УДК 004.085

СТАН І ПРОБЛЕМИ СТВОРЕННЯ ДНК-ПАМ'ЯТІ

А. А. Крючин, Є. В. Беляк, Є. А. Крючина, А. В. Потебня

Інститут проблем реєстрації інформації НАН України

Представлено результати аналізу стану розробок систем пам'яті на ДНК-молекулах. Показані значні потенційні можливості таких систем пам'яті для організації довготермінового зберігання великих обсягів інформації. Визначені умови широкого використання пам'яті на ДНК. Показано, що ключовим моментом використання пам'яті типу WORM на ДНК-молекулах є суттєве підвищення швидкості секвенування записаних послідовностей нуклеотидів. Наведені дані про умови зберігання чипів пам'яті на ДНК, при яких забезпечується довготермінове зберігання великих обсягів інформації.

Ключові слова: ДНК-пам'ять, секвенування, кодування даних, big data.

СОСТОЯНИЕ И ПРОБЛЕМЫ СОЗДАНИЯ ДНК-ПАМЯТИ

А. А. Крючин, Е. В. Беляк, Е. А. Крючина, А. В. Потебня

Институт проблем регистрации информации НАН Украины

Представлены результаты анализа состояния разработок систем памяти на ДНК-молекулах. Показаны значительные потенциальные возможности таких систем памяти для организации долгосрочного хранения больших объемов информации. Определены условия широкого использования памяти на ДНК. Показано, что ключевым моментом использования памяти типа WORM на ДНК-молекулах является существенное повышение скорости секвенирования записанных последовательностей нуклеотидов. Приведенные данные об условиях хранения чипов памяти на ДНК, при которых обеспечивается долгосрочное хранение больших объемов информации.

Ключевые слова: ДНК-память, секвенирование, кодирование данных, big data.

STATUS AND PROBLEMS OF DNA MEMORY CREATING

A. Kryuchyn, Ye. V. Belyak, Ye. A. Kryuchyna, A. V. Potebnya

Institute for Information Recording of NAS of Ukraine

The results of the analysis of the development of storage systems on DNA molecules are given. The considerable potential for such storage systems for the organization of long-term storage of large volumes of information is shown. The conditions for the extensive use of memory on DNA are determined. It is shown that the key to the use of WORM-type memory on DNA molecules is a significant increase in the speed of sequencing nukleatidov recorded sequences. Data on conditions of storage memory chips on DNA, which provide long-term storage of large volumes of information is presented.

Key words: DNA memory, sequencing, data encoding, big data.

Вступ. Обсяги інформації, яка підлягає довготерміновому зберіганню, постійно збільшуються. За оцінками Міжнародної корпорації даних (IDC) обсяг даних, накопичених по всьому світу в 2013 році, становив 4,4 ЗБ і має збільшиться до 40 ЗБ у 2020 році. В останні роки спостерігається 40 % щорічне зростання обсягів інформації [1]. Особливо швидко зростають обсяги розшифрованої генетичної інформації, яка підлягає довготривалому зберіганню [2]. Проблема довготермінового зберігання даних вирішується створенням спеціальних оптичних носіїв інформації, в яких використовується запис даних у вигляді мікрорельєфних структур на поверхні високостабільних матеріалів [3–5]. Використання для довготермінового зберігання даних тільки оптичних носіїв не може вирішити проблему зберігання великих інформаційних масивів (big data). Швидкозростаючий потік даних, які потребують довготермінового зберігання, вимагає розробки принципово нових систем запису та зберігання даних. Останнім часом вважають, що довготермінове зберігання великих

інформаційних масивів (big data) може здійснюватися з використанням запису на молекули ДНК [6–11]. ДНК має багато потенціальних переваг для організації довготермінового зберігання великих обсягів інформації. Головні переваги пам'яті на ДНК-молекулах представлено на рисунку 1.

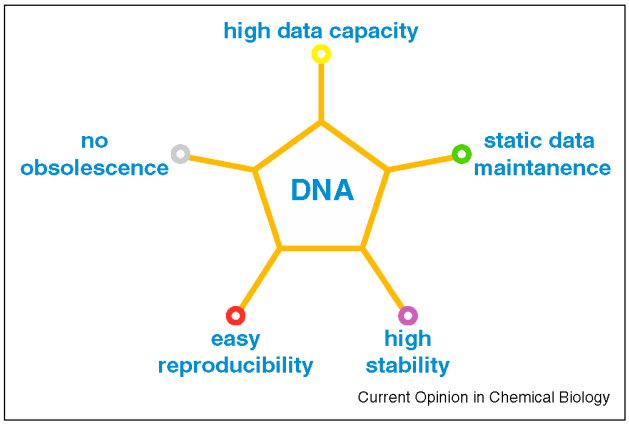

Рис. 1. Головні особливості пам'яті на ДНК, що дозволяють вважати її революційною технологією зберігання даних [12].

