Кластер-анализ пространственных контактов аминокислотных остатков белков с нуклеотидами ДНК*

***Е. Н. Кузнецов*[1], *А. А. Анашкина*[2], *Н. Г. Есипова*[2], *В. Г. Туманян*[2]**

nastya@eimb.ru

[1]Институт проблем управления им. В. А. Трапезникова РАН, Россия, Москва 117997, ул. Профсоюзная, 65; [2]Институт молекулярной биологии им. В. А. Энгельгардта РАН, Россия, Москва 119991, ул. Вавилова, 32

Предлагается классификация аминокислотных остатков по признакам контактов аминокислот белков с нуклеотидами ДНК. Аминокислотные остатки обладают множеством различных свойств и функций и могут одновременно принадлежать к разным классам, поэтому в работе рассматриваются классификации с разными типами размытости. Для определения количества и площади контактов каждой аминокислоты с каждым нуклеотидом в 1937 комплексах использовали разбиение Вороного–Делоне. Задача классификации аминокислотных остатков с разными типами размытости решалась с помощью общего вариационного подхода. Было показано, что около 30% всех контактов между аминокислотами и нуклеотидами в комплексах белок–ДНК являются неслучайными. Методами четкой классификации показано существование инвариантов кластеризации аминокислот. Методами размытой классификации показано, что классификация аминокислот на шесть классов является оптимальной для задачи белок-нуклеинового распознавания.

Ключевые слова: кластер-анализ; размытая классификация; контакты аминокислота-нуклеотид; разбиение Вороного–Делоне; свойства аминокислотных остатков

Cluster analysis for spatial contacts of amino acid residues of proteins with DNA nucleotides*

***E. N. Kuznetsov*[1], *A. A. Anashkina*[2], *N. G. Esipova*[2], *and V. G. Tumanyan*[2]**

[1]Trapeznikov Institute of Control Sciences RAS, 65 Profsoyuznaya Str., Moscow 117997, Russia; [2]Engelhardt Institute of Molecular Biology RAS, 32 Vavilov Str., Moscow 119991, Russia

Background: Amino acids are classified on the basis of protein–DNA contacts geometry and statistics. Amino acid residues have a variety of properties and can simultaneously belong to different classes. So, it was interesting to use the classification of amino acids with different types of fuzzing.

Methods: Voronoi–Delaunay tessellation was used to determine the spatial relationship between the amino acids of proteins and DNA nucleotides from 1937 protein–DNA complexes. General variation approach was used for the classification of amino acids with different types of fusion.

Results: It was shown that about 30% of all contacts between amino acids and nucleotides in protein–DNA complexes are not random. Crisp classification methods showed the existence of clustering invariants of amino acids at the lowest level of association. It was shown by fuzzy classification methods that six classes are optimal for protein–DNA recognition task.

Concluding Remarks: Fuzzy classification of amino acids data can be used to construct the substitution matrix for DNA-binding protein sequences and protein-DNA binding analysis.

*Работа выполнена при финансовой поддержке РФФИ, проекты № 14-04-00639-а и 12-07-00634-а.

Keywords: *cluster analysis; crisp classification; fuzzy classification; protein–DNA interactions*

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 [3, 4, 5].

Вычислительные методы, опирающиеся на кристаллографические исследования, широко и успешно использующиеся для оценки энергии взаимодействий белок-лиганд [6, 7, 8, 9], должны быть применимы и для понимания формирования белок-ДНК комплексов. Исследователи [10, 11] пытались оценить вклад каждой пары аминокислотный остаток/нуклеотид в общую аффинность белка к ДНК. Другой подход, предложенный [12], предполагал, что общая оценка, отражающая комплементарность между белком и его специфической ДНК, может быть вычислена методами статистического анализа частоты взаимодействия между парой аминокислотный остаток/нуклеотид, таким образом подразумевая аддитивность в энергии связывания. Другие попытки качественно или количественно описать взаимодействие между белком и ДНК [13, 14, 15, 16, 17, 18, 19, 20, 21] также опираются на доступные трехмерные кристаллические структуры белков, связанных с ДНК. Таким образом, 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 [22, 23], на анализе точечных мутаций и кластеризации матриц замен [24], на анализе соседних по последовательности аминокислотных остатков [25] и т. д. При этом используется большое разнообразие методов кластер-анализа и автоматической классификации, в том числе методы иерархической классификации [26, 27], методы типа k-средних, вариационные методы классификации, методы многомерного шкалирования [22] и др. Очевидно, что универсальной классификации

аминокислот не существует, и каждая классификация предназначается для целей определенного исследования [28]. Это означает, что имеет смысл говорить о контекст-зависимой классификации для решения конкретной задачи.

