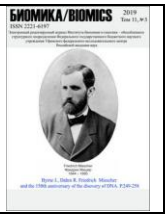
Биомика, 2019, Том 11 № 3, 344-377

БИОМИКА/BIOMICS

ISSN 2221-6197 http://biomicsj.ru



НЕБИОЛОГИЧЕСКОЕ ПРИМЕНЕНИЕ МОЛЕКУЛ ДНК

Сахабутдинова А.Р.[1], Михайленко К.И.[2], Гарафутдинов Р.Р.[1],
Кирьянова О.Ю.[3], Сагитова М.А.[4], Сагитов А.М.[5], Чемерис А.В.[1]

[1]Институт биохимии и генетики – обособленное структурное подразделение Федерального государственного бюджетного научного учреждения Уфимского федерального исследовательского центра Российской академии наук, Россия, 450054, Уфа, пр. Октября, 71, chemeris@anrb.ru
[2]Институт механики – обособленное структурное подразделение Федерального государственного бюджетного научного учреждения Уфимского федерального исследовательского центра Российской академии наук, Россия, 450054, Уфа, пр. Октября, 71
[3]Уфимский государственный нефтяной технический университет, Россия, 450062, Уфа, ул. Космонавтов 1
[4]Московский физико-технический институт, Россия, 141701, г. Долгопрудный, Институтский пер., 9
[5]НКО «Фонд развития промышленности Республики Башкортостан», 450101, Уфа, ул. К.Маркса, 3

Резюме

О том, что существует ДНК известно уже 150 лет, с 1944 года стало ясно, что в ДНК кодируется наследственная информация, передающаяся потомкам, а в 1953 году было выяснено, что ДНК имеет двухцепочечную структуру, удерживаемую водородными связями, возникающими между комплементарными азотистыми основаниями. За последующие годы осуществлены огромные прорывы в познании организации и функционирования ДНК как биологической макромолекулы, включая определение ее первичной структуры. Получены убедительные доказательства, что ДНК за счет ее огромного (био)разнообразия и бесчисленных перестановок нуклеотидов можно считать истинно цифровой молекулой. Однако возможность небиологического применения ДНК и попытки его реализации насчитывают меньше трех-четырех десятилетий и основной толчок им дал все же подход с молекулярными вычислениями или иначе ДНК-компьютинг. Появившиеся затем ДНК-криптография и ДНК-стеганография привлекли значительное внимание исследователей по всему миру, и было предложено немало способов кодировки азотистыми основаниями небиологической информации в виде букв английского алфавита и прочих символов, большая часть которых рассмотрена в данной статье. Другое интересное направление небиологического использования молекул ДНК представляет собой разработку различных способов кодирования информации для ее долговременного хранения в молекулах ДНК, чему в данной статье уделено значительное внимание. Описаны также исторические аспекты давних предложений по использованию молекул ДНК в качестве носителей компьютерной памяти, где благодаря М.С. Нейману приоритет за нашей страной.

Ключевые слова: ДНК, ДНК-компьютинг, ДНК-криптография, ДНК-стеганография, ДНК хранение

Цитирование: Сахабутдинова А.Р., Михайленко К.И., Гарафутдинов Р.Р., Кирьянова О.Ю., Сагитова М.А., Сагитов А.М., Чемерис А.В. Небиологическое применение молекул ДНК // *Биомика*. 2019. Т.11(3). С. 344-377. DOI: 10.31301/2221-6197.bmcs.2019-28

NON-BIOLOGICAL APPLICATION OF DNA MOLECULES

Sakhabutdinova A.R.[1], Mikhailenko K.I.[2], Garafutdinov R.R.[1],
Kiryanova O.Yu.[3], Sagitova M.A.[4], Sagitov A.M.[5], Chemeris A.V.[1]

[1]Institute of Biochemistry and Genetics, Ufa Federal Research Center, Russian Academy of Sciences, 71 Prospekt Oktyabrya, 450054, Ufa, Russia, chemeris@anrb.ru
[2]Institute of Mechanics, Ufa Federal Research Center, Russian Academy of Sciences, 71 Prospekt Oktyabrya, 450054, Ufa, Russia, const.mkh@gmail.com
[3]Ufa State Petroleum Technological University, 1 Kosmonavtov str., Ufa, 450062, Russia

[4]The Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia
[5]NCO "Industrial development Fund of the Republic of Bashkortostan», 3 K.Marksa str., Ufa, 450101, Russia,

Resume

The existence of DNA has been known for 150 years, since 1944 it became clear that DNA encodes hereditary information transmitted to descendants, and in 1953 it was found that DNA has a double-stranded structure, held by hydrogen bonds arising between complementary nitrogenous bases. In the following years, huge breakthroughs were made in the knowledge of the organization and functioning of DNA as a biological macromolecule including a determination of their primary structure. There is convincing evidence that DNA due to its huge (bio)diversity and countless permutations of nucleotides can be considered a truly digital molecule. However, the possibility of non-biological application of DNA and attempts to implement it are less than three to four decades and the main impulse was given to them by the approach of DNA computing. Then appeared DNA cryptography and DNA steganography attracted considerable attention of researchers around the world, and it was proposed many ways to encode non-biological information in the form of letters of the English alphabet and other symbols by nitrogen bases, most of which are discussed in this article. Another interesting area of non-biological use of DNA molecules is the development of different ways of encoding different information for its long-term storage in DNA molecules, which in this article is given considerable attention. The historical aspects of long-standing proposals on the use of DNA molecules as computer memory carriers, where thanks to M. S. Neiman the priority for our country, are also considered.

Keywords: DNA, DNA computing, DNA cryptography, DNA steganography, storage in DNA

Citation: Sakhabutdinova A.R., Mikhailenko K.I., Garafutdinov R.R., Kiryanova O.Yu., Sagitova M.A., Sagitov A.M., Chemeris A.V. Non-biological application of DNA molecules. *Biomics.* 2019. V.11(3). P. 344-377. DOI: 10.31301/2221-6197.bmcs.2019-28 (In Russian)

«DNA is essentially digital»
Leonard Adleman

Введение

Только в середине 1940-х гг. [Avery et al., 1944] через 75 лет после открытия в 1869 г. нуклеина – неизвестного до той поры природного вещества [Miescher, 1871; Byrne, Dahm, 2019], названного позже ДНК, стала ясна сверхважная биологическая роль этого типа биоорганических кислот, выполняющих роль наследственного материала. Но потребовалось еще несколько десятилетий до того как молекулы ДНК попытались использовать в небиологических целях в виде неких самособирающихся наноструктур [Seeman, 1982], для молекулярных вычислений, получивших название «ДНК-компьютинг» [Adleman, 1994], а также для ДНК-криптографии и ДНК-стеганографии [Cleland et al., 1999], а также хранения различных данных [Banckroft et al., 2001]. При этом главную роль в небиологическом применении ДНК играет биологический принцип комплементарности азотистых оснований, согласно которому в двойной спирали ДНК аденин спаривается с тимином, а гуанин – с цитозином.

Как раз Л. Адлеману – известному американскому математику-шифровальщику[1] принадлежит взятое в качестве эпиграфа к данной статье высказывание, в несколько вольном переводе звучащее как «ДНК по сути, цифровая». Достаточно долго бытовавшее представление о данном биополимере, как состоящем из монотонно чередующейся четверки нуклеотидов (аденина, гуанина, тимина и цитозина) не позволяло допустить существование того огромного разнообразия, которое ему присуще. Произвольное же чередование данных нуклеотидов на самом деле обеспечивает поистине огромнейшее число комбинаций из них, которое в свою очередь приводит к великому биоразнообразию всего живого на Планете. Чтобы не быть голословными, пожалуй, следует привести ряд примеров, подтверждающих правоту высказывания Л. Адлемана и поясняющих почему мы ему беспрекословно вторим.

Теоретические количества возможных комбинаций нуклеотидных последовательностей для олиго- или полинуклеотидов различной длины легко подсчитать, возводя в соответствующую степень четверку нуклеотидов по формуле 4^n, где n – длина

[1] Л. Адлеман является одним из авторов весьма популярного продукта в виде алгоритма шифрования с открытым ключом, широко используемого в приложениях компьютерной безопасности, включая протокол HTTPS, получившего название RSA, и под данной аббревиатурой скрываются фамилии его разработчиков – Rivest, Shamir, Adleman.

анализируемого участка ДНК. Количество всех возможных комбинаций, например, октануклеотидов определяется как число размещений с повторениями из 4 элементов по 8, что составляет 65536 (4^8) вариантов (AAAAAAAA AAAAAAAc AAAAAAAg AAAAAAAt AAAAAAcA AAAAAAgA AAAAAAtA AAAAAcAA AAAAAgAA AAAAAtAA AAAAcAAA AAAAgAAA AAAAtAAA .... tttttttt, где строчными буквами показаны отличающиеся нуклеотиды от таковых в условно первом 8-ми звенном олигонуклеотиде AAAAAAAA). Более длинным молекулам или участкам ДНК присущи планомерно возрастающие количества размещений четверки азотистых оснований. Так, для декануклеотида таковых будет уже за миллион (4^{10}), для олигонуклеотида длиной 15 звеньев – миллиард (биллион)[2], для 20 звеньев – триллион, для олигонуклеотида из 30 звеньев – квинтиллион (10^{18}), для участков ДНК из 60 или 100 азотистых оснований количества комбинаций превысят соответственно ундециллион (10^{36}) и нонадециллион (10^{60}). Гугол[3] комбинаций (10^{100}) будет превышен, если иметь дело с ДНК длиной 167 нуклеотидов.

Если еще не убедили этими числами в цифровом характере ДНК[4], то приведем сведения по неким условным микроорганизмам, имеющими геномы размерами по 2 миллиона пар нуклеотидов[5], в которых может иметься $4^{2000000}$ вариантов перебора нуклеотидов, что приблизительно соответствует $10^{1200000}$. При этом считается, что во Вселенной имеется всего то около 10^{80} элементарных частиц. Впрочем, эти варианты разнообразия геномов существуют только теоретически. Ввиду общности происхождения всего живого, и, как показывают уже имеющиеся геномные данные, эти бактерии (вообще любые) будут непременно иметь некие совпадающие гомологичные участки, заметно уменьшающие число реальных комбинаций перестановок четырех нуклеотидов у таких организмов. Более того, некоторая гомология нуклеотидных последовательностей для эволюционно консервативных генов характерна даже для представителей всех трех ветвей жизни на Земле: архей, бактерий и эукариот. И если провести аналогию с языками разных народов, то можно отметить, что, как во многих из них есть сходные однокоренные слова, так и в геномах различных организмов (подчас даже далеко отстоящих друг от друга на эволюционной лестнице) есть некие совпадающие высококонсервативные последовательности, что свидетельствует, во-первых, об эволюционировании организмов от некоего единого пра(пра)$_n$прапредка, а во-вторых, позволяет предполагать выполнение ими одинаковых или близких функций, отражая дивергенцию и конвергенцию признаков. Причем, как и в языках, так и в ДНК такие слова/мотивы могут сохраняться и передаваться потомству не полностью, а лишь своими корнями/фрагментами. Поэтому у разных биологических объектов в их ДНК все же нет совсем произвольного чередования азотистых оснований. Из всего многообразия вариантов перемежения нуклеотидов существует достаточное большое число последовательностей, которые можно назвать «запрещенными», при наличии которых функционирование биологических систем невозможно. Однако для небиологического применения молекул ДНК, к описанию которого скоро перейдем, такие ограничения отсутствуют, и практически все комбинации возможны[6], и это еще

[2] Здесь мы привели наименование больших чисел фактически по двум системам, действующим в разных странах. По так называемой длинной шкале после миллиона идет миллиард, затем чередуя «...лионы» и «...лиарды», следуют биллион, биллиард и т.д. По короткой шкале за миллионом следует биллион, затем триллион, квадриллион, квинтиллион, секстиллион и так далее, что сразу дает возможность понять о каком классе чисел идет речь. При использовании длинной шкалы это сделать несколько сложнее. Но на самом деле нечасто возникает необходимость оперировать столь большими числами и в быту уже произошло некоторое смешение систем, так как в России, например, после миллиарда идет триллион, тогда как должен был быть сначала биллион, потом биллиард и только затем триллион.

[3] Число «гугол» (googol), равное 10^{100} не надо путать со сходным по произношению названием поисковика Google, что представляет собой «игру слов» и подразумевает претензию последнего на огромную информативность.

[4] Собственно не только в этом заключается «цифровизна» молекул ДНК для небиологического их применения, но об этом будет говориться дальше.

[5] Такие размеры геномов являются довольно типичными для многих бактерий.

[6] Определенные ограничения в разнообразии и соответственно в числе комбинаций для использования ДНК в небиологических целях накладываются возникающими (нежелательными) вторичными структурами молекул ДНК, зависящими от последовательностей нуклеотидов в них, а также практической невозможностью, например химического синтеза олигонуклеотидов, содержащих протяженные гомополимерные участки из остатков гуанина. Но даже за такими исключениями молекулы ДНК все равно позволяют оперировать с фактически бесконечным количеством вариантов перебора азотистых оснований в них.

больше укрепляет в мысли, что ДНК – поистине цифровая молекула, придуманная самой Природой.

Основными предпосылками к небиологическому использованию молекул ДНК являются следующие. Во-первых, упоминаемое выше огромное число комбинаций перестановок нуклеотидов обеспечивает уникальность практически любых «отрезков» ДНК, начиная, например с 20 звеньев, поскольку количество их вариаций превышает триллион и для большинства целей этого вполне достаточно. Причем при необходимости повышения уникальности нуклеотидных последовательностей они могут быть удлинены на требуемое число нуклеотидов, причем увеличение длины на 10 звеньев приводит к увеличению числа комбинаций в 10^6 раз или по-другому – на два класса чисел, поскольку 4^{30} превысит уже квинтиллион. Во-вторых, важным является биологический принцип комплементарности нуклеотидов, согласно которому аденин спаривается с тимином, а цитозин с гуанином. Таким образом, можно запрограммировать взаимодействие разных цепей ДНК при формировании ими вторичной структуры в виде двойной спирали, поскольку ДНК всегда стремится стать двуцепочечной молекулой, по крайней мере в участках, где возникнет достаточное количество водородных связей между комплементарными азотистыми основаниями, обеспечивающими при определенных условиях (в первую очередь температурных) поддержание (природной) целостности этой молекулы.

Если вышеупомянутые особенности молекул ДНК являются непременным атрибутом этого типа биополимеров, то еще ряд возможностей, необходимых, в том числе для небиологического использования ДНК, опять-таки основываясь на природных процессах, являются экспериментальными разработками. Так, третьим важным обстоятельством работы с молекулами ДНК является их «размножаемость» в системах *in vitro*, происходящая под действием соответствующих ДНК полимераз в присутствии магния и «строительных блоков» в виде дезоксинуклеозидтрифосфатов. Этот процесс называется амплификацией и осуществляется разными способами, наиболее часто используемым среди которых является полимеразная цепная реакция (ПЦР). В-четвертых, установление последовательностей нуклеотидов (секвенирование), по которому можно оценивать (био)разнообразие этих молекул, также является краеугольным камнем современной физико-химической биологии и без него небиологическое применение молекул ДНК будет неполноценным. Причем методов секвенирования уже достаточно много и стоимость этого процесса с появлением методов новых поколений очень сильно снизилась. Наконец, разработаны методы химического синтеза олигонуклеотидов с заданными последовательностями и длиной до 200 звеньев, что является еще одним ключевым моментом для небиологического применения ДНК.

Исторические аспекты

Прежде чем приступить к изложению основного материала статьи, возможно, следует напомнить о размерах молекул ДНК, двойная спираль которых в наиболее типичной В-конформации имеет диаметр около 2 нм, а шаг спирали из 10,5 нуклеотидов – приблизительно 3,4 нм. При этом длина нативных молекул ДНК может достигать весьма внушительных размеров. Так, самая большая хромосома человека, состоящая из 245 миллионов пар оснований, содержит ДНК протяженностью около 8,3 см, что приводит к гигантскому отношению длины к толщине. Но это при плотной упаковке в хромосоме, тогда как изолированная ДНК не может сохранить подобную целостность и практически как попало рвется, в том числе, вследствие гидродинамических воздействий. Благодаря своим основным размерам, лежащим в нанометровом диапазоне, а также за счет способности формирования (в том числе запрограммированного) двуцепочечных структур по принципу комплементарности азотистых оснований, молекулы ДНК находят в современной нано(био)технологии определенное применение, которое вне всякого сомнения будет шириться.

Впрочем, история нанотехнологии ведет начало не от биологических молекул. В декабре 1959 г. американский физик Р. Фейнман – будущий Нобелевский лауреат 1965 г., получивший премию за фундаментальные работы в области квантовой электродинамики, на ежегодном собрании Американского физического общества в Калифорнийском технологическом институте сделал доклад [Feynman, 1960], полное название которого «There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics» на русский язык можно перевести как «Внизу премного места: Приглашение в новую область физики», в котором заглянул в технологическое будущее нашей цивилизации и высказал предположение, что со временем многие материалы и устройства будут изготавливаться на молекулярном и атомарном уровнях, то есть станет возможным механическое перемещение атомов с помощью специального манипулятора соответствующего размера. Однако в первую очередь он думал тогда о металлах. Р. Фейнман подсчитал, что вся информация, которую человечество накопило во всех книгах мира, к тому моменту составляла 10^{15} бит информации, и допустил для одного бита достаточность использования всего 100 атомов. Забегая вперед, скажем, что он оказался недалек от количества атомов, кодирующих один бит при нынешнем использовании в качестве хранения информации молекул ДНК. В качестве примера

возможностей миниатюризации несложными расчетами Р. Фейнман показал, что принципиально достижимо разместить все тома энциклопедии Британика на кончике булавки и для этого требуется масштабирование в сторону уменьшения всего в 25 тысяч раз. В своем эпохальном докладе Р. Фейнман не оставил без внимания и биологию, перечислив некоторые ее центральные и фундаментальные проблемы того времени, и поставил на первое место вопрос о последовательности нуклеотидов в ДНК. И сам ответил на него следующим образом – «... Вы будете видеть порядок оснований в цепочке ...», имея в виду цепочку ДНК и для этого, по его мнению, требовалось значительно повысить разрешение микроскопов, которые оставались достаточно «грубы».

Однако приоритет в предложении использовать ДНК для целей хранения небиологической информации все же принадлежит известному советскому ученому М.С. Нейману, опубликовавшему в журнале «Радиотехника» в 1964 – 1965 гг. серию статей, где им были высказаны оригинальные идеи и принципиальные соображения о радикальной миниатюризации элементов записи, хранения и извлечения дискретной информации на молекулярно-атомном уровне, в том числе при использовании биологических структур, а именно нуклеиновых кислот. Причем ряд его высказываний, вне всякого сомнения, заслуживает приведения их полностью. Так, в статье, посвященной микроминиатюризации, рассматривая функционирование молекул ДНК, М.С. Нейман указывает, что при этом используется «четырехзначный код химического, а значит электронного характера» [Нейман (Neiman), 1964]. Он также замечает, что такие процессы поддаются техническому управлению и со временем «могут быть установлены методы искусственного построения информационных машин с микроэлементами, состоящими из отдельных молекул или атомов». В своей следующей статье [Нейман (Neiman), 1965], развивающей идеи микроминиатюризации на молекулярно-атомном уровне, включая вопросы надежности и быстродействия таких систем, М.С. Нейман обращает внимание, что запись генетической информации в ДНК имеет чрезвычайно высокую степень миниатюризации. В другой статье М.С. Нейман (M.S.Neiman) [1965a] отмечает, что в предыдущих публикациях [Нейман (Neiman), 1964; 1965] им были рассмотрены самые общие вопросы, связанные с идеей радикальной миниатюризации дискретных элементов информационных систем, а экспериментальные подходы оставались неясными, в связи с чем могло создаться представление об их полной нереализуемости и поэтому в этой третьей работе [Нейман (Neiman), 1965a] делается попытка обсудить некоторые моменты на примере задач записи и считывания информации в молекулярных системах памяти, в том числе принимая во внимание возможные мутации, которые, по его мнению сопоставимы с внесением изменений в хранящуюся информацию. Значительная часть статьи посвящена вопросам как производить нужные изменения в записи на генетическом уровне и как потом считывать такую информацию. При этом надо учесть, что уровень молекулярно-биологических знаний и возможностей в середине 1960-х гг. был достаточно невысок. Тем не менее, М.С. Нейман считал, что «проблема радикальной миниатюризации систем памяти информационных машин оказывается очень близкой к биологической проблеме управления наследственностью». То есть, говоря другими словами, желание внести в молекулу ДНК, как хранительницу некой информации конкретные изменения тесно сопряжено со стремлением в биологии «получить ... заранее заданные изменения наследственности». Продолжая эту мысль, М.С. Нейман пишет, что «проблема направленных мутаций, имеющая крупнейшее значение для селекционной работы, а также в медицине для борьбы наследственными и вирусными, а, возможно, и раковыми заболеваниями уже ставится в биологии, хотя и рассматривается пока биологами как весьма отдаленная». В этих его словах можно узреть и прогнозирование нынешнего геномного CRISPR/Cas-редактирования, «задержавшегося», правда, на полвека, и которому мы посвятили отдельный номер журнала «Биомика» [Чемерис (Chemeris), 2017] и целую серию статей в нем [Кулуев и др. (Kuluev et al.), 2017; Баймиев и др. (Baymiev et al.), 2017; Чемерис и др. (Chemeris et al.), 2017; Вершинина и др. (Vershinina et al.), 2017] и некоторые другие обзорные статьи [Чемерис и др. (Chemeris et al.), 2018; Кирьянова и др. (Kiryanova et al.), 2019; Кулуев и др. (Kuluev et al.), 2019; 2019a; Геращенков и др. (Gerashchenkov et al.), 2020].

