ДНК – КОМБИНАТОРНОЕ ОБЪЯСНЕНИЕ ЗАПИСИ ИНФОРМАЦИИ, ПРЕОДОЛЕНИЕ ФУНДАМЕНТАЛЬНОГО ПОРОГА СЖАТИЯ

Филатов О.В. Email: Filatov17155@scientifictext.ru

Филатов Олег Владимирович - инженер-программист,
ЗАО «Научно технический центр «Модуль», г. Москва

Аннотация: *в настоящее время происходит объединение знаний о ДНК, полученных разными науками. В статье, с позиций информатики (которая рассматривает ДНК как информационный носитель) и «Комбинаторики длинных последовательностей», приводится объяснение существования двух нитей ДНК как способа достижения большей информационной ёмкости. Анализируя возможные способы записи бинарной информации в ДНК, с позиции комбинаторики длинных последовательностей, было обнаружено, что природная бинарная запись ДНК информации на семнадцать процентов более плотная, чем та которую сейчас обеспечивают современные алгоритмы сжатия информации, то есть природа в ДНК преодолела фундаментальный порог сжатия информации, который присущ современной вычислительной технике. В статье раскрывается способ сверхплотной бинарной записи информации в ДНК с позиций информатики и комбинаторики длинных последовательностей.*
Ключевые слова: *мтДНК, ДНК, комбинаторика, КДП, цуга, составное событие.*

DNA IS A COMBINATORIAL EXPLANATION FOR RECORDING INFORMATION, CROSSING THE FUNDAMENTAL THRESHOLD OF COMPRESSION

Filatov O.V.

Filatov Oleg Vladimirovich - Software Engineer,
SCIENTIFIC AND TECHNICAL CENTER «МОДУЛЬ», MOSCOW

Abstract: *at present, the knowledge about DNA obtained by different sciences is being combined, in the article, from the standpoint of informatics (which considers DNA as an information carrier) and "Combinatorics of Long Sequences", an explanation of the existence of two DNA strands is given as a way to achieve a greater information capacity. Analyzing possible ways of recording binary information in DNA, from the perspective of combinatorics of long sequences, it was found that the natural binary recording of DNA information is seventeen percent denser than that which is now provided by modern information compression algorithms, that is, nature in DNA has overcome the fundamental threshold of information compression , which is inherent in modern computing. The article reveals a method for superdense binary recording of information in DNA from the standpoint of informatics and combinatorics of long sequences.*
Keywords: *mtDNA, DNA, combinatorics, CDP, train, compound event.*

УДК 51; 28.21.19; 34.23.00

Сокращения:
КДП – Комбинаторика длинных последовательностей;
Пос-ть – последовательность.

Введение

Две нити ДНК - как способ сверх сжатия бинарной информации.

На рис. 1 схематично представлены две нити (спирали) ДНК. На каждой нити отображены буквенные последовательности (пос-ти), из множества букв {A, C; G; T}. Для получения бинарной пос-ти, предмета нашего исследования, заменим каждую букву двоичным кодом: A→ «00»; C→ «01»; G→ «10»; T→ «11». Эта перекодировка показывает возможность исследования ДНК в бинарном пространстве методами «Комбинаторики длинных последовательностей» [5; 6].

Каждая буква на одной нити ДНК однозначно связана с буквой на другой нити ДНК, рис.1. Так как для информатики ДНК буквы – это элементы информации, то эта особенность хранения информации в ДНК (по парная, однозначная, связанность букв на разных нитях ДНК) стала причиной исследования по записи бинарной информацию подобным образом. Но по какому признаку надо делить единую информацию на две части? За такой признак была принята модель деления случайных пос-ей на составные события и цуги в «Комбинаторике длинных последовательностей» КДП [1 - 4]. КДП методы начинают работать и сжимать пос-ти (информацию) тогда, когда все современные алгоритмы сжатия становятся не эффективны, и перестают работать. Аналогом составных событий КДП в физике являются атомы, а в биологии - молекулы. Напомним, что такое составные события в КДП, которые обозначаются буквой *S* [1 - 4].

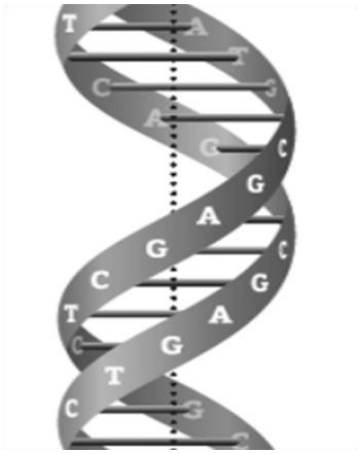

Рис. 1. Двойная спираль ДНК

Рассмотрим пос-ть $F1_b$: «11100111101010100000001111111», эта пос-ть не содержит форматированных кодов символов, такие пос-ти появляются при случайном подбрасывании монеты или как результат максимального сжатия данных. Разобьём $F1_b$ на одиннадцать составных событий: ($_1^3S$ = «111»); ($_2^2S$ = «00»); ($_3^4S$ = «1111»); ($_4^1S$ = «0»); ($_5^1S$ = «1»); ($_6^1S$ = «0»); ($_7^1S$ = «1»); ($_8^1S$ = «0»); ($_9^1S$ = «1»); ($_{10}^7S$ = «0000000»); ($_{11}^7S$ = «1111111»). В физике цепочки волн называют цугами, в КДП цепочки одинаковых составных событий назвали так же [1 - 4]. Шесть составных событий ^1S повторяются друг за другом (выпали цугой, образовали цугу): $^{S=1}C_{W=6}$=«010101»; два составных события ^7S составили цугу: $^{S=7}C_{W=2}$= «00000001111111». Все остальные составные события - это цуги C_1, из одного составного события. Запишем фрагмент $F1_b$ в цуговых символах $F1_C$: «11100111101010100000001111111» → $^3C_1$; $^2C_1$; $^4C_1$; $^1C_6$; $^7C_2$. В левом верхнем углу символа C указана длина составного события, в правый нижний символ обозначает число повторов этого составного события (длину цуги).

