Миляев Д.Р. Хранение данных в ДНК // Международный журнал информационных технологий и энергоэффективности.– 2024. – Т. 9 № 11(49) с. 94–99

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.94

ХРАНЕНИЕ ДАННЫХ В ДНК

Миляев Д.Р.
ФГАОУ ВО "САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ "ЛЭТИ" ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)" , Санкт-Петербург, Россия (197022, город Санкт-Петербург, ул Профессора Попова, д. 5 литера Ф), e-mail: milyaev.dmitry00@mail.ru

Статья посвящена проблеме постоянно растущего объёма данных в современном мире и новому способу их хранения. Рассматриваются особенности технологии, а также присущие достоинства и недостатки. В первой части статьи описана проблематика предмета рассмотрения. Далее приведено описание процесса хранения данных в ДНК. Статья завершается описанием плюсов и минусов технологии, а также будущим перспективам направления.

Ключевые слова: Хранение данных, информационные технологии, технологии будущего, ДНК, кодирование информации, исследования, информация.

DATA STORAGE IN DNA

Milyaev D.R.
ST. PETERSBURG STATE ELECTROTECHNICAL UNIVERSITY "LETI". V.I. ULYANOVA (LENINA), St. Petersburg, Russia (197022, St. Petersburg, Professora Popova str., 5 letter F), e-mail: milyaev.dmitry00@mail.ru

The article is devoted to the problem of the constantly growing volume of data in the modern world and a new way of storing it. The features of the technology are considered, as well as the inherent advantages and disadvantages. The first part of the article describes the problems of the subject of consideration. The following is a description of the process of storing data in DNA. The article concludes with a description of the pros and cons of the technology, as well as the future prospects of the direction.

Keywords: Data storage, information technology, future technologies, DNA, information coding, research, information..

Необходимость нового метода хранения данных

С ростом интернета количество создаваемой человеком информации увеличивается быстрыми темпами. По прогнозам, за следующие три года общий объем цифровых данных утроится и достигнет 175 зетабайт (175 миллиардов терабайт). Существующие сейчас технологии хранения данных, такие как жесткие диски и магнитные ленты, не обеспечат надежное и долговечное сохранение такого объема информации.

Исследователи считают, что строительство новых центров обработки с применением сегодняшних технологий не спасет человечество от переизбытка информации. Хранение данных будет становиться все более дорогостоящим, что будет тормозить развитие по всем передовым направлениям.

Одно из перспективных решений — хранение данных в искусственных молекулах ДНК. Всего один грамм ДНК способен хранить в себе до 215 петабайт данных. Это означает, что весь существующий сегодня интернет-контент мог бы поместиться в небольшую коробку.

Впервые идею хранить информацию в таком формате более 60 лет назад предложил американский физик и нобелевский лауреат Ричард Фейнман. На тот момент идея звучала крайне неоднозначно — но появившиеся в XXI веке методы создания полностью синтетической ДНК-молекулы сделали эту технологию реальностью.

Рынок разработок в области цифровой ДНК-памяти в прошлом году достиг $105.5 млн, и в дальнейшем, будет расти на 69,8% в год. Исследованиями в этой области занимаются технологические компании, научные институты и даже Агентство национальной безопасности США. [1]

Хранение данных в ДНК

ДНК представляет собой последовательность нуклеотидов. Их четыре: аденин, гуанин, тимин, цитозин.

Для кодирования информации каждому из них приписывают цифровой код. Например, тимин — 0, гуанин — 1, аденин — 2, цитозин — 3.

Эта система аналогична тому, как хранит данные компьютер — с той лишь разницей, что цифровые данные зашифрованы в виде последовательностей из нулей и единиц.[1]

Последовательность нуклеотидов позволяет кодировать информацию о различных типах РНК. Все эти типы РНК синтезируются на матрице ДНК за счет копирования последовательности ДНК в последовательность РНК, синтезируемой в процессе транскрипции, и принимают участие в биосинтезе белков.

