BioSystems 237 (2024) 105136

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

A novel approach to encode melodies in DNA

Olga Yu Kiryanova [a,*], Ravil R. Garafutdinov [b], Irek M. Gubaydullin [a], Alexey V. Chemeris [b]

[a] Institute of Petrochemistry and Catalysis, Ufa Federal Research Center, Russian Academy of Sciences, Prosp. Oktyabrya, 141, 450075, Ufa, Bashkortostan, Russian Federation

[b] Institute of Biochemistry and Genetics, Ufa Federal Research Center, Russian Academy of Sciences, Prosp. Oktyabrya, 71, 450054, Ufa, Bashkortostan, Russian Federation

ARTICLE INFO

Keywords:
DNA data storage
Non-biological information
Nucleotide sequences
Melody encoding
Oligotheca
NYRN-Oligonucleotides

ABSTRACT

DNA data storage has gained more attention last decades. DNA molecules can be used for encoding of non-biological information and as promising carriers due to greater data capacity, higher duration of the storage, and better technical failures stability. Here we propose a new method for encoding of notes and music in DNA. The encoding technique takes into account the duration and tonality of each note, enabling to encode all seven octaves by assigning a nucleotide sequence to each key. A certain set of short sequences is suggested to define the duration of note sound. The proposed method allows to encode more complicated melodies compared to the approach based on Huffman algorithm.

1. Introduction

To date, humanity has accumulated a huge amount of information (>20 zettabytes) that requires a new type of carriers for its long-term storage. It is predicted that by 2040 the amount of information will exceed three yottabytes, which will require the production of $>10^9$ kg of high-purity silicon (Zhirnov et al., 2016). However, most likely, only 10^7–10^8 kg of such silicon will be produced by 2040. This demand causes the search for new ways for data storage including using other data carriers, among which DNA molecules are of considerable interest. Natural DNAs are long polymeric molecules consisting of two complementary polynucleotide chains. One of the advantages of DNA molecules as a data carrier is the ability to store a lot of information in a minimal physical volume. It is considered that 1 g of DNA can keep of about 500 exabytes ($5*10^{20}$) of information, which is equivalent to 100 billion DVDs (Church et al., 2012). A huge number of nucleotide combinations (namely, 4^n, where n is the length of the sequence) makes it possible to encode all the current data accumulated.

Since the 1980s different methods for data encoding in nucleotide sequences have been proposed (Sakhabutdinova et al., 2019). However, the most optimal way is the data conversion into a binary code and, then, into nucleotides. Simple texts including hidden messages have been successfully encoded in DNA (Hodgson, 1990; Clelland et al., 1999). The ways for encoding of images and any graphic information were proposed as well. The first one was the rune translated into DNA sequence as a part of the Microvenus project (Davis, 1996). The binary code consists of 7 lines with 5 characters visually represents of the Microvenus image (See Fig. 1). For nucleotide encoding, nucleotides were ranked by size and assigned Arabic numerals - C = 1; T = 2; A = 3; G = 4. After transformation the incoding system looks like C = X; T = XX; A = XXX; G = XXXX, where X is the presence of the corresponding nucleotide depending on the number of actual rows of such X in the binary sequence. By using a digitizing phase-change code method the nucleotide sequence was synthesized. Later, as oligonucleotide synthesis improved, full texts, audio and video files, and computer programs were encoded in DNA (Goldman et al., 2013a; Blawat et al., 2016a; Erlich and Zielinski, 2017; Organick et al., 2018a; Antkowiak et al., 2020).

