© 2023 г. А.И. МИХАЛЬСКИЙ, д-р биол. наук, канд. техн. наук (ipuran@yandex.ru),
Ж.А. НОВОСЕЛЬЦЕВА, канд. техн. наук (novoselc.janna@yandex.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва),
А.А. АНАШКИНА, канд. физ.-мат. наук (a_anastasya@inbox.ru)
(Институт молекулярной биологии им. В.А. Энгельгардта РАН, Москва),
А.Н. НЕКРАСОВ, канд. физ.-мат. наук (a_nnekrasov@mail.ru)
(Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН, Москва)

ВЕРОЯТНОСТНАЯ ОЦЕНКА ВЛИЯНИЯ СОСТАВА ПЕНТАПЕПТИДА НА ЕГО УСТОЙЧИВОСТЬ

Изучается влияние расположения аминокислотных остатков в пентапептиде на его устойчивость. Строится прогноз устойчивости пентапептида с помощью метода градиентного бустинга, позволяющего оценить влияние каждого признака на стабильность пентапептида. Выявлены комбинации расположения аминокислот в пентапептиде, вносящие существенный вклад в его стабильность. Показано, что использование таких комбинаций позволяет сократить количество данных, необходимых для получения достоверного прогноза стабильности пентапептида.

Ключевые слова: аминокислотный остаток, пентапептид, градиентный бустинг, предсказание, достаточность информации.

DOI: 10.31857/S0005231023120048, **EDN:** NFXRAG

1. Введение

Проблема предсказания пространственной структуры белков является одной из приоритетных задач в области математико-биологического моделирования, ведущей к практическому применению – конструированию новых белков с полезными медицинскими свойствами. На текущий момент существует инструмент для предсказания третичной структуры белка по его аминокислотной последовательности AlphaFold [1], показавший невероятную точность предсказания структуры, сравнимую с точностью рентгеноструктурного анализа на CASP [2]. Однако этот инструмент создан на основе глубокой нейронной сети и принципы, обеспечивающие укладку, остаются неизученными. Понимание того, какие аминокислоты и в какой комбинации способствуют повышению устойчивости фрагмента белка, позволит создать метод проектирования структуры белка. Цель работы заключается в том, чтобы на основе экспериментальных данных об устойчивости пентапептидов выделить потенциальные маркеры устойчивости (комбинации и позиции аминокислот в молекуле).

Изучение энтропийных характеристик фрагментов последовательностей белков показало, что для пяти последовательно расположенных остатков наблюдается пониженный уровень информационной энтропии и, следовательно, блоки именно такого размера необходимо рассматривать как элементарные единицы последовательности. Это приближение позволило разработать метод, выявляющий иерархическую структуру в последовательностях белков, – метод анализа информационной структуры (метод АНИС) [3]. Анализ конформационной стабильности пентапептидов методом молекулярной динамики показал, что все пентапептиды можно условно разделить на три типа [4]: конформационно-стабильные (находящиеся в преимущественной топологии более 80% времени моделирования), триггерные (имеющие две преимущественные топологии, в каждой из которых пептид находился не менее 40% времени моделирования) и лабильные. Молекулярная динамика – это метод, в котором временная эволюция системы взаимодействующих атомов или частиц отслеживается интегрированием их уравнений движения. Для описания атомов или частиц и их движения применяется классическая механика. Закон движения частиц находят при помощи аналитической механики, а силы межатомного взаимодействия представляются в форме классических потенциальных сил (как градиент потенциальной энергии системы).

2. Данные

В работе использованы 44 860 пентапептидов, последовательности которых созданы по определенному правилу, и 4885 ранее изученных пентапептидов из реальных белков, устойчивость которых определялась методом молекулярно-динамического моделирования. В полученном наборе из 49 745 пентпептидов лишь 1705 пентапептидов оказались устойчивы, что составило около 3,43% от общего числа.