Единую бинарную пос-ть (информацию, файл) делим по подобию двух нитей ДНК на две связные пос-ти (файла). В одном файле собраны базовые длины составных событий цуг, а во втором содержаться связанные с этими составными события численности колен цуг. Составные события собираем в одном *S* - файле: 3; 2; 4; 1; 7, а число повторов в другом, *W*- файле: 1; 1; 1; 6; 2. Для восстановления пос-ти из пар чисел *SW*, цуговой записи $F1_C$: $^3C_1$; $^2C_1$; $^4C_1$; $^1C_6$; $^7C_2$ надо сохранить значение первого члена пос-ти (в файле 3), для $F1_b$ это значение равно: «1».

Современная процессорная техника имеет такой параметр как разрядность. Этим же фиксированным параметром обладают все цифровые данные (которые хранят информацию) Но, ДНК пос-ти не имеют фиксированной разрядной сетки, поэтому при записи в файлы составных событий и связанных с ними цуг, нужно исключить привязку записываемой информации к фиксированной разрядной сетке.

Основная часть

Рассмотрим механизм сжатия, который по подобию двух нитей ДНК делит информации на две части и преодолевает информационный предел «не сжимаемости даже на один». Результат - гарантированное сжатие минимум на 17 %. Так пос-ть из 10^9 бит, которую невозможно сжать «на один» бит, будет сжата, на: $0.17 \cdot 10^9 = 1.7 \cdot 10^8$ бит (много больше, чем на один бит).

Опишем метод сжатия каждого числа в любой паре чисел *SW* на нитях ДНК, рис.2. Так как метод сжатия один, для *S* и *W* , то будем обозначать любое сжимаемое число буквой *V*: $V \in \{S; W\}$, а любое сжатое число V буквой G. Бинарный вид числа G обозначим - $_R^bg$, где R - число бит. В таблице 1 представлен принцип геномного сжатия чисел: $V \rightarrow G$.

Строка 1 содержит десятичные числа *V*. Число ноль невозможно, так как не существует составных событий *S* нулевой длины, и цуг *W* из них.

Таблица 1. Геномное сжатие: $V \rightarrow G$ в R – разрядной позиционной системе

1	*V(S; W)*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	*G;* $_X^dg$	0	1	0	1	2	3	0	1	2	3	4	5	6	7
3	$_{R=1}^bg$	0	1												
4	$_{R=2}^bg$			00	01	10	11								
5	$_{R=3}^bg$							000	001	010	011	100	101	110	111
6	^bR	1		2				3							

7	$n(R)$	2^1	$6 = 2^1 + 2^2$	$14 = 2^1 + 2^2 + 2^3 = 2^{R+1} - 2$; ф.1.7;1.8
8	$\{{}^{d}_{R}g\} = f(2^R)$	2^1	2^2	2^3
В строке 1 числа принадлежат НЕ десятичной системе, а R – РАЗРЯДНОЙ системе. На одном разряде R – системы возможны две комбинации, которые отданы под числа: 1; 2. На 2-х R – разрядах возможны 4-е комбинации, которые отданы под числа: 3; 4; 5; 6.				

Применим таблицу 1 для сжатия данного выше бинарного фрагмента $F1_b$: «11100111101010100000001111111». Поделим пос-ть $F1_b$ на S - файл

составных событий и W - файл цуг. Для этого переведём $F1_b$ в цуговый вид $F1_C$: ${}^{S=3}C_{W=1}$; ${}^{S=2}C_{W=1}$; ${}^{S=4}C_{W=1}$; ${}^{S=1}C_{W=6}$; ${}^{S=7}C_{W=2}$ и сожмём все S числа: $S \rightarrow G$. Для этого, в таблице 1, в строке 1 ищем число V равное числу S. строке 2, под найденным числом S, дано сжатое число $G = f(S)$. Число S может быть записано в любой допустимой компьютерной разрядной сетке, а разрядная сетка R каждого сжатого числа G сама несёт информацию о числе G. Число разрядов R, каждого числа G, является информацией для перевода числа G из сжатого состояния в исходное число S.

В $F1_b$ первое составное событие $S(\text{«}111\text{»}) = 3$ оно сжимается в G = 0, строки 1; 2, таблицы 1: $(V(S;\ W) = 3) \rightarrow (G = 0)$.

В $F1_b$ второе составное событие $S(\text{«}00\text{»}) = 2$ оно сжимается в G = 1, строки 1; 2, таблицы 1: $(V(S;\ W) = 2) \rightarrow (G = 1)$.

