УДК 512.6:577.2

*Ю. И. Журавлёв*[1,2], *К. В. Рудаков*[1,2], *И. Ю. Торшин*[2]

[1]Вычислительный центр им. А. А. Дородницына РАН
[2]Московский физико-технический институт (государственный университет)

# Алгебраические критерии локальной разрешимости и регулярности как инструмент исследования морфологии аминокислотных последовательностей

В рамках алгебраического подхода к проблеме синтеза корректных алгоритмов построение алгоритмов распознавания основано в том числе на фундаментальных критериях разрешимости и регулярности исследуемой задачи. В настоящей работе проведён анализ критериев локальной разрешимости и регулярности одной из задач биоинформатики — задачи распознавания вторичной структуры белка. Показано, что регулярность (и, следовательно, разрешимость) локальной формы задачи определяется тупиковыми множествами наиболее информативных мотивов заданной размерности и протяжённости. Приведены результаты экспериментов, проведённых на выборке всех известных на сегодняшний день аминокислотных последовательностей. Установлены тупиковые множества мотивов, обеспечивающие регулярность локальной формы задачи при произвольном множестве прецедентов.

**Ключевые слова:** распознавание, классификация, алгебраический подход, корректные алгоритмы, разрешимость, регулярность, локальность, аминокислотные последовательности.

## 1. Введение

Для задач синтеза корректных алгоритмов в течение нескольких последних десятилетий развит алгебраический подход, позволяющий точным образом формулировать и решать проблемы разрешимости, регулярности задач и полноты моделей алгоритмов и семейств корректирующих операций [1–7]. В настоящей работе алгебраические конструкции применяются для нового специального класса задач.

В биоинформатике имеется отдельный класс задач распознавания, связанных с обработкой символьных последовательностей. В биологии символьные последовательности используются для описания химической структуры биологических макромолекул (прежде всего, белков и нуклеиновых кислот). Данные об аминокислотных последовательностях (отражающих химическую структуру белков) и о нуклеотидных последовательностях (описывающих структуру ДНК и РНК), сочетаясь с данными о биологических и биофизических ролях соответствующих биомакромолекул, порождают целый спектр задач распознавания и классификации [8, 9].

Распознавание вторичной структуры белка на основе его аминокислотной последовательности представляет особый интерес, так как является одним из важных шагов к установлению взаимосвязи между химической и пространственной уровнями структуры белка. Задача рассматривается как перевод последовательности символов из одного алфавита в другой, а накопленный материал о третичном и вторичном уровнях структуры белка — как основа для построения непротиворечивых множеств прецедентов [10]. В работах [10–12] предложен формализм для анализа разрешимости и локальности данной задачи распознавания. Введение ключевых понятий для анализа локальной разрешимости задачи (окрестность, система масок, объект, множество мотивов, монотонность и тупиковость по системам масок и множествам мотивов) позволило провести эксперименты по установлению тупиковых множеств аминокислотных мотивов с наибольшей информативностью по отношению ко вторичной структуре белка.

При практическом применении рассматриваемого формализма следует принимать во внимание, что объем данных по первичной структуре белка (миллионы аминокислотных последовательностей) в сотни раз превышает массив имеющихся прецедентов «первичная структура — вторичная структура». Следовательно, при переносе закономерностей, установленных в ходе анализа множеств прецедентов, возникает вопрос о возможности обобщения установленных закономерностей на все имеющиеся аминокислотные последовательности. В настоящей работе данный вопрос исследуется на основе критериев разрешимости и регулярности локальной формы рассматриваемой задачи.

## 2. Критерии локальной разрешимости и регулярности на множествах мотивов

Одним из основных результатов работ [10–12] является формулировка критериев разрешимости исследуемой задачи распознавания. Используются два алфавита: алфавит A для описания первичной структуры белка («верхнего слова») и алфавит B для описания вторичной структуры («нижнего слова»). Пусть $\mathrm{A} = \{a_1, a_2, \ldots, a_{n(\mathrm{A})}\}$, $n(\mathrm{A}) = |\mathrm{A}| > 0$ и $\mathrm{B} = \{b_1, b_2, \ldots, b_{n(\mathrm{B})}\}$, $n(\mathrm{B}) = |\mathrm{B}| > 0$. Алфавит A (однобуквенные обозначения аминокислот) обычно определяется как $\mathrm{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, R, S, T, V, W, Y\}$. Алфавит B может быть определён существенно различными способами [10]; для целей настоящей работы вполне приемлем трёхбуквенный алфавит $\mathrm{B} = \{S, H, L\}$, описывающий три принципиально различных вида вторичной структуры: «стрэнды» ($S$, англ. strand), «спирали» ($H$, helix) и «петли» ($L$, loop).

Произвольное слово в алфавите A будем обозначать $V = v_1 v_2 \ldots v_{n(V)}$, в алфавите B — $W = w_1 w_2 \ldots w_{n(W)}$, $n(V)$ и $n(W)$ — длины слов.

