УДК 501.1

Классификация и распознавание структур генетических последовательностей

В. А. Твердохлебов, Д. А Карякин

Твердохлебов Владимир Александрович, доктор технических наук, профессор, главный научный сотрудник, Институт проблем точной механики и управления РАН, Россия, 410028, г. Саратов, ул. Рабочая, д. 24, tverdokhlebovva@list.ru

Карякин Денис Алексеевич, аспирант, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, Россия, 410012, г. Саратов, ул. Астраханская, д. 83, dakariakin@gmail.com

Для решения проблемы определения связей свойств организмов со свойствами соответствующих им генетических последовательностей в статье рассматривается метод классификации последовательностей и распознавание принадлежности исследуемой последовательности конкретному классу. Впервые предлагается классификация последовательностей на основе числовых показателей рекуррентных и Z-рекуррентных форм, определяющих структуры функциональных связей элементов последовательностей. Для числовых показателей рекуррентных и Z-рекуррентных форм вводится классификация, которая распространяется на классификацию генетических последовательностей. Каждому рассматриваемому в задаче распознавания классу последовательностей, имеющему содержательную интерпретацию в приложениях, сопоставляется числовая характеристика, обобщающая числовые показатели рекуррентной или Z-рекуррентной формы, определяющих структуру последовательностей класса. При распознавании полученная числовая характеристика класса сравнивается с числовой характеристикой рекуррентной или Z-рекуррентной формы, соответствующей исследуемой генетической последовательности. При классификации последовательностей на основе числовых показателей рекуррентной и Z-рекуррентной форм, определяющих структуры функциональных связей элементов в последовательностях, причинно-следственные связи в генетических последовательностях, заменяются формальными функциональными зависимостями между элементами последовательностей. Задача распознавания рассматривается в двух формах: в форме принадлежности последовательности заданному конкретному классу последовательностей и в форме определения, какому из заданных классов последовательностей принадлежит исследуемая последовательность. Основные математические трудности при решении указанных задач распознавания связаны с определением рекуррентных и Z-рекуррентных форм, по числовым показателям которых исследуемая последовательность и классы последовательностей различаются. Для преодоления этих трудностей разработан спектр числовых показателей рекуррентных и Z-рекуррентных форм, с использованием которого рекуррентно и Z-рекуррентно определены последовательности. Классификация и распознавание иллюстрируются примером, в котором рассматриваются три класса генетических кодов организмов, каждый из которых представлен пятью генетическими последовательностями. Для уточнения и расширения классификации последовательностей и повышения эффективности методов распознавания вводится Z-рекуррентное определение последовательностей.

Ключевые слова: последовательность, генетическая последовательность, рекуррентное определение последовательности, Z-рекуррентное определение последовательности, рекуррентная форма, Z-рекуррентная форма, классификация последовательностей, распознавание последовательностей.

Поступила в редакцию: 12.04.2018 / Принята: 22.02.2019 / Опубликована: 31.08.2019

DOI: https://doi.org/10.18500/1816-9791-2019-19-3-338-350

ВВЕДЕНИЕ

При анализе структур и функций генетического материала в решениях проблем наследственности и патологий живых организмов, а также в профилактике и лечении наследственных патологий явно или неявно используются классификации и задачи распознавания генетических последовательностей. Существенное применение имеет математический аппарат.

В данной статье предлагается новый подход построения классификации на основе числовых показателей структур генетических последовательностей и излагаются методы распознавания генетических последовательностей по числовым показателям их структур. Предлагаемые классификация и методы распознавания структур основываются на различных вариантах рекуррентных и Z-рекуррентных определений последовательностей, интерпретируемых как генетические последовательности [1,2]. Примеры использования рекуррентных определений последовательности в задачах распознавания содержатся в работе [3].

Первоначальными характеристиками генетических последовательностей являются порядки (числовые показатели) рекуррентных форм в рекуррентном определении последовательностей. Числовые показатели рекуррентных определений последовательностей систематизированы в 5-уровневый спектр. На каждом из уровней спектра генетическая последовательность определяется числовой структурой (числом, набором чисел, набором наборов чисел), соответствующей взаиморасположению нуклеотидов в последовательности. С использованием числовых показателей спектра определяются формальные классы генетических последовательностей, которые в ряде случаев могут совмещаться с классами генетических кодов, определяемыми свойствами, имеющими интерпретацию в генетике. Определение классов генетических последовательностей по числовым показателям рекуррентных определений является компактным, классы определяются с использованием простых вычислительных процедур и при решении задач определения класса, которому принадлежит исследуемая генетическая последовательность, применяется простая вычислительная процедура. Практическая эффективность разработанных числовых моделей генетических последовательностей и методов распознавания таких последовательностей по их числовым показателям зависит от меры совпадения формальных классов и представляемых ими классов содержательно определенных в генетике классов. Впервые разработанные основные положения, модели и методы, изложенные в данной статье, апробированы только на некоторых примерах, которые следует рассматривать с точки зрения их логической непротиворечивости и принципиальной возможности использования.

Предлагаемые для исследования свойств генетических последовательностей модели и методы следует рассматривать как разработку формального аппарата, позво-

ляющего строить числовые характеристики генетических последовательностей и множеств генетических последовательностей, а также разрабатывать достаточно простые и логически понятные алгоритмы вычисления формальных показателей, соответствующих генетическим последовательностям и множествам генетических последовательностей.