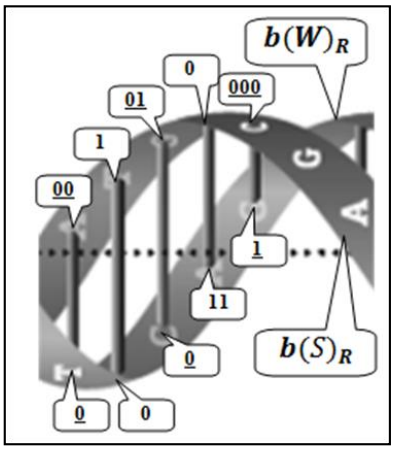

Рис. 2. Размещение $F1_b$ на двойной спирали ДНК

В $F1_b$ третье составное событие $S(\text{«}1111\text{»}) = 4$ оно сжимается в G = 1, стр. 1; 2, тбл.1: $(V(S;\ W) = 4) \rightarrow (G = 1)$.

Далее, в $F1_b$ идёт цуга ${}^{S=1}C_{W=6}$ из шести единичных составных событий ${}^1S$. Для каждого ${}^1S$: $(V(S;\ W) = 1) \rightarrow (G = 0)$.

Далее, в $F1_b$ идёт цуга ${}^{S=7}C_{W=2}$ из двух составных событий: ${}^7S$. Для каждого ${}^7S$: $(V(S;\ W) = 7) \rightarrow (G = 0)$.

Для пос-ти $F1_b$ составные события S_i: {3; 2; 4; 1; 1; 1; 1; 1; 1; 7; 7} переведём в множество сжатых бинарных чисел $\{{}^{b}_{R}g\}$, строки 3-5, таблицы 1: (3→«00»); (2→«1»); (4→«01»); (1→«0») – шесть раз; (7→«000»); (7→«000»).

Сжатые бинарные числа составных событий $\{{}^{b}_{R}g(S)\}$: {00; 1; 01; 0; 0; 0; 0; 0; 0; 000; 000}, они не демонстрируют эффекта сжатия информации. Для очевидного сжатия информации нужно получить ещё и цуговые бинарные числа: $b(W)_R$ из приводимой выше записи $F1_C$: ${}^3C_1$; ${}^2C_1$; ${}^4C_1$; ${}^1C_6$; ${}^7C_2$, выпишем числа цуг W из записи $F1_C$: 1; 1; 1; 6; 2. Теперь надо сжать в бинарный вид эти числа: 1; 1; 1; 6; 2 применяя таблицу 1: (1→«0»); (1→«0»); (1→«0»); (6→«11»); (2→«1»). Запишем искомое бинарное цуговое множество $\{{}^{b}_{R}g(W)\}$: {0;0;0;11;1}.

Перепишем цуговую запись $F1_b$, записав в угловых скобках полученные сжатые бинарные значения: ${}^{3\rightarrow\text{«}00\text{»}}C_{1\rightarrow\text{«}0\text{»}}$; ${}^{2\rightarrow\text{«}1\text{»}}C_{1\rightarrow\text{«}0\text{»}}$; ${}^{4\rightarrow\text{«}01\text{»}}C_{1\rightarrow\text{«}0\text{»}}$; ${}^{1\rightarrow\text{«}0\text{»}}C_{6\rightarrow\text{«}11\text{»}}$; ${}^{7\rightarrow\text{«}000\text{»}}C_{2\rightarrow\text{«}1\text{»}}$.

В файл составных событий $b(S)_R$ записываем пять верхних бинарных чисел из цуговой записи $F1_b$: $b(S)_R$ = {00; 1; 01; 0; 000}.

В файл цуг $b(W)_R$ записываем пять нижних бинарных чисел из цуговой записи $F1_b$: $b(W)_R$ = {0; 0; 0; 11; 1}.

Организуя связи между парами чисел: $b(S)_R \leftrightarrow b(W)_R$ мы получаем генетически сжатую информацию, рис.2.

В строке b, таблицы 2, дана исходная пос-ть $F1_b$. В строке b_{S+W} показан результат сжатия $F1_b$. В строке $b(S)_R$ дано содержимое файла содержащего информацию о составных событиях $F1_b$. В строке $b(W)_R$ дано содержимое файла содержащего информацию о цугах составных событий $F1_b$. Отобразим на рис. 2 возможную запись в двух нитях ДНК пос-ти $F1_b$.

Таблица 2. Не сжимаемые файлы сжаты, по типу данных в ДНК, на 17 %

b	1	1	1	0	0	1	1	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1
b_{S+W}	**0**	**0**	1	**0**	**1**	0	1	**0**	**0**	**0**	+	**0**	**0**	**0**	**1**	**1**	**1**												
$b(S)_R$	**00**			1		**01**				**0**						**000**													
$b(W)_R$	**0**			**0**		**0**				**11**						**1**													
А	N – исходных (бит)									5000000 бит						20 000 000 бит													
Б	Button252									эксперимент						эксперимент									теория				
В	Счётчик $b(S)_R$ (бит)									2278029						9113499									9114380				
Г	Счётчик $b(W)_R$ (бит)									1866674						7467543									7468248				
Д	Счётчик S + W (бит)									4144703						16581042									16582628				
Е	Сжатие: (S + W) / N									0,82894						0,8290521									0,8291314				
Ё	**Сжато ($N - S -$ W) на:**									**855297 бит**						**3418958 бит**									**3417372 бит**				
Суммарный размер в битах сжатых файлов - 83 % от исходного размера файла «не сжимаемого на один». Файл с информацией о составных событиях $b(S)_R$ всегда длиннее файла с информацией о цугах $b(W)_R$ [10].																													

