# Кластерный анализ ДНК-белковых пространственных контактов с использованием процедуры Вороного-Делоне*

Е.Н. Кузнецов, А.А. Анашкина, А.А. Дорофеюк, Ю.А. Дорофеюк,
Н.Г. Есипова, А.Г. Спиро, В.Г. Туманян

**Аннотация.** Предлагается классификация аминокислотных остатков по признакам контактов аминокислот белков с нуклеотидами ДНК, рассматриваются классификации с разными типами размытости. Для определения количества и площади контактов каждой аминокислоты с каждым нуклеотидом использовалось разбиение Вороного-Делоне. Показано существование инвариантов кластеризации аминокислот, а также то, что размытая классификация аминокислот на 6 классов является оптимальной для задачи белок-нуклеинового распознавания.

**Ключевые слова:** *кластерный анализ, размытая классификация, контакты аминокислота–нуклеотид, разбиение Вороного–Делоне, свойства аминокислотных остатков.*

## 1. Введение

Проблема специфичности взаимодействия ДНК-белок лежит в основе понимания механизмов экспрессии генов, а, следовательно, механизмов реализации генетической информации на различных уровнях строения биообъектов. Различают специфическое и неспецифическое связывание нуклеиновых кислот белком. Под первым понимается избирательное взаимодействие определенного участка нуклеиновой кислоты с определенным белком, под вторым – равновероятное взаимодействие белка с различными последовательностями нуклеиновых кислот в различных участках генома [1, 2].

Из анализа первых рентгеновских структур белок-нуклеиновых комплексов стало очевидно, что в создание комплекса вносят свой вклад множество различных факторов: водородные связи, опосредованные водой контакты, взаимные конформационные перестройки, изгибы и искажения, высвобождение ионов, электростатика, Ван-дер-Ваальсовые взаимодействия, гидрофобный эффект. Все многообразие информации о правилах, управляющих биомолекулярным распознаванием, получено из структурных данных, в основном из рентгеноструктурного анализа и ЯМР.

Белок и ДНК различаются структурно и химически. В комплексах белок-ДНК молекулярные интерфейсы пространственно комплементарны, и распознавание является точным структурным процессом. Стереохимическая ориентация взаимодействующих поверхностей партнеров определяет комплементарность химических контактов и неизбежно влечет за собой существование молекул с комплементарными водородными донорными и акцепторными группами. Это означает химическое распознавание.

В данной работе мы задались целью найти способ классификации аминокислот, наиболее интегрально учитывающий факторы, определяющие образование специфических комплексов ДНК-белок. Известны различные классификации аминокислотных остатков, основанные, в частности, на их физико-химических свойствах [3, 4], на анализе точечных мутаций и кластеризации матриц замен [5], на анализе соседних по последовательности аминокислотных остатков [6] и так далее. При этом используется большое разнообразие методов кластер-анализа и автоматической классификации, в том числе методы иерархической классификации [7, 8], методы типа *k*-средних, вариационные методы классификации, методы многомерного шкалирования [4] и другие. Очевидно, что универсальной классификации аминокислот не существует, и каждая классификация предназначается для целей определенного исследования [9]. Это означает, что имеет смысл говорить о контекст-зависимой классификации для решения конкретной задачи.

Для поиска конкретных способов реализации белок-нуклеинового узнавания, мы решили создать классификацию аминокислот на основе анализа геометрических характеристик структур комплексов белок-ДНК. Аминокислотные остатки в составе белков, взаимодействующих с ДНК,

* Работа выполнена при частичной финансовой поддержке РФФИ: гранты 14-07-00463-а, 15-04-99605-а, 16-07-00895-а, 16-29-07433-офи.

образуют пространственные контакты с нуклеиновыми основаниями и сахарофосфатным остовом ДНК. Мы провели анализ пространственного взаимного расположения аминокислотных остатков и нуклеотидов на большой выборке комплексов белок-ДНК (1937 комплексов, т.е. все известные структуры белок-ДНК в базе данных Protein DataBank на момент исследования). Для расчета количества и площади контактов использовался подход, основанный на пространственном разбиении Вороного-Делоне [10, 11].

Для проверки надежности предлагаемого подхода к решению общей проблемы белок-нуклеинового узнавания доступная выборка пространственных структур белок-ДНК была разбита на две подвыборки в 987 и 950 комплексов. Было показано совпадение результатов классификаций для каждой подвыборки и выборки в целом для иерархических методов классификации, а для вариационных – совпадение с точностью до задания начальных условий.

В данной работе впервые в основу классификации аминокислот положены геометрические характеристики структур комплексов белок-ДНК. Для конкретного представления пространственного взаимодействия аминокислотных остатков белков и нуклеотидов ДНК использовано разбиение Вороного-Делоне. В качестве признаков для применения методов кластер-анализа использованы как статистика контактов, так и статистика площадей контактов между аминокислотными остатками и нуклеотидами белок-нуклеиновых комплексов.

## 2. Разбиение Вороного-Делоне

Для любого центра из системы центров можно указать область пространства, все точки которой ближе к данному центру, чем к любому другому. Такая область называется многогранником Вороного или областью Вороного. Разбиение Вороного разделяет пространство между набором центров. Каждый центр системы посредством граней многогранника Вороного определяет своих геометрических соседей. Те, в свою очередь, определяют своих соседей и т.д. Таким образом, данный метод распределяет пространство внутри белковой глобулы между всеми ее атомами по следующему принципу: разделяющая плоскость проводится между двумя соседними атомами через середину отрезка, соединяющего эти атомы, и перпендикулярно ему. Такие плоскости образуют вокруг каждого атома выпуклый многогранник произвольного вида, называемый полиэдром Вороного. Область внутри многогранника лежит ближе к данному атому, чем к любому другому. Таким образом, контакт между двумя атомами существует, если у этих атомов есть общая грань полиэдра Вороного с площадью, отличной от нуля. Следовательно, контакт между двумя аминокислотами определяется как совокупность общих граней полиэдров Вороного составляющих их атомов. Площадь такого контакта определяется как сумма площадей граней составляющих его атомарных контактов.

С помощью программы, реализующей трехмерное разбиение Вороного–Делоне для координат атомов структур в формате PDB, исследовали полученные на основе данных рентгеноструктурного анализа комплексы ДНК–белок. При отборе рассматривали только структуры, содержащие одновременно как белковые цепи, так и ДНК, и исключали структуры, содержащие РНК или ДНК/РНК–гибриды. Всего исследовали 1937 структур.

Контакты между белками и ДНК были вычислены на основе анализа координат атомов пространственных структур белок–ДНК методом разбиения Вороного–Делоне [12]. Помимо информации о контактах, в результате применения этого метода мы имеем данные о площади общей грани полиэдров соседних атомов. Таким образом, результатом проведенного разбиения Вороного–Делоне являются таблицы контактов как между атомами аминокислот и атомами нуклеотидов, так и между более крупными пространственными единицами – аминокислотными остатками и нуклеотидами, как по числу контактов, так и по суммарной площади. Ранее мы применили это разбиение для анализа белок–белковых и белок–нуклеиновых взаимодействий [10, 11]. Программа для построения разбиения написана на языке C++.