Спектр числовых показателей рекуррентных определений последовательностей изложен в работах В. А. Твердохлебова [1, 2] и параграфах 1, 2 данной статьи. Алгоритмы и программы, позволяющие определять числовые показатели по трем уровням спектра, разработаны Д. А. Карякиным. Для иллюстрации моделей и методов получения числовых показателей Д. А. Карякиным проведен вычислительный эксперимент, результаты которого изложены в параграфах 3, 4. Анализ полученных результатов написан В. А. Твердохлебовым и Д. А. Карякиным.

1. ЧИСЛОВЫЕ ПОКАЗАТЕЛИ РЕКУРРЕНТНЫХ ОПРЕДЕЛЕНИЙ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Генетические последовательности рассматриваются как последовательности знаков без их содержательной интерпретации. Свойства генетических последовательностей представляются числовыми показателями структур последовательностей в форме функциональных зависимостей в формальных представлениях последовательностей. Для этого рассматриваются варианты рекуррентных определений последовательностей и показатели используемых рекуррентных форм. Каждое рекуррентное определение последовательности определяет класс последовательностей, в который входят последовательности с взаимно однозначными переобозначениями (с взаимно однозначным преобразованием) элементов. На основе этого предположения формальный класс последовательностей, соответствующий множеству генетических последовательностей, имеет характеристику точно одной числовой структуры, состоящей из целого положительного числа, или набора целых положительных чисел, или набора наборов целых положительных чисел.

Спектр Ω числовых показателей рекуррентных определений последовательностей, предложенный в работах [1, 2], определяется числовыми показателями на пяти уровнях Ω_0, Ω_1, Ω_2, Ω_3, Ω_4. Характеристика последовательности на каждом из следующих уровней спектра является более точной относительно предшествующих уровней.

В спектре числовыми значениями представлены порядки рекуррентных форм, длины отрезков последовательности, определяемые отдельными рекуррентными формами, и количество смен рекуррентных форм.

По определению $\Omega_0(\xi) = m_0(\xi)$, где $m_0(\xi)$ — наименьший порядок рекуррентной формы, определяющей всю последовательность ξ. На уровне $\Omega_1(\xi)$ спектра $\Omega(\xi)$ расположено m_0 чисел ($m_0 \in \mathbb{N}^+$), определяющих для порядков от 1 до m_0 размеры наибольших определяемых начальных отрезков последовательности ξ.

Уровень $\Omega_2(\xi)$ содержит m_0 чисел, показывающих, сколько раз для рассматриваемого порядка рекуррентных форм потребовалось заменять рекуррентные формы при определении последовательности ξ. На уровне $\Omega_3(\xi)$ каждое число смен рекуррентных форм, показанное на уровне $\Omega_2(\xi)$, заменено последовательностью чисел, представляющих длины отрезков, определяемых отдельными рекуррентными формами.

По построению спектр динамических показателей определения последовательности состоит из числовых значений:

– наименьшего порядка $m_0(\xi)$ рекуррентной формы, определяющей всю последовательность ξ;

– набора наименьших длин $d^1(\xi), d^2(\xi), \ldots, d^{m_0}(\xi)$, префиксов последовательности ξ, задаваемых рекуррентными формами соответственно порядков $1, 2, \ldots, m_0$;

– набора чисел $r^1(\xi), r^2(\xi), \ldots, r^{m_0}(\xi)$, смен рекуррентных форм порядков $1, 2, \ldots, m_0$, задающих всю последовательность;

– набора наборов длин

$$d_1^1(\xi), d_2^1(\xi), \ldots, d^1_{r^1(\xi)+1}(\xi);$$

$$d_1^2(\xi), d_2^2(\xi), \ldots, d^2_{r^2(\xi)+1}(\xi);$$

$$\ldots\ldots\ldots\ldots\ldots$$

$$d_1^{m_0}(\xi) = |\xi|$$

отрезков последовательности ξ, где $d_j^m(\xi)$ — длина j-го отрезка в определении рекуррентной формой порядка m последовательности ξ.

Для любой последовательности $\overline{\xi} \in U^v$ наименьший порядок рекуррентной формы, определяющей последовательность $\overline{\xi}$, будем обозначать $m_0(\overline{\xi})$. Для любой последовательности $\overline{\xi} \in U^v$ и $m \in \mathbb{N}^+$, где $1 \leqslant m \leqslant m_0(\overline{\xi})$, наибольшую длину начального отрезка последовательности $\overline{\xi}$, определяемого рекуррентной формой порядка m, будем обозначать $d^m(\overline{\xi})$. Для любой последовательности $\overline{\xi} \in U^v$ и $m \in \mathbb{N}^+$, где $1 \leqslant m \leqslant |\overline{\xi}| - 1$, число смен рекуррентных форм порядка m, требующихся при определении последовательности $\overline{\xi}$, будем обозначать $r^m(\overline{\xi})$. Для любой последовательности $\overline{\xi} \in U^v$ и $m \in \mathbb{N}^+$, где $1 \leqslant m \leqslant m_0(\overline{\xi})$, и j, где $1 \leqslant j \leqslant r^m(\overline{\xi})$, длину j-го отрезка в определении последовательности $\overline{\xi}$ будем обозначать $d_j^m(\overline{\xi})$.