В строке «А» даны длины (в битах) «не сжимаемых на один» пос-ей. В строке «Ё» показаны числа бит, на которые были сжаты «не сжимаемые на один» файлы. Файл с информацией о составных событиях $b(S)_R$ всегда длиннее файла с информацией о цугах $b(W)_R$ [10].

Формулы КДП для расчёта ДНК сжатия информации.

Для сжатия данных геномным способом нужно ввести особую систему счисления. Опишем основные свойства этой системы счисления. В таблице 1, строка 1, дан неразрывный ряд исходных чисел, который продолжен в таблице 3 в рядах: «${}^d_R V_{min} = 2^R - 1$» и «${}^d_R V_{max} = 2^{R+1} - 2$». Так на пересечении ряда «${}^d_R V_{min}$» и столбца 2, находится число три, а на пересечении ряда «${}^d_R V_{max}$» и столбца 2, находится число шесть - это минимальное и максимальное значения непрерывного диапазона: 3; 4; 5; 6.

Таблица 3. Распределение чисел ${}^d_R V$ по числу байтовых разрядов R

R - разряды байта	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
${}^d_R V_{min} = 2^R - 1$	1	3	7	15	31	63	127	255	511	1023	2047	4095	8191	16383	32767
${}^d_R V_{max} = 2^{R+1} - 2$	2	6	14	30	62	126	254	510	1022	2046	4094	8190	16382	32766	65534
$\Delta({}^d_R V) = 2^R$	2	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32768
Таблица 3 расширяет диапазоны значений V и □□ таблицы 1.															

Из таблицы 3 видно, что для обеспечения сжатия по геномному типу, исключено число ноль, и все целые десятичные числа ${}^d_R V$ разбиты на диапазоны, зависящие от двойки в степени R. В каждом R - диапазоне есть максимальное число ${}^d_R V_{max}$ ф.1.1 и минимальное число ${}^d_R V_{min}$, ф.1.2:

$$ {}^d_R V_{max} = 2^{R+1} - 2 \qquad \text{Ф.1.1} $$

$$ {}^d_R V_{min} = \frac{{}^d_R V_{max}}{2} = \frac{2^{R+1} - 2}{2} = 2^R - 1 \qquad \text{Ф.1.2} $$

Число чисел $\Delta({}^d_R V)$ в R - диапазоне два в степени R, ф.1.3, таблица 3:

$$ \Delta({}^d_R V) = {}^d_R V_{max} - {}^d_R V_{min} = 2^R \qquad \text{Ф.1.3} $$

Для десятичных чисел: ${}^d_R V > 2;$ ($R >$1) верно неравенство, ф1.4:

$$\sum_{r=1}^{R-1} 2^r < {}_R^dV \le \sum_{r=1}^{R} 2^r;\ \ {}_R^dV > 2 \qquad \text{Ф.1.4}$$

Замечаем, что: $\sum_{r=1}^{R-1} 2^r = 2^R - 2$, и: $\sum_{r=1}^{R} 2^r = 2^{R+1} - 2$, поэтому перепишем ф.1.4 в виде ф.1.5:

$$2^R - 2 < {}_R^dV \le 2^{R+1} - 2;\ \ {}_R^dV > 2 \qquad \text{Ф.1.5}$$

Пример на ф.1.5. При $R = 5$: ${}_R^dV > 2^5 - 2 = 30$; ${}_R^dV \le 2^6 - 2 = 62$. Действительно: 30 < (31; …; 62) ≤ 62, таблица 3.

Для: $0 < {}_R^dV \le 2$, ф.1.5 принимает вид ф.1.6, таблицы 2, 3:

$$0 < {}_R^dV \le 2^{R=1};\quad 0 < {}_R^dV \le 2 \qquad \text{Ф.1.6}$$

В строке 7, таблицы 1, показано, что количество $n(R)$ всех чисел ${}_R^dV$ вошедших в диапазоны: ${}^bR_1 = 1$: ${}^bR_2 = 2$; …; ${}^bR_R = R$ - равно сумме основания два, в степенях R_i, ф.1.7:

$$n(R) = \sum_{r=1}^{R} 2^r \qquad \text{Ф.1.7}$$

Сумма в ф.1.7 равна: $\sum_{r=1}^{R} 2^r = 2^{R+1} - 2$, поэтому запишем ф.1.8:

$$n(R) = \sum_{r=1}^{R} 2^r = 2^{R+1} - 2 \qquad \text{Ф.1.8}$$

В предложенной системе счисления величина десятичного значения ${}_R^dV$ зависит от числа двоичных разрядов R, по ф.1.5: ${}_R^dV = f(R)$ и наоборот, число двоичных разрядов R зависит от числа ${}_R^dV$: $R = f({}_R^dV)$.

Выяснив особенности связи десятичных чисел ${}_R^dV$ с числом двоичных разрядов R, перейдём к рассмотрению сжатых десятичных чисел ${}^dg$, где: g - сжатое десятичное число, d – символ десятичного формата.