Завдяки високій щільності запису інформації в пам'яті на ДНК-молекулах можливе досягнення ємності носіїв, яку важко реалізувати на відомих типах носіїв. На відміну від більшості цифрових носіїв інформації, зберігання даних на ДНК не обмежено одним планарним шаром. При зберіганні даних ДНК може кодувати два біти на нуклеотид що забезпечує 455 ексабайт на один грам одноланцюгової ДНК. У живій природі пам'ять на молекулах ДНК забезпечує надійне зберігання величезних обсягів інформації. Молекула ДНК людини включає в себе близько 3 мільярдів пар нуклеотидів і тому в ній закодована вся інформація про організм людини: його зовнішність, здоров'я або схильність до хвороб, здібності тощо. ДНК-пам'ять має потенційні можливості для забезпечення високонадійного довготермінового зберігання даних. Для живих організмів основною роллю ДНК є довготермінове зберігання даних і успадкованої генетичної інформації. Зокрема, цьому сприяє статичний характер організації зберігання даних, на відміну від більшості магнітних і оптичних систем зберігання даних. Другою перевагою ДНК-пам'яті є стабільність: інформація, яка записана в ДНК з використанням хімічних зв'язків. Генетична інформація може зберігатися десятки тисяч років, що вже дозволяє зчитувати генотипи деяких вимерлих в незапам'ятні часи тварин і рослин, а з подальшим розвитком технології – і повертати їх до життя. Універсальний ензимний механізм запису і зчитування інформації, відшліфований за мільярди років еволюції живої речовини, дозволяє розглядати ДНК-пам'ять в якості майбутнього потенційного стандарту зберігання та зчитування даних. Важливою перевагою зберігання даних на ДНК є те, що ДНК – це біологічна молекула, яка завжди зможе бути біологічно прочитаною без спеціального обладнання, яке може швидко застарівати [6]. Для ДНК-пам'яті не існує проблеми старіння системи зберігання даних, яка є одним з обмежень при створенні систем довготермінового зберігання даних на існуючих магнітних, оптичних носіях та флеш-пам'яті. Серед можливих переваг ДНК-пам'яті необхідно також відзначити технологічність копіювання записаної інформації [12]. Методи запису і кодування даних на ДНК молекулах дозволяють ефективно використовувати технологію стеганографії для захисту інформації. Ще у 1999 році був розроблений метод стеганографії на основі ДНК для кодування секретних повідомлень при реєстрації даних в ДНК-пам'яті [13]. Про високі потенційні можливості ДНК-пам'яті із зберігання великих обсягів інформації свідчить наведена на рисунку 2 порівняльна оцінка ємності носіїв інформації різних типів. Слід зазначити, що прогнози щодо довготермінового зберігання великих обсягів даних на ДНК ґрунтуються на дослідженнях археологів і генетиків. Зовсім недавно була розшифрована 300000-річна мітохондріальна ДНК ведмедів і людини [14, 15].

Відомі випадки, коли в природних умовах ДНК зберігалася 700 тис. років і була доступна після цього для вивчення. Аналізуються можливості збільшення терміну зберігання даних в ДНК до декількох мільйонів років. Наприклад, 2 млн років можна добитися при зберіганні ДНК при температурі -18 градусів Цельсія. Показано, що з підвищенням температури термін придатності генетичних носіїв інформації швидко знижується. При температурі в 10 градусів Цельсія інформація буде зберігатися лише близько 2 тис. років [6]. Більшість розроблених досі методів зберігання даних на основі ДНК базуються на полімеразній ланцюговій реакції (ПЛР) для кодування та зчитування інформації. У процедурі зберігання даних цими методами послідовність даних спочатку перетвориться в послідовності ДНК з використанням ключів шифрування і зворотна процедура дозволяє встановити послідовність закодованих інформаційних одиниць. Дві унікальних області додаються до кожного кінця області даних ДНК для дешифрування даних (рис. 3).