*Критерий локальной разрешимости* с использованием отдельных масок (выражение (6″) в работе [10]) был сформулирован следующим образом:

$$\underset{\mathrm{Pr}}{\forall}(V^1, W^1), (V^2, W^2) \forall (i, j) \left( \overset{|\mathrm{M}|}{\underset{k=1}{\forall}} \hat{m}_k : \eta(i, \hat{m}_k, V^1) = \eta(j, \hat{m}_k, V^2) \right) \Rightarrow w_i^1 = w_j^2, \qquad (1)$$

$$L(\mathrm{M}) < i \leqslant \left|V^1\right| - R(\mathrm{M}), L(\mathrm{M}) < j \leqslant \left|V^2\right| - R(\mathrm{M}), i \neq j,$$

где $(V^1, W^1)$ и $(V^2, W^2)$ — произвольные элементы *множества прецедентов* Pr, $i$, $j$ — *ведущие позиции* в прецедентах, $\mathrm{M} = \left\{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_{|\mathrm{M}|}\right\}$ — *множество (система) масок*, $\hat{m}_k = \left\{\mu_1^k, \mu_2^k, \ldots, \mu_{m(k)}^k\right\}$ — $k$-я *маска* ($\mu_i^k \in \mathbb{Z}$, $\mu_1^k < \mu_2^k < \cdots < \mu_{m(k)}^k$, $k = 1 \ldots |\mathrm{M}|$), $\mu_i^k$ — $i$-я *позиция* $k$-й маски, $m(k) = |\hat{m}_k|$ — *размерность маски* $\hat{m}_k$, $\eta$ — *оператор выбора подслова*, а $L(\mathrm{M})$ и $R(\mathrm{M})$ — границы для описания краевых эффектов, $L(\mathrm{M}) = \max\left(-\min_{k=1,N} \mu_1^k, 0\right)$, $R(\mathrm{M}) = \max(\max_{k=1,N} \mu_{|m_k|}^k, 0)$. В дальнейшем предполагается выполнение указанных в (1) ограничений по $L(\mathrm{M})$ и $R(\mathrm{M})$ на значения $i$ и $j$. Будем также использовать $[\hat{m}_k]$ — *протяжённость* маски $\hat{m}_k$, $[\hat{m}_k] = \mu_{m(k)}^k - \mu_1^k + 1$. Заметим, что утверждение (1) соответствует локальной форме задачи распознавания вторичной структуры, т. е. существованию функции $f : \mathrm{A}^{|\hat{m}_\Sigma(\mathrm{M})|} \to \mathrm{B}$, где $\hat{m}_\Sigma(\mathrm{M})$ — *объединённая маска* M, $\hat{m}_\Sigma(\mathrm{M}) = \bigcup_{k=1}^{|\mathrm{M}|} \hat{m}_k$ (см. [10]).

Очевидно, что каждая маска описывает конкретный способ выбора подпоследовательности в заданной последовательности символов. Каждую маску можно рассматривать как признак. В работах [11, 12] был осуществлён переход от анализа разрешимости на множествах признаков (масок) к анализу разрешимости на множествах значений признаков (т. н. «мотивов») и от множеств прецедентов — к множествам объектов.

*Элементарными объектами* $q$ (далее — просто «объектами») назовём элементы множества $Q = \mathrm{A}^{|\hat{m}_\Sigma(\mathrm{M})|} \times \mathrm{B}$. Элементами наблюдаемых множеств объектов $Q(\mathrm{Pr}, \mathrm{M})$ являются пары $q_i^j = \left(\eta(i, \hat{m}_\Sigma(\mathrm{M}), V^j), w_i^j\right)$; каждая пара есть совокупность подслова, выбранного по

$\hat{m}_\Sigma(\mathrm{M})$ в $i$-й ведущей позиции верхнего слова ($V^j = v_1^j v_2^j \dots v_{n(V^j)}^j$) и $i$-й литеры нижнего слова ($W^j = w_1^j w_2^j \dots w_{n(W^j)}^j$) $j$-го прецедента.

*Элементарные мотивы* $\kappa$ (далее — просто «мотивы») — элементы множества

$$\mathsf{K} = \left\{ (\hat{m}, V) \,\middle|\, \hat{m} \in M, n(V) = |\hat{m}| \right\}.$$

Элементарный мотив $\kappa = (\hat{m}, V')$ *присутствует в объекте* $q = (V, w)$, если $\eta(L(\mathrm{M}) + 1, \hat{m}, V) = V'$. Обозначим принадлежность мотива объекту $q$ как $\kappa \in^* q$. Мотив $\kappa$ назовём *отличающим для* произвольной пары объектов $q_1$ и $q_2$, если $\kappa$ присутствует в одном из объектов и отсутствует во втором. Условие локальной разрешимости задачи выполнено тогда и только тогда, когда для каждой пары объектов $q_i = (V_i, w_i)$ и $q_j = (V_j, w_j)$ при $w_i \neq w_j$ во множестве мотивов $\mathsf{K}$ существует хотя бы один отличающий мотив (теорема 1 в работе [12]):

$$\underset{Q}{\forall}(i,j) : w_i \neq w_j \Rightarrow \underset{\mathsf{K}}{\exists}\kappa : (\kappa \in^* V_i) \neq (\kappa \in^* V_j). \tag{2}$$

Пусть $r_1 = N_{(2)}/N \cdot (N-1)$, где $N = |Q|$, а $N_{(2)}$ — множество пар объектов, на котором выполнено условие (2). В разрешимой задаче $Z(Q, \mathsf{K})$ всегда $r_1 = 1$.

Критерий разрешимости на множествах мотивов имеет принципиальное значение для дальнейшего развития разрабатываемого формализма и для практических приложений. Утверждение (2) соответствует переходу от задачи $Z(\mathrm{Pr}, \mathrm{M})$ [10] к эквивалентной задаче $Z(Q, \mathsf{K})$ [12], в которой в качестве параметров выступают наблюдаемое множество объектов $Q = Q(\mathrm{Pr}, \mathrm{M})$ и множество мотивов $\mathsf{K} = \mathsf{K}(Q, \mathrm{M})$, порождённое системой масок M. В частности, условие (2) позволяет сформулировать критерий локальной регулярности на множествах мотивов.

