Особое мнение

THE REVERSE TRANSLATION FROM ENGLISH TO DNA: HOW TO ENCRYPT MESSAGES IN THE GENOME

P.Y. Andreev, *Student*
I.S. Ilyina, *Student*
Voronezh N.N. Burdenko State Medical University
(Russia, Voronezh)

DOI: 10.24411/2500-1000-2020-10832

Abstract. *The year 2010 was significant for biology in the light of complete artificial synthesis of genome and its transplantation, which was realized in the Craig Venter institute, Rockville, Maryland. A research group, led by Nobel laureate H. O. Smith, J.C. Venter and C.A. Hutchison, transplanted a de novo synthesized chromosome of Mycoplasma mycoides into Mycoplasma capricolum. In order to verify the success of the experiment, the donor genome was edited by inserting the nucleotide sequence of four watermarks into different loci of the genome, antibiotic resistance gene and lacZ gene. Thus, the efficiency of transplantation could be evaluated using three approaches:*

1. Control sequencing of the synthetic genome and identification of watermarks.

2. Testing viability of bacterial culture, treating their growth medium with antibiotics.

3. Turning the colony bright blue in the presence of an organic compound X-gal, metabolizing by a product of the lacZ gene.

These three approaches confirmed success in transplantation, proclaiming that the first ever species was created by human. The linguistic interest of this great scientific breakthrough lies in watermarks, which were inserted into the donor genome. Each additional sequence encodes a message in English. In this light, the transplantation of de novo synthesized chromosome into the living cells became the first ever precedent where the human language has been translated to the language of nucleic acids. The aim of this study is to biologically and linguistically analyze the adaptation of the English alphabet to a genetic code, which is the first ever trial to encode a formally non-biologic information into a living cell, mapping the loci of watermarks and deciphering encoded data.

Keywords: *synthetic biology, synthetic chromosome, molecular biology, molecular genetics, biotechnology, gene, genome, proteome, transcriptome, bioinformatics, linguistics, English language.*

The uprise of biosphere is believed to be launched with a relatively short prebiotic period [1]. This conception announces the very beginning of life in the form of molecular evolution of RNA, leading to all current life samples and species on the Earth, based on DNA genomes [2]. These early stages of evolution support the idea, that life is a chemical reaction [3]. Nevertheless, there is no sedimentary evidence of such aeonian molecular fossils, which could be identified both as the first genomic DNA and cells either as any direct signs of ancestral ribonucleic systems. The reason of such absence is instability of nucleic acids in major of milieu, making a big problem for paleogenenomics [4]. The idea of ribonucleic priority has come via two assumed features, which are critical for progressing ancient polymer systems. The first one is about storing information about itself and the second lies in its ability to replicate and implement all heredity features to offsprings [5]. Though these two abilities, which are believed to be unique only to the RNA, the question about the prime molecular precursor of cellular life kept unclear up to the 80s, when T. R. Cech et al. first ever revealed self-splicing RNA [6] and S. Altman et al. discovered catalytic features of this molecule [7]. In the light of such catalytic ac-

tivity some RNA molecules are called "ribozymes" [8], thus the ribonucleic machinery was seemed to be able to organize primitive replicating systems, which are believed to be an ancestor of life. This prebiotic period was christened "RNA-world" [9] and has an unidentified genesis time hitherto because the origin day of the first-ever polymerized RNA remains mysterious. Moreover, there are some other arguments in favor of RNA priority in molecular evolution, which are highly conservative and reflected in modern species. Among them, synthesis of ribonucleic primer before replication; TRNA as adapter molecules, involved in the transferring of aminoacids for peptide chain elongation; ATP [10], NAD [11], CoA [12], which are based on ribonucleosides etc. However, there are many other questions, which are of interest in molecular evolution. Taking into account that RNA is a rather complicated structure, which is composed of alternating monomers, there is nothing surprising, that there was the idea of an even earlier predecessor of the RNA. Some living fossils of the ancient RNA period in current life samples in the form of ribonucleoside-built molecules raise the question of possible echo of pre-RNA era. This epoch could be based on alternative, more primitive structures, which were able to store information and its aftersound in form of so-called peptide nucleic acid was recorded in Cyanobacteria [13]. As we know, all modern life on earth is based on DNA genomes and only viral genomes can be built out of DNA or RNA, including bacteriophages [14, 15] and viroids [16]. According to this data, one day the heredity function was first evolutionary delegated from pre-RNA to RNA and then from RNA to DNA, but there are no detailed insights of how these steps of molecular evolution were realized. At the same time, we know for sure, that the earliest biosignatures of cellular life are as old as 3,5 billion years [17] or even older – 3,77 billion – 4,28 billion years [18]. In this light, at least after 3,5 billion years of molecular evolution and macroevolution, FOXP2-mutative [19] primate appeared, which could use its own vocal sounds for creating comprehended words and in consequence was able to design a set of symbols, to transcript these words into information for other individuals and offsprings. This way, life required a long-term odyssey – from the very first genetic replicative system to the first cell, from newborn compartmentalized Eukaryote to the multicellularity until the first human stepped on Earth and in 286 thousand ± 32 thousand - 315 thousand ± 34 thousand years [20] to become the only one of planetary species, who could make a breakthrough of its greatest enigma possible: from the first ever DNA extraction by F. Miescher in 1869 to acceptation of heredity functions of this molecule; from investigations of molecular structure of nucleic acids and postulation of central dogma of molecular biology to revealing studies of genetic code features; from understanding the gene structure to the DNA sequencing epoch: from phages, viruses and bacteria to the Human Genome Project and finally to the era of synthetic biology. The aim of this study is to make a linguistic and biological overview of four watermarks in the first ever synthetic genome, revealing their exact location in the chromosome sequence and deciphering encoded data.

