”,. “фракционный состав” или более детализированный признак – “источник происхождения и фракционный состав”). Согласно принятому в нашей работе подходу к численному описанию строения гумусовых кислот в качестве количественных характеристик для классификации использовали наборы интегральных дескрипторов состава, соответствующие трем иерархическим уровням структурной организации органических объектов – элементному, фрагментному и молекулярному. Применение комплекса описанных методов для анализа гумусовых кислот позволило определить следующие интегральные дескрипторы состава: содержание элементов (% масс) и их атомные соотношения, содержание С и Н в составе структурных групп в процентах от общего С и Н, соответственно, и их отношения, средние ММ (их отношения. Данный набор интегральных дескрипторов состава дополняли за счет интегрального дескриптора свойств – массового коэффициента поглощения ε*.

6.1 Характеристика полученного массива данных

Спецификой полученного массива данных являлось то, что полный набор интегральных дескрипторов был получен лишь для малого количества препаратов (10), что не позволяло поставить задачу классификации по химическому строению. В тоже время размеры выборок с известным набором интегральных дескрипторов одного уровня – элементного, фрагментного или молекулярного, полученные с использованием одного метода, – были гораздо

больше и составляли 30 (эксклюзионная хроматография – ЭХ), 40 (ПМР), 60 (^{13}С ЯМР) и 80 (элементный анализ – ЭЛАН) препаратов. В набор ЭХ входили: M_w, M_n, M_z, M_w/M_n, M_z/M_w, M_p; в ПМР-набор – содержание Н в составе COOH, ArOH, ArH, AlkOH, α-CH, CH_n, а также содержание Н в составе ароматических (H_{Ar}) и углеводных (H_{Carb}) фрагментов в процентах от общего содержания скелетных протонов; в ^{13}С ЯМР-набор – содержание С в составе C=O, COO, C_{Ar}, $C_{Ar}O$, CHO, CH_2O, CH_3O, CH_n, а также суммарное содержание углерода в составе ароматических (ΣC_{Ar}) и углеводных фрагментов (ΣC_{Carb}).

Первичная классификация препаратов указанных выборок с использованием интегральных дескрипторов одного уровня (элементного и структурно-группового) (рис. 6.1), отчетливо показала тенденцию к образованию кластеров в соответствии с источником происхождения и фракционным составом гумусовых кислот. Так, рассмотрение диаграммы Ван Кревелена (рис. 6.1а) показывает, что все исследованные препараты гумусовых кислот расположились в коридоре значений Н/С от 0.4 до 1.2, а О/С – от 0.3 до 1.0. По величине Н/С препараты различного происхождения, но сходного фракционного состава образуют следующий ряд: уголь ≤ чернозем < торф < дерново-подзолистые и серые лесные почвы < воды. Для О/С этот же ряд выглядит почти аналогично: уголь ≤ чернозем < дерново-подзолистые и серые лесные почвы < торф < воды. Это говорит о максимальной ненасыщенности препаратов угля и наибольшей окисленности гумусовых кислот вод. Изменение Н/С и О/С по фракционному составу происходит сходным образом для торфа и почв: при довольно близкой величине Н/С для ГК характерны значительно меньшие значения О/С по сравнению с ФК. Это указывает на обогащенность молекулярной структуры ФК кислородсодержащими группами. Нефракционированные препараты ГФК (как видно на примере обширной выборки препаратов торфа) занимают промежуточное положение между ГК и ФК.

Рассмотрение двумерной диаграммы распределения исследованных препаратов гумусовых кислот по содержанию С в составе ароматических Σ(Ar, ArO) и углеводных Σ(OCO, CHO, CH2O) фрагментов показывает, что они расположились в коридоре значений ΣC_{Ar} от 0.25 до 0.72 и ΣC_{Carb} – от 0.02 до 0.37. Наиболее уникальный структурно-групповой состав характерен для ГК угля – максимальное содержание ароматического углерода (до 70 %) при практически полном отсутствии углеводных фрагментов. Это позволяет предположить наличие поликонденсированных структур в составе данных соединений и их высокую гидрофобность. Высокое содержание

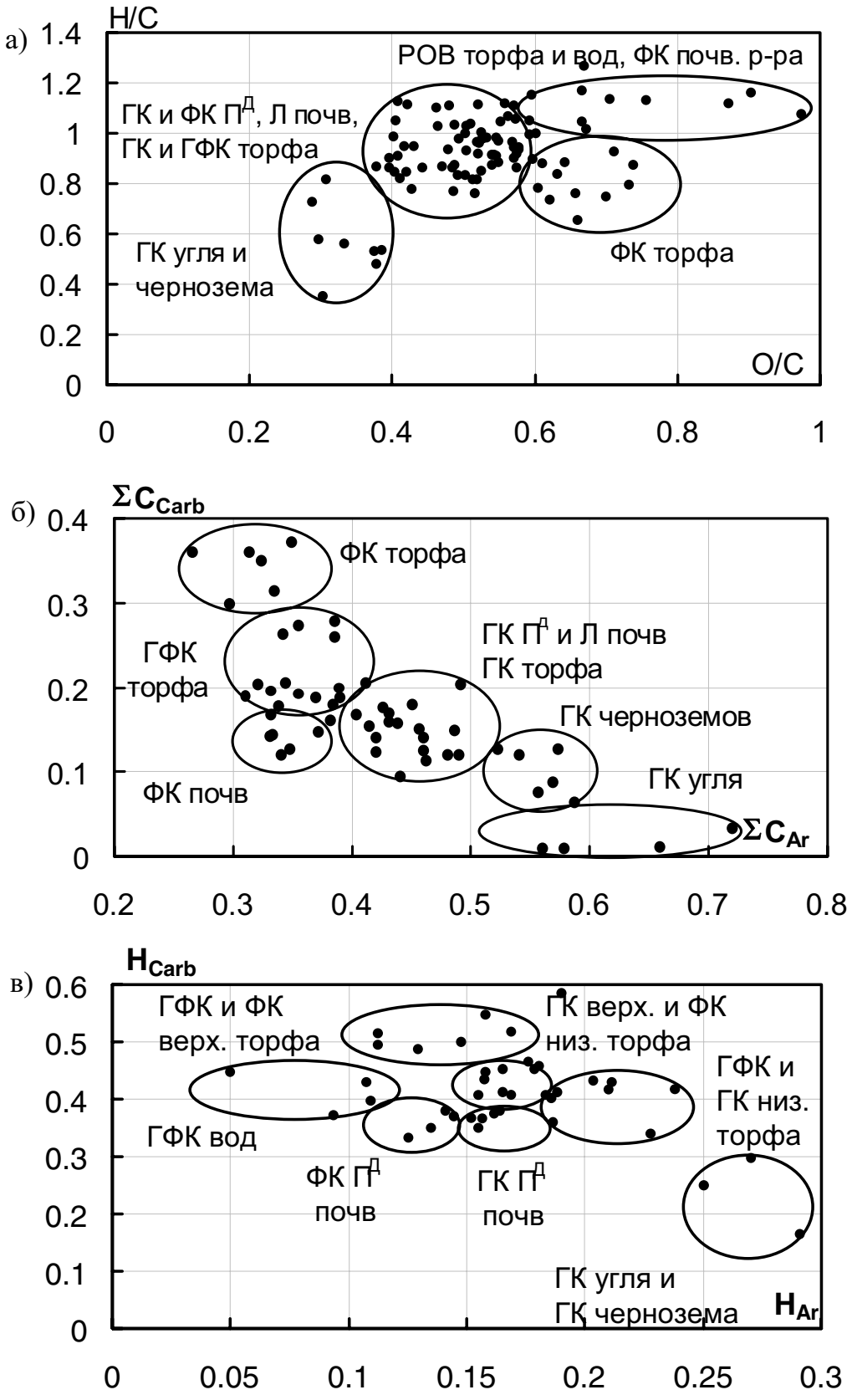


Рис. 6.1. Распределение препаратов по атомным соотношениям (а) и суммарному содержанию С (б) и Н (в) в составе ароматических и углеводных фрагментов углеродного скелета гумусовых кислот.