Используя введенные обозначения, определим спектр параметров, характеризующих последовательность, как следующую структуру:

$\Omega_0(\overline{\xi}) = \langle m_0(\overline{\xi})\rangle$;
$\Omega_1(\overline{\xi}) = \langle d^1(\overline{\xi}), d^2(\overline{\xi}), \ldots, d^\alpha(\overline{\xi})\rangle$;
$\Omega_2(\overline{\xi}) = \langle r^1(\overline{\xi}), r^2(\overline{\xi}), \ldots, r^\alpha(\overline{\xi})\rangle$;
$\Omega_3(\overline{\xi}) = \langle \Omega_3^1(\overline{\xi}), \Omega_3^2(\overline{\xi}), \ldots, \Omega_3^\alpha(\overline{\xi})\rangle$;
$\Omega_4(\overline{\xi}) = \Theta(\Omega_3(\overline{\xi}))$,

где $\alpha = m_0(\overline{\xi})$ и $\Omega_3^j(\overline{\xi}) = \langle d_1^j(\overline{\xi}), d_2^j(\overline{\xi}), \ldots, d_{n_j}^j(\overline{\xi})\rangle$ (n_j — номер последнего отрезка в определении последовательности $\overline{\xi}$ как последовательности отрезков, определяемых отдельными рекуррентными формами порядка j), Θ — оператор замены в $\Omega_3(\overline{\xi})$ величин длин отрезков весами использованных рекуррентных форм для определения отрезков.

Четвертый уровень $\Omega_4(\overline{\xi})$ спектра $\Omega(\overline{\xi})$ к характеристике последовательности $\overline{\xi}$ по количеству изменений правил, определяющих взаиморасположение элементов в последовательности, и величинам областей действия правил, представленной на уровнях $\Omega_1(\overline{\xi}) - \Omega_3(\overline{\xi})$, добавляет оценки сложности правил и величины области использования правил. В достаточно общем случае можно вводить веса правил (рекуррентных форм) и веса реализации правил, используемых при определении отрезка. Например, для каждого шага применения рекуррентной формы $F(z_1^0, z_2^0, \ldots, z_m^0) = z_{m+1}^0$, т. е. для набора $(z_1^0, z_2^0, \ldots, z_m^0)$, задается вес $\Theta(z_1^0, z_2^0, \ldots, z_m^0)$

в числовой форме и сумма весов всех шагов применения рекуррентной формы для последовательности полагается весом последовательности.

Первые четыре уровня $\Omega_0(\xi), \Omega_1(\xi), \Omega_2(\xi)$ и $\Omega_3(\xi)$ спектра $\Omega(\xi)$ характеризуют алгоритмические свойства определения последовательности ξ и ее строение, так как рекуррентные формы являются правилами построения порядка следования элементов. Эти отдельные, базовые, правила сменяют одно другое по общему критерию достижения границы применимости рекуррентной формы.

Расширим спектр Ω до спектра Ω', где Ω' содержит уровни $\Omega_0, \Omega_1, \Omega_2, \Omega_3, \Omega_4$ и уровни $\Omega_5(\xi) = \Omega_0(\xi^{-1})$, $\Omega_6(\xi) = \Omega_1(\xi^{-1})$, $\Omega_7(\xi) = \Omega_2(\xi^{-1})$, $\Omega_8(\xi) = \Omega_3(\xi^{-1})$, $\Omega_9(\xi) = \Omega_4(\xi^{-1})$. Проведенный анализ показал, что для использования в представлении неравенства последовательностей $\xi_1 \neq \xi_2$ числовых показателей последовательностей могут выполняться следующие отношения: $\xi_1 \neq \xi_2$, $\Omega_i(\xi_1) = \Omega_i(\xi_2)$ и $\Omega_i(\xi_1^{-1}) \neq \Omega_i(\xi_2^{-1})$. Для определения числовых показателей последовательностей по уровням спектра существуют простые алгоритмы.

2. Z-РЕКУРРЕНТНОЕ ОПРЕДЕЛЕНИЕ ГЕНЕТИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Классическое рекуррентное определение последовательностей, устанавливающее функциональную связь каждого $(m+1)$-го элемента последовательности от предшествующих m элементов, является простейшим и допускает принципиальное усовершенствование.

Z-рекуррентное определение последовательностей предполагает использование двух процедур: процедуру покрытия последовательностей набором элементов и процедуру установления в каждом наборе покрытия функциональных связей.

Как процедура покрытия, так и определение структуры функциональных связей в каждом наборе из покрытия представляют большие возможности для интерпретации в области приложений причинно-следственных связей элементов в генетических последовательностях.

Z-рекуррентные определения последовательностей, рассматриваемых как формальные структуры, обладают свойствами, которые можно представить в следующих утверждениях.

Теорема 1. *Для последовательности длины C, $C \in \mathbb{N}^+$, и величин $\alpha + \beta \leqslant C$ существует D вариантов возможных Z-рекуррентных форм рангов вида $(i_1, i_2, \ldots, i_\alpha, j_1, j_2, \ldots, j_\beta)$ для Z-рекуррентного определения, где*

$$D = \prod_{t=1}^{\alpha+\beta-1} (C - (\alpha + \beta - t)).$$

Теорема 2. *Если для двух последовательностей ξ_1, ξ_2 длины $C, C \in \mathbb{N}^+$, выполняется условие $\xi_1 \neq \xi_2$, то во множестве вариантов Z-рекуррентных форм существует рекурретная форма, Z-рекуррентно определяющая только одну из последовательностей.*

Генетические последовательности живых организмов определяются последовательностями элементов из множества знаков $M = \{A, T, G, C\}$.

Ряд важных свойств живых организмов определяются на основе анализа таких последовательностей, включая рассмотрение структур генетической последовательности, т. е. с использованием взаиморасположения элементов в последовательности.

В данной статье предлагаются модели и метод определения числовых показателей, характеризующих варианты рекуррентного определения последовательностей, соответствующих генетическим последовательностям.