Materials and methods. Analysis of publications, including the Russian edition of C. Venter's "Life at the speed of light" in order to get the cryptography between the nucleotide code of watermarks in the synthetic chromosome.

1. Building the tagging sequences (TSs) for pairwise alignment in order to identify watermarks in the synthetic chromosome. TSs were created without punctuation marks, except commas and spaces as .txt files and converted into FASTA format by MEGA-X (Version: 10.1.7) tool.

2. Complete sequence of synthetic Mycoplasma mycoides JCVI-syn1.0 clone sMmYCp235-1 was found in NCBI Nucleotide database (GenBank: CP002027.1) and downloaded in FASTA. The sequence was divided into 5 fragments (F0-4), which correspond to the approximate loci of watermarks in the synthetic chromosome using DNASTAR Lasergene 17.01.1 (2020) MegAlign Pro.

3. PSA (Pairwise Sequence Alignment), completed by DNASTAR Lasergene 17.01.1 (2020) MegAlign Pro with MAUVE algorithm [Default seed weight (15) and seed fam-

ilies were used] in order to complete a pairwise alignment of created TSs and corresponding fragments of JCVI-syn1.0 chromosome, where we expected to find these watermarks. This alignment made it possible to identify each watermark.

4. An extra PSA was completed by DNASTAR Lasergene 17.01.1 (2020) MegAlign Pro, using the Smith-Waterman DNA alignment algorithm [Matrix: "NUC44"; Gap penalty: 10; Gap extension penalty: 1] in order to make sure that the first trial with MAUVE was successful.

5. Deciphering the watermarks codon by codon, moving to the 5'-end and 3'-end.

6. Overview of decoded watermarks.

Results. Analysis of the book "Life at the speed of light", authored by C. Venter. This book tells an exciting story about the long odyssey of the creation of the first ever synthetic genome and the first ever synthetic species – Synthetic Mycoplasma mycoides JCVI-syn1.0. Also it contains an invaluable data of numerous insights of four watermarks, which were inserted in JCVI-syn1.0 and the way to decode them. The chapter 8 "Synthesis of the M. mycoides Genome" has the key for deciphering the nucleotide cryptography, used in watermarks. A series of characters, including the English alphabet of 26 letters, newline, space, numbers (0-9), various punctuation marks, such as full stop, comma, hyphen etc, mathematical signs, such as plus, minus and some additional symbols were represented in the designed trinucleotide code. The analysis of "Life at the speed of light" made it possible to make a summary and building a table of cryptography between series of characters used to encode JCVI-syn1.0 watermarks and trinucleotides. It is well known that genetic information in species is encoded in DNA molecule using a combination of trinucleotides of 4 types: A (adenine), G (guanine), T (thymine) and C (cytosine), so that 43 give 64 possible combinations. There are 20 proteinogenic amino acids and 3 stop-codons, encoded by the genetic code. In this light one amino acid can be encoded by several trinucleotides in DNA, e.g. such amino acids as leucine, serine and arginine are encrypted by 6 different trinucleotide combinations (n=3; 3x6=18); glycine, alanine, valine, proline and threonine by 4 (n=5; 5x4=20); isoleucine by 3 (n=1; 1x3=3) and the most numerous group of amino acids are encrypted by 2 trinucleotides: phenylalanine, tyrosine, cysteine, histidine, glutamine, glutamic acid, asparagine, aspartic acid and lysine (n=9; 9x2=18). There are only two amino acids, that have a single DNA-codon: methionine and tryptophan (n=2; 2x1=2). Therefore, there are 61 DNA-codons for amino acids and the rest (ATT, ACT, ATC) are the stop-codons. It is easy to refer to each amino acid by an acronym of 3 letters or a single Latin letter: tryptophan is generally marked with the acronym "Trp" or Latin letter "W" to make the visualization of polypeptide like NH2-MKKTANKVVL…-COOH simpler (first 10 amino acids of M. laboratorium L-lactate dehydrogenase). In this light, the real genetic code is considerably restricted to make the idea of enciphering the English language in the genome possible due to the fact, that the protein language of 20 amino acids composes an alphabet of 20 characters: A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V with absence of 6 letters (B, J, O, U, X and Z), punctuation marks, mathematical signs and other symbols, which are essential to create a message with semantic load. The prevalence of the absent letters in WM1 could be demonstrated after its analysis. Letters B, J, O, U, X and Z were used 4, 5, 13, 4, 1 and 2 times respectively. This data shows the necessity to create an alternative cryptography between the language of nucleic acids and anthropogenic characters to enable codification of messages in the genome, including punctuation marks and other symbols. The next step after analyzing the book is to build a table of cryptography between JCVI-syn1.0, DNA and RNA codons and a real genetic code.

Table 1. Comparison between the language of watermarks in JCVI-syn1.0 and the language of amino acids