Для поиска конкретных способов реализации белок-нуклеинового узнавания авторы 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. Был проведен 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 Data Bank на момент исследования). Для расчета количества и площади контактов использовался подход, основанный на пространственном разбиении Вороного–Делоне [29, 30].

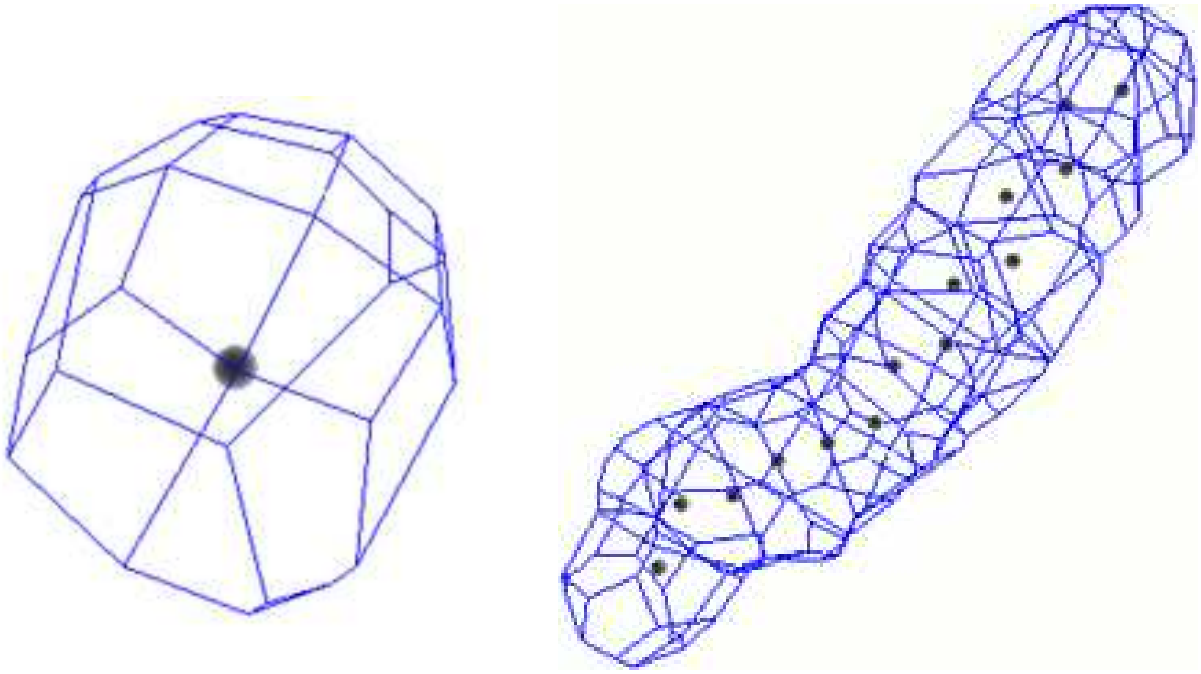
Для проверки надежности предлагаемого подхода к решению общей проблемы белок-нуклеинового узнавания доступная выборка пространственных структур белок-ДНК была разбита на две подвыборки в 987 и 950 комплексов. Было показано совпадение результатов классификаций для каждой подвыборки и выборки в целом для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

В данной работе впервые (1) в основу классификации аминокислот положены геометрические характеристики структур комплексов белок-ДНК; (2) для конкретного представления пространственного взаимодействия аминокислотных остатков белков и нуклеотидов ДНК использовано разбиение Вороного-Делоне; (3) 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.