ароматического углерода (иногда сопоставимое с ГК угля) характерно для ГК черноземов. Однако, в отличие от угля, в них присутствуют углеводные фрагменты, в состав которых входит от 8 до 13% углерода. ГК дерново-подзолистых ($П^{д}$) и серых лесных (Л) почв существенно отличаются от черноземов. Содержание ароматического углерода в них не превышает 45%, тогда как содержание углерода углеводных фрагментов достигает 15-20%. Почвенным ГК весьма близки по распределению углерода ГК и ГФК торфа. ФК почв весьма отличны от ГК и характеризуются самым высоким содержанием углерода карбоксильных (сложноэфирных) групп – до 22%, содержание ароматического углерода не превышает 40%, а в состав углеводных фрагментов входит 20-25% углерода. ФК торфа содержат еще больше углеводных фрагментов, чем ФК почв: для верховых торфов величина C_{AlkO} достигает максимума – 40%. В тоже время для низинных торфов этот показатель составляет 24-25%, будучи на уровне почвенных ФК. Тем самым ФК верхового торфа образуют второй экстремум – по содержанию углеводного углерода, и должны характеризоваться высокой гидрофильностью. Следовательно, можно сделать вывод о том, что ГК угля и ФК торфа представляют собой два граничных класса гумусовых кислот с максимальным содержанием ароматических и углеводных фрагментов, соответственно. Все остальные классы гумусовых кислот занимают по этим показателям промежуточное положение. Для ГФК вод наблюдается самое высокое содержание алкильного углерода (до 25%), довольно высокие значения этой величины (до 22%) характерны и для двух других низкомолекулярных, гидрофильных классов гумусовых кислот – для ФК почв и торфа.

Расположение препаратов на диаграмме в координатах содержание Н в составе “ароматических (H_{Ar})” и “углеводных (H_{Carb})” фрагментов углеродного скелета (рис. 6.1в) позволило определить коридор значений этих параметров для исследованных препаратов гумусовых кислот как 0.05-0.3 для H_{Ar} и 0.15-0.60 для H_{Carb}. Характер распределения водорода в углеродном скелете гумусовых кислот верховых и низинных торфов свидетельствует о существенном различии в их строении. В скелете гумусовых кислот верховых торфов до 60 % водорода находится в составе неразложившихся олиго- или полисахаридных цепочек, тогда как для низинных торфов характерен высокий вклад водорода ароматических фрагментов. Это указывает на большую степень трансформации углеводно-пептидного комплекса гумусовых кислот низинных торфов и, следовательно, на их более глубокую гумификацию. Помимо различия между торфами, можно отметить весьма специфическую структуру ГК угля и черноземов, углеводный комплекс которых претерпел наибольшую деградацию – в углях он практически полностью отсутствует. В

тоже время для ГФК вод характерна наибольшая степень замещения ароматических структур. Наряду с максимальной окисленностью, это указывает на наибольшую степень деградации ароматического каркаса в водных гумусовых кислотах.

Описанные особенности строения гумусовых кислот различного происхождения и фракционного состава позволили сформулировать задачу классификационного анализа как установление принадлежности препарата к классам “источник происхождения” (вода, почва, торф, уголь), “фракционный состав” (ГК, ФК, ГФК) или “источник происхождения и фракционный состав”. Наборы интегральных дескрипторов состава для проведения соответствующей классификации создавали с учетом критерия специфичности, требующего использования комплекса дескрипторов разного уровня для численного описания строения гумусовых кислот. При этом из-за обсужденных выше ограничений, данную проблему решали, дополняя наборы дескрипторов фрагментного или молекулярного уровней дескрипторами элементного состава, которые были определены для препаратов всех выборок. Кроме того, набор дескрипторов молекулярно-массового состава дополняли значениями ε*, найденными для той же выборки препаратов. Наборы интегральных дескрипторов разных уровней называли смешанными и обозначали их как ЭЛАН+ЭХ, ЭЛАН+ПМР, ЭЛАН+^{13}С ЯМР, ε*+ЭХ и ЭЛАН+ЭХ+ε*.

До проведения классификации с использованием указанных дескрипторов, необходимо было ответить на вопрос о принципиальной воспроизводимости свойств препаратов гумусовых кислот при их повторном выделении из сходного источника по стандартной методике. Данная проблема имеет особое значение, так как гумусовые кислоты, присутствующие в природном объекте, представляют собой открытую динамическую систему, обменивающуюся веществом и энергией с окружающей средой. Помимо естественной изменчивости условий гумусообразования, важнейшим фактором, влияющим на свойства наблюдаемых гумусовых кислот, является процедура их выделения из природного объекта. Поэтому если в стандартных условиях выделения из сходных источников получаются препараты гумусовых кислот с воспроизводимыми свойствами, то можно говорить о типологизируемости систем гумусовых кислот [Степин, 198&&], присутствующих в различных природных объектах. Это, в свою очередь, означает принципиальную возможность получения воспроизводимых результатов для данных объектов исследования.

Для ответа на поставленный вопрос было проведено сопоставление атомных соотношений, полученных для препаратов, выделенных из сходных источников по стандартным методикам в разные годы (рис. 6.2). Как видно,

для большинства препаратов наблюдается довольно хорошая воспроизводимость данных. Максимальный разброс характерен для H, на результаты определения которого большое влияние оказывает атмосферная влажность, тогда как для C и для N результаты разных лет хорошо согласуются между собой. На рис. 6.3 приведено распределение атомов углерода с различным химическим окружением (данные ^{13}C ЯМР) для препаратов, выделенных из сходных источников в разные годы. Данные характеризуются хорошей воспроизводимостью, что свидетельствует о воспроизводимости дескрипторов структурно-группового состава.

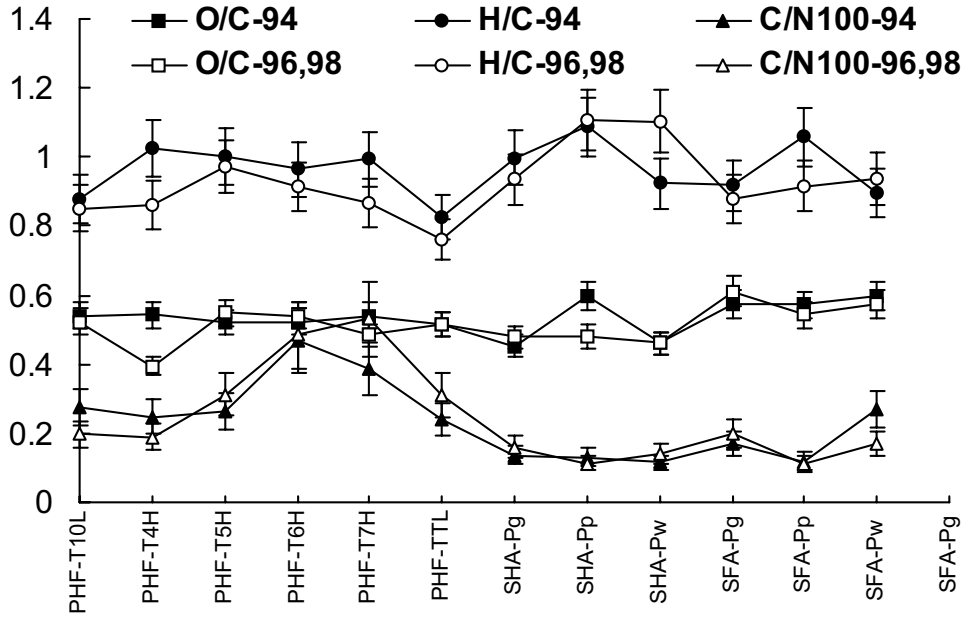


Рис. 6.2. Атомные соотношения препаратов гумусовых кислот, выделенных из сходных источников в разные годы (n = 3, P = 0.95).