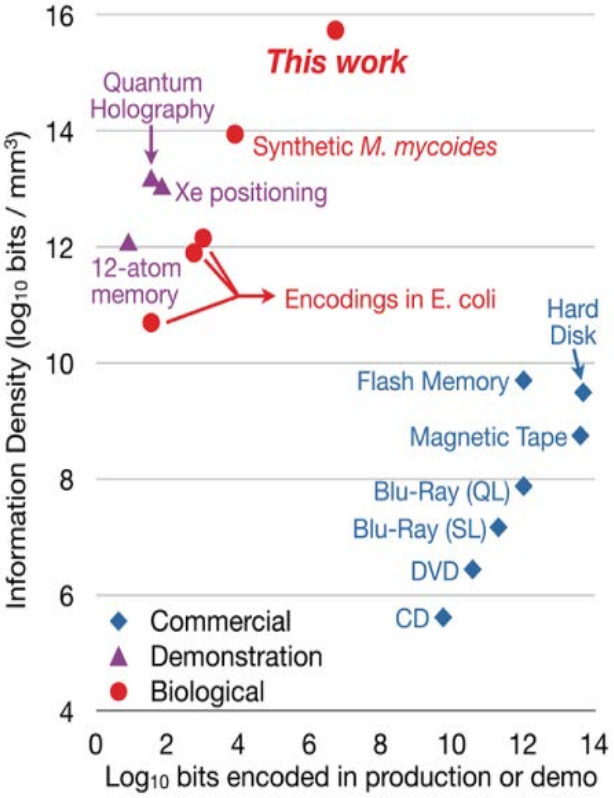

Рис. 2. Щільність запису на різних типах носіїв [7].

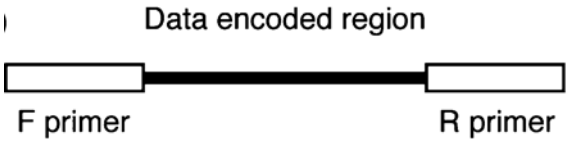

Рис. 3. Структура даних, які зберігаються на синтетичній ДНК [16].

ДНК з послідовністю кодованих даних та прямими і зворотними праймерами вставляється у геномну ДНК. Для зчитування кодованих даних ДНК і обох послідовностей праймерів, ділянка кодованих даних ДНК підсилюється за допомогою ПЛР та декодується з використанням секвентора ДНК.

Ідея і загальні міркування про можливості запису, зберігання і зчитування інформації на ДНК-молекулах були спочатку зроблені М. С. Нейманом [17]. Л. М. Адлеман (Adleman L. M.) визначив шляхи обробки інформації в молекулярному масштабі і побудови на її основі комп'ютера загального призначення. Він визначив, що молекулярні структури здатні вирішувати набагато більше коло завдань, наприклад створювати ефективні пристрої пам'яті [18]. Перший пристрій ДНК-пам'яті було продемонстровано ще в 1988 році, коли за допомогою ДНК-молекул вдалося закодувати 7920 біт даних [19]. Технологія зберігання цифрових даних в ДНК, яка зараз використовується у експериментальних системах ДНК-пам'яті була вперше запропонована і описана у роботах [13, 20].

Кодування у ДНК пам'яті. Молекула ДНК складається з двох скручених один з одним в спіраль ланцюгів, побудованих з чотирьох нуклеотидів: A, G, T і C, які утворюють генетичний алфавіт. Аденін (A) і гуанін (G) відносяться до класу пуринів, а до числа піримідинів – цитозин (C) і тимін (T). Молекула ДНК зберігає інформацію в четверичній системі числення, за кількістю нуклеотидів (0 = A аденін, 1 = T тимін, 2 = C цитозин, 3 = G гуанін). Це компактний контейнер з щільністю запису в тисячі разів більше, ніж в існуючих носіїв. Однак, щоб технологія перейшла від наукових випробувань до комерційного використання, потрібно вирішити ряд проблем. Одна з них – специфіка цифрової інформації, в якій одні й ті ж біти можуть багаторазово повторюватися (CCCCCCCCCCCCCCC). Якщо багато разів повторювати один і той же нуклеотид в молекулі ДНК, то це негативно впливає на стабільність кластера, тому інформація може бути втрачена, навіть при використанні надлишкового дублювання і корекції помилок. Зберігання можна організувати використовуючи нуклеотиди: A і C в якості 0, а T і G -одиниці. Для підвищення надійності зберігання буде використовуватись надлишкове кодування і механізм виправлення помилок Ріда–Соломона [6]. Використання такого методу кодування та сучасних методів секвенування дозволило записати та відтворити 5.27-мегабітний блок даних. Блок даних був закодований у 54,898 нуклеотидній послідовності довжиною 159 біт, яка складається з 96-бітного блоку даних, у 19-бітній адресі із зазначенням місця розташування блоку даних і 22 бітів службової інформації для полегшення секвенування на кінцях послідовності. Блок даних був прочитаний 3,000 раз з використанням сучасного секвенсора блоку даних та були виявлені 10 біт помилок (більшість помилок переважно розташовані на кінцях послідовностей) [6]. Для значного зменшення кількості помилок і, як наслідок, збільшення обсягів даних, що зберігаються в ДНК-пам'яті ~2.2 ПБ/г ДНК запропонований спосіб, який полягає в тому, що пропонують відмовитися від четверичної системи (Base-4) на користь троїчної (Base-3), а четвертий