The storage of music as nucleotide sequences is of particular interest. It should be noted that mathematical principles are widely used for melodies creation (e.g., for choosing a particular rhythm or adding an accent to a specific note) (Hart; Yamuna et al., 2013). Some composers tried to translate various functions and numbers into melodies. However, it is impossible to determine the exact mathematical component of hearing a sound. Nonetheless, the regularity of a rhythm can be digitized. In a computer, melodies are converted into a binary format. An analog-to-digital converter (ADC) samples the sound wave at regular intervals and assigns a binary code to the wave amplitude. The obtained binary record is stored in a computer, and a digital-to-analog converter

* Corresponding author.
E-mail addresses: olga.kiryanova27@gmail.com (O.Y. Kiryanova), garafutdinovr@gmail.com (R.R. Garafutdinov), irekmars@mail.ru (I.M. Gubaydullin), chemeris@anrb.ru (A.V. Chemeris).

https://doi.org/10.1016/j.biosystems.2024.105136
Received 18 July 2023; Received in revised form 17 November 2023; Accepted 2 February 2024
Available online 3 February 2024

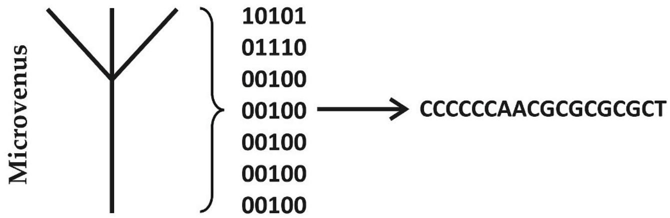

Fig. 1. The encoded rune (Microvenus project) (Davis, 1996).

(DAC) can convert it back to a sound wave. It is important how many bits are used to store a sound for better encoding quality. For 8 bits, the maximum number is available (11111111 = 255). Thus, 256 binary values from 0 to 255 can be assigned to the amplitude (Binary Code and Storing Music).

In 2013 the encoding of five files was reported upon: all 154 sonnets of W. Shakespeare (ASCII format), Watson-Crick's article about DNA structure (Watson and Crick, 1953) (PDF format), 26-s fragment of Martin Luther King's speech « I have a dream» (MP3 format), color photo of the European Bioinformatics Institute (JPG format), and description of the Huffman algorithm (ASCII format) (Goldman et al., 2013b). To encode this information, 153335 oligonucleotides with the length of 117 nt each, were synthesized. The 22 MB video was stored in DNA as well (Blawat et al., 2016b). For this, 900000 oligonucleotides consisting of two parts (190 nt encoding part and two 20 nt flanks) were obtained using the microarray method. Two-bit encoding was used as follows: 00 corresponds to A, 01 – to C, 10 – to G, and 11 – to T. Considerable attention was paid to errors elimination during oligonucleotides synthesis and DNA sequencing. More than 200 MB of information in 35 files of different types (TXT, PDF, MP3, MP4, and JPG), was stored in DNA (Organick et al., 2018b). In 2017 Twist Bioscience and Microsoft Corporation reported on storage of two audio records (songs "Smoke on the Water" by Deep Purple and "Tutu" by Miles Davis) followed by its retrieval with 100% accuracy (DNA-Based Digital Storage). These DNA-saved files were added to UNESCO's Memory of the World Archive (Service, 2017).

2. Keyboards of modern keyboard instruments. Some definitions

Modern pianos contain 88 keys grouped by octaves (Fig. 2). Other keyboard musical instruments, e.g., electronic synthesizers, may have more or fewer keys.

White keys represent the basic notes. Black keys indicate semitones (major and minor). A major tone raises the sound of a note by half of a tone, a minor tone lowers it, respectively. Duration of a note varies from the whole note to its parts, as it is shown on Fig. 3. It is possible to divide note duration infinitely. However, the substantial division of notes is not common, since melodies with virtuoso speed of sound are rare, and few musicians are able to play at a high speed.

All melodies are recorded on the musical staff, consisting of five main lines. The notes locate both on the lines and between them. The notes may have certain notations showing the tempo and the volume of the melody parts (Fig. 4).