Полученные результаты могут служить убедительным положительным ответом как на вопрос о типологизируемости систем гумусовых кислот в природных объектах, так и о воспроизводимости интегральных дескрипторов состава препаратов гумусовых кислот при их повторном выделении из сходного объекта по стандартной методике. Это позволило перейти к решению поставленной задачи классификации с точки зрения источника происхождения и фракционного состава гумусовых кислот с использованием интегральных дескрипторов состава.

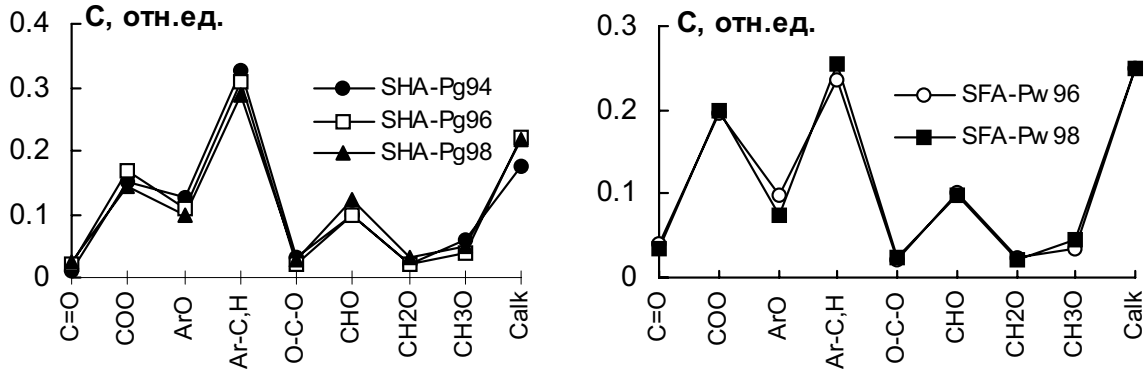


Рис. 6.3. Распределение углерода в структуре гумусовых кислот, выделенных из сходных источников в разные годы (по данным ^{13}C ЯМР).

6.2 Классификация с использованием интегральных дескрипторов состава

Ввиду отсутствия сведений о характере связи между принадлежностью препарата к определенному классу и значениями его интегральных дескрипторов состава и свойств были использованы различные методы многомерной классификации, а именно, ***линейный дискриминантный анализ (ЛДА), К ближайших соседей (КБС)*** и ***нейронных сетей***. Достоинство ЛДА – возможность вероятностных оценок полученной классификации, недостаток – неприменимость в случае сильного отличия разграничительных поверхностей между классами от линейности. Метод КБС, напротив, хорошо работает при наличии границ между классами сложной формы, он не требует априорных предположений о распределении элементов внутри класса. Недостаток – отсутствие вероятностных оценок классификационного правила. Расчеты по методу КБС проводили с использованием оригинальной программы Regr (автор А. В. Кудрявцев), алгоритм которой предусматривает кросс-валидацию и перебор дескрипторов. Метод нейронных сетей обладает наиболее мощным универсальным алгоритмом и позволяет оценить значимость вклада каждого признака. Недостаток – сложность реализации и ограниченная применимость для работы с малыми массивами данных – легкость “переобучения”.

Для оценки дискриминирующей способности дескрипторов одного уровня проводили классификацию с точки зрения источника происхождения и фракционного состава с использованием наборов ЭЛАН, ЭХ, ПМР и ^{13}C ЯМР методами ЛДА и КБС. Полученные результаты приведены в табл. 6.1 и 6.2, соответственно.

Таблица 6.1

Классификация препаратов гумусовых кислот методом ЛДА с использованием дескрипторов одного уровня

Набор данных	Источник					Фракционный состав			
	Правильно классифицировано, %								
	Торф	Почва	Уголь	Вода	Общее	ГФК	ГК	ФК	Общее
ЭЛАН	94	84	67	56	83	77	77	55	72
ЭХ	100	100	–	80	**96**	85	57	85	77
^{13}С ЯМР	79	96	60	–	84	60	82	70	72
ПМР	96	91	–	–	94	85	79	100	**87**

Таблица 6.2

Классификация препаратов гумусовых кислот методом КБС с использованием дескрипторов одного уровня

Набор данных	Источник					Фракционный состав			
	Правильно классифицировано, %								
	Торф	Почва	Уголь	Вода	Общее	ГФК	ГК	ФК	Общее
ЭЛАН	97	84	50	22	80	74	73	55	69
ЭХ	100	93	–	80	**93**	85	100	86	**89**
^{13}С ЯМР	82	88	40	–	81	75	82	40	72
ПМР	93	82	–	–	90	69	79	58	69

Как видно из полученных результатов, проведение ЛДА с применением указанных наборов дескрипторов состава одного уровня дает довольно высокий уровень правильных классификаций, хотя ни в одном случае 100 % правильной классификации достичь не удавалось. Наибольшее количество правильных отнесений обеспечивалось с использованием дескрипторов ММ состава (ЭХ-набор).

Классификации с использованием ***смешанных наборов дескрипторов*** показали их гораздо более высокую дискриминирующую способность как для метода ЛДА (табл. 6.3), так и КБС (табл. 6.4). Применение наборов ЭЛАН+ЭХ и ЭЛАН+ЭХ+ε* позволило достичь 100 % правильных классификаций по всем трем априорным признакам – по источнику происхождения, по фракционному составу и по детализированному признаку – источник происхождения и фракционный состав – для обоих методов. Это говорит о

высокой дискриминирующей способности указанных наборов дескрипторов. На рис. 6.4 показан пример такой классификации. Следует отметить, что и остальные смешанные наборы – ЭЛАН+^{13}С ЯМР, ЭЛАН+ПМР и ЭХ+ε* обеспечивали довольно высокий процент правильных классификаций по всем трем категориям.

Таблица 6.3

Характеристики классификации препаратов гумусовых кислот методом ЛДА (Q – общее количество правильных классификаций, %)

Наборы данных	Источник		Фракционный состав		Источник и фракционный состав	
	Q	Дескрипторы	Q	Дескрипторы	Q	Дескрипторы
ЭЛАН+ ^{13}С ЯМР	94	N, C, ΣC_{Carb}, C/N, CH_3O, $C_{Ar}O$, CH_n	88	O, CHO, COO C/N, N, C, H/C, CH_2O, ΣC_{Ar}	94	N, ΣC_{Carb}, C/N, C, O, H/C, CH_2O, C_{Ar}, CH_3O
ЭЛАН+ ПМР	97	N, AlkH, AlkOH, COOH, ArOH, H/C, C/N	97	O/C, ArOH, COOH, ΣH_{Carb}, AlkOH, N, C/N	94	N, C/N, COOH, ArOH, ΣH_{Carb}, C, O/C, ArH, ΣH_{Ar}, AlkH
ЭХ+ε*	96	M_w/M_n, M_z/M_w, M_w, M_n, M_p, ε*	96	ε*, M_w/M_n, M_w, M_p, M_n	**100**	M_p, ε*, M_n, M_z/M_w, M_w/M_n, M_w
ЭЛАН+ ЭХ	**100**	N, M_p, M_w/M_n, M_z/M_w, O, C, M_w, M_n, C/N, H	**100**	N, M_z, C/N, M_n, O, O/C, C, M_p, H, H/C, M_w/M_n, M_w	**100**	N, M_p, O, O/C, C, M_n, M_w/M_n, M_w, M_z/M_w, H, H/C
ЭЛАН+ ЭХ+ε*	**100**	ε*, M_p, N, M_w, M_z, O, O/C, M_z/M_w, C, H, H/C	**100**	N, ***ε****, M_w/M_n, M_z, C/N, O, O/C, C, M_p, H, H/C	**100**	M_p, N, M_w, M_z, O, O/C, M_z/M_w, C, H, H/C