Разбиение Вороного-Делоне

Для любого центра из системы центров можно указать область пространства, все точки которой ближе к данному центру, чем к любому другому центру системы. 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, можно говорить о графе, вершинами которого являются центры системы, а связность определена через геометрическое соседство. В трехмерном пространстве область Вороного для произвольного центра системы является выпуклым многогранником (полиэдром). Области Вороного для каждого центра системы образуют «сеть» полиэдров, называемую разбиением Вороного [29]. Если внутри описанной сферы тетраэдра, определенного четырьмя центрами, нет других центров системы, такой тетраэдр называется симплексом Делоне. Совокупность всех симплексов Делоне системы заполняет пространство без наложений и щелей, т. е. подобно многогранникам Вороного реализует разбиение пространства, но на этот раз на тетраэдры. Это разбиение называется разбиением Делоне. Как методом пустого шара Делоне, так и с помощью плоскостей Вороного мы выявляем одну и ту же систему центров. Итак, можно говорить о едином разбиении Вороного–Делоне, в котором мы видим одновременно как мозаику многогранников Вороного, так и симплексов Делоне. Каждый симплекс Делоне соответствует определенной вершине Вороного, и, наоборот, каждой вершине Вороного со-

ответствует симплекс Делоне. Эти разбиения являются дуальными, и являются топологически эквивалентными. 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 (рис. 1). 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.



(а) А. Полиэдр Вороного для одного атома (б) Полиэдры Вороного для цепочки атомов

Рис. 1. (*а*) Полиэдр Вороного, построенный вокруг одного атома. В общем случае он является выпуклым многогранником с произвольным числом граней разного размера, зависящем от расположения соседних атомов. Атомы-соседи не показаны. (*б*) Разбиение Вороного цепочки атомов. Атомы-соседи, окружающие цепочку, не показаны

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 [31]. Помимо информации о контактах, в результате применения этого метода мы имеем данные о площади общей грани полиэдров соседних атомов. Таким образом, результатом прове-

денного 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 [29, 30]. Программа для построения разбиения написана на языке C++, ее исходный код доступен по запросу авторам статьи через электронную почту.

Модели случайно и неслучайно контактирующих химических единиц (аминокислот/нуклеотидов)

Для полноценной интерпретации полученных данных нам необходимо опираться на статистическую математическую модель контактирующих аминокислот/нуклеотидов, для того, чтобы оценить и выявить отклонения от случайных явлений. Сделаем несколько принципиальных приближений. Первое состоит в том, что мы будем рассматривать область взаимодействия белковых (белок-нуклеиновых) молекул как поверхность, образованную гранями полиэдров Вороного пар атомов, один из которых принадлежит одной молекуле, а другой – второй молекуле. Второе предположение заключается в том, что все контакты на уровне аминокислотных остатков (остаток/нуклеотид) можно рассматривать как совокупность случайных и неслучайных контактов. Под неслучайными, специфическими контактами подразумеваются контакты, возникающие на участках пространственных структур между определенными химическими группами и/или вследствие определенных типов взаимодействий, сопровождающихся выигрышем в энергии, между элементами на определенных местах в структурах. Такие контакты могут иметь характерную площадь контакта (или несколько, в случае нескольких возможных взаимных расположений). Случайные контакты, в свою очередь, образуются как следствие пространственного сближения двух остатков по причине формирования неслучайных контактов. Таким образом, каждый тип контакта между двумя аминокислотными остатками, например Arg-Glu, может образовывать как случайные, так и неслучайные контакты. В этом приближении оценим распределение площади контакта между остатками на поверхности белок-белкового интерфейса.

Случайные контакты

Предположим, что два круга бросают на некоторую область случайным образом, и каждый раз фиксируют площадь перекрывания. Для упрощения предположим, что эти круги одинаковые с радиусом r, а бросание производится на квадратную область с длиной стороны, равной R. Площадь каждого круга πr^2, тогда площадь их перекрывания лежит в диапазоне $[0, \pi r^2]$. Определим зависимость площади пересечения S от расстояния между центрами кругов L. Очевидно, что если $L \geqslant 2r$, то $S = 0$. Требуется вычислить площадь сектора $AOBs$ и площадь треугольника AOB для вычисления площади сегмента, ограниченного дугой s и отрезком AB. Площадь пересечения кругов будет:

$$S(L) = 2\left(r^2 \arcsin\left(\sqrt{1 - \frac{L^2}{4r^2}}\right) - \frac{L\sqrt{r^2 - L^2/4}}{2}\right) \tag{1}$$

где L расстояние между центрами кругов. Эта формула верна, если $L \in [0, 2r]$.