Наряду с разрешимостью в современной теории распознавания [1–7] изучается также *регулярность* задач. В общем случае под регулярностью задачи понимается разрешимость задачи, сопровождающаяся разрешимостью задач из некоторой её окрестности в изучаемом множестве задач, так что точное определение понятие регулярности определяется способом задания окрестности задачи. Следуя идеологии научной школы академика Ю. И. Журавлёва, определим окрестность задачи $Z(Q, \mathsf{K})$ со множеством объектов $Q = \{(V_1, w_1), (V_2, w_2), \dots, (V_i, w_i), \dots\}$ как множество задач $Z'$ со множеством объектов $Q\prime = \{(V_1, w_1'), (V_2, w_2'), \dots, (V_i, w_i'), \dots\}$ при произвольных $w_1', w_2', \dots, w_i', \dots$. Отсюда следует, что задача $Z$ будет регулярной на множестве объектов $Q$ тогда и только тогда, когда выполняется следующее *условие регулярности на множестве мотивов*:

$$\underset{Q}{\forall} q_i, q_j, i \neq j \Rightarrow \underset{\mathsf{K}}{\exists}\kappa : (\kappa \in^* V_i) \neq (\kappa \in^* V_j). \tag{3}$$

Если задача $Z(Q, \mathsf{K})$ — регулярна, то будем называть $Q$ регулярным множеством объектов, а $\mathsf{K}$ — *регулярным множеством мотивов*. Пусть $r_0 = N_{(3)}/N \cdot (N-1)$, где $N = |Q|$, а $N_{(3)}$ — множество пар объектов, на котором выполнено условие (3). В регулярной задаче $Z(Q, \mathsf{K})$ выполнено $r_0 = 1$.

Для исследования вопроса о возможности обобщения установленных закономерностей на все имеющиеся аминокислотные последовательности представляет практический интерес нахождение некоторых минимальных множеств мотивов, гарантирующих регулярность на произвольном множестве объектов. Регулярное множество мотивов назовём *тупиковым*, если условие (3) выполнено для $\mathsf{K}$, но не выполнено для любого $\mathsf{K}' \subset \mathsf{K}$.

**Теорема 1**. *В задаче* $Z(Q, \mathsf{K})$ *тупиковое множество мотивов* $\mathsf{K}_1$*, обеспечивающее разрешимость, является подмножеством тупикового множества мотивов* $\mathsf{K}_0$*, обеспечивающего регулярность.*

**Доказательство.** В выражениях (2) и (3) мотивы $\kappa$ необходимы для попарного различения объектов из $Q$. В общем случае чем больше пар объектов сравнивается, тем большее число мотивов необходимо для различения всех пар объектов. Рассмотрим задачу с алфавитом

из двух литер, $\mathrm{B} = \{B_1, B_2\}$. При этом, $|Q| = N = n(B_1) + n(B_2)$. При проверке условия (2), проводится сравнение $N_1 = n(B_1) \cdot n(B_2)$ пар объектов. При проверке условия (3) на том же множестве объектов $Q$ проводится сравнение $N_0 = N \cdot (N-1)$ пар объектов, включая $N_1$ пар объектов в условии (2). Очевидно, что при любых $n(B_1)$ и $n(B_2)$ $N_0 \geqslant N_1$, так что $\mathsf{K}_1 \subseteq \mathsf{K}_0$. Доказательство для произвольного числа классов проводится аналогично. Теорема доказана.

**Замечание.** В вычислительных экспериментах множества мотивов $\mathsf{K}_1$ и $\mathsf{K}_0$ могут, вообще говоря, определятся по различным выборкам, приблизительно и т. д. Для оценки соответствия этих двух множеств используются вводимые ниже параметры $\Delta_{1,0}$ и $r_{1,0}$.

**Следствие 1.** Пусть $\Delta_{1,0} = 1 - |\mathsf{K}_1 \cap \mathsf{K}_0|/|\mathsf{K}_1|$ — параметр, описывающий соответствие множества $\mathsf{K}_1$ множеству $\mathsf{K}_0$. В условиях теоремы $\Delta_{1,0} = 0$.

**Следствие 2.** Пусть $r_{1,0} = r_1(\mathsf{K}_1 \cap \mathsf{K}_0)$. В условиях теоремы $r_{1,0} = 1$.

Таким образом, множество мотивов, удовлетворяющее критерию регулярности (3), содержит в себе множество мотивов, обеспечивающее разрешимость задачи распознавания. Важно, что тестирование регулярности может проводиться без какой-либо информации о вторичной структуре белка, т. е. на таких множествах объектов, как $Q' = \{(V_1, \Delta), (V_2, \Delta), \ldots, (V_i, \Delta), \ldots\}$.

Определение тупиковых множеств мотивов, соответствующих системе масок M — задача, разрешимая полным перебором. Если $\mathrm{M}_n^m$ — система масок, образованная всеми сочетаниями $m$ позиций из $n$ возможных в соответствующей объединенной маске (т. е. $m$ — размерность каждой маски из $\mathrm{M}_n^m$, а $n$ — протяженность объединенной маски), то $|\mathrm{M}_n^m| = \mathrm{C}_n^m$. Полный перебор подмножеств множества из $|\mathsf{K}| = \mathrm{C}_n^m \cdot |A|^m$ мотивов не представляется возможным практически, поэтому необходима редукция множества мотивов. Редукция множества мотивов $\mathsf{K}(Q, \mathrm{M})$, наблюдаемого при заданных $Q$ и M, с целью нахождения тупиковых $\mathsf{K}$ может рассматриваться как частный случай выделения подкласса «наиболее информативных признаков» во множестве всех значений всех исследуемых признаков. Для этого вводятся эвристические оценки информативности мотивов.

