УДК 004.074
doi:10.31799/1684-8853-2021-3-39-52

Научные обзоры
Surveys

Теоретико-информационные проблемы ДНК-памяти

С. А. Круглик[а,б], *младший научный сотрудник, orcid.org/0000-0001-9557-5197 stanislav.kruglik@skoltech.ru*
Г. А. Кучеров[а,в], *канд. физ.-мат. наук, ведущий научный сотрудник, orcid.org/0000-0001-5899-5424*
К. Н. Назирханова[г], *аспирант, orcid.org/0000-0002-7447-9857*
М. Е. Филитов[а], *магистрант, orcid.org/0000-0003-2421-0777*
[а]*Сколковский институт науки и технологий, Большой б-р, 30, стр. 1, Москва, 121205, РФ*
[б]*Московский физико-технический институт, Институтский пер., 9, Долгопрудный, Московская обл., 141701, РФ*
[в]*Национальный центр научных исследований, Университет Густава Эйфеля, 77454 Марн-ля-Валле, Франция*
[г]*Стэнфордский университет, 94305 Стэнфорд, Калифорния, США*

Введение: *взрывной рост объемов производимой человечеством информации ставит новые фундаментальные задачи, связанные с ее эффективным хранением и доступом к ней. Широко используемые при этом магнитные, оптические и полупроводниковые устройства хранения имеют ряд существенных недостатков, связанных, прежде всего, с ограничениями на объем и долговечность хранения. Одной из возможных альтернатив, активно исследуемой в последние годы, является хранение данных с помощью молекул ДНК.* ***Цель:*** *обзор текущего состояния методов хранения информации с помощью молекул ДНК и связанных теоретико-информационных проблем.* ***Результаты:*** *сделан обзор современного состояния дел в разработке систем ДНК-памяти. Проведен анализ типов ошибок, возникающих в таких системах, и корректирующих кодов для выявления и исправления этих ошибок. Показаны недостатки предложенных на сегодня кодов и указаны возможные направления их улучшения. Приведен анализ существующих теоретико-информационных моделей каналов для систем ДНК-памяти и присущих им ограничений. В заключении обзора сформулированы основные проблемы на пути создания практических систем ДНК-памяти, решению которых послужит дальнейшее развитие теоретико-информационных методов, рассмотренных в настоящем обзоре.*

Ключевые слова *— системы хранения информации, ДНК-память, каналы передачи информации, пропускная способность канала, ошибки замены, ошибки вставки, ошибки выпадения.*

Для цитирования: Круглик С. А., Кучеров Г. А., Назирханова К. Н., Филитов М. Е. Теоретико-информационные проблемы ДНК-памяти. *Информационно-управляющие системы*, 2021, № 3, с. 39–52. doi:10.31799/1684-8853-2021-3-39-52
For citation: Kruglik S. A., Kucherov G. A., Nazirkhanova K. N., Filitov M. E. Information-theoretic problems of DNA-based storage systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 3, pp. 39–52 (In Russian). doi:10.31799/1684-8853-2021-3-39-52

Введение

Мы живем в эпоху цифровых технологий, в которой множество различных устройств ежедневно генерируют огромное количество данных. При этом как число таких устройств, так и объемы генерируемой ими информации растут с экспоненциальной скоростью [1]. По имеющимся оценкам, суммарный объем производимой человечеством информации достигает миллиардов терабайт в год [2, 3]. Столь быстрый рост поднимает множество вопросов, связанных, прежде всего, с хранением данных и управлением ими. В частности, остро стоит проблема увеличения емкости существующих хранилищ [4]. В настоящее время большая часть данных хранится на магнитных и оптических устройствах, таких как компакт-диски, жесткие диски и магнитные ленты. Еще недавно магнитные устройства являлись наиболее популярным и доступным решением, но затем их заменили оптические устройства. Теперь же и они вытесняются более функциональными и дешевыми устройствами флэш-памяти. Однако все вышеперечисленные носители имеют ряд общих недостатков. Во-первых, их емкости ограничены. Например, максимальный объем, который может хранить магнитная лента, это эксабайт данных, но такое хранилище может стоить весьма дорого в обслуживании и занимает значительное пространство [5]. Кроме того, все эти устройства имеют низкую плотность хранения, как правило, не превышающую тысячи гигабайт на квадратный миллиметр [6]. Еще одной проблемой, связанной с хранением данных на существующих носителях, является возможность потери данных с течением времени. Все это требует разработки принципиально новых способов хранения информации [7, 8].

Одним из таких методов, активно изучаемых в последнее время, является хранение информации с использованием молекул ДНК. Отметим, что вскоре после открытия структуры ДНК

в 1953 г. [9] некоторые известные ученые высказали предположения о возможном использовании ДНК для хранения произвольной цифровой информации. Об этих перспективах говорил выдающийся физик Ричард Фейнман в своей лекции «There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics» в 1959 г. В 60-х годах подобные идеи высказывал Норберт Винер [10], а в Советском Союзе — физик и радиоинженер М. С. Нейман [11, 12]. Последний в своих работах изложил соображения о возможных способах реализации данных систем и некоторые предварительные расчеты, которые, однако, в то время были далеки от практической реализации. Появлению и раннему развитию идей ДНК-памяти посвящен исторический обзор [13].

Первые успешные попытки сохранить информацию с использованием ДНК относятся к 1988 г., когда коллективу ученых удалось вставить в плазмидную ДНК бактерии *Escherichia coli* искусственный фрагмент из 28 нуклеотидов, из которых 18 кодировали простой символ-пиктограмму, а оставшиеся 10 содержали метаинформацию для декодирования. Впоследствии этот фрагмент был успешно извлечен из бактериальной ДНК с помощью секвенирования [14]. Интересно, что целью этого эксперимента было создание нового типа художественного объекта. Подобные эксперименты проводились и позже, в конце 90-х: так, например, в работе [15] описан эксперимент по передаче секретных сообщений, закодированных в растворе ДНК, а в [16] — эксперимент по кодированию в ДНК коротких предложений на естественном языке. Однако эти эксперименты имели целью кодирование в ДНК лишь очень небольшого объема информации, измеряемого десятками байтов, и не допускали масштабирования.

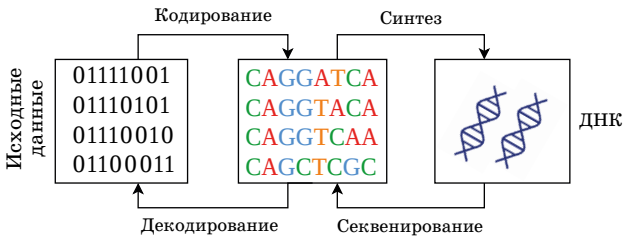
Ситуация изменилась в 2012-м, когда с использованием новых технологий синтеза и секвенирования ДНК был закодирован набор из 643 Кбайт данных, состоящий из книги, 11 изображений JPG и одной программы JavaScript [17]. Годом позже другим коллективом ученых была представлена схема хранения 739 Кбайт произвольной цифровой информации с использованием ДНК [18]. Эти работы открыли новый этап в развитии систем ДНК-памяти, послужив началом серии экспериментов по кодированию в молекулах ДНК все больших объемов информации. Так, в 2018 г. авторам работы [19] удалось сохранить 200 Мбайт пользовательских данных, а уже в 2019-м авторы работы [20] смогли сохранить 16 Гбайт англоязычной Википедии.

В последние годы к тематике ДНК-памяти прикован большой интерес исследовательского сообщества [21, 22], и многие ученые рассматривают ее как перспективный носитель для долговременного хранения гигантских объемов информации, с которыми сегодня имеет дело человечество. Скорость размещения информации в ДНК-памяти и доступа к ней вряд ли в обозримом будущем сможет конкурировать с ныне используемыми носителями, однако по плотности информации, стоимости хранения и долговечности ДНК-память может их превзойти качественным образом.

Главным препятствием для практической реализации ДНК-памяти являются ошибки, т. е. модификации в последовательностях ДНК, неизбежно возникающие в процессе манипуляции с ДНК. В связи с этим в последние несколько лет произошел всплеск работ по применению идей теории кодирования для надежной реализации ДНК-памяти. Анализу типов ошибок и методам их коррекции уделено главное внимание в настоящем обзоре. В качестве основных источников информации использованы труды ведущих конференций по теории информации и теоретической информатике последних лет, а также специализированная научная периодика.

В упрощенном виде процедура хранения информации с использованием ДНК может быть представлена следующим образом: первоначально исходная информация (двоичный код) преобразуется в четверичный алфавит, соответствующий четырем нуклеотидам, составляющим ДНК. Затем в целях борьбы с возможными ошибками к полученной таким образом последовательности применяется помехоустойчивое кодирование. После чего сгенерированная синтетическая ДНК помещается в специальное хранилище или же встраивается в существующую ДНК живого организма. Для извлечения информации применяются процедуры секвенирования и декодирования имеющейся нуклеотидной последовательности [23–26]. Эти шаги схематически представлены на рис. 1 и будут рассмотрены более подробно в разделе «Технологии систем ДНК-памяти» данного обзора.