В этой же третьей своей статье, посвященной проблемам радикальной миниатюризации элементов хранения и обработке информации [Нейман, 1965a], М.С. Нейман также цитирует фрагменты последнего интервью «отца кибернетики» Н. Винера, данного им 24 февраля 1964 г. незадолго до смерти последнего, и при этом отмечает, что независимо и фактически одновременно в СССР и США возникли приблизительно одинаковые идеи в плане совершенствования носителей компьютерной памяти и серьезной миниатюризации компьютерной техники. Здесь следует заметить, что упоминаемая выше статья М.С. Неймана (M.S.Neiman) [1964], поступившая в редакцию 20 сентября 1963 г., вышла затем в

январском номере 1964 г. и таким образом приоритет М.С. Неймана неоспорим.

При этом нельзя не коснуться того интервью Н. Винера [Interview, 1964] более подробно – оно заслуживает того по ряду причин. Так, на просьбу журналиста заглянуть в компьютерное будущее человечества Н. Винер предположил, что произойдет миниатюризация и очень важной составляющей этого процесса должно стать появление новых типов компьютерной памяти – транзисторов и прочих подобных элементов. При этом при ответе на тот вопрос Н. Винер плавно перешел на генетическую память, обеспечиваемую нуклеиновыми кислотами, и выразил некую надежду, что в следующем десятилетии или несколько позднее ее использование может стать технически доступно и подчеркнул, что он не один так думает. И на последовавший прямой вопрос – считает ли он, что вместо магнитной ленты базовыми элементами памяти компьютера станут гены Н. Винер ответил, что как их называть – вопрос скорее фразеологический, но это будут вещества того же рода. Однако до первых экспериментальных работ в области ДНК-компьютинга оставалось еще три десятилетия. То интервью было озаглавлено вопросительно как «Машины умнее людей?» и что весьма примечательно имело подзаголовок-цитату «...Советы впереди нас в теории автоматизации». На вопросы корреспондента – «Не заметили ли вы во время вашей последней поездки в Россию, что Советы уделяют большое внимание компьютеру?» и «В полной ли мере они используют эту науку, сравнимо с нами?» Н. Винер сказал, что этому в СССР уделяется много внимания. Имеются институты в Москве, Киеве, Ленинграде, Ереване, Тбилиси, Самарканде, Ташкенте и Новосибирске. Возможно и в других местах. При этом имеется некоторое отставание в оборудовании, но оно не безнадежно, однако они опережают США в теоретизации автоматизации. И это был 1964 год и интервью давал основоположник кибернетики, которую в СССР в первой половине 1950-х гг. считали лженаукой, но за относительно короткий период смогли практически догнать и, по признанию самого Н. Винера, даже в чем-то перегнать Америку [Interview, 1964].

ДНК-компьютинг

Прежде чем начать описывать как производится ДНК-компьютинг следует напомнить, что математические проблемы делятся на несколько классов сложности. Вычислительная сложность задач определяется временем, которое необходимо затратить на их решение на воображаемых детерминированных или недетерминированных машинах Тьюринга (ДМТ и НДМТ соответственно). В ДМТ используются детерминированные алгоритмы и порядок операций, выполняющихся последовательно, предопределен заранее. В НДМТ применяются недетерминированные алгоритмы, имеющие узловые точки, в которых происходит ветвление производящихся вычислений, ведущихся затем параллельно. На основании того, известен ли для конкретной задачи эффективный алгоритм действий для ДМТ или НДМТ классы сложности подразделяют на «P» (polynomial) задачи, «NP» (nondeterministic polynomial) и «NP-полная» (NP-complete – NPC) задачи. К первому типу относятся задачи, решаемые на ДМТ за полиномиальное время. NP задачи решаются за экспоненциальное время на ДМТ и за полиномиальное время на НДМТ. Наиболее сложными считаются NP-полные задачи, для которых считается, что полиномиальных детерминированных алгоритмов не существует. Одной из таких задач является задача о коммивояжере[7], при решении которой используется алгоритм отыскания Гамильтонова пути, названного по имени ирландского математика W.R. Hamilton, описавшего ее еще в 1857 г.

Именно такую задачу задумал решить Л. Адлеман с помощью молекулярных вычислений с использованием ДНК. По его признанию эта идея пришла к нему летом 1993 г., когда он читал книгу Дж. Уотсона и соавт. «Molecular Biology of the Gene», но лишь к Рождеству пришло понимание как ее выполнить. И еще через 5 месяцев 27 мая 1994 г. готовая статья поступила в редакцию журнала Science. 19 сентября 1994 г. она оказалась принятой к опубликованию и увидела свет 11 ноября того же года [Adleman, 1994]. Причем на само проведение экспериментов, как сообщается в статье, ушло всего 7 дней. Вскоре после этой пионерной работы в англоязычной литературе появились термины «Molecular computing» (молекулярный компьютинг) и «DNA computing» (ДНК-компьютинг). Л. Адлеман использовал набор 20-звенных олигонуклеотидов, отождествляющих как вершины графа, так и их грани. При этом 10-ти нуклеотидные части одних олигонуклеотидов совпадали или были гомологичны другим. Фактически первые 10 азотистых оснований на 5’-конце у части таких олигонуклеотидов имитировали собой первые половины пути от одной вершины к другой, тогда как 10 последующих, прилегающих к 3’-концу имитировали вторые половины пути. И так по всем граням. А дополнительные олигонуклеотиды, условно принадлежавшие вершинам графа, были подобраны так, что оказывались комплементарны и 3’- и 5’-

[7] Задача, в которой коммивояжер должен посетить N городов, побывав в каждом из них ровно один раз, и завершив путешествие в том городе, с которого он начал. В какой последовательности ему нужно обходить города, чтобы общая длина его пути была наименьшей?

концам олигонуклеотидов, подходящим к одной вершине с разных направлений, с тем расчетом, что последние гибридизировались с ними и у Т4 ДНК-лигазы появлялась возможность произвести лигирование соседних олигонуклеотидов, имитирующих грани, формируя тем самым соответствующий путь уже между тремя вершинами. После того как возникающие за счет молекулярной гибридизации комплементарных участков всех этих олигонуклеотидов различные двухцепочечные структуры оказывались пролигировавшими, проводилась ступенчатая ПЦР с фланкирующими праймерами, соответствующими начальной и конечной точкам пути, и на гель-электрофоретической картине проявлялись полосы ДНК разного размера, по которым можно было судить о произошедших событиях и восстанавливать предположительные маршруты.

Построение такого маршрута между четырьмя-пятью городами можно выполнить вручную на листке бумаги, тогда как при возрастании числа городов ее сложность резко возрастает. Так, если увеличить число городов до ста, то она становится практически нерешаемой, поскольку, посчитано, что используя стандартный алгоритм, будет необходимо выполнить 10^{147} действий, на которые суперкомпьютер с производительностью в триллион операций в секунду должен будет затратить таковых 10^{135}, тогда как наша планета существует всего-то около 10^{18} секунд. Л. Адлеман в своем исследовании остановился на 7 условных городах, соединенных между собой 10 «односторонними» и 2 «двусторонними» путями. Несмотря на то, что некоторые вычислительные шаги в ДНК-компьютинге (молекулярная гибридизация, лигирование, ПЦР, гель-электрофорез) достаточно длительны, и по скорости операций обычные компьютеры опережают их на несколько порядков, благодаря массивному параллелизму, обеспечиваемому тем, что в реакционной пробирке манипуляции со всеми молекулами ДНК (которых может быть 10^{14} и более) совершаются одновременно, молекулярный компьютер по общему числу операций легко превосходит нынешние суперкомпьютеры, не говоря уже о том, что последние объединяют в себе тысячи процессоров, тратят массу электроэнергии и занимают значительные площади в сотни квадратных метров, тогда как непосредственно для ДНК-компьютинга достаточно обычного лабораторного стола. Так, например, Л. Адлеман упоминает, что при ДНК-компьютинге на 10^{19} операций тратится около 1 джоуля энергии, притом, что суперкомпьютеры за такое же количество энергии совершают лишь 10^{9} операций.

Надо сказать, что после опубликования этой работы Л. Адлемана последовала мощная критика, которая сводилась в основном к тому, что, если число городов будет увеличено, то необходимые количества олигонуклеотидов для таких молекулярных вычислений вырастут неимоверно. При этом оценки разных авторов также весьма сильно различались. Так, например для 70 городов (вершин графа) общий вес необходимых олигонуклеотидов одни посчитали, что составит 10^{25} кг! [Linial M., Linial N., 1995]. Другая группа авторов решила, что даже для 23 городов каждого олигонуклеотида потребуется более килограмма весом [Lo et al., 1995]. Эти же авторы подвергли сомнению затраты энергии при молекулярном компьютинге, приведенные Л. Адлеманом, отметив, что надо также считать расход энергии, затрачиваемой на работу ДНК-термоциклера при проведении ПЦР, гель-электрофорез и прочие процедуры. В еще одном возражении [Bunow, 1995] отмечалось, что для подобного вычисления необходимо использовать олигонуклеотиды большей протяженности (100–200 азотистых оснований) и их число вообще может составить 10^{70}, тогда как уже говорилось выше считается, что вся наша Вселенная состоит из приблизительно 10^{80} элементарных частиц. В ответном сообщении Л. Адлеман обратил внимание публики и оппонентов на то, что молекулярный компьютинг находится в зародышевом состоянии и вопрос сможет ли он конкурировать с электронными компьютерами надо считать открытым [Adleman, 1995]. При этом он также заметил, что при проведении подобных работ не следует упускать из виду их первопричину, заключающуюся, по его мнению, в выяснении фундаментальных вопросов вычислений и биологии и именно это должно вселять оптимизм. Что касается количеств используемых олигонуклеотидов в ДНК-компьютинге, то Л. Адлеман в своей статье 1994 г. упоминал, что теоретически можно вести вычисления, работая с единичными молекулами. Позже было показана принципиальная возможность проведения ДНК-компьютинга на основе детекции единичных молекул олигонуклеотидов с помощью флуоресцентной корреляционной спектроскопии [Schmidt et al., 2004].

Несмотря на столь критические замечания, интерес к ДНК-компьютингу не пропал и продолжается до сих пор, и вскоре вслед за работой Л. Адлемана вышла еще одна статья [Lipton, 1995], посвященная молекулярным вычислениям, где была показана применимость такого подхода для решения не только комбинаторных задач. Так, R. Lipton показал принципиальную возможность решения с помощью ДНК-компьютинга и других NP- и NPC-проблем, относящихся к труднорешаемым задачам с нечеткой логикой.

Л. Адлеман проводил свои эксперименты в реакционной пробирке с олигонуклеотидами,

находящимися в растворе, однако использование фиксированных нуклеиновых кислот на некоей поверхности имеет целый ряд преимуществ в виде, например, удобства обращения с ними, хотя есть и некоторые недостатки из-за стехиометрических взаимодействий сорбированных олигонуклеотидов, а также дополнительных усилий по подготовке твердофазных систем. Но преимуществ все же больше, и не удивительна серия статей одной группы авторов, посвященных молекулярным вычислениям на поверхности стекла или золота, позволившая говорить о разработке ДНК-компьютера на поверхности [Frutos et al., 1997; Liu et al., 1998; 2000; Smith et al., 1998; Wang et al., 2001; Su, Smith, 2004]. Центральным элементом такого ДНК-компьютинга служит набор из «ДНК-слов», представляющий собой 16-ти звенные олигонуклеотиды, состоящие из фиксированных (F) и вариабельных (v) последовательностей нуклеотидов, схему организации которых можно изобразить как 5'FFFFvvvvvvvvFFFF3'. Именно вариабельные нуклеотиды и определяют специфичность действия таких «ДНК-слов». Во избежание больших различий по температурам плавления данных олигонуклеотидов на их вариабельные части накладывались ограничения в виде необходимости соблюдения 50%-ного GC-содержания. Принцип действия такого молекулярного компьютера заключается в выполнении ряда стадий (опуская подготовительные процедуры по синтезу и иммобилизации подходящих олигонуклеотидов), получивших следующие названия – MARK (молекулярная гибридизация «ДНК-слов» с анализируемыми олигонуклеотидными последовательностями, маркирующая конкретные «ДНК-слова» как «true» для произошедшей гибридизации и «false» без оной, что соответствует компьютерным 1 и 0); DESTROY (ферментативное разрушение под действием экзонуклеазы I *E. coli* не вступивших в гибридизацию одноцепочечных «ДНК-слов», фиксированных на твердой фазе); UNMARK (денатурация сформировавшихся дуплексов); READOUT (считывание оставшихся «ДНК-слов», соответствующих значениям «true» или «1») [Wang et al., 1998; 1999]. При этом три первых стадии могут неоднократно повторяться при использовании новых вводных олигонуклеотидов. В одной из работ этих авторов для решения более сложных вычислительных задач было предложено проводить лигирование «ДНК слов», фиксированных на твердой фазе с помощью T4 ДНК лигазы [Frutos et al., 1998]. Было также описано некоторое улучшение ДНК-компьютинга на поверхности, включая уменьшение числа необходимых для молекулярного компьютинга олигонуклеотидов [Wu, 2001]. Другими авторами отмечены ошибки при подобных вычислениях и рекомендовано добавить в конце еще одну проверочную стадию, названную «verify» [Li et al., 2005; 2006].

В одной из работ при ДНК-компьютинге предложено оперировать вместо последовательностей нуклеотидов некими абстрактными символами [Nishikawa et al., 2001]. Так, например, вместо «ATC», «CTTAGC» и «AAGCCGGAT» использовались «X», «ZW» и «zyx» соответственно, что, по мнению авторов, может иметь определенные преимущества и дало им основания называть подобные симуляции как «VNA – Virtual Nucleic Acids».

Молекулярному или ДНК-компьютингу посвящено достаточно большое число обзорных, теоретических и проблемных публикаций, часть которых следует упомянуть [Landweber, Kari, 1999; Fu, 2007; Melkikh, 2008; Tagore et al., 2010; Zhang et al., 2019 и др.], поскольку из них при необходимости можно извлечь дополнительные сведения.

ДНК- цифровизация

Под ДНК-цифровизацией применительно к небиологическому использованию молекул ДНК следует понимать различные процессы, приводящие к отображению нуклеотидных последовательностей в виде некоего набора цифр в двоичном или в ином счислении, а также обратные действия по превращению цифровых записей в последовательности нуклеотидов. Подобные преобразования требуют присвоения отдельным нуклеотидам соответствующих цифровых кодов, которых уже на самом деле предложено немало. Отчасти это оправданно тем, что перевод азотистых оснований «в цифру» и обратно решает разные задачи как то: долговременное хранение небиологической информации в молекулах ДНК (олигонуклеотидах); криптографическая передача данных, включая стеганографию и др. В данном разделе все эти направления будут рассмотрены, соблюдая по возможности хронологический порядок. Начнем с ДНК-криптографии и ДНК-стеганографии, но прежде некоторое внимание уделим истории вопроса.

Впервые небиологические данные (графическое изображение) были переведены в последовательность ДНК еще в 1988 г. в рамках проекта *Microvenus* [Davis, 1996]. *Microvenus* представляла собой иконку (рис. 1), соответствующую древнегерманским рунам, олицетворявшим саму жизнь и ее женское начало.

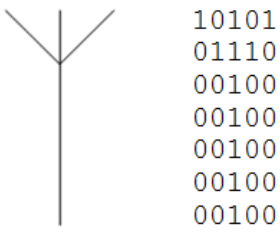


Рис.1. Иконка *Microvenus* и ее оцифровка в двоичном коде

Для отображения данной иконки с помощью четырех азотистых оснований требовался некий перевод одних данных в другие. Удобным оказался двоичный код в виде нулей и единиц, расположенный в 7 строк и содержащий 5 знаков в строке, визуально несколько напоминающий само изображение *Microvenus*. Для цифрового кодирования нуклеотидов они сначала были ранжированы по размеру и им были присвоены арабские цифры – C = 1; T = 2; A = 3; G = 4. Затем эта запись подверглась некоей трансформации и стала выглядеть как C = X; T = XX; A = XXX; G = XXXX, где X – не конкретно «0» или «1», а присутствие соответствующего нуклеотида в зависимости от количества имеющихся подряд таких X в двоичной последовательности. Используя метод оцифровки, названный phase-change code, заключающийся в смене нуклеотида, если меняется двоичное число (фаза), но при этом учитывая количество одинаковых цифр после такой смены для выбора того или иного нуклеотида, перемежающиеся нули и единицы иконки *Microvenus*, были преобразованы в последовательность нуклеотидов CCCCCCAACGCGCGCGCT.

Ввиду чрезвычайной сложности такого перевода двоичных чисел в нуклеотиды, необходимо подробнее пояснить его принцип, тем более что далее в этой статье мы опять вернемся к разным подходам к оцифровке азотистых оснований при их преобразовании в «цифру», в том числе в двоичные числа. Поскольку первая строка цифрового отображения (не кодировки) *Microvenus* выглядела как 10101, то смена фазы (нуклеотида) была произведена сразу при переходе ко второй цифре (от 1 к 0), после чего также произошла смена фазы (от 0 к 1) и так далее. На рис. 2 приведен порядок перевода цифр, соответствующих иконке *Microvenus*, в последовательность нуклеотидов.

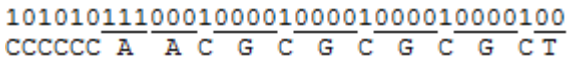


Рис. 2. Phase-change принцип оцифровки нуклеотидов. Подчеркнуты те цифры, которые не вызвали сразу после первого, второго или третьего знаков переключения фаз.

Таким образом, в случае с иконкой *Microvenus* для хранения 35 двоичных чисел потребовалось использовать 18 нуклеотидов. Кроме довольно непростого перевода двоичного кода в последовательность азотистых оснований ДНК, еще одним недостатком такого способа служило то, что одна и та же нуклеотидная последовательность могла кодироваться иным набором нулей и единиц, о чем упоминается в цитируемой статье [Davis, 1996]. Так, помимо уже приведенного выше двоичного числа 10101011100010000100001000010000100, данная последовательность нуклеотидов CCCCCCAACGCGCGCGCT может полностью соответствовать двоичному числу 01010100011101111011110111011111011, которое является как бы зеркальным отображением первого, когда нулям противопоставляются единицы. Также в статье не поясняется – а как быть, если нулей или единиц подряд будет более 4, например 5 или большее количество. Видимо после четырех «1» или «0» переключение фаз (нуклеотидов) должно производиться обязательно, а там - какому стоять нуклеотиду дальше будет также зависеть от числа нулей и единиц поочередного переключения фазы.

Как бы ни был сложен и плох описываемый выше фазовый принцип кодирования нуклеотидов с помощью двоичного кода для небиологических задач, он оказался первым, намного (больше, чем на десятилетие) опередив своих последователей. При этом цифровое кодирование *Microvenus* и перевод двоичного кода в нуклеотидную последовательность, включая ее клонирование, были выполнены в 1988 г. (секвенирование в 1990 г.), но статья, описывающая эти результаты [Davis, 1996], была опубликована только в 1996 г. после появления уже упоминавшейся выше эпохальной статьи Л. Адлемана [Adleman, 1994], сдвинувшей с мертвой точки интерес к небиологическому использованию молекул ДНК для различных целей и задач, среди которых значительную долю составляют ДНК-криптография/стеганография.