## 3. Эвристические оценки информативности мотивов и критерий регулярности

В духе теории классификации значений признаков [13] можно сказать, что следует оставлять мотивы с «высокой информативностью» и удалять мотивы с «достаточно низкой» информативностью так, что регулярность задачи (3) не нарушена. Оценка информативности мотивов $D\colon \mathsf{K} \to \mathbb{R}_+$ может быть введена различными способами так, чтобы бóльшая «информативность» мотива соответствовала бóльшим значениям $D$.

Отметим принципиальное различие между оценками информативности, используемыми при тестрировании условий разрешимости (2) и регулярности (3). В случае разрешимости «более информативными» являются мотивы, которые (а) характеризуются наибольшей частотой встречаемости и (б) выделяют «достаточно много» объектов $l$-го класса и «достаточно мало» объектов всех остальных классов. При отборе мотивов следует учитывать оба эти фактора, так что построение функции $D$, адекватной для тестирования разрешимости, представляет собой нетривиальную задачу. Практически полезным является использование функций $D_1(\alpha) = \sum_{l=1}^m D_l^\alpha$ и $D_2(\alpha) = N_\Sigma^\alpha \sum_{l=1}^m D_l^\alpha$, $m = |\mathrm{B}|$; $D_l^\alpha$ определяется в соответствии с (4):

$$D_l^\alpha = \begin{cases} 1 - \frac{\nu_l^\alpha}{\nu_l^0} \text{ при } \nu_l^\alpha \leqslant \nu_l^0, \\ \frac{\nu_l^\alpha - \nu_l^0}{1 - \nu_l^0} \text{ при } \nu_l^\alpha > \nu_l^0, \end{cases} \tag{4}$$

где $\nu_l^\alpha = \frac{N_l^\alpha}{N_\Sigma^\alpha}$ — *частота встречаемости* значения $b_l \in \mathrm{B}$, а $\nu_l^0$ — частоты встречаемости литеры $b_l \in \mathrm{B}$ во всем множестве объектов $Q$ [12].

В случае условия регулярности (3), которое не включает информации о нижних словах объектов, необходимым условием наибольшей информативности мотива является только частота его встречаемости. Поэтому оценка информативности мотива $\kappa_\alpha \in \mathsf{K}(Q, \mathrm{M})$, который входит в состав $N_\Sigma^\alpha$ объектов из $Q$, может быть определена просто как $D_{reg}(\alpha) = \frac{N_\Sigma^\alpha}{|Q|}$.

**Замечание.** Функцию $D_1(\alpha)$ можно использовать для оценки степени непротиворечивости нерегулярных множеств объектов (т. е. фактически реальных наборов экспериментальных данных). В данном случае объект рассматривается как мотив, построенный по одноэлементной системе масок $\mathrm{M}_n^n$, где $n$ — протяженность объекта. Очевидно, что *объект непротиворечив* тогда и только тогда, когда $D_1(\alpha) = |\mathrm{B}|$. Пусть $Q_{\text{нп}} = \{q_\alpha \subset Q \,\big|\, D_1(\alpha) = |\mathrm{B}|\}$ – подмножество непротиворечивых объектов, тогда *долей непротиворечивых объектов* во множестве $Q$ назовём отношение $|Q_{\text{нп}}|/|Q|$.

Эвристические оценки информативности мотивов необходимы, прежде всего, для нахождения тупиковых множеств мотивов на основе критериев разрешимости регулярности. Функция $D : \mathsf{K} \to \mathbb{R}_+$ ставит в соответствие каждому мотиву множества $\mathsf{K}$ его информативность из определенного подмножества $\mathbb{R}_+$. Отношение порядка на $\mathbb{R}_+$ порождает линейно *упорядоченный список* $I(\mathsf{K})$ на множестве мотивов $\mathsf{K}$. Введение линейного порядка на множестве мотивов позволяет использовать данные об информативности мотивов при тестировании условия (3). Принцип отбора мотивов состоит в том, что для каждой пары объектов из $Q$ находится различающий мотив с наивысшей информативностью. Отобранные таким образом мотивы образуют некоторое множество различающих мотивов $\mathsf{K}_0$ с наивысшей информативностью, такое что $\mathsf{K}_0 \subseteq \mathsf{K}(\mathrm{Pr}, \mathrm{M})$.

Перенумеруем все элементы $\mathsf{K} = \mathsf{K}(Q, \mathrm{M})$ так, чтобы порядок мотивов в линейно упорядоченном списке $I(\mathsf{K})$ соответствовал убыванию значений $D$: $\kappa_1, \kappa_2, \ldots, \kappa_\alpha, \ldots, \kappa_{|\mathsf{K}|}$, $D(\kappa_\alpha) \geqslant$ $\geqslant D(\kappa_{\alpha+1})$. Пусть на исходном множестве мотивов $\mathsf{K}$ выполнено условие регулярности (3). Определим функцию $K_f(i, j)$, находящую единственный мотив с максимальным $D$ (и, следовательно, с минимальным номером мотива $\alpha$), который позволит различить $i$-й и $j$-й объекты из $Q$:

$$K_f(i, j) = \min_{1 \ldots |K|} \alpha \,:\, (\kappa_\alpha \in^* V_i) \neq (\kappa_\alpha \in^* V_j). \tag{5}$$

Тогда минимальное множество мотивов $\mathsf{K}_0$, на котором сохраняется регулярность, определяется *характеристической функцией* $T(\alpha, Q)$:

$$T(\alpha, Q) = \begin{cases} 1 \equiv \underset{Q}{\exists}(i, j) : (K_f(i, j) = \alpha), \\ 0 \text{ в противном случае.} \end{cases} \tag{6}$$

Множество мотивов $\mathsf{K}_0$, полученное при вычислении (5), будет регулярным только тогда, когда в $Q$ существует хотя бы одна пара объектов, для которой данный мотив — единственный различающий. Последнее обеспечено определением $K_f(i, j)$ — ведь в соответствии с выражением (5) выбирается единственный различающий мотив для произвольной пары объектов. $\mathsf{K}_0$ не может не быть тупиковым, когда последнее утверждение справедливо для всех мотивов. По теореме 1 такое множество мотивов будет содержать подмножество мотивов, обеспечивающее разрешимость на том же множестве объектов.