Перевод десятичного целого числа ${}^dV$ в десятичное сжатое ${}^dg$ осуществим по ф.2.1, таблица 1:

$${}_X^dg = {}_R^dV - 1 - \sum_{r=1}^{R-1} 2^r;\quad V > 0;\ \ X \in \{S; W\} \qquad \text{Ф.2.1}$$

Где: r; R - число бинарных разрядов необходимое для бинарного представления ${}^dg$, иными словами: $R = f(V)$ - число бинарных разрядов диапазона, в котором находится число V; $S; W$ - уточняющие символы (S - расчёт длин составных событий, W - расчёт длин цуг).

Пример на ф.2.1. При V = 8, R = 3, поэтому: ${}^dg = 8 - 1 - \sum_{r=1}^{R=3} 2^r$ = 7 - (2 + 4) = 1, смотри таблицу 1.

Так как по ф.1.8: $\sum_{r=1}^{R} 2^r = 2^{R+1} - 2$, то заменим ф.2.1 на ф.2.2:

$${}_R^dg = {}_R^dV - 1 - (2^R - 2) = {}_R^dV - 2^R + 1 \qquad \text{Ф.2.2}$$

Где: R - число бинарных разрядов десятичного числа ${}_R^dV$, таблица 1.

Пример на ф.2.2. Переведём ${}_{R=3}^dV$ = 12 в сжатое целое ${}_{R=3}^dg$. Так как: $R = 3$, то подставив значения в ф.2.2 получим: ${}_{R=3}^dg = 12 - 2^3 + 1$ = 5.

Для любого R сжатые десятичные величины: ${}_R^dG_{min} = 0$; ${}_R^dG_{max}$ - ф.2.3:

$${}_R^dG_{max} = 2^R - 1 \qquad \text{Ф.2.3}$$

Множество сжатых чисел $\{{}_R^dg\}$ на R бинарных разрядах содержит $1 + 2^R - 1 = 2^R$ чисел: $\{{}_R^dg\} = \{{}_R^dG_{min}\} \cup \{1;\ \ldots; {}_R^dG_{max}\} = \{0; 1;\ \ldots; {}_R^dG_{max}\}$.

Перевод сжатых десятичных чисел ${}_R^dg$ в сжатые бинарные числа ${}_R^bg$.

При переводе десятичных чисел ${}_R^dg$ в двоичные числа ${}_R^bg$, разрядность двоичного числа ${}_R^bg$ равна R, даже если его можно записать при помощи меньшего числа двоичных разрядов, таблица 1.

Нахождение числа разрядов ${}^{b}R$ в сжатых бинарных словах ${}^{b}_{R}g$.

Для нахождения из не сжатой десятичной величины ${}^{d}V$ числа разрядов ${}^{b}R$, сжатого бинарного слова ${}^{b}_{R}g$, нужно найти целое число R, при котором верно неравенство: ${}^{d}_{R}V_{min} \le {}^{d}V \le {}^{d}_{R}V_{max}$, где ${}^{d}_{R}V_{min}$ - ф.1.2; ${}^{d}_{R}V_{max}$ - ф.1.1, таблицы 1, 3. Раскрывая: ${}^{d}_{R}V_{min}$ и ${}^{d}_{R}V_{max}$ получаем неравенство ф.2.4:

$$2^R - 1 \le {}^{d}V \le 2^{R+1} - 2 \qquad \text{Ф.2.4}$$

Где: R - число разрядов R, сжатого бинарного слова ${}^{b}_{R}g$.

Перевод сжатых бинарных слов ${}^{b}_{R}g$ в не сжатые десятичные числа ${}^{d}V$.

Перевод ${}^{b}_{R}g$ в ${}^{d}V$ производится по ф.2.5:

$${}^{d}V = 1 + {}^{b}_{R}g + \sum_{r=1}^{R-1} 2^r \qquad \text{Ф.2.5}$$

Так как: $\sum_{r=1}^{R-1} 2^r = 2^R - 2$, то запишем ф.2.5 в виде ф.2.6:

$${}^{d}V = {}^{d}_{R}g({}^{b}_{R}g) + 2^R - 1 \qquad \text{Ф.2.6}$$

Пример для ${}^{b}_{R}g$ = «011». В «011» три разряда: R = 3. Переводим «011» в десятичное: ${}^{b}_{R}g \rightarrow {}^{d}_{R}g$ = 3. Ставим ${}^{d}_{R}g$ в ф.2.6: ${}^{d}V$= 3 + 2^3 – 1 = 10, таблица 1.

Перевод десятичных чисел ${}^{d}V$ сжатые бинарные слова ${}^{b}_{R}g$.

Перепишем ф.2.6 относительно ${}^{b}_{R}g$, получим ф.2.7:

$${}^{b}_{R}g({}^{d}_{R}g) = {}^{d}V - 2^R + 1 \qquad \text{Ф.2.7}$$

Пример для ${}^{d}V$ = 10. Ищем по неравенству ф.2.4 число разрядов R в сжатом бинарном слове ${}^{b}_{R}g$. Для этого начнём поочерёдно подставлять числа R = 1; 2; 3; …, в ф.2.4, до получения требуемого результата в неравенстве. Ф.2.4 истинно при R = 3: $2^3 - 1 \le 10 \le 2^{3+1} - 2$; действительно: $7 <$ (${}^{d}V$ = 10) < 14. Ставим: ${}^{d}V$ =10 и $R = 3$ в ф.2.7: ${}^{d}_{R}g$ = 10 - 2^3 + 1 = 3. Переводим десятичное сжатое ${}^{d}_{R}g$ в бинарное сжатое ${}^{b}_{R}g$: 3 → «011», таблица 1.