## 3. Классификация аминокислотных остатков на основе сравнительного анализа контактов в структурах комплексов белок–ДНК и их специфические взаимодействия

Белок–нуклеиновое распознавание представляется сложным многоступенчатым процессом, и найти соответствие между типами аминокислотных остатков и типами распознаваемых ими нуклеиновых оснований, т.е. так называемый «код» ДНК–белкового узнавания, было и остается мечтой множества исследователей. Спустя годы поисков стало понятно, что простого, единственного кода белок–нуклеинового узнавания не существует. Существует ли вырожденный код или несколько таких кодов, когда одной группе нуклеотидов соответствует определенная группа ами-

нокислотных остатков? Чтобы развить подход к решению такого сложного вопроса, мы задались целью найти способ классификации аминокислот, наиболее интегрально включающей признаки, определяющие образование специфических комплексов ДНК–белок. Известны различные классификации аминокислотных остатков, основанные, в частности, на их физико–химических свойствах, на анализе точечных мутаций, на анализе соседних по последовательности аминокислотных остатков, кластеризации матриц замен и так далее. Это означает, что имеет смысл говорить о контекст–зависимой классификации для решения конкретной задачи. В нашем случае, для поиска способов реализации белок–нуклеинового узнавания, мы решили создать классификацию аминокислот на основе анализа геометрических характеристик структур комплексов белок–ДНК. Аминокислотные остатки в составе белков, взаимодействующих с ДНК, образуют пространственные контакты с нуклеиновыми основаниями и сахарофосфатным остовом ДНК. Для построения независимой классификации аминокислотных остатков, наилучшим образом применимой для установления вырожденного кода узнавания белком ДНК, можно использовать статистику контактов аминокислот с нуклеотидами. Мы провели анализ поведения аминокислотных остатков по отношению к нуклеотидам на основе представительной статистики, полученной нами в данной работе с помощью разбиения Вороного–Делоне. При этом впервые в основу классификации аминокислот положены как статистика контактов, так и статистика площадей контактов между аминокислотными остатками и нуклеотидами белок–нуклеиновых комплексов.

Статистика контактов и площадей контактов аминокислота/нуклеотид, полученная с использованием процедуры Вороного–Делоне, представлена в табл. 1.

**Таблица 1**

Количество и суммарные площади контактов между аминокислотными остатками белка и нуклеотидами ДНК в белок-нуклеиновых комплексах. Получены с помощью анализа пространственных координат атомов остатков и нуклеотидов в 1937 комплексах методом разбиения Вороного-Делоне

| | **Количество контактов, шт.** | | | | **Площади контактов, $A^2$** | | | |
|---|---|---|---|---|---|---|---|---|
| | **A** | **T** | **G** | **C** | **A** | **T** | **G** | **C** |
| **ALA** | 2408 | 2764 | 2553 | 2461 | 16966.05 | 24384.55 | 18979.93 | 20319.00 |
| **ARG** | 11039 | 11319 | 12667 | 9013 | 134455.83 | 138697.44 | 166185.09 | 100840.31 |
| **ASN** | 3285 | 3936 | 3275 | 2980 | 33582.22 | 40044.17 | 32957.42 | 28341.64 |
| **ASP** | 1376 | 1065 | 2060 | 1747 | 10820.57 | 7528.08 | 15287.42 | 13140.64 |
| **CYS** | 328 | 352 | 340 | 341 | 2750.04 | 3128.09 | 1968.50 | 3369.63 |
| **GLN** | 2959 | 2802 | 2687 | 2720 | 29097.04 | 27407.50 | 30136.37 | 28434.73 |
| **GLU** | 1702 | 1776 | 2037 | 2079 | 12592.60 | 13869.09 | 16019.02 | 16700.26 |
| **GLY** | 3561 | 4144 | 3597 | 3494 | 27282.42 | 35127.11 | 29074.57 | 27561.70 |
| **HIS** | 1895 | 2372 | 1951 | 1356 | 19315.16 | 24701.81 | 21231.09 | 12459.42 |
| **ILE** | 2004 | 2169 | 2026 | 1790 | 17583.49 | 21376.97 | 21771.69 | 14961.05 |
| **LEU** | 1907 | 2203 | 1812 | 1698 | 15658.60 | 22519.82 | 17680.12 | 14674.31 |
| **LYS** | 7964 | 8156 | 8184 | 7123 | 79052.97 | 83339.78 | 85176.25 | 67237.77 |
| **MET** | 741 | 1128 | 1008 | 742 | 7884.58 | 12799.33 | 10824.30 | 8368.55 |
| **PHE** | 1458 | 1847 | 1643 | 1461 | 20434.49 | 25979.69 | 19290.77 | 17668.08 |
| **PRO** | 1905 | 2070 | 1610 | 1586 | 17408.12 | 17352.59 | 12599.16 | 11623.36 |
| **SER** | 3998 | 4897 | 4596 | 3496 | 37826.26 | 52214.91 | 44773.53 | 31770.22 |
| **THR** | 4066 | 4902 | 4095 | 3517 | 40021.98 | 54137.06 | 40657.42 | 34666.27 |
| **TRP** | 544 | 625 | 680 | 790 | 7120.88 | 10196.77 | 8758.85 | 10829.64 |
| **TYR** | 2476 | 2906 | 2992 | 2389 | 29171.89 | 39001.30 | 38277.34 | 31241.26 |
| **VAL** | 2263 | 2483 | 2097 | 1733 | 23049.29 | 20393.23 | 18528.68 | 14429.04 |

Эта статистика является промежуточным результатом в рамках способа классификации аминокислот применительно к процессам белок–нуклеинового взаимодействия. Числа в таблице отражают количество случаев (число событий), когда аминокислота, соответствующая строке, образует контакт (пространственно сближена) с нуклеотидом, соответствующим столбцу. Определение сходства аминокислотных остатков путем анализа матриц контактов и площадей контактов. Измерение близости между аминокислотами мы проводили на основе сравнения соответствующих строк в матрице контактов. В качестве меры близости (сходства) строк были использованы девять мер расстояния: D1–D4, D7–D10, D12 [13]. Далее анализ расстояний между соответствующими строками матрицы контактов проводили при помощи различных методов кластеризации.

Иерархические методы кластеризации включали метод ближней связи, дальней связи и средней связи. Неиерархические методы включали метод *k*-средних и метод Уорда с заданным числом классов от четырех до семи и с разными критериями кластеризации: Trace(W), Trace(W)/Median и Wilks' Lambda, где W – внутрикластерная ковариационная матрица [14].

В результате применения девяти методов оценки расстояний и трех иерархических методов кластеризации было получено 27 иерархических деревьев, отражающих взаимосвязь между характеристиками аминокислотных остатков. Методом сравнения 27 деревьев получено следующее суммарное дерево, наиболее полно отражающее результаты кластеризации (рис. 1).