С использованием так полученных числовых показателей рекуррентных определений последовательностей ставятся и решаются следующие задачи:

– задача 1 классификации генетических последовательностей по значениям числовых показателей рекуррентных определений последовательностей;

– задача 2 определения принадлежности рассматриваемой генетической последовательности к классу последовательностей по значениям числовых показателей рекуррентного определения рассматриваемой последовательности;

– задача 3 разработки числовых показателей рекуррентного определения последовательностей, представляющих организм, специфических, т.е. разграничивающих классы организмов.

Эффективность для приложений решений задач 1–3 ограничивается полнотой и точностью, с которыми содержательные свойства организмов представлены формальными структурами последовательностей.

Работы [4–14] дают характеристику области возможных приложений полученных в статье результатов для решения задач, связанных с проблемой определения связей свойств организмов со свойствами соответствующих им генетических последовательностей.

На основе исходного рекуррентного определения последовательности числовыми показателями разработан спектр Ω вариантов числовых показателей рекуррентных определений.

Спектр Ω в работах [1,2] определен как 5-уровневый спектр $\Omega = \{\Omega_0, \Omega_1, \Omega_2, \Omega_3, \Omega_4\}$.

Простейшей числовой характеристикой расположений элементов в последовательности полагается порядок рекуррентной формы, используемой для рекуррентного определения последовательности.

Числовые структуры следующих уровней спектра строятся на основе рекуррентных определений частей последовательностей. Классификация генетических последовательностей определяется на основе принадлежностей числовых показателей рекуррентных определений последовательностей выбранным интервалом изменений числовых показателей рекуррентных определений.

Для развития средств числового представления структур генетических последовательностей разработано Z-рекуррентное определение последовательностей.

Для моделей и методов, используемых в решениях задач 1–3, Д. А. Карякиным разработаны алгоритмы и программы, реализующие методы, а также проведены вычислительные эксперименты, иллюстрирующие решения задач в частных случаях.

Решения задач 1–3 иллюстрируется на примерах, в которых рассматриваются 15 генетических последовательностей имеющих следующую содержательную классификацию:

– последовательности $\xi^n_{1,1}, \xi^n_{1,2}, \ldots, \xi^n_{1,5}$, представляющие префиксы длины n генетических последовательностей, соответствующие бактериям;

– последовательности $\xi^n_{2,1}, \xi^n_{2,2}, \ldots, \xi^n_{2,5}$, представляющие префиксы длины n генетических последовательностей, соответствующие вирусам;

– последовательности $\xi^n_{3,1}, \xi^n_{3,2}, \ldots, \xi^n_{3,5}$, представляющие префиксы длины n генетических последовательностей, соответствующие животным.

С использованием алгоритмов и программ задача 1 решена в вариантах:

– для уровней $\Omega \in \{\Omega_0, \Omega_1, \Omega_2, \Omega_3\}$ определено число n_0, для которого выполняются неравенства $r_i(\xi_\nu^{n_0}) \neq r_i(\xi_j^{n_0})$, где $1 \leqslant \nu, j, \leqslant 5, \nu \neq j$ и $r_i(\xi)$ — числовой показатель, соответствующий i уровню.

Рассмотрим задачу построения числовых показателей рекуррентного определения последовательности из множеств последовательностей M_1, M_2, M_3, по которым элементы разных множеств не пересекаются. Анализ процессов распознавания последовательностей по числовым показателям их рекуррентного определения проведен в двух направлениях:

– построением с фиксированием уровня спектра и увеличением длины последовательностей;

– с переходом вычислений показателей по уровням спектра.

Для того, чтобы решать задачи распознавания генетических последовательностей по числовым показателям их рекуррентных и Z-рекуррентных определений, требуется найти эффективную рекуррентную или Z-рекуррентную форму. Общий подход предполагает первоначальное использование рекуррентных форм нулевого уровня. Если для конкретных анализируемых множеств генетических последовательностей числовые показатели нулевого уровня спектра оказываются не достаточными, то проверке на эффективность подвергаются числовые показатели следующих первого, второго, третьего уровней. При этом установлено, что имеются случаи, когда распознавание генетических последовательностей может быть осуществлено на основе применения числовых показателей рекуррентных форм для обращений исследуемых последовательностей. Более сложная методика требуется для поиска Z-рекуррентных форм, позволяющих с их использованием распознавать генетические последовательности. В случае поиска эффективной Z-рекуррентной формы можно пользоваться следующим методом.

Для исследования генетической последовательности ξ рассмотрим варианты покрытия последовательности подпоследовательностями длин $2, 3, \ldots, C$, где C — длина исследуемой последовательности. В множестве подпоследовательностей длины K, $K < C$, образующих покрытие ξ, проводится анализ с целью поиска определения наборов целых положительных чисел $\langle i_1, i_2, \ldots, i_a\rangle$ и $\langle j_1, j_2, \ldots, j_b\rangle$, удовлетворяющих условиям.

В каждой подпоследовательности с первым элементом a_{t+i_1} рассмотрим покрытия последовательности ξ. Элементы последовательности $a_{t+j_1}, a_{t+j_2}, \ldots, a_{t+j_b}$ функционально соответствуют (функционально связаны) элементам $a_{t+i_1}, a_{t+i_2}, \ldots, a_{t+i_a}$. Если для последовательности ξ имеется Z-рекуррентное определение Z-рекуррентной формой порядка $\langle i_1, i_2, \ldots, i_a, j_1, j_2, \ldots, j_b\rangle$, то этот порядок рассматривается как числовая форма характеристики, по которой определяется последовательность ξ для решения задач распознавания.