Пусть координаты центров кругов (x_1, y_1) и (x_2, y_2) соответственно. Тогда расстояние между центрами кругов $l = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. Если мы поместим область для

бросания в начале координат, то координаты центров кругов могут принимать значения из интервала $[r, R-r]$. Найдем вероятность того, что площадь пресечения S упавших кругов равна нулю. Выражаясь математическим языком, расстояние между центрами кругов должно быть больше или равно суммы их радиусов:

$$l \geqslant 2r\,. \tag{2}$$

Как известно, такая вероятность равна отношению объема пространства, удовлетворяющего этому условию, ко всему объему пространства, которое могут принимать значения координат центров кругов. Объем пространства, которое могут принимать значения координат центров кругов $(R-2r)^4$. Для вычисления объема пространства, в котором площадь пересечения кругов будет равна нулю, произведем замену $u_1 = (x_1 - x_2)/\sqrt{2}$, $u_2 = (x_1 + x_2)/\sqrt{2}$, $v_1 = (y_1 - y_2)/\sqrt{2}$, $v_2 = (y_1 + y_2)/\sqrt{2}$.

Тогда матрица перехода будет выглядеть так:

$$\begin{pmatrix} u_1 \\ u_2 \\ v_1 \\ v_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix}\,.$$

Неравенство (2) можно теперь записать в виде

$$u_1^2 + v_1^2 \geqslant 2r^2\,, \tag{3}$$

где u_1 и v_1 могут принимать значения из интервала $[(-R+2r)/\sqrt{2}; (R-2r)/\sqrt{2}]$. Заметим, что неравенство (3) представляет собой пространство вне круга радиусом $\sqrt{2}r$ и внутри квадрата со стороной $R-2r$. Таким образом, вероятность того, что площадь пересечения кругов будет равна нулю

$$P = \frac{((R-2r)^2 - 2\pi r^2)(R-2r)^2}{(R-2r)^4} = \frac{((R-2r)^2 - 2\pi r^2)}{(R-2r)^2} = 1 - \frac{2\pi r^2}{(R-2r)^2}\,. \tag{4}$$

Таким образом, вероятность того, что расстояние между центрами кругов больше L, выражается формулой:

$$P = 1 - \frac{\pi L^2}{2(R-2r)^2}\,. \tag{5}$$

Для нахождения плотности вероятности dP/dS от площади пересечения S можно переписать уравнение в параметрическом виде, поскольку нельзя выразить L как функцию от S в явном виде:

$$\left.\begin{aligned} \frac{dP}{dS} = \frac{dP}{dL}\frac{dL}{dS} = \left(-\frac{\pi L}{(R-2r)^2}\right)\left(-\frac{1}{2\sqrt{r^2 - L^2/4}}\right) = \frac{\pi L}{r(R-2r)^2\sqrt{1 - L^2/(4r^2)}}\,; \\ S(L) = 2\left(r^2 \mathrm{acrsin}\left(\sqrt{1 - \frac{L^2}{4r^2}}\right) - \frac{Lr}{2}\sqrt{1 - \frac{L^2}{4r^2}}\right)\,. \end{aligned}\right\} \tag{6}$$

Зависимость (6) показана на рис. 2, кривая А. По мере увеличения площади контакта число контактов резко уменьшается. Следовательно, среднее распределения близко к нулю. Другими словами, из распределения для случайных контактов видно наличие большого числа малых по площади контактов.

Неслучайные контакты

Логично предположить, что специфические контакты обладают некоторой, отличной от нуля средней площадью контакта, обусловленной физико-химической природой взаимодействия остатков. Предположим, что специфические взаимодействия стремятся образовать максимально большой возможный контакт. В этом случае распределение расстояний между центрами кругов подчиняется нормальному распределению, напоминая задачу о стрельбе по мишени. Выразим распределение площадей неслучайных контактов также в параметрическом виде:

$$\left.\begin{aligned} f(L) &= \frac{1}{\sigma\sqrt{2\pi}} e^{(L-a)^2/(2\sigma^2)}; \\ S(L) &= 2\left(r^2 \operatorname{acrsin}\left(\sqrt{1-\frac{L^2}{4r^2}}\right) - \frac{Lr}{2}\sqrt{1-\frac{L^2}{4r^2}}\right). \end{aligned}\right\} \tag{7}$$

Зависимость (7) показана на рис. 2, кривая B. Кривая имеет куполообразную форму, несимметричная, с некоторым, существенно отличным от нуля средним значением. Распределение для специфических контактов отражает существование некоторой характерной площади контакта. В общем случае уравнения (6) и (7) должны входить в суммарное уравнение, отражающее общее распределение, с некоторыми весовыми функциями, отражающими пропорцию между специфическими и случайными контактами. Площадь под суммарной кривой должна равняться 1.