ДНК-криптография / ДНК-стеганография

Криптография (греч. κρυπτός – «скрытый» и γράφω – «пишу») – обеспечение конфиденциальности написанного и невозможности прочтения информации посторонними с помощью различных способов шифрования. Стеганография (греч. στεγανός «скрытый» и γράφω «пишу») представляет собой способ передачи или хранения информации с учетом сохранения в тайне самого факта такой передачи/хранения. Синонимом стеганографии является тайнопись. Про обычные криптографию и стеганографию написано так много, что здесь будет излишним уделять этому внимание, тогда как ДНК-криптография и ДНК-стеганография являются относительно новыми направлениями в области защиты и тайной передачи информации и еще не оправдали возлагаемых на них надежд, поскольку не все вопросы, с ними связанные, решены. Впрочем, и ДНК-криптографии и ДНК-стеганографии также посвящено уже немало обзоров [Xiao et al., 2006; Kaundal, Verma, 2014].

Под ДНК-криптографией и ДНК-стеганографией понимается использование при подобных записях для кодирования информации

молекул ДНК (олигонуклеотидов). Но для того, чтобы с помощью четырехбуквенного кода ДНК шифровать небиологическую информацию требуется соответствующий перевод одних текстов в другие. И таких способов кодировки разработано уже немало.

В 1999 г. было сообщено о зашифрованном с помощью ДНК послании, которое было в виде микрокапли нанесено на печатную точку в бумажном письме, для расшифровки которого проводилась ПЦР и затем секвенирование ДНК [Clelland et al., 1999]. Для этого был синтезирован олигонуклеотид длиной 109 звеньев, из которых в 69 содержался текст послания, а остальные 40 представляли собой места отжига (по 20 нуклеотидов) прямого и обратного праймеров. Для осуществления стеганографической передачи данной информации в ту же точку с информационным олигонуклеотидом дополнительно вносилась в довольно большом количестве фрагментированная ультразвуком (до 50–150 звеньев) денатурированная ДНК человека, что по мнению авторов полностью исключало возможность установления информационной последовательности, которую можно было только амплифицировать, зная последовательности праймеров и «ключи» для дешифровки последовательности нуклеотидов в обычный английский текст. Помимо алфавита из 26 букв, с помощью триплетов ДНК были кодированы также арабские цифры и некоторые знаки препинания, включая пробел между словами, приведенными здесь в виде таблицы (см. таблицу 1).

Таблица 1

Таблица шифрования с помощью ДНК букв английского алфавита, арабских цифр, знаков препинания, пробела [Clelland et al., 1999]

A=CGA	K=AAG	U=CTG	0=ACT
B=CCA	L=TGC	V=CCT	1=ACC
C=GTT	M=TCC	W=CCG	2=TAG
D=TTG	N=TCT	X=CTA	3=GAC
E=GGC	O=GGA	Y=AAA	4=GAG
F=GGT	P=GTG	Z=CTT	5=AGA
G=TTT	Q=AAC	_=ATA	6=TTA
H=CGC	R=TCA	,=TCG	7=ACA
I=ATG	S=ACG	.=GAT	8=AGG
J=AGT	T=TTC	:=GCT	9=GCG

Зашифрованным в цитируемой работе [Clelland et al., 1999] оказалось послание «JUNE 6 INVASION : NORMANDY». На данный способ стенографической передачи данных этими авторами 6 ноября 2001 г. получен патент США US 6,312,911 [Bancroft et al., 2001].

Справедливости ради, здесь следует вспомнить оставшуюся практически незамеченной одну статью под кратким, но броским названием «A DNA Text Code», опубликованную на десятилетие раньше в журнале BioTechniques [Hodgson, 1990]. В этой статье впервые было предложено кодирование азотистыми основаниями арабских цифр, английских букв, довольно большого числа символов, нескольких слов и даже одной хорошо известной молекулярным биологам аббревиатуры температуры плавления ДНК (Tm), используя от одного до трех нуклеотидов (см. таблицу 2).

Таблица 2

Кодирование букв английского алфавита, арабских цифр, различных символов, отдельных слов и аббревиатур нуклеотидами и их последовательностями (по Hodgson, 1990)

Символ	Код	Символ	Код
(space)	A	,	TAC
0	C	.	TAG
1	G	:	TAT
2	T	;	TCA
3	AA	‘	TCT
4	AC	"	TGA
5	AG	&	TGT
6	AT	•	TTA
7	CA	(	TTC
8	CC	)	TTG
9	CT	-	TTT
A	GA	_	CTA
B	GC	+	CTC
C	GG	=	CTG
D	GT	?	GTT
E	TA	!	GAA
F	TC	@	GAC
G	TG	©	GAG
H	TT	$	GAT
I	AAA	%	GCA
J	AAC	#	GCT
K	AAG	\	GGA
L	AAT	/	GTA
M	ACA	<	GTC
N	ACC	>	GTG
O	ACT	[	GTT
P	AGA	]	CG
Q	AGT	^	ACG
R	ATA	{	AGC
S	ATC	&	AGG
T	ATG	}	CCA
U	ATT	~	CCC
V	CAA	Tm	CCT
w	CAC	patent	GCC
X	CAG	patent	GGC
Y	CAT	pending	
Z	TAA		

Несмотря на то, что в этой публикации [Hodgson, 1990] приводится кодировка текста «© DNACO 1990» в виде нуклеотидной последовательности

GAGAGTACCGAGGACTAGCTCTC, вероятно все же автор ограничился только теоретической проработкой данного вопроса и экспериментального исследования не провел, иначе бы столкнулся с невозможностью применения на практике предложенного им способа кодирования текстовой информации с помощью одного-трех нуклеотидов из-за неоднозначности установления чему соответствует, например А - кодирует ли пробел или являются частью кодов АА, АС, AG, AT, AAA, AAC и т.д., ответственных за иную информацию. Подобная гетероразмерная кодировка несколько напоминает алгоритм Хаффмана, к рассмотрению которого вскоре перейдем, но забегая вперед скажем, что в том коде эти вопросы тщательно продуманы. При этом в данной статье рассмотрены возможные применения такого кодирования разной информации в виде соответствующих нуклеотидных последовательностей, которые могут использоваться, например, для защиты авторских прав при обнаружении новых генов, при создании рекомбинантных организмов. Также в статье говорится об использовании таких ДНК-кодов для предупреждения плагиата до публикации научной или художественной композиции, но не поясняется как это может производиться, отмечая, что шифрование является интуитивно очевидным способом идентификации генетических изобретений. Также автор пишет (в несколько вольном переводе), что «хотя эти вопросы, концептуально связанные с ДНК-фингерпринтированием, уже обсуждались ранее как **анекдот**[8], но это, по-видимому, первое сообщение, предлагающее для шифрования стандартный код из четверок нуклеотидов ДНК». По всей видимости, это так и есть. Удивительно, но данная работа (пусть и не полностью корректная), явно опередившая свое время, была согласно базе данных Scopus, процитирована всего однажды в 1992 г. в обзорной статье, посвященной вопросам автоматизации детекции специфических фрагментов ДНК, где ей уделено совсем мало внимания в связи с обсуждением возможности обнаружения внедренных фрагментов ДНК, несущих некую небиологическую информацию, при их использования в качестве меток в трансгенных организмах [Landegren, 1992]. Еще одно цитирование этой работы нам удалось встретить в патенте США за номером 6,537,747 от 25 марта 2003 г., в котором описывается способ с использованием твердой фазы передачи информации, кодированной в 40-звенных олигонуклеотидах, состоящих из трех частей, две из которых по краям служат местами отжига праймеров [Mills, Yurke, 2003].

Однако трехнуклеотидное кодирование букв, цифр и различных знаков все же ограничено 64 комбинациями (в работе Clelland и соавт. [1999] было использовано всего 40 таких комбинаций), что не дает возможности кодирования большего числа символов, необходимых для скрытой передачи более полноценной информации. По этой причине было предложено использовать для такой цели комбинации из четырех нуклеотидов, которых может быть 256 вариантов, но не все из них из-за особенностей их последовательностей пригодны для кодирования. Причем в работах разных авторов можно встретить различные комбинации четверок нуклеотидов.

Так, 95 кодонов было использовано в одной из работ, 52 из которых соответствовали прописным и строчным буквам английского алфавита (таблица 3), а оставшиеся 43 кодировали различные символы, включая «пробел» (таблица 4) [Agrawal et al., 2012].

Таблица 3

Кодировка букв английского алфавита четырьмя нуклеотидами по Agrawal et al., 2012.

буквы английского алфавита	кодоны	буквы английского алфавита	кодоны
A	ATCG	a	CCAG
B	ATGC	b	CCGA
C	AGTC	c	CCAT
D	AGCT	d	CCTA
E	ACGT	e	GGAT
F	ACTG	f	GGTA
G	CATG	g	GGCT
H	CAGT	h	GGTC
I	CGAT	i	GGAC
J	CGTA	j	GGCA
K	CTAG	k	TTGA
L	CTGA	l	TTAG
M	TACG	m	TTCA
N	TAGC	n	TTAC
O	TCAG	o	TTCG
P	TCGA	p	TTGC
Q	TGAC	q	TTTT
R	TGCA	r	GGGG
S	GTCA	s	AAAA
T	GTAC	t	CCCC
U	GATC	u	TTTA
V	GACT	v	TTTG
W	GCTA	w	TTTC
X	GCAT	x	GGGA
Y	AACG	y	GGGT
Z	AAGC	z	GGGC

[8] Жирным шрифтом выделено нами; при этом в оригинале использовано слово «anecdotally».

Таблица 4

Кодировка различных символов четырьмя нуклеотидами (по Agrawal et al., 2012)

Символы	кодоны	Символы	кодоны
1	CCCA	+	CGCG
2	CCCG	=	AATT
3	CCTT	{	AAAC
4	CCGG	}	AAAG
5	CCAA	[	AAGT
6	TTCC	]	AACT
7	TTAA	|	AAAT
8	TTGG	\	AATG
9	AAGG	;	CACA
0	AACC	:	TCTC
!	GGAA	“	TGTG
@	GGTT	‘	TATA
#	GAGA	<	TAAA
$	GTGT	>	CAAA
%	GCGC	,	ATTT
^	AATC	.	CTTT
&	ACAC	?	GAAA
*	AGAG	/	GTTT
(	ATAT	_	CCTG
)	CTCT	space	AGGG
-	AACC	~	GGAA
`	CCGT		

В одной из статей [Jimenez-Sanchez, 2013] для перевода букв и некоторых символов в тетраплеты нуклеотидов была задействована ASCII кодировка, приведенная в таблице 5. Автор ввел новые термины tyte – tetramer byte, соответствующий четырем азотистым основаниям, каждое из которых дает 2 бита информации, что как раз составляет 1 байт, и tet – tetramer bit.

Таблица 5

Буквы, символы и ASCII кодировка и нуклеотиды (по Jimenez-Sanchez, 2013 с некоторыми изменениями)

Буквы и символы	ASCII	tyte	Буквы и символы	ASCII	tyte	Буквы и символы	ASCII	tyte
NULL	0	AAAA						
space	32	ACAA	@	64	TAAA	`	96	TCAA
!	33	ACAT	A	65	TAAT	a	97	TCAT
"	34	ACAC	B	66	TAAC	b	98	TCAC
#	35	ACAG	C	67	TAAG	c	99	TCAG
$	36	ACTA	D	68	TATA	d	100	TCTA
%	37	ACTT	E	69	TATT	e	101	TCTT
&	38	ACTC	F	70	TATC	f	102	TCTC
'	39	ACTG	G	71	TATG	g	103	TCTG
(	40	ACCA	H	72	TACA	h	104	TCCA
)	41	ACCT	I	73	TACT	i	105	TCCT
*	42	ACCC	J	74	TACC	j	106	TCCC
+	43	ACCG	K	75	TACG	k	107	TCCG
,	44	ACGA	L	76	TAGA	l	108	TCGA
-	45	ACGT	M	77	TAGT	m	109	TCGT
	46	ACGC	N	78	TAGC	n	110	TCGC
/	47	ACGG	O	79	TAGG	o	111	TCGG
0	48	AGAA	P	80	TTAA	p	112	TGAA
1	49	AGAT	Q	81	TTAT	q	113	TGAT
2	50	AGAC	R	82	TTAC	r	114	TGAC
3	51	AGAG	S	83	TTAG	s	115	TGAG
4	52	AGTA	T	84	TTTA	t	116	TGTA
5	53	AGTT	U	85	TTTT	u	117	TGTT
6	54	AGTC	V	86	TTTC	v	118	TGTC
7	55	AGTG	W	87	TTTG	w	119	TGTG
8	56	AGCA	X	88	TTCA	x	120	TGCA
9	57	AGCT	Y	89	TTCT	y	121	TGCT
:	58	AGCC	Z	90	TTCC	z	122	TGCC
;	59	AGCG	[	91	TTCG	{	123	TGCG
<	60	AGGA	\	92	TTGA	}	124	TGGA
=	61	AGGT	]	93	TTGT	~	125	TGGT
>	62	AGGC	^	94	TTGC			
?	63	AGGG	_	95	TTGG		255	GGGG

Однако в предложенном варианте оказалось задействовано всего 96 комбинаций четырех нуклеотидов из 256 возможных, при этом отмечается, что данный принцип обеспечивает униформное кодирование в виде 4 нуклеотидов на букву или символ, а также «пробел». В статье приводятся и другие преимущества подобного способа кодирования небиологической информации азотистыми основаниями, в том числе, позволяющими выявлять ошибки при кодировании и декодировании данных. Несколько позже другие авторы также использовали 96 комбинаций четырех нуклеотидов [UbaidurRahman et al., 2015] (см. таблицу 6).

Таблица 6

Шифрование букв латинского алфавита и арабских цифр с помощью четырехнуклеотидного кода ДНК [UbaidurRahman et al., 2015]

ACAT – a	AAAA – y	ATAA – W	AGAG – {
ACTG – b	AATT – z	ATTT – X	AGTA – [
ACCC – c	AACC – A	ATCG – Y	AGCG – }
ACGA – d	AAGG – B	ATGC – Z	AGGG –]
TCAT – e	TAAT – C	TTAA – 0	TGAA – |
TCTG – f	TATG – D	TTTT – 1	TGTT – \
TCCG – g	TACC – E	TTCC – 2	TGCG – +
TCGT – h	TAGA – F	TTGG – 3	TGGC – =
CCAG – i	CAAT – G	CTAT – 4	CGAA – _
CCTA – j	CATG – H	CTTG – 5	CGTT – -
CCCG – k	CACG – I	CTCC – 6	CGCC –)
CCGG – l	CAGT – J	CTGA – 7	CGGG – (
GCAA – m	GAAG – K	GTAT – 8	GGAT – *
GCTT – n	GATA – L	GTTG – 9	GGTG – &
GCCG – o	GACG – M	GTCG – <	GGCC – ^
GCGC – p	GAGG – N	GTGT – >	GGGA – %
ACTC – q	AATA – O	ATTA – ,	AGTT – $
ACCG – r	AACG – P	ATCC – .	AGCC – #
TCTC – s	TATC – Q	TTTA – ?	TGTA – @
TCCC – t	TACG – R	TTCG – /	TGCC – !
CCTT – u	CATC – S	CTTC – :	CGTA – ~
CCCC – v	CACC – T	CTCG – ;	CGCG –
GCTA – w	GATT – U	GTTC – “	GGTC – €
GCCC – x	GACC – V	GTCC – ‘	GGCG – £

В еще одной статье предложена несколько иная кодировка английского алфавита и различных символов Rashid et al., 2017 (см. таблицу 7).

Как можно видеть из таблиц 6 и 7 с кодированием четырьмя нуклеотидами букв английского алфавита и различных символов, единого подхода нет и возможно это даже оправдано, поскольку такое кодирование используется для ДНК-криптографии, фактически еще больше запутывающей тех, кому послание не предназначается, а прочесть его злоумышленникам хочется. Существуют и иные способы кодирования нуклеотидами текстовой и прочей информации.

Таблица 7

Кодировка букв английского алфавита и символов четырьмя нуклеотидами по Rashid et al., 2017.

ACAT–a	AAAA–y	ATAA–W	AGAG–{
ACTG–b	AATT–z	ATTT–X	AGTA–[
ACCC–c	AACC–A	ATCG–Y	AGCG-}
ACGA–d	AAGG–B	ATGC–Z	AGGG-]
TCAT–e	TAAT–C	TTAA–0	TGAA-|
TCTG–f	TATG–D	TTTT–1	TGTT-\
TCCG–g	TACC–E	TTCC–2	TGCG-+
TCGT–h	TAGA–F	TTGG–3	TGGC-=
CCAG–i	CAAT–G	CTAT–4	CGAA-_
CCTA–j	CATG–H	CTTG–5	CGTT—
CCCG–k	CACG–I	CTCC–6	CGCC-)
CCGG–l	CAGT–J	CTGA–7	CGGG–(
GCAA–m	GAAG–K	GTAT–8	GGAT-*
GCTT–n	GATA–L	GTTG–9	GGTG-&
GCCG–o	GACG–M	GTCG-<	GGCC-^
GCGC–p	GAGG–N	GTGT->	GGGA-%
ACTC–q	AATA–O	ATTA-,	AGTT-$
ACCG–r	AACG–P	ATCC-.	AGCC-#
TCTC–s	TATC–Q	TTTA-?	TGTA-@
TCCC–t	TACG-R	TTCG-/	TGCC-!
CCTT–u	CATC–S	CTTC-:	CGTA–~
CCCC–v	CACC–T	CTCG-;	CGCG–‘
GCTA–w	GATT–U	GTTC–“	GGTC-€
GCCC–x	GACC–V	GTCC–‘	GGCG-£

Еще в начале 1950-х гг. был предложен довольно оригинальный «жадный» способ кодирования арабских цифр и некоторых чисел в двоичном счислении [Huffman, 1952]. Его особенностью можно считать увеличивающееся число двоичных знаков. В частности, в приведенном в той статье примере (фрагмент которого представлен в таблице 8) для арабских чисел от 1 до 13 потребовалось от 2 до 6 нулей и единиц со средней длиной 3.42 двоичных знака на одно арабское число.

Таблица 8

Бинарное кодирование арабских цифр и чисел по Huffman [1952] с изменениями

i	*Code*	*L(i)*
1	10	2
2	000	3
3	011	3
4	110	3
5	111	3
6	0101	4
7	00100	5
8	00101	5
9	01000	5
10	01001	5
11	00110	5
12	001110	6
13	001111	6

Позже этот принцип кодирования двоичными числами по алгоритму Хаффмана был применен вместо арабских чисел и двоичных чисел для четырех азотистых оснований и 26 букв английского алфавита [Smith et al., 2003], в котором принимались во внимание их частоты встречаемости в английском языке, что в несколько измененном виде представлено нами в таблице 9. Здесь необходимо заметить, что в цитируемой статье под понятием «кодон» подразумевается участок ДНК (даже из одного нуклеотида), обеспечивающий кодирование одной буквы алфавита.

Таблица 9

Кодирование букв английского алфавита нуклеотидами ДНК по алгоритм Хаффмана [Huffman, 1952]

Letter	Frequency (%)	Codon	Length
e	12.7	T	1
t	9.1	AG	2
a	8.2	AT	2
o	7.5	GA	2
i	7	GG	2
n	6.7	GC	2
s	6.3	GT	2
h	6.1	CA	2
r	6	CG	2
d	4.3	CT	2
l	4	AAA	3
c	2.8	AAG	3
u	2.8	AAC	3
w	2.4	AAT	3
m	2.4	ACA	3
f	2.2	ACG	3
y	2	ACC	3
g	2	ACT	3
p	1.9	CCA	3
b	1.5	CCG	3
v	1	CCT	3
k	0.8	CCCA	4
j	0.2	CCCG	4
x	0.2	CCCC	4
q	0.1	CCCTA	5
z	0.1	CCCTG	5

Как можно видеть из данной таблицы, для кодирования разных букв этим способом потребовалось использовать от одного до 5 нуклеотидов, и таким образом средняя длина кодона составила приблизительно 2,2 нуклеотида. С помощью алгоритма Хаффмана фамилия Huffman будет однозначно кодироваться как CA AAC ACG ACG ACA AT GC или слитно как CAAACACGACGACAATGC. Причем, какое другое прочтение (кодирование и декодирование) при использовании такого принципа кодировки невозможно.