Ошибки, возникающие в системах ДНК-памяти, не ограничиваются простой заменой символов, а включают также выпадения и вставки, когда некоторые из передаваемых символов «те-



■ ***Рис. 1.*** Общая архитектура системы ДНК-памяти

■ ***Fig. 1.*** General architecture of DNA-based storage systems

ряются» или же, наоборот, в последовательность вставляются символы, ей не принадлежащие. Встречаются также определенные комбинации таких ошибок, в частности дупликации и пакетные выпадения или вставки [27–29]. Обзору наиболее часто встречающихся в системах ДНК-памяти ошибок, а также кодов, их исправляющих, посвящен раздел «Ошибки в системах ДНК-памяти и коды, их исправляющие» настоящего исследования.

Пользовательские данные представляются в системах ДНК-памяти в виде большого числа коротких последовательностей нуклеотидов, также называемых олигонуклеотидами. При этом в процессе манипуляций с ними, помимо вышеописанных ошибок на уровне символов, могут также происходить ошибки на уровне самих последовательностей, такие как изменение числа или удаление некоторых из них. Это, в свою очередь, приводит к изменению в пуле последовательностей на выходе системы ДНК-памяти по отношению ко входу. Рассмотрению возникающих математических моделей каналов передачи информации, а также построению кодов для них посвящен раздел «Модели каналов для систем ДНК-памяти».

Технологии систем ДНК-памяти

Рассмотрим более подробно этапы хранения информации на основе ДНК (см. рис. 1). Перед началом процедуры кодирования данные необходимо привести в формат, соответствующий четырем нуклеотидам (A, C, G, T), образующим молекулы ДНК. В теории информации такая процедура называется кодированием источника [30, 31]. Так, например, в своей ранней работе [15] авторы использовали простое отображение букв английского алфавита, знаков препинания и цифр в последовательности из нуклеотидов по заранее определенному правилу, по которому буква D английского алфавита преобразуется в последовательность TTG, буква N — в TCT, буква A — в CGA. В результате слово DNA преобразуется в последовательность нуклеотидов TTGTCTCGA. Позднее другими учеными [17] была предложена схема кодирования с помощью ДНК относительно больших объемов разнородной информации, изначально представленной в html-файле. При этом символ 0 представлялся в виде нуклеотида A или C, выбираемого случайным образом. Символ же 1 представлялся в виде нуклеотида T или G, также выбираемого случайно. В данном случае двоичная последовательность 0100 представлялась в виде AGAC. В работе [18] к данным перед их преобразованием в последовательность нуклеотидов авторы применили троичный алгоритм Хаффмана для кодирования источника. При этом кодирование в алфавит нуклеотидов было определено таким образом, чтобы избежать повторения подряд одного нуклеотида в целевой последовательности ДНК. Это достигалось с помощью специальной таблицы преобразования, определяющей правило отображения текущего троичного символа в нуклеотид в зависимости от значения предыдущего нуклеотида. В частности, при предыдущем нуклеотиде A символ 0 отображался в C, а при C — уже в G. При этом отображение первого символа определялось по правилу, при котором предыдущим нуклеотидом является A. Например, последовательность 0020 представлялась в виде CGCG. Данное требование вызвано повышенной вероятностью возникновения ошибки, свойственной современным технологиям секвенирования, при секвенировании гомополимерных (состоящих из одного нуклеотида) участков ДНК [32]. Другим важным для кодирования обстоятельством, влияющим на уровень ошибок и надежность хранения, является доля GC нуклеотидов [33]. Этот фактор учитывался, например, в работе [23], где кодирование было организовано таким образом, чтобы избежать гомополимерных участков длины больше 3, а также олигонуклеотидов с GC-содержанием больше 55 % либо меньше 45 %. Это достигалось путем предварительного применения преобразования Луби с различными псевдослучайными параметрами и отбрасывания неподходящих последовательностей. При этом преобразование бит в последовательность нуклеотидов осуществлялось по заранее определенному правилу $00 \to A$, $01 \to C$, $10 \to G$, $11 \to T$. В таком случае строка 0100 представлялась в виде CA.

Для борьбы с ошибками, возникающими в процессе хранения данных с использованием ДНК, в информационную последовательность вносят дополнительную избыточность. Различают физическую и логическую избыточность. Физическая избыточность предполагает увеличение «покрытия», или, иными словами, числа копий молекул ДНК, хранящих информацию об одном и том же участке исходной последовательности. Например, в работе [18] использовалось четырехкратное покрытие, тогда как в работе [34] для кодирования одной и той же информации использовалось несколько молекул ДНК, полученных путем «сдвига фазы» в процессе преобразования исходной двоичной последовательности в последовательность нуклеотидов. К сожалению, подобные методы не позволяют полностью застраховать информацию от возникающих ошибок. Другой независимый способ исправления возникающих ошибок — использование логической избыточности, задаваемой с помощью помехоустойчивых кодов. Этот метод

требует значительно меньшего объема дополнительной информации, что, в свою очередь, увеличивает итоговую плотность хранения. Коды, используемые для обнаружения и коррекции возникающих ошибок, в частности ошибок типа вставки и выпадения, будут рассмотрены в следующем разделе.

После приведения исходной последовательности к виду, соответствующему четырем нуклеотидам, ее необходимо преобразовать в нуклеотидную последовательность и поместить соответствующую молекулу ДНК в некоторое хранилище. Большинство существующих экспериментальных систем оперируют с искусственно синтезированными олигонуклеотидами, хранящимися в виде раствора, однако ДНК простейших живых организмов (как правило, бактерий или других микроорганизмов) также потенциально может быть использована для хранения синтезированной ДНК. Такой подход применен в ранней работе [14]. В дальнейшем было установлено [19], что бактерия способна нести в себе около одного мегабайта пользовательской информации, что сопоставимо с информацией, хранящейся в ее собственном геноме [35]. Несмотря на это кодирование информации in vivo, помимо очевидных технологических ограничений, вряд ли может быть масштабировано на большие объемы информации. В настоящем обзоре мы сосредоточим наше внимание на кодировании in vitro.

Современные технологии синтеза ДНК позволяют синтезировать одновременно на одном микрочипе множество коротких одноцепочечных олигонуклеотидов, представленных во многих экземплярах. Количество различных синтезируемых олигонуклеотидов может достигать нескольких миллионов, а каждый олигонуклеотид может быть представлен десятками или сотнями тысяч копий, однако из-за технологических погрешностей количество копий для разных олигонуклеотидов может существенно различаться [36–38]. С другой стороны, длины этих фрагментов очень малы, порядка 200 нуклеотидов. При этом сам процесс синтеза подвержен ошибкам типа замены, вставки и выпадения нуклеотидов, в результате которых экземпляры одного и того же олигонуклеотида могут слегка различаться. Отдельно отметим появление новых перспективных технологий синтеза, которые могут существенно сократить его стоимость и увеличить скорость, а также уменьшить число возникающих ошибок в ближайшем будущем [39].

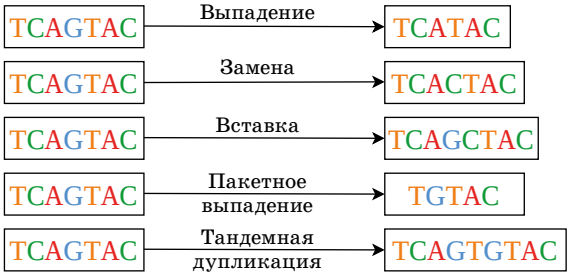
Для представления исходной последовательности в виде набора олигонуклеотидов каждый олигонуклеотид должен содержать информацию, обычно называемую индексом, о позиции соответствующего фрагмента во входной последовательности. Кроме того, по краям олигонуклеотида добавляются специальные последовательности — праймеры для полимеразной цепной реакции (ПЦР). С помощью ПЦР синтезированные олигонуклеотиды затем амплифицируются, т. е. «размножаются» до большего числа копий. Этот этап также позволяет отфильтровать «неправильно синтезированные» олигонуклеотиды, не содержащие праймеров ПЦР, однако вносит дополнительную неравномерность в число копий олигонуклеотидов из-за неравномерной амплификации.

Для извлечения информации из ДНК-памяти применяются стандартные технологии секвенирования ДНК [40]. С появлением так называемых секвенаторов нового поколения в середине 2000-х годов эти технологии вышли на качественно новый уровень, позволяя считывать фрагменты ДНК общим объемом миллиарды баз (нуклеотидов) за один цикл работы. Одними из наиболее распространенных на сегодня являются секвенаторы компании Illumina. По сравнению с другими высокопроизводительными технологиями, секвенаторы Illumina характеризуются относительно низким уровнем ошибок, порядка 0,1 %. Подавляющее большинство этих ошибок являются заменами, причем частота вставок и выпадений примерно на два порядка ниже частоты замен. При этом, как отмечалось, уровень ошибок зависит от последовательности: он может быть существенно выше для последовательностей с очень низким или, наоборот, очень высоким GC-содержанием, а также для последовательностей, содержащих длинные (шесть нуклеотидов и более) гомополимерные фрагменты [37].

Ошибки в системах ДНК-памяти и коды, их исправляющие

Типичными ошибками в процессе манипуляции с молекулами ДНК являются замены, вставки и выпадения нуклеотидов. Кроме того, случаются пакетные вставки и выпадения, т. е. одновременные вставки или выпадения коротких фрагментов, а также дупликации, когда в последовательность вставляется копия некоторого его фрагмента. Примеры ошибок представлены на рис. 2.