#	CHARACTER IN JCVI-syn1.0 cryptography	DNA	RNA	REAL GC
1	A	TAG	AUC	Ile, I
2	B	AGT	UCA	Ser, S
3	C	TTT	AAA	Lys, K
4	D	ATT	UAA	Stop-codon
5	E	TAA	AUU	Ile, I
6	F	GGC	CCG	Pro, P
7	G	TAC	AUG	Met, M
8	H	TCA	AGU	Ser, S
9	I	CTG	GAC	Asp, D (Aspartic acid)
10	J	GTT	CAA	Gln, Q (Glutamine)
11	K	GCA	CGU	Arg, R
12	L	AAC	UUG	Leu, L
13	M	CAA	GUU	Val, V
14	N	TGC	ACG	Thr, T
15	O	CGT	GCA	Ala, A
16	P	ACA	UGU	Cys, C
17	Q	TTA	AAU	Asn, N (Asparagine)
18	R	CTA	GAU	Asp, D (Aspartic acid)
19	S	GCT	CGA	Arg, R
20	T	TGA	ACU	Thr, T
21	U	TCC	AGG	Arg, R
22	V	TTG	AAC	Asn, N (Asparagine)
23	W	GTC	CAG	Gln, Q (Glutamine)
24	X	GGT	CCA	Pro, P
25	Y	CAT	GUA	Val, V
26	Z	TGG	ACC	Thr, T
27	Paragraph	GGG	CCC	Pro, P
28	Space	ATA	UAU	Tyr, Y
29	0	TCT	AGA	Arg, R
30	1	CTT	GAA	Glu, E (Glutamic acid)
31	2	ACT	UGA	Stop-codon
32	3	AAT	UUA	Leu, L
33	4	AGA	UCU	Ser, S
34	5	GCG	CGC	Arg, R
35	6	GCC	CGG	Arg, R
36	7	TAT	AUA	Ile, I
37	8	CGC	GCG	Ala, A
38	9	GTA	CAU	His, H
39	#	TTC	AAG	Lys, K
40	@	TCG	AGC	Ser, S
41	)	CCG	GGC	Gly, G
42	(	GAC	CUG	Leu, L
43	-	CCC	GGG	Gly, G
44	+	CCT	GGA	Gly, G
45	\	CTC	GAG	Glu, E (Glutamic acid)
46	=	CCA	GGU	Gly, G
47	/	CAC	GUG	Val, V
48	:	CAG	GUC	Val, V
49	<	CGG	GCC	Ala, A
50	;	TGT	ACA	Thr, T
51	>	AGC	UCG	Ser, S
52	$	ATC	UAG	Stop-codon
53	&	ACC	UGG	Trp, W
54	}	AAG	UUC	Phe, F
55	{	AAA	UUU	Phe, F

56	*	ATG	UAC	Tyr, Y
57	]	AGG	UCC	Ser, S
58	”	GGA	CCU	Pro, P
59	[	ACG	UGC	Cys, C
60	%	GAT	CUA	Leu, L
61	!	GAG	CUC	Leu, L
62	’	GAA	CUU	Leu, L
63	.	CGA	GCU	Ala, A
64	,	GTG	CAC	His, H

The cryptography between the genetic code and the English alphabet was achieved in JCVI-syn1.0 by engaging all 64 possible codons to encode 26 letters, figures from 0 to 9 and such symbols as comma, full stop, hyphen etc, as it is indicated in Table 1. There is a total mismatch between the alphabet and the one-letter code of the amino acid. The only exception is in position 12: the letter L of the JCVI-syn1.0 alphabet corresponds to leucine (Leu, L).

Building the tagging sequences for pairwise alignment. The identification of watermarks in the synthetic genome required building the sequences to complete pairwise alignment. Since the whole sequence of watermarks remained unpublished, 4 specific tagging sequences (TSs) were created in order to identify corresponding fragments of each watermark in the chromosome. The TSs were built according to the only available data from “The life at the speed of light”. Chapter 8, “Synthesis of M. mycoides genome”, fragmentally provides some insights of what has been encoded in each watermark. Thus, among encrypted data in WM1 there is “Craig Venter Institute”, the quote from James Joyce: “To live, to err, to fall, to triumph, to recreate life out of life” in WM2, the quote of Robert Oppenheimer: “See things not as they are, but as they might be” in WM3 and Richard Feynman’s quote: “What I cannot build, I cannot understand” in WM4. Considering these insights from chapter 8 we can hypothesize that “Craig Venter Institute” and the quotes above are suitable to be tagging sequences for PSA. Using the cryptographic features from Table 1 we can see that the TSs were created with no punctuation marks, except commas and spaces as .txt files and converted into the FASTA format by MEGA-X (Version: 10.1.7) tool.

Further preparations for PSA. A complete sequence of synthetic Mycoplasma mycoides JCVI-syn1.0 clone sMmYCp235-1 was found in the NCBI Nucleotide database (GenBank: CP002027.1) and downloaded in FASTA. The sequence was divided into 5 fragments, which correspond to the approximate loci of the watermarks in the chromosome, which are schematically represented in the original paper “Creation of a bacterial cell controlled by a chemically synthesized genome” [21]. The sequence was divided using DNASTAR Lasergene 17.01.1 (2020) MegAlign Pro into the 5 fragments: F0 (1→200000 nucleotides), F1 (200001→400000), F2 (400001→600000), F3 (600001→800000) and F4 (800001→1078809) as seen in Figure 1. This manipulation was completed in order to simplify and speed up the pairwise sequence alignment.

Figure 1. Fragmentation of the complete sequence of synthetic Mycoplasma mycoides JCVI-syn1.0 clone sMmYCp235-1.

Table 2. The tactics of pairwise sequence alignment: fragments of the chromosome and tagging sequences

Fragments and corresponding watermarks	**Number of nucleotides**	**Tagging sequence**	**English context of watermarks**
F0	200000	-	No WM in F0
F1 (WM3)	200000	TS3	See things not as they are, but as they might be
F2 (WM1)	200000	TS1	Craig Venter Institute
F3 (WM2)	200000	TS2	To live, to err, to fall, to triumph, to recreate life out of life
F4 (WM4)	278809	TS4	What I cannot build, I cannot understand