Помимо алгоритма Хаффмана в этой статье [Smith et al., 2003] приводятся еще два вида кодов – Comma code и Alternating Code. Если для алгоритма Хаффмана существует 26 кодонов, то для Comma Code и Alternating Code таковых может быть 80 и 64 варианта соответственно. При этом длины кодонов для них фиксированы и содержат по 6 нуклеотидов, тогда как в алгоритме Хаффмана от 1 до 5 для приведенного фрагмента такого кода. Comma Code представляет собой гексануклеотидный участок ДНК, всегда начинающийся с гуанина, что удобно тем, что всегда известно начало места кодирования, после которых следует 5 азотистых оснований из двух цитозинов и трех аденинов или тиминов. То есть последовательности таких кодонов для Comma Code могут быть выражены как GCWWWC или GWCWWC и т.д., где под «и т.д.» понимается все остальные возможные комбинации (78) таких последовательностей, а W соответствует общепринятому однобуквенному кодированию аденинов или тиминов, нахождение которых допускается в одном месте. Alternating Code представляет собой шестерку перемежающихся пуринов и пиримидинов, числом по три для каждого типа азотистых оснований. Таким образом, кодоны в Alternating Code могут быть RYRYRY, RYYRRY и т.д. (еще 62 комбинации), где R и Y соответствуют однобуквенному обозначению пуринов и пиримидинов соответственно.

С целью увеличения емкости для небиологических данных, хранимых в молекулах ДНК, был предложен оригинальный подход с использованием ПЦР со вложенными праймерами, получивший название NPMM (Nested Primer Molecular Memory) и заключавшийся в весьма сложной организации олигонуклеотидов, используемых для хранения информации, состоящих из средней информационной части, фланкированной составными последовательностями [Kashiwamura et a., 2005]. В своей следующей работе эти авторы [Yamamoto et al., 2008] еще больше усложнили адресующие последовательности, состоящие из трех блоков каждый и расположенные по краям информационных последовательностей размерами в 20, 40 и 60 звеньев, что привело к формированию почти 17 млн. уникальных «адресов».

Помимо секвенирования для извлечения из ДНК закодированной в ней небиологической информации применяются и другие подходы. Так, в одной из работ были сконструированы самособирающиеся олигонуклеотидные наноконструкции, имеющие в зависимости от числа молекул, вступивших в молекулярную гибридизацию, восемь различных состояний, которые авторы оцифровали как 000, 100, 010, 001, 110, 101, 011, 111

[Shin, Pierce, 2004]. Причем все эти состояния созданных наноконструкций можно было контролировать как с помощью гель-электрофореза, так и детектируя изменения флуоресценции, поскольку используемые олигонуклеотиды были мечены разными флуорохромами. Цифровизация различного состояния молекул ДНК оценивалась также с помощью рестрикционного расщепления неких фрагментов ДНК с преобразованием результатов гель-электрофорез в «нули» и «единицы». Так, были сконструированы олигонуклеотиды по 77 звеньев длиной, содержащие среднюю информационную часть и фланкирующие ее участки с адресацией, потенциально несущие два сайта узнавания рестрикционной эндонуклеазы *Eco*RI, в результате ферментативного действия которой происходило расщепление или обоих или только первого из них или только второго сайтов, либо не происходило вовсе, что выяснялось с помощью гель-электрофореза и при переводе результатов электрофоретического разделения в «цифру» давало соответственно такие двоичные числа – «00», «01», «10» и «11» [Skinner et al., 2007]. Другими авторами использовалось частичное (неполное) расщепление ДНК рестрикционными эндонуклеазами, также сопровождавшееся разделением фрагментов ДНК разного размера гель-электрофорезом, после чего выявляемые полосы ДНК подвергали оцифровке двоичными числами в виде «1» и «0» в зависимости от их размеров [Portney et al., 2008]. В этой работе было закодировано слово «MEMO», для чего потребовалось 12 бит и 110 азотистых оснований. Авторы отметили, что их подход более дешев, поскольку не требует определения последовательности нуклеотидов, однако для создания однозначной элекрофоретической картины после недорасщепления ДНК пришлось использовать концевое мечение анализируемого фрагмента ДНК радиоактивным изотопом.

С ДНК-стеганографией довольно тесно соприкасаются использование так называемых «ДНК-водяных знаков» в виде внедренных в геномы тех или иных генно-модифицированных организмов неких фрагментов ДНК, призванных охранять права авторов на такие объекты их интеллектуальной собственности, и хотя эти фрагменты ДНК также предлагается определенным образом оцифровывать, поскольку в этих случаях речь идет о живых организмах, эта информация выходит за рамки данной статьи и поэтому ей должна быть посвящена отдельная публикация.

Помимо непосредственного перевода текстов в нуклеотидные последовательности существует также возможность предварительной кодировки написанного в виде двоичных кодов, которые затем преобразуются в нуклеотидные последовательности, используя соответствующие кодировки, которых также имеется немало, но к их описанию мы приступим при рассмотрении вопросов хранения в молекулах ДНК (олигонуклеотидах) небиологической информации.

Долговременное хранение небиологической информации в молекулах ДНК (олигонуклеотидах)

За несколько тысяч лет своей истории человечество накопило большое количество материалов, которыми так или иначе следует пользоваться и при этом необходимо их надежно хранить. После скальных, глиняных, папирусных, берестяных, бумажных и иных поверхностей в конце 1920-х гг. первым высокоэффективным хранилищем звуковой и затем прочей информации стала магнитная лента, широко применяющаяся до сих пор. Позже появились и другие носители информации, каждый из которых имеет свои преимущества и недостатки. Но прежде чем перейти к их краткому рассмотрению необходимо перечислить объемы памяти (таблица 10), которыми человечество уже широко оперирует и/или через некоторый довольно небольшой отрезок времени подойдет к таковым вплотную.

Таблица 10

Характеристики компьютерной памяти

Объем памяти	Сокращение	Кол-во
байт*	байт, Б	10^0
килобайт	кбайт, Кб	10^3
мегабайт	Мбайт, Мб	10^6
гигабайт	Гбайт, Гб	10^9
терабайт	Тбайт, Тб	10^{12}
петабайт	Пбайт, Пб	10^{15}
экзабайт	Эбайт, Эб	10^{18}
зеттабайт	Збайт, Зб	10^{21}
йоттабайт	Йбайт, Йб	10^{24}

* 1 байт = 8 бит

Как можно видеть из сноски к этой таблице сейчас принято, что один байт состоит из 8 бит или бинарных цифр (binary digit). Винчестеры персональных компьютеров еще не так давно (по крайней мере, ряд авторов данной статьи их еще застали) имели емкость всего 5 Мб и оперативную память 640 Кб, а сейчас один видеофайл может иметь размер, гораздо больший, чем те пресловутые 5 Мб. В настоящее время терабайтным винчестером уже никого не удивишь, но более высокие емкости памяти в обыденной жизни не используются и являются уделом серверов, различных дата-центров.

Подсчитано, что вся накопленная человечеством информация составляет более 20 зеттабайт, и при этом растет с каждым годом приблизительно наполовину. По другим оценкам человечество в 2025 году произведет 160 зеттабайт информации. И ее объем продолжит бурно расти. Прогнозируется, что к 2040 году накопленные данные

составят более 3 гептиллионов байт или 3 йоттабайт, для хранения которых потребуется более 10^9 кг кремния особой чистоты и это может оказаться критичным, поскольку считается, что к тому времени его будет произведено всего 10^7–10^8 кг [Zhirnov et al., 2018]. При этом получение такого высокоочищенного кремния является весьма грязным производством, потребляющим немалые сырьевые и энергетические ресурсы [Williams et al., 2002].

Поиск новых носителей информации и способов хранения данных в виде двоичных чисел активно ведется и среди них одно из первых мест занимают молекулы ДНК, имеющие как определенные недостатки так и важные преимущества. Среди последних – возможность долговременного хранения и огромный объем данных. Главным недостатком является очень большое время доступа к хранящейся информации, значительно уступающей по этому параметру остальным носителям компьютерной информации, сведенным в таблицу 11. Кроме этих машиночитаемых носителей значительное количество хранимой информации представлено на бумаге и микрофишах, срок службы которых составляет по некоторым оценкам около 2 тысяч лет и 700 лет соответственно.

Таблица 11

Некоторые параметры основных носителей компьютерной информации

Тип носителя	Емкость	Время доступа	Срок службы
SSD накопитель	до 1 терабайта	микросекунды	несколько месяцев или лет
CD; DVD; Blue-Ray диски	до 128 гигабайт	десятки микросекунд	несколько десятилетий
HDD винчестер	до 10 терабайт	десятки микросекунд	около десятилетия
Магнитная лента	до 100 терабайт	минуты	несколько десятков лет
ДНК (олигонуклеотиды)	экзабайты	часы/дни	несколько сотен лет

Помимо этих параметров следует остановиться еще на одном – физическом объеме хранимой информации. Так, подсчитано, что 1 г одноцепочечной ДНК (олигонуклеотидов) может хранить около 500 квинтиллионов байт или 500 экзабайт (10^{18}) информации, что эквивалентно приблизительно 100 млрд. обычных DVD дисковв. На самом деле такой подсчет (фигурирующий во многих статьях и обзорах) не совсем корректен, поскольку в расчет берется обращение с единичными молекулами ДНК, тогда как в силу ряда причин это технически невозможно и вряд ли когда-нибудь станет возможно оперирование для целей долговременного хранения негенетической информации менее чем 100–1000 копиями одинаковых молекул ДНК (олигонуклеотидов). Поэтому необходимо уменьшать такие числа на два-три порядка, но даже при таком пересчете емкость молекул ДНК намного превосходит все остальные носители информации.

Как уже отмечалось выше первым, кто обратил внимание на ДНК, как на потенциальный носитель небиологической информации был отечественный ученый М.С. Нейман [1964], но его высказывание было довольно абстрактным. В другой эпохальной статье на эту тему L. Adleman [1994] также довольно кратко коснулся возможности хранения некой информации в ДНК, отметив при этом, что для одного бита будет достаточно всего 1 нм^3. Впервые компьютерные нули и единицы в связи с хранением в ДНК небиологических данных были упомянуты в теоретической статье, посвященной вопросу организации машинной памяти в молекулах ДНК [Baum, 1995]. Однако и в этой статье 0 и 1 предлагалось присваивать неким подпоследовательностям ДНК, а не отдельным нуклеотидам.

Пожалуй, первой работой, ставившей четкой целью долговременное хранение в ДНК текстовой информации можно считать статью [Bancroft et al., 2001], принадлежащую той же группе авторов, ранее закодировавших в ДНК упоминавшееся выше короткое послание «JUNE 6 : INVASIONS : NORMANDY» и переславших его в виде ДНК-микроточки на фильтровальной бумаге [Clelland et al., 1999]. В этот раз они закодировали фразы из повести Ч.Диккенса «IT WAS THE BEST OF TIMES IT WAS THE WORST OF TIMES. IT WAS THE AGE OF FOOLISHNESS IT WAS THE EPOCH OF BELIEF», акцентируя внимание на четырехкратном повторении слов «it was the», послужившим неким показателем возможности оперирования для таких целей повторяющейся ДНК. Для кодирования данной текстовой информации была разработана особая конструкция олигонуклеотидов, состоявших из центральной части, фланкированной участками, служившими местами отжига прямого и обратного праймеров (до 20 звеньев длиной), и некоего маленького спейсера (3–4 нуклеотида), служащего указателем начала кодирующего участка. При этом кодирующая последовательность формировалась из трех нуклеотидов – A, C и T, тогда как в состав фланкирующих участков для отжига праймеров входили и гуанины с таким расчетом, чтобы они оказывались в них каждыми четвертыми. Тем самым была достигнута ситуация, когда праймеры не могли эффективно отжигаться на информационных участках, что очень важно. К сожалению, авторы не сообщили

используемый ими код, но в дополнительной информации к статье была приведена расшифровка секвенированной автоматическим секвенатором нуклеотидной последовательности с наложенным на нее вышеупомянутым текстом, из чего можно восстановить каким триплетам соответствуют используемые в данных фразах 17 букв английского алфавита, а также «пробел», приведенные в таблице 12 заглавными буквами. Строчными буквами в данной таблице приведена восстановленная уже нами кодировка оставшихся 9 букв, исходя из выявленной закономерности.

Таблица 12

Кодирование английских букв нуклеотидами (по Bancroft et al., 2001)

Буква	Код	N	CCC
A	AAA	O	CCT
B	AAC	P	CTA
C	AAT	Q	ctc
D	aca	R	CTT
E	ACC	S	TAA
F	ACT	T	TAC
G	ATA	U	tat
H	ATC	V	tca
I	ATT	W	TCC
J	caa	X	tct
K	cac	Y	tta
L	CAT	Z	ttc
M	CCA	space	TTT

Здесь также можно заметить, что использование трехнуклеотидного кода для 26 букв английского алфавита и пробела как раз удачно соответствуют числу возможных комбинаций перестановок трех нуклеотидов – $3^3 = 27$. Однако в этой работе прямого соответствия компьютерных «нулей» и «единиц» азотистым основаниям продемонстрировано не было.

Следующей работой, описывающей долговременное хранение в ДНК небиологической информации, в которой авторы также обошлись без кодирования в виде «0» и «1» нуклеотидов напрямую, а прибегли к ASCII кодам, стала статья [Wong et al., 2003]. В таблице 13 приведены используемые в цитируемой статье кодировки букв, цифр и знаков азотистыми основаниями. Особенностью данной работы было то, что закодированные отрывки детской песенки «It's a Small World» были внедрены для хранения в бактерии *Escherichia coli* (как промежуточный хозяин; в плазмидном векторе) и затем в геном *Deinococcus radiodurans* (взятого как вид, устойчивый к высоким дозам радиации). Для этого были использованы 7 химически синтезированных фрагментов ДНК размерами от 57 до 99 п.н., фланкированные по краям участками для отжига праймеров, что позволило амплифицировать клонированные фрагменты ДНК и затем их просеквенировать.

Таблица 13

Кодировка букв, цифр, знаков (по Wong et al., 2003)

AAA – 0	AAC – 1	AAG – 2	AAT – 3	ACA – 4	ACC – 5	ACG – 6	ACT – 7
AGA – 8	AGC – 9	AGG – A	AGT – B	ATA – C	ATC – D	ATG – E	ATT – F
CAA – G	CAC – H	CAG – I	CAT – J	CCA – K	CCC – L	CCG – M	CCT – N
CGA – O	CGC – P	CGG – Q	CGT – R	CTA – S	CTC – T	CTG – U	CTT – V
GAA – W	GAC – X	GAG – Y	GAT – Z	GCA – SP	GCC – :	GCG – ,	GCT – -
GGA – .	GGC – !	GGG – (	GGT –)	GTA – `	GTC – ‘	GTG – “	GTT – "
TAA – ?	TAC – ;	TAG – /	TAT – [	TCA –]	TCC –	TCG –	TCT –
TGA –	TGC –	TGG –	TGT –	TTA –	TTC –	TTG –	TTT –

Как можно видеть из данной таблицы из 64-х возможных комбинаций трех нуклеотидов не все оказались задействованы для кодирования.

Несколько иная (шестибитная) кодировка арабских цифр, букв английского алфавита и ряда знаков (таблица 14) была предложена в работе [Arita, Ohashi, 2004], целью которой было создать так называемые «водяные знаки» в геноме *Bacillus subtilis* путем сайт-направленного мутагенеза ряда кодонов в одном из белков, благодаря чему в гене этого белка оказалось закодировано слово KEIO по названию мест работы авторов – Keio University.

Таблица 14

Шестибитная кодировка букв английского алфавита и некоторых символов по [Arita, Ohashi, 2004]

000001 ‘‘	010011 R	001110 F	110100 J
000010 E	100011 H	010110 G	111000 Q
000100 T	001101 D	011010 W	011111 Z
001000 A	010101 L	011100 Y	101111 ′
010000 O	100101 C	100110 B	110111 ,
100000 S	011001 M	101010 V	111011 &
000111 N	101001 U	101100 K	111101
001011 I	110001 P	110010 X	111110

Не указанный в таблице код 000000 соответствовал пробелу.

В одной из теоретических работ, посвященной вопросам хранения в ДНК небиологической информации, упоминается вариант однобитного кодирования нуклеотидов для пар азотистых оснований, находящихся на комплементарных цепях и кодируемых как «1» и «0» соответственно для AT-пар и GC-пар [Arita, 2004]. Таким образом, нуклеотиды A и T кодировались «1», а G и C – «0». Дальнейшего развития этот принцип кодирования комплементарных оснований в двухцепочных структурах ДНК не получил.

Также было предложено отображать нуклеотидную последовательность ДНК в виде двоичного кода с помощью неких индикаторов [Nair, Mahalakshmi, 2005]. Суть такого подхода, который вряд ли можно признать экономящим дисковое пространство, заключается в применении индикаторов i_A, i_G, i_C, i_T, каждый из которых помещается в начале своей строки, после чего следует перемежающиеся нули («0») и единицы («1»). В качестве примера авторы привели последовательность AGTTCTACCGAGC и ее бинарное кодирование:

i_A = 1000001000100; i_G = 0100000001010;

i_C = 0000100110001; i_T = 0011010000000.

То есть данные индикаторы (i_A, i_G, i_C, i_T) указывают на присутствие в последовательности нуклеотида, за который они «отвечают», в виде компьютерной «1», тогда как «0» соответствуют в каждой такой строке трем любым другим нуклеотидам.

Задумав внедрить в геном Bacillus subtilis текстовую информацию «E=mc^2 1905!», авторы [Yachie et al., 2007] сначала перевели ее в код сканирования клавиатуры, затем в шестнадцатиричный код и потом в двоичные числа. 16-ти комбинациям динуклеотидов были присвоены двоичные четырехразрядные коды, приведенные в таблице 15.

Таблица 15

Кодирование динуклеотидов двоичными четырехразрядными числами

AA = 0000	AG = 1000
CA = 0001	CG = 1001
GA = 0010	GG = 1010
TA = 0011	TG = 1011
AC = 0100	AT = 1100
CC = 0101	CT = 1101
GC = 0110	GT = 1110
TC = 0111	TT = 1111

В результате для ДНК-кодирования текстовой информации «E=mc^2 1905!» потребовалось создать 4 кассеты из химически синтезированных олигонуклеотидов размерами 64, 62, 62, 62 звена, которые в составе подготовленных конструкций были внедрены в геном бациллы и после успешно секвенированы.

Также с целью маркирования рекомбинантных организмов в геном *Bacillus subtilis*, в частности был внедрен фрагмент ДНК, с закодированной с помощью ASCII в нем аббревиатурой ICSP08 (International Conference on Signal Processing) для чего потребовалось 48 бит, которые были переведены в последовательность ДНК, исходя из правила кодирования по которому нуклеотиды A, C, G, T соответствовали двоичным числам 00, 01, 10 и 11 [Jiao, Goutte, 2008].

Нельзя не задержать внимание на статье, в которой в одной молекуле ДНК оказались закодированы сразу текст, мелодия детской песенки «Mary Had a Little Lamb», а также графическое изображение овечки в виде простенького детского рисунка [Ailenberg, Rotstein, 2009]. Для этого потребовалось создать фрагмент ДНК размером 844 п.н., который был клонирован в плазмидном векторе. Для его секвенирования (для извлечения текстовых, музыкальных и графических данных) использовались две пары праймеров, образующие ампликоны из 500 и 344 п.н. Для кодирования этой небиологической информации был предложен модифицированный способ кодирования азотистыми основаниями различных букв, знаков, символов, нот по алгоритму Хаффмана, отличающийся от его варианта, ранее переложенного для ДНК [Smith et al., 2003]. В частности, было уменьшено число G и C нуклеотидов с целью уменьшить вероятность образования нежелательных вторичных структур. Причем авторы отмечают, что хранить плазмиды отдельно в виде ДНК, а не в живом организме более предпочтительно, поскольку можно не опасаться мутаций. Ниже приведены три таблицы (16, 17 и 18), в которых представлены способы кодирования азотистыми основаниями музыкальной, текстовой и графической информации.