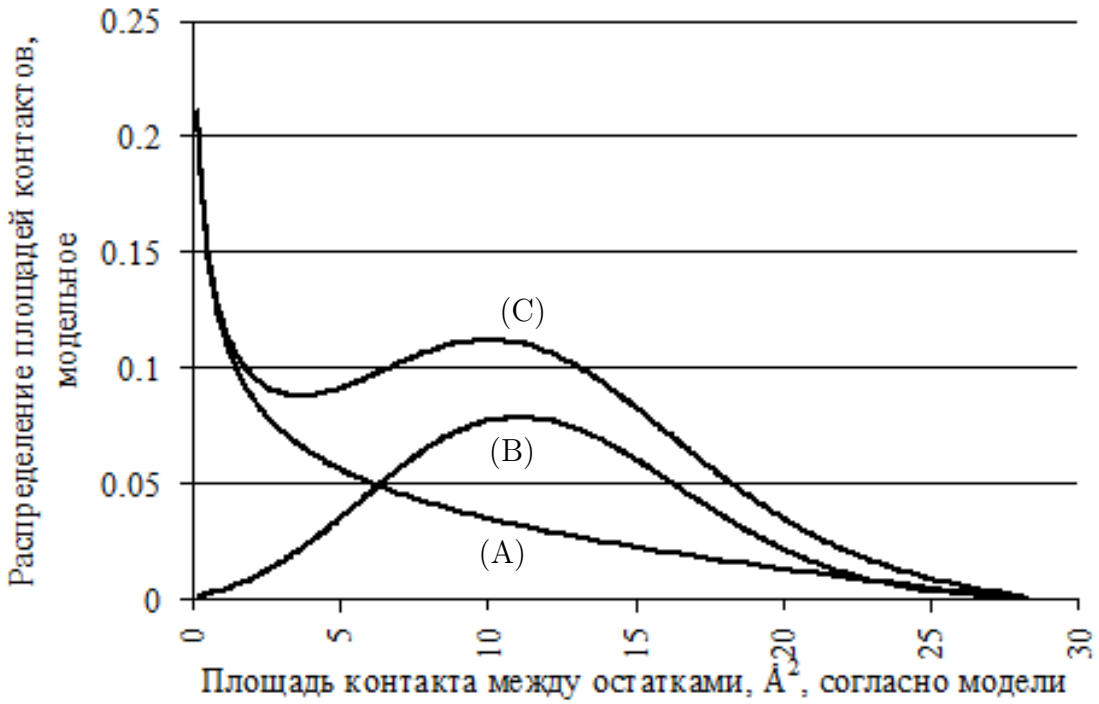


Рис. 2. Графики, отражающие системы (6) и (7), моделируют распределения площадей случайных (A) и специфических (B) контактов: A —график системы (6) в параметрической форме. График отражает распределение площади случайных контактов; B — график системы (7) в параметрической форме. График отражает распределение площади специфических контактов; C — сумма графиков A и B. Параметры, использованные в данном случае: $R = 20$, $r = 3$, $a = 3$, $\sigma = 1$

Классификация аминокислотных остатков на основе сравнительного анализа контактов в структурах комплексов белок-ДНК и специфические взаимодействия ДНК-белок

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 аминокислотных 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 в данной работе на выборке из 1937 белок-ДНК комплексов методом Вороного–Делоне, представлена в табл. 1.

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: D_1–D_4, D_7–D_{10}, D_{12} [32]. Далее анализ расстояний между соответствующими строками матрицы контактов проводили при помощи различных методов кластеризации. Иерархические методы кластеризации включали метод ближней связи, дальней связи и метод 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 [33].