Pairwise sequence alignment. The MAUVE algorithm was used in order to identify watermarks in JCVI-syn1.0. This algorithm is suitable for aligning long sequences, including genomes [22]. The input data was submitted by the complete sequence of synthetic Mycoplasma mycoides JCVI-syn1.0 clone sMmYCp235-1 and fragmented sequence of each watermark (F1-F4). The first WM was identified via alignment of **F2** and **TS1**: 5'-TTTCTATAGCTGTACATATTGTAATGC TGATAACTAATACTGTGCGCTTGACTG TGATCCTGATAA-3', which corresponds to "Craig Venter Institute". WM2 was found via alignment of **F3** and **TS2**: 5'-TGACGTAACCTGTTGTAATGACGTTAA CTACTATGACGTGGCTAGAACAACTGA CGTTGACTACTGTCCCAAACATCATGA CGTCTATAATTTCTATAATAGTGATAA AACCTGGGCTAACGTTCCTGACGTGGC AACCTGGGCTAA-3'. WM3 was identified via **TS3**: 5'-GCTTAATAAATATGATCACTGTGCTAC GCTATATGCCGTTGAATATAGGCTATA TGATCATAACATATATAGCTATAAGTG ATAAGTTCCTGAATATAGGCTATATGA TCATAACATATACAACTGTACTCATGA ATAAGTTAA-3', which corresponds to "See things not as they are, but as they might be", - the quote of Robert Oppenheimer, who is often called "Father of the atomic bomb". The last watermark (WM4) was detected via **TS4**: 5'-GTCTCATAGTGAATACTGATATTTTAG TGCTGCCGTTGAATAAGTTCCCTGAAC ATTGTGATACTGATATTTTAGTGCTGC CGTTGAATATCCTGCATTTAACTAGCT TGATAGTGCATT-3', encoding "What I cannot build, I cannot understand" of Richard Feynman, an American physicist. After PSA was prepared with MAUVE a verification

check was completed via the Smith-Waterman algorithm, which is suitable for local alignment and to search for similar repertoires between two DNA sequences [23]. Each of the four watermarks was determined by their TS, as mentioned above.

Table 3. Identification of the watermarks via alignment of tagging sequences and fragments of JCVI-syn1.0. Control PSA of tagging sequences and the whole chromosome

Tagging sequence	Aligned locus in JCVI-syn1.0 fragment	Corresponding locus in the whole JCVI-syn1.0 (control PSA)
Craig Venter Institute (TS1)	F2: 165527→165592	565527→565592
To live, to err, to fall, to triumph, to recreate life out of life (TS2)	F3: 126295→126492	726295→726492
See things not as they are, but as they might be (TS3)	F1: 190264→190407	390264→390407
What I cannot build, I cannot understand (TS4)	F4: 159487→159606	959487→959606

PSA indicated that “Craig Venter Institute” is encoded in position 565527→565592 of the chromosome; “To live, to err, to fall, to triumph, to recreate life out of life” in 726295→726492; “See things not as they are, but as they might be” in 390264→390407; and “What I cannot build, I cannot understand” was found in 959487→959606.

Deciphering the watermarks. The next step of this study lies in decoding the full sequence of each watermark after they are identified in JCVI-syn1.0. The process of deciphering is to move up to the 5’ end and down to the 3’ end from alignment, e.g. after the WM4 was identified via “What I cannot build, I cannot understand” in position 959487→959606, deciphering started first from 959486 up to the 5’ end codon by codon: GGA (“), GGG (paragraph), CGA (.), TCA (H), TAG (A), TGC (N), TGC (N), TAG (A), ATA (SPACE), TGA (T), TAA (E), GCA (K), AAC (L), TAA (E), TAG (A), AGT (B) etc. until the meaningless text. Here is the decoding of Hanna Tekleab demonstrated, who participated in the project. After the 5’ end of this WM has been decoded, the 3’ end was completed, beginning from 959607 position. This strategy was used for deciphering other watermarks.

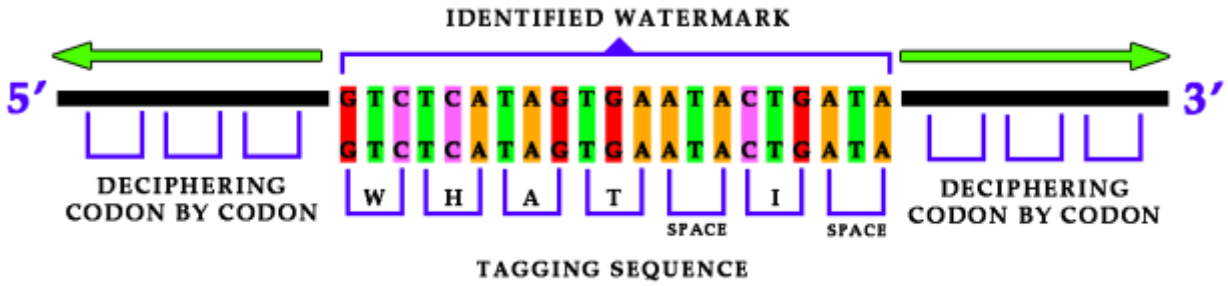

Figure 2. Deciphering nucleotide context codon by codon 3’→5’ and 5’→3’ after identification of watermarks via tagging sequences.