нуклеотид використовувати у службових цілях для розбиття довгих ланцюжків (CCCACCCACCCAC-CCACCC). В таблиці 1 наведено зразок троїчної системи кодування в ДНК-пам'яті [20].

Таблиця 1. Зразок троїчної системи кодування в ДНК-пам'яті

Codons encoding the English alphabet.

Alphabet	Ternary Value	DNA Codon
A	000	AAA
B	001	AAC
C	002	AAT
D	010	ACA
E	011	ACC
F	012	ACT
G	020	ATA
H	021	ATC
I	022	ATT
J	100	CAA
K	101	CAC
L	102	CAT
M	110	CCA
N	111	CCC
O	112	CCT
P	120	CTA
Q	121	CTC
R	122	CTT
S	200	TAA
T	201	TAC
U	202	TAT
V	210	TCA
W	211	TCC
X	212	TCT
Y	220	TTA
Z	221	TTC
Space	222	TTT

При переході з Base-4 на Base-3 втрачається приблизно 25 % інформаційної ємності, але навіть у такому варіанті інформаційна щільність запису становить 2,2 петабайта на 1 грам біологічного матеріалу. Схема описаного процесу кодування представлена на рисунку 4 [10].

Враховуючи темпи вибухового зростання обсягів даних, зберігання даних на ДНК стає незамінною технологією зберігання даних завдяки низькій вартості обслуговування, високій щільності даних, екологічності і довговічності. Розробка ефективних методів кодування та алгоритми декодування є актуальним завданням DNACloud, що може розглядатися як потенційний інструмент, щоб конвертувати дані файли у ДНК і навпаки. Планується розширення можливостей програмного забезпечення для кодування великого розміру даних, здійснюючи більш досконале кодування з урахуванням методів корекції помилок [10].

Аналіз методів запису інформації на молекулах ДНК. У процесі запису даних на штучні ДНК-молекули використовується синтез олігонуклеотидів – синтез відносно коротких фрагментів нуклеїнових кислот із заданою хімічною структурою (послідовністю). Процес відтворення базується на секвенуванні ДНК – визначення нуклеотидної послідовності. В результаті секвенування отримують формальний опис первинної структури лінійної макромолекули у вигляді послідовності мономерів в текстовому вигляді. Процес секвенування виявляється достатньо повільним, тому що ДНК повинні