Отметим, что при вычислении тупиковых множеств мотивов по формуле (6), так что $\mathsf{K}_0$ вычисляется с использованием $D$, а $\mathsf{K}_1$ — с использованием $D'$, *выполнение* $\mathsf{K}_1 \subseteq \mathsf{K}_0$ *в условиях теоремы 1 гарантировано только при* $D = D'$. Действительно, вычисление характеристической функции $T$ зависит от линейных порядков в списках $I(\mathsf{K}_0)$ и $I(\mathsf{K}_1)$. При различных $D$ и $D'$ линейный порядок в списке $I(\mathsf{K}_0)$ может отличаться от порядка в списке $I(\mathsf{K}_1)$ так, что $\mathsf{K}_1 \not\subset \mathsf{K}_0$. Заметим также, что параметр $\Delta_{1,0}$ отражает длину наибольшей общей подпоследовательности в списках $I(\mathsf{K}_1)$ и $I(\mathsf{K}_0)$.

В основе разрабатываемого формализма лежат два принципиальных допущения, анализ которых представляет собой отдельные направления дальнейших исследований: а) использование определённых функций $D$, эвристических оценок информативности мотивов и

б) произвол в выборе мотива при $D(\kappa_\alpha) = D(\kappa_{\alpha+1}) = D(\kappa_{\alpha+2}) = \ldots$. Вследствие а) и б) использование различных выборок объектов даже при фиксированной системе масок может нарушать отношение порядка на мотивах. Поэтому при экспериментальном тестировании регулярности и разрешимости проводится усреднение вычисляемых $\mathsf{K}_1$ и $\mathsf{K}_0$ по различным выборкам объектов одного размера. При этом контролируется значение определённого ранее параметра $\Delta_{1,0}$.

Определим $z_\alpha$ — *заполненность* элементарного мотива $\kappa_\alpha$ при тестировании $n$ выборок объектов как $z_\alpha = \sum_{i=1\ldots n} T(\alpha, Q_i)/n$. Во множестве мотивов $\mathsf{K}$ (это может быть множество $\mathsf{K}_1$ или $\mathsf{K}_0$), усреднённом по $n$ выборкам $Q$, информативным назовём мотив $\kappa_\alpha$ с заполненностью $z_\alpha \geqslant z_{\min}$. Очевидно, что при заданной $D$ наиболее информативны мотивы с $z_{\min} = 1$. Так как при снижении значения параметра $z_{\min}$ ($z_{\min} = 0{,}9, 0{,}8$ и т. д.) в $\mathsf{K}$ войдёт большее число различающих мотивов, то параметры разрешимости $r_1$ и регулярности $r_0$ увеличатся.

Размер выборки объектов $|Q|$ является важным параметром, определяющим значения $z_\alpha$ конкретных мотивов при данной системе масок. Величины $|Q|$ и $\mathsf{K}(Q, \mathrm{M})$ определяют мощности множеств мотивов $\mathsf{K}_1$ и $\mathsf{K}_0$ так, что выполнена

**Теорема 2.** *При фиксированном* $\mathsf{K} = \mathsf{K}(Q, \mathrm{M})$, $|\mathsf{K}_0| = f(|\mathsf{K}|, |Q|)$, *причём* $f$ *— монотонна по* $Q$.

**Доказательство.** Произвольная функция $D$ индуцирует линейный порядок в списке $I(\mathsf{K})$. При добавлении объекта к $Q$ может потребоваться дополнительный распознающий мотив в $I(\mathsf{K})$, при этом $\mathsf{K}_0$ увеличится, так что $f$ не убывает при увеличении $|Q|$. При удалении объекта из $Q$ некоторый мотив может быть исключён, если он был различающим только для пар объектов, образованных этим объектом, так что $f$ не возрастает при уменьшении $|Q|$. Таким образом, $f$ монотонна по $|Q|$. Теорема доказана.

**Следствие 1.** $r_0$ монотонно возрастает с увеличением $|Q|$.

**Следствие 2.** Справедливо соответствующее утверждение для разрешимости.

**Замечание.** Утверждение теоремы относится к $\mathsf{K}_0$, полученному по (6) и обеспечивающему $r_0 = 1{,}0$. Монотонность может нарушаться в подмножествах $\{z_\alpha \in \mathsf{K}_0 \,|\, z_\alpha = 1\}$, полученных при усреднении по $n$ выборкам $Q$, в которых $r_0 \leqslant 1{,}0$ вследствие исключения редко встречающихся мотивов.

Согласно теореме 2, можно ожидать возрастания $|\mathsf{K}_0|$ и $r_0$ ($|\mathsf{K}_1|$ и $r_1$) при увеличении числа объектов в тестируемых выборках. Исследование динамики роста значений $|\mathsf{K}_0|$ и $r_0$ в ходе экспериментов представляет интерес с точки зрения нахождения тупиковых множеств наиболее информативных мотивов на произвольной выборке объектов.

## 4. Экспериментальное тестирование условий разрешимости и регулярности

Выражения (3–6) позволяют вычислять тупиковые множества мотивов для данных $Q$ и M. Эксперименты по тестированию разрешимости (2) и регулярности (3) проводились на общедоступных экспериментальных данных по первичной, вторичной и третичной структурам белков (PDB, Protein Data Bank), суммарно включающих более 150 000 последовательностей и структур [14]. Использовались различные протяжённости объектов с $n = 4 \ldots 16$; доля непротиворечивых объектов при $n \geqslant 6$ составила не менее 0,87 (рис. 1).