Модель, когда ошибки ограничиваются только заменами символов, хорошо изучена в теории кодирования, и для нее известно много конструкций кодов с хорошими корректирующими свойствами, в том числе конструкции оптимальных кодов. Модель ошибок типа вставки и выпадения, исследование которой началось в работах [27–29], оказывается гораздо более сложной. В этом разделе мы сосредоточимся на кодах, исправляющих именно такие ошибки, а также их



■ ***Рис. 2.*** Примеры ошибок в молекулах ДНК
■ ***Fig. 2.*** Examples of errors in DNA-molecules

комбинации. Отметим, что некоторые существующие модели систем ДНК-памяти ограничиваются рассмотрением кодов, исправляющих лишь ошибки замены [41], хотя эта модель ошибок плохо соответствует практическим системам.

В некоторых работах для исправления вставок/выпадений применялись различные эвристики, например, в работе [42] в центр каждой последовательности ДНК вставлялся специальный маркер, позволяющий локализовать данную ошибку и тем самым привести ее к ошибке замены. В качестве корректирующих кодов в работе [42] рассматривались коды с малой плотностью проверок на четность длины 256. В работах [43–46] применялись коды Рида — Соломона, а в [23] применены фонтанные коды. Основным недостатком данных конструкций является то, что результирующие коды гарантируют исправление лишь небольшого числа ошибок вставки/выпадения, что делает затруднительным применение этих кодов в системах ДНК-памяти.

Рассмотрим теперь ошибки типа вставки и выпадения. К первой конструкции кодов для их исправления можно отнести коды Варшамова — Тененгольца [28, 47]. Такие коды состоят из всех двоичных векторов длины n, поэлементная сумма которых принимает некоторое фиксированное значение по модулю n. Отметим, что данные коды исправляют одну ошибку вставки или выпадения и являются асимптотически оптимальными. Систематическая версия данных кодов представлена в работе [48], а на их основе были построены асимптотически оптимальные коды, исправляющие две последовательные вставки или выпадения [49].

Обобщение кодов Варшамова — Тененгольца на случай большего числа вставок и выпадений является сложной научной задачей. Большинство предложенных явных кодовых конструкций, в частности конструкция из работы [50], обобщающая их на случай исправления до пяти вставок и выпадений, являются неоптимальными с точки зрения кодовой скорости, а также не обладают эффективными алгоритмами кодирования и декодирования. Коды, исправляющие до двух вставок и выпадений и улучшающие конструкцию [50], были получены с помощью построения кодовой книги путем перебора [51]. Конструкции, не использующие перебор, получены в работе [52].

Первая конструкция кодов для исправления фиксированного числа вставок и выпадений, обладающая малой избыточностью и эффективными алгоритмами кодирования и декодирования, была представлена в работе [53]. Эта конструкция обобщает идею разделения кодового слова на блоки фиксированного размера и последующего отделения их друг от друга путем добавления длинных последовательностей из нулей [54]. Процедура декодирования исходного кодового слова при этом осуществляется по мажоритарному правилу. Впоследствии полученный результат был улучшен [55, 56].

Все вышеописанные конструкции подразумевают исправление фиксированного числа вставок и выпадений, в то время как наиболее интересным с практической точки зрения является случай фиксированной доли вставок и выпадений. Первые коды для этого случая были построены в работе [57]. Данная конструкция основывается на каскадной схеме, подразумевающей использование внутренних кодов, полученных с помощью перебора и внешних кодов Рида — Соломона, и обладает полиномиальными алгоритмами кодирования и декодирования. Верхняя граница на скорость кодов, обладающих данным свойством, получена в работе [58]. Там же показано, что коды из работы [57] являются асимптотически оптимальными.

Конструкции кодов, исправляющих вставки и выпадения для двух наиболее интересных случаев: высокой кодовой скорости и большой доли вставок/выпадений, — были предложены в работе [54]. Эти конструкции также основываются на каскадной схеме, состоящей из кода с хорошими корректирующими свойствами с точки зрения исправления ошибок замены (в частности, кода Рида — Соломона и кодов Парвареша — Варди) и кодов для исправления вставок/выпадений со значительно меньшей длиной блока. Последнее свойство позволяет декодировать данный код методом простого перебора. При этом в случае высокой доли вставок/выпадений необходимо специальным образом учитывать влияние неправильного декодирования внутреннего кода на внешний код. В случае высокой кодовой скорости, в свою очередь, необходимо разделять блоки внутренних кодовых слов с помощью блоков из последовательных нулей. Ключевой особенностью предложенного подхода является определение параметров декодирования внутреннего кода, позволяющее внешнему коду восстановить исходное сообщение с помощью списочного де-

кодера. Улучшение данной конструкции для исправления большей доли выпадений предложено в работе [59]. Обобщение на случай фиксированной доли как вставок, так и выпадений представлено в работе [60].

Еще одним методом, позволяющим исправлять вставки и выпадения, является преобразование корректирующих кодов для ошибок замены с помощью строк синхронизации [61]. Его основная идея состоит в добавлении к каждому символу исходной кодовой последовательности специальным образом построенной синхронизирующей строки над малым алфавитом и исправлении произошедших ошибок типа вставки/выпадения с помощью декодера исходного кода.

Детерминированный способ построения синхронизирующих строк предложен в работе [62]. В случае если после передачи по каналу со вставками/выпадениями получена синхронизирующая строка S', то алгоритм приведения ошибок типа вставки/выпадения к ошибкам замены последовательно находит наибольшие соответствия между S и S'. Если какой-либо элемент исходной строки находится во всех полученных ранее соответствиях, то он поступает на вход декодера исходного кода. В работе [61] приведены оценки на число исправляемых таким образом вставок и выпадений, связанные с корректирующей способностью применяемого кода и свойствами синхронизирующей строки.

Последние результаты для кодов, исправляющих фиксированную долю вставок и выпадений, получены в работах [63, 64] в рамках исследования связанной задачи об обмене документами, популярной в области теоретической информатики. В этом контексте кодирование происходит с помощью добавления к исходной строке дополнительного сообщения — скетча (sketch), с помощью которого подверженная ошибкам исходная строка может быть впоследствии декодирована.

Отметим, что приведенные выше коды способны эффективно исправлять фиксированную долю ошибок лишь определенного типа, тогда как в системах ДНК-памяти могут одновременно происходить как вставки и выпадения, так и замены символов. При этом различные типы ошибок имеют различные вероятности, определяемые структурой конкретной системы.

Важным обобщением ошибок вставки и выпадения индивидуальных символов, возникающих в системах ДНК-памяти, являются вставки или выпадения последовательных символов, также называемых пакетными. Впервые такая постановка задачи введена в работе [49], в которой рассматривались группы, состоящие не более чем из двух последовательных выпадений, для которых были предложены корректирующие коды с оптимальной избыточностью. Отметим различие между кодами, исправляющими до k последовательных вставок или выпадений, и кодами, исправляющими ровно k последовательных вставок или выпадений. Применение последних, в свою очередь, не гарантирует исправление меньшего числа ошибок в общем случае.

В работе [65] исследованы коды, исправляющие k последовательных вставок или выпадений (но не комбинацию их), обобщенные также на случай не более k последовательных вставок или выпадений. Результат для случая ровно k последовательных вставок или выпадений был в дальнейшем улучшен в работе [66], а для случая не более k таких ошибок — в работах [67, 68]. В частности, в работе [68] получены коды с асимптотически оптимальным значением избыточности. При этом все отмеченные выше конструкции основываются на различных модификациях кодов Варшамова — Тененгольца. Отметим, что возможный случай нескольких пакетных выпадений ограниченной длины является менее изученной проблемой. Для решения последней в работе [69] предложен метод синдромной компрессии (syndrome compression) в применении к конструкции кодов из работы [53]. При этом полученные таким образом коды не являются асимптотически оптимальными.

Другим распространенным типом вставок являются дупликации, когда некоторый фрагмент последовательности копируется в позицию справа от него. Ошибки дупликации были впервые рассмотрены в работе [70], в которой исследован случай дупликации одного символа в двоичном алфавите и получена асимптотическая верхняя граница на мощность такого кода. Асимптотически оптимальные коды для исправления фиксированного числа ошибок дупликации одного символа, а также коды с эффективными алгоритмами кодирования и декодирования предложены в работах [71, 72]. В [73] выведена оптимальная конструкция для исправления неограниченного числа дупликаций фиксированной длины. Кроме того, в этой работе проведено обобщение на случай тандемных дупликаций, при которых фрагмент преобразуется в две последовательные копии. Для этого случая была представлена конструкция, исправляющая тандемные дупликации длины до 1, 2 или 3. Оптимальность этих кодов показана в работе [74]. В работе [75] рассмотрена задача исправления одной тандемной дупликации, но уже ограниченной длины. Авторами предложены границы на мощность таких кодов, а также явные конструкции на основе кодов Варшамова — Тененгольца, не являющиеся оптимальными.

В заключение раздела отметим, что в случае систем ДНК-памяти конфигурации и доля возможных ошибок определяются их характери-

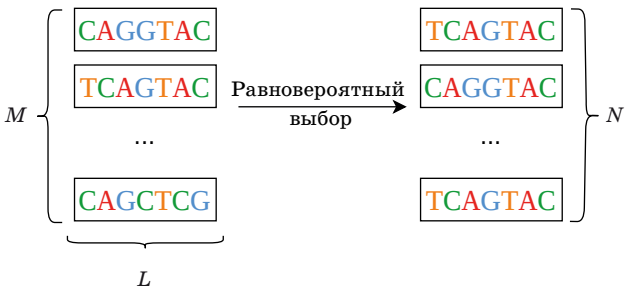
стиками. Это ставит задачу построения для них оптимальных кодов, обладающих вычислительно-эффективными алгоритмами кодирования и декодирования. Помимо этого, важной является задача дальнейшего улучшения существующих кодовых конструкций, в частности кодов, исправляющих несколько пакетных выпадений фиксированной длины, и кодов, исправляющих тандемные дупликации ограниченной длины.