Помимо кодирующих последовательностей, ДНК клеток содержит последовательности, выполняющие регуляторные и структурные функции. Кроме того, в геноме эукариот часто встречаются участки, принадлежащие «генетическим паразитам», например, транспозонам.

Для того, чтобы использовать ДНК как носитель информации, необходимо решить две основные задачи: как кодировать и декодировать информацию в ДНК, и как синтезировать и стабилизировать ДНК. Для кодирования и декодирования информации в ДНК можно использовать простой алгоритм, который основан на бинарной системе счисления. В бинарной системе счисления информация представляется в виде последовательности нулей и единиц, называемых битом. Каждый бит соответствует одному из двух состояний: включено или выключено, да или нет, истина или ложь и т. д. Например, число 13 в десятичной системе счисления записывается как 1101 в бинарной системе счисления.

Для того, чтобы перевести бинарную информацию в ДНК, можно использовать следующее правило: 00 соответствует А, 01 соответствует Т, 10 соответствует Г, 11 соответствует Ц. Таким образом, число 1101 в бинарной системе счисления будет закодировано в ДНК как ГЦАТ. Для того, чтобы перевести информацию из ДНК в бинарную систему счисления, нужно использовать обратное правило: А соответствует 00, Т соответствует 01, Г соответствует 10, Ц соответствует 11. Таким образом, последовательность ГЦАТ в ДНК будет декодирована в бинарную систему счисления как 1101.[2]

Процесс передачи данных с помощью ДНК

В основе новой технологии хранения информации лежит способность ДНК записывать и сохранять информацию. Для создания искусственной ДНК-молекулы требуется определить нужную последовательность нуклеотидов. Затем строительные блоки будущей ДНК, производные от отдельных нуклеотидов, добавляются в раствор и соединяются в общую цепочку.

Этот процесс синтеза был автоматизирован в начале 1980-х годов, но существующие технологии синтезируют относительно короткие цепочки ДНК - не более 200 нуклеотидов, что занимает много времени.

В сфере хранения данных точность не столь критична, как в медицине, где любая ошибка может иметь серьезные последствия. Поэтому разработчики ищут новые подходы к синтезу ДНК.

Стартап Catalog предложил использовать набор из 100 коротких фрагментов ДНК, которые заранее готовятся в большом количестве. Эти фрагменты соединяются не случайным образом, а в строго определенной последовательности, заданной компьютером. Робот добавляет фрагменты в раствор, а затем ферменты соединяют их в единую цепочку. Прототип Catalog способен синтезировать ДНК со скоростью 125 ГБ в сутки, а будущая модель будет в тысячу раз быстрее.

Важный этап в разработке - это преобразование двоичных данных в четырехзначную систему, которая подходит для записи на ДНК. Современные методы позволяют сжимать данные без потери качества и записывать их с максимальной плотностью. Исследователи из Иллинойского Института Бекмана расширили алфавит ДНК, добавив в него 7 новых символов, увеличив тем самым вместимость ДНК-носителя.

Искусственная ДНК-молекула создается путем сборки последовательностей нуклеотидов в соответствии с закодированной информацией. Процесс напоминает работу струйного принтера, где информация "печатается" на стекле.

Для долговременного хранения ДНК используют специальные растворы, которые минимизируют физический износ. Существуют компании, использующие для хранения ДНК живые организмы, например, бактерии. Чтобы прочитать данные с ДНК, ее необходимо секвенировать - определить последовательность нуклеотидов. Современные методы секвенирования позволяют одновременно считывать несколько участков ДНК, что значительно ускоряет процесс.

Таким образом, ДНК-носитель является перспективной технологией хранения данных. Ученые работают над усовершенствованием методов синтеза, секвенирования и хранения, чтобы сделать эту технологию доступной и эффективной.

После секвенирования данные подвергаются декодированию. Выведенная последовательность нуклеотидов переводится обратно в двоичный код и собирается в формат, поддерживаемый компьютером.