При исследовании все данные были разделены случайным образом на «обучающую», «контрольную» и «валидационную» выборки в пропорциях 0,66, 0,17, 0,17 с сохранением исходного баланса классов. Обучающая и контрольная выборки использовались на этапе обучения. Обучающая выборка использовалась и на этапе интерпретации результатов

2.1. Кодировки данных

В исходном наборе данных каждый пентапептид закодирован последовательностью из пяти букв, означающих аминокислотные остатки, входящие в пентапептид. Порядок следования букв соответствует последовательности аминокислотных остатков в молекуле пентапептида. Для формального численного анализа данных пятибуквенное представление кодировалось с помощью трех различных представлений. Рассматривались бинарная кодировка (One Hot Encoding), непрерывное строковое представление (n-грамма), разрывное строковое представление (разрывная n-грамма). Каждый из рассмотренных способов кодирования позволяет по-своему оценивать вклады и стро-

ить суждения о влиянии тех или иных сочетаний аминокислотных остатков на стабильность пентапептида.

2.2. Бинарная кодировка (ONE)

One hot encoding – кодировка, при которой наличие каждой аминокислоты на своей позиции задается положением единицы в векторе, остальные координаты которого равны нулю. Число элементов вектора равно 20 – числу типов аминокислот. В результате каждый пентапептид кодируется матрицей из 20 строк и 5 столбцов. Столбец соответствует позиции аминокислоты в молекуле пентапептида, а строка – аминокислоте. Например, при классификации аминокислот по первой букве названия пентапептид DKLNV будет закодирован матрицей, в которой в первом столбце в третьей строке стоит 1, остальные элементы равны нулю, во втором столбце в девятой строке стоит 1, остальные элементы равны нулю и т.д. При вычислениях каждый пентапептид представляется вектором в 100-мерном пространстве.

2.3. Непрерывное строковое представление (n-грамма)

n-грамма – непрерывное строковое представление последовательности аминокислот в пентапетиде. В зависимости от числа букв, входящих в строку, различают n-граммы порядка 1, 2 и более. Например, пептид DKLNV кодируется пятью n-граммами порядка 1 D, K, L, N, V, четырьмя n-граммами порядка 2 DK, KL, LN, NV, тремя n-граммами порядка 3 DKL, KLN, LNV. В проведенном анализе использовались n-граммы от 1 до 3. Как и при кодировке ONE вся совокупность n-грамм, кодирующих пентапептиды, представляется в виде таблицы, состоящей из нулей и единиц. В каждом столбце таблицы на определенной строке стоит единица, а остальные элементы – нули.

2.4. Разрывное строковое представление (рваная n-грамма)

Рваная n-грамма является обобщением непрерывной n-граммы и является строковым представлением последовательности аминокислот в пентапептиде, при котором между группами аминокислот есть разрыв от одного до трех символов. При формировании рваной n-граммы указываются аминокислоты, входящие в n-грамму, указывается позиция первой аминокислоты из n-граммы в молекуле пентапептида, число позиций между каждой из аминокислот, входящих в n-грамму. Например, для пентапептида DKLNV существуют рваная n-грамма второго порядка 12DN, где 1 – позиция первой аминокислоты, 2 – число позиций между аминокислотами, DN – перечень аминокислот, входящих в рваную n-грамму. Для этого пентапептида существует всего шесть рваных n-грамм порядка 2, а именно 11DL, 21KN, 31LV, 12DN, 22KV, 13DV. В исследовании рассматривались рваные n-граммы только порядка 2.

3. Алгоритм классификации

Для классификации пентапептидов на устойчивые и неустойчивые в работе использован алгоритм градиентного бустинга над решающими деревьями (gradient boosted decision trees) [5]. Алгоритм построен согласно принципу, по которому относительно слабый алгоритм машинного обучения можно усилить тем же алгоритмом, который будет «уточнять» предсказания предыдущего алгоритма, основываясь на его ошибках. При применении этого принципа для классификации методом случайного леса первый ряд деревьев обучается на реальных данных, предсказывая метку класса для каждого объекта. Второй ряд деревьев обучается на тех же данных, но придавая большее значение объектам, на которых были совершены ошибки деревьями первого ряда, и исправляя их. Деревья третьего ряда обучаются, исправляя ошибки деревьев второго ряда и т.д. В настоящее время градиентный бустинг над решающими деревьями является одним из самых популярных алгоритмов машинного обучения, потому что при малых затратах на обучение обеспечивает высокую точность, защиту от переобучения за счет того, что используется случайный лес из решающих деревьев. При этом признаки и подвыборка перемешиваются для построения нового дерева. Кроме того, полученный результат легко интерпретируется.