Регулярность (3) также тестировалась на множествах объектов, полученных на основе всех известных аминоксилотных последовательностей в базе данных UNIPROT [15], в которой присутствует 15 млн попарно различных последовательностей общей длиной в $5 \cdot 10^9$ литер. Из данных БД UNIPROT были сформированы выборки объектов длиной в 4, 6, 8 и 10 литер (таблица 1), при этом объекты с частотой встречаемости менее $10^{-7}$ были исключены.

В качестве оценок информативности мотивов использовались $D_{reg}(\alpha)$ и $D_2(\alpha)$.

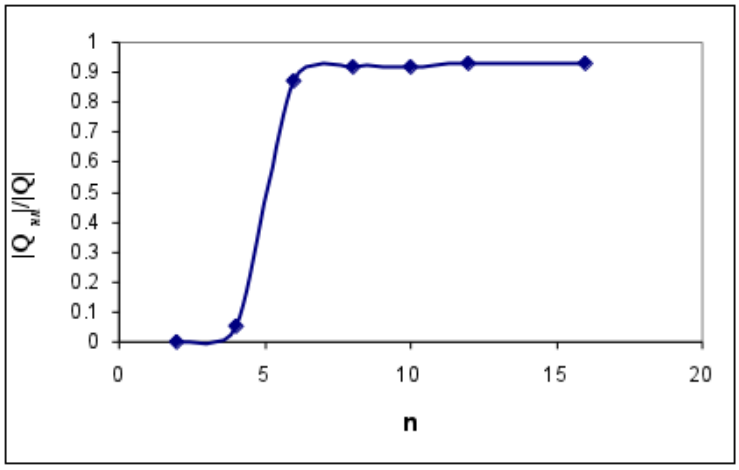


Рис. 1. Доля непротиворечивых объектов при различных длинах объекта (система масок $\mathrm{M}_n^n$)

Таблица 1

**Объекты, полученные на основе данных БД UNIPROT**

| Мн-во объектов $Q$ | Длина объекта | $\lvert Q\rvert$ | Покрытие $Q$ выборки UNIPROT |
|---|---|---|---|
| $Q4$ | 4 | $1,60 \cdot 10^5$ | 100% |
| $Q6$ | 6 | $2,07 \cdot 10^7$ | 57% |
| $Q8$ | 8 | $2,33 \cdot 10^7$ | 20% |
| $Q10$ | 10 | $1,92 \cdot 10^7$ | 18% |

Каждая из *использованных систем масок* имела фиксированную размерность всех масок. Были исследованы системы масок с размерностью всех масок равной $m = 2$ (системы $\mathrm{M}_n^2$) и $m = 3$ ($\mathrm{M}_n^3$), полученные полным перебором по $m$ позиций из $n = 6, 8, 10, 12, 16$.

Далее, последовательно рассматриваются частоты встречаемости мотивов в последовательностях из PDB и UNIPROT, результаты тестирования разрешимости и регулярности $Z(Q, \mathsf{K})$ с использованием $D_{reg}(\alpha)$ и $D_2(\alpha)$, результаты исследования выполнения условия регулярности на $\mathsf{K}_0$ (UNIPROT) при множествах мотивов $\{z_\alpha = 1\}$ и, наконец, морфология аминокислотных последовательностей с учётом данных $\mathsf{K}_0$ (UNIPROT) при $\{z_\alpha = 1\}$.

*Сравнение частот встречаемости мотивов в последовательностях из UNIPROT и PDB* даёт общее представление о различиях в морфологии последовательностей, для которых были получены трёхмерные структуры (кристаллизованные белки, структура которых определена дифракционными методами), и всех остальных последовательностей (15 млн). Значимость в различиях частот ($p$) оценивалась кумулятивной функцией гипергеометрического распределения. Анализ показал, что «значимые» ($p < 0{,}05$) различия наблюдались для большинства мотивов. Например, для системы масок $\mathrm{M}_8^3$ $p < 0{,}05$ наблюдалось для 104 000 мотивов из 176 500. При той же системе масок, однако, отношение частот более чем 2,0 наблюдалось всего для 1790 мотивов, причём эти мотивы содержали заряженные и гидрофильные аминокислоты. Мотивы, содержащие т. н. «гистидиновые тэги» (HHS, EHH, HHH, MHH и др., литера «H» означает заряженную аминокислоту гистидин) и триптофан (WMC, WCP, WCW, MCW, WMW, WWC и др., литера «W» — триптофан), встречались в 10 раз и более чаще среди кристаллизованных белков. Известно, что триптофан, заряженные и гидрофильные аминокислоты способствуют взаимодействию молекул белков, что облегчает формирование кристалла. Схожие результаты (значимые различия в частотах $H$- и $W$-содержащих мотивов) были получены и при анализе мотивов с другими системами масок.