Расчёт эффективности генетически подобного способа сжатия.

С помощью предложенного мной КДП - способа сжатия, сжимаемы пос-ти которые нельзя сжать даже на один бит известными методами. У пос-тей сжатых современными архиваторами и у случайных бинарных пос-тей одна и та же структура [7], рассчитаем размер их КДП досжатия.

Расчёт размера сжатого $L(W)$ – файла цуг.

При числе событий случайной бинарной пос-ти: N, число составных событий: $N/2$ [8], число цуг: $N/3$ [9]. Для расчёта длины W – файла, который содержит числа цуговых колен, их нужно сосчитать. Пример: в цуге «10101» - 5ть колен, в «110011» - 3и колена, в цугах: «1»; «00»; «000»; «0»; «11..11» - по одному колену, в «000111» - 2а колена. Из примера видно, что цуги с разными базовыми длинами составных событий имеют одинаковое число колен. Число цуг ${}^{n}_{N}C_w$ образованных составными событиями равной длины n, с числом колен w рассчитывается по ф.3.1 [9], из ф.3.1 получаем ф.3.2.

$${}^{n}_{N}C_w = \frac{(2^n - 1)^2}{2^{n(w+2)+1}} N \qquad \text{Ф.3.1}$$

Число цуг C_{wN} с числом колен w в пос-ти из N равновероятных бинарных событий рассчитываю по ф.3.2 (привожу без вывода):

$$C_{wN} = \sum_{n=1}^{\infty} C_w = \frac{N}{2} \cdot \left(\frac{1}{2^w - 1} - \frac{2}{2^{w+1} - 1} + \frac{1}{2^{w+2} - 1} \right) \qquad \text{Ф.3.2}$$

Где: n - длина составных событий в цугах; w - число колен цуг.

Примеры расчёта C_{wN} по ф.3.2, при $N = 2 \cdot 10^7$: w[1] = 4761905 (эксперимент: 4760325); w[2] = 1142857 (экс-т: 1143459); w[3] = 417819; w[4] = 180236. При суммировании цепочек всех колен C_{wN}, получим полное число цуг случайной бинарной пос-ти из N бит: $C_N = \sum_{w=1}^{w \to \infty} (C_{wN}) = \frac{N}{3}$.

Для расчёта $L(W)$ – длины, W – файла, необходимо найти число значимых бит, которое содержит каждое число цуговых полуволн: w, и умножить это число бит на каждую величину C_{wN} из ф.3.2. При

получении числа bit, для сжатия данных, будем вычитать единицу из каждого w: $bit(w-1)$, таблица 1. Теоретический расчёт длины $L(W)$ - файла производим по ф.3.3:

$$L(W) = \sum_{w=1}^{\infty} (bit(w-1) \cdot C_{wN}) \qquad \text{Ф.3.3}$$

При расчёте ф.3.3, для $N = 2 \cdot 10^7$, получено: $L(W)$ = 7468248 бит, таблица 2. Из таблиц 1, 3 и ф.2.4 видно, что: W = 1; 2 кодируется одним битом, W = 3; .. 6 - двумя битами, W = 7; .. 14 - тремя битами, и т.д.

Расчёт размера сжатого $L(S)$ – файла составных событий.

Для расчёта теоретической длины второго сжатого файла, ищу число цуг[1]: ${}^nC_{0N}$ - цепочек из составных событий равной длины по ф.3.4 [1-4]:

$${}^nC_{0N} = \frac{2^n - 1}{2^{2n+1}} N \qquad \text{Ф.3.4}$$

Где, n - длина составных событий ${}^nS$ цуги; N – число бит пос-ти.

Результаты расчёта ${}^nC_{0N}$ по ф.3.4, при $N = 2 \cdot 10^7$: n[1] = 2500000; n[2] = 4375000; n[3] = 6562500; n[4] = 7734375. Сумма всех ${}^nC_{0N}$ даёт полное число цуг C_{0N} пос-ти: $C_{0N} = \sum_{n=1}^{n\to\infty}({}^nC_{0N}) = \frac{N}{3}$.

Для расчёта $L(S)$ - длины S – файла необходимо найти число значимых бит, которое содержит каждое число n, которым обозначается длина составных событий ${}^nS$ в каждой цуговой цепочки ${}^nC_{0N}$, и умножить это число бит на каждую величину C_{0N} из ф.3.4. При получении числа bit, для сжатия данных, будем вычитать единицу из каждого n: $bit(n-1)$, таблица 1. Теоретический расчёт длины $L(S)$ - файла производим по ф.3.5:

$$L(S) = \sum_{n=1}^{\infty} (bit(n-1) \cdot {}^nC_{0N}) \qquad \text{Ф.3.5}$$

Необходимый формат записи информации для $L(W)$ и $L(S)$ файлов.

Сжатые $b(S)$ и $b(W)$ бинарные пос-ти, рис. 2, основаны на иной системе счисления (нет жёсткой разрядной сетки) и, поэтому, они не могут быть записаны в память современных вычислительных систем с получением эффекта экономии места. Компьютер выдаёт эффект сжатия только как расчётный результат при моделировании рассмотренного генетического сжатия. Для получения эффекта сверхсжатия надо перейти на другие принципы цифровой бинарной записи информации, в такой записи длина бинарного слова должна быть не фи+ксирована, как в молекулах ДНК.