Преимущества и недостатки технологии

Потенциально у синтетической ДНК-молекулы множество преимуществ по сравнению с традиционными хранилищами данных, но есть и свои ограничения.

Преимущества

Новая технология хранения информации в ДНК обладает рядом преимуществ, которые делают ее привлекательной альтернативой традиционным методам.

- Вместимость:

ДНК-молекула способна хранить информацию в 1009 раз плотнее, чем самый компактный жесткий диск. Это достижение свидетельствует о громадном потенциале ДНК как носителя информации. Разработчики постоянно работают над увеличением плотности записи, чтобы еще больше повысить эффективность этой технологии.

- Долговечность:

ДНК-молекула отличается невероятной устойчивостью. Состав ДНК остается неизменным десятки тысяч лет, что позволяет ученым расшифровывать информацию, содержащуюся в останках древних организмов. ДНК-молекулы сохраняют свою целостность в течение долгих периодов - ученым удалось извлечь геном из зубов сибирского мамонта, возраст которого составляет миллион лет. Для усиления сохранности ДНК-молекул рекомендуется хранить их при низких температурах. При 10°C информация сохраняется около 2000 лет, а при -20°C - 2000 столетий. В сравнении с этим, магнитные ленты - один из самых надежных носителей информации - сохраняют работоспособность в течение 30 лет.

- Постоянство:

Цифровые технологии стремительно развиваются, но в то же время быстро устаревают, что затрудняет доступ к информации, записанной на старых устройствах. [2] Структура ДНК остается неизменной уже 3 миллиарда лет. Это означает, что ДНК-хранилище не подвержено устареванию, и человечество всегда сможет расшифровать информацию, записанную с помощью этой технологии.

- Экологичность:

Современные серверные центры потребляют огромные объемы электроэнергии, что негативно влияет на окружающую среду. ДНК-хранилище работает без электричества, что делает его более экологичным решением.

Недостатки

Несмотря на то, что заголовки научных изданий прочат ДНК-памяти большое будущее, речи о массовом применении технологии пока не идет.

- Дороговизна

Хранение информации в ДНК сегодня очень дорогое. Текущая стоимость загрузки одного мегабайта — около $1. Компании, конечно, активно работают над ее снижением. Так, например, перспективная, но пока не воплощенная в жизнь многослойная модель ДНК-хранилища от французского стартапа Biomemory позволит снизить стоимость до $1 за терабайт (то есть в 1 млн раз).

При этом самой дорогой составляющей технологии, по мнению ученых, остается синтез самой искусственной ДНК-молекулы.

В развитии и удешевлении технологии синтеза ДНК заинтересованы далеко не только те, кто хочет хранить в ней данные. Последние 20 лет она развивается как самостоятельная отрасль и имеет огромное значение для биологии, медицины и генетики. На основе таких молекул создаются полезные бактерии, вакцины и биологическое топливо.[3]

Развитие технологий приводит к тому, что стоимость создания ДНК-молекулы уменьшается. За последние 30 лет оно подешевело в 10 миллионов раз. Возможно не за горами момент, когда цена станет достаточно низкой для массового производства.

Другой способ решения проблемы с дороговизной — использование натуральных ДНК-молекул вместо искусственных. Именно это сделали ученые из Гарвардского университета в 2017 году, записав короткую анимацию на ДНК живых бактерий. Для записи использовался механизм CRISPR, который позволяет бактериям вырабатывать иммунитет, накапливая память о встреченных вирусах. Но есть серьезная проблема — в отличие от синтезированной, натуральная молекула ДНК склонна к мутации, что сильно снижает надежность хранения данных.

- Низкая скорость загрузки

Вторая слабая сторона всех текущих разработок связана с первой: при высокой стоимости у технологии крайне низкая скорость работы.