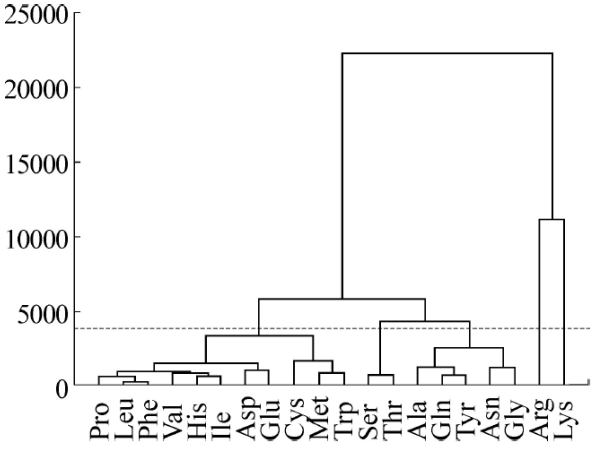


**Рис. 1.** Суммарное дерево, полученное в результате анализа статистики распределения контактов девятью мерами расстояния и тремя иерархическими методами кластеризации. Данное дерево совпадает с результатами применения метода средней связи к данным, полученным с помощью расстояния D1.

Это дерево полностью совпадает с деревом, полученным в результате оценки расстояний по формуле D1 и использования метода кластеризации «среднего расстояния». Во всех 27 случаях было найдено, что следующие аминокислоты группируются в пары: Leu и Phe; Ser и Thr; Met и Trp; Asn и Gly; Gln и Tyr. His и Ile группируются в пары в 19 случаях из 27, Arg и Lys в 21 случае из 27, в 20 случаях из 27 Cys входит в группу с Met и Trp. Ala входит в группу Gln и Tyr в 25 случаях. В 17 случаях из 27 образуется группа из аминокислот Leu, Phe, Pro, His, Ile, Val. В остальных случаях все эти шесть аминокислот располагаются в одном классе, дополняясь глютаминовой кислотой. По результатам иерархической кластеризации можно выделить шесть классов аминокислот: I – Leu, Phe, Pro, His, Ile, Val. II – Asp, Glu. III – Met, Trp, Cys. IV – Ser, Thr. V – Gln, Tyr, Asn, Gly, Ala. VI – Arg, Lys.

Главным физико–химическим свойством, объединяющим аминокислоты класса I, является большое количество неполярных групп, входящих в боковые радикалы этих аминокислот. Это характерно и для гистидина, невзирая на его заряд. Класс II наблюдается только в девяти случаях из 27, в остальных случаях отрицательно заряженные аспарагиновая и глютаминовая кислоты могут присоединяться к классу I или оставаться в фоновом классе. В классе III аминокислоты являются неполярными. В девяти случаях из 27 цистеин образует свой собственный класс. Обратим внимание на то, что метионин и цистеин являются серосодержащими аминокислотами, а триптофан содержит ароматическое кольцо. Таким образом, аминокислоты этого класса обладают большими боковыми радикалами и поэтому занимают значительное пространство в интерфейсе белок–ДНК. Класс IV образуют аминокислоты, обладающие очень близкими физико–химическими свойствами: содержат одинаковые функциональные группы и имеют одинаковую длину бокового радикала. Класс V интересен в нескольких аспектах. Во–первых, он содержит всегда составляющие пару две разных аминокислоты – аспарагин и глицин. Причем аспарагин и глицин образуют пару при использовании любого способа вычисления расстояния и применении любого иерархического метода кластеризации. Именно эти аминокислоты могут принимать конформации, запрещенные для остальных аминокислот. Например, они входят в состав некоторых бета–изгибов II′–типа [15]. Аминокислоты этой группы (глутамин, аспарагин, аланин) часто входят в состав левой спирали типа PPII. Класс VI включает в себя аргинин и ли-

зин, которые в большинстве случаев образуют самостоятельные классы и поэтому могут являться самостоятельными объектами для возникновения кодовых комбинаций, важных при развитии ДНК–белкового узнавания.

Мы приводим ниже только несколько примеров результатов неиерархической классификации, допускающих интерпретацию и определенное сравнение с иерархическими методами. Например, в результате анализа евклидового расстояния (D2) методом средней связи Кинга и методом Уорда мы получили, что образуется четыре класса: A – Ala, Gln, Asn, Ser. B – Arg. C – Asp, Glu, His, Cys. D – Lys.

Остальные аминокислоты остаются в фоновом классе. Однако это искупается возможностью ясной интерпретации выявленных классов. Класс A содержит аминокислоты, соответствующие левой спирали типа PPII, о которой мы говорили при обсуждении класса V иерархической кластеризации. Класс B характерен для протаминов, класс C – для альфа–спиральных структур, а класс D – для гистонов. Методом к–средних Мак–Куина частично воспроизведены результаты иерархических методов: 1 – Gln, Tyr, Asn, Gly, Ser, Thr. 2 – Arg, Lys. 3 – Ala, Asp, Glu, Met, Trp, Cys, Leu, Phe, Pro, His, Ile, Val.

Здесь первый класс, по сути, объединяет классы четыре и пять, полученные иерархическими методами классификации, второй класс есть класс шесть, а третий класс объединяет классы 1–3. Исключение составляет лишь аланин. Метод к–средних, если в качестве критерия кластеризации взять нормированную суммарную внутриклассовую дисперсию (Trace(W)/Median), с заданным числом классов от трех до шести, дает противоречивые результаты. Очевидно, что существует не один, а несколько близких по эффективности способов группирования аминокислот в контексте определенной проблемы, при этом признаки, определяющие классификацию, могут быть непосредственно не связанными с их физико–химическими свойствами. Поэтому кластеризация, созданная на основе признаков, выявленных при ДНК–белковом узнавании, не будет адекватной, если мы попытаемся использовать ее в рамках проблемы узнавания белок–белок. Используя различные методы оценки расстояния и способы объединения аминокислот в группы, можно выявить инварианты кластеризации аминокислот. Результаты, полученные с помощью иерархических методов кластеризации, имеют общие характерные черты. Надо еще раз подчеркнуть, что следующие аминокислоты группируются в пары вне зависимости от способа вычисления расстояния и метода кластеризации: Leu и Phe; Ser и Thr; Met и Trp; Asn и Gly; Gln и Tyr. Из пяти пар бинарной классификации только две пары находят четкое физико–химическое и структурное толкование, но все пять пар связаны с различными типами локальных структур полипептидной цепи. Дополнительно отметим подобие структур лейцина и фенилаланина. На верхних уровнях организации состав классов также практически неизменен. Физические свойства аминокислот, такие как гидрофобность, заряд, наличие гидроксильной группы, проявляют себя во взаимодействиях с ДНК не в полной мере, что отражается на классификации этих аминокислот.

Сходства химической структуры боковых радикалов также оказалось недостаточно для разделения аминокислот по группам. Любопытно выглядит объединение в один класс таких разных по физико–химическим свойствам аминокислот, как глютамин, аспарагин, тирозин, глицин и аланин. Метионин, триптофан и цистеин, образующие класс III, также обладают очень разными физико–химическими свойствами. Метионин и цистеин являются серосодержащими аминокислотами, в то время как триптофан имеет большую ароматическую группу. Цистеин в семи классификациях из 27 образует собственный класс. В физико–химическом смысле он не имеет аналогов среди аминокислот.

## 4. Вариационный подход к задаче классификации аминокислотных остатков

Проведенный выше классификационный анализ аминокислот, с нашей точки зрения, не полностью описывает все многообразие их свойств. Кроме того, описанные выше методы не позволяют изучить группировку аминокислот в матрицах эволюционных замен. Эти матрицы характерны тем, что замена аминокислоты на аминокислоту того же типа характеризуется некоторым, отличным от нуля, числом. Таким образом, нарушается требование, что сходство объекта с самим собой абсолютно. Для решения этой задачи мы воспользовались общим вариационным подходом к задаче классификационного анализа.