Таблица 1. Количество и суммарные площади контактов между аминокислотными остатками белка и нуклеотидами ДНК в белок-нуклеиновых комплексах. Данные получены с помощью анализа пространственных координат атомов аминокислотных остатков и нуклеотидов в 1937 комплексах белок-ДНК методом разбиения Вороного–Делоне

	Количество контактов, шт.				Площади контактов, Å^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

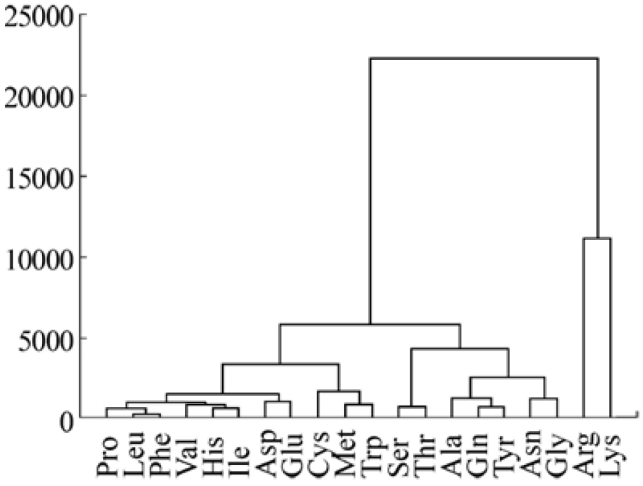


Рис. 3. Иерархическое дерево, полученное в результате применения метода средней связи и расстояния D_1 к данным табл. 1 о количестве контактов аминокислотных остатков белков с нуклеотидами ДНК, рассчитанным при помощи разбиения Вороного–Делоне

В результате применения девяти методов оценки расстояний и трех иерархических методов кластеризации было получено 27 иерархических деревьев, отражающих структуру взаимосвязи между аминокислотными остатками на основе пространственного взаимодействия аминокислотных остатков с нуклеотидами ДНК. На рис. 3 для примера приведено иерархическое дерево, полученное в результате применения кластеризации по методу средней связи и расстояния $D1$ (манхэттенское расстояние или «сити-блок»). 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 [34]. Аминокислоты этой группы (глутамин, аспарагин, аланин) часто входят в состав левой спирали типа PPII. Класс VI включает в себя аргинин и лизин, которые в некоторых случаях образуют отдельные 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 (D_2) 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 k-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.

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 [35].

1. Классифицируемое множество объектов. В нашей задаче классифицируемое множество объектов состоит из $N = 20$ типов аминокислотных остатков. Обозначим это множе-

ство 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 принимает значения от нуля до единицы, а сумма принадлежностей объекта i ко всем классам равна единице: $\sum_{k=1}^{K} h_{ik} = 1$, $0 \leqslant h_{ik} \leqslant 1$. Можно рассматривать эту матрицу как вектор-функцию размерности K от номера объекта, при этом принадлежность объекта i всем классам задается вектор-строкой $H_i = \{h_{i1}, ..., h_{iK}\}$ [36].

3. Способ описания класса. Считается, что объекты k-го класса должны хорошо описываться некоторой моделью (эталоном) этого класса [35]. В соответствии с этим вводится в рассмотрение множество возможных эталонов классов T. Между элементами множества объектов и элементами множества эталонов T вводится некоторая мера близости $S(i, t)$, $(i \in X, t \in T, S(i, t) \geqslant 0)$. 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 к соответствующему эталону класса k.

4. Критерий качества классификации. Критерий качества классификации в соответствии с методом обобщенного среднего строится следующим образом:

$$F(H, T) = \sum_{k=1}^{K} \sum_{i=1}^{N} S(i, a_k)\varphi(h_{ik})\,. \tag{8}$$

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 $\varphi(h_{ik})$ [36, 37, 38]. Выбор этой функции и ограничения, накладываемые на функцию принадлежности объекта к классу h_{ik} , определяет конкретный тип размытости классификации [36]. Для классификации с фоновым классом, фоновому классу присваивается значение $k = 0$, соответственно функция h_{i0} описывает принадлежность объекта i к фоновому классу.

Четкая классификация

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

Размытая классификация

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

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

$0 \leqslant h_{ik} \leqslant 1,\ k = 0, ..., K;\ \sum\limits_{k=0}^{K}(b - h_{ik})^2 = (K-1)b^2 + (b-1)^2$, 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

$$\begin{cases} h_{i0} = 1;\ h_{ik} = 0;\ k = 1, \ldots, K\,; \\ h_{i0} = 0;\ 0 \leqslant h_{ik} \leqslant 1;\ k = 1, \ldots, K;\ \sum\limits_{k=1}^{K}(h_{ik})^{\lambda} = 1\,. \end{cases}$$

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

$$\begin{cases} h_{i0} = 1;\ h_{ik} = 0;\ k = 1, \ldots, K\,; \\ h_{i0} = 0;\ 0 \leqslant h_{ik} \leqslant 1;\ k = 1, \ldots, K;\ \sum\limits_{k=1}^{K}(b - h_{ik})^2 = (K-1)b^2 + (b-1)^2\,, \end{cases}$$

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 $\hat{h}_i = \sum\limits_{k=1}^{K} h_{ik}$ и ограничения накладывать на $\hat{h}_i$ и h_{i0}, как на функции принадлежности для классификации на два класса:

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

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

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 [36] 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).