Контроль качества обучения проводился с использованием метрики F_1, задаваемой формулой

$$F_1 = 2\frac{precision * recall}{precision + recall}.$$

При этом один класс рассматривается как класс «положительных объектов», например класс устойчивых пентапептидов, а другой – класс «отрицательных объектов». Метрика *precision* определяет долю правильно опознанных положительных объектов среди всех объектов, отнесенных к положительным. Метрика *recall* определяет долю правильно опознанных положительных объектов среди всех положительных объектов. Метрика F_1 применяется для оценки качества классификации в случае данных, в которых классы существенно не сбалансированы.

Настройка параметров алгоритма классификации проводилась для каждого использованного метода кодировки с помощью процедуры кроссвалидации в пространстве высокой размерности [6] с помощью пакета hyperopt. В табл. 1 приведены результаты классификации, достигнутые при найденных параметрах настройки.

Таблица 1. Результаты классификации стабильности пентапептидов при различных способах кодировки

Кодировка	Метрика		
	precision	*recall*	F_1
OHE	0,39	0,54	0,45
n-грамма	0,39	0,41	0,40
разрывная n-грамма	0.32	0.54	0.40

Наилучшее качество по метрикам F_1 достигается при использовании кодировки OHE. Для кодировок n-грамма и рваная n-грамма качество ниже. Это объясняется малостью выборки и, характерным для кодировки дискретных признаков с помощью n-грамм, большим числом признаков.

4. Вероятностная оценка значимости положения аминокислот в пентапептиде

Кроме оценки качества классификации, большой интерес представляет оценка важности отдельных признаков в стабильности пентапептидов. Для построения такой оценки при использовании градиентного бустинга в настоящем исследовании применялся алгоритм SHAP (SHapley Additive exPlanations) [7], который позволяет оценить вероятностный вклад каждого сочетания аминокислот в вероятность классификации пентапептида как стабильного, учитывая при этом взаимодействие факторов (аминокислот и их положения) между собой. Этот метод вычисляет важность конкретного признака путем сравнения результатов, полученных с учетом этого признака и без его учета. При построении правила классификации в виде дерева на результат может влиять порядок, в котором используются элементы обучающей выборки. Чтобы устранить такое влияние на оценку важности признака, элементы обучающей выборки поступают на обучение многократно в случайной последовательности.

Метод SHAP получил свое обоснование в теории кооперативных игр, когда участники игры могут объединяться в коалиции для достижения наилучшего результата. Выигрыш каждого игрока равен его среднему по всем коалициям вкладу в общий выигрыш при случайном равновероятном упорядочивании участников. Эта величина называется индексом Шепли [7] и вычисляется путем суммирования по всем наборам признаков, не включающим признак i, взвешенного эффекта от использования исключенного признака. Под эффектом использования признака i в данном случае понимается разность точности классификации пентапептида с учетом признака i и без его учета. Индекс Шепли вычисляется по формуле

$$\Phi_i = \sum_{S \in F \setminus i} \frac{n_S!\,(n_F - n_S - 1)!}{n_F!} \left(f_{S \bigcup i} - f_S\right),$$

здесь F обозначает множество всевозможных наборов признаков, $F \setminus i$ обозначает множество наборов признаков, не включающих признак i, S – набор признаков без признака i, $S \bigcup i$ – набор признаков S с добавлением признака i, f_S и $f_{S \bigcup i}$ – точность классификации при использовании наборов признаков S и $S \bigcup i$ соответственно, n_F и n_S – число наборов признаков в множествах F и S соответственно. Значимость признака определяется абсолютной величиной соответствующего ему индекса Шепли.

5. Интерпретация результатов

Ниже приводятся результаты интерпретации с помощью метода SHAP результатов классификации устойчивости пентапептидов алгоритмом градиентного бустинга при использовании трех различных кодировок.

5.1. Бинарная кодировка (ONE)

В табл. 2 представлен пример оценки влияния положения аминокислот в пентапептиде DRNAA на его стабильность. Важно отметить, что на устойчивость пентапептида влияет не только наличие аминокислоты в какой-либо позиции, но и ее отсутствие.

