УДК 519.725

ВИКОРИСТАННЯ КОДІВ РІДА-СОЛОМОНА ДЛЯ РЕАЛІЗАЦІЇ ЗБЕРІГАННЯ ДАНИХ НА ДНК

М. А. Замкова[1, a]

[1] *Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Фізико-технічний інститут*

Анотація

У роботі розглянуто задачу кодування інформації для подальшого її запису на ДНК та, відповідно, декодування відновленої з ДНК послідовності даних. Для вирішення даної задачі запропоновано код на основі кодів Ріда-Соломона, що враховує особливості носія інформації.

Ключові слова: кодування інформації, ДНК, коди з надлишковістю, коди Ріда-Соломона

Вступ

В останній час спостерігається значний зріст кількості інформації у світі. Так, згідно доповіді компанії IDS до 2025 року об'єм всіх даних світу буде складати 163 зетабайти (ЗБ), що, наприклад, вдесятеро більше за об'єм інформації у 2016 [1]. Сучасні сховища даних не надають такого обсягу. Так, максимальна ємність SSD наразі скаладає 30,72 ТБ [2]. Час користування такими накопичувачами також залежить від того, як швидко пам'ять заповниться, а строк їх життя нараховує лише сотні років. Отже, нові сховища даних повинні мати велику ємність. Одним з таких рішень є ДНК. Теоретично, в 1 г ДНК можна записати 1 КБ інформації. До того ж, дані на ДНК можуть зберігатися до 2000 років без значного погіршення [3]. Отже, ДНК має значні переваги перед сучасними накопичувачами. Вже використано ДНК рослин [4] та штучно синтезованих молекул [5, 6].

При цьому постає задача кодування інформації та вибору алгоритму кодування/декодування. Так, у роботах попередніх досілдників використано кодування з одноразовим шифрблокнотом [6] та на основі кодів Ріда-Соломона [7].

У даній роботі запропоновано алгоритм кодування та декодування інформації на основі коду Ріда-Соломона, з параметрами, обраними на основі біологічних особливостей ДНК.

1. Постановка задачі

1.1. Біологічні передумови

Для запису інформації на ДНК звичайний бінарний файл необхідно перевести в послідовність нуклеотидів. Для зчитування процедура є оберненою.

Нуклеотидна послідовність ДНК визначається завдяки секвенаторам. Існує три типи секвенаторів: флуоресцентні, напівпровідникові та нанопорові [8]. Через відносно низьку вартість секвенаторів останнього типу будемо орієнтуватися на них. Точність таких секвенаторів становить 90 %, тобто похибка зчитування складає 10 %. Далі для надійності будемо вважати, що похибка становить 12%.

ДНК складається з чотирьох видів нуклеотидів, а саме Аденіну (А), Тиміну (Т), Цитозіну (Ц) і Гуаніну (Г). Кожні три утворють амінокислоту. Тобто, виходить $4^3 = 64$ варіантів амінокислот.

1.2. Вибір коду та його параметрів

Через значну похибку використовуємо коди з надлишковістю, а саме код Ріда-Соломона [9]. Параметри кода вибираємо наступні:

- Довжина послідовності нуклеотидів n:
$$n = 64 - 1 = 63.$$
- Кодова відстань $d = 2t + 1 = 17$,
де $t = 8$ – це кількість помилок, що виправляються кодом. Таке t беремо з урахуванням того, що через похибку
$$n_{wrong} = n * 12/100 = 7.56$$
символів послідовності зчитається неправильно.
- Кількість інформативних символів k:
$$k = n - d + 1 = 47.$$

2. Побудова коду

Код будуємо над полем $GF(64)$.
Незвідний поліном: $P(x) = x^6 + x + 1$.
Породжуючий поліном
$$g(x) = \prod_{i=1}^{16}(x - a^i),$$
де a – певний примітивний елемент поля.

Перемноживши, отримуємо $g(x)$ у вигляді поліному 16 степеня, коефіцієнти якого мають наступний вигляд: (див. табл. 1).