В табл. 6.3. и 6.4 приведены также дескрипторы, использование которых обеспечивает наилучшие классификации методами ЛДА и КБС, соответственно. Значимость дескрипторов и частота их встречаемости в наилучших дискриминирующих функциях метода ЛДА позволили сделать вывод о максимальной дискриминирующей способности N, M_p, ε*, M_w/M_n, CHO. Анализ частоты встречаемости дескрипторов в наилучших классификациях методом КБС показал весьма сходный ряд параметров: N, C, M_p, M_w/M_n, CHO и ΣC_{carb}. Следует отметить, что использование набора исходных дескрипторов, расширенного за счет их отношений и обратных

величин, для проведения классификаций методом КБС показало высокую дискриминирующую способность $1/\varepsilon^*$ и $1/M_p$.

Таблица 6.4

Характеристики классификации препаратов гумусовых кислот методом КБС (Q – общее количество правильных классификаций, %)

Наборы данных	Источник		Фракционный состав		Источник и фракционный состав	
	Q	Дескрипторы	Q	Дескрипторы	Q	Дескрипторы
ЭЛАН+ $+^{13}$С ЯМР	98	N/C, C, N, ΣC_{Carb}	83	C/H, C, CHO, CH_3O	71	H/C, N/C, CHO, ΣC_{Carb}
ЭЛАН+ +ПМР	94	N/C, 1/AlkH	73	1/C, N, $AlkOH/CH_nO$	72	O/C, N/C, $CH_n/AlkOH$
ЭЛАН+ЭХ	100	$1/M_p$, H, N, M_w/M_n	100	M_p, N, M_z/M_w	100	$1/M_p$, N, M_w/M_n
ЭХ+ε*	100	$1/\varepsilon^*$, $1/M_p$, $1/M_n$	100	$1/\varepsilon^*$, $1/M_p$, $1/M_n$	100	$1/\varepsilon^*$, M_p, $1/M_n$
ЭЛАН+ +ЭХ+ε*	100	$1/\varepsilon^*$, M_n, M_p	100	$1/\varepsilon^*$, M_n, M_p	100	$1/\varepsilon^*$, M_n, M_p

Метод нейронных сетей применяли только для наилучшего из смешанных наборов дескрипторов – ЭЛАН+ЭХ+ε*. Тестирование сетей различных топологий показало, что для данного набора дескрипторов достаточно 4 нейронов и 13 синапсов для полной классификации 27 препаратов по детализированному признаку – источник и фракционный состав. Это свидетельствует о высокой достоверности классификационного правила, получаемого с использованием указанного набора дескрипторов. Анализ значимости дескрипторов еще раз подтвердил максимальную дискриминирующую способность содержания N, M_w/M_n и M_p.

Представляло интерес сопоставить полученные результаты с набором пяти диагностических признаков гумусовых кислот, выработанных Д.С. Орловым, который подробно описывался в обзоре литературы [Орлов, 1974]. Данные признаки были выбраны автором с точки зрения их наибольшей характеристичности для гумусовых кислот как класса химических соединений с целью создания основ классификации гумусовых кислот по химическому строению. В нашей работе решали другую задачу классификации гумусовых кислот, однако установленный в результате набор дескрипторов, обладающих самой высокой дискриминирующей способностью по происхождению и фракционному составу гумусовых кислот, может также рассматриваться как набор наиболее характеристичных

признаков данных соединений. Поэтому ниже приводится сопоставление указанных наборов признаков гумусовых кислот.

Два из пяти диагностических признаков, предложенных Д. С. Орловым, описывают общее содержание и распределение N между гидролизуемой и негидролизуемой частью гумусовых кислот. Это хорошо согласуется с максимальной дискриминирующей способностью содержания N, выявленной тремя различными методами классификации, использованными в нашей работе. Содержание N, как правило, гораздо выше в почвенных гумусовых кислотах и ниже в торфяных и водных. При этом в ФК его меньше, чем в ГК. Первые два признака по Д.С. Орлову включают и содержание С, что также согласуется с полученными нами результатами. Третий признак характеризует величину массового коэффициента оптического поглощения при 465 нм, – нами показана высокая дискриминирующая способность аналогичного показателя – ε* при 254 нм. Величина ε* гораздо выше для почвенных препаратов, чем для водных, и возрастает при переходе от ФК к ГК. Четвертый и пятый признаки описывают выход бензолполикарбоновых кислот и наличие характеристических полос поглощения в ИК-спектре гумусовых кислот. Оба указанных признака призваны охарактеризовать специфику строения углеродного скелета гумусовых кислот. Проведенное нами исследование показало, что содержание С в составе углеводных фрагментов является более мощным дискриминирующим признаком, чем содержание С в составе ароматических фрагментов (выход бензолполикарбоновых кислот). Величина данного параметра тесным образом связана с глубиной разложения углеводного комплекса гумусовых кислот. Его минимальные значения наблюдаются для угля и чернозема, характеризующихся самой высокой степенью гумификации. Максимальные значения характерны для ФК верховых торфов, содержащих в своем составе (как было показано с помощью двумерной спектроскопии ЯМР, Глава 4) цепочки нетрансформированных олигосахаридов. Кроме того, данный параметр определяет гидрофильно-гидрофобный баланс молекул гумусовых кислот, их растворимость в воде. Принципиально новым результатом нашей работы является установление максимальной дискриминирующей способности дескрипторов ММ состава по источнику и происхождению гумусовых кислот. Это дает основания рассматривать средние ММ и полидисперсность как важнейшие диагностические признаки, количественной характеристике которых должно быть уделено особое внимание при создании классификации гумусовых кислот по химическому строению.