Таблица 2. Вероятностный вклад аминокислот и их позиций на стабильность пентапептида DRNA

Аминокислота		позиция	Вероятностный вклад
наличие	отсутствие		
D		1	0,048
R		2	0,018
	A	1	0,010
A		4	0,0040
A		5	–0,0083
N		3	–0,0096

В табл. 2 строки упорядочены по мере уменьшения вероятностного вклада аминокислот и их позиций на стабильность пентапептида. Отрицательные значения означают негативное влияние на стабильность. Из таблицы следует, что наличие на первой позиции аминокислоты D на пятой позиции повышает вероятность того, что пентапептид стабилен, а отсутствие аминокислоты A на первой позиции повышает вероятность стабильности пентапептида только на 1%. Наличие же на последней позиции аминокислоты A на 0,8% понижает вероятность стабильности пентапептида. При этом предполагается, что признаки влияют на стабильность пентапептида независимо друг от друга.

Если провести подобный вероятностный анализ для множества пентапептидов, то совокупный результат можно представить в виде диаграммы вероятностных вкладов аминокислот и их положений в стабильность. На рис. 1 представлена диаграмма для наиболее значимых признаков. В силу больших вычислительных трудностей, связанных с необходимостью решения задачи классификации для всевозможных наборов признаков, вычисления проводились для 1000 случайно выбранных пентапептидов. На диаграмме отдельная точка соответствует результату анализа отдельного пентапептида.

Наличие признака (присутствие аминокислоты на указанном месте) изображается открытым символом, а отсутствие – закрытым.

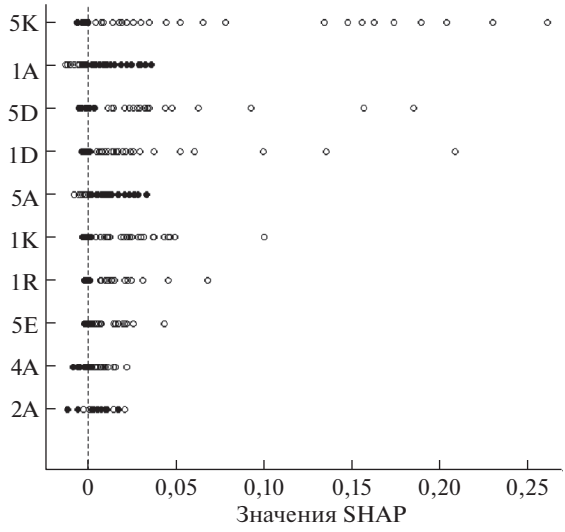

Рис. 1. Диаграмма вероятностных вкладов признаков в стабильность 1000 случайно выбранных пентапептидов при кодировке OHE, построенная с помощью алгоритма SHAP.

Из рисунка видно, что при наличии в пентапептиде аминокислоты K на пятой позиции оказывает самое большое положительное влияние на его стабильность. Обратный эффект – наиболее сильное отрицательное влияние на стабильность оказывает аминокислота A на первой позиции.

5.2. Непрерывное строковое представление

При кодировке с использованием n-грамм величина оценки вероятностного вклада в стабильность отдельного признака оказывается меньше, чем при кодировке OHE. Это является следствием того, что при использовании n-грамм до третьего порядка число признаков в 256 раз больше, чем при OHE кодировке. В табл. 3 приведены примеры оценок вероятностного вклада в стабильность пептида DRNAA.

Таблица 3. Примеры оценки вероятностного вклада в стабильность пентапептида DRNAA при кодировке с помощью n-граммы

Сочетание аминокислот		позиция	Вероятностный вклад
наличие	отсутствие		
D		1	0,0030
R		2	0,0022
	K	2	–0,00004
	EK	1	–0,00008
	T	5	–0,000028
	R	5	–0,000029
A		5	–0,0001

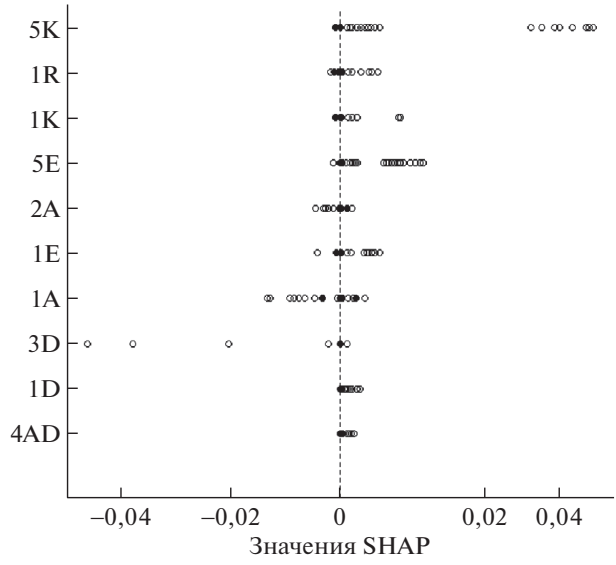

Рис. 2. Диаграмма вероятностных вкладов признаков в стабильность 1000 случайно выбранных пентапептидов при кодировке с помощью n-грамм, построенная по алгоритму SHAP.