*Тестирование разрешимости и регулярности* $Z(Q, \mathsf{K})$ проводилось на непротиворечи-

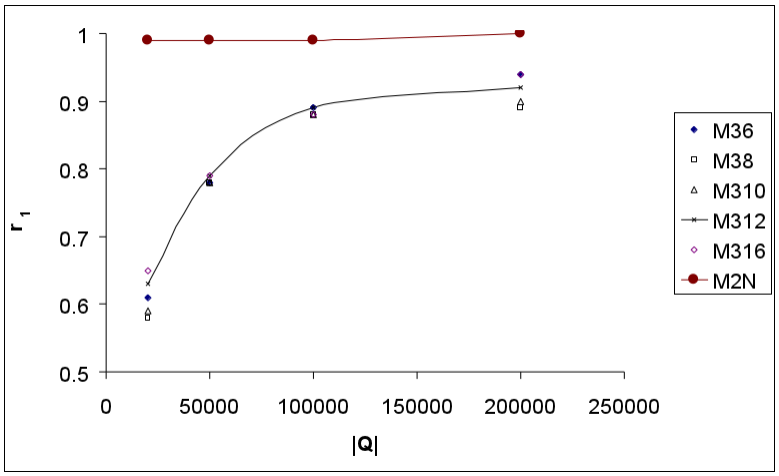


Рис. 2. Зависимость разрешимости ($r_1$) от размера выборки объектов (объекты PDB, подмножество $\{z_\alpha \in \mathsf{K}_1 \,|\, z_\alpha = 1\}$). Оценка информативности $D_{reg}(\alpha)$, M36 — система масок $\mathrm{M}_6^3$ и т. д. Разброс значений $r_1$ в каждой исследованной точке, обусловленный тестированием на различных $Q$ одного размера, не превышал 0,02. Кривые для систем масок $\mathrm{M}_n^2$, $n = 6 \ldots 16$, полностью совпадают

вых выборках объектов из PDB [14] с использованием $D_{reg}(\alpha)$ и $D_2(\alpha)$ в качестве эвристических оценок информативности. Были исследованы выборки размером $2 \cdot 10^4$, $5 \cdot 10^4$, $10^5$ и $2 \cdot 10^5$ объектов, сформированные путём случайного отбора объектов без возвращения, по 10 выборок для каждого из приведенных выше значений $|Q|$. Для каждого размера выборки вычислялись множества мотивов $\mathsf{K}_1$ (подмножество $\{z_\alpha = 1\}$) и $\mathsf{K}_0(\{z_\alpha = 1\})$; рассчитывались значения показателей $r_1, r_0, r_{1,0}$ и $\Delta_{1,0}$. Затем сравнивались множества мотивов, полученные с использованием различных оценок информативности.

Как видно из рис. 2, наибольшие различия в значениях $r_1$ для масок $\mathrm{M}_n^3$ наблюдались при малых размерках выборок ($2 \cdot 10^4$ объектов); наилучший результат показала система масок $\mathrm{M}_6^3$ ($r_1 = 0{,}94 \pm 0{,}01$; $|Q| = 2 \cdot 10^5$). Для данной системы масок $r_1 \geqslant 0{,}99$ достигалось при $z_{\min} = 0{,}7$. Для системы масок $\mathrm{M}_n^2$ практически полная разрешимость ($r_1 \geqslant 0{,}99$) достигалась при любых значениях $n = 6 \ldots 16$ даже на малых выборках объектов (20000–50000), причём пересечение множеств $\{\kappa_\alpha \in \mathsf{K}_1(D_{reg}(\alpha)) \,\big|\, z_{\alpha=1}\}$ и $\{\kappa_\alpha \in \mathsf{K}_1(D_2(\alpha)) \,\big|\, z_{\alpha=1}\}$ обеспечивало $r_1 \geqslant 0{,}98$. Сравнение результатов для $D_{reg}(\alpha)$ и $D_2(\alpha)$ в системе масок $\mathrm{M}_6^3$ показало, что множества мотивов $\{\kappa_\alpha \in \mathsf{K}_1(D_{reg}(\alpha)) \,\big|\, z_\alpha = z_{\min}\}$ и $\{\kappa_\alpha \in \mathsf{K}_1(D_2(\alpha)) \,\big|\, z_\alpha = z_{\min}\}$ содержат общее подмножество, обеспечивающее $r_1 \geqslant 0{,}99$ при $z_{\min} = 0{,}7$ и $|Q| = 2 \cdot 10^5$.

Результаты тестирования *выполнимости условия регулярности* на выборках из PDB и UNIPROT с использованием $D_{reg}(\alpha)$ показало, что параметр $\Delta_{1,0}$ имел наименьшее значение ($\Delta_{1,0} = 0{,}005 \pm 0{,}003$) в системах масок $\mathrm{M}_6^3$ и $\mathrm{M}_6^2$. При этом множества вида $\{\kappa_\alpha \in (\mathsf{K}_1 \cap \mathsf{K}_0) \,\big|\, z\alpha = 1\}$ обеспечивали различение 0,9995 пар объектов по критерию регулярности (3). В системе масок $\mathrm{M}_6^3$ значение $r_0$ достигало 0,99 при $|Q| = 7 \cdot 10^5$, в то время как $r_{1,0}$ показало более медленный рост (рис. 3). Замедленный рост $r_{1,0}$ связан с тем, что мотивы, содержащиеся в структурах кристаллизованных белков, встречаются намного реже в произвольной выборке аминокислотных последовательностей.

**Морфология аминокислотных последовательностей.** Множества мотивов, получаемые в результате тестирования (3, 6), характеризуют морфологию или в некотором смысле «структуру» аминокислотных последовательностей. В соответствие с (5), для каждой пары из $i$-го и $j$-го объектов множества $Q$ функция $K_f(i, j)$ находит наиболее информативный различающий мотив. Для всех таких мотивов $T(\alpha) = 1$, т. е. эти мотивы образуют $\mathsf{K}_0$. После вычисления $T(\alpha)$ для всех пар объектов из $Q$ каждому $i$-му объекту из $Q$ соответствует $n_i^{rm}$ различающих мотивов из $\mathsf{K}_0$, $n_i^{rm} = |\{T(\alpha) = 1\}_i|$. Более чем для 90% объектов эти мотивы покрывают не все позиции объекта, выделяя тем самым некоторые «информативные» позиции аминокислотной последовательности, соответствующие «информативным»