Таблица 16

Кодирование нот и музыкальных символов нуклеотидами по алгоритму Хаффмана

Кодон ДНК	Музыкальный код
G	Quarter rote (1/4)
TT	Half note (1/2)
TA	Whole note (1)
AT	Eighth note (1/8)
CT	Sixteenth note (1/16)
TC	dot (.)
TG	A
AC	B
AG	C
CG	D
AAT	E
AAC	F
AAG	G
CAT	2/4 (meter)
CAA	3/4 (meter)
CAC	4/4 (meter)
CAG	(
CCA	(
CCT	x
CCG	2
CCC	3
AAAT	4

Таблица 17

Кодирование арабских цифр, английских букв и прочих символов азотистыми основаниями по алгоритму Хаффмана [Huffman, 1952] для хранения в ДНК текстовой информации

	Группа 1 Код – **G***		Группа 2 Код - **TT**		Группа 3 Код - **TA**	
№№	Символ	Кодон ДНК	Символ	Кодон ДНК	Символ	Кодон ДНК
1	Space	AT	n	AT	.	AT
2	e	CT	s	CT	u	CT
3	shift	TC	h	TC	,	TC
4	t	TG	r	TG	w	TG
5	a	AC	d	AC	m	AC
6	o	AG	l	AG	f	AG
7	i	CG	c	CG	y	CG
8	g	AAT	3	AAT	;	AAT
9	p	AAC	4	AAC	q	AAC
10	b	AAG	S	AAG	z	AAG
11	v	CAT	6	CAT	<	CAT
12	-	CAA	7	CAA	=	CAA
13	(	CAC	a	CAC	%	CAC
14	)	CAG	9	CAG	+	CAG
15	к	CCA	j	CCA	*	CCA
16	0	CCT	x	CCT	?	CCT
17	1	CCG	/	CCG	>	CCG
I8	2	CCC	:	CCC	tab	CCC
19	return	AAAT	$	AAAT	{	AAAT
20	^	AAAA	&	AAAA	}	AAAA
21	_	AAAC	~	AAAC	“	AAAC
22	#	AAAGC	[	AAAGC	\	AAAGC
23	@	AAAGT	]	AAAGT	|	AAAGT
24	!	GTCGCCG				
25	page break	GTCTACCC				

* - Коды групп (G, TT и TA) означают, что для всех используемых кодонов в данных группах должны предшествовать эти нуклеотиды. Таким образом, например, «e» кодируется GCT; «s» – TTCT; «u» TACT. Аналогично следует «прибавлять» эти групповые коды и для кодирования остальных букв, цифр, символов.

Таблица 18

Кодирование графического изображения азотистыми основаниями по алгоритму Хаффмана

Кодон ДНК	Код графического изображения
G	;
TT	.
TA	0
AT	i
CT	2
TC	3
TG	4
AC	5
AG	6
CG	7
AAT	8
AAC	9
AAG	S (s;x1;y1;a)
CAT	R (l;w;x1;y1;a)
CAA	L (x1;y1;x2;y2)
CAC	C (r;x1;y1)
CAG	P (n;x1;y1;x2;y2;x3;y3)
CCT	Tri (s1;an;s2;x1;y1; a)
CCA	E (x;y;l1;l2,a)

где S – квадрат, R – прямоугольник, L – линия, C – круг, P – многоугольник, Tri – треугольник, E – эллипс.

В том же 2009 г. в мартовском номере журнала Nature было опубликовано интервью [Zala, 2009], в котором канадский поэт Christian Bok, вдохновленной работой Wong и соавт. [2003], сообщил, что намерен одну из своих поэм перевести в код ДНК и поместить созданную ДНК также в бактерию *Deinococcus radiodurans* с целью прочтения ее потомками, и планировал потратить на эту задачу пару лет. Однако через несколько номеров в том же журнале Nature была опубликована заметка [Gustafsson, 2009], в которой автор сообщал, что еще в 2005 г. их фирма (DNA2.0 Inc.[9]) закодировала поэму Tomtem (автор V.Rydberg) из 50 слов, которую они сначала перевели в аминокислотную последовательность, используя однобуквенное кодирование (заменив букву O на Q - глутаминовая кислота), и затем в нуклеотидную, после чего этот «ген» был клонирован в *E.coli* в плазмидном векторе. ДНК с этой поэмой была лиофилизирована и отправлена в рождественской открытке по почте. А нуклеотидная последовательность этого гена (точнее поэмы) хранится в GenBank под номером EU600200.1 Synthetic construct Tomten gene, complete cds, 776 bp.

[9] Ныне фирма Atum.

Там же сообщается, что информация об этом кодировании поэмы под авторством Gustafsson C., Minshull J., Ness J.E. and Govindarajan S. с названием «Seasons greetings from DNA2.0» была опубликована в 2005 г. в их фирменном журнале DNA2.0 Newsletter (No 4, P. 1-4). Здесь (забегая вперед) можно еще добавить, что в конце 2018 г. сотрудники фирмы Atum «опубликовали» в несуществующем журнале шутливую статью[10], в которой допустили, что в свойственных приматам Alu-повторах (а именно в последовательности ДНК), закодировано древнее послание, при прочтении которого им удалось углядеть в нем новогоднее поздравление от Санта Клауса.

Следующей серьезной публикацией, посвященной хранению в ДНК небиологической информации, стала еще одна эпохальная статья [Church et al., 2012], даже название которой (Next-Generation Digital Information Storage in DNA) свидетельствовало о прорыве в этой области. В данной работе было применено однобитное кодирование нуклеотидов, согласно которому «0» соответствовали аденин и цитозин, а «1» - гуанин и тимин. Несмотря на некоторое снижение плотности записи (один нуклеотид – один бит информации), открывшаяся возможность использовать при кодировании данных в отдельных случаях аденин вместо цитозина или наоборот (равно как и гуанин с тимином можно было заменять один другим) позволила при конструировании олигонуклеотидов выбирать для них наиболее подходящие последовательности, лишенные гомополимерных участков (длиной более трех нуклеотидов), а также нежелательной вторичной структуры при обеспечении их уникальности. Используя такой подход, авторы смогли закодировать в последовательностях ДНК HTML версию чернового варианта книги «Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves»[11], состоящую из 53426 слов, 11 jpg файлов, а также одну JavaScript программу на что им потребовалось в общей сложности 5,27 млн. бит и 54898 олигонуклеотидов протяженностью 159 звеньев каждый. Причем синтез этих олигонуклеотидов в отличие от всех предыдущих работ велся не колоночным способом, а с помощью микроэррейного ДНК-синтезатора в результате чего образовался пул олигонуклеотидов в пикомолярном количествах. Каждый олигонуклеотид состоял из 4 частей – центральной кодирующей части длиной 96 нуклеотидов (12 байт информации), сопровождаемой 19 нуклеотидами, выполняющими адресную функцию (нумерованных, начиная с 0000000000000000001), и фланкированных двумя участками по 22 нуклеотида, служащими местами отжига праймеров. Таким образом, итоговая плотность записи составила 96 бит информации на 159 нуклеотидов или 0,6 бит/нуклеотид. При этом авторы отметили, что стоимость синтеза олигонуклеотидов в том числе микроэррейным способом слишком высока, чтобы хранение информации в ДНК становилось массовым. Да и с массовым секвенированием ДНК в тот период дело обстояло несколько хуже, чем сейчас. Также в статье говорилось, что в будущих работах надо думать о компрессии данных, проверках четности, снижении числа ошибок и безопасности. Касательно числа ошибок в данной работе, следует заметить, что при секвенировании хранящихся данных с помощью флуоресцентного секвенатора Illumina HiSeq 2000 выявилось 22 расхождения, притом, что 20 из которых пришлись на концевые участки с однократным покрытием.

Несколько месяцев спустя появилась другая знаковая работа [Goldman et al., 2013] в процессе выполнения которой было закодировано в ДНК 5 различных файлов – все 154 сонета У. Шекспира (в ASCII форматт), pdf-файл статьи из журнала Nature, оповестившей мир о том, что молекулы наследственности представляют собой двойную спираль [Watson, Crick, 1953], 26-ти секундный фрагмент знаменитой речи Мартина Лютера Кинга «I have a dream», сделанной им в 1963 г. (MP3 формат), цветная фотография здания Европейского биоинформатического института (JPG формат среднего разрешения) и описание алгоритма Хаффмана (ASCII формат). В общей сложности для кодирования в ДНК всей этой информации потребовалось синтезировать почти 18 миллионов нуклеотидов в виде 153335 олигонуклеотидов длиной 117 звеньев, хотя в дополнительной информации к цитируемой статье говорится, что велся микроэррейный синтез олигонуклеотидов длиной 183 звена, поскольку основную последовательность фланкировали одинаковые участки по 33 нуклеотида, необходимые для секвенирования с помощью секвенатора Illumina HiSed 2000, и на сайте https://www.ebi.ac.uk/goldman-srv/DNA-storage/features.txt именно такие представлены. Помимо этих фланкирующих участков остальная последовательность данных олигонуклеотидов тоже была составной: 1 + 100 (25 x 4) + 14 + 1 + 1. Два

[10] Claes S., Ness N.E., El J., Sridhar V., Medinivy, Navidad F. Xmas message from extraterrestrial intelligence found in transposon encode DNA // Proc. Natl. Acad. NorthPole. 2018. No. 12. P. 25-31.

[11] Изданная в октябре 2012 г. данная книга за авторством George M. Church и Ed Regis, объемом 304 страницы, может быть приобретена, например, в интернет-магазине www.amazon.com.

крайних однонуклеотидных участка как бы окаймляли информационную и служебную последовательности, еще один нуклеотид служил контролем четности. 14 нуклеотидный участок представлял собой индексную последовательность, служащую для правильности сборки, а оставшаяся 100 нуклеотидная часть представляла собой участки по 25 нуклеотидов, три из которых были гомологичны следующим за ним олигонуклеотидам, фактически многократно дублируя одну и туже информацию.

Для перевода намеченных для хранения в ДНК данных потребовалось несколько конвертаций. Так, сначала двоичные числа преобразовывались в троичный код из 0, 1, 2, согласно алгоритму Хаффмана, что затем при переводе таких тритов (троичный аналог битов) в последовательность нуклеотидов позволило исключить гомополимерные участки. После успешного кодирования в последовательностях ДНК этих пяти файлов, они затем были декодированы путем секвенирования. Однако в pdf-файле статьи Уотсона и Крика обнаружилась недодача в виде двух 25 нуклеотидных участков. После анализа нуклеотидной последовательности выяснилось, что потерявшиеся участки, должны были быть среди нескольких одинаковых повторов из 20 нуклеотидов каждый, что видимо, повлияло негативным образом. Изменив эти участки при повторном кодировании на лишенные подобных недостатков, при новом декодировании проблем уже не возникло.

Несколько позже подход, использованный Goldman и соавт. [2013], был видоизменен за счет того, что вместо Huffman кодирования авторы [Limbachiya et al., 2015] применили код Голея, а также архитектура и размер олигонуклеотидов оказались иными. Так, в их работе 99-ти звенные олигонуклеотиды помимо протяженного информационного участка несли две индексных последовательности, имеющих разное предназначение и дополнительно один нуклеотид, отвечающий за контроль четности. С помощью данного подхода для кодирования слова из двух букв «DA» потребовалось 22 нуклеотида – CATGATGCTGAGTCTCGTAGTC.

Поскольку ошибки при хранении информации в ДНК и ее извлечении могут происходить не только на стадиях кодирования (синтеза олигонуклеотидов) и декодирования (секвенирования ДНК), но и во время самого хранения в виде апуринизации, дезаминирования или гидролиза, то было предложено хранить ДНК в неких капсулах из кремния, исключающего попадание влаги и заметно продлевающего время жизни молекул ДНК [Grass et al., 2015]. В этой работе, помимо решения вопроса улучшенного хранения, были также закодированы в ДНК отрывки текстов из старинных произведений – Швейцарской федеративной хартии 1291 года и английского перевода книги древнегреческого математика Архимеда «Методы механических теорем» общим объемом 83 Кб для чего потребовалось микроэррейным способом синтезировать 4991 олигонуклеотид длиною каждый 158 звеньев. Имитация длительного хранения проводилась с использованием высоких температур, и было показано, что после недельного хранения синтезированной ДНК при 70^{o}C (что приблизительно соответствует двум тысячам лет при обычных температурных условиях) извлеченная информация с помощью полногеномного секвенатора Illumina MiSeq не содержала ошибок. Для кодирования данных в ДНК был использован Reed-Solomon код, согласно которому было построено некое «колесо ДНК кодонов», состоящее из 48 секторов, в которых были расположены 47 триплетов ДНК, более подходящих для кодирования цифровой информации и позволивших исключить протяженные гомополимерные участки (более трех одинаковых нуклеотидов подряд). Синтезированные олигонуклеотиды состояли из центральной информационной части в виде 117 звеньев, которую фланкировали служебные последовательности – адапторы общей длиной 41 нуклеотид. Позже этим авторам удалось повысить плотность упаковки молекул ДНК в кремниевых частицах [Chen et al., 2019a]. Процедура состояла в покрытии Fe/C-наночастиц полиэтиленимином, придавшим им положительный заряд, с помощью которого происходило их покрытие отрицательно заряженными молекулами ДНК, завершавшаяся их помещением в раствор тетраэтилортосиликата, обеспечивающего образование капсул. Проведя тщательный анализ возникновения ошибок эта же группа авторов пришла к выводу, что большинство из них все же происходят при кодировании (синтезе) и декодировании (секвенировании), а не при хранении [Heckel et al., 2019].

В том же году была опубликована статья [Yazdi et al., 2015], в которой сообщалось о возможности не только хранения в ДНК неких данных, но и их переписывании. Так, авторы выбрали для хранения информацию из Wikipedia в ASCII кодах о 6 университетах США объемом 17 Кб, которую они закодировали в 27 блоках по 1000 нуклеотидов каждый, фланкируемых адресными последовательностями по 20 звеньев, а остальные 960 нуклеотидов представляли собой 12 субблоков по 80 нуклеотидов, несших в себе информацию о 126 битах.

Все блоки по 1000 нуклеотидов хранились вместе, и для извлечения информации проводилась ПЦР с последующим секвенированием протяженных ампликонов по методу Сэнгера. Причем

разработанный авторами алгоритм перевода информации в последовательности ДНК не исключал наличие гомополимерных последовательностей ввиду того, что такие участки трудны для секвенирования лишь некоторыми методами. Для сборки и редактирования выбранных участков применялись подходы с использованием CRISPR/Cas9 технологии редактирования последовательностей, а также ПЦР с перекрывающимися праймерами [Bryksin, Matsumara, 2010].

В своей следующей работе эта же группа авторов внесла коррективы в устройство тысяченуклеотидных блоков, которые стали состоять из 984 информативных звеньев и 16-ти нуклеотидной адресной последовательности [Yazdi et al., 2017]. С помощью таких блоков в ДНК были закодированы два рисунка – черно-белая реклама фильма 1941 г. «Citizen Kane» и смайлик «Улыбающееся лицо», общим объемом 10894 байт, которые были «ужаты» до 3633 байт. Для их хранения были сконструированы 17 блоков, для извлечения информации из которых использовалось нанопоровое секвенирование с помощью секвенатора MinIon фирмы Oxford Nanopore Technologies. Сообщается о неверном прочтении гомополимерных участков – последовательности AAAAAAA были принята за AAAAA, а AAAAA – за AAAA. Однако примененный подход *homopolymer check codes* смог устранить данные ошибки.

При выполнении еще одной масштабной работы было закодировано в ДНК более 2 Мебибайт (2146816 байт) информации, в которую вошли шесть различных файлов, включая операционную систему, pdf-файл, графическое изображение, видеофайл и др. [Erlich, Zielinski, 2017]. Причем видеофайл представлял собой снятый братьями Люмьер знаменитый немой документальный короткометражный фильм 1896 г. под названием «Прибытие поезда на вокзал Ла-Сьота». Для кодирования всей этой информации в последовательности ДНК было применено специальное кодирование DNA Fountain, обеспечивающее подбор олигонуклеотидов с GC-составом от 45 до 55% и отсутствием протяженных гомополимерных участков с допустимыми тремя одинаковыми нуклеотидами подряд. Длина олигонуклеотидов, синтезированных с помощью микроэррейного метода, составляла 200 звеньев, из которых 48 приходились на фланкирующие участки, необходимые для отжига адапторов при секвенировании. Оставшиеся 152 нуклеотида кодировали намеченную для хранения информацию (128 бит), а также некие служебные последовательности, включая контроль четности. Декодирование с помощью секвенатора Illumina MiSeq показало отсутствие ошибок. На данную технологию хранения небиологической информации в ДНК подана заявка на патент США под номером 2019/0020353 [Erlich, 2019].

Заметно увеличенный объем информации (22 Мб) в виде некоего видеофайла был сохранен в ДНК в другой работе [Blawat et al., 2016]. Причем сообщалось, что это первая часть проекта по сохранению в ДНК 1 Гб информации. Для хранения этих 22 Мб было синтезировано 900 тысяч олигонуклеотидов микроэррейным методом, из которых было сформировано 4 библиотеки по 225 тысяч каждая. Синтезированные олигонуклеотиды имели длину по 230 звеньев. Из них за кодирование информации «отвечали» 190 нуклеотидов, а фланкирующие участки (по 20 звеньев) представляли собой адапторы для секвенирования с помощью Illumia HiSeq 2500. Использовалось двухбитное кодирование по которому двоичное число 00 соответствовало A, 01 – C, 10 – G и 11 – T. Причем в данной работе значительный упор был сделан на исключении ошибок при синтезе и секвенировании.

На порядок больший объем данных (более 200 Мб или точнее 200268740 байт), содержащий информацию о 35 различных файлах (размерами от 29 Кб до 44 Мб), был сохранен в молекулах ДНК в работе целой группы ученых преимущественно из Вашингтонского университета и корпорации Майкрософт [Organick et al., 2018]. Закодированные файлы были представлены большим разнообразием их типов – txt, pdf, MP3, MP4, jpg, а также ряд архивированных. Среди наиболее крупных и примечательных следует отметить «Декларацию прав человека» на более чем 100 языках, видеоклип музыкальной рок-группы с песней «This Too Shall Pass», база данных семян, хранящихся в Глобальном хранилище семян на Шпицбергене. Синтез 13448372 олигонуклеотидов осуществлялся фирмой Twist Bioscience на синтезаторе микроэррейного типа. Их длина равнялась 150 звеньям, за исключением тех, что кодировали файл под номером 33, для которого такие олигонуклеотиды были чуть длиннее – 154 звена. Так как по краям данных информационных олигонуклеотидов располагались места для отжига 20-звенных праймеров, то общие размеры олигонуклеотидов были 190 и 194 звена. Подбор праймеров велся многостадийно. Сначала были сгенерированы случайным образом с учетом ряда ограничений 19480 последовательностей длиной 20 нуклеотидов, после чего шла отбраковка неподходящих, оставившая 9869, а итоговый комплект оказался равным 3240 праймерам, лишенным протяженных гомополимерных участков (не более трех одинаковых нуклеотидов подряд для A и T и не более двух таких для C и G. Также принималось во внимание отсутствие значительной

интергомологии и GC-состав находится в пределах от 45 до 55%. Самым главным было отсутствие гомологии с информационными последовательностями для соответствия которых двоичным числам использовался тот же подход, что и в их предыдущей работе [Blawat et al., 2016] согласно которому 00 – A, 01 – C, 10 – G и 11 – T. Для извлечения хранимой информации проводилось секвенирование с помощью флуоресцентного секвенатора Illumina NextSeq и нанопорового секвенатора ONT MinIon.

В своей предыдущей статье [Bornholt et al., 2017] данный коллектив авторов в несколько уменьшенном составе сообщил, что для кодирования цифровых данных азотистыми основаниями, бинарный код сначала переводился в третичный и затем аналогично тому как это осуществлялось в работе Goldman и соавт. [2013] преобразовывался в нуклеотиды. Размер синтезированных олигонуклеотидов был несколько короче – от 120 до 150 звеньев. При этом весьма подробно рассмотрена его архитектура. Так, центральная часть представляла собой информационную последовательность, которую с одной стороны окаймлял единичный нуклеотид, служащий указателем направления, а с другой стороны (по направлению к 3') такому же единичному нуклеотиду с той же функцией предшествовала адресная последовательность. Всю эту конструкцию фланкировали места для праймеров.