В табл. 3 строки упорядочены по мере уменьшения вероятностного вклада аминокислот и их позиций на стабильность пентапептида. Из таблицы видно, что при кодировке с помощью n-грамм совместный вклад аминокислот D и R, находящихся в первой и второй позициях, в стабильность пентапептида оценивается около 0,5%, тогда как при кодировке OHE оценка составляет 6%.

На рис. 2 приведен пример диаграммы вероятностных вкладов аминокислот и их положений в стабильность 1000 случайно выбранных пентапептидов. Из рисунка видно, что наибольшую значимость имеют единичные комбинации аминокислот, самым мощным положительным эффектом по оценке обладает аминокислота K на пятой позиции, а отрицательным – аминокислота A на первой позиции.

5.3. Разрывное строковое представление

В табл. 4 представлен результат оценивания вероятностного вклада в стабильность отдельного признака на примере пентапептида DRNAA при кодировке разрывной n-граммой. На рис. 3 приведен пример диаграммы вероятностных вкладов аминокислот и их положений в стабильность 1000 случайно выбранных пентапептидов при той же кодировке.

В табл. 4 строки упорядочены по мере уменьшения вероятностного вклада сочетания аминокислот и их позиций на стабильность пентапептида. Из таблицы следует, что наибольший эффект на стабильность пентапептида DRNAA оказывает сочетание аминокислот R во второй позиции и A в четвертой или в пятой позициях. Отсутствие аминокислоты A в первой позиции и

Таблица 4. Примеры оценки вероятностного вклада в стабильность пентапептида DRNAA при кодировке разрывной n-граммой

Сочетание аминокислот		позиция	Вероятностный вклад
наличие	отсутствие		
R–A		2	0,0093
R–A		2	0,0041
	A–A	1	0,0031
D–A		1	0,0020
	A–A	1	0,0013
	A–K	2	–0,0014
D–N		1	–0,0030

одновременно в четвертой или пятой позициях также повышает вероятность стабильности пентапептида DRNAA, но в меньшей мере.

Из рис. 3 видно, что наибольшую значимость для стабильности имеют комбинации с аминокислотой A на второй и K на пятой позиции. Присутствие же в пентапептиде двух аминокислот A с двумя или тремя пропусками между ними, наоборот, является признаком его нестабильности.

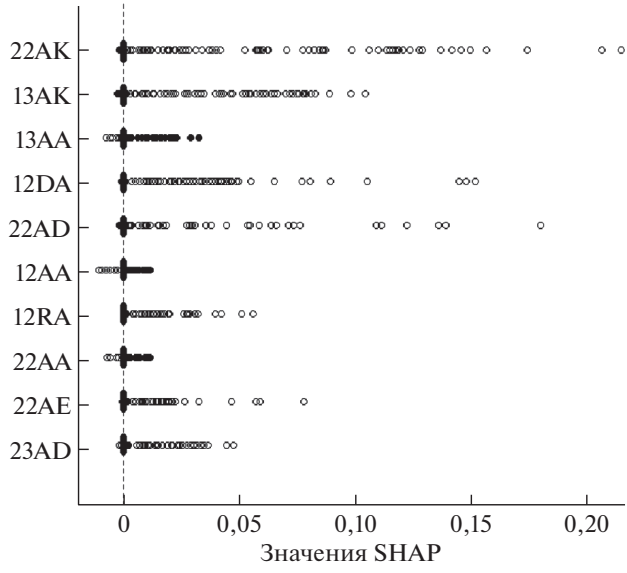

Рис. 3. Диаграмма вероятностных вкладов признаков в стабильность 1000 случайно выбранных пентапептидов при кодировке разрывной n-граммой, построенная с помощью алгоритма SHAP.