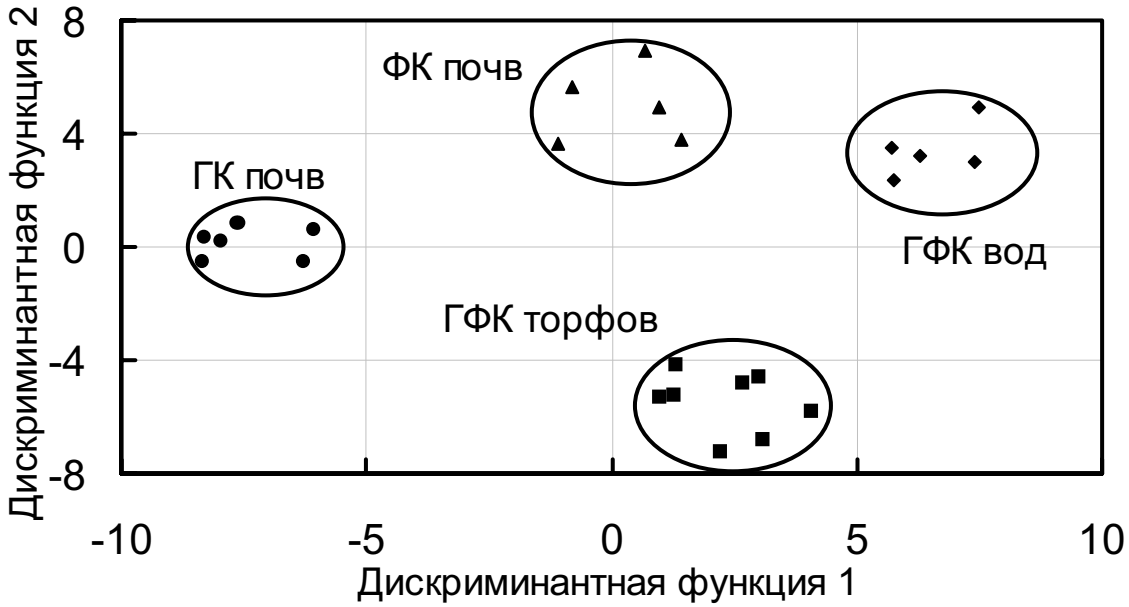


Рис. 6.4. Классификация гумусовых кислот по источнику происхождения и фракционному составу методом ЛДА с использованием набора дескрипторов “ЭЛАН+ЭХ”.

Таким образом, проведенный классификационный анализ показал, что смешанные наборы дескрипторов обладают гораздо более высокой дискриминирующей способностью, чем индивидуальные. Это подтверждает справедливость выбранного в нашей работе подхода к повышению специфичности численного описания структуры с помощью дескрипторов состава за счет использования комплекса дескрипторов, отвечающих разным уровням структурной организации.

В связи с тем, что интегральные дескрипторы ММ состава показали максимальную дискриминирующую способность по источнику происхождения и фракционному составу гумусовых кислот, то была поставлена задача классификации гумусовых кислот по указанным категориям с использованием набора интегральных дескрипторов ММ состава, расширенного за счет численных характеристик трех ЭХ-распределений: кривой элюирования в шкале K_d, ММР и распределения ε* по ММ.

6.3 Классификация с использованием расширенного набора дескрипторов молекулярно-массового состава

Расширенные наборы дескрипторов ММ состава включали в себя около 20 параметров, описывающих особенности, характерные как для кривых ММР, так и эксклюзионных хроматограмм в шкале K_d и распределения ε* по ММ (табл. 5.4). Для их расчета использовали программу Geltreat. В связи с весьма ограниченными размерами выборки препаратов – 27, прибегали к

искусственному приему ее увеличения, рассматривая каждую повторность как новый препарат того же класса. Допустимость использования такого приема обосновывали с помощью дисперсионного анализа, который показал, что дисперсия внутри повторностей одного препарата несущественно отличалась от общей дисперсии внутри повторностей класса (в среднем для всех классов в 1.5 раза).

Полученная описанным способом выборка содержала 116 препаратов: ГФК вод (31), ГФК торфа (35), ФК почв (32) и ГК почв (18), что позволило провести классификацию с контрольной выборкой, разделив исходную выборку случайным образом на обучающую и контрольную. Для классификации использовали методы, не накладывающие ограничений на формы классов в пространстве признаков – метод КБС и нейронных сетей.

6.3.1 Классификация по методу К-ближайших соседей

Каждый из исходных наборов ЭХ распределений – "K_d", "ММР" и "ε* по ММ" состоял из около двадцати дескрипторов, в которые входили средние ММ, полидисперсность, коэффициенты асимметрии и эксцесса, верхний и нижний квартили, интегралы по заданному диапазону и др. (полный список дескрипторов приведен в табл. 5.1). Для поиска максимально "полезных" дескрипторов в столь обширном массиве данных была предусмотрена процедура их отбора при реализации алгоритма КБС. Отбор проводили путем перебора всех возможных сочетаний дескрипторов с оценкой качества получаемых классификаций. Помимо исходных, в расчет вводили комбинированные дескрипторы: произведения или отношения исходных дескрипторов (не выше второй степени) и их обратные величины. Расчет классификационного правила проводили с участием двух комбинированных или трех-четырех исходных дескрипторов. На число и максимальный "порядок" дескрипторов накладывали ограничения как вычислительные мощности, так и сложность интерпретации классификационных правил с участием комбинированных дескрипторов. Для улучшения качества классификаций все дескрипторы нормировали на стандартное отклонение по всей обучающей выборке хроматограмм (образцов) и масштабировали с помощью дисперсионных весов w_1, w_2 и w_3, рассчитываемых по уравнениям, приведенным в Приложении 6.1. Масштабирование придает больший вес дескриптору, обладающему более высокой дискриминирующей способностью

Качество классификаций оценивали методом кросс-валидации. Для этого из обучающей выборки последовательно исключали образец и проводили его отнесение к тому или иному классу по оставшимся образцам. Если полученный класс совпадал с оригинальным, то классификация признавалась успешной. По окончании данной процедуры рассчитывали коэффициент качества классификации – $Q_{обуч}$, представляющий собой

отношение успешных классификаций к общему числу образцов. Затем отбирали 11 лучших классификаций, качество которых проверяли с использованием контрольной выборки. Для этой цели рассчитывали параметр качества $Q_{контр}$ – отношение правильно отнесенных хроматограмм с использованием данного классификационного правила к общему числу хроматограмм в контрольной выборке. Для оценки влияния масштабирования усредняли $Q_{обуч}$ и $Q_{контр}$ для 11 лучших классификаций, а для оценки классифицирующей способности различных наборов дескрипторов проводили усреднение рассчитываемых оценок качества – $Q_{обуч}$ и $Q_{контр}$ – по каждому из наборов.

Характеристики лучших классификационных правил, рассчитанных с использованием трех рассматриваемых наборов ММ дескрипторов, приведены в табл. 6.5. В ней даны оценки качества классификаций Q (усредненные по 11 лучшим) с участием трех-четырех исходных или двух комбинированных дескрипторов, которые были подвергнуты различным процедурам нормирования и масштабирования.

Таблица 6.5

Характеристики качества классификаций гумусовых кислот по происхождению и фракционному составу с использованием трех ЭХ-наборов дескрипторов ММ состава

Масштабирование	Набор дескрипторов					
	K_d		ММР		ε* по ММ	
	$Q_{контр}$	$Q_{обуч}$	$Q_{контр}$	$Q_{обуч}$	$Q_{контр}$	$Q_{обуч}$
2 комбинированных дескриптора						
без масштабирования	0.837	0.925	0.803	0.923	0.760	0.987
w_1	0.835	0.918	0.832	0.917	0.804	0.983
w_2	0.851	0.920	0.817	0.915	–	–
w_3	0.846	0.926	0.818	0.914	0.846	0.983
В среднем	*0.842*	*0.922*	*0.817*	*0.917*	*0.804*	*0.984*
3-4 исходных дескриптора						
без масшт., 3 дескр.	0.840	0.903	0.777	0.906	0.864	0.987
без масшт., 4 дескр	0.804	0.920	0.788	0.918	0.842	1.000
В среднем по всем	*0.836*	*0.919*	*0.806*	*0.916*	*0.823*	*0.988*