В 2021 году ученым из Технологического исследовательского института Джорджии удалось создать прототип ДНК-чипа, потенциально способного параллельно записывать до 20 Гб данных в день благодаря одновременному созданию нескольких цепочек. Но пока его работа недостаточно стабильна.

- Низкая скорость поиска и выгрузки

Большинство цифровых данных предполагают постоянный доступ к ним. Низкая скорость поиска и выгрузки данных на ДНК-носителе делает работу с ними крайне неэффективной.

Этот вызов пытаются преодолеть компания Catalog. Ее особенность в быстрой системе поиска данных по ключевым словам. Для поиска данных в записанном учеными отрывке из «Гамлета» в 17 000 слов системе понадобилось всего несколько минут.

Этот невысокий показатель, но все дело в самом принципе. Разработанный химический метод позволяет сразу осуществлять поиск в том участке, где содержится нужная информация, не анализируя структуру целиком. Ученые уверяют, что в будущем этот метод ускорится примерно в тысячу раз.

- Недостающая компактность

Устройства для записи и считывания информации с ДНК совсем не так компактны, как сами молекулы. Так, разработка Catalog под названием Shannon, как говорят ее создатели, занимает объем среднестатистической комнаты. Для решения проблемы ученые объединили усилия с компанией Seagate, лидером в сфере современных систем хранения. По словам технического директора Catalog Дэвида Турека, совместно они будут стремиться уменьшить объем в тысячу раз. Конечная цель — создать «лабораторию на чипе», содержащую десятки резервуаров для хранения молекул ДНК.

Сами разработчики настроены оптимистично. «Мы не видим никаких радикальных препятствий для успеха этой технологии», — говорит Адам Мейер, старший научный сотрудник Гарварда. В пример он приводит магнитную ленту для хранения данных, которая совершенствовалась в течение 60 лет, прежде чем стать передовым способом хранения информации.[3]

Будущее технологии

ДНК давно признана идеальным носителем информации благодаря своей плотности, долговечности и способности к копированию. Несмотря на успешные эксперименты по записи на ДНК различных данных, включая литературные произведения и музыку, массовое применение этой технологии сдерживается сложностью и дороговизной процесса.

Компания Catalog под руководством Хьенджуна Парка работает над созданием революционной машины, способной записывать терабайты данных на ДНК ежедневно.

Цель компании - предложить корпоративным клиентам, включая IT-компании, индустрию развлечений и государственные учреждения, услуги хранения данных на ДНК.

Успех этой технологии позволит решить проблему переизбытка информации, которая остро стоит в XXI веке.

Магнитные ленты, используемые для хранения цифровых архивов, требуют замены каждые 10 лет. Переход к ДНК-хранилищу в первую очередь станет доступен крупным

клиентам, но в долгосрочной перспективе предполагается полная замена магнитных накопителей.

Развитие генетики и синтетической биологии может ускорить этот процесс, позволив людям получить доступ к данным, хранящимся в их собственной ДНК.

Список литературы

1. DNA Chips: The Billion Gigabyte Storage Solution of Tomorrow [электронный ресурс] URL: https://scitechdaily.com/dna-chips-the-billion-gigabyte-storage-solution-of-tomorrow/ (дата обращения 10.09.24).
2. Дж. Уотсон. Двойная спираль. Издательство Харвест, 2013. – 400 с.
3. От флешек к ДНК: разбираемся в новой технологии хранения данных[электронный ресурс] URL: https://habr.com/ru/companies/itglobalcom/articles/743248/ (дата обращения 11.09.24).

References

1. DNA Chips: The Billion Gigabyte Storage Solution of Tomorrow [electronic resource] URL: https://scitechdaily.com/dna-chips-the-billion-gigabyte-storage-solution-of-tomorrow / (accessed 09/10/24).
2. J. Watson. The double helix. Harvest Publishing House, 2013. – 400 p.
3. From flash drives to DNA: we understand the new technology of data storage [electronic resource] URL: https://habr.com/ru/companies/itglobalcom/articles/743248 / (accessed 11.09.24).