Watermark 1. Complete nucleotide sequence of the first watermark was revealed: 5’-

TTAACTAGCTAAGTTCGAATATTTCTA
TAGCTGTACATATTGTAATGCTGATAA
CTAATACTGTGCGCTTGACTGTGATCC
TGATAAATAACTTCTTCTGTAGGGTAG
AGTTTTATTTAAGGCTACTCACTGGTT
GCAAACCAATGCCGTACATTACTAGCT
TGATCCTTGGTCGGTCATTGGGGGATA
TCTCTTACTAATAGAGCGGCCTATCGC
GTATTCTCGCCGGACCCCCCTCTCCCA
CACCAGCGGTGTAGCATCACCAAGAA
AATGAGGGGAACGGATGAGGAACGAG
TGGGGGCTCATTGCTGATCATAATGAC
TGTTTATATACTAATGCCGTCAACTGT
TTGCTGTGATACTGTGCTTTCGAGGGC
GGGAGATTCGTTTTTGACATACATAAA
TATCATGACAAAACAGCCGGTCATGA
CAAAACAGCCGGTCATAATAGATTAG

CCGGTGACTGTGAAACTAAAGCTACTA
ATGCCGTCAATAAATATGATAATAGCA
ACGGCACTGACTGTGAAACTAAAGCC
GGCACTCATAATAGATTAGCCGGAGTC
GTATTCATAGCCGGTAGATATCACTAT
AAGGCCCAGGATCATGATGAACACAG
CACCACGTCGTCGTCCGAGTTTTTTTG
CTGCGACGTCTATACCACGGAAGCTGA
TCATAAATAGTTTTTTTGCTGCGGCAC
TAGAGCCGGACAAGCACACTACGTTT
GTAAATACATCGTTCCGAATTGTAAAT
AATTTAATTTCGTATTTAAATTATATG
ATCACTGGCTATAGTCTAGTGATAACT
ACAATAGCTAGCAATAAGTCATATATA
ACAATAGCTGAACCTGTGCTACATATC
CGCTATACGGTAGATATCACTATAAGG
CCCAGGACAATAGCTGAACTGACGTC
AGCAACTACGTTTAGCTTGACTGTGGT
CGGTTTTTTTGCTGCGACGTCTATACG
GAAGCTCATAACTATAAGAGCGGCAC
TAGAGCCGGCACACAAGCCGGCACAG
TCGTATTCATAGCCGGCACTCATGACA
AAACAGCGGCGCGCCTTAACTAGCT-
3’.

Decoded data:

Q2>EJ. CRAIG VENTER INSTITUTE 2009

[PARAGRAPH]

ABCDEFGHIJKLMNOPQRSTUVWXYZ

[PARAGRAPH]

0123456789#@)(-+\=/:<;>$&}{*]”[%!’.,

[PARAGRAPH]

SYNTHETIC GENOMICS, INC.

[PARAGRAPH]

<!DOCTYPE HTML><HTML><HEAD><TITLE>GENOME TEAM</TITLE></HEAD><BODY><A HREF=HTTP://WWW.JCVI.ORG/”>THE JCVI</A><P>PROVE YOU’VE DECODED THIS WATERMARK BY EMAILING US <A HREF=MAILTO:MROQSTIZ@JCVI.ORG” >HERE!</A></P></BODY></HTML>F5+ERS

Watermark 2. Complete nucleotide sequence of the second watermark was revealed:

5’-
TTAACTAGCTAACAACTGGCAGCATAA
AACATATAGAACTACCTGCTATAAGTG
ATACAACTGTTTTCATAGTAAAACATA
CAACGTTGCTGATAGTACTCCTAAGTG
ATAGCTTAGTGCGTTTAGCATATATTG
TAGGCTTCATAATAAGTGATATTTTAG
CTACGTAACTAAATAAACTAGCTATGA
CTGTACTCCTAAGTGATATTTTCATCCT
TTGCAATACAATAACTACTACATCAAT
AGTGCGTGATATGCCTGTGCTAGATAT
AGAACACATAACTACGTTTGCTGTTTT
CAGTGATATGCTAGTTTCATCTATAGA
TATAGGCTGCTTAGATTCCCTACTAGC
TATTTCTGTAGGTGATATACGTCCATT
GCATAAGTTAATGCATTTAACTAGCTG
TGATACTATAGCATCCCCATTCCTAGT
GCATATTTTCATCCTAGTGCTACGTGA
TATAATTGTACTAATGCCTGTAGATAA
TTTAATGCCTGGCTCGTTTGTAGGTGA
TAATTTAGTGCCTGTAAAACATATACC
TGAGTGCTCGTTGCGTGATAGTTCGTT
CATGCATATACAACTAGGCTGCTGTGA
TATGGTCACTGCCCTTACTGTGCTACA
TATTACTGCGAGGGGGATGACGTATA
AACCTGTTGTAAGTGATATGACGTATA
TAACTACTAGTGATATGACGTATAGGC
TAGAACAACGTGATATGACGTATATG
ACTACTGTCCCAAACATCAGTGATATG
ACGTATACTATAATTTCTATAATAGTG
ATAAATAAACCTGGGCTAAATACGTTC
CTGAATACGTGGCATAAACCTGGGCTA
ACGAGGAATACCCATAGTTTAGCAAT
AAGCTATAGTTCGTCATTTTTAAGGCG
CGCCTTAACTAGCT-3’. Decoded data: Q2>EMIKKEL ALGIRE, MICHAEL MONTAGUE, SANJAY VASHEE, CAROLE LARTIGUE, CHUCK MERRYMAN, NINA ALPEROVICH, NACYRA ASSAD-GARCIA, GWYN BENDERS, RAY-YUAN CHUANG, EVGENIA DENISOVA, DANIEL GIBSON, JOHN GLASS, ZHI-QING QI.