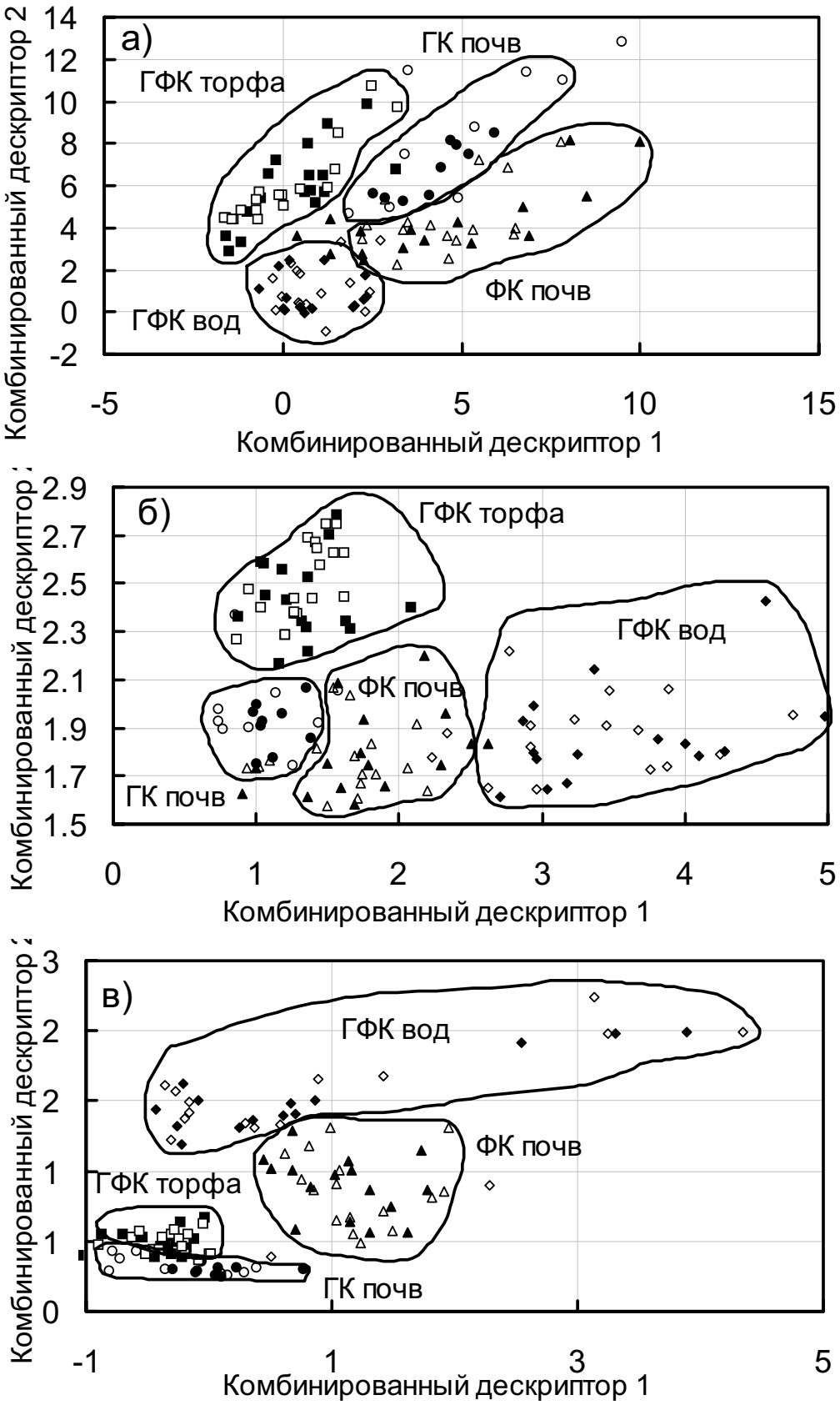


Рис. 6.5. Наилучшие классификации гумусовых кислот по происхождению и фракционному составу, полученные методом КБС с использование наборов дескрипторов K_d (а), ММР (б) и ε^* по ММ (в).

Как видно, дискриминирующая способность наборов дескрипторов увеличивается в ряду ММР < K_d < ε* по ММ. Наибольшее количество правильных отнесений достигнуто для классификационных правил, построенных с использованием набора дескрипторов распределения ε* по ММ при участии трех исходных дескрипторов ($Q_{контр}$ = 0.864). На рис. 6.5 приведены лучшие классификации, полученные с использованием каждого из трех наборов дескрипторов.

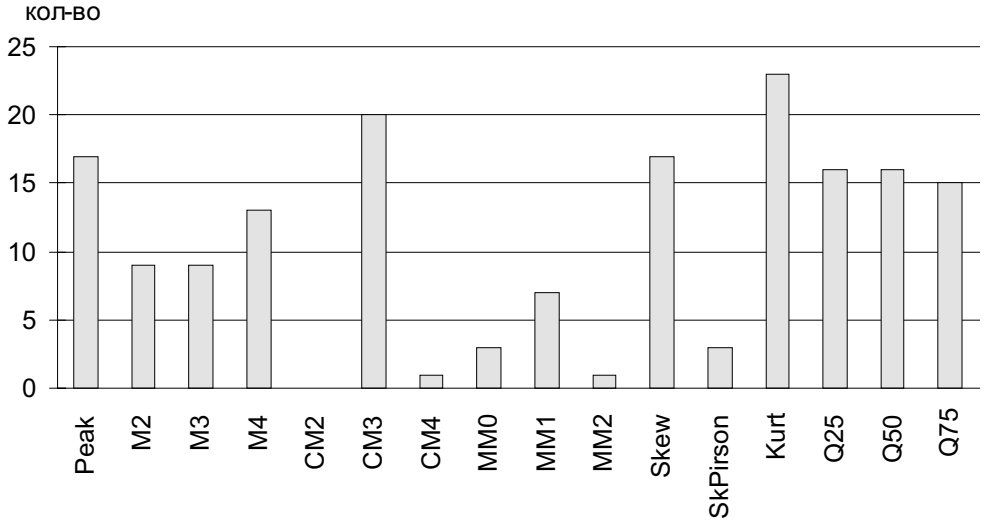
Для оценки дискриминирующей способности индивидуальных дескрипторов, входящих в состав рассматриваемых наборов, был проведен анализ частоты их встречаемости в лучших классификациях, полученных с использованием каждого набора. Результаты приведены на рис. 6.6.

Как видно из приведенных диаграмм, из набора дескрипторов кривых элюирования в шкале K_d чаще всего встречаются $(K_d)_p$, m_4, cm_3, A, E, Q_{25}, Q_{50} и Q_{75}; из дескрипторов ММР – M_z/M_w, M_p, Q_{25}, Q_{50} и Q_{75}; из дескрипторов распределения ε* по ММ – I_{100}, A, Q_{25}, $I_{0\text{-}10}$, $I_{10\text{-}20}$, $I_{60\text{-}70}$. Тем самым каждый из упомянутых дескрипторов характеризуется высокой дискриминирующей способностью по признакам “источник происхождения” и “фракционный состав” гумусовых кислот.

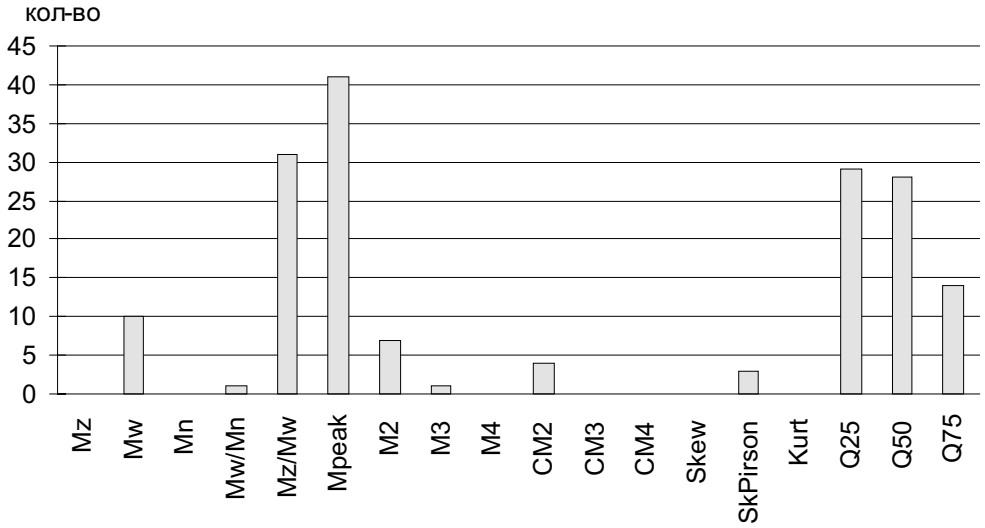
Представлялось целесообразным объединить выявленные “максимально полезные” дескрипторы в новый, расширенный набор дескрипторов ММ состава, с потенциально гораздо более высокой дискриминирующей способностью, чем каждый из отдельных наборов, и провести классификацию с его использованием.