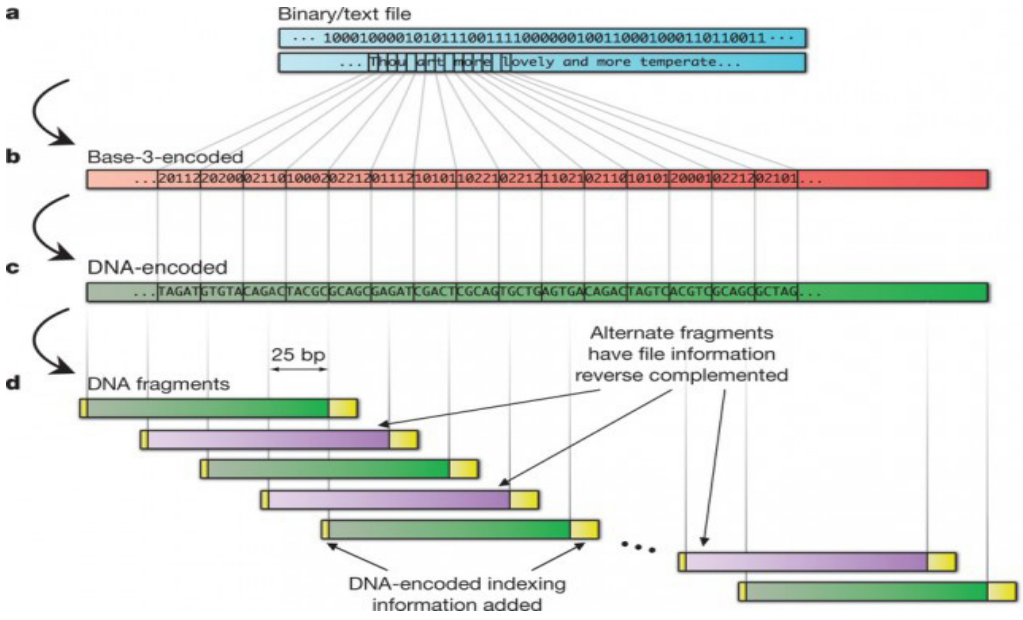

Рис. 4. Схема процесу кодування в ДНК-пам'яті.

бути упорядковані з метою отримання даних, і тому метод призначений для використання з низькою швидкістю доступу до них. Узагальнена схема процесу зберігання даних представлена на рисунку 5.

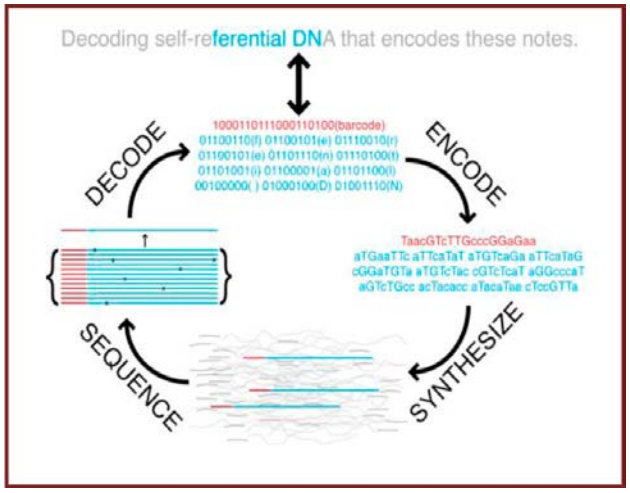

Рис. 5. Загальна схема запису даних на ДНК-молекулах [21].

Ключовими операціями цього процесу є синтез молекул ДНК з визначеною послідовністю нуклеотидів та операція секвенування (визначення послідовності нуклеотидів у записаних даних). Виконання цих операцій вимагає наявності спеціального технологічного обладнання і методик для їх реалізації. В усіх розроблених системах ДНК-пам'яті використовується однаковий процес запису і відтворення інформації: масив даних спочатку ділиться на блоки розміром трохи більше ста біт, потім перекодовується в буквенну послідовність нуклеотидів, на основі якої синтезуються короткі ДНК-ланцюжки (рис. 6).

Зчитування інформації з масиву здійснювалося за допомогою автоматизованої полімеразно-ланцюгової реакції і паралельних ДНК-секвенаторів новітнього покоління: ДНК-ланцюжки багаторазово клонували, далі, одночасно коригуючи помилки, отримані коди з'єднували в масиви даних відповідно до адресних міток, розташованих на кінцях ланцюжків. Для створення ДНК-памяті використовується технологія секвенування нового покоління (СНП) – техніка визначення нуклеотидної послідовності ДНК і РНК для отримання формального опису її первинної структури, яка дозволяє «прочитати» одноразово відразу декілька ділянок геному, що є головною відмінністю від більш ранніх методів секвенування. На рисунку 7 наведено приклад визначення нуклеотидної послідовності ДНК.

СНП здійснюється за допомогою повторюваних циклів подовження ланцюга, індукованого полімеразою, або багаторазового лігування олігонуклеотидів. У ході СНП можуть генеруватися до сотень Мегабайт і навіть Гігабайт нуклеотидних послідовностей за один робочий цикл. Описана процедура не може бути використана для перезапису даних, але може використовуватися для тривалого зберігання даних. Реалізована в роботі [6] схема зберігання даних на молекулах ДНК наведена на рисунку 8.