[a] mashazamk@gmail.com

Табл. 1. Коефіцієнти поліному $g(x)$

Степінь	**Коефіцієнт**
x^{16}	1
x^{15}	$a^4 + a^3 + a^2$
x^{14}	$a^4 + a^3 + a^2 + 1$
x^{13}	$a^5 + a^4 + a^2 + a$
x^{12}	$a^5 + a^4 + a + 1$
x^{11}	$a^3 + a^2 + 1$
x^{10}	$a^4 + a^3 + a^2$
x^9	$a^5 + 1$
x^8	$a^5 + a^4 + a^3$
x^7	$a^5 + a^3 + a^2 + 1$
x^6	$a^5 + a^3 + a^2 + a + 1$
x^5	$a^5 + a^4 + a^3 + a^2 + 1$
x^4	$a^5 + a$
x^3	a
x^2	$a^5 + a^4 + a^3 + a^2 + 1$
x^1	0
x^0	a^4

Далі нам необхідно отримати породжуючу матрицю G_{47x63}. Вона має вигляд:

$$G_{47x63} = \begin{bmatrix} 1 & 1 & 1 & ... & 1 \\ 1 & a & a^2 & ... & a^{62} \\ 1 & a^2 & a^4 & ... & a^{124} \\ ... & & & & \\ 1 & a^{46} & a^{92} & ... & a^{2852} \end{bmatrix},$$

де всі степені елемента a приведені за модулем 63 і виконані перетворення для кожного степеня згідно незвідному поліному $P(x)$.

Тепер можемо побудувати алгоритми кодування і декодування.

2.1. Алгоритм кодування

Нехай в нас є певне вхідне слово, задано у поліноміальному вигляді:

$$U = (x) = u_0 + u_1 x + ... + u_{46} x^{46},$$

коефіцієнти при степенях x якого відомі.

Кодове слово C знаходимо як

$$C = (u_0, u_1, ..., u_{46}) * G$$

У вигляді поліному:

$$C(x) = c_0 + c_1 x + ... + c_{62} x^{62},$$

де $c_i = U(a^i)$

2.2. Алгоритм декодування

Маємо кодове слово C:

$$C(x) = c_0 + c_1 x + ... + c_{62} x^{62},$$

коефіцієнти при степенях x якого відомі.

Вхідне слово:

$$U = (c_0, c_1, ..., c_{62}) * G^{-1}$$

У вигляді поліному:

$$U(x) = u_0 + u_1 x + ... + u_{46} x^{46},$$

де $u_i = C(a^{-i})$.

Висновки

У даній роботі було розглянуто проблематику зберігання даних на ДНК. Для кодування оптимальним варіантом є використання кодів з надлишковістю. Аналізуючи біологічні передумови, було вирішено використовувати код Ріда-Соломона, підібрані його параметри та побудовані необхідні для алгоритмів кодування і декодування об'єкти.

Перелік використаних джерел

1. The Digitization of the World: From Edge to Core / Reinsel D., Gantz J., Rydning J. An IDC White Paper — `https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf`
2. Samsung Electronics Begins Mass Production of Industry's Largest Capacity SSD – 30.72TB – for Next-Generation Enterprise Systems (2018) – `https://news.samsung.com/global/samsung-electronics-begins-mass-production-of-industrys-largest-capacity-ssd-30-72tb-for-next-generation-enterprise-systems`
3. DNA Data Storage – Setting the Data Density Record with DNA Fountain (2017) — `https://twistbioscience.com/company/blog/twistbiosiencednastoragefountain`
4. O'Hare R. A living library you can water: Plan to store data in the DNA of plants could see all of the world's archives secured in a box of SEEDS (2016) — `https://www.dailymail.co.uk/sciencetech/article-3406604/A-living-library-water-Plan-store-data-DNA-plants-world-s-archives-secured-box-SEEDS.html`
5. Weinberger M. Microsoft is buying 10 million molecules of custom DNA from a San Francisco startup (2016). — `https://www.businessinsider.com/microsoft-buys-dna-from-twist-bioscience-2016-4`
6. Demonstration of End-to-End Automation of DNA Data Storage / Takahashi C.N., Nguyen B.H., Strauss K., Ceze L. Scientific Reports. (2019) 9:4998. `https://doi.org/10.1038/s41598-019-41228-8`
7. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes // R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, W. J. Stark. Angew. Chem. Int. Ed. 2015, 54, 1–5.
8. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers / Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Swerdlow H.P., Gu Y // BMC Genomics. — 13(1): 341. — 2012.
9. Wicker S.B., Bhargava V.K. An Introduction to Reed-Solomon Codes // Reed-Solomon Codes and Their Applications. – New York: IEEE Press, 1994, pp. 1-15.