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 нуклео-тидами 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 [37]. 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 порог 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 ами-

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)

	Числа контактов				Площади 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

нокислоты 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 (см. 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 аминокислот насчитывается более десяти, всего же на данный момент в базе данных AAindex содержится 544 различных свойств для каждого типа аминокислотного остатка [39].

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.

Таблица 3. Результаты размытой и четкой классификации количества и площади контактов аминокислот белков с нуклеотидами ДНК при заданном числе классов 4 и коэффициенте размытости $\lambda = 2$

	Числа контактов								Площади 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

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». 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. На основе классических методов кластер-анализа (иерархических, типа k-средних, и других) и с применением различных мер близости построены классификации аминокис-

Таблица 4. Результаты размытой и четкой классификации количества и площади контактов аминокислот белков с нуклеотидами ДНК при заданном числе классов 6 и коэффициенте размытости $\lambda = 2$

	Числа контактов												Площади контактов											
	Размытая кластеризация						Четкая кластеризация						Размытая кластеризация						Четкая кластеризация					
№ класса	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

лотных остатков и 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 то же 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. P. 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. P. 932–946.

[3] *Jordan S. R., Pabo C. O.* Structure of the lambda complex at 2.5 A resolution: details of the repressor-operator interactions // *Science*, 1988. Vol. 242, No. 4880. P. 893–899.

[4] *Brennan R. G., Roderick S. L., Takeda Y., Matthews B. W.* Protein-DNA conformational changes in the crystal structure of a lambda Cro-operator complex // *Proc. Natl. Acad. Sci. U.S.A.*, 1990. Vol. 87, No. 20. P. 8165–8169.

[5] *Schultz S. C., Shields G. C., Steitz T. A.* Crystal structure of a CAP-DNA complex: The DNA is bent by 90 degrees // *Science*, 1991. Vol. 253, No. 5023. P. 1001–1007.

[6] *Bohm H. J.* The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure // *J. Comput. Aided Mol. Des.*, 1994. Vol. 8, No. 3. P. 243–256.

[7] *Aqvist J., Fothergill M.* Computer simulation of the triosephosphate isomerase catalyzed reaction // *J. Biol. Chem.*, 1996. Vol. 271, No. 17. P. 10010–10016.

[8] *Eldridge M. D., Murray C. W., Auton T. R., Paolini G. V., Mee R. P.* Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes // *J. Comput. Aided Mol. Des.*, 1997. Vol. 11, No. 5. P. 425–445.

[9] *Cozzini P., Fornabaio M., Marabotti A., Abraham D. J., Kellogg G. E., Mozzarelli A.* Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water // *J. Med. Chem.*, 2002. Vol. 45, No. 12. P. 2469–2483.

[10] *Lesser D. R., Kurpiewski M. R., Jen-Jacobson L.* The energetic basis of specificity in the Eco RI endonuclease–DNA interaction // *Science*, 1990. Vol. 250, No. 4982. P. 776–786.

[11] *Draper D. E.* Protein–DNA complexes: The cost of recognition // *Proc. Natl. Acad. Sci. U.S.A.*, 1993. Vol. 90, No. 16. P. 7429–7430.

[12] *Mandel-Gutfreund Y., Margalit H.* Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites // *Nucleic Acids Res.*, 1998. Vol. 26, No. 10. P. 2306–2312.

[13] *Mandel-Gutfreund Y., Schueler O., Margalit H.* Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: In search of common principles // *J. Mol. Biol.*, 1995. Vol. 253, No. 2. P. 370–382.

[14] *Choo Y., Klug A.* Physical basis of a protein-DNA recognition code // *Curr. Opin. Struct. Biol.*, 1997. Vol. 7, No. 1. P. 117–125.