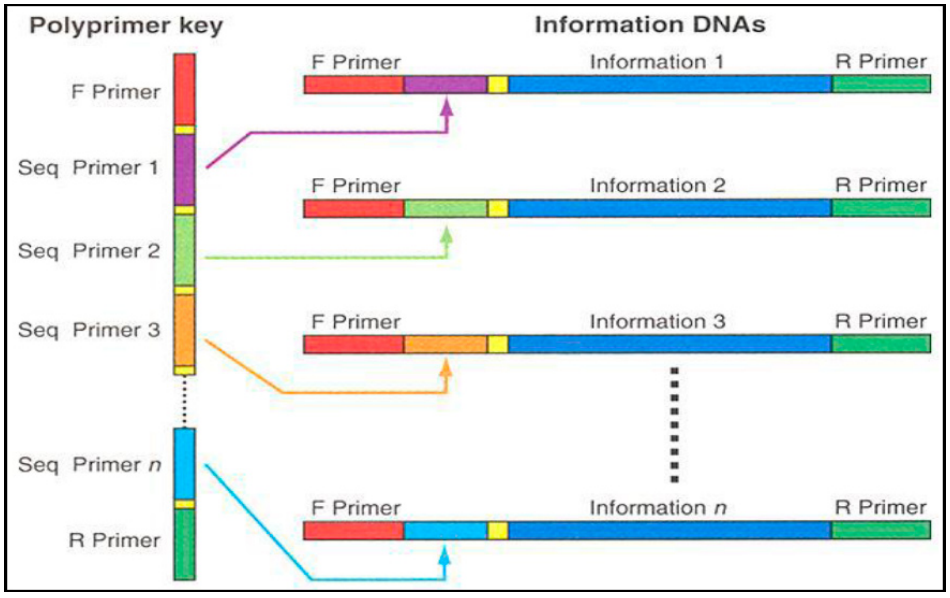

Рис. 6. Структура ДНК-молекул, що використовуються для зберігання даних [21].

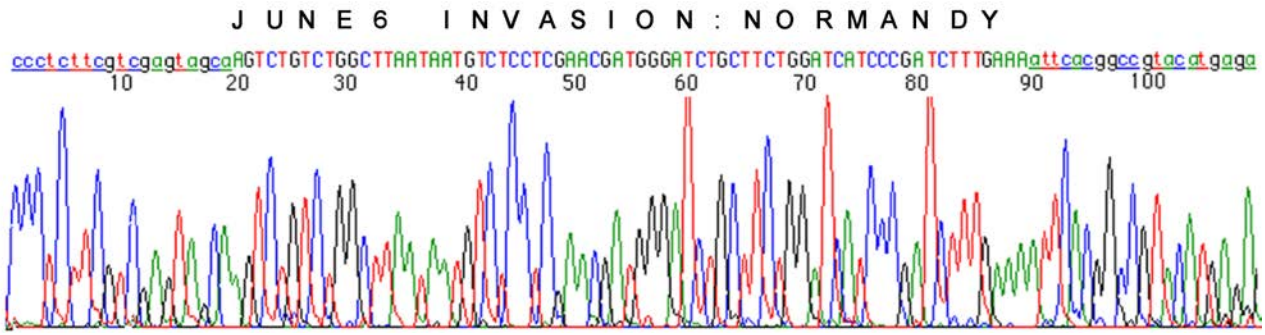

Рис. 7. Приклад визначення нуклеотидної послідовності ДНК [13].

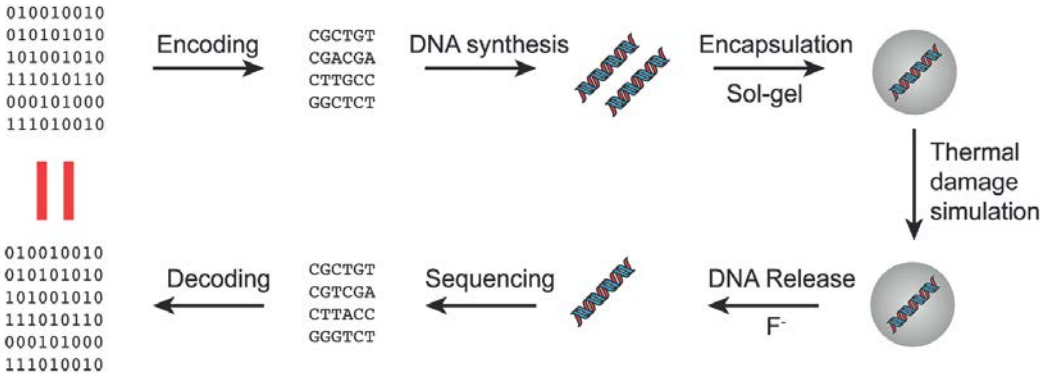

Рис. 8. Схема зберігання даних на молекулах ДНК.