Модели каналов для систем ДНК-памяти

Как отмечалось ранее, при хранении информации с использованием ДНК пользовательские данные преобразуются в большое число олигонуклеотидов. В ходе манипуляций с ними, а именно их синтеза и чтения (секвенирования), возникают ошибки, что приводит к появлению новых олигонуклеотидов. При этом даже в процессе хранения олигонуклеотиды могут быть подвержены химическим преобразованиям, которые приводят к «потере» некоторых из них на этапе чтения [37]. Наконец, в силу специфических свойств ПЦР распределение числа копий олигонуклеотидов весьма неравномерно: некоторые могут быть представлены намного большим числом копий (возможно, содержащих ошибки), чем другие. Это приводит к задаче моделирования всего процесса хранения информации с помощью ДНК-памяти, т. е. построению соответствующих ей теоретико-информационных моделей каналов, а также изучению их характеристик и кодов для них.

В одной из первых работ в этом направлении была рассмотрена модель, в которой на вход канала поступает набор из M последовательностей длины L. На выход, в свою очередь, поступает набор из N последовательностей, равновероятно выбранных из входных [76]. Такой канал проиллюстрирован на рис. 3.

Авторами работы получена оценка на пропускную способность данного канала. Отметим, что такая постановка задачи исключает возможные ошибки в последовательностях, а предположение о равновероятном выборе не вполне адекватно отражает реальность. В дальнейшем эта постановка была расширена в работе [77], в которой в последовательностях могли происходить ошибки замены. Схема кодов, достигающих пропускной способности, основанная на случайном кодировании и декодере, группирующем последовательности в соответствии с расстоянием Хэмминга между ними, представлена в работе [78].



■ ***Рис. 3.*** Пример модели канала ДНК-памяти
■ ***Fig. 3.*** Example of DNA-storage channel model

Упрощенная постановка вышеописанных задач рассмотрена в работах [79, 80], где последовательности лишь перемешивались без проведения процедуры случайного выбора. При этом были рассмотрены варианты данных каналов с ошибками замены и стираниями, получены значения пропускной способности для них, а также приведены схемы, с помощью которых они достигаются. В последних применялись кодирования порядка с помощью индексов, а также коды, достигающие пропускной способности для двоично-симметричного канала или же двоично-стирающего канала. Отметим, что в этих работах не рассматривались ошибки типа вставки/выпадения, для которых до сих пор неизвестны точные значения пропускной способности для всех областей возможных параметров [81], а также способы получения выходных последовательностей, отличные от равновероятного выбора.

Еще одной возможной моделью каналов для ДНК-памяти является модель, представленная в работе [82]. В ней на вход канала поступает набор из M последовательностей длины L. На выход канала поступает набор из $M - s$ последовательностей, из которых в t последовательностях происходит до e ошибок типа вставки/выпадения, а также замены. Параметры данной модели являются фиксированными величинами. Интересным фактом является то, что в рассматриваемых каналах код, исправляющий выпадения, не способен исправлять и вставки, и выпадения. Это отличает рассматриваемую задачу от задачи исправления ошибок в символах последовательностей, рассмотренной в предыдущем разделе. Авторами работы [82] выведены границы на избыточность данных кодов, а также представлены конструкции на основе индексации и кодирования всех информационных подпоследовательностей кодом с максимально достижимым кодовым расстоянием. Также предложен способ дальнейшего уменьшения избыточности с помощью использования наиболее значимых бит в индексах. Отметим, что данная модель не учитывает неравномерность распределения числа копий различных последовательностей, а также различия в вероятностях разных типов ошибок.

Интересной теоретико-информационной моделью каналов для ДНК-памяти также являет-

ся модель, представленная в работе [83]. В ней, как и в предыдущем случае, на вход поступает набор из M последовательностей длины L. На выход же канала поступает набор из T последовательностей, в которых произошло K замен, где $M - K \leq T \leq M$. При этом считается, что возможная потеря последовательностей является следствием произошедших в них ошибок замены. С помощью границы для упаковки сфер авторами получена граница для существования кодов, исправляющих ошибки в таком канале. Кроме того, была представлена явная конструкция, исправляющая несколько произошедших замен, основанная на идее конкатенации последовательностей в специальном порядке и применении кода Рида — Соломона. Существенным недостатком данной модели является отсутствие учета возможных ошибок типа вставки/выпадения.

Кроме описанных, еще одной недавно предложенной моделью канала для ДНК-памяти [84] является модель, в которой на вход канала поступает набор из M последовательностей длины L, а на выходе получается набор из уже $\tilde{M}$ последовательностей, часть из которых добавлена дополнительно. При этом часть исходных последовательностей может отсутствовать на выходе канала, а в части могут происходить ошибки замены. Отметим, что, в отличие от предыдущих постановок, M не обязательно является фиксированной величиной. Для работы в данном канале авторами была введена новая метрика, называемая последовательно-множественным расстоянием, и предложены способы исправления ошибок в ней на основе кодов постоянного веса и кодирования индексов. Подчеркнем, что, как и в предыдущем случае, авторы не рассматривают возможные ошибки типа вставки/выпадения.

В заключение добавим, что построение новых теоретико-информационных моделей каналов ДНК-памяти, более точно учитывающих неравномерность числа копий различных последовательностей и возможные типы ошибок в них, по-прежнему является важной теоретической задачей.

Заключение

Системы ДНК-памяти являются перспективным способом хранения пользовательской информации, но на сегодня они еще далеки от практического внедрения. Этапы «записи» и «чтения» информации пока остаются более медленными и более дорогими в сравнении с традиционными носителями. Кроме того, проблему борьбы с различными ошибками, возникающими на этапах синтеза, хранения и секвенирования ДНК, нельзя считать полностью решенной. Однако на этом пути в последние годы достигнут ощутимый прогресс, что и явилось главной темой настоящего обзора. Как часто случается, возникновение новой области приложений дает импульс для появления и развития новых моделей, алгоритмов и теоретических результатов, мотивированных новыми постановками задач. В данном случае это относится к теории информации и теории кодирования. Зачастую эти результаты представляют независимый интерес, что мы также попытались отразить в настоящем обзоре. Скажем в заключение, что обзор посвящен теоретико-информационным аспектам и не претендует на полное освещение этой новой области на стыке теории информации и биоинформатики.

Финансовая поддержка

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов 19-01-00364, 20-07-00652 и 20-17-50170.

Литература

1. **Aftab U., Siddiqui G. F.** Big data augmentation with data warehouse: a survey. *Proceedings of 2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 2775–2784. doi:10.1109/BigData.2018.8622182
2. **Lee S. Y.** DNA data storage is closer than you think. 2019. https://www.scientificamerican.com/article/dna-data-storage-is-closer-than-you-think/ (дата обращения: 12.03.2021).
3. **Reinsel D., Gantz J., Rydning J.** The digitization of the world from edge to core. *An IDC White Paper US44413318*, 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (дата обращения: 12.03.2021).
4. **Bohannon J.** DNA: The ultimate hard drive. *Science*, 2012. https://www.sciencemag.org/news/2012/08/dna-ultimate-hard-drive (дата обращения: 12.03.2021).
5. **Colen C.** DNA data storage — setting the data density record with DNA fountain. 2017. https://www.twistbioscience.com/blog/perspectives/dna-data-storage-setting-data-density-record-dna-fountain (дата обращения: 12.03.2021).
6. **Ceze L., Nivala J., Strauss K.** Molecular digital data storage using DNA. *Nature Reviews Genetics*, 2019, vol. 20, pp. 456–466. doi:10.1038/s41576-019-0125-3
7. **Extance A.** How DNA could store all the world's data. *Nature*, 2016, vol. 537, pp. 22–24. doi:10.1038/537022a
8. **Zhirnov V., Zadegan R. M., Sandhu G. S., Church G. M., Hughes W. L.** Nucleic acid memory.