Общий вариационный подход к задаче классификационного анализа формулируется при помощи четырех основных категорий: классифицируемое множество объектов, класс допустимых классификаций, способ описания класса и функционал качества разбиения.

1. *Классифицируемое множество объектов*. В нашей задаче классифицируемое множество

объектов состоит из ***N***=20 типов аминокислотных остатков. Обозначим это множество как $X = \{x_1,...,x_N\}$. Каждый объект ***i*** описывается через коэффициенты матрицы замен аминокислот.

2. *Класс допустимых классификаций*. Пусть требуется разбить множество объектов на K классов. Обозначим принадлежность любого объекта *i* классу *k* через $h_{ik}$. Тогда, в общем случае, размытая классификация нашего множества $X = \{x_1,...,x_N\}$ на K классов описывается матрицей $H(X,K) = \{h_{ik}\}$ размерности *N*K*, отражающей принадлежность каждого объекта *i* к каждому из классов *k*. Вводятся естественные ограничения на значения элементов матрицы. Принадлежность объекта к любому классу $h_{ik}$ принимает значения от нуля до единицы, а сумма принадлежностей объекта *i* ко всем классам равна единице: $\sum_{k=1}^{K} h_{ik} = 1,\ 0 \le h_{ik} \le 1$. Можно рассматривать эту матрицу как вектор–функцию размерности *K* от номера объекта, при этом принадлежность объекта *i* всем классам задается вектор–строкой $H_i = \{h_{i1},...,h_{iK}\}$ [16].
3. *Способ описания класса*. Считается, что объекты *k*-го класса должны хорошо описываться некоторой моделью (эталоном) этого класса [17]. В соответствии с этим вводится в рассмотрение множество возможных эталонов классов *T*. Между элементами множества объектов X и элементами множества эталонов *T* вводится некоторая мера близости $S(i,t)$, $(i \in X, t \in T, S(i,t) \ge 0)$. Таким образом, любой набор из *K* классов описывается вектором *A* эталонов размерности *K*, $A = (a_1,...,a_K)$, $(a_k \in T)$. Тогда близость объекта *i* к классу *k* определяется его близостью $S(i,t_k)$ к соответствующему эталону $t_k$ класса *k*.
4. *Критерий качества классификации* в соответствии с методом обобщенного среднего строится следующим образом

$$(10)\quad F(H,T) = \sum_{k=1}^{K}\sum_{i=1}^{N} S(i,a_k)\varphi(h_{ik}).$$

То есть этот функционал представляет собой суммарную близость всех объектов ко всем классам, представленным их эталонами, с учетом степени принадлежности. Задача состоит в максимизации критерия (10) по вектор–функции $H(X,K) = \{H_i\}$ принадлежности объектов классам и по вектору эталонов классов $A = (a_1,...,a_K)$, $(a_k \in T)$.

Здесь $\varphi(h_{ik})$ – монотонно возрастающая функция, отображающая отрезок [0,1] на себя, причем $\varphi(0) = 0$ и $\varphi(1) = 1$. В литературе рассматривались различные примеры функции $\varphi(h_{ik})$ [16, 18, 19]. Выбор этой функции и ограничения, накладываемые на функцию принадлежности объекта к классу $h_{ik}$, определяет конкретный тип размытости классификации [16]. Для классификации с фоновым классом, фоновому классу присваивается значение ***k***=0, соответственно функция $h_{i0}$ описывает принадлежность объекта ***i*** к фоновому классу.

*Четкая классификация:*

$$0 \le h_{ik} \le 1,\ k = 0,...,K;\ h_{i0} + \sum_{k=1}^{K} h_{ik} = 1.$$

*Размытая классификация:*

$$0 \le h_{ik} \le 1,\ k = 0,...,K;\ (h_{i0})^{\lambda} + \sum_{k=1}^{K} (h_{ik})^{\lambda} = 1,\quad \lambda > 1.$$

В данном случае каждый объект ***i*** в оптимальной классификации принадлежит с ненулевым весом ко всем классам, в том числе и к фоновому. Причем мера его принадлежности к фоновому классу тем больше, чем «дальше» объект от нефоновых классов.

*Классификация с размытой границей:*

$$0 \le h_{ik} \le 1,\ k = 0,...,K;$$

$$\sum_{k=0}^{K} (b - h_{ik})^2 = (K-1)b^2 + (b-1)^2,$$

где *b* – коэффициент размытости границы. Этот случай является промежуточным между двумя предыдущими случаями: оптимальная классификация выделяет области однозначного отнесения к одному из классов (как к обычному, так и к фоновому), а между ними оказываются зоны неоднозначного отнесения, т.е. размываются только границы классов.

*Размытая классификация с четким фоновым классом:*

$$\begin{cases} h_{i0} = 1; h_{ik} = 0; k = 1,...,K \\ h_{i0} = 0; 0 \le h_{ik} \le 1; k = 1,...,K; \sum_{k=1}^{K} (h_{ik})^{\lambda} = 1. \end{cases}$$

Использование такого ограничения приводит к тому, что фоновый класс – четкий, а разбиение на обычные классы – размытое.

*Классификация с размытыми границами между обычными классами и четким фоновым классом:*

$$\begin{cases} h_{i0} = 1; h_{ik} = 0; k = 1,...,K \\ h_{i0} = 0; 0 \le h_{ik} \le 1; k = 1,...,K; \sum_{k=1}^{K} (b - h_{ik})^2 = (K-1)b^2 + (b-1)^2, \end{cases}$$

где ***b*** – коэффициент размытости.

*Классификация с четкими обычными классами и размытым фоном:*

Для того, чтобы размытость была только между фоном и обычными классами, а между классами были четкие границы, нужно ввести единую функцию принадлежности ко всем обычным классам $\hat{h}_i = \sum_{k=1}^{K} h_{ik}$ и ограничения накладывать на $\hat{h}_i$ и $h_{i0}$, как на функции принадлежности для классификации на два класса:

$$0 \le h_{i0} \le 1, 0 \le \hat{h}_i \le 1, (h_{i0})^{\lambda} + (\hat{h}_i)^{\lambda} = 1.$$

Тогда размытость будет только между фоновым классом и объединенным классом, а внутри объединенного класса объект будет относиться к тому классу, к эталону которого он ближе.

*Классификация с размытой границей между обычными классами и фоновым классом:*

$$0 \le h_{i0} \le 1, 0 \le \hat{h}_i \le 1, (b - h_{i0})^2 + (b - \hat{h}_i)^2 = b^2 + (b-1)^2.$$

Для нашей задачи, когда каждый объект можно, исходя из его физических и биологических свойств, отнести одновременно к нескольким классам, интересно воспользоваться классификацией с разными типами размытости.

В работе [16] доказана теоретическая сходимость алгоритма при всех вариантах конкретных функций. Для начала мы исследовали размытую классификацию на разное число классов и со значением показателя размытости $\lambda = 2$ (т.е. фактически размытый вариант кластер-анализа *k*-средних).