Если в своей работе 2018 г. Organick и соавт. [2018] для извлечения данных преимущественно использовали флуоресцентный секвенатор и лишь в небольшом числе случаев нанопоровый, то в работе 2019 г. [Lopez et al., 2019] был задействован исключительно нанопоровый секвенатор ONT MinIon, что с учетом длинных прочтений позволило вести сборку олигонуклеотидов в протяженные последовательности. Для этого синтезированные микроэррейным способом олигонуклеотиды длиной 150 звеньев (из которых информационная часть составляла 110 нуклеотидов, фланкированных участками для отжига праймеров по 20) с помощью двух способов [Gibson et al., 2009; Bryksin, Matsumura, 2010] объединялись в блоки до 24 штук длиной 4590 пар нуклеотидов. В рамках этой работы в синтетической ДНК в виде 111499 олигонуклеотидов были закодированы три графических файла общим объемом 1,67 Мб. Они представляли собой фотографию космического шаттла, снимок Земли из космоса и известный рисунок Леонардо да Винчи Витрувианского человека, причем для кодирования этого рисунка применялись олигонуклеотиды с гомополимерными участками, а для других файлов таковые не использовались. Еще одним переведенным в последовательность ДНК файлом была кулинарная книга 1908 г. «365 Foreign Dishes». Извлечение информации путем нанопорового секвенирования показало немало ошибок, большинство которых сводилось к плохой дифференциации аденинов и гуанинов (среди пуринов) и такой же плохой различимости цитозинов от тиминов (среди пиримидинов).

Тем не менее, данный подход с извлечением хранимых данных с помощью нанопорового секвенатора ONT MinIon позволил создать первую автоматизированную станцию размером по площади со стандартный лабораторный стол для обеспечения хранения в ДНК небиологических данных, о чем было сообщено весной 2019 г. [Takahashi et al., 2019]. В ее состав вошли три основных модуля: модуль кодирования/декодирования с соответствующим программным обеспечением; модуль микроэррейного синтеза олигонуклеотидов; модуль обращения с молекулами ДНК, включая их секвенирование. Ориентировочная стоимость данного устройства по оценкам авторов составила 10 тысяч долларов США с возможным снижением до 3–4 тысяч. В качестве демонстрации возможностей этого оборудования с его помощью было кодировано и декодировано 5-ти байтовое слово «HELLO» (01001000 01000101 01001100 01001100 01001111). При этом использованный принцип кодирования информации был таким – A = 0, C = 1, G = 2, T = 3 и поэтому требовалось соответствующее преобразование данных.

Относительно недавно было предложено для хранения информации в ДНК переводить ее сначала в четвертичный код в виде цифр - 0, 1, 2, 3, которым было решено, что будут соответствовать нуклеотиды A, T, C, G [Zhong et al., 2018]. Данная кодировка ДНК получила название BitDNA и с ее помощью древнее китайское произведение "The Analects of Confucius" объемом в 21505 слов было конвертировано в 176724 нуклеотида, после чего разбито на короткие блоки, состоящие из информационных 44 нуклеотидов, фланкированных 8-ми нуклеотидными индексами и 20-ти нуклеотидными участками Flank-L и Flank-R, в результате каждый такой олигонуклеотид имел общую длину 92 звена. Ранее аналогичное кодирование азотистых оснований четырьмя цифрами описано в другой работе [Schouhamer Immink et al., 2017], однако выбор нуклеотидов был несколько иной: A – 0; C – 1; G – 2; T – 3. В этой работе также определенное внимание было уделено исключению гомополимерных участков.

В одной из работ было предложено использовать для хранения информации в ДНК (фрагмента аннотации статьи 2001 г., связанной с завершением глобального проекта секвенирования генома человека, размером 377 бит) синтетические

олигонуклеотиды, содержащие 8-ми нуклеотидный участок, представленный вырожденными последовательности как NNNNNNNN [Hwang, Bang, 2016]. Причем эти 8 нуклеотидов делились на два блока из 4 нуклеотидов каждый – «адресный» и «информационный». Фланкирующие последовательности подбирались с таким расчетом, чтобы в них не было более 4 подряд одинаковых нуклеотидов и общий GC-состав находился в пределах от 40 до 60%. Извлечение данных осуществлялось с помощью пиросеквенатора 454 Junior и флуоресцентных секвенаторов Illumina HiSeq и MiSeq.

С целью снижения стоимости синтеза олигонуклеотидов для хранения данных в ДНК также было предложено синтезировать их вырожденные последовательности [Choi et al., 2019]. Так, помимо четырех отдельных азотистых оснований (A, C, G, T), существует еще 11 комбинаций их объединений – K (G, T), M (A, C), R (A, G), S (C, G), W (A, T), Y (C, T), B (C, G, T), D (A, G, T), H (A, C, T), V (A, C, G), N (A, C, G, T). Наработка таких вырожденных олигонуклеотидов достигалась смешиванием на этапе синтеза соответствующих фосфорамидитов. Применяемые для хранения олигонуклеотиды состояли из центральной информационной части длиной 42 звена, трех нуклеотидов адресного участка и двух фланкирующих последовательностей по 20 нуклеотидов каждая. В рамках данной работы были закодированы в ДНК один тестовой и один графический файлы объемом 854 и 135 Кб.

Предложенный оригинальный способ цифровизации ДНК [Chen et al., 2019] заключался в детекции (дифференциации) искусственно созданных шпилечных структур двух размеров (8 и 16 пар нуклеотидов, которым соответственно были присвоены двоичные числа «0» и «1») при прохождении молекулы ДНК фага M13mp18, представляющей в норме одноцепочечную последовательность длиной 7228 нуклеотидов, но в которой через промежутки в 114 нуклеотидов располагались эти самые шпильки. Для этого были синтезированы соответствующие олигонуклеотиды длинами 61 и 77 звеньев, 38 из которых были комплементарны участкам ДНК фаг M13 и формировали с ней двойную спираль, а оставшиеся 23 и 39 нуклеотидов образовывали шпилечные структуры со стебельками из 8 и 16 пар нуклеотидов. При прохождении такой одно-/двухцепочечной молекулы ДНК через нанопору в кварцевом стекле диаметром 5 нм менялся ток в зависимости от того какая шпилька это отверстие перекрывала. Сообразно уровням сигналов происходило двоичное кодирование в виде нулей и единиц. Авторы сообщают, что их способ цифровизации молекул ДНК может при создании соответствующих библиотек из подходящих молекул ДНК обеспечить до 2^{112} комбинаций.

Предложен также способ хранения небиологической информации в ДНК, извлечение которой не связано с секвенированием [Nguyen et al., 2018]. В этой работе хранение информации в виде синтезированных олигонуклеотидов происходило путем их фиксации на чипе, а детекция производилась регистрацией флуоресценции с помощью микроскопа. Особенностью этого подхода было также то, что молекулярная гибридизация фиксированных олигонуклеотидов (Capture strand) с находящимися в растворе мечеными олигонуклеотидами (Partner strand) происходила в условиях смещения цепей еще одного типа олигонуклеотидов (Displacement strand). Наличие флуоресцентного сигнала давало двоичное число «1», а отсутствие такового – «0». При выполнении цитируемой работы в ДНК было закодировано краткое название Института, в котором выполнялась данное исследование – KRIBB (Korean Research Institute of Bioscience and Biotechnology), потребовавшее для каждой буквы 8 бит – K (01001011), R (01010010), I (01001001), B (01000010).

Одним из наиболее простых и удобных кодировок азотистых оснований двоичными числами является представленная в таблице 19. При этом в разных статьях встречаются разные варианты такого кодирования, поскольку в пределах двухбитного двоичного кодирования в целом имеется 24 (4!) варианта, и все они, следуя некой логике, имеют в той или иной степени «право на существование». Такое двоичное кодирование обеспечивает плотность записи 2 бита на нуклеотид. Сейчас в большинстве публикаций аденин (A) кодируется как 00, тогда как азотистые основания G и T имеют по три варианта кодировок из четырех возможных для двоичных двухразрядных чисел, а цитозин (C) кодируется всеми вариантами – 00, 01, 10 и 11. Это свидетельствует о том факте, что в настоящее время единого подхода к кодировке нуклеотидов двоичными числами не существует, что вносит определенные трудности и разночтения при ДНК-цифровизации данных, хотя для ДНК-криптографии наоборот может быть преимуществом.

Здесь можно вспомнить Р. Фейнмана, в 1959 году допустившего, что в будущем для кодирования одного бита информации будет достаточно 100 атомов [Feynman, 1960]. При таком подходе, когда 2 бита кодируются одним нуклеотидом, состоящим при его нахождении в составе молекулы ДНК из 54, 56, 57 или 58 атомов в зависимости от типа азотистого основания, можно считать, что один бит может быть записан приблизительно 27–29 атомами.

Таблица 19

Некоторые способы цифровых кодировок отдельных нуклеотидов двоичными числами

Нуклеотиды	A (аденин)				C (цитозин)				G (гуанин)				T (тимин)			
Числа / Варианты	00	01	10	11	00	01	10	11	00	01	10	11	00	01	10	11
1	00					01					10					11
2	00						10			01						11
3	00						10					11		01		
4	00							11			10			01		
5	00					01						11			10	
6			10		00							11		01		
7		01			00							11			10	

Кроме кодирования нуклеотидов ДНК двоичными числами предложен также их перевод в азбуку Морзе [Murugan, Thilagavathy, 2017], представленный в таблице 20.

Таблица 20

Способ кодировок отдельных нуклеотидов с помощью азбуки Морзе

нуклеотид	код Морзе
A (аденин)	**.-** *(точка тире)*
C (цитозин)	**..** *(точка точка)*
G (гуанин)	**-.** *(тире точка)*
T (тимин)	**--** *(тире тире)*

Собственно вариантов кодирования азотистых оснований азбукой Морзе может быть также 24 (4!) комбинаций точек и тире для четверки нуклеотидов.

Вопросам хранения небиологической информации в ДНК посвящено большое количество обзорных работ [De Silva, Ganegoda, 2016; Zhirnov et al., 2016; Akram et al., 2018; Bhat et al., 2018; Panda et al., 2018; Ceze et al., 2019; Ping et al., 2019 и др.], но ни в одной из них не сделано столь подробного анализа принципов и способов кодирования цифровых данных азотистыми основаниями, как здесь нами.

Заключение

Л. Адлеман, чьи слова мы взяли в качестве эпиграфа к данной статье, безусловно, прав, утверждая, что ДНК это цифровая молекула. Всевозможные перестановки нуклеотидов обеспечивают абсолютную уникальность относительно протяженных фрагментов ДНК, что важно и для небиологического применения молекул ДНК, в частности в виде ДНК-криптографии и ДНК-стеганографии, а также долговременного хранения любой информации в ДНК. Причем в их основе лежит биологический принцип комплементарности азотистых оснований, согласно которому аденин спаривается с тимином, а гуанин с цитозином, что позволяет запрограммировать взаимодействие разных одноцепочечных фрагментов ДНК для формирования ими вторичной структуры в виде двойной спирали, поскольку ДНК всегда стремится стать двухцепочечной молекулой за счет возникновения водородных связей между комплементарными нуклеотидами, обеспечивающими при определенных условиях (в первую очередь температурных) поддержание (природной) целостности этой молекулы. Дополнительными основаниями использовать молекулы ДНК для небиологических целей служат возможность амплификации *in vitro* определенных участков ДНК, обычно ограниченных праймерами в виде специально синтезированных химическим путем олигонуклеотидов с заданной последовательностью, а также секвенирование ДНК, в том числе высокопроизводительное новых поколений.

Несмотря на то, что начало небиологическому применению молекул ДНК было положено работами по созданию запрограммированных наноструктур из олигонуклеотидов, все же основной толчок это направление получило после публикации эпохальной статьи Л. Адлемана по ДНК-компьютингу [Adleman, 1994]. При этом надо признать, что до сих пор никакого практического воплощения молекулярных вычислений с помощью ДНК не наблюдается. Однако через несколько лет возникло новое направление в виде ДНК-криптографии и ДНК-стеганографии, которые также пока находятся в стадии исследований. При этом необходимо отметить, что почти одновременно стали разрабатываться способы долговременного хранения в ДНК небиологической информации с чем тесно соприкасаются вопросы ДНК-цифровизации, и кодировки короткими последовательностями нуклеотидов букв английского алфавита и прочих символов чему в данной статье уделено довольно большое внимание. Причем, возможно именно это направление раньше других небиологических применений ДНК дойдет до использования в широких масштабах и при этом наиболее перспективным выглядит универсальное кодирование последовательностей битов и/или байтов посредством цепочек азотистых оснований.

Интерес к подобному использованию молекул ДНК отчасти вызван грантом РФФИ № 20-07-00222.

Литература

1. Баймиев Ан.Х., Кулуев Б.Р., Вершинина З.Р. Князев А.В., Чемерис Д.А., Рожнова Н.А., Геращенков Г.А., Михайлова Е.В., Баймиев Ал.Х., Чемерис А.В. CRISPR/Cas редактирование геномов (растений) и общество // *Биомика*. 2017. Т.9. С.183-202.
2. Баймиев Ан.Х., Чемерис Д.А., Кирьянова О.Ю., Матниязов Р.Т., Валеев А.Ш., Баймиев Ал.Х., Губайдуллин И.М., Чемерис А.В. Биоинформатические ресурсы для in silico поиска CRISPR локусов в геномах прокариот // *Биомика*. 2017а. Т.9. С.229-244.
3. Вершинина З.Р., Кулуев Б.Р., Геращенков Г.А., Князев А.В., Чемерис Д.А., Гумерова Г.Р., Баймиев Ал.Х., Чемерис А.В. Эволюция методов редактирования геномов // *Биомика*. 2017. Т.9. С.245-270.
4. Геращенков Г.А, Рожнова Н.А, Кулуев Б.Р., Кирьянова О.Ю., Гумерова Г.Р., Князев А.В., Вершинина З.Р., Михайлова Е.В., Чемерис Д.А., Матниязов Р.Т., Баймиев Ан.Х., Губайдуллин И.М., Баймиев Ал.Х., Чемерис А.В. Дизайн РНК-гидов для CRISPR/CAS редактирования геномов растений // *Молекулярная биология*. 2020. Т.54(1). С. 29-50. DOI: 10.1134/S0026898420010061
5. Кулуев Б.Р., Баймиев Ан.Х., Чемерис Д.А., Матниязов Р.Т., Геращенков Г.А., Никоноров Ю.М., Баймиев Ал.Х., Чемерис А.В. Применение CRISPR-локусов не для редактирования геномов // *Биомика.* 2017. Т.9. С.271-283.
6. Кулуев Б.Р., Геращенков Г.А., Рожнова Н.А., Баймиев Ан.Х., Вершинина З.Р., Князев А.В., Матниязов Р.Т., Гумерова Г.Р., Михайлова Е.В., Никоноров Ю.М., Чемерис Д.А., Баймиев Ал.Х., Чемерис А.В. CRISPR/Cas редактирование геномов растений // *Биомика*. 2017а. Т.9. С.155-182.
7. Кулуев Б.Р., Кирьянова О.Ю., Геращенков Г.А., Рожнова Н.А., Гумерова Г.Р., Вершинина З.Р., Матниязов Р.Т., Ахметзянова Л.У., Князев А.В., Михайлова Е.В., Гарафутдинов Р.Р., Баймиев Ан.Х., Губайдуллин И.М., Баймиев Ал.Х., Чемерис А.В. Некоторые новшества в CRISPR/Cas геномном редактировании и в смежных областях // *Биомика*. 2019. Т.11(3). С. 315-343. DOI: 10.31301/2221-6197.bmcs.2019-27
8. Кулуев Б.Р., Гумерова Г.Р., Михайлова Е.В., Геращенков Г.А., Рожнова Н.А., Вершинина З.Р., Князев А.В., Матниязов Р.Т., Баймиев Ан.Х., Баймиев Ал.Х., Чемерис А.В. Доставка CRISPR/CAS-компонентов в клетки высших растений для редактирования их геномов. // *Физиология растений*. 2019. Т.66(5). С.339-353. DOI: 10.1134/S0015330319050117
9. Нейман М.С. Некоторые принципиальные вопросы микроминиатюризации // *Радиотехника*. 1964. Т.19(1), с. 3-12.
10. Нейман М.С. О связях между надежностью, быстродействием и степенью микроминиатюризации на молекулярно-атомном уровне // *Радиотехника*. 1965. Т.20(1). С. 1-9.
11. Нейман М.С. О молекулярных системах памяти и о направленных мутациях // *Радиотехника*. 1965а. Т.20(6). С. 1-8.
12. Чемерис А.В. CRISPR/Cas системы (специальный тематический выпуск журнала) // *Биомика*. 2017. Т.9(3). С. 148-154.
13. Чемерис А.В., Рожнова Н.А., Геращенков Г.А. Некоторые недавние улучшения методов геномного редактирования // *Известия Уфимского научного центра РАН*. 2018. №3(5). С. 86–93.
14. Чемерис Д.А., Кирьянова О.Ю., Геращенков Г.А., Кулуев Б.Р., Рожнова Н.А., Матниязов Р.Т., Баймиев Ан.Х., Баймиев Ал.Х., Губайдуллин И.М., Чемерис А.В. Биоинформатические ресурсы для CRISPR/Cas редактирования геномов // *Биомика*. 2017. Т.9. С.203-228.
15. Adleman LM. Molecular computation of solutions to combinatorial problems. // *Science.* 1994. V.266(5187). P.1021-1024. DOI: 10.1126/science.7973651
16. Adleman LM. Response. // *Science*. 1995 Apr 28;268(5210):483-4. DOI: 10.1126/science.268.5210.483
17. Agrawal A. Bhopale A., Sharma J., Shizan Ali M., Gautam D. Implementation of DNA algorithm for secure voice communication. // *International Journal of Scientific & Engineering Research*. 2012. V.3(6). P.1-5.
18. Ailenberg M, Rotstein O. An improved Huffman coding method for archiving text, images, and music characters in DNA. // Biotechniques. 2009. V.47(3). P.747-754. doi: 10.2144/000113218.
19. Akram F., Haq I.U., Ali H., Laghari A.T. Trends to store digital data in DNA: an overview. // *Mol. Biol. Rep*. 2018. V.45(5). P.1479-1490. doi: 10.1007/s11033-018-4280-y
20. Arita M. Writing information into DNA // In: Jonoska N., Păun G., Rozenberg G. (eds) Aspects of Molecular Computing. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2004. V. 2950. P. 23-35. DOI: 10.1007/978-3-540-24635-0_2
21. Arita M, Ohashi Y. Secret signatures inside genomic DNA. // *Biotechnol Prog*. 2004. V.20(5). P.1605-1607. DOI: 10.1021/bp049917i
22. Avery O.T., MacLeod C.M., McCarty M. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of

transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. // *J. Exp. Med.* 1944. V.79(2). P.137–158. DOI: 10.1084/jem.79.2.137
23. Bancroft C., Bowler T., Bloom B., Clelland C.T. Long-term storage of information in DNA // *Science*. 2001. V.293(5536). P.1763-1765. DOI: 10.1126/science.293.5536.1763c
24. Bancroft F.C., Cleland C. DNS - based steganography Patent. US 6,312,911 B1. Data of patent Nov. 6, 2001
25. Baum E.B. Building an associative memory vastly larger than the brain. // *Science*. 1995. V.268(5210). P.583-585. DOI: 10.1126/science.7725109
26. Bhat W.A. Bridging data-capacity gap in big data storage. // *Future Generation Computer Systems*. 2018. V. 87. P. 538-548. doi: 10.1016/j.future.2017.12.066
27. Blawat M., Gaedke K., Huetter I., Chen X-M., Turczyk B., Inverso S., Pruitt B., Church G. Forward Error Correction for DNA Data Storage. // *Procedia Computer Science*. 2016. V.80. P. 1011-1022. doi: 10.1016/j.procs.2016.05.398
28. Bornholt J., Lopez R., Carmean D., Ceze L., Seelig G., Strauss K. A DNA-Based Archival Storage System // *ASPLOS '16 Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*. 2017. P. 637-649. doi: 10.1145/2872362.2872397
29. Bornholt J., Lopez R., Carmean D.M., Ceze L., Seelig G., Strauss K. Toward a DNA-Based Archival Storage System. // *IEEE Micro*. 2017a. V.37(3). P.98-104. DOI: 10.1109/MM.2017.70
30. Bryksin AV, Matsumura I. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. // Biotechniques. 2010 Jun;48(6):463-5. doi: 10.2144/000113418.
31. Bunow B. On the potential of molecular computing. // *Science*. 1995 Apr 28;268(5210):482-3. DOI: 10.1126/science.7725087
32. Byrne J., Dahm R. Friedrich Miescher and the 150th anniversary of the discovery of DNA. // *Biomics*. 2019. V.11(3). P. DOI: 10.31301/2221-6197.bmcs.2019-
33. Ceze L, Nivala J, Strauss K. Molecular digital data storage using DNA. // *Nat. Rev. Genet*. 2019. V.20(8). P.456-466. doi: 10.1038/s41576-019-0125-3
34. Chen K., Kong J., Zhu J., Ermann N., Predki P., Keyser U.F. Digital data storage using DNA nanostructures and solid-state nanopores // *Nano Lett*. 2019a. V.19(2). P.1210-1215. doi: 10.1021/acs.nanolett.8b04715
35. Chen W.D., Kohll A.X., Nguyen B.H., Koch J., Heckel R., Stark W.J., Ceze L., Strauss K., Grass R.N. Combining Data Longevity with High Storage Capacity – Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles. // *Adv. Func. Mater*. 2019. V. 29(28). 1901672. doi: 10.1002/adfm.201901672
36. Choi Y., Ryu T., Lee A.C., Choi H., Lee H., Park J., Song S-H., Kim S., Kim H., Park W., Kwon S. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. // *Sci. Rep*. 2019. V.9(1):6582. doi: 10.1038/s41598-019-43105-w
37. Church G.M., Gao Y., Kosuri S. Next-generation digital information storage in DNA // *Science*. 2012. V.337(6102). P.1628. DOI: 10.1126/science.1226355
38. Clelland C.T., Risca V., Bancroft C. Hiding messages in DNA microdots. // *Nature*. 1999. V.399(6736). P.533-534. DOI: 10.1038/21092
39. Davis J. Microvenus. // *Art Journal*. 1996. V. 55(1). P. 70-74. DOI: 10.2307/777811
40. De Silva PY, Ganegoda GU. New Trends of Digital Data Storage in DNA. // *Biomed Res Int*. 2016.:8072463. DOI: 10.1155/2016/8072463
41. Erlich Y. Efficient encoding of data for storage in polymers such as DNA. US Patent Application No 2019/0020353. January 17, 2019
42. Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. // *Science*. 2017. V.355(6328). P.950-954. doi: 10.1126/science.aaj2038.
43. Feynman R.P. There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics // *Engineering and Science (California Institute of Technology)*. 1960. V23. P.22-36.
44. Frutos AG, Liu Q, Thiel AJ, Sanner AM, Condon AE, Smith LM, Corn RM. Demonstration of a word design strategy for DNA computing on surfaces. // *Nucleic Acids Res*. 1997. V.25(23). P.4748-4757. DOI: 10.1093/nar/25.23.4748
45. Frutos A.G., Smith, L.M., Corn R.M. Enzymatic ligation reactions of DNA "words" on surfaces for DNA computing // *Journal of the American Chemical Society*. 1998. V. 120. P. 10277-10282.
46. Fu P. Biomolecular computing: is it ready to take off? *Biotechnol J.* 2007. V.2(1). P.91-101. DOI: 10.1002/biot.200600134
47. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA 3rd, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. // *Nat Methods*. 2009. V.6(5). P.343-345. doi: 10.1038/nmeth.1318
48. 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. V.494(7435). P.77-80. doi: 10.1038/nature11875.
49. 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. // *Angew Chem Int Ed Engl*. 2015. V.54(8). P.:2552-2555. doi: 10.1002/anie.201411378.
50. Gustafsson C. For anyone who ever said there's no such thing as poetic gene. // *Nature*. 2009. V.458. P. 703.
51. Heckel R., Mikutis G., Grass R.N. A Characterization of the DNA Data Storage Channel. // *Sci Rep*. 2019. V.9(1):9663. doi: 10.1038/s41598-019-45832-6.
52. Hodgson C.P. A DNA text code // *Biotechniques*. 1990. V. 9(3). P. 312.
53. Huffman D.A. A Method for the Construction of Minimum-Redundancy Codes // *Proceedings of the IRE*. 1952. V. 40(9). P. 1098 – 1101. DOI: 10.1109/JRPROC.1952.273898
54. Hwang B, Bang D. Toward a new paradigm of DNA writing using a massively parallel sequencing platform and degenerate oligonucleotide. // *Sci. Rep*. 2016. V.6:37176. doi: 10.1038/srep37176
55. Interview. Machines smarter than men? // *U.S. News & World Report*. 1964. Feb., 24. P.84-86.
56. Jiao S., Goutte R. Code for encryption hiding data into genomic DNA of living organisms. *9th International Conference on Signal Processing*. 2008. DOI: 10.1109/ICOSP.2008.4697576
57. Jimenez-Sanchez A. A proposal for a DNA-based computer code // *International Invention Journal of Biochemistry and Bioinformatics*. 2013. V.1(1) P. 1-4.
58. Kashiwamura S., Yamamoto M., Kameda A., Shiba T., Ohuchi A. Potential for enlarging DNA memory: the validity of experimental operations of scaled-up nested primer molecular memory. // *Biosystems*. 2005. V.80(1). P.99-112. DOI: 10.1016/j.biosystems.2004.10.007
59. Landegren D.A. DNA probes and automation. // *Curr. Opin. Biotechnol*. 1992. V.3(1). P.12-17. doi: 10.1016/0958-1669(92)90119-4
60. Landweber LF, Kari L. The evolution of cellular computing: nature's solution to a computational problem. // Biosystems. 1999. V.52(1-3). P.3-13. DOI: 10.1016/s0303-2647(99)00027-1
61. Li D., Li X., Huang H., Li X. The surface-based approach for DNA computation is unreliable for SAT // Biosystems. 2005. V. 82. P. 20 - 25. doi:10.1016/j.biosystems.2005.05.007
62. Li D., Li X., Huang H., Li X. Scalability of the surface-based DNA algorithm for 3-SAT // *Biosystems*. 2006. V. 85. P. 95-98. doi: 10.1016/j.biosystems.2005.12.002
63. Limbachiya D. et al. On optimal family of codes for archival DNA storage. *2015 Seventh International Workshop on Signal Design and its Applications in Communications (IWSDA)*. 2015. DOI: 10.1109/IWSDA.2015.7458386
64. Linial M, Linial N. On the potential of molecular computing. // *Science*. 1995. V.268(5210). P.481. DOI: 10.1126/science.7725085
65. Lipton RJ. DNA solution of hard computational problems. // *Science*. 1995. V.268(5210). P.542-545. DOI: 10.1126/science.7725098
66. Liu Q, Frutos AG, Thiel AJ, Corn RM, Smith LM. DNA computing on surfaces: encoding information at the single base level. // *J Comput Biol*. 1998. V.5(2). P.269-278. DOI: 10.1089/cmb.1998.5.269
67. Liu Q, Wang L, Frutos AG, Condon AE, Corn RM, Smith LM. DNA computing on surfaces. // *Nature*. 2000. V.403(6766). P.175-179. DOI: 10.1038/35003155
68. Lo YM, Yiu KF, Wong SL. On the potential of molecular computing. // *Science*. 1995. V.268(5210). P.481-482. DOI: 10.1126/science.7725086
69. Lopez R., Chen Y-J., Ang S.D, Yekhanin S., Makarychev K., Racz M.Z., Seelig G., Strauss K., Ceze L. DNA assembly for nanopore data storage readout. // *Nat. Commun*. 2019. V.10(1):2933. doi: 10.1038/s41467-019-10978-4
70. Melkikh AV. DNA computing, computation complexity and problem of biological evolution rate. // *Acta Biotheor*. 2008. V.56(4). P.285-295. doi: 10.1007/s10441-008-9055-8
71. Miescher F. Ueber die chemische Zusammensetzung der Eiterzellen // *Medicinisch-chemische Untersuchungen*. 1871. V.4. P. 441–460.
72. Mills A.P., Yurke B. Data trnsmission using DNA oligomers. Patent US 6, 537,747 B1. Data of patent Mar. 25, 2003
73. Murugan A., Thilagavathy R. Securing cloud data using DNA and Morse code: A triple encryption scheme // *International Journal of Control Theory and Applications*. 2017. V.10(23). P. 31-38.
74. Nair ACC, Mahalakshmi T. Visualization of genomic data using inter-nucleotide distance signals. // *Conference Proceedings.* 2005. (https://www.researchgate.net/publication/228643588_Visualization_of_genomic_data_using_inter-nucleotide_distance_signals)
75. Nguyen H.H., Park J., Hwang S., Kwon O.S, Lee C., Shin Y., Ha T.H, Kim M. On-Chip fluorescence switching system for constructing a rewritable random-access data storage device // *Sci Rep*. 2018. V.8. P. 337. doi: 10.1038/s41598-017-16535-7
76. Nishikawa A., Yamamura M., Hagiya M. DNA computation simulator based on abstract bases // *Soft*

Computing. 2001. P. 25-38. doi: 10.1007/s005000000062
77. Organick L., Ang S.D., Chen Y.J., Lopez R., Yekhanin S., Makarychev K., Racz M.Z., Kamath G., Gopalan P., Nguyen B., Takahashi C.N., Newman S., Parker H.Y., Rashtchian C., Stewart K., Gupta G., Carlson R., Mulligan J., Carmean D., Seelig G., Ceze L., Strauss K. Random access in large-scale DNA data storage. // *Nat. Biotechnol*. 2018. V. 36(3). P.242-248. doi: 10.1038/nbt.4079. Erratum: *Nat. Biotechnol*. 2018. V.36(7). P.660. doi: 10.1038/nbt0718-660c
78. Panda D., Molla K.A., Baig M.J., Swain A., Behera D., Dash M. DNA as a digital information storage device: hope or hype? // *3 Biotech*. 2018. V.8(5):239. doi: 10.1007/s13205-018-1246-7.
79. Ping Z., Ma D., Huang X., Chen S., Liu L., Guo F., Zhu S.J., Shen Y. Carbon-based archiving: current progress and future prospects of DNA-based data storage. // *Gigascience*. 2019. V.8(6). pii: giz075. doi: 10.1093/gigascience/giz075
80. Portney N.G., Wu Y., Quezada L.K., Lonardi S., Ozkan M. Length-based encoding of binary data in DNA // *Langmuir*. 2008. V.24(5). P. 1613-1616. doi: 10.1021/la703235y
81. Rashid O.F., Othman Z.A., Zainitdin S. A novel DNA sequence approach for network intrusion detection system based on cryptography encoding method // *International Journal on Advanced Science, Engineering and Information Technology*. V. 7(1). P. 183-189. doi: 10.18517/ijaseit.7.1.1569
82. Schmidt KA, Henkel CV, Rozenberg G, Spaink HP. DNA computing using single-molecule hybridization detection. // *Nucleic Acids Res*. 2004. V.32(17). P.4962-2968. DOI: 10.1093/nar/gkh817
83. Schouhamer Immink K.A., Cai K. Design of Capacity-Approaching Constrained Codes for DNA-Based Storage Systems. // *IEEE Communications Letters*. 2018. V. 22(2). P. 224 – 227. DOI: 10.1109/LCOMM.2017.2775608
84. Seeman N.C. Nucleic acid junctions and lattices. // *J. Theor. Biol*. 1982. V.99(2). P.237-247. DOI: 10.1016/0022-5193(82)90002-9
85. Shin J., Pierce N. Rewritable memory by controllable nanopatterning of DNA // *Nano Letters*. 2004. V. 4(5). P. 905-909. doi: 10.1021/nl049658r
86. Skinner G.M., Visscher K., Mansuripur M. Biocompatible Writing of Data into DNA // *J. Bionanoscience*. 2007.V. 1(1). P. 1-5. doi: 10.1166/jbns.2007.005
87. Smith LM, Corn RM, Condon AE, Lagally MG, Frutos AG, Liu Q, Thiel AJ. A surface-based approach to DNA computation. // *J Comput Biol*. 1998. V.5(2). P.255-267. DOI: 10.1089/cmb.1998.5.255
88. Smith GC, Fiddes CC, Hawkins JP, Cox JP. Some possible codes for encrypting data in DNA. // *Biotechnol Lett*. 2003. V.25(14). P.1125-1130. 10.1023/A:1024539608706
89. Su X, Smith LM. Demonstration of a universal surface DNA computer. // *Nucleic Acids Res*. 2004. V.32(10). P.3115-3123. DOI: 10.1093/nar/gkh635
90. Tagore S., Bhattacharya S., Islam M.A., Islam M.L. DNA Computation: Applications and Perspectives // *Journal of Proteomics & Bioinformatics*. 2010. V. 3. P. 234-243. doi:10.4172/jpb.1000145
91. Wang L., Liu Q., Frutos A.G., Gillmor S.D., Thiel A.J., Strother T.C., Condon A.E., Corn R.M., Lagally M.G., Smith L.M. Surface-based DNA computing operations: DESTROY and READOUT // *Biosystems*. 1999. V. 52. P. 189 - 191. doi: 10.1016/S0303-2647(99)00046-5
92. Takahashi C.N., Nguyen B.H., Strauss K., Ceze L. Demonstration of End-to-End Automation of DNA Data Storage. // *Sci. Rep*. 2019. V.9(1):4998. doi: 10.1038/s41598-019-41228-8
93. UbaidurRahman N.H., Balamurugan C., Mariappan R. A Novel DNA Computing Based Encryption and Decryption Algorithm // *Procedia Computer Science*. 2015. V. 46. P.463-475. doi: 10.1016/j.procs.2015.02.045
94. Wang L, Liu Q, Frutos AG, Gillmor SD, Thiel AJ, Strother TC, Condon AE, Corn RM, Lagally MG, Smith LM. Surface-based DNA computing operations: DESTROY and READOUT. // *Biosystems*. 1999. V.52(1-3). P.189-191. DOI: 10.1016/s0303-2647(99)00046-5
95. Wang L, Liu Q, Frutos AG, Gillmor SD, Thiel AJ, Strother TC, Condon AE, Corn RM, Lagally MG, Smith LM. Surface-based DNA computing operations: DESTROY and READOUT. // *J Comput Biol*. 1998. V.5(2). P.269-278. DOI: 10.1089/cmb.1998.5.269
96. Watson J.D., Crick F.H.C. A structure for deoxyribose nucleic acid. // *Nature*. 1953. V. 171(4356). P. 737-738. doi:10.1038/171737a0
97. Williams E.D., Ayres R.U., Heller M. The 1.7 kilogram microchip: energy and material use in the production of semiconductor devices. // *Environ Sci Technol*. 2002. V.36(24). P.5504-5510. DOI: 10.1021/es025643o
98. Wong P.C., Wong K-K., Foote H. Organic data memory using the DNA approach // *Communications of the ACM*. 2003. V.46(1). P.95-98. DOI: 10.1145/602421.602426
99. Wu H. An improved surface-based method for DNA computation // *BioSystems*. 2001. V. 59(1). P. 1-5. doi: 10.1016/s0303-2647(00)00133-7

100. Xiao G., Lu M., Qin L., Lai X. New field of cryptography: DNA cryptography. Chinese Sci Bull 2006. V.51. P., 1413–1420. doi:10.1007/s11434-006-2012-5
101. Yachie N., Sekiyama K., Sugahara J., Ohashi Y., Tomita M. Alignment-based approach for durable data storage into living organisms. // *Biotechnol Prog*. 2007. V.23(2). P.501-505. DOI: 10.1021/bp060261y
102. Yamamoto M., Kashiwamura S., Ohuchi A., Furukawa M. Large-scale DNA memory based on the nested PCR // *Natural Computing*. 2008. V. 7(3). P.335–346. doi: 10.1007/s11047-008-9076-x
103. Yazdi S.M.H.T., Yuan Y., Ma J., Zhao H., Milenkovic O. A Rewritable, Random-Access DNA-Based Storage System. // *Sci. Rep*. 2015. V.5:14138. doi: 10.1038/srep14138.
104. Yazdi S.M.H.T., Gabrys R., Milenkovic O. Portable and Error-Free DNA-Based Data Storage. // *Sci. Rep*. 2017. V.7(1):5011. doi: 10.1038/s41598-017-05188-1.
105. Zala K. Poetry in the genes. // *Nature*. 2009. V.485. P.35.
106. Zhirnov V., Zadegan R.M., Sandhu G.S., Church G.M., Hughes W.L. Nucleic acid memory. // *Nat Mater*. 2016. V.15(4). P.366-370. doi: 10.1038/nmat4594.
107. Zhong Y., Qi S., Sheng F., Tian J., Zhu P., Yang P., Cai X. A new digital information storing and reading system based on synthetic DNA. // *Sci China Life Sci*. 2018. V.61(6). P.733-735. doi: 10.1007/s11427-017-9131-7

References

1. Adleman LM. Molecular computation of solutions to combinatorial problems. *Science*. 1994. V.266(5187). P.1021-1024. DOI: 10.1126/science.7973651
2. Adleman LM. Response. *Science*. 1995 Apr 28;268(5210):483-4. DOI: 10.1126/science.268.5210.483
3. Agrawal A. Bhopale A., Sharma J., Shizan Ali M., Gautam D. Implementation of DNA algorithm for secure voice communication. *International Journal of Scientific & Engineering Research*. 2012. V.3(6). P.1-5.
4. Ailenberg M, Rotstein O. An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques*. 2009. V.47(3). P.747-754. doi: 10.2144/000113218.
5. Akram F., Haq I.U., Ali H., Laghari A.T. Trends to store digital data in DNA: an overview. *Mol. Biol. Rep*. 2018. V.45(5). P.1479-1490. doi: 10.1007/s11033-018-4280-y
6. Arita M. Writing information into DNA // In: Jonoska N., Păun G., Rozenberg G. (eds) Aspects of Molecular Computing. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2004. V. 2950. P. 23-35. DOI: 10.1007/978-3-540-24635-0_2
7. Arita M, Ohashi Y. Secret signatures inside genomic DNA. *Biotechnol Prog*. 2004. V.20(5). P.1605-1607. DOI: 10.1021/bp049917i
8. Avery O.T., MacLeod C.M., McCarty M. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J. Exp. Med.* 1944. V.79(2). P.137–158. DOI: 10.1084/jem.79.2.137
9. Bancroft C., Bowler T., Bloom B., Clelland C.T. Long-term storage of information in DNA. *Science*. 2001. V.293(5536). P.1763-1765. DOI: 10.1126/science.293.5536.1763c
10. Bancroft F.C., Cleland C. DNS - based steganography Patent. US 6,312,911 B1. Data of patent Nov. 6, 2001
11. Baum E.B. Building an associative memory vastly larger than the brain. *Science*. 1995. V.268(5210). P.583-585. DOI: 10.1126/science.7725109
12. Baymiev An.Kh., Kuluev B.R., Vershinina Z.R., Knyazev A.V., Chemeris D.A., Rozhnova N.A., Gerashchenkov G.A., Mikhailova E.V., Baymiev Al.Kh., Chemeris A.V. CRISPR/Cas genome editing (plants) and society. *Biomics*. 2017. V. 9(3). P. 183-202. (In Russian)
13. Baymiev An.Kh., Chemeris D.A., Kiryanova O.Yu., Matniyazov R.T., Valeev A.Sh., Baymiev Al.Kh., Gubaydullin I.M., Chemeris A.V. Bioinformatic resources for in silico search of the CRISPR loci in the genomes of prokaryotes. *Biomics.* 2017. V.9(3). P. 229-244. (In Russian)
14. Bhat W.A. Bridging data-capacity gap in big data storage. *Future Generation Computer Systems*. 2018. V. 87. P. 538-548. doi: 10.1016/j.future.2017.12.066
15. Blawat M., Gaedke K., Huetter I., Chen X-M., Turczyk B., Inverso S., Pruitt B., Church G. Forward Error Correction for DNA Data Storage. *Procedia Computer Science*. 2016. V.80. P. 1011-1022. doi: 10.1016/j.procs.2016.05.398
16. Bornholt J., Lopez R., Carmean D., Ceze L., Seelig G., Strauss K. A DNA-Based Archival Storage System. *ASPLOS '16 Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*. 2017. P. 637-649. doi: 10.1145/2872362.2872397
17. Bornholt J., Lopez R., Carmean D.M., Ceze L., Seelig G., Strauss K. Toward a DNA-Based Archival Storage System. *IEEE Micro*. 2017a. V.37(3). P.98-104. DOI: 10.1109/MM.2017.70