Обсуждение

Оставаясь в рамках бинарного представления данных, но используя принципы природной упаковки информации в ДНК, можно сильно увеличить ёмкость носителей информации (микросхем памяти, цифровой памяти). Два основных приёма, которые применила природа для достижения большей ёмкости записи, это: использование переменной разрядности бинарных слов, и разделение записываемой бинарной информации на две нити в ДНК. Если взять «несжимаемую на один» пос-ть, то ни один архиваторов не может сжать её даже на один бит. Но, применив к ней, два приёма упаковки информации «подсмотренных» в ДНК, она будет сжата минимум на 17%.

Колмогоров предложил оценивать сложность бинарных пос-тей через их сжимаемость («сложность»). Речь идёт о сжатии информации без потери данных, когда из сжатого файла можно заново развернуть исходную пос-ть большей длины. Если бы не существовало фундаментального предела сжатия информации, то все пос-ти сжимались бы до длины в один бит и восстанавливались в исходную пос-ть. Фундаментальный предел сжатия информации (пос-ти) характеризуется тем, что структура сжатой пос-ти становится идентичной структуре случайной бинарной пос-ти [7]. По мнению Колмогорова, сложные пос-ти не сжимаемы даже «на один», а простые пос-ти сжимаемы. С удовольствием воспользуюсь этой общепризнанной отечественными математиками идеей для защиты практического применения предложенного мной способа генетического сжатия, основанного на «Комбинаторики длинных пос-тей», но не с позиции сжимаемости, а с точки зрения распознавания математиками (сокращённо: м-ми) индивидуальных пос-тей. То есть, м-ки не выставляют Колмогорову претензий по форме записи рассматриваемых пос-ей. М-ки не выставляют претензий по типу носителя пос-ти: доска, бумага, ЭЛТ, ЖК монитор, ПЗУ, ОЗУ, магнитный или оптический носитель, …; не выставляют претензий по способу

[1] Расчёт в: Graph2 \ «Размер номеров цуг C0 Btn268».

записи и химического состава пос-ти: мел, чернила, водные / масляные краски, магнитные или пространственные свойства материи, …; не выставляют претензий по способу разделения членов пос-ти друг от друга: пробел на доске /бумаге (не)установленной длины между членами, знак между членами, пропадание (появление) некоторых физических свойств материи между членами, измерение строго (не)фиксированного временного интервала, или изменение мощности излучения (лазера), смена фазы волны…

Рассмотрим пос-ть $F1_b$ в строке «b», таблицы 2, и используем слово «очевидно», для решения задач распознавания. М-ик смотря на $F1_b$ видит в ней первую, левую «1» и понимает, что это первый член, пос-ти $F1_b$. М-ик определяет последний член пос-ти $F1_b$ (то же «1») и что $F1_b$, содержит 29 бинарных членов. М-ку очевидно, как отличить один член пос-ти $F1_b$ от другого его члена, как отличить «0» от «1». Конечно, программа искусственного интеллекта способна то же сделать, но для описания заложенных в ней алгоритмов потребуется не один книжный том, поэтому применим по Колмогорову «рассмотрим пос-ть» и «очевидно» и потребуем от м-ков не придераться к распознаванию пос-тей в строках: b; $b(S)_R$; $b(W)_R$ таблицы 2.

Для м-ка, очевидно, сколько бинарных членов находится в ячейках строк: $b(S)_R$ и $b(W)_R$ таблицы 2. М-ку достаточно объяснить, что ячейки одного столбца логически связаны и представляют сжатое описание более длинной бинарной пос-ти и дать правила восстановления из сжатой пос-ти. И, самое главное то, что м-ик будет легко различать число бинарных членов в каждой ячейке строк: $b(S)_R$ и $b(W)_R$, и только эти бинарные члены (их упорядоченное множество) для м-ка несут информацию. Очевидно, что для м-ков допустимо любую «несжимаемую на один» пос-ть представить в виде строки b, таблицы 2, и осуществить сжатие информации приведённым выше способом, с записью в строки: $b(S)_R$ и $b(W)_R$. При подсчёте бинарных чисел в строках $b(S)_R$ и $b(W)_R$ их число окажется на 17% меньше числа бинарных цифр в несжимаемой на один пос-ти (строка b). Строки: $b(S)_R$, $b(W)_R$ таблицы 2, являются первой электронной (при просмотре на мониторе) или материальной (при просмотре на бумажном носителе) реализацией сжатия бинарного кода генетическим способом.

Рассмотрим рис.1, 2, на них видно, что природа использует связи между двумя нитями (спиралями) ДНК. Для связи друг с другом двух фрагментов информации природа использует молекулярные мосты (на рисунках эти связи обозначены прямыми перемычками). Я предполагаю, что за счёт таких связей между двумя нитями ДНК, природа достигает более плотной записи информации, записывая в более длинной нити $b(S)_R$ длины составных событий, а в более короткой нити $b(W)_R$ записывая число цуг w. Простые составные события не имеют определённого значения («0»; «1») и для восстановления информации, природа должна сохранить значение первого бита S - пос-ти, он задаст полярность первому восстанавливаемому составному событию [1 – 4]. На рис.3 условно показан фрагмент одной нити ДНК которая реализована на гипотетической нейронной памяти.