Nature Material, 2016, vol. 15, pp. 366–370. doi:10.1038/nmat4594
9. **Watson J. D., Crick F. H.** Molecular structure of nucleic acids. *Nature*, 1953, vol. 171, pp. 737–738. doi:10.1038/171737a0
10. **Wiener N.** Interview: machines smarter than men? *US News World Rep*, 1964, vol. 56, pp. 84–86.
11. **Нейман М. С.** Некоторые фундаментальные вопросы микроминиатюризации. *Радиотехника*, 1964, т. 12, № 1, с. 3–12. https://2a008ed5-a-62cb3a1a-s-sites.googlegroups.com/site/msneiman1905/Neiman-1964_Micromini.pdf (дата обращения: 12.03.2021).
12. **Нейман М. С.** О молекулярных системах памяти и направлениях мутаций. *Радиотехника*, 1965, т. 20, № 6, с. 1–8. https://2a008ed5-a-62cb3a1a-s-sites.googlegroups.com/site/msneiman1905/Neiman-1965_Molecul.pdf (дата обращения: 12.03.2021).
13. **Реброва И. М., Реброва О. Ю.** Запоминающие устройства на основе искусственной ДНК: рождение идеи и первые публикации. *Вопросы истории естествознания и техники*, 2019, т. 41, № 4, с. 666–676. doi:10.31857/S020596060013006-8
14. **Davis J.** Microvenus. *Art Journal*, 1996, vol. 55, no. 1, pp. 70–74. doi:10.2307/777811
15. **Clelland C. T., Risca V., Bancroft C.** Hiding messages in DNA microdots. *Nature*, 1999, vol. 399, pp. 533–534. doi:10.1038/21092
16. **Bancroft C., Bowler T., Bloom B.** Long-term storage of information in DNA. *Science*, 2001, vol. 293, iss. 5536, pp. 1763–1765. doi:10.1126/science.293.5536.1763c
17. **Church G. M., Gao Y., Kosuri S.** Next-generation digital information storage in DNA. *Science*, 2012, vol. 337, pp. 1628. doi:10.1126/science.1226355
18. **Goldman N., Bertone P., Chen S., Dessimoz C., LeProust E. M., Sipos B., Birney E.** Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 2013, vol. 494, pp. 77–79. doi:10.1038/nature11875.
19. **Organick L., Ang S. D., Chen Y.,** et al. Random access in large-scale DNA data storage. *Nature Biotechnology*, 2018, vol. 36, pp. 242–248. doi:10.1038/nbt.4079
20. **Shankland S.** Startup packs all 16GB of Wikipedia onto DNA strands to demonstrate new storage tech. *Cnet*, 2019. https://www.cnet.com/news/startup-packs-all-16gb-wikipedia-onto-dna-strands-demonstrate-new-storage-tech/ (дата обращения: 12.03.2021).
21. **Extance A.** How DNA could store all the world's data. *Nature*, 2016, vol. 537, pp. 22–24. doi:10.1038/537022a
22. **Service R.** DNA could store all of the world's data in one room. *Science*, 2017. doi:10.1126/science.aal0852. https://www.sciencemag.org/news/2017/03/dna-could-store-all-worlds-data-one-room (дата обращения: 12.03.2021).
23. **Erlich Y., Zielinski D.** DNA fountain enables a robust and efficient storage architecture. *Science*, 2017, vol. 355, iss. 6328, pp. 950–954. doi:10.1126/science.aa j2038
24. **Yazdi H. T., Kiah H. M., Garcia-Ruiz E., Ma J., Zhao H., Milenkovic O.** DNA-based storage: trends and methods. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2015, vol. 1, no. 3, pp. 230–248. doi:10.1109/TMBMC.2016.2537305
25. **Yiming D., Fajia S., Zhi P., Qi O., Long Q.** DNA storage: research landscape and future prospects. *National Science Review*, 2020, vol. 7, iss. 6, pp. 1092–1104. doi:10.1093/nsr/nwaa007
26. **Yim S. S., McBee R. M., Song A. M., Huang Y., Sheth R. U., Wang H. H.** Robust direct digital-to-biological data storage in living cells. *Nature Chemical Biology*, 2021, vol. 17, pp. 246–253. doi:10.1038/s41589-020-00711-4
27. **Gilbert E.** Synchronization of binary messages. *IRE Transactions on Information Theory*, 1960, vol. 6, no. 4, pp. 470–477. doi:10.1109/TIT.1960.1057587
28. **Левенштейн В. И.** Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады АН СССР*, 1965, т. 163, с. 845–848.
29. **Sellers F.** Bit loss and gain correction code. *IRE Transactions on Information Theory*, 1962, vol. 8, no. 1, pp. 35–38. doi:10.1109/TIT.1962.1057684
30. **Xiong Z.** Multiterminal source coding: theory, code design and applications. *Proceedings of the Fifth International Workshop on Signal Design and its Applications in Communications*, IEEE, 2011, pp. 3–3. doi:10.1109/IWSDA.2011.6159430
31. **Luby M.** LT codes. *The 43rd Annual IEEE Symposium on Foundations of Computer Science*, IEEE, 2002, pp. 271–280. doi:10.1109/SFCS.2002.1181950
32. **Niedringhaus T. P., Milanova D., Kerby M. B., Snyder M. P., Barron A. E.** Landscape of next-generation sequencing technologies. *Analytical Chemistry*, 2011, vol. 83, no. 12, pp. 4327–4341. doi:10.1021/ac2010857
33. **Schwartz J. J., Lee C., Shendure J.** Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nature Methods*, 2012, vol. 9, pp. 913–915. doi:10.1038/nmeth.2137
34. **Yachie N., Sekiyama K., Sugahara J., Ohashi Y., Tomita M.** Alignment-based approach for durable data storage into living organisms. *Biotechnol Progress*, 2007, vol. 23, no. 2, pp. 501–505. doi:10.1021/bp060261y
35. **Itaya M., Tsuge K., Koizumi M., Fujita K.** Combining two genomes in one cell: stable cloning of the synechocystis PCC6803 genome in the bacillus subtilis 168 genome. *Proceedings of the National Academy of Sciences of the USA*, 2005, vol. 102, no. 44, pp. 15971–15976. doi:10.1073/pnas.0503868102
36. **Kosuri S., Church G. M.** Large-scale de novo DNA synthesis: technologies and applications. *Nature Methods*, 2014, vol. 11, no. 5, pp. 499–507. doi:10.1038/nmeth.2918
37. **Heckel R., Mikutis G., Grass R. N.** A characterization of the DNA data storage channel. *Science Reports*, 2019, vol. 9, pp. 1–10. doi:10.1038/s41598-019-45832-6

38. **Schmidt T., Beliveau B., Uca Y., Theilmann M., Cruz F. D., Wu C.-T., Shih W. M.** Scalable amplification of strand subsets from chip-synthesized oligonucleotide libraries. *Nature Communications*, 2015, vol. 6, pp. 1–11. doi:10.1038/ncomms963439.
39. **Lee H. H., Kalhor R., Goela N., Bolot J., Church G. M.** Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nature Communications*, 2019, vol. 10, pp. 1–12. doi:10.1038/s41467-019-10258-1
40. **Kulski J. K.** Next-generation Sequencing — an overview of the history, tools, and "omic" applications. Next generation sequencing — advances, applications and challenges. *IntechOpen*, 2016, pp. 1–15. doi:10.5772/61964
41. **Organick L., Chen Y.-J., Ang S. D., Lopez R., Liu X., Strauss K., Ceze L.** Probing the physical limits of reliable DNA data retrieval. *Nature Communications*, 2020, vol. 11, pp. 1–12. doi:10.1038/s41467-020-14319-8
42. **Chandak S., Tatwawadi K., Lau B., Mardia J., Kubit M., Neu J., Griffin P., Wootters M., Weissman T., Ji H.** Improved read/write cost tradeoff in DNA-based data storage using LDPC codes. *Proceedings of 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2019, pp. 147–156. doi:10.1109/ALLERTON.2019.8919890
43. **Grass R. N., Heckel R., Puddu M., Paunescu D., Stark W. J.** Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie*, 2015, vol. 54, pp. 2552–2555. doi:10.1002/anie.201411378
44. **Anavy L., Vaknin I., Atar O., Amit R., Yakhini Z.** Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nature Biotechnology*, 2019, vol. 37, pp. 1229–1236. doi:10.1038/s41587-019-0240-x
45. **Meiser L. C., Antkowiak P. L., Koch J., Chen W. D., Kohll A. X., Stark W. J., Heckel R., Grass R. N.** Reading and writing digital data in DNA. *Nature Protocols*, 2020, vol. 15, pp. 86–101. doi:10.1038/s41596- 019-0244-5
46. **Lopez R., Chen Y.-J., Ang D. S., Yekhanin S., Makarychev K., Racz M. Z., Seelig G., Strauss K., Ceze L.** DNA assembly for nanopore data storage readout. *Nature Communications*, 2019, vol. 10, pp. 1–9. doi:10.1038/s41467-019-10978-4
47. **Варшамов Р. Р., Тенengольц Г. М.** Код, исправляющий одиночные несимметричные ошибки. *Автоматика и телемеханика*, 1965, т. 26, № 2, с. 286–290.
48. **Saowapa K., Kaneko H., Fujiwara E.** Systematic binary deletion/insertion error-correcting codes capable of correcting random bit errors. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2000, vol. E83-A, no. 12, pp. 2699–2705.
49. **Левенштейн В. И.** Асимптотически оптимальный двоичный код с исправлением выпадений одного или двух соседних символов. *Проблемы кибернетики*, 1967, т. 19, с. 298–304.
50. **Helberg A. S. J., Ferreira H. C.** On multiple insertion/deletion correcting codes. *IEEE Transactions on Information Theory*, 2002, vol. 48, no. 1, pp. 305–308. doi:10.1109/18.971760
51. **Swart T. G., Ferreira H. C.** A note on double insertion/deletion correcting codes. *IEEE Transactions on Information Theory*, 2003, vol. 49, no. 1, pp. 269–273. doi:10.1109/TIT.2002.806155
52. **Guruswami V., Hastad J.** Explicit two-deletion codes with redundancy matching the existential bound. *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, SIAM, 2021, pp. 1–8. doi:10.1137/1.9781611976465.2
53. **Brakensiek J., Guruswami V., Zbarsky S.** Efficient low-redundancy codes for correcting multiple deletions. *IEEE Transactions on Information Theory*, 2018, vol. 64, no. 5, pp. 3403–3410. doi:10.1109/TIT.2017.2746566
54. **Guruswami V., Wang C.** Deletion codes in the high-noise and high-rate regimes. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 4, pp. 1961–1970. doi:10.1109/TIT.2017.2659765
55. **Sima J., and Bruck J.** On optimal k-deletion correcting codes. *IEEE Transactions on Information Theory*, 2021. doi:10.1109/TIT.2020.3028702
56. **Sima J., Gabrys R., and Bruck J.** Optimal systematic t-deletion correcting codes. *Proceedings of 2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 769–774. doi:10.1109/ISIT44484.2020.9173986
57. **Schulman L. J., Zuckerman D.** Asymptotically good codes correcting insertions, deletions, and transpositions. *IEEE Transactions on Information Theory*, 1999, vol. 45, no. 7, pp. 2552–2557. doi:10.1109/18.796406
58. **Kulkarni A. A., Kiyavash N.** Nonasymptotic upper bounds for deletion correcting codes. *IEEE Transactions on Information Theory*, 2013, vol. 59, no. 8, pp. 5115–5130. doi:10.1109/TIT.2013.2257917
59. **Bukh B., Guruswami V., and Hastad J.** An improved bound on the fraction of correctable deletions. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 1, pp. 93–103. doi:10.1109/TIT.2016.2621044
60. **Guruswami V., and Li R.** Efficiently decodable insertion/deletion codes for high-noise and high-rate regimes. *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2016, pp. 620–624. doi:10.1109/ISIT.2016.7541373
61. **Haeupler B., Shahrasbi A.** Synchronization strings: codes for insertions and deletions approaching the Singleton bound. *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2017)*, ACM, 2017, pp. 33–46. doi:10.1145/3055399.3055498
62. **Cheng K., Haeupler B., Li X., Shahrasbi A., Wu K.** Synchronization strings: highly efficient determinis-