Если генетические последовательности ξ_1 и ξ_2 имеют Z-рекуррентное определение Z-рекуррентной формой одного и того же порядка $\langle i_1, i_2, \ldots, i_a, j_1, j_2, \ldots, j_b\rangle$, то поиск новой пригодной для распознавания последовательности ξ_1 и ξ_2 Z-рекуррентной формой может быть продолжен как поиск новых наборов, составляющих порядок $\langle i_1, i_2, \ldots, i_a, j_1, j_2, \ldots, j_b\rangle$, так и на основе поиска новых покрытий последовательностей.

3. РАСПОЗНАВАНИЕ ГЕНЕТИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ПО ЧИСЛОВЫМ ПОКАЗАТЕЛЯМ РЕКУРРЕНТНЫХ ОПРЕДЕЛЕНИЙ

Рассмотрим примеры распознавания генетических последовательностей по числовым показателям их рекуррентных определений. Для этого из классов генетических последовательностей, соответствующих бактериям, вирусам, животным, выберем по пять представителей длины 10 000 и построим числовые показатели их рекуррентных определений их префиксов длины 100, 1000, 5000, 10 000.

Исследованные последовательности взяты из банка генетических последовательностей NCBI Genome [15]. В связи с большими размерами рассматриваемых генетических последовательностей и их префиксов их конкретная форма в данной статье не приводится.

Для проведения вычислительного эксперимента разработаны алгоритмы и составлены программы, с использованием которых получены результаты, представленные в табл. 1–9.

Используются следующие обозначения: K_1, K_2, K_3 — классы генетических последовательностей, $K = K_1 \cup K_2 \cup K_3$ — универсум генетических последовательностей.

Таблица 1 / Table 1

Числовые показатели рекуррентного определения последовательностей по Ω_1 из K_1

Numeric indicators of recurrent definition of sequences by Ω_1 from K_1

Длина	$\xi_{1,1}$	$\xi_{1,2}$	$\xi_{1,3}$	$\xi_{1,4}$	$\xi_{1,5}$
100	8	8	8	8	15
1000	15	12	23	12	15
5000	20	49	81	23	25
10 000	64	49	81	23	41

Таблица 2 / Table 2

Числовые показатели рекуррентного определения последовательностей по Ω_1 из K_2

Numeric indicators of recurrent definition of sequences by Ω_1 from K_2

Длина	$\xi_{2,1}$	$\xi_{2,2}$	$\xi_{2,3}$	$\xi_{2,4}$	$\xi_{2,5}$
100	7	9	6	8	8
1000	12	11	10	14	16
5000	14	12	14	15	16
10 000	16	15	15	73	16

Таблица 3 / Table 3

Числовые показатели рекуррентного определения последовательностей по Ω_1 из K_3

Numeric indicators of recurrent definition of sequences by Ω_1 from K_3

Длина	$\xi_{3,1}$	$\xi_{3,2}$	$\xi_{3,3}$	$\xi_{3,4}$	$\xi_{3,5}$
100	7	7	12	14	8
1000	12	250	19	14	14
5000	20	378	19	19	17
10 000	40	378	60	25	17

Таблица 4 / Table 4

Числовые показатели рекуррентного определения последовательностей по Ω_2 из K_1

Numeric indicators of recurrent definition of sequences by Ω_2 from K_1

Длина	$\xi_{1,1}$	$\xi_{1,2}$	$\xi_{1,3}$	$\xi_{1,4}$	$\xi_{1,5}$
100	5, 7, 26, 31, 100	6, 17, 44, 100	4, 6, 11, 24, 58, 78, 100	4, 8, 20, 23, 38, 62, 100	4, 8, 14, 56, 100

Таблица 5 / Table 5

Числовые показатели рекуррентного определения последовательностей по Ω_2 из K_2

Numeric indicators of recurrent definition of sequences by Ω_2 from K_2

Длина	$\xi_{2,1}$	$\xi_{2,2}$	$\xi_{2,3}$	$\xi_{2,4}$	$\xi_{2,5}$
100	4, 9, 12, 33, 56, 100	5, 9, 16, 46, 78, 100	3, 10, 57, 100	2, 5, 14, 38, 75, 92, 100	3, 7, 17, 19, 35, 100

Таблица 6 / Table 6

Числовые показатели рекуррентного определения последовательностей по Ω_2 из K_3
Numeric indicators of recurrent definition of sequences by Ω_2 from K_3

Длина	$\xi_{3,1}$	$\xi_{3,2}$	$\xi_{3,3}$	$\xi_{3,4}$	$\xi_{3,5}$
100	5, 16, 28, 66, 100	3, 8, 11, 40, 100	3, 8, 10, 14, 23, 100	3, 5, 9, 39, 90, 100	5, 17, 58, 86, 100

Таблица 7 / Table 7

Числовые показатели рекуррентного определения последовательностей по Ω_3 из K_1
Numeric indicators of recurrent definition of sequences by Ω_3 from K_1

Длина	$\xi_{1,1}$	$\xi_{1,2}$	$\xi_{1,3}$	$\xi_{1,4}$	$\xi_{1,5}$
100	65, 61, 36, 10, 7, 3, 1, 0	67, 59, 34, 19, 8, 2, 1, 0	49, 52, 45, 29, 14, 7, 2, 0	76, 67, 32, 12, 4, 1, 0	60, 52, 31, 16, 6, 2, 1, 0