[PARAGRAPH]

“TO LIVE, TO ERR, TO FALL, TO TRIUMPH, TO RECREATE LIFE OUT OF LIFE.” – JAMES JOYCEF5+ERS

Watermark 3. Complete nucleotide sequence of the third watermark was revealed:

5’-
TTAACTAGCTAATTTAACCATATTTAA
ATATCATCCTGATTTTCACTGGCTCGTT
GCGTGATATAGATTCTACTGTAGTGCT
AGATAGTTCTGTACTAGGTGATACTAT
AGATTTCATAGATAGCACTACTGGCTT
CATGCTAGGCATCCCAATAGCTAGTGA
TAGTTTAGTGCATACAACGTCATGTGA

TACAACGTTGCTGGCTGTAGATACAAC
GTCGTATTCTGTAAGTGATACAATAGC
TATTGCTGTGCATAGGCCTATAGTGGC
TGTAACTAGTGATATCACGTAACAACC
ATATAAGTTAGATTTAATGCCCCTGAC
TGAACGCTCGTTGCGTGATAGTTTAGG
CTCGTTGCATACAACTGTGATTTTCAT
AAAACAACGTGATAATTTAGTGCTAG
ATAAGTTCCGCTTAGCAAGTGATAGTT
TCCGCTTGACTGTGCATAGTTCGTTCA
TGCGCTCGTTGCGTGATAAACTAGGCA
GCTTCACAACTGATAATTTAATTGCTG
ATATTGCTGGCTGTCTAGTGCTAGTGA
TCATAGTGCGTGATAGTTTAAGCTGCT
CTGTTTTAGATATCACGTGCTTGATAA
TGAAACTAACTAGTGATACTACGTAGT
TAACTATGAATAGGCCTACTGTAAATT
CAATAGTGCGTGATATTGAACTAGATT
CTGCAACTGCTAATATGCCGTGCTGCA
CGTTTGGTGATAGTTTAGCATGCTTCA
CTATAATAAATATGGTAGTTGTAACTA
CTGCGAATAGGGGGAGCTTAATAAAT
ATGATCACTGTGCTACGCTATATGCCG
TTGAATATAGGCTATATGATCATAACA
TATATAGCTATAAGTGATAAGTTCCTG
AATATAGGCTATATGATCATAACATAT
ACAACTGTACTCATGAATAAGTTAACG
AGGA-3’.

Decoded data: Q2>ECLYDE HUTCHISON, ADRIANE JIGA, RADHA KRISHNAKUMAR, JAN MOY, MONZIA MOODIE, MARVIN FRAZIER, HOLLY BADEN-TILSON, JASON MITCHELL, DANA BUSAM, JUSTIN JOHNSON, LAKSHMI DEVI VISWANATHAN, JESSICA HOSTETLER, ROBERT FRIEDMAN, VLADIMIR NOSKOV, JAYSHREE ZAVERI.

[PARAGRAPH]

“SEE THINGS NOT AS THEY ARE, BUT AS THEY MIGHT BE.”

Watermark 4. Complete nucleotide sequence of the fourth watermark was revealed:

5’-
TTAACTAGCTAATTTCATTGCTGATCA
CTGTAGATATAGTGCATTCTATAAGTC
GCTCCCACAGGCTAGTGCTGCGCACGT
TTTTCAGTGATATTATCCTAGTGCTAC
ATAACATCATAGTGCGTGATAAACCTG
ATACAATAGGTGATATCATAGCAACTG
AACTGACGTTGCATAGCTCAACTGTGA
TCAGTGATATAGATTCTGATACTATAG
CAACGTTGCGTGATATTTTCACTACTG
GCTTGACTGTAGTGCATATGATAGTAC
GTCTAACTAGCATAACTAGTGATAGTT
ATATTTCTATAGCTGTACATATTGTAA
TGCTGATAACTAGTGATATAATCCAAC
TAGATAGTCCTGAACTGATCCCTATGC
TAACTAGTGATAAACTAACTGATACAT
CGTTCCTGCTACGTGATAGCTTCACTG
AGTTCCATACATCGTCGTGCTTAAACA
TCAGTGATAACACTATAGAGTTCATAG
ATACTGCATTAACTAGTGATATGACTG
CAAATAGCTTGACGTTTTGCAGTCTAA
AACAACGTGATAATTCTGTAGTGCTAG
ATACTATAGATTTCCTGCTAAGTGATA
AGTCTACTGATTTACTAATGAATAGCT
TGGTTTTGGCATACACTGTGCGCTGCA
CTGGTGATAGCTTTTCGTTGATGAATA
ATTTCCCTAGCACTGTGCGTGATATGC
TAGATTCTGTAGATAGGCTAAATTCGT
CTACGTTTGTAGGTGATAGTTTAGTTG
CTGTAACTAATATTATCCCTGTGCCGT
TGCTAAGCTGTGATATCATAGTGCTGC
TAGATATGATAAGCAAACTAATAGAG
TCGAGGGGGAGTCTCATAGTGAATACT
GATATTTTAGTGCTGCCGTTGAATAAG
TTCCCTGAACATTGTGATACTGATATT
TTAGTGCTGCCGTTGAATATCCTGCAT
TTAACTAGCTTGATAGTGCATTCGAGG
AATACCCATACTACTGTTTTCATAGCT
AATTATAGGCTAACATTGCCAATAGTG
CGGCGCGCCTTAACTAGCT-3’.

Decoded data: Q2>ECYNTHIA ANDREWS-PFANNKOCH, QUANG PHAN, LI MA, HAMILTON SMITH, ADI RAMON, CHRISTIAN TAGWERKER, J CRAIG VENTER, EULA WILTURNER, LEI YOUNG, SHIBU YOOSEPH, PRABHA IYER, TIM STOCKWELL, DIANA RADUNE, BRIDGET SZCZYPINSKI, SCOTT DURKIN, NADIA FEDOROVA, JAVIER QUINONES, HANNA TEKLEAB.