tic constructions over small alphabets. *Proceedings of the 2019 Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM, 2019, pp. 2185–2204. doi:10.1137/1.9781611975482.132

63. **Cheng K., Jin Z., Li X., and Wu K.** Deterministic document exchange protocols, and almost optimal binary codes for edit errors. *Proceedings of 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 2018, pp. 200–211. doi:10.1109/FOCS.2018.00028
64. **Haeupler B.** Optimal document exchange and new codes for insertions and deletions. *Proceedings 2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 2019, pp. 334–347. doi:10.1109/FOCS.2019.00029
65. **Schoeny C., Wachter-Zeh A., Gabrys R., Yaakobi E.** Codes correcting a burst of deletions or insertions. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 4, pp. 1971–1985. doi:10.1109/TIT.2017.2661747
66. **Saeki T., Nozaki T.** An improvement of non-binary code correcting single b-burst of insertions or deletions. *Proceedings of 2018 International Symposium on Information Theory and its Applications (ISITA)*, IEICE, 2018, pp. 6–10. doi:10.23919/ISITA.2018.8664217
67. **Gabrys R., Yaakobi E., Milenkovic O.** Codes in the damerau distance for deletion and adjacent transposition correction. *IEEE Transactions on Information Theory*, 2018, vol. 64, no. 4, pp. 2550–2570. doi:10.1109/TIT.2017.2778143
68. **Lenz A., Polyanskii N.** Optimal codes correcting a burst of deletions of variable length. *2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 757–762. doi:10.1109/ISIT44484.2020.9174288
69. **Sima J., Gabrys R., and Bruck J.** Syndrome compression for optimal redundancy codes. *Proceedings of 2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 751–756. doi:10.1109/ISIT44484.2020.9174009
70. **Левенштейн В. И.** Двоичные коды с исправлением вставок и выпадений символа 1. *Проблемы передачи информации*, 1965, т. 1, № 1, с. 8–17.
71. **Dolecek L., Anantharam V.** Repetition error correcting sets: explicit constructions and prefixing methods. *SIAM Journal on Discrete Mathematics*, 2010, vol. 23, no. 4, pp. 2120–2146. doi:10.5555/1958171.1958195
72. **Mahdavifar H., Vardy A.** Asymptotically optimal sticky-insertion correcting codes with efficient encoding and decoding. *Proceedings of 2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 2688–2692. doi:10.1109/ISIT.2017.8007016
73. **Jain S., Farnoud F., Schwartz M., Bruck J.** Duplication-correcting codes for data storage in the DNA of living organisms. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 8, pp. 4996–5010. doi:10.1109/ISIT.2016.7541455
74. **Kovacevic M.** Codes correcting all patterns of tandem-duplication errors of maximum length 3. https://arxiv.org/pdf/1911.06561.pdf (дата обращения: 12.03.2021).
75. **Nazirkhanova K., Medova L., Kruglik S., and Frolov A.** Codes correcting bounded length tandem duplication. *Proceedings of 2020 International Symposium on Information Theory and its Applications (ISITA)*, IEICE, 2020, pp. 299–303. doi:10.34385/proc.65.B06-6
76. **Heckel R., Shomorony I., Ramchandran K., Tse D. N. C.** Fundamental limits of DNA storage systems. *Proceedings of 2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 3130–3134. doi:10.1109/ISIT.2017.8007106
77. **Lenz A., Siegel P. H., Wachter-Zeh A., Yaakobi E.** An upper bound on the capacity of the DNA storage channel. *2019 IEEE Information Theory Workshop (ITW)*, IEEE, 2019, pp. 1–5. doi:10.1109/ITW44776.2019.8989388
78. **Lenz A., Siegel P. H., Wachter-Zeh A., Yaakobi E.** Achieving the capacity of the DNA storage channel. *Proceedings of ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal*, IEEE, 2020, pp. 8846–8850. doi:10.1109/ICASSP40776.2020.9053049
79. **Shin S., Heckel R., Shomorony I.** Capacity of the erasure shuffling channel. *Proceedings of ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal*, IEEE, 2020, pp. 8841–8845. doi:10.1109/ICASSP40776.2020.9053486
80. **Shomorony I., Heckel R.** Capacity results for the noisy shuffling channel. *Proceedings of 2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2019, pp. 762–766. doi:10.1109/ISIT.2019.8849789
81. **Mitzenmacher M.** A survey of results for deletion channels and related synchronization channels. *Probability Surveys*, 2009, vol. 6, pp. 1–33. doi:10.1214/08-PS141
82. **Lenz A., Siegel P. H., Wachter-Zeh A., Yaakobi E.** Coding over sets for DNA storage. *IEEE Transactions on Information Theory*, 2020, vol. 66, no. 4, pp. 2331–2351. doi:10.1109/TIT.2019.2961265
83. **Sima J., Raviv N., and Bruck J.** On coding over sliced information. *IEEE Transactions on Information Theory*, 2021. doi:10.1109/TIT.2021.3063709
84. **Song W., Cai K., Schouhamer Immink K. A.** Sequence-subset distance and coding for error control in DNA-based data storage. *IEEE Transactions on Information Theory*, 2020, vol. 66, no. 10, pp. 6048–6065. doi:10.1109/TIT.2020.3002611

UDC 004.074
doi:10.31799/1684-8853-2021-3-39-52

Information-theoretic problems of DNA-based storage systems

S. A. Kruglik[a,b], Junior Researcher, orcid.org/0000-0001-9557-5197,
stanislav.kruglik@skoltech.ru
G. A. Kucherov[a,c], PhD, Phys.-Math., Leading Researcher, orcid.org/0000-0001-5899-5424
K. N. Nazirkhanova[d], Post-Graduate Student, orcid.org/0000-0002-7447-9857
M. E. Filitov[a], Master Student, orcid.org/0000-0003-2421-0777
[a]Skolkovo Institute of Science and Technology, bld. 1, 30, Bolshoy Boulevard, 121205, Moscow, Russian Federation
[b]Moscow Institute of Physics and Technology, 9, Institutskiy Per., 141701, Dolgoprudny, Moscow region, Russian Federation
[c]Centre National de Recherche Scientifique, Université Gustave Eiffel, 77454 Marne-la-Vallée, France
[d]Stanford University, 94305 Stanford, CA, USA

Introduction: Currently, we witness an explosive growth in the amount of information produced by humanity. This raises new fundamental problems of its efficient storage and processing. Commonly used magnetic, optical, and semiconductor information storage devices have several drawbacks related to small information density and limited durability. One of the promising novel approaches to solving these problems is DNA-based data storage. **Purpose:** An overview of modern DNA-based storage systems and related information-theoretic problems. **Results:** The current state of the art of DNA-based storage systems is reviewed. Types of errors occurring in them as well as corresponding error-correcting codes are analized. The disadvantages of these codes are shown, and possible pathways for improvement are mentioned. Proposed information-theoretic models of DNA-based storage systems are analyzed, and their limitation highlighted. In conclusion, main obstacles to practical implementation of DNA-based storage systems are formulated, which can be potentially overcome using information-theoretic methods considered in this overview.

Keywords — data storage systems, DNA-based memory, communication channel, channel capacity, substitution errors, insertion errors, deletion errors.

For citation: Kruglik S. A., Kucherov G. A., Nazirkhanova K. N., Filitov M. E. Information-theoretic problems of DNA-based storage systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 3, pp. 39–52 (In Russian). doi:10.31799/1684-8853-2021-3-39-52

Funding
The reported study was funded by RFBR, projects 19-01-00364, 20-07-00652 and 20-17-50170.