[15] *Jones S., van Heyningen P., Berman H. M., Thornton J. M.* Protein-DNA interactions: A structural analysis // *J. Mol. Biol.*, 1999. Vol. 287, No. 5. P. 877–896.

[16] *Oda M., Nakamura H.* Thermodynamic and kinetic analyses for understanding sequence-specific DNA recognition // *Genes Cells*, 2000. Vol. 5, No. 5. P. 319–326.

[17] *Pabo C. O., Nekludova L.* Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? // *J. Mol. Biol.*, 2000. Vol. 301, No. 3. P. 597–624.

[18] *Benos P. V., Lapedes A. S., Stormo G. D.* Is there a code for protein–DNA recognition? Probab(ilistical)ly // *Bioessays*, 2002. Vol. 24, No. 5. P. 466–475.

[19] *Luscombe N. M., Thornton J. M.* Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity // *J. Mol. Biol.*, 2002. Vol. 320, No. 5. P. 991–1009.

[20] *Benos P. V., Lapedes A. S., Stormo G. D.* Probabilistic code for DNA recognition by proteins of the EGR family // *J. Mol. Biol.*, 2002. Vol. 323, No. 4. P. 701–727.

[21] *Gorfe A. A., Jelesarov I.* Energetics of sequence-specific protein–DNA association: Computational analysis of integrase Tn916 binding to its target DNA // *Biochemistry*, 2003. Vol. 42, No. 40. P. 11568–11576.

[22] *Venkatarajan M. S., Braun W.* New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties // *J. Mol. Model.*, 2001. Vol. 7, No. 12. P. 445–453.

[23] *Shen B., Bai J., Vihinen M.* Physicochemical feature-based classification of amino acid mutations // *Protein Eng. Des. Sel.*, 2008. Vol. 21, No. 1. P. 37–44.

[24] *Kosiol C., Goldman N., Buttimore N. H.* A new criterion and method for amino acid classification // *J. Theor. Biol.*, 2004. Vol. 228, No. 1. P. 97–106.

[25] *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. P. 459–463.

[26] *May A. C.* Towards more meaningful hierarchical classification of amino acid scoring matrices // *Protein Eng.*, 1999. Vol. 12, No. 9. P. 707–712.

[27] *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.

[28] *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. P. 111–122.

[29] *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. P. 1060–1077.

[30] *Anashkina A. A., Tumanyan V. G., Kuznetsov E. N., Galkin A. V., Esipova N. G.* Geometrical analysis of protein-DNA interactions on the basis of the Voronoi-Delaune tessellation // *Biofizika*, 2008. Vol. 53, No. 3. P. 402–406. (in Russ.)

[31] *Medvedev N. N.* Voronoi-Delaunay Method in Noncrystal Systems Investigations. Novosibirsk: SO RAS, 2000. (in Russ.)

[32] *Raushenbakh G. V.* Proximity and similarity measures // in Non-numerical information analysis in social science. M.: Nauka, 1985. P. 169–203. (in Russ.)

[33] *Mirkin B. G.* Cluster Analysis for Decision Making: Review. M.: HSE, 2011. (in Russ.)

[34] *Gunasekaran K., Ramakrishnan C., Balaram P.* Disallowed Ramachandran conformations of amino acid residues in protein structures // *J. Mol. Biol.*, 1996. Vol. 264, No. 1. P. 191–198.

[35] *Diday E.* Data analysis methods. (Trans.from fr. Diday E. et collaborateurs. Optimisation en classification automatique. Paris: Institut national de rechercher en informatique et en automatique, 1979) Moscow: Finansy i Statistika, 1985. 357 p. (in Russ.)

[36] *Bauman E. V., Bludjan N. O.* Metody nahozhdenija global'nyh jekstremumov funkcionalov v zadache klassifikacionnogo analiza dannyh // *Trudy Instituta problem upravlenija RAN*, 2001. Vol. XIII. P. 129–136. (in Russ.)

[37] *Zadeh L. A.* Fuzzy sets as a basis for a theory of possibility // *Fuzzy Sets Systems*, 1978. Vol. 1. P. 3–28.

[38] *Bezdek J. C.* A convergence theorem for the fuzzy ISODATA clusters algorithms // *IEEE Trans. Pattern Analysis Machine Intelligence*, 1980. P. 1–8.

[39] *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. Database issue. P. D202–D205.