Для этого проводили нормирование, а затем масштабирование указанных дескрипторов, умножая их на дисперсионные веса w_1-w_3 (Приложение 6.1). Классификационные правила, рассчитанные с использованием расширенного набора дескрипторов ММ состава, характеризовались гораздо более высоким качеством: средние $Q_{обуч}$ по 11 лучшим классификациям при участии двух комбинированных дескрипторов составляли 1.0, для 3-4-х исходных дескрипторов – 0.993. Соответствующие $Q_{контр}$ составляли 0.92 и 0.92. На рис. 6.7 для примера показано классификационное поле, построенное для двух комбинированных дескрипторов, обеспечивающих наилучшее отнесение препаратов.

а)



б)



в)

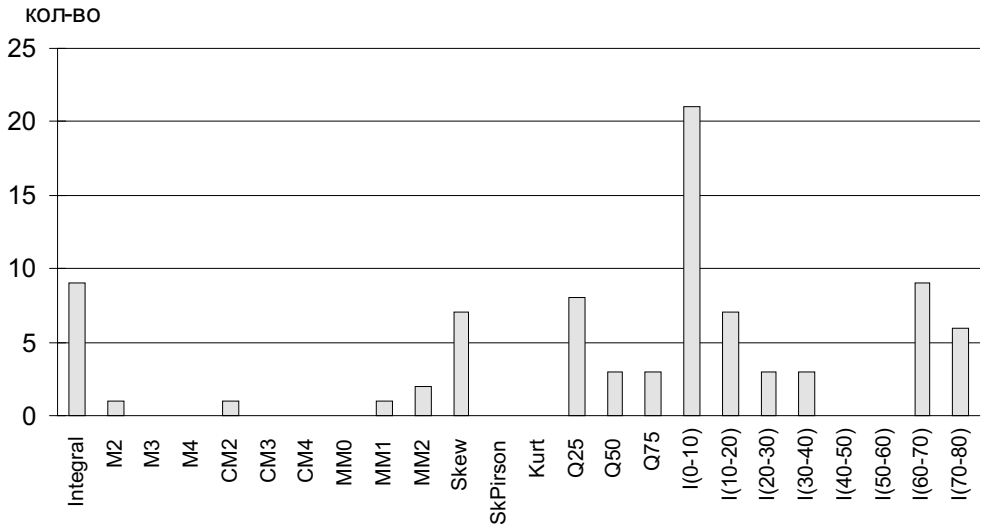


Рис. 6.6. Частота встречаемости дескрипторов кривых элюирования в шкале K_d (а), ММР (б) и распределения ε* по ММ (в). Skew – A, Kurt – E, Peak – K_d пика.

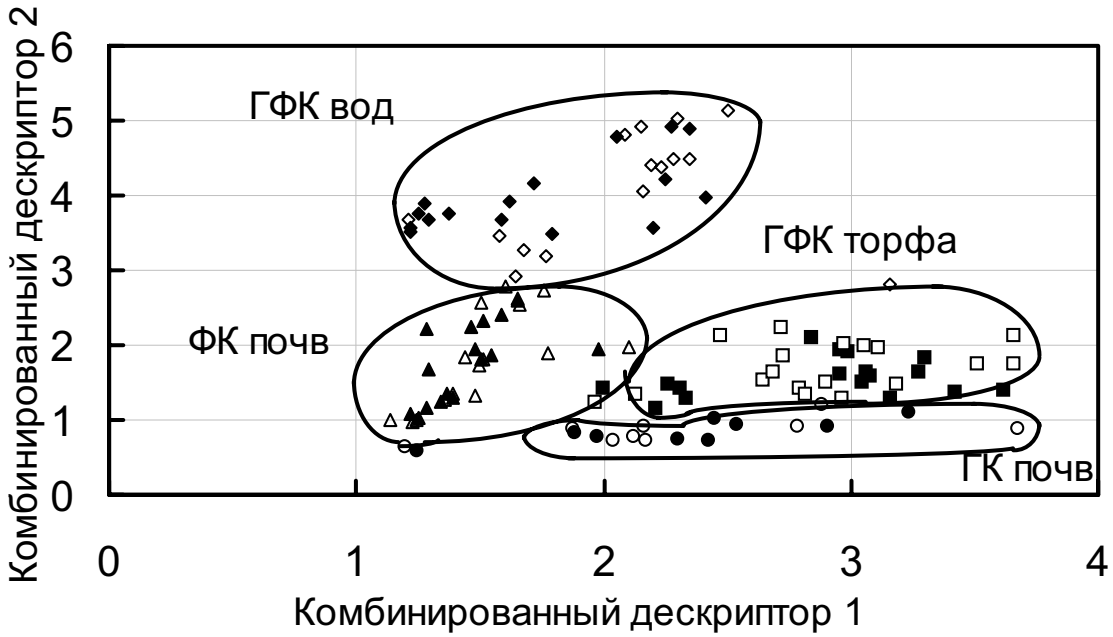


Рис. 6.7. Наилучшая классификация гумусовых кислот по происхождению и фракционному составу, полученная методом КБС с использование расширенного набора ММ дескрипторов.

Таким образом, использование расширенного набора дескрипторов ММ состава (включающего дескрипторы распределений K_d, ММР и ε* по ММ), обладающих высокой дискриминирующей способностью по детализированному признаку “источник происхождения и фракционный состав”, позволило решить задачу классификации гумусовых кислот с применением дескрипторов одного уровня На этом основании данный набор дескрипторов использовали и при построении соответствующего классификационного правила с применением нейронных сетей.

6.3.2 Классификация с использованием нейросетей

В связи с тем, что нейронные сети сравнительно недавно вошли в практику компьютерного моделирования, мы сочли уместным кратко изложить основные принципы этого самого мощного метода классификации и прогнозирования и ввести термины, употребляемые для описания результатов моделирования с использованием нейросетей.

Под искусственной нейронной сетью понимается некоторое вычислительное устройство обработки информации, состоящее из большого числа параллельно работающих простых процессорных элементов – нейронов, связанных между собой линиями передачи информации – синапсами. У нейронной сети выделена группа синапсов, по которым она получает входную информацию, и группа синапсов, с которых снимается

информация на выходе сети. Нейронная сеть обучается решению задачи с помощью обучающей выборки [Горбань, 1990; Горбань и Россиев, 1996]. На рис. 6.8 представлена схема нейрона.

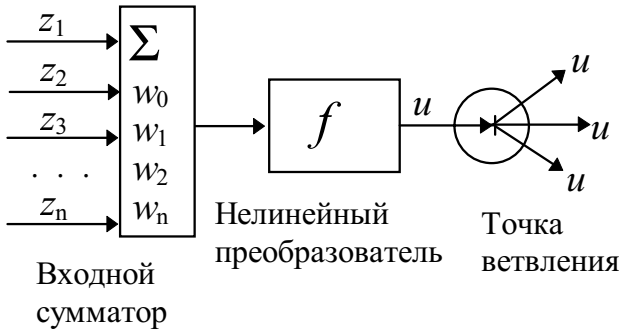


Рис. 6.8. Схема нейрона.