[PARAGRAPH]

“WHAT I CANNOT BUILD, I CANNOT UNDERSTAND.” – RICHARD FEYNMANF5+ERS

Overview of the decoded watermarks. In order to provide verification procedures after the transplantation of a synthetic chromosome into M. capricolum cells, three approaches were applied to check the success of the experimental work by inserting additional sequences into the chromosome. Among them

are two genes: antibiotic (tetracycline) resistance gene (tetM), β-galactosidase (lacZ) gene and 4 fragments, containing encoded messages in the English language, so-called "watermarks". In this light, the efficiency of transplantation could be evaluated using three approaches:

1. Control sequencing of the synthetic genome and identification of watermarks.
2. Testing viability of bacterial culture, treating their growth medium with antibiotic.
3. Turning the colony bright blue in the presence of an organic compound X-gal, metabolizing by a product of lacZ gene.

There are 4 watermarks in the first ever synthetic genome with a total length of ≈3924 base pairs. The total length of each WM does not seem through and through evident. After these cryptographic elements were identified in the synthetic chromosome, the further strategy of decoding was to move codon by codon in the following directions: 3'→5' and 5'→3' from alignment until the beginning of a meaningless text in English. The features of this strategy can be illustrated, using the WM1 as an example. Moving from the "Craig Venter Institute" (5'-TTTCTATAGCTGTACATATTGTAATGC TGATAACTAATACTGTGCGCTTGACTG TGATCCTGATAA-3') in the 5'-end direction (3'→5') the following nucleotide context was identified: 5'-…CAT (Y) TCT (0) TAT (7) TTT (C) TGT (;) TAA (E) AAT (3) TTT (C) TTA (Q) ACT (2) AGC > TAA (E) GTT (J) CGA (.) ATA (space) TTT (C) CTA (R) TAG(A) CTG (I) TAC (G) ATA (space) TTG (V) TAA (E) TGC (N) TGA (T) TAA(E)CTA (R) ATA (space) CTG (I) TGC(N) GCT (S) TGA (T) CTG (I) TGA (T) TCC (U) TGA (T) TAA (E)…-3'. As shown above, the rest of the proximal characters are hard to interpret and they are most likely senseless. Contrarily, it was noticed that each watermark begins with 4 characters "Q2>E" (TTAACTAGCTAA) with no space between "Q2>E" and the context of WM. In addition, in WM1 and in WM2 there is TTT codon before the "Q2>E". At the same time WM1, WM2 and WM4 end with "F5+ERS" with different distal context (GGCGCGCCTTAACTAGCT), but WM3 ends with "E2>ETN2%9;!C". In this light, there are following borders in watermarks.

Table 4. Borders of the watermarks

Watermark	**Beginning, including "Q2>E"**	**End, including F5+ERS (WM1,2,4)**	**Length, bp**	**Number of encoded characters**
Watermark 1	565506	566573	1068	356
Watermark 2	725653	726558	909	303
Watermark 3	389493	390413*	921	307
Watermark 4	958641	959684	1044	348

**The last codon (CGA) of the quotation "See things not as they are, but as they might be." is considered as the 3'-end border of this watermark because of hardly interpreted text after the CGA trinucleotide, which differs from other watermarks.*

The first watermark is composed of 1068 base pairs and is the largest one. Encoded data: involved organizations and year of origin (Craig Venter institute, Synthetic Genomics, 2009); contains the key for deciphering (the series of used characters) and HTML-code for emailing after decoding the watermark.

The second watermark is composed of 909 base pairs (303 characters) and is the smallest one. It is localized in the region 725653→ 726558, encoding names of the scientists, which participated in the project (n=13) and the quote from James Joyce's "A Portrait of the Artist as a Young Man": "To live, to err, to fall, to triumph, to recreate life out of life".

The third watermark is composed of 921 base pairs (307 characters) and is localized in the region 389493 → 390413, encoding names of the scientists, which participated in the project (n=15) and Robert Oppenheimer's quote: "See things not as they are, but as they might be."

The fourth watermark is composed of 1044 base pairs (348 characters) and localized in the region 958641 → 959684, encoding names of the scientists, which participated in the project (n=18) and Richard Feynman's quote: "What I cannot build, I cannot understand."

Conclusion. This study shows the complete nucleotide sequence of watermarks and their loci in JCVI-syn1.0. The data encoded in these watermarks has been fully deciphered.

References

1. Добрецов Н.Л. О ранних стадиях зарождения и эволюции жизни // Вестник ВОГиС. – 2005. – Т. 9, № 1.

2. Joyce, G. The antiquity of RNA-based evolution // Nature. – 2002. – № 418, P. 214–221. doi:10.1038/418214a

3. Martin, William F. "Early evolution without a tree of life." Biology direct vol. 6 36. 30 Jun. 2011, doi:10.1186/1745-6150-6-36

4. Мовсесян А.А. Палеогеномика: достижения, проблемы, перспективы // Вестник Московского университета. Серия 23: Антропология. – 2010. – Т. 1. – С. 58-65.

5. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002.

6. Kruger K., Grabowski P.J., Zaug A.J., Sands J., Gottschling D.E., Cech T.R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Cell 31, 147–157. 10.1016/0092-8674(82)90414-7

7. Altman S, Baer MF, Bartkiewicz M, Gold H, Guerrier-Takada C, Kirsebom LA, et al. Catalysis by the RNA subunit of RNase P--a minireview. Gene. 1989;82:63–4. doi: 10.1016/0378-1119(89)90030-9.

8. Walter, Nils G, and David R Engelke. Ribozymes: catalytic RNAs that cut things, make things, and do odd and useful jobs // Biologist (London, England). Vol. 49, 5 (2002): 199-203.

9. Marie-Christine Maurel, Anne-Lise Haenni. The RNA world: hypotheses, facts and experimental results / M. Gargaud, B. Barbier, H. Martin, J. Reisse // Lectures in Astrobiology. Vol. 1, SpringerVerlag, pp. 571-594, 2005, copyright Springer-Verlag. ffhal-00008134f

10. Galimov, Erik M. Concept of sustained ordering and an ATP-related mechanism of life's origin // International journal of molecular sciences. – Vol. 10, 5 2019-30. 6 May. 2009, doi:10.3390/ijms10052019

11. Kirschning, A. Coenzymes and their role in the evolution of Life.Angew Chem Int Ed Engl. 2020 Jan 16. doi: 10.1002/anie.201914786.