18. Bryksin AV, Matsumura I. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. *Biotechniques.* 2010 Jun;48(6):463-5. doi: 10.2144/000113418.
19. Bunow B. On the potential of molecular computing. *Science*. 1995 Apr 28;268(5210):482-3. DOI: 10.1126/science.7725087
20. Byrne J., Dahm R. Friedrich Miescher and the 150th anniversary of the discovery of DNA. *Biomics*. 2019. V.11(3). P. DOI: 10.31301/2221-6197.bmcs.2019-
21. Chemeris A.V. CRISPR/Cas SYSTEMS (Special thematic issue). *Biomics*. 2017. V.9(3). P. 148-154. (In Russian)
22. Chemeris A.V., Rozhnova N.A., Gerashchenkov G.A. Some recent improvements in genome editing techniques. *Proceedings of the RAS Ufa Scientific Centre*. 2018. №3(5). P. 86–93. (In Russian)
23. Chemeris D.A., Kiryanova O.Yu., Gerashchenkov G.A., Kuluev B.R., Rozhnova N.A., Matniyazov R.T., Baymiev An.Kh., Baymiev Al.Kh., Gubaidullin I.M., Chemeris A.V. Bioinformatic resources for CRISPR/Cas genome editing, *Biomics*, 2017. V.9(3). P. 203-228. (In Russian)
24. Ceze L, Nivala J, Strauss K. Molecular digital data storage using DNA. *Nat. Rev. Genet*. 2019. V.20(8). P.456-466. doi: 10.1038/s41576-019-0125-3
25. Chen K., Kong J., Zhu J., Ermann N., Predki P., Keyser U.F. Digital data storage using DNA nanostructures and solid-state nanopores. *Nano Lett*. 2019a. V.19(2). P.1210-1215. doi: 10.1021/acs.nanolett.8b04715
26. Chen W.D., Kohll A.X., Nguyen B.H., Koch J., Heckel R., Stark W.J., Ceze L., Strauss K., Grass R.N. Combining Data Longevity with High Storage Capacity – Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles. *Adv. Func. Mater*. 2019. V. 29(28). 1901672. doi: 10.1002/adfm.201901672
27. Choi Y., Ryu T., Lee A.C., Choi H., Lee H., Park J., Song S-H., Kim S., Kim H., Park W., Kwon S. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Sci. Rep*. 2019. V.9(1):6582. doi: 10.1038/s41598-019-43105-w
28. Church G.M., Gao Y., Kosuri S. Next-generation digital information storage in DNA. *Science*. 2012. V.337(6102). P.1628. DOI: 10.1126/science.1226355
29. Clelland C.T., Risca V., Bancroft C. Hiding messages in DNA microdots. *Nature*. 1999. V.399(6736). P.533-534. DOI: 10.1038/21092
30. Davis J. Microvenus. *Art Journal*. 1996. V. 55(1). P. 70-74. DOI: 10.2307/777811
31. De Silva PY, Ganegoda GU. New Trends of Digital Data Storage in DNA. *Biomed Res Int*. 2016.:8072463. DOI: 10.1155/2016/8072463
32. Erlich Y. Efficient encoding of data for storage in polymers such as DNA. US Patent Application No 2019/0020353. January 17, 2019
33. Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. *Science*. 2017. V.355(6328). P.950-954. doi: 10.1126/science.aaj2038.
34. Feynman R.P. There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics // *Engineering and Science (California Institute of Technology)*. 1960. V23. P.22-36.
35. Frutos AG, Liu Q, Thiel AJ, Sanner AM, Condon AE, Smith LM, Corn RM. Demonstration of a word design strategy for DNA computing on surfaces. *Nucleic Acids Res*. 1997. V.25(23). P.4748-4757. DOI: 10.1093/nar/25.23.4748
36. Frutos A.G., Smith, L.M., Corn R.M. Enzymatic ligation reactions of DNA "words" on surfaces for DNA computing. *Journal of the American Chemical Society*. 1998. V. 120. P. 10277-10282.
37. Fu P. Biomolecular computing: is it ready to take off? *Biotechnol J*. 2007. V.2(1). P.91-101. DOI: 10.1002/biot.200600134
38. Gerashchenkov G.A., Rozhnova N.A., Kuluev B.R., Kiryanova O.Yu., Gumerova G.R., Knyazev A.V., Vershinina Z.R., Mikhailova E.V., Chemeris D.A., Matniyazov R.T., Baimiev An.Kh., Gubaidullin I.M., Baimiev Al.Kh., Chemeris A.V. Design of guide RNA for CRISPR/Cas plant genome editing. *Molecular Biology*. 2020. V.54(1).
39. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA 3rd, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009. V.6(5). P.343-345. doi: 10.1038/nmeth.1318
40. 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. V.494(7435). P.77-80. doi: 10.1038/nature11875.
41. 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. *Angew Chem Int Ed Engl*. 2015. V.54(8). P.:2552-2555. doi: 10.1002/anie.201411378.
42. Gustafsson C. For anyone who ever said there's no such thing as poetic gene. *Nature*. 2009. V.458. P. 703.
43. Heckel R., Mikutis G., Grass R.N. A Characterization of the DNA Data Storage Channel. *Sci Rep*. 2019. V.9(1):9663. doi: 10.1038/s41598-019-45832-6.
44. Hodgson C.P. A DNA text code. *Biotechniques*. 1990. V. 9(3). P. 312.

45. Huffman D.A. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*. 1952. V. 40(9). P. 1098 – 1101. DOI: 10.1109/JRPROC.1952.273898
46. Hwang B, Bang D. Toward a new paradigm of DNA writing using a massively parallel sequencing platform and degenerate oligonucleotide. *Sci. Rep*. 2016. V.6:37176. doi: 10.1038/srep37176
47. Interview. Machines smarter than men? *U.S. News & World Report*. 1964. Feb., 24. P.84-86.
48. Jiao S., Goutte R. Code for encryption hiding data into genomic DNA of living organisms. *9th International Conference on Signal Processing*. 2008. DOI: 10.1109/ICOSP.2008.4697576
49. Jimenez-Sanchez A. A proposal for a DNA-based computer code. *International Invention Journal of Biochemistry and Bioinformatics*. 2013. V.1(1) P. 1-4.
50. Kashiwamura S., Yamamoto M., Kameda A., Shiba T., Ohuchi A. Potential for enlarging DNA memory: the validity of experimental operations of scaled-up nested primer molecular memory. *Biosystems*. 2005. V.80(1). P.99-112. DOI: 10.1016/j.biosystems.2004.10.007
51. Kuluev B.R., Baymiev An.Kh., Chemeris D.A., Matniyazov R.T., Gerashchenkov G.A., Nikonorov Yu.M., Baymiev Al.Kh., Chemeris A.V. The application of the CRISPR loci not for editing of genomes. *Biomics*. 2017. V.9(3). P.271-283. (In Russian)
52. Kuluev, B.R., Gerashchenkov, G.A., Rozhnova, N.A., Bayimiev, An.Kh., Vershinina, Z.R., Knyazev, A.V., Matniyazov, R.T., Gumerova, G.R., Mikhailova, E.V., Nikonorov, Yu.M., Chemeris, D.A., Baymiev, Al.Kh., and Chemeris, A.V., CRISPR/Cas genome editing of plants. *Biomics*, 2017a. V.9(3). P. 155-182. (In Russian)
53. Kuluev B.R., Kiryanova O.Yu., Gerashchenkov G.A., Rozhnova N.A., Gumerova G.R., Vershinina Z.R., Matniyazov R.T., Akhmetzyanova L.U., Knyazev A.V., Mikhailova, E.V., Garafutdinov R.R., Baymiev An.Kh., Gubaidullin I.M., Baymiev Al.Kh., Chemeris A.V. Some novelties in CRISPR/Cas genome editing and related areas. *Biomics.* 2019. V. 11(3). P. 315-343 DOI: 10.31301/2221-6197.bmcs.2019-27
54. Kuluev B.R., Gumerova G.R., Mikhailova E.V., Gerashchenkov G.A., Rozhnova N.A., Vershinina Z.R., Knyazev A.V., Matniyazov R.T., Baymiev An.Kh. Baymiev Al.Kh., Chemeris A.V. Delivery of CRISPR/Cas Components into Higher Plant Cells for Genome Editing. *Russian Journal of Plant Physiology*. 2019. V.66(5). P.694-706. DOI: 10.1134/S102144371905011X
55. Landegren D.A. DNA probes and automation. *Curr. Opin. Biotechnol*. 1992. V.3(1). P.12-17. doi: 10.1016/0958-1669(92)90119-4
56. Landweber LF, Kari L. The evolution of cellular computing: nature's solution to a computational problem. Biosystems. 1999. V.52(1-3). P.3-13. DOI: 10.1016/s0303-2647(99)00027-1
57. Li D., Li X., Huang H., Li X. The surface-based approach for DNA computation is unreliable for SAT. *Biosystems.* 2005. V. 82. P. 20 - 25. doi:10.1016/j.biosystems.2005.05.007
58. Li D., Li X., Huang H., Li X. Scalability of the surface-based DNA algorithm for 3-SAT. *Biosystems*. 2006. V. 85. P. 95-98. doi: 10.1016/j.biosystems.2005.12.002
59. Limbachiya D. et al. On optimal family of codes for archival DNA storage. *2015 Seventh International Workshop on Signal Design and its Applications in Communications (IWSDA)*. 2015. DOI: 10.1109/IWSDA.2015.7458386
60. Linial M, Linial N. On the potential of molecular computing. *Science*. 1995. V.268(5210). P.481. DOI: 10.1126/science.7725085
61. Lipton RJ. DNA solution of hard computational problems. *Science*. 1995. V.268(5210). P.542-545. DOI: 10.1126/science.7725098
62. Liu Q, Frutos AG, Thiel AJ, Corn RM, Smith LM. DNA computing on surfaces: encoding information at the single base level. *J Comput Biol*. 1998. V.5(2). P.269-278. DOI: 10.1089/cmb.1998.5.269
63. Liu Q, Wang L, Frutos AG, Condon AE, Corn RM, Smith LM. DNA computing on surfaces. *Nature*. 2000. V.403(6766). P.175-179. DOI: 10.1038/35003155
64. Lo YM, Yiu KF, Wong SL. On the potential of molecular computing. *Science*. 1995. V.268(5210). P.481-482. DOI: 10.1126/science.7725086
65. Lopez R., Chen Y-J., Ang S.D, Yekhanin S., Makarychev K., Racz M.Z., Seelig G., Strauss K., Ceze L. DNA assembly for nanopore data storage readout. *Nat. Commun*. 2019. V.10(1):2933. doi: 10.1038/s41467-019-10978-4
66. Melkikh AV. DNA computing, computation complexity and problem of biological evolution rate. *Acta Biotheor*. 2008. V.56(4). P.285-295. doi: 10.1007/s10441-008-9055-8
67. Miescher F. Ueber die chemische Zusammensetzung der Eiterzellen. *Medicinisch-chemische Untersuchungen*. 1871. V.4. P. 441–460.
68. Mills A.P., Yurke B. Data trnsmission using DNA oligomers. Patent US 6, 537,747 B1. Data of patent Mar. 25, 2003
69. Murugan A., Thilagavathy R. Securing cloud data using DNA and Morse code: A triple encryption

scheme // *International Journal of Control Theory and Applications*. 2017. V.10(23). P. 31-38.
70. Nair ACC, Mahalakshmi T. Visualization of genomic data using inter-nucleotide distance signals. *Conference Proceedings.* 2005. (https://www.researchgate.net/publication/228643588_Visualization_of_genomic_data_using_inter-nucleotide_distance_signals)
71. Neiman M.S. Some fundamental issues of microminiaturisation. *Radiotekhnika.* 1964. V.19(1). P. 3-12. (In Russian)
72. Neiman M.S. On the relationships between the reliability, performance and degree of microminiaturisation at the molecular-atomic level. *Radiotekhnika*. 1965. V.20(1). P. 1-9. (In Russian)
73. Neiman M.S. On the molecular memory systems and the directed mutations. *Radiotekhnika*. 1965a. V.20(6). P. 1-8. (In Russian)
74. Nguyen H.H., Park J., Hwang S., Kwon O.S, Lee C., Shin Y., Ha T.H, Kim M. On-Chip fluorescence switching system for constructing a rewritable random-access data storage device. *Sci Rep*. 2018. V.8. P. 337. doi: 10.1038/s41598-017-16535-7
75. Nishikawa A., Yamamura M., Hagiya M. DNA computation simulator based on abstract bases. *Soft Computing*. 2001. P. 25-38. doi: 10.1007/s005000000062
76. Organick L., Ang S.D., Chen Y.J., Lopez R., Yekhanin S., Makarychev K., Racz M.Z., Kamath G., Gopalan P., Nguyen B., Takahashi C.N., Newman S., Parker H.Y., Rashtchian C., Stewart K., Gupta G., Carlson R., Mulligan J., Carmean D., Seelig G., Ceze L., Strauss K. Random access in large-scale DNA data storage. *Nat. Biotechnol*. 2018. V. 36(3). P.242-248. doi: 10.1038/nbt.4079. Erratum: *Nat. Biotechnol*. 2018. V.36(7). P.660. doi: 10.1038/nbt0718-660c
77. Panda D., Molla K.A., Baig M.J., Swain A., Behera D., Dash M. DNA as a digital information storage device: hope or hype? *3 Biotech*. 2018. V.8(5):239. doi: 10.1007/s13205-018-1246-7.
78. Ping Z., Ma D., Huang X., Chen S., Liu L., Guo F., Zhu S.J., Shen Y. Carbon-based archiving: current progress and future prospects of DNA-based data storage. *Gigascience*. 2019. V.8(6). pii: giz075. doi: 10.1093/gigascience/giz075
79. Portney N.G., Wu Y., Quezada L.K., Lonardi S., Ozkan M. Length-based encoding of binary data in DNA. *Langmuir*. 2008. V.24(5). P. 1613-1616. doi: 10.1021/la703235y
80. Rashid O.F., Othman Z.A., Zainitdin S. A novel DNA sequence approach for network intrusion detection system based on cryptography encoding method. *International Journal on Advanced Science, Engineering and Information Technology*. V. 7(1). P. 183-189. doi: 10.18517/ijaseit.7.1.1569
81. Schmidt KA, Henkel CV, Rozenberg G, Spaink HP. DNA computing using single-molecule hybridization detection. *Nucleic Acids Res*. 2004. V.32(17). P.4962-2968. DOI: 10.1093/nar/gkh817
82. Schouhamer Immink K.A., Cai K. Design of Capacity-Approaching Constrained Codes for DNA-Based Storage Systems. *IEEE Communications Letters*. 2018. V. 22(2). P. 224 – 227. DOI: 10.1109/LCOMM.2017.2775608
83. Seeman N.C. Nucleic acid junctions and lattices. *J. Theor. Biol*. 1982. V.99(2). P.237-247. DOI: 10.1016/0022-5193(82)90002-9
84. Shin J., Pierce N. Rewritable memory by controllable nanopatterning of DNA. *Nano Letters*. 2004. V. 4(5). P. 905-909. doi: 10.1021/nl049658r
85. Skinner G.M., Visscher K., Mansuripur M. Biocompatible Writing of Data into DNA. *J. Bionanoscience*. 2007.V. 1(1). P. 1-5. doi: 10.1166/jbns.2007.005
86. Smith LM, Corn RM, Condon AE, Lagally MG, Frutos AG, Liu Q, Thiel AJ. A surface-based approach to DNA computation. *J Comput Biol*. 1998. V.5(2). P.255-267. DOI: 10.1089/cmb.1998.5.255
87. Smith GC, Fiddes CC, Hawkins JP, Cox JP. Some possible codes for encrypting data in DNA. *Biotechnol Lett*. 2003. V.25(14). P.1125-1130. 10.1023/A:1024539608706
88. Su X, Smith LM. Demonstration of a universal surface DNA computer. *Nucleic Acids Res*. 2004. V.32(10). P.3115-3123. DOI: 10.1093/nar/gkh635
89. Tagore S., Bhattacharya S., Islam M.A., Islam M.L. DNA Computation: Applications and Perspectives. *Journal of Proteomics & Bioinformatics*. 2010. V. 3. P. 234-243. doi:10.4172/jpb.1000145
90. Wang L., Liu Q., Frutos A.G., Gillmor S.D., Thiel A.J., Strother T.C., Condon A.E., Corn R.M., Lagally M.G., Smith L.M. Surface-based DNA computing operations: DESTROY and READOUT. *Biosystems*. 1999. V. 52. P. 189 - 191. doi: 10.1016/S0303-2647(99)00046-5
91. Takahashi C.N., Nguyen B.H., Strauss K., Ceze L. Demonstration of End-to-End Automation of DNA Data Storage. *Sci. Rep*. 2019. V.9(1):4998. doi: 10.1038/s41598-019-41228-8
92. UbaidurRahman N.H., Balamurugan C., Mariappan R. A Novel DNA Computing Based Encryption and Decryption Algorithm. *Procedia Computer Science*. 2015. V. 46. P.463-475. doi: 10.1016/j.procs.2015.02.045
93. Wang L, Liu Q, Frutos AG, Gillmor SD, Thiel AJ, Strother TC, Condon AE, Corn RM, Lagally MG, Smith LM. Surface-based DNA computing

operations: DESTROY and READOUT. *Biosystems*. 1999. V.52(1-3). P.189-191. DOI: 10.1016/s0303-2647(99)00046-5

94. Vershinina Z.R., Kuluev B.R., Gerashchenkov G.A., Knyazev A.V., Chemeris D.A., Gumerova G.R., Baymiev Al.Kh., Chemeris A.V. Evolution of methods for genome editing. *Biomics*. 2017. V.9(3). P. 245-270. (In Russian)
95. Wang L, Liu Q, Frutos AG, Gillmor SD, Thiel AJ, Strother TC, Condon AE, Corn RM, Lagally MG, Smith LM. Surface-based DNA computing operations: DESTROY and READOUT. *J Comput Biol*. 1998. V.5(2). P.269-278. DOI: 10.1089/cmb.1998.5.269
96. Watson J.D., Crick F.H.C. A structure for deoxyribose nucleic acid. *Nature*. 1953. V. 171(4356). P. 737-738. doi:10.1038/171737a0
97. Williams E.D., Ayres R.U., Heller M. The 1.7 kilogram microchip: energy and material use in the production of semiconductor devices. *Environ Sci Technol*. 2002. V.36(24). P.5504-5510. DOI: 10.1021/es025643o
98. Wong P.C., Wong K-K., Foote H. Organic data memory using the DNA approach. *Communications of the ACM*. 2003. V.46(1). P.95-98. DOI: 10.1145/602421.602426
99. Wu H. An improved surface-based method for DNA computation. *BioSystems*. 2001. V. 59(1). P. 1-5. doi: 10.1016/s0303-2647(00)00133-7
100. Xiao G., Lu M., Qin L., Lai X. New field of cryptography: DNA cryptography. Chinese Sci Bull 2006. V.51. P., 1413–1420. doi:10.1007/s11434-006-2012-5
101. Yachie N., Sekiyama K., Sugahara J., Ohashi Y., Tomita M. Alignment-based approach for durable data storage into living organisms. *Biotechnol Prog*. 2007. V.23(2). P.501-505. DOI: 10.1021/bp060261y
102. Yamamoto M., Kashiwamura S., Ohuchi A., Furukawa M. Large-scale DNA memory based on the nested PCR. *Natural Computing*. 2008. V. 7(3). P.335–346. doi: 10.1007/s11047-008-9076-x
103. Yazdi S.M.H.T., Yuan Y., Ma J., Zhao H., Milenkovic O. A Rewritable, Random-Access DNA-Based Storage System. *Sci. Rep*. 2015. V.5:14138. doi: 10.1038/srep14138.
104. Yazdi S.M.H.T., Gabrys R., Milenkovic O. Portable and Error-Free DNA-Based Data Storage. *Sci. Rep*. 2017. V.7(1):5011. doi: 10.1038/s41598-017-05188-1.
105. Zala K. Poetry in the genes. *Nature*. 2009. V.485. P.35.
106. Zhirnov V., Zadegan R.M., Sandhu G.S., Church G.M., Hughes W.L. Nucleic acid memory. *Nat Mater*. 2016. V.15(4). P.366-370. doi: 10.1038/nmat4594.
107. Zhong Y., Qi S., Sheng F., Tian J., Zhu P., Yang P., Cai X. A new digital information storing and reading system based on synthetic DNA. *Sci China Life Sci*. 2018. V.61(6). P.733-735. doi: 10.1007/s11427-017-9131-7