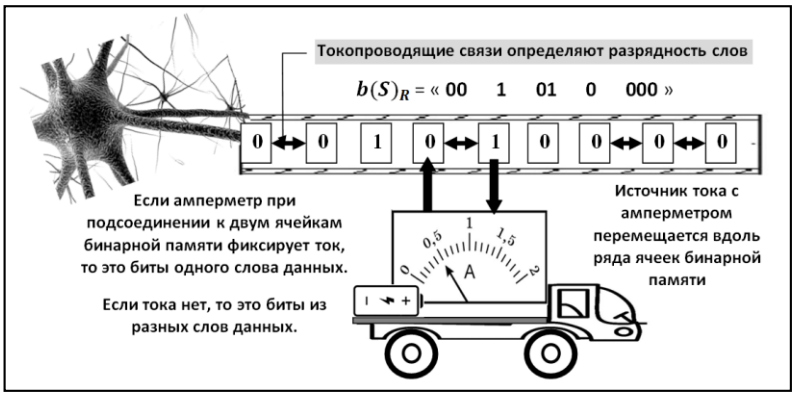

Рис. 3. Сжатие бинарной информации на основе нейронной сети

Геномное сжатие бинарной информации на основе нейронной сети возможно тогда, когда токопроводящие связи объединяют в слова переменной разрядности не только биты в каждой из нитей ДНК, но и соединяют попарно в единую электропроводную сеть связанные $b(S)_R \leftrightarrow b(W)_R$ участки двух нитей ДНК, рис.2. В электротехнике, для выявления электрически связанных участков цепи применяется «прозвонка» (которая является стандартным режимом в электрическом тестере). Такая же электрическая прозвонка, но перенесённая на молекулярный уровень, позволит определить все связанные биты информации в попарно связанных участках: $b(S)_R \leftrightarrow b(W)_R$, на нитях ДНК, рис. 2; 3. Предположим, что биологические ячейки памяти каждого слова данных, отращивают и соединяются последовательно друг с другом токопроводящими нитями, рис. 3, а пос-сть слов организуется лежащими на разных нитях парами токопроводящих участков. Эти токопроводящие участки разных нитей соединены друг с другом перемычками, рис. 1, 2.

Выводы

1) В статье приведена гипотеза, что в ДНК информация хранится в сжатом виде. Для большего сжатия природа использует разделение информации по двум нитям ДНК, связь между двумя фрагментами одной информации, хранящейся на двух нитях сразу, осуществляется одной перемычкой между фрагментами, при помощи которой каждая пара информационных фрагментов соединяется в отдельные, целостные данные.

2) Дан теоретический расчёт сжатия КДП, методом, преодолевающим современный порог «не сжимания на один», приведены результаты экспериментальных данных, которые хорошо совпали с теоретическим КДП расчётом.

3) Разработан КДП метод гарантированного сжатия на 17 % «не сжимаемых на один» последовательностей, что позволит увеличить ёмкость памяти перспективных цифровых устройств.

Список литературы / References

1. *Филатов О.В., Филатов И.О., Макеева Л.Л. и др.* «Потоковая теория: из сайта в книгу». Москва, «Век информации», 2014. С. 200.
2. *Филатов О.В., Филатов И.О.* «Закономерность в выпадении монет – закон потоковой последовательности». Германия, Издательский Дом: LAPLAMBERT Academic Publishing, 2015. С. 268.
3. *Филатов О.В., Филатов И.О.* Статья «О закономерностях структуры бинарной последовательности». «Журнал научных публикаций аспирантов и докторантов», 2014. № 5 (95). С. 226–233.
4. *Филатов О.В., Филатов И.О*. Статья «О закономерностях структуры бинарной последовательности (продолжение)». «Журнал научных публикаций аспирантов и докторантов», 2014. № 6 (96). С. 236–245.
5. *Филатов О.В.* Статья «Применение структур случайных последовательностей для описания свойств мтДНК и определения принадлежности отдельных мтДНК к их хозяйской группе животных», «Проблемы современной науки и образования». № 5 (150), 2020. С. 6-12.
6. *Филатов О.В.* Статья «ДНК комбинаторика, применение мтДНК матриц для расчёта родственных связей. Теорема о равенстве нулю корректирующей мтДНК матрицы», «Проблемы современной науки и образования». № 8 (153), 2020. С. 5-11, DOI: 10.24411/2304-2338-2020-10801.
7. *Филатов О.В.* Статья «Числовая оценка Колмогоровской сложности. Определение вероятности через смену событий», «Проблемы современной науки и образования». № 8 (38), 2015. С. 17-29, DOI: 10.20861/2304-2338-2015-38-001.
8. *Филатов О.В.* Статья «Теорема «О амплитудно-частотной характеристике идеальной бинарной случайной последовательности», «Проблемы современной науки и образования», 2015. № 1 (31) С. 5–11, DOI: 10.20861/2304-2338-2014-31-001.
9. *Филатов О.В.* Статья «Доказательство теоремы: «Формула для цуг из составных событий, образующих случайную бинарную последовательность», «Проблемы современной науки и образования», 2017. № 20 (102). С. 6 –12, DOI: 10.20861/2304-2338-2017-102-003.
10. *Филатов О.В., Филатов И.О.* Статья «Эффект Арнольда – Филатова. Золотое, серебряное сечения. Альтернативная запись бесконечно сложной последовательности. Аргументация по фундаментальности «Потоковой теории». «Журнал научных публикаций аспирантов и докторантов», 2014. № 12 (102). С. 124–130.