## 5. Результаты применения вариационного подхода к кластеризации аминокислотных остатков

В результате применения кластер-анализа аминокислот по геометрическим признакам контактов аминокислот с нуклеотидами в белок-нуклеиновых комплексах (статистике контактов и площадям контактов, вычисленных с помощью разбиения Вороного-Делоне) были получены следующие основные результаты. Для удобства описания результатов размытой классификации мы будем говорить об отнесении аминокислоты к некоторому классу, если значение ее функции принадлежности к этому классу значительно превышает ее принадлежность к другим классам. Введем в качестве меры отличия размытой и четкой классификации сумму модулей разности принадлежностей, нормированную на число классов и число классифицируемых элементов. В табл. 2 приведены результаты размытой и четкой классификации аминокислотных остатков на 2 класса по признакам контактов и площадей контактов с нуклеотидами ДНК. В отдельный класс попали аминокислотные остатки ARG и LYS (класс 1). Положительно заряженные аминокислотные остатки аргинин и лизин играют ключевую роль во взаимодействиях с отрицательно заряженной ДНК. Эти остатки могут формировать контакты сразу с несколькими нуклеотидами одновременно. Также эти остатки ответственны за сближение и посадку белков на ДНК [20]. Второй класс образован из остальных 18 аминокислот, и объединяет алифатические аминокислоты, серосодержащие аминокислоты, отрицательно заряженные и слабо заряженный положительно гистидин. Как известно, четкая классификация не позволяет учесть многообразие свойств и их проявлений в тех или иных типах контактов. В результатах размытой классификации видно, что для серина и треонина принадлежность к обоим классам практически одинакова, и отнесение их к какому-то одному классу, как требует четкая классификация, весьма условно. Эти остатки, обладающие гидроксильной группой в боковом радикале, участвуют в образовании водородных связей с нуклеотидами ДНК. Отличие для размытой и четкой классификации по признакам контактов составило 0,1805, по признакам площадей 0,149.

В табл. 3 и 4 приведены результаты размытой и четкой классификации аминокислотных остатков на 4 и 6 классов соответственно, по признакам контактов и площадей контактов с нуклеотидами ДНК.

В табл. 3 положительно заряженные аминокислоты аргинин и лизин по-прежнему образуют отдельный класс (3 класс). В то же время, результаты размытой классификации указывают на многообразие свойств лизина, входящего с принадлежностью не менее 0,15 во все классы. В отдельный класс объединяются аминокислоты, образующие водородные связи с ДНК: аспарагин, глутамин, глицин, серин, треонин (класс 1). Размытая классификация объединяет гидрофобные остатки Ile, Val, Leu, Ala, а также His, Gln и Tyr в один класс (класс 4). Класс 2 включает в себя отрицательно заряженные аминокислоты Asp и Glu, серусодержащие аминокислоты метионин и цистеин, а также фенилаланин, пролин и триптофан. Четкая классификация не полностью воспроизводит результаты размытой классификации. Мера отличия составляет 0,238. Так, можно увидеть, что классы 1 и 3 совпадают в обеих классификациях, а в классах 2 и 4 наблюдаются различия. Также, задав некий по-

**Таблица 2**

Результаты размытой и четкой классификации количества и площади контактов аминокислот белков с нуклеотидами ДНК при заданном числе классов 2 и коэффициенте размытости $\lambda = 2$ (методом k-средних)

| | Число контактов | | | | Площадь контактов | | | |
|---|---|---|---|---|---|---|---|---|
| | Размытая кластеризация | | Четкая кластеризация | | Размытая кластеризация | | Четкая кластеризация | |
| № класса | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| **ALA** | 0.09 | 0.91 | 0 | 1 | 0.04 | 0.96 | 0 | 1 |
| **ARG** | 0.66 | 0.34 | 1 | 0 | 0.66 | 0.34 | 1 | 0 |
| **ASN** | 0.28 | 0.72 | 0 | 1 | 0.22 | 0.78 | 0 | 1 |
| **ASP** | 0.13 | 0.87 | 0 | 1 | 0.14 | 0.86 | 0 | 1 |
| **CYS** | 0.22 | 0.78 | 0 | 1 | 0.20 | 0.80 | 0 | 1 |
| **GLN** | 0.15 | 0.85 | 0 | 1 | 0.14 | 0.86 | 0 | 1 |
| **GLU** | 0.07 | 0.93 | 0 | 1 | 0.10 | 0.90 | 0 | 1 |
| **GLY** | 0.35 | 0.65 | 0 | 1 | 0.15 | 0.85 | 0 | 1 |
| **HIS** | 0.07 | 0.93 | 0 | 1 | 0.05 | 0.95 | 0 | 1 |
| **ILE** | 0.03 | 0.97 | 0 | 1 | 0.04 | 0.96 | 0 | 1 |
| **LEU** | 0.05 | 0.95 | 0 | 1 | 0.06 | 0.94 | 0 | 1 |
| **LYS** | 0.82 | 0.18 | 1 | 0 | 0.95 | 0.05 | 1 | 0 |
| **MET** | 0.18 | 0.82 | 0 | 1 | 0.14 | 0.86 | 0 | 1 |
| **PHE** | 0.10 | 0.90 | 0 | 1 | 0.03 | 0.97 | 0 | 1 |
| **PRO** | 0.07 | 0.93 | 0 | 1 | 0.10 | 0.90 | 0 | 1 |
| **SER** | 0.48 | 0.52 | 0 | 1 | 0.37 | 0.63 | 0 | 1 |
| **THR** | 0.45 | 0.55 | 0 | 1 | 0.37 | 0.63 | 0 | 1 |
| **TRP** | 0.20 | 0.80 | 0 | 1 | 0.15 | 0.85 | 0 | 1 |
| **TYR** | 0.13 | 0.87 | 0 | 1 | 0.23 | 0.77 | 0 | 1 |
| **VAL** | 0.04 | 0.96 | 0 | 1 | 0.06 | 0.94 | 0 | 1 |

рог отсечения, можно включать одни и те же аминокислотные остатки одновременно в два и более класса. Результаты классификации по контактам и суммарным площадям контактов также немного различаются между собой. Мера отличия четкой и размытой классификаций по признакам площадей контактов составила 0,225.

В табл. 4 по результатам размытой классификации контактов между аминокислотными остатками и нуклеотидами, положительно заряженные аминокислоты аргинин и лизин по-прежнему образуют отдельный класс (1 класс). Аминокислоты, участвующие в образовании водородных связей, оказались рассредоточены по классам 2, 3, 5. Отрицательно заряженные аспарагиновая и глютаминовая кислоты попали в класс с гидрофобными аминокислотами (класс 4). Отдельный класс образовали достаточно редкие аминокислоты цистеин, триптофан и метионин (класс 6). Здесь также результаты размытой классификации отличаются от результатов четкой классификации.