6. Заключение

В статье рассмотрен результат применения трех различных кодировок структуры пентапептида при прогнозе его стабильности через реше-

ние задачи классификации. Рассматривались бинарная кодировка (One Hot Encoding), непрерывное строковое представление (n-грамма), разрывное строковое представление (рваная n-грамма). Каждая из кодировок порождает пространства признаков различной размерности: 100 при бинарной кодировке OHE, 25 600 при кодировке с помощью n-грамм не выше третьего порядка и 10 400 при использовании разрывной n-граммы. При этом возникает различная степень разреженности данных. Задача классификации пентапептидов на устойчивые и неустойчивые решалась методом градиентного бустинга (LGBM). В исследовании использовался набор из 49 745 пентапептидов, среди которых устойчивыми были 3,43%. Данные были разделены случайным образом на «обучающую», «тестовую» и «валидационную» выборки в пропорциях с сохранением исходного баланса классов. После обучения результаты проверки на контрольной выборке на каждой из кодировок показали примерно одинаковую величину метрики качества F_1, равную 0,45 для бинарной кодировки и 0,40 при использовании различных n-грамм.

Оценка важности признаков для прогноза стабильности пентапептидов выделила наиболее важные признаки. Каждый из способов кодировки обладает своей особенностью. При кодировке OHE оценивается важность расположения конкретной аминокислоты на определенной позиции. Кодировка при использовании n-грамм позволяет оценить важность сочетания аминокислот на соседних позициях, а при использовании рваных n-грамм оценивается важность расположения аминокислот на удаленных друг от друга позициях. Кодировка с использованием рваных n-грамм позволяет выделять эффект влияния комбинации аминокислот, расположенных в разных позициях молекулы пентапептида.

Вопрос о структурной стабильности пентапептидов рассматривался в [8]. В этой работе при бинарной кодировке OHE применялся метод снижения размерности задачи, основанный на вычислении взаимной информации между признаком стабильности и описанием пентапептида. Выяснилось, что снижение размерности с помощью взаимной информации позволяет применять для прогноза стабильности «простой» метод классификации «К ближайших соседей». При этом качество результата в терминах метрик «точность» и «полнота» практически совпадает с результатом применения метода «случайный лес», требующего значительно больших вычислительных и временных затрат. Вероятностная оценка влияния состава пентапептида на его устойчивость в этом исследовании не проводилась. В настоящей работе акцент ставился на оценку влияния состава пентапептида и приведены результаты такой оценки для 1000 случайно выбранных пентапептидов, что связано с большими требованиями к необходимым вычислительным мощностям.

СПИСОК ЛИТЕРАТУРЫ

1. *Senior A. W., Evans R., Jumper J. et al.* Improved protein structure prediction using potentials from deep learning // Nature. 2020. V. 577. P. 706–710.

2. *Pereira J., Simpkin A.J., Hartmann M.D. et al.* High accuracy protein structure prediction in CASP14 // Proteins Structure Function and Bioinformatics. 2021. V. 89. No. 12. P. 1687–1699. https://doi.org/10.1002/prot.26171
3. *Nekrasov A.N., Kozmin Yu.P., Kozyrev S.V. et al.* Hierarchical structure of protein sequence // Int. J. Mol. Sci. 2021. V. 22. No. 15. 8339. https://doi.org/10.3390/ijms22158339
4. *Anashkina A.A., Nekrasov A.N., Alekseeva L.G. et al.* A minimum set of stable blocks for rational design of polypeptide chains // Biochimie. 2019. V. 160. P. 88–92.
5. *Ke G., Meng Q., Finley T., Wang T. et al.* A Highly Efficient Gradient Boosting Decision Tree // Proc. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach. 2017. P. 3149–3157.
6. *Bergstra J., Yamins D., Cox D.D.* Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures // Proc. of the 30th International Conference on Machine Learning (ICML). 2013. P. 115–123.
7. *Lundberg S.M., Lee S.I.* A unified approach to interpreting model predictions // Proc. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach. 2017. P. 4765–4774.
8. *Mikhalskii A.I., Petrov I.V., Tsurko V.V., Anashkina A.A. et al.* Application of mutual information estimation for prediction the structural stability of pentapeptides // Rus. J. Numer. Anal. Math. Model. 2020. V. 35. No. 5. P. 263–271.

Статья представлена к публикации членом редколлегии А.А. Галяевым.

Поступила в редакцию 31.05.2023

После доработки 12.09.2023

Принята к публикации 30.09.2023