Особливістю наведеної схеми є те, що для підвищення надійності зберігання даних фрагменти ДНК-ланцюжків капсулюються в силікатну оболонку [6].

Експериментальні результати зі створення пам'яті на ДНК. У першому пристрої ДНК-пам'яті, який було продемонстровано у 1988 році, вдалося записати 7920 біт даних. В таблиці 2 наведено дані про розроблені та випробувані системи ДНК-пам'яті.

Таблиця 2. Характеристика системи ДНК-пам'яті

Розробники системи пам'яті	Обсяг інформаційного блоку, кбайт	Метод кодування даних	Система відтворення даних	Джерело інформації
Гарвардська медична школа, факультет генетики Керівник рубрики G. Church	659	Бінарне	Oligonucleotide Platform: Agilent OLS	[7]
Європейська лабораторія молекулярної біології Керівник розробки N. Goldman	157	Троїсте	Oligonucleotide Platform: Agilent OLS	[8]
Швейцарський Федеральний технологічний інститут Керівник розробки R. Grass	83	Бінарне (використаний код Ріда–Соломона для корекції помилок)	Oligonucleotide Platform: CustomArray	[6]

Сучасний рівень створення ДНК-пам'яті характеризується досягненнями групи дослідників під керівництвом Джорджа Черча (G. Church) з факультету генетики Гарвардської медичної школи. Результатом їх досліджень став запис на масив одноланцюгових ДНК цілої книги у форматі HTML

об’ємом 53 426 слів, та, крім того, 11 зображень у форматі JPG и одну програму, написану мовою Java. Загальний обсяг даних, записаних за допомогою ДНК, становив 5,27 Мбайт. Для запису було використано **54 898 159 нуклеотидів, які були організовані в 115-бітні олігонуклеотидні блоки.** Єдина істотна відмінність між системами памяті, які наведені в таблиці, полягає в схемі кодування двійкового потоку в послідовність нуклеотидів: якщо у [6] використовували просту бінарну схему кодування то у [6] використовували більш складний алгоритм, перетворення бітового потоку у троїсний за допомогою алгоритму Хаффмана. Останнє дозволило стиснути дані і знизити ймовірність помилок, виключивши з ДНК-масиву гомополімерні ланцюжки. Ще одним способом підвищення стійкості до помилок було чотирикратне дублювання 117-бітових ланцюжків з регулярним зміщенням коду на 25 біт, притому кожен другий дубль кодувався в зворотній послідовності. При такій схемі ймовірність виникнення однакових помилок відразу в декількох ланцюжках стає мізерно малою [8]. Практичне використання розробленого методу зберігання даних на ДНК-молекулах суттєво обмежене громіздкістю використаного обладнання, значною тривалістю циклу запис / відтворення і, звичайно, вартістю. Вартість розшифровки ДНК щорічно падає приблизно в 5–12 разів – набагато швидше, ніж вартість цифрового електронно-оптичного мегабайта, так що у технології ДНК-пам'яті, безумовно, є велике майбутнє. Виходячи з нинішнього технологічного прогресу в галузі синтезу і секвенування, носії ДНК для запису інформації повинні з'явитися у відкритому продажу протягом десяти років. Хоча ДНК дозволяє зберігати інформацію тисячоліттями, перші комерційні носії будуть продаватися з гарантією до 50-ти років. На сьогодні вартість кодування інформації в ДНК оцінюється приблизно в $ 12 400 за мегабайт, вартість зчитування – $ 220 за мегабайт. Протягом десятиліття ціни повинні впасти на кілька порядків. Наприклад, синтез ланцюжка ДНК, що містить 100 мільйонів пар нуклеотидів в 2001 році коштував US $ 10000 і тільки 10 центів сьогодні. Вартість синтезу ДНК для цілей зберігання інформації і систем відтворення даних потребує зменшення на 6–8 порядків для широкого використання ДНК-пам’яті [9].

Висновки. 1. ДНК-пам'ять має значні потенційні можливості для створення систем довготермінового зберігання великих обсягів інформації (десятки і навіть сотні ексабайт даних).

2. ДНК-зберігання інформації, яке сприймалося як фантастика всього кілька років тому, завдяки створенню новітніх технологій секвенування ДНК наблизило до створення реальних систем пам’яті.

3. Основним типом систем ДНК-пам’яті в найближні роки будуть системи пам’яті типу WORM, тому створення новітніх швидкодіючих систем секвенування буде визначати прогрес в галузі створення ДНК-пам’яті для довготермінового зберігання даних.