Результаты классификации суммарных площадей в целом повторяют результаты классификации контактов, с некоторыми отличиями. Отличие для размытой и четкой классификации по признакам контактов составило 0,169, по признакам площадей 0,231. Преимущество размытой классификации наглядно видно на примере лизина (Табл. 4). Видно, что лизин входит во все классы с принадлежностью не менее 0,1. В действительности, лизин участвует во всех возможных взаимодействиях с ДНК – образовании ионных мостиков, водородных связей, ван-дер-ваальсовых взаимодействий. Таким образом, размытая классификация позволяет учесть многообразие свойств и проявлений этих свойств аминокислот. Интерпретация результатов классификации зачастую представляет самостоятельную задачу, поскольку только базовых свойств

**Таблица 3**

Результаты размытой и четкой классификации количества и площади контактов аминокислот белков с нуклеотидами ДНК при числе классов 4 и коэффициенте размытости $\lambda = 2$

| | Число контактов | | | | | | | | Площадь контактов | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Размытая кластеризация | | | | Четкая кластеризация | | | | Размытая кластеризация | | | | Четкая кластеризация | | | |
| № класса | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| ALA | 0.19 | 0.27 | 0.03 | 0.51 | 0 | 0 | 0 | 1 | 0.08 | 0.28 | 0.01 | 0.63 | 0 | 0 | 0 | 1 |
| ARG | 0.12 | 0.09 | 0.69 | 0.10 | 0 | 0 | 1 | 0 | 0.05 | 0.04 | 0.86 | 0.05 | 0 | 0 | 1 | 0 |
| ASN | 0.61 | 0.15 | 0.04 | 0.20 | 1 | 0 | 0 | 0 | 0.59 | 0.16 | 0.03 | 0.22 | 1 | 0 | 0 | 0 |
| ASP | 0.10 | 0.55 | 0.03 | 0.33 | 0 | 0 | 0 | 1 | 0.11 | 0.57 | 0.03 | 0.29 | 0 | 1 | 0 | 0 |
| CYS | 0.17 | 0.43 | 0.06 | 0.33 | 0 | 1 | 0 | 0 | 0.17 | 0.46 | 0.05 | 0.33 | 0 | 1 | 0 | 0 |
| GLN | 0.28 | 0.27 | 0.04 | 0.41 | 0 | 0 | 0 | 1 | 0.31 | 0.26 | 0.03 | 0.40 | 1 | 0 | 0 | 0 |
| GLU | 0.07 | 0.48 | 0.02 | 0.43 | 0 | 0 | 0 | 1 | 0.06 | 0.69 | 0.01 | 0.23 | 0 | 1 | 0 | 0 |
| GLY | 0.83 | 0.07 | 0.02 | 0.08 | 1 | 0 | 0 | 0 | 0.37 | 0.23 | 0.03 | 0.36 | 1 | 0 | 0 | 0 |
| HIS | 0.08 | 0.41 | 0.02 | 0.50 | 0 | 0 | 0 | 1 | 0.09 | 0.31 | 0.01 | 0.59 | 0 | 0 | 0 | 1 |
| ILE | 0.04 | 0.21 | 0.01 | 0.74 | 0 | 0 | 0 | 1 | 0.07 | 0.35 | 0.01 | 0.56 | 0 | 0 | 0 | 1 |
| LEU | 0.06 | 0.40 | 0.01 | 0.53 | 0 | 0 | 0 | 1 | 0.07 | 0.49 | 0.01 | 0.43 | 0 | 0 | 0 | 1 |
| LYS | 0.23 | 0.15 | 0.46 | 0.16 | 0 | 0 | 1 | 0 | 0.32 | 0.21 | 0.25 | 0.22 | 0 | 0 | 1 | 0 |
| MET | 0.13 | 0.50 | 0.04 | 0.33 | 0 | 1 | 0 | 0 | 0.11 | 0.56 | 0.03 | 0.30 | 0 | 1 | 0 | 0 |
| PHE | 0.04 | 0.78 | 0.01 | 0.17 | 0 | 0 | 0 | 1 | 0.06 | 0.17 | 0.01 | 0.76 | 0 | 0 | 0 | 1 |
| PRO | 0.06 | 0.55 | 0.01 | 0.37 | 0 | 0 | 0 | 1 | 0.07 | 0.65 | 0.01 | 0.26 | 0 | 0 | 0 | 1 |
| SER | 0.65 | 0.13 | 0.06 | 0.16 | 1 | 0 | 0 | 0 | 0.71 | 0.11 | 0.03 | 0.14 | 1 | 0 | 0 | 0 |
| THR | 0.71 | 0.11 | 0.05 | 0.13 | 1 | 0 | 0 | 0 | 0.69 | 0.12 | 0.04 | 0.15 | 1 | 0 | 0 | 0 |
| TRP | 0.15 | 0.46 | 0.05 | 0.33 | 0 | 1 | 0 | 0 | 0.12 | 0.55 | 0.03 | 0.30 | 0 | 1 | 0 | 0 |
| TYR | 0.25 | 0.27 | 0.04 | 0.44 | 0 | 0 | 0 | 1 | 0.61 | 0.15 | 0.03 | 0.21 | 1 | 0 | 0 | 0 |
| VAL | 0.07 | 0.22 | 0.01 | 0.70 | 0 | 0 | 0 | 1 | 0.09 | 0.37 | 0.02 | 0.53 | 0 | 0 | 0 | 1 |

аминокислот насчитывается более десяти, всего же на данный момент в базе данных AAindex содержится 544 различных свойств для каждого типа аминокислотного остатка [21].

Размытая классификация при увеличении числа классов более 6 создает дублирующиеся классы, с одинаковым составом, что указывает на нецелесообразность дальнейшего разделения. Тем самым, позволяет определить естественное максимальное число классов.

## 6. Заключение

Для широкого круга биоинформатических исследований представляет большой интерес уменьшение сложности описания 20 стандартных аминокислот путем их разбиения на группы и создания так называемого «вырожденного алфавита» [22]. Хотя не существует универсального способа классификации аминокислот, имеются многочисленные примеры использования различных методов и алгоритмов кластер-анализа, с одной стороны и различных типов исходной информации для такой группировки (физико-химические свойства, мутации, эволюционные замены, и т.д.), с другой стороны. Впервые в данной работе в качестве исходной информации для классификации аминокислот используются данные о пространственных контактах между аминокислотными остатками и нуклеотидами в структурах комплексов белок-ДНК. При этом для определения таких контактов применяется метод пространственного разбиения Вороного-Делоне. Кроме того, впервые учитывается площадь контакта между соседними атомами. При помощи математической модели показан неслучайный характер таких контактов, а именно около 30% всех контактов между аминокислотами и нуклеотидами в комплексах белок-ДНК являются неслучайными.