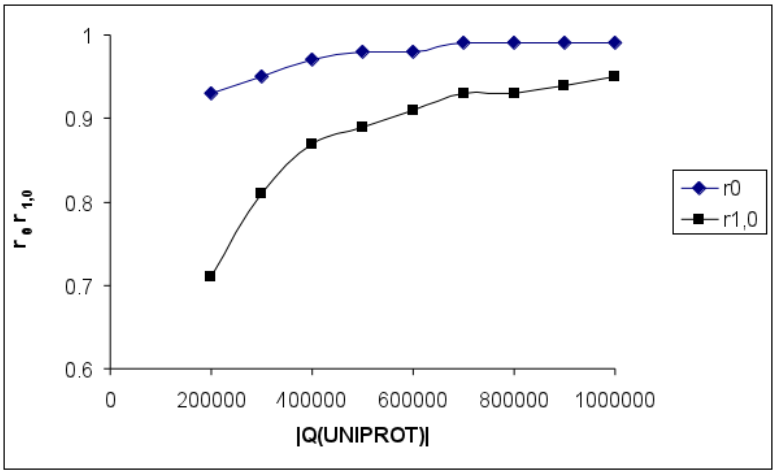


Рис. 3. Исследование зависимости параметров разрешимости/регулярности от размера выборки объектов (эксперименты на выборке Q6 БД UNIPROT, система масок $M_6^3$)

| | | | | |
|---|---|---|---|---|
| а) | RKSGnS L | SEKfRE L | qAGfhD L | EmVnDA H |
| | ARVhEy H | VATTGE H | DPVhKA H | hrSySc S |
| б) | RKSGNS<br>212403 | SEKFRE<br>243032 | QAGFHD<br>022002 | EMVNDA<br>203023 |
| | ARVHEY<br>323020 | VATTGE<br>554554 | DPVHKA<br>223034 | HRSYSC<br>001010 |

Рис. 4. Примеры структур объектов с учётом позиций, выбранных по тупиковому множеству мотивов $\mathsf{K}_0$. а) Информативные позиции выделены заглавными А-литерами, указаны литеры вторичной структуры объектов. б) Численная оценка информативности соответствующей позиции показана под каждой литерой

мотивам. Отметим, что использование мотивов $\mathsf{K}_0$ также позволяет провести численную оценку «информативности» каждой из позиций объекта (рис. 4).

## 5. Заключение

Показано, что регулярность (и, следовательно, разрешимость) локальной формы задачи гарантирована тупиковыми множествами наиболее информативных мотивов заданной размерности и протяжённости. Приведены результаты экспериментов, проведённых на выборке всех известных на сегодняшний день аминокислотных последовательностей. Установлены тупиковые множества мотивов, обеспечивающие регулярность локальной формы задачи при произвольном множестве прецедентов. Показано, что тупиковые множества мотивов могут быть использованы для выделения информативных позиций в аминокислотной последовательности. Установление тупиковых множеств мотивов и информативных позиций в первичной структуре необходимо не только для синтеза корректных алгоритмов распознавания вторичной структуры белка, но и для решения ряда других задач биоинформатики, связанных с анализом символьных последовательностей.

Работа выполнена при поддержке грантов РФФИ 09–07–12098, 09–07–00212–а и 09–07–00211–а и контракта Минобрнауки РФ № 07.514.11.4001.

## Литература


1. *Журавлев Ю. И.* Теоретико-множественные методы в алгебре логики. // Проблемы кибернетики — 1962. — Т. 8(1). — 25–45.
2. *Журавлев Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. I // Кибернетика. — 1977. — № 4. — С. 5–17.
3. *Журавлев Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. II // Кибернетика. — 1977. — № 6. — С. 21–27.
4. *Журавлев Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. III // Кибернетика. — 1978. — № 2. — С. 35–43.
5. *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — Вып. 33. — М.: Наука, 1978. — С. 5–68.
6. *Журавлев Ю. И., Рудаков К. В.* Об алгебраической коррекции процедур обработки (преобразования) информации // Проблемы прикладной математики и информатики. — М.: Наука, 1987. — С. 187–198.
7. *Рудаков К. В.* Универсальные и локальные ограничения в проблеме коррекции эвристических алгоритмов // Кибернетика. — 1987. — № 2. — С. 30–35.
8. *Torshin I. Y.* Bioinformatics in the Post-Genomic Era: The Role of Biophysics // NY: Nova Biomedical Books, 2006. — ISBN: 1–60021–048.
9. *Torshin I. Yu.* Sensing the change from molecular genetics to personalized medicine // "Bioinformatics in the Post-Genomic Era" series. — Nova Biomedical Books, NY, USA, 2009. — ISBN 1-60692-217-0.
10. *Рудаков К. В., Торшин И. Ю.* Вопросы разрешимости задачи распознавания вторичной структуры белка // Информатика и её применения. — Т. 4, № 2. — 2010. — С. 25–35.
11. *Рудаков К. В., Торшин И. Ю.* О разрешимости формальной задачи распознавания вторичной структуры белка // ММРО–14, Суздаль, 21–25 сентября, 2009. — С. 596–597.
12. *Рудаков К. В., Торшин И. Ю.* Анализ информативности мотивов на основе критерия разрешимости в задаче распознавания вторичной структуры белка // Информатика и её применения. — 2011. — № 4.
13. *Рудаков К. В.* О проблемах классификации значений признаков в задачах распознавания // Международная конференция «Интеллектуализация обработки информации» (ИОИ–8), Кипр, г. Пафос, 17–23 октября 2010 г.
14. *Berman H. M., Henrick K., Nakamura H.* Announcing the worldwide Protein Data Bank // Nature Structural Biology. — 2003. — V. 10, N 12. — P. 980–982.
15. The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource // Nucleic Acids Res. 39: D214-D219. — 2011.


*Поступила в редакцию 20.09.2011*