Таблица 8 / Table 8

Числовые показатели рекуррентного определения последовательностей по Ω_3 из K_2
Numeric indicators of recurrent definition of sequences by Ω_3 from K_2

Длина	$\xi_{2,1}$	$\xi_{2,2}$	$\xi_{2,3}$	$\xi_{2,4}$	$\xi_{2,5}$
100	69, 60, 36, 17, 4, 2, 0	67, 63, 38, 7, 2, 1, 0	67, 64, 35, 15, 5, 0	66, 56, 34, 19, 9, 4, 1, 0	69, 51, 31, 20, 10, 6, 2, 0

Таблица 9 / Table 9

Числовые показатели рекуррентного определения последовательностей по Ω_3 из K_3
Numeric indicators of recurrent definition of sequences by Ω_3 from K_3

Длина	$\xi_{3,1}$	$\xi_{3,2}$	$\xi_{3,3}$	$\xi_{3,4}$	$\xi_{3,5}$
100	65, 56, 37, 17, 9, 3, 0	62, 53, 41, 21, 7, 2, 0	49, 48, 47, 34, 21, 10, 8, 3, 1, 0	62, 47, 32, 15, 10, 8, 4, 1, 0	53, 55, 38, 25, 11, 5, 1, 0

4. АНАЛИЗ РЕЗУЛЬТАТОВ ВЫЧИСЛИТЕЛЬНОГО ЭКСПЕРИМЕНТА

Проведем анализ результатов вычислительного эксперимента с классами K_1, K_2, K_3, представленных в табл. 1–9. Числовые показатели рекуррентного определения префиксов длины 100 представлены первыми строчками в табл. 1–3. В этом случае числовые показатели префиксов длины 100 последовательностей имеют пересечения, например по числовому показателю 8. Это означает, что, например, по числовому показателю уровня Ω_0 последовательности $\xi_{1,1}$, $\xi_{1,2}$, $\xi_{1,3}$, $\xi_{1,4}$, $\xi_{2,4}$, $\xi_{2,5}$, $\xi_{3,5}$ не различимы.

Аналогичный вывод следует для префиксов последовательности длины 1000. На префиксах длины 5000 числовые показатели рекуррентного определения последовательностей совпадают только для последовательностей $\xi_{1,1}$, $\xi_{3,1}$. Следовательно, по числовым показателям рекуррентного определения префиксов длины 5000 генетических последовательностей не различимыми являются только последовательности $\xi_{1,1}$, $\xi_{3,1}$. Если для генетических последовательностей рассматриваемых классов вычисляются числовые показатели уровня Ω_0 для префиксов длины 10 000, то каждая из рассматриваемых последовательностей имеет свой, принадлежащий только ей числовой показатель: классу K_1 соответствуют числовые показатели 23, 41, 49, 64, 81; классу K_2 — показатели 15, 16, 73; классу K_3 — показатели 17, 25, 40, 60, 378.

Числовые показатели следующего уровня Ω_1 дают более точную и полную характеристику рекуррентного определения последовательностей.

Для того чтобы числовые показатели по уровню Ω_0 для классов K_1, K_2, K_3 различались, потребовалось рассматривать префиксы длины 5000. Существенно большую полноту и точность представления структур генетических последовательностей дают числовые показатели уровня Ω_1. В табл. 4–6 приведены числовые показатели для префикса длины 100. Как видно из данных таблиц, чтобы отличать генетические последовательности по их принадлежности конкретным классам, достаточно вычислить и сравнить числовые показатели префиксов длины 100.

Анализ по числовым показателям уровня Ω_2. Рассмотрим случай распознавания генетических последовательностей как элементов в одном и том же классе. Для класса K_1 числовые показатели отдельных последовательностей имеют пересечения для префиксов длин 100 и 1000. Для префиксов длины 5000 числовые показатели уровня Ω_0 для последовательностей $\xi 1, 1, ..$ пересечений не имеют. Следует предполагать, что последовательности в классе K_2 по числовым показателям функциональных зависимостей элементов в последовательностях имеют бóльшие различия, так как пересечения числовых показателей на префиксах длины 1000 не имеется. Анализ числовых показателей в табл. 3 показывает, что индивидуальные, присущие каждой последовательности из класса K_3 числовые показатели не пересекаются на префиксах длины 10 000.

Числовые показатели рекуррентного определения последовательностей уровня Ω_1 достаточны, чтобы по их показателям распознавать каждую последовательность в каждом из классов K_1, K_2, K_3. Действительно, наборы числовых показателей для префиксов длины 100 последовательностей в табл. 4–6 в каждом из классов K_1, K_2, K_3 попарно различны.

ЗАКЛЮЧЕНИЕ

В исследовании генетических последовательностей одной из основных задач является определение причинно-следственных связей между свойствами генетических последовательностей и свойствами организмов, которым соответствуют последовательности.

Любые конкретные свойства генетических последовательностей, представленные формальными показателями, определяют класс последовательностей, причинно-следственно связанный с классами живых организмов, свойствами живых организмов, видами патологий организмов и т.д.

Для исследований генетических последовательностей в данной статье предлагается разработанный формальный аппарат, который позволяет:

– определять последовательность в виде числовых структур (целых положительных чисел, наборов целых положительных чисел, наборов наборов целых положительных чисел);

– строить модели для конкретных генетических последовательностей и модели для любых конечных множеств генетических последовательностей;

– разрабатывать эффективные алгоритмы для построения моделей генетических последовательностей и моделей множеств генетических последовательностей в форме числовых структур;

– выделять из универсума генетических последовательностей с использованием моделей классы генетических последовательностей и на этой основе строить классификацию;

– с использованием разработанных моделей генетических последовательностей

решать задачи проверки равенства генетических последовательностей и задачу проверки принадлежности рассматриваемой генетической последовательности конкретному классу последовательностей;

– сравнивать классы генетических последовательностей с определением непустоты области их пересечения.