12. Baddiley, J., Thain, E., Novelli, G. et al. Structure of Coenzyme A. Nature 171, 76 (1953) doi:10.1038/171076a0

13. Banack, Sandra Anne et al. Cyanobacteria produce N-(2-aminoethyl) glycine, a backbone for peptide nucleic acids which may have been the first genetic molecules for life on Earth // PloS one Vol. 7,11 (2012): e49043. doi: 10.1371/journal.pone.0049043

14. Ganaie, Safder S, and Jianming Qiu. Recent Advances in Replication and Infection of Human Parvovirus B19 // Frontiers in cellular and infection microbiology. Vol. 8, 166. 5 Jun. 2018, doi:10.3389/fcimb.2018.00166

15. Greninger, Alexander L, and Joseph L DeRisi. Draft Genome Sequences of Leviviridae RNA Phages EC and MB Recovered from San Francisco Wastewater // Genome announcements. Vol. 3, 3 e00652-15. 25 Jun. 2015, doi:10.1128/genomeA.00652-15.

16. Diener TO. (2001). The viroid: biological oddity or evolutionary fossil? Adv. Virus. Res. 57, 137-184.

17. Schopf JW. Fossil evidence of Archaean life. Philos Trans R Soc Lond B Biol Sci. 2006;361(1470):869–885. doi:10.1098/rstb.2006.1834

18. Dodd M.S., Papineau D., Grenne T., Slack J.F., Rittner M., Pirajno F., O'Neil J., and Little C.T.S. (2017) Evidence for early life in Earth's oldest hydrothermal vent precipitates // Nature. 543:60–64

19. Meredith L.J., Wang C.M., Nascimento L., Liu R., Wang L., Yang W.H. The Key Regulator for Language and Speech Development, FOXP2, is a Novel Substrate for SUMOylation. J Cell Biochem. 2016;117(2):426–438. doi:10.1002/jcb.25288

20. Richter, D., Grün, R., Joannes-Boyau, R. et al. The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age // Nature. 546, 293–296 (2017). https://doi.org/10.1038/nature22335

21. Gibson D. G., Glass J. I., Lartigue C., et al. Creation of a bacterial cell controlled by a chemically synthesized genome // Science. 2010;329:52–56. doi: 10.1126/science.1190719.

22. Darling, Aaron C E et al. Mauve: multiple alignment of conserved genomic sequence with rearrangements // Genome research. Vol. 14, 7 (2004): 1394-403. doi:10.1101/gr.2289704

23. Rucci, Enzo et al. SWIFOLD: Smith-Waterman implementation on FPGA with OpenCL for long DNA sequences // BMC systems biology. Vol. 12, Suppl 5 96. 20 Nov. 2018, doi:10.1186/s12918-018-0614-6

ОБРАТНАЯ ТРАНСЛЯЦИЯ С АНГЛИЙСКОГО ЯЗЫКА НА ДНК: КАК ЗАКОДИРОВАТЬ ПОСЛАНИЕ В ГЕНОМЕ

П.Ю. Андреев, *студент*
И.С. Ильина, *студент*
Воронежский государственный медицинский университет им. Н.Н. Бурденко (Россия, г. Воронеж)

Аннотация. *2010 год стал знаменательным для биологической науки в свете завершения проекта по полному химическому синтезу генома и его трансплантации, реализованного в Институте Крейга Вентера, Роквилл, штат Мэриленд. Исследовательская группа, возглавляемая нобелевским лауреатом Х. Смитом, К. Вентером и К. Хатчисоном успешно осуществила трансплантацию генома M. Mycoides, синтезированного de novo, в клетки Mycoplasma capricolum. Для того, чтобы подтвердить успех экспериментальной работы, в донорский геном были добавлены дополнительные репертуары, среди которых четыре т.н. водяных знака и два гена: TetM, продукт которого обеспечивает резистентность к антибиотику тетрациклину, а также lacZ, кодирующий фермент β-галактозидазу. Таким образом, верификация эффективности трансплантации могла быть осуществлена тремя способами:*

1. Контрольное полногеномное секвенирование – идентификация водяных знаков.

2. Добавление в культуральные среды тетрациклина – лизис клеток с интактным геном, не экспрессирующих tetM. Отсутствие бактериостатического влияния на клетки с синтетическим геномом.

3. Обработка культуральных сред органическим соединением X-gal – колонии окрашиваются в ярко синий цвет на фоне экспрессии lacZ, продукт которого расщепляет это соединение.

В конечном счёте, успешность экспериментальной работы была подтверждена, ознаменовав создание первого в истории синтетического организма, созданного человеком. Лингвистический интерес этого выдающегося прорыва состоит в водяных знаках, содержащихся в геноме синтетических бактерий, в которых зашифрованы послания на английском языке. Таким образом, трансплантация генома, синтезированного de novo, в живые клетки является первым в истории прецедентом, когда человеческий язык был переведён на язык нуклеиновых кислот. Целью настоящего исследования является биологический и лингвистический анализ адаптации английского алфавита к генетическому коду, что является первой в истории попыткой кодирования формально негенетической информации в живой клетке; картирование локусов водяных знаков в синтетической хромосоме и их полная расшифровка.

Ключевые слова: *синтетическая биология, синтетическая хромосома, молекулярная биология, молекулярная генетика, биотехнология, ген, геном, протеом, транскриптом, биоинформатика, лингвистика, английский язык.*