References

1. Aftab U., Siddiqui G. F. Big data augmentation with data warehouse: a survey. *Proceedings of 2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 2775–2784. doi:10.1109/BigData.2018.8622182
2. Lee S. Y. *DNA data storage is closer than you think*. 2019. Available at: https://www.scientificamerican.com/article/dna-data-storage-is-closer-than-you-think/ (accessed 12 March 2021).
3. Reinsel D., Gantz J., Rydning J. The digitization of the world from edge to core. *An IDC White Paper US44413318*, 2018. Available at: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (accessed 12 March 2021).
4. Bohannon J. DNA: The ultimate hard drive. *Science*, 2012. Available at: https://www.sciencemag.org/news/2012/08/dna-ultimate-hard-drive (accessed 12 March 2021).
5. Colen C. *DNA data storage — setting the data density record with DNA fountain*. 2017. Available at: https://www.twistbioscience.com/blog/perspectives/dna-data-storage-setting-data-density-record-dna-fountain (accessed 12 March 2021).
6. Ceze L., Nivala J., Strauss K. Molecular digital data storage using DNA. *Nature Reviews Genetics*, 2019, vol. 20, pp. 456–466. doi:10.1038/s41576-019-0125-3
7. Extance A. How DNA could store all the world's data. *Nature*, 2016, vol. 537, pp. 22–24. doi:10.1038/537022a
8. Zhirnov V., Zadegan R. M., Sandhu G. S., Church G. M., Hughes W. L. Nucleic acid memory. *Nature Material*, 2016, vol. 15, pp. 366–370. doi:10.1038/nmat4594
9. Watson J. D., Crick F. H. Molecular structure of nucleic acids. *Nature*, 1953, vol. 171, pp. 737–738. doi:10.1038/171737a0
10. Wiener N. Interview: machines smarter than men? *US News World Rep*, 1964, vol. 56, pp. 84–86.
11. Neiman M. S. Some fundamental issues of microminiaturization. *Radiotekhnika*, 1964, vol. 12, no. 1, pp. 3–12 (In Russian). Available at: https://2a008ed5-a-62cb3a1a-s-sites.googlegroups.com/site/msneiman1905/Neiman-1964_Micromini.pdf (accessed 12 March 2021).
12. Neiman M. S. On the molecular memory systems and the directed mutations. *Radiotekhnika*, 1965, vol. 20, no. 6, pp. 1–8 (In Russian). Available at: https://2a008ed5-a-62cb3a1a-s-sites.googlegroups.com/site/msneiman1905/Neiman-1965_Molecul.pdf (accessed 12 March 2021).
13. Rebrova I. M., Rebrova O. Y. Synthesized DNA-based data storage devices: the birth of the idea and the first publications. *Studies in the History of Science and Technology*, 2019, vol. 41, no. 4, pp. 666–676 (In Russian). doi:10.31857/S020596060013006-8
14. Davis J. Microvenus. *Art Journal*, 1996, vol. 55, no. 1, pp. 70–74. doi:10.2307/777811
15. Clelland C. T., Risca V., Bancroft C. Hiding messages in DNA microdots. *Nature*, 1999, vol. 399, pp. 533–534. doi:10.1038/21092
16. Bancroft C., Bowler T., Bloom B. Long-term storage of information in DNA. *Science*, 2001, vol. 293, iss. 5536, pp. 1763–1765. doi:10.1126/science.293.5536.1763c
17. Church G. M., Gao Y., Kosuri S. Next-generation digital information storage in DNA. *Science*, 2012, vol. 337, pp. 1628. doi:10.1126/science.1226355
18. Goldman N., Bertone P., Chen S., Dessimoz C., LeProust E. M., Sipos B., Birney E. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 2013, vol. 494, pp. 77–79. doi:10.1038/nature11875.
19. Organick L., Ang S. D., Chen Y., et al. Random access in large-scale DNA data storage. *Nature Biotechnology*, 2018, vol. 36, pp. 242–248. doi:10.1038/nbt.4079
20. Shankland S. Startup packs all 16GB of Wikipedia onto DNA strands to demonstrate new storage tech. *Cnet*, 2019. Available at: https://www.cnet.com/news/startup-packs-all-16gb-wikipedia-onto-dna-strands-demonstrate-new-storage-tech/ (accessed 12 March 2021).
21. Extance A. How DNA could store all the world's data. *Nature*, 2016, vol. 537, pp. 22–24. doi:10.1038/537022a
22. Service R. DNA could store all of the world's data in one room. *Science*, 2017. doi:10.1126/science.aal0852 Available at: https://www.sciencemag.org/news/2017/03/dna-

could-store-all-worlds-data-one-room (accessed 12 March 2021).
23. Erlich Y., Zielinski D. DNA fountain enables a robust and efficient storage architecture. *Science*, 2017, vol. 355, iss. 6328, pp. 950–954. doi:10.1126/science.aa j2038
24. Yazdi H. T., Kiah H. M., Garcia-Ruiz E., Ma J., Zhao H., Milenkovic O. DNA-based storage: trends and methods. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2015, vol. 1, no. 3, pp. 230–248. doi:10.1109/TMBMC.2016.2537305
25. Yiming D., Fajia S., Zhi P., Qi O., Long Q. DNA storage: research landscape and future prospects. *National Science Review*, 2020, vol. 7, iss. 6, pp. 1092–1104. doi:10.1093/nsr/nwaa007
26. Yim S. S., McBee R. M., Song A. M., Huang Y., Sheth R. U., Wang H. H. Robust direct digital-to-biological data storage in living cells. *Nature Chemical Biology*, 2021, vol. 17, pp. 246-253. doi:10.1038/s41589-020-00711-4
27. Gilbert E. Synchronization of binary messages. *IRE Transactions on Information Theory*, 1960, vol. 6, no. 4, pp. 470–477. doi:10.1109/TIT.1960.1057587
28. Levinstein V. I. Binary codes for the correction of deletions insertions, and changes of symbols. *Reports of the USSR Academy of Sciences*, 1965, vol. 163, pp. 845–848.
29. Sellers F. Bit loss and gain correction code. *IRE Transactions on Information Theory*, 1962, vol. 8, no. 1, pp. 35–38. doi:10.1109/TIT.1962.1057684
30. Xiong Z. Multiterminal source coding: theory, code design and applications. *Proceedings of the Fifth International Workshop on Signal Design and its Applications in Communications*, IEEE, 2011, pp. 3-3. doi:10.1109/IWSDA.2011.6159430
31. Luby M. LT codes. *The 43rd Annual IEEE Symposium on Foundations of Computer Science*, IEEE, 2002, pp. 271–280. doi:10.1109/SFCS.2002.1181950
32. Niedringhaus T. P., Milanova D., Kerby M. B., Snyder M. P., Barron A. E. Landscape of next-generation sequencing technologies. *Analytical Chemistry*, 2011, vol. 83, no. 12, pp. 4327–4341. doi:10.1021/ac2010857
33. Schwartz J. J., Lee C., Shendure J. Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nature Methods*, 2012, vol. 9, pp. 913–915. doi:10.1038/nmeth.2137
34. Yachie N., Sekiyama K., Sugahara J., Ohashi Y., Tomita M. Alignment-based approach for durable data storage into living organisms. *Biotechnol Progress*, 2007, vol. 23, no. 2, pp. 501–505. doi:10.1021/bp060261y
35. Itaya M., Tsuge K., Koizumi M., Fujita K. Combining two genomes in one cell: stable cloning of the synechocystis PCC6803 genome in the bacillus subtilis 168 genome. *Proceedings of the National Academy of Sciences of the USA*, 2005, vol. 102, no. 44, pp. 15971–15976. doi:10.1073/pnas.0503868102
36. Kosuri S., Church G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nature Methods*, 2014, vol. 11, no. 5, pp. 499–507. doi:10.1038/nmeth.2918
37. Heckel R., Mikutis G., Grass R. N. A characterization of the DNA data storage channel. *Science Reports*, 2019, vol. 9, pp. 1–10. doi:10.1038/s41598-019-45832-6
38. Schmidt T., Beliveau B., Uca Y., Theilmann M., Cruz F. D., Wu C.-T., Shih W. M. Scalable amplification of strand subsets from chip-synthesized oligonucleotide libraries. *Nature Communications*, 2015, vol. 6, pp. 1–11. doi:10.1038/ncomms9634
39. Lee H. H., Kalhor R., Goela N., Bolot J., Church G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nature Communications*, 2019, vol. 10, pp. 1–12. doi:10.1038/s41467-019-10258-1
40. Kulski J. K. Next-generation Sequencing — an overview of the history, tools, and "omic" applications. Next generation sequencing — advances, applications and challenges. *IntechOpen*, 2016, pp. 1–15. doi:10.5772/61964
41. Organick L., Chen Y.-J., Ang S. D., Lopez R., Liu X., Strauss K., Ceze L. Probing the physical limits of reliable DNA data retrieval. *Nature Communications*, 2020, vol. 11, pp. 1–12. doi:10.1038/s41467-020-14319-8
42. Chandak S., Tatwawadi K., Lau B., Mardia J., Kubit M., Neu J., Griffin P., Wootters M., Weissman T., Ji H. Improved read/write cost tradeoff in DNA-based data storage using LDPC codes. *Proceedings of 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2019, pp. 147–156. doi:10.1109/ALLERTON.2019.8919890
43. Grass R. N., Heckel R., Puddu M., Paunescu D., Stark W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie*, 2015, vol. 54, pp. 2552–2555. doi:10.1002/anie.201411378
44. Anavy L., Vaknin I., Atar O., Amit R., Yakhini Z. Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nature Biotechnology*, 2019, vol. 37, pp. 1229–1236. doi:10.1038/s41587-019-0240-x
45. Meiser L. C., Antkowiak P. L., Koch J., Chen W. D., Kohll A. X., Stark W. J., Heckel R., Grass R. N. Reading and writing digital data in DNA. *Nature Protocols*, 2020, vol. 15, pp. 86–101. doi:10.1038/s41596- 019-0244-5
46. Lopez R., Chen Y.-J., Ang S. G., Yekhanin S., Makarychev K., Racz M. Z., Seelig G., Strauss K., Ceze L. DNA assembly for nanopore data storage readout. *Nature Communications*, 2019, vol. 10, pp. 1–9. doi:10.1038/s41467-019-10978-4
47. Varshamov R. R., Tenengolts G. M. Codes which correct single asymmetric errors. *Automation and Remote Control*, 1965, vol. 26, no. 2, pp. 286–290.
48. Saowapa K., Kaneko H., Fujiwara E. Systematic binary deletion/insertion error-correcting codes capable of correcting random bit errors. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2000, vol. E83-A, no. 12, pp. 2699–2705.
49. Levenshtein V. I. Asymptotically optimum binary codes with correction for losses of one or two adjacent bits. *Problems of Cybernetics*, 1967, vol. 19, pp. 298–304.
50. Helberg A. S. J., Ferreira H. C. On multiple insertion/deletion correcting codes. *IEEE Transactions on Information Theory*, 2002, vol. 48, no. 1, pp. 305–308. doi:10.1109/18.971760
51. Swart T. G., Ferreira H. C. A note on double insertion/deletion correcting codes. *IEEE Transactions on Information Theory*, 2003, vol. 49, no. 1, pp. 269–273. doi:10.1109/TIT.2002.806155
52. Guruswami V., Hastad J. Explicit two-deletion codes with redundancy matching the existential bound. *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, SIAM, 2021, pp. 1–8. doi:10.1137/1.9781611976465.2
53. Brakensiek J., Guruswami V., Zbarsky S. Efficient low-redundancy codes for correcting multiple deletions. *IEEE Transactions on Information Theory*, 2018, vol. 64, no. 5, pp. 3403–3410. doi:10.1109/TIT.2017.2746566
54. Guruswami V., Wang C. Deletion codes in the high-noise and high-rate regimes. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 4, pp. 1961–1970. doi:10.1109/TIT.2017.2659765
55. Sima J., and Bruck J. On optimal k-deletion correcting codes. *IEEE Transactions on Information Theory*, 2021. doi:10.1109/TIT.2020.3028702
56. Sima J., Gabrys R., and Bruck J. Optimal systematic t-deletion correcting codes. *Proceedings of 2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 769–774. doi:10.1109/ISIT44484.2020.9173986
57. Schulman L. J., Zuckerman D. Asymptotically good codes correcting insertions, deletions, and transpositions. *IEEE Transactions on Information Theory*, 1999, vol. 45, no. 7, pp. 2552–2557. doi:10.1109/18.796406
58. Kulkarni A. A., Kiyavash N. Nonasymptotic upper bounds for deletion correcting codes. *IEEE Transactions on Information Theory*, 2013, vol. 59, no. 8, pp. 5115–5130. doi:10.1109/TIT.2013.2257917
59. Bukh B., Guruswami V., and Hastad J. An improved bound on the fraction of correctable deletions. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 1, pp. 93–103. doi:10.1109/TIT.2016.2621044
60. Guruswami V., and Li R. Efficiently decodable insertion/deletion codes for high-noise and high-rate regimes. *Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2016, pp. 620–624. doi:10.1109/ISIT.2016.7541373
61. Haeupler B., Shahrasbi A. Synchronization strings: codes for insertions and deletions approaching the Singleton bound. *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2017)*, ACM, 2017, pp. 33–46. doi:10.1145/3055399.3055498
62. Cheng K., Haeupler B., Li X., Shahrasbi A., Wu K. Synchronization strings: highly efficient deterministic constructions over small alphabets. *Proceedings of the 2019 Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM, 2019, pp. 2185–2204. doi:10.1137/1.9781611975482.132