На вход нейрона подаются значения z_1, z_2, ... z_n (значения дескрипторов или выходные сигналы с других нейронов), которые преобразуются по формуле:

$$u = f(w_0 + z_1w_1 + z_2w_2 + z_3w_3 + ...+ z_nw_n) \qquad (6.1.)$$

где f – нелинейная (обычно S-образная) пороговая функция.

Выходное значение u передается на входы других нейронов или является выходным значением нейросети. Параметры w представляют собой веса синапсов сети. Они образуют набор адаптивных параметров, настраивая которые, нейронная сеть обучается решению задачи. В ходе обучения проводится минимизация отклонения (чаще всего минимизируется сумма квадратов отклонений) рассчитанных значений свойств от известных.

Для построения классификационных правил или прогностических моделей нейроны объединяются в нейросети. Наиболее распространены слоистые сети (рис. 6.9), где нейроны расположены в несколько слоев. Нейроны первого слоя получают входные сигналы, преобразуют их и через точки ветвления передают нейронам второго слоя. Далее срабатывает второй слой, и т.д. до k-го слоя, который выдает выходные сигналы. Если не оговорено особо, то каждый выходной сигнал i-го слоя подается на вход всех нейронов i+1-го. Число нейронов в каждом слое может быть любым и никак заранее не связано с количеством нейронов в других слоях. Стандартный способ подачи входных сигналов: каждый нейрон первого слоя получает все входные сигналы.

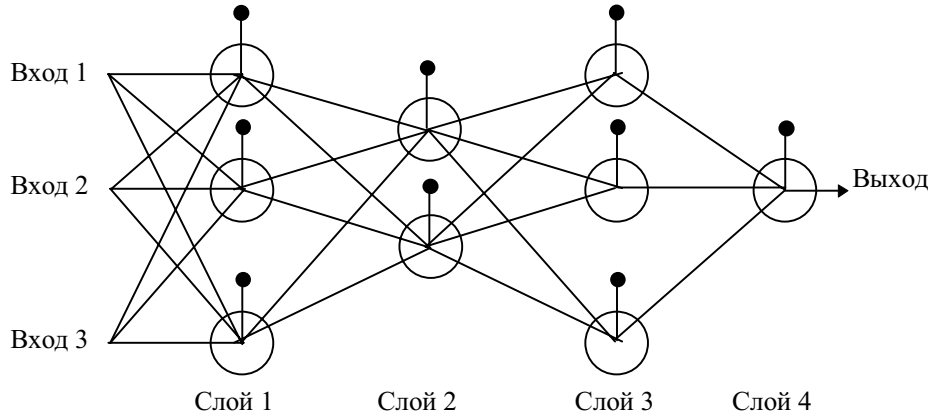


Рис. 6.9. Слоистая нейронная сеть.

К преимуществам метода относится: (1) хорошо работает при наличии сложных разграничительных поверхностей между классами, (2) при правильном построении и обучении нейросети устраняется влияние посторонних дескрипторов, т.е. выбирается оптимальный набор дескрипторов; (3) позволяет оценивать значимость дескрипторов.

К недостаткам метода относится: (1) сложность реализации, (2) большое количество настраиваемых параметров и неочевидный выбор оптимальной топологии, (3) весьма продолжительный процесс обучения, (4) склонность к “переобучению”, т.е. нейросеть может научиться “узнавать” образец из обучающего набора, а не выявлять зависимость между свойствами и дескрипторами, (5) получаемая сложная нелинейная классифицирующая функция или прогностическая модель затрудняет интерпретацию результатов.

Для построения классификационного правила, которое бы позволяло относить исследуемые эксклюзионные хроматограммы, или “препараты”, к классам по источнику происхождения и фракционному составу гумусовых кислот использовали слоистую нейронную сеть. Для выбора оптимальной топологии нейросети тестировали пять различных вариантов с числом нейронов от 2 до 10 и слоев – от 1 до 2. На вход подавали значения всех дескрипторов. Число выходных сигналов соответствовало количеству задаваемых классов. В ходе обучения был задан люфт (т.е. допустимый предел отклонения) ±0.1 для требуемых значений на выходах сети при этом в большинстве случаев количество синапсов и входов сокращалось с сохранением 100% правильной классификации обучающей выборки. Характеристики протестированных топологий и качества полученных классификаций приведены в табл. 6.6.

Таблица 6.6

Характеристики классификаций гумусовых кислот по источнику происхождения и фракционному составу методом нейронных сетей с использованием расширенного набора дескрипторов ММ состава

Кол-во слоев	Кол-во нейронов	Кол-во входов	Кол-во синапсов	$Q_{обуч}$	$Q_{контр}$
1	10	3	62	1	0.95
1	5	5	40	1	0.95
1	3	9	37	1	0.95
1	2	20	54	0.98	0.72
2	8	3	35	1	0.95

Как видно, качество классификации контрольной выборки было лучше, чем в среднем для метода КБС: величина $Q_{контр}$ составляла 95% для 4 из 5 протестированных топологий. Исходя из принципа, что простая модель более предпочтительна при равном качестве классификаций, оптимальной можно считать топологии с одним слоем и 3 нейронами и с двумя слоями и 8 нейронами. Однако в последнем случае число используемых входов снизилось до 3, а количество синапсов – до 35. Стоит заметить, что топология с одним слоем нейронов (не считая сумматора на выходе) реализует линейную дискриминантную модель, а с двумя слоями – нелинейную

Для четырех топологий нейросетей была рассчитана относительная значимость дескрипторов, представляющая собой нормированную сумму модулей весовых коэффициентов синапсов, относящихся к соответствующему входу. Результаты приведены на рис. 6.10.

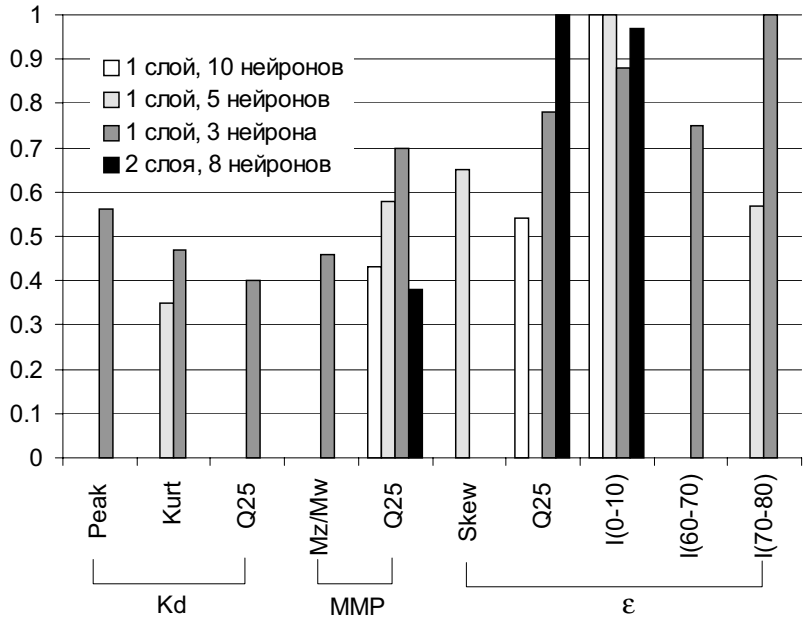


Рис. 6.10. Относительная значимость дескрипторов, использованных в нейросетях различных топологий.