В статье результаты вычислительного эксперимента иллюстрируются простейшими примерами.

Библиографический список

1. *Твердохлебов В. А.* Геометрическая форма автоматных отображений, рекуррентное и Z-рекуррентное определение последовательностей // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика. 2016. Т. 16, вып. 2. С. 232–241. DOI: https://doi.org/10.18500/1816-9791-2016-16-2-232-241
2. *Твердохлебов В. А.* Z-рекуррентное определение последовательностей в задачах контроля и диагностирования процессов в системах // Докл. Акад. воен. наук. 2016. № 2 (70). С. 43–47.
3. *Карякин Д. А.* Анализ генетических кодов по показателям сложности взаиморасположения нуклеотидов // Компьютерные науки и информационные технологии : материалы междунар. науч. конф. Саратов : ИЦ «Наука», 2016. С. 190–193.
4. *Льюин Б.* Гены. М. : БИНОМ, Лаборатория знаний, 2011. 896 с.
5. *Уотсон Д.* Двойная спираль. Воспоминания об открытии структуры ДНК. М. : Мир, 1969. 152 с.
6. *Hogeweg P.* The Roots of Bioinformatics in Theoretical Biology // PLoS. Computational Biology. 2011. Vol. 7, iss. 3. Art. ID e1002021. DOI: https://doi.org/10.1371/journal.pcbi.1002021
7. *Wattam A. R., Abraham D., Dalay O., Disz T. L., Driscoll T., Gabbard J. L., Gillespie J. J., Gough R., Hix D., Kenyon R., Machi D., Mao C., Nordberg E. K., Olson R., Overbeek R., Pusch G. D., Shukla M., Schulman J., Stevens R. L., Sullivan D. E., Vonstein V., Warren A., Will R., Wilson M. J., Yoo H. S., Zhang C., Zhang Y., Sobral B. W.* PATRIC, the bacterial bioinformatics database and analysis resource // Nucleic Acids Res. 2014. Vol. 42, iss. D1. P. D581–D591. DOI: https://doi.org/10.1093/nar/gkt1099
8. *Barnett D. W., Garrison E. K., Quinlan A. R., Stromberg M. P., Marth G. T.* BamTools: a C++ API and toolkit for analyzing and managing BAM files // Bioinformatics. 2011. Vol. 27, iss. 12. P. 1691–1692. DOI: https://doi.org/10.1093/bioinformatics/btr174
9. *Plieskatt J., Rinaldi G., Brindley P. J., Jia X., Potriquet J., Bethony J., Mulvenna J.* Bioclojure: a functional library for the manipulation of biological sequences // Bioinformatics. 2014. Vol. 30, iss. 17. P. 2537–2539. DOI: https://doi.org/10.1093/bioinformatics/btu311
10. *Goto N., Prins P., Nakao M., Bonnal R., Aerts J., Katayama T.* BioRuby: bioinformatics software for the Ruby programming language // Bioinformatics. 2010. Vol. 26, iss. 20. P. 2617–2619. DOI: https://doi.org/10.1093/bioinformatics/btq475
11. *de Brevern A. G., Meyniel J. P., Fairhead C., Neuvéglise C., Malpertuy A.* Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies // BioMed Research International. Vol. 2015. Article ID 904541, 15 p. DOI: http://dx.doi.org/10.1155/2015/904541
12. *Schuster S. C.* Next-generation sequencing transforms today's biology // Nature Methods. 2008. Vol. 5, iss. 1. P. 16–18. DOI: https://doi.org/10.1038/nmeth1156
13. *Сингер М., Берг П.* Гены и геномы. М. : Мир, 1998. 391 с.
14. *Berg J. M., Tymoczko J. L., Stryer L.* DNA, RNA, and the Flow of Genetic Information // Berg J. M., Tymoczko J. L., Stryer L. Biochemistry. 5th ed. N. Y. : W. H. Freeman and Company, 2002. 1515 p.

15. NCBI Genome List. URL: http://www.ncbi.nlm.nih.gov/genome/browse/ (дата обращения: 18.12.2017).

Образец для цитирования:
Твердохлебов В. А., Карякин Д. А. Классификация и распознавание структур генетических последовательностей // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика. 2019. Т. 19, вып. 3. С. 338–350. DOI: https://doi.org/10.18500/1816-9791-2019-19-3-338-350

Classification and Recognition of Structures of Genetic Sequences

V. A. Tverdokhlebov, D. A. Kariakin

Vladimir A. Tverdokhlebov, https://orcid.org/0000-0002-2629-441X, Institute of Precision Mechanics and Control, RAS, 24 Rabochaya St., Saratov 410028, Russia, tverdokhlebovva@list.ru

Denis A. Kariakin, https://orcid.org/0000-0002-0670-3407, Saratov State University, 83 Astrakhanskaya St., Saratov 410012, Russia, dakariakin@gmail.com