**Таблица 4**

Результаты размытой и четкой классификации количества и площади контактов аминокислот белков с нуклеотидами ДНК при заданном числе классов 6 и коэффициенте размытости $\lambda = 2$

| № класса | Число контактов | | | | | | | | | | | | Площадь контактов | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Размытая кластеризация | | | | | | Четкая кластеризация | | | | | | Размытая кластеризация | | | | | | Четкая кластеризация | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
| **ALA** | 0.01 | 0.79 | 0.05 | 0.08 | 0.03 | 0.03 | 0 | 1 | 0 | 0 | 0 | 0 | 0.01 | 0.50 | 0.10 | 0.16 | 0.06 | 0.17 | 0 | 1 | 0 | 0 | 0 | 0 |
| **ARG** | 0.84 | 0.03 | 0.04 | 0.03 | 0.04 | 0.03 | 1 | 0 | 0 | 0 | 0 | 0 | 0.92 | 0.01 | 0.02 | 0.02 | 0.02 | 0.01 | 1 | 0 | 0 | 0 | 0 | 0 |
| **ASN** | 0.02 | 0.14 | 0.58 | 0.08 | 0.13 | 0.05 | 0 | 0 | 1 | 0 | 0 | 0 | 0.01 | 0.06 | 0.61 | 0.17 | 0.10 | 0.04 | 0 | 0 | 0 | 1 | 0 | 0 |
| **ASP** | 0.02 | 0.18 | 0.10 | 0.34 | 0.07 | 0.28 | 0 | 0 | 0 | 0 | 1 | 0 | 0.01 | 0.20 | 0.07 | 0.10 | 0.05 | 0.57 | 0 | 0 | 0 | 0 | 0 | 1 |
| **CYS** | 0.03 | 0.13 | 0.09 | 0.19 | 0.07 | 0.49 | 0 | 0 | 0 | 0 | 0 | 1 | 0.03 | 0.23 | 0.12 | 0.14 | 0.09 | 0.39 | 0 | 0 | 0 | 0 | 0 | 1 |
| **GLN** | 0.02 | 0.49 | 0.17 | 0.15 | 0.09 | 0.07 | 0 | 1 | 0 | 0 | 0 | 0 | 0.01 | 0.12 | 0.20 | 0.52 | 0.08 | 0.07 | 0 | 0 | 0 | 1 | 0 | 0 |
| **GLU** | 0.02 | 0.20 | 0.08 | 0.50 | 0.06 | 0.14 | 0 | 0 | 0 | 0 | 1 | 0 | 0.01 | 0.27 | 0.07 | 0.10 | 0.05 | 0.50 | 0 | 0 | 0 | 0 | 1 | 0 |
| **GLY** | 0.01 | 0.09 | 0.65 | 0.05 | 0.16 | 0.04 | 0 | 0 | 1 | 0 | 0 | 0 | 0.01 | 0.09 | 0.24 | 0.53 | 0.08 | 0.06 | 0 | 0 | 0 | 1 | 0 | 0 |
| **HIS** | 0.01 | 0.17 | 0.07 | 0.57 | 0.05 | 0.12 | 0 | 0 | 0 | 1 | 0 | 0 | 0.01 | 0.56 | 0.09 | 0.13 | 0.06 | 0.16 | 0 | 0 | 1 | 0 | 0 | 0 |
| **ILE** | 0.01 | 0.11 | 0.04 | 0.75 | 0.03 | 0.06 | 0 | 0 | 0 | 1 | 0 | 0 | 0.01 | 0.66 | 0.06 | 0.09 | 0.04 | 0.13 | 0 | 0 | 1 | 0 | 0 | 0 |
| **LEU** | 0.01 | 0.08 | 0.03 | 0.81 | 0.02 | 0.05 | 0 | 0 | 0 | 1 | 0 | 0 | 0.01 | 0.64 | 0.06 | 0.09 | 0.04 | 0.17 | 0 | 0 | 0 | 0 | 1 | 0 |
| **LYS** | 0.25 | 0.14 | 0.17 | 0.12 | 0.20 | 0.11 | 1 | 0 | 0 | 0 | 0 | 0 | 0.15 | 0.14 | 0.19 | 0.17 | 0.23 | 0.13 | 1 | 0 | 0 | 0 | 0 | 0 |
| **MET** | 0.01 | 0.05 | 0.03 | 0.08 | 0.02 | 0.82 | 0 | 0 | 0 | 0 | 0 | 1 | 0.01 | 0.19 | 0.07 | 0.09 | 0.05 | 0.59 | 0 | 0 | 0 | 0 | 0 | 1 |
| **PHE** | 0.02 | 0.15 | 0.08 | 0.47 | 0.06 | 0.22 | 0 | 0 | 0 | 0 | 1 | 0 | 0.01 | 0.52 | 0.10 | 0.16 | 0.06 | 0.14 | 0 | 1 | 0 | 0 | 0 | 0 |
| **PRO** | 0.01 | 0.13 | 0.06 | 0.63 | 0.04 | 0.12 | 0 | 0 | 0 | 1 | 0 | 0 | 0.01 | 0.34 | 0.08 | 0.11 | 0.05 | 0.41 | 0 | 0 | 0 | 0 | 1 | 0 |
| **SER** | 0.01 | 0.05 | 0.12 | 0.04 | 0.76 | 0.03 | 0 | 0 | 1 | 0 | 0 | 0 | 0.01 | 0.05 | 0.13 | 0.08 | 0.70 | 0.04 | 0 | 0 | 0 | 1 | 0 | 0 |
| **THR** | 0.02 | 0.06 | 0.16 | 0.05 | 0.68 | 0.03 | 0 | 0 | 1 | 0 | 0 | 0 | 0.01 | 0.05 | 0.12 | 0.08 | 0.71 | 0.04 | 0 | 0 | 0 | 1 | 0 | 0 |
| **TRP** | 0.02 | 0.10 | 0.06 | 0.15 | 0.05 | 0.63 | 0 | 0 | 0 | 0 | 0 | 1 | 0.02 | 0.19 | 0.08 | 0.10 | 0.06 | 0.56 | 0 | 0 | 0 | 0 | 0 | 1 |
| **TYR** | 0.01 | 0.59 | 0.12 | 0.14 | 0.07 | 0.06 | 0 | 1 | 0 | 0 | 0 | 0 | 0.01 | 0.08 | 0.53 | 0.19 | 0.13 | 0.05 | 0 | 0 | 0 | 1 | 0 | 0 |
| **VAL** | 0.01 | 0.27 | 0.09 | 0.47 | 0.06 | 0.10 | 0 | 0 | 0 | 1 | 0 | 0 | 0.01 | 0.52 | 0.09 | 0.14 | 0.06 | 0.18 | 0 | 0 | 1 | 0 | 0 | 0 |

На основе классических методов кластер-анализа (иерархических, типа *k*-средних, и других) и с применением различных мер близости построены классификации аминокислотных остатков, проанализированы их свойства и выявлены инварианты кластеризации аминокислот. В некоторых случаях объединение аминокислот в классы по признакам пространственных контактов с нуклеотидами совпадает с результатами кластеризации на основе физико-химических свойств аминокислот. Это является еще одним подтверждением адекватности предлагаемого подхода. Было показано совпадение результатов классификаций для выборки в целом и двух ее подвыборок. В тоже время, единое жесткое разбиение аминокислот на фиксированные группы не может отразить сложный характер взаимодействия аминокислот белка и нуклеотидов ДНК, существующий в природе. В связи с этим, предложено использовать вариационные методы для построения различных типов размытой классификации аминокислот (размытая классификация, классификация с перекрывающимися классами, классификация с размытыми границами и с фоновым классом), позволяющие учесть разные аспекты взаимодействий ДНК-белок. Показано, что применение размытой классификации позволяет более адекватно описывать разные аспекты белок-нуклеинового взаимодействия.