63. Cheng K., Jin Z., Li X., and Wu K. Deterministic document exchange protocols, and almost optimal binary codes for edit errors. *Proceedings of 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 2018, pp. 200–211. doi:10.1109/FOCS.2018.00028
64. Haeupler B. Optimal document exchange and new codes for insertions and deletions. *Proceedings 2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 2019, pp. 334–347. doi:10.1109/FOCS.2019.00029
65. Schoeny C., Wachter-Zeh A., Gabrys R., Yaakobi E. Codes correcting a burst of deletions or insertions. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 4, pp. 1971–1985. doi:10.1109/TIT.2017.2661747
66. Saeki T., Nozaki T. An improvement of non-binary code correcting single b-burst of insertions or deletions. *Proceedings of 2018 International Symposium on Information Theory and its Applications (ISITA)*, IEICE, 2018, pp. 6–10. doi:10.23919/ISITA.2018.8664217
67. Gabrys R., Yaakobi E., Milenkovic O. Codes in the damerau distance for deletion and adjacent transposition correction. *IEEE Transactions on Information Theory*, 2018, vol. 64, no. 4, pp. 2550–2570. doi:10.1109/TIT.2017.2778143
68. Lenz A., Polyanskii N. Optimal codes correcting a burst of deletions of variable length. *2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 757–762. doi:10.1109/ISIT44484.2020.9174288
69. Sima J., Gabrys R., and Bruck J. Syndrome compression for optimal redundancy codes. *Proceedings of 2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 751–756. doi:10.1109/ISIT44484.2020.9174009
70. Levenshtein V. I. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems on Information Transmission*, 1965, vol. 1, no. 1, pp. 8–17.
71. Dolecek L., Anantharam V. Repetition error correcting sets: explicit constructions and prefixing methods. *SIAM Journal on Discrete Mathematics*, 2010, vol. 23, no. 4, pp. 2120–2146. doi:10.5555/1958171.1958195
72. Mahdavifar H., Vardy A. Asymptotically optimal sticky-insertion correcting codes with efficient encoding and decoding. *Proceedings of 2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 2688–2692. doi:10.1109/ISIT.2017.8007016
73. Jain S., Farnoud F., Schwartz M., Bruck J. Duplication-correcting codes for data storage in the DNA of living organisms. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 8, pp. 4996–5010. doi:10.1109/ISIT.2016.7541455
74. Kovacevic M. *Codes correcting all patterns of tandem-duplication errors of maximum length 3*. Available at: https://arxiv.org/pdf/1911.06561.pdf (accessed 12 March 2021).
75. Nazirkhanova K., Medova L., Kruglik S., and Frolov A. Codes correcting bounded length tandem duplication. *Proceedings of 2020 International Symposium on Information Theory and its Applications (ISITA)*, IEICE, 2020, pp. 299–303. doi:10.34385/proc.65.B06-6
76. Heckel R., Shomorony I., Ramchandran K., Tse D. N. C. Fundamental limits of DNA storage systems. *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 3130–3134. doi:10.1109/ISIT.2017.8007106
77. Lenz A., Siegel P. H., Wachter-Zeh A., Yaakobi E. An upper bound on the capacity of the DNA storage channel. *2019 IEEE Information Theory Workshop (ITW)*, IEEE, 2019, pp. 1–5. doi:10.1109/ITW44776.2019.8989388
78. Lenz A., Siegel P. H., Wachter-Zeh A., Yaakobi E. Achieving the capacity of the DNA storage channel. *Proceedings of ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal*, IEEE, 2020, pp. 8846–8850. doi:10.1109/ICASSP40776.2020.9053049
79. Shin S., Heckel R., Shomorony I. Capacity of the erasure shuffling channel. *Proceedings of ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal*, IEEE, 2020, pp. 8841–8845. doi:10.1109/ICASSP40776.2020.9053486
80. Shomorony I., Heckel R. Capacity results for the noisy shuffling channel. *Proceedings of 2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2019, pp. 762–766. doi:10.1109/ISIT.2019.8849789
81. Mitzenmacher M. A survey of results for deletion channels and related synchronization channels. *Probability Surveys*, 2009, vol. 6, pp. 1–33. doi:10.1214/08-PS141
82. Lenz A., Siegel P. H., Wachter-Zeh A., Yaakobi E. Coding over sets for DNA storage. *IEEE Transactions on Information Theory*, 2020, vol. 66, no. 4, pp. 2331–2351. doi:10.1109/TIT.2019.2961265
83. Sima J., Raviv N., and Bruck J. On coding over sliced information. *IEEE Transactions on Information Theory*, 2021. doi:10.1109/TIT.2021.3063709
84. Song W., Cai K., Schouhamer Immink K. A. Sequence-subset distance and coding for error control in DNA-based data storage. *IEEE Transactions on Information Theory*, 2020, vol. 66, no. 10, pp. 6048–6065. doi:10.1109/TIT.2020.3002611