For solving problems of determining the relationships between the properties of organisms and the properties of the corresponding genetic sequences, we proposed a classification of genetic sequences based on numerical indicators of recurrent and Z-recurrent shapes, which define the structure of functional relationships of elements in sequences. For numerical indicators of recurrent and Z-recurrent shapes, we introduce a method of classification of genetic sequences. We compared a numerical characteristic that generalizes numerical values with a numerical characteristic of recurrent or Z-recurrent shapes which determine the structure of a sequence for each sequence of a biological rank considered in the recognition problem, which has a meaningful interpretation in the application area. The problem of recognition is considered from two points of view: when we determine belonging of a sequence to a specific rank of sequences, and when we determine which group of sequences contains the experimental sequence. Basic mathematical difficulties in solving these recognition problems are associated with the search difference in numerical representation of recurrent and Z-recurrent shapes of experimental sequences. To overcome these difficulties we created a spectrum of numerical indicators of recurrent and Z-recurrent shapes. Classification and recognition of sequences are illustrated by an example with three ranks of genetic codes of organisms, each of them represented by 5 sequences. Z-recurrent shape is introduced to define and extend the classification of sequences and increase the efficiency of recognition methods.

Keywords: sequence, genetic sequence, recurrent definition of a sequence, Z-recursive definition of a sequence, recurrent shape, Z-recurrent shape, classification of sequences, recognition of sequences.

Received: 12.04.2018 / Accepted: 22.02.2019 / Published: 31.08.2019

References

1. Tverdokhlebov V. A. Geometric Shape Automaton Mappings, Recurrent and Z-recurrent Definition Sequences. *Izv. Saratov Univ. (N. S.), Ser. Math. Mech. Inform.*, 2016, vol. 16, iss. 2, pp. 232–241 (in Russian). DOI: https://doi.org/10.18500/1816-9791-2016-16-2-232-241

2. Tverdokhlebov V. A. *Z*-recurrent definition sequences in the tasks of monitoring and diagnosing processes in systems. *Reports of the Academy of Military Sciences*, 2016, no. 2 (70), pp. 43–47 (in Russian).
3. Kariakin D. A. Analysis of genetic codes by indicators interposition of nucleotides. In: *Komp'yuternye nauki i informatsionnye tekhnologii* [Computer Science and Information Technology: Proc. Int. Sci. Conf.]. Saratov, Publ. Center "Nauka", 2016, pp. 190–193 (in Russian).
4. Lewin B. *Geny* [Genes]. Moscow, BINOM, Laboratoriya znanij Publ., 2011. 896 p. (in Russian).
5. Watson D. *Dvojnaya spiral'. Vospominaniya ob otkrytii struktury DNK* [Double helix. Memories of the discovery of the structure of DNA]. Moscow. Mir, 1969. 152 p. (in Russian).
6. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology. *PLoS. Computational Biology*, 2011, vol. 7, iss. 3, art. ID e1002021. DOI: https://doi.org/10.1371/journal.pcbi.1002021
7. Wattam A. R., Abraham D., Dalay O., Disz T. L., Driscoll T., Gabbard J. L., Gillespie J. J., Gough R., Hix D., Kenyon R., Machi D., Mao C., Nordberg E. K., Olson R., Overbeek R., Pusch G. D., Shukla M., Schulman J., Stevens R. L., Sullivan D. E., Vonstein V., Warren A., Will R., Wilson M. J., Yoo H. S., Zhang C., Zhang Y., Sobral B. W. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, 2014, vol. 42, iss. D1, pp. D581–D591. DOI: https://doi.org/10.1093/nar/gkt1099
8. Barnett D. W., Garrison E. K., Quinlan A. R., Stromberg M. P., Marth G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 2011, vol. 27, iss. 12, pp. 1691–1692. DOI: https://doi.org/10.1093/bioinformatics/btr174
9. Plieskatt J., Rinaldi G., Brindley P. J., Jia X., Potriquet J., Bethony J., Mulvenna J. Bioclojure: a functional library for the manipulation of biological sequences. *Bioinformatics*, 2014, vol. 30, iss. 17, pp. 2537–2539. DOI: https://doi.org/10.1093/bioinformatics/btu311
10. Goto N., Prins P., Nakao M., Bonnal R., Aerts J., Katayama T. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*, 2010, vol. 26, iss. 20, pp. 2617–2619. DOI: https://doi.org/10.1093/bioinformatics/btq475
11. de Brevern A. G., Meyniel J. P., Fairhead C., Neuvéglise C., Malpertuy A. Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies. *BioMed Research International*, vol. 2015, art. ID 904541, 15 p. DOI: http://dx.doi.org/10.1155/2015/904541
12. Schuster S. C. Next-generation sequencing transforms today's biology. *Nature Methods*, 2008, vol. 5, iss. 1, pp. 16–18. DOI: https://doi.org/10.1038/nmeth1156
13. Singer M., Berg P. *Geny i genomy* [Genes and genomes]. Moscow, Mir, 1998. 391 p. (in Russian).
14. Berg J. M., Tymoczko J. L., Stryer L. DNA, RNA, and the Flow of Genetic Information. In: *Berg J. M., Tymoczko J. L., Stryer L. Biochemistry*. 5th. ed. New York, W. H. Freeman and Company, 2002. 1515 p.
15. NCBI Genome List. Available at: http://www.ncbi.nlm.nih.gov/genome/browse/ (accessed 18 Desember 2017).

Cite this article as:
Tverdokhlebov V. A, Kariakin D. A. Classification and Recognition of Structures of Genetic Sequences. *Izv. Saratov Univ. (N. S.), Ser. Math. Mech. Inform.*, 2019, vol. 19, iss. 3, pp. 338–350 (in Russian). DOI: https://doi.org/10.18500/1816-9791-2019-19-3-338-350