## Литература


1. *Gurskii G. V., Tumanian V. G., Zasedatelev A. S., Zhuze A. L., Grokhovskii S. L., Gottikh B. P.* A code governing specific binding of regulatory proteins to DNA and structure of stereospecific sites of regulatory proteins // Mol Biol (Mosk), 1975. Vol. 9, No. 5. pp. 635-651.
2. *Gurskii G. V., Zasedatelev A. S.* Precise relationships for calculating the binding of regulatory proteins and other lattice ligands in double-stranded polynucleotides // Biofizika, 1978. Vol. 23, No. 5. pp. 932-946.
3. *Shen B., Bai J., Vihinen M.* Physicochemical feature-based classification of amino acid mutations // Protein Eng Des Sel, 2008. Vol. 21, No. 1. pp. 37-44.
4. *Venkatarajan M. S., Braun W.* New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties // Journal of Molecular Modeling, 2001. Vol. 7, No. 12. pp. 445-453.
5. *Kosiol C., Goldman N., Buttimore N. H.* A new criterion and method for amino acid classification // J Theor Biol, 2004. Vol. 228, No. 1. pp. 97-106.
6. *Rogov S. I., Nekrasov A. N.* A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences // Protein Eng, 2001. Vol. 14, No. 7. pp. 459-463.
7. *Davies M.N., Secker A., Halling-Brown M., Moss D. S., Freitas A. A., Timmis J., Clark E., Flower D. R.* GPCRTree: online hierarchical classification of GPCR function // BMC Res Notes, 2008. Vol. 1, P. 67.
8. *May A. C.* Towards more meaningful hierarchical classification of amino acid scoring matrices // Protein Eng, 1999. Vol. 12, No. 9. pp. 707-712.
9. *Davies M. N., Secker A., Freitas A. A., Clark E., Timmis J., Flower D. R.* Optimizing amino acid groupings for GPCR classification // Bioinformatics, 2009. Vol. 11, No. 1. pp. 111-122.
10. *Anashkina A., Kuznetsov E., Esipova N., Tumanyan V.* Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces // Proteins, 2007. Vol. 67, No. 4. pp. 1060-77.
11. *Анашкина А.А., Туманян В.Г., Кузнецов Е.Н., Галкин А.В., Есипова Н.Г.* Геометрический анализ ДНК-белковых взаимодействий на основе метода Вороного-Делоне // Биофизика, 2008. Т. 53, №. 3. С. 402-406.
12. *Медведев Н.Н.* Метод Вороного-Делоне в исследовании структуры некристаллических систем. Новосибирск: НИЦ ОИГГМ СО РАН. – 2000
13. *Раушенбах Г.В.* Меры близости и сходства // Анализ нечисловой информации в социологических исследованиях. М.: Наука – 1985.. С. 169–203.
14. *Миркин Б.Г.* Методы кластер-анализа для поддержки принятия решений: обзор. // М.: ВШЭ. – 2011.
15. *Gunasekaran K., Ramakrishnan C., Balaram P.* Disallowed Ramachandran conformations of amino acid residues in protein structures // J Mol Biol, 1996. Vol. 264, No. 1. pp. 191-8.
16. *Бауман Е.В.* Методы размытой классификации (вариационный подход) // Автоматика и телемеханика, 1988. Вып.12. с.143-156.
17. *Дидэ Э.* Методы анализа данных: Подход, основанный на методе динамических сгущений. Пер. с фр. -М.: Финансы и статистика, 1985. – 357 с.
18. *Zadeh L. A.* Fuzzy sets as a basis for a theory of possibility // Fuzzy sets and systems, 1978. Vol. 1, pp. 3-28.
19. *Bezdek J. C.* A convergence theorem for the fuzzy ISODATA clusters algorithms // IEEE Transac-

tions on pattern analysis and machine intelligence. PAMI-2., 1980. pp. 1-8.
20. *Pabo C. O, Sauer R. T.* Protein-DNA recognition // Annu Rev Biochem, 1984. Vol. 53, pp. 293-321.
21. *Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M.* AAindex: amino acid index database, progress report 2008 // Nucleic Acids Res, 2008. Vol. 36, No. Database issue. pp. D 202-205.
22. *von Hippel P. H.* Protein-DNA recognition: new perspectives and underlying themes // Science, 1994. Vol. 263, No. 5148. pp. 769-70.

**Кузнецов Евгений Николаевич.** Старший научный сотрудник ИПУ РАН им. В.А.Трапезникова. К.т.н. Окончил в 1976 г. Московский Институт нефти и газа им. И.М.Губкина. Количество печатных работ: 52. Область научных интересов: интеллектуальные методы анализа данных, классификация, распознавание образов, анализ изображений и сигналов, технологии математического моделирования.
E-mail: enken54@mail.ru.

**Анашкина Анастасия Андреевна.** Научный сотрудник Института молекулярной биологии им. В.А. Энгельгардта РАН (ИМБ РАН), г. Москва, ул. Профсоюзная, д.32. Кандидат физико-математических наук. В 1999 г. окончила Московский физико-технический институт (МФТИ). Количество печатных работ: 21. Область научных интересов: биофизика, системная биология, интеллектуальные методы анализа данных. E-mail: nastya@eimb.ru.

**Дорофеюк Александр Александрович.** Главный научный сотрудник ИСА ФИЦ ИУ РАН. Зав. лабораторией ИПУ РАН им. В.А.Трапезникова. Профессор НИУ ВШЭ. Д.т.н., профессор. Окончид в 1965 г. МФТИ. Количество печатных работ: 232 (в т.ч. 15 монографий). Область научных интересов: математическая статистика, функциональный анализ, интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений, системный анализ.
E-mail: daa2@mail.ru.

**Дорофеюк Юлия Александровна.** Старший научный сотрудник ИПУ РАН им. В.А.Трапезникова. К.т.н. Окончила в 2007 г. МИЭМ, ГТУ. Количество печатных работ: 124 (в т.ч. 2 монографии). Область научных интересов: интеллектуальные методы анализа данных, математическое моделирование в организационных, социальных, экономических, медико-биологических и технических системах; технологии интеллектуальной поддержки принятия решений; структурное прогнозирование.
E-mail: dorofeyuk_julia@mail.ru.

**Есипова Наталья Георгиевна.** Ведущий научный сотрудник Института молекулярной биологии им. В.А. Энгельгардта РАН (ИМБ РАН), г. Москва, ул. Профсоюзная, д.32. Кандидат физико-математических наук. В 1957 г. закончила физический факультет МГУ им. М.В.Ломоносова. Количество печатных работ: 298. Область научных интересов: биофизика, молекулярная биология, биоинформатика.
E-mail: isinfo@eimb.ru.

**Спиро Арнольд Григорьевич**. Старший научный сотрудник ИПУ РАН им. В.А.Трапезникова. К.т.н. Окончил в 1958 г. Ростовский политехнический институт (РосПИ). Количество печатных работ: 43. Область научных интересов: интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений. E-mail: spiro35@mail.ru.

**Туманян Владимир Гайевич**. Заведующий лабораторией Института молекулярной биологии им. В.А. Энгельгардта РАН (ИМБ РАН), г. Москва, ул. Профсоюзная, д.32. Доктор физико-математических наук, профессор. В 1961 г. Закончил физический факультет МГУ им. М.В.Ломоносова. Количество печатных работ: 172. Область научных интересов: биофизика, молекулярная биология, биоинформатика.
E-mail: tuman@imb.imb.ac.ru.