Література

1. EMC: The digital universe of opportunities // *Infobrief*, 2014. – P. 1–17. Market report analyzing the global digital data landscape.

2. Landenmark H. K. E. An Estimate of the Total DNA in the Biosphere / H. K. E. Landenmark, D. H. Forgan, C. S. Cockell // PLoS One. Published: June 11, 2015 DOI 10.1371/journal.pbio.1002168

3. Оптические диски для долговременного хранения информации / В. В. Петров, В. М. Пузиков, А. А. Крючин, И. В. Горбов // Наносистеми, наноматеріали, нанотехнології Nanosystems, Nanomaterials, Nanotechnologies. − 2009. − Т. 7, № 3. – С. 825–832.

4. Analysis of properties of optical carriers after long-term storage / V. V. Petrov, A. A. Kryuchin, I. V. Gorbov [et al.] // Semiconductor Physics, Quantum Electronics and Optoelectronics. – 2009. – 12(4). – P. 399–402.

5. Method of aberration compensation in sapphire optical discs for the long term data storage / V. V. Petrov, V. P. Semynozhenko, V. M.Puzikov [et al.] // Functional Materials. – 2014. – Vol. 21, № 1. – P. 105–111.

6. Robust Chemical Preservation of Digital Information on DNA in Silicawith Error-Correcting Codes / R. N. Grass, R. Heckel, M. Puddu [et al.] // Angew. Chem. Int. Ed. – 2015. – № 54. – P. 1 – 5 DOI 10.1002/anie.201411378

7. Church G. M. Next-Generation Digital Information Storage in DNA / G. M. Church, Y. Gao, S. Kosuri // SCIENCE. – 2012. – Vol. 337. – P. 1627–1628.

8. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA / N. Goldman, P. Bertone, S. Chen [et al.] // Nature. – 2013. – Vol. 494. – P. 77–80 DOI:10.1038/nature11875

9. Castillo M. From Hard Drives to Flash Drives to DNA Drives / M. Castillo // Am. J. Neuroradiol Jan. – 2014. – Vol. 35, № 1. – P. 1–2.

10. Shah S., Limbachiya D. Manish K. DNACloud: A Tool for Storing Big Data on DNA arXiv:1310.6992v2[cs.Et]16 May 2014.

11. O' Driscoll A. Synthetic DNAThe next generation of big data storage / O' Driscoll, R. D. Sleator // Bioengineered. – 2013. – Vol. 4(3): – P. 123–125. DOI 10.4161/bioe.24296
12. Zakeri B. DNA nanotechnology: new adventures for an oldwarhorse / B. Zakeri, T. K. Lu // Current Opinion in Chemical Biology. – 2015. – № 28. – P. 9–14 DOI. org/10.1016/j.cbpa.2015.05.020
13. Clelland C. T. Hiding messages in DNA microdots / C. T. Clelland, V. Risca, C. Bancroft // Nature. 1999. – Vol. 399, – № 1033. – P. 533–534.
14. A mitochondrial genome sequence of a hominin from Sima de los Huesos / M. Meyer, Q. Fu, A. Aximu-Petri [et al.] // Nature. – 2014. – Vol. 505, № 7483. – P. 403–406. DOI 10.1038/nature12788
15. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments / J. Dabney, M. Knapp, I. Glocke [et al.] // Proc. Natl. Acad. Sci. USA. – 2013. – Vol. 110. – № 39. – P. 15758–15763.
16. Stabilizing synthetic data in the DNA of living organisms / Yachie N. Ohashi, Y. Tomita M. // Syst Synth Biol. – 2008. – № 2 (1–2). – P. 19–25. DOI 10.1007/s11693-008-9020-5.
17. Neiman M. S. On the molecular memory systems and the directed mutations // Radiotekhnika. – 1965. – № 6. – P. 1–8.
18. Adleman L. M. Molecular computation of solutions to combinatorial problems // Science. – 1994. – Vol. 266, № 5187. – P. 1021–1024. DOI 10.1126/science.7973651.
19. Creation of bacterial cell controlled by a chemically synthesized genome / Gibson D. G., Glass J. I., Lartigue C. [et al.] // Science. – 2010. – Vol. 329, № 5987. – P. 52–56.
20. Long-term storage of information in DNA / C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland // Science. – 2001. – Vol. 93, № 5536. – P. 1763–1765.
21. Shrivastava S. Data Storage in DNA / S. Shrivastava, R. B. Birla // International Journal of Electrical Energy. – 2014. – Vol. 2, № 2. – P. 119–124.