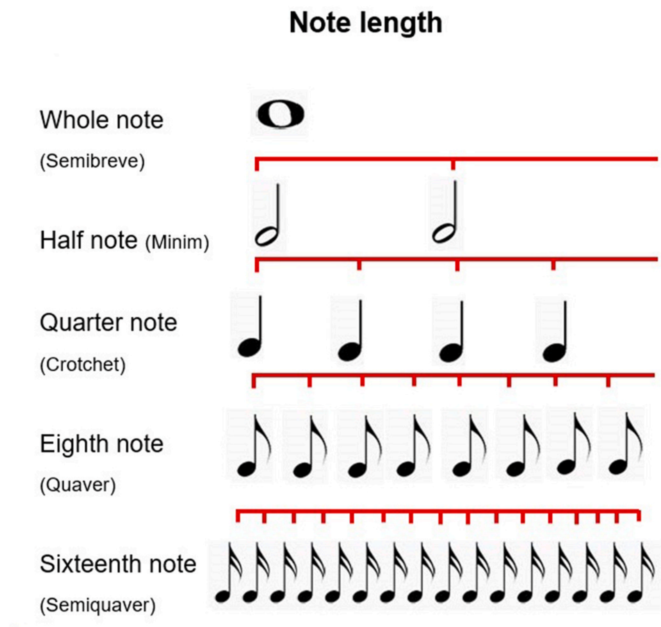

Fig. 3. Duration of notes.

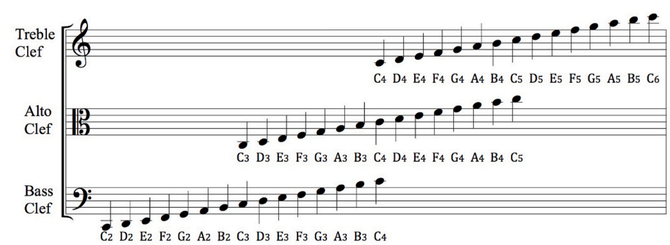

Fig. 4. The arrangement of the notes on the stave and clefs.

3. Huffman encoding

In 2009 an approach for melody conversion into a nucleotide sequence based on the modified Huffman algorithm was proposed (Ailenberg and Rotstein, 2009). According to the approach, the number of dG and dC nucleotides was reduced to decrease the formation of undesirable secondary structures. It was suggested to use three-column Huffmans coding to denote a larger number of notes. Table 1 indicates the designation of the notes for one octave. The principle was demonstrated on the children's song "Mary Had a Little Lamb" encoded in one DNA molecule. A 844 bp DNA fragment was obtained and cloned in a plasmid vector. In order to recover the melody two amplification products with the sizes 500 and 344 bp were obtained, followed by amplicons sequencing. The melody of the song "Mary Had a Little Lamb" is shown on Fig. 5. Encoding of the song part in nucleotide sequences is shown in Table 2.

The abovementioned method is quite convenient and provides a compact record, since each note corresponds to a short nucleotide

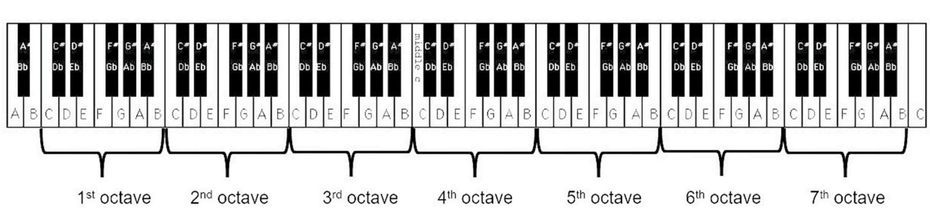

Fig. 2. Octaves of piano keyboard.

Table 1
Encoding of notes and musical symbols with nucleotides according to the Huffman algorithm.

Music code	DNA codon
Quarter note (1/4)	G
Half note (1/2)	TT
Whole note (1)	TA
Eighth note (1/8)	AT
Sixteenth note (1/16)	CT
Dot (.)	TC
A	TG
B	AC
C	AG
D	CG
E	AAT
F	AAC
G	AAG
2/4 (meter)	CAT
3/4 (meter)	CAA
4/4 (meter)	CAC
((repeat)	CAG
((repeat)	CCA
X (repeat)	CCT
2 (repeat)	CCG
3 (repeat)	CCC
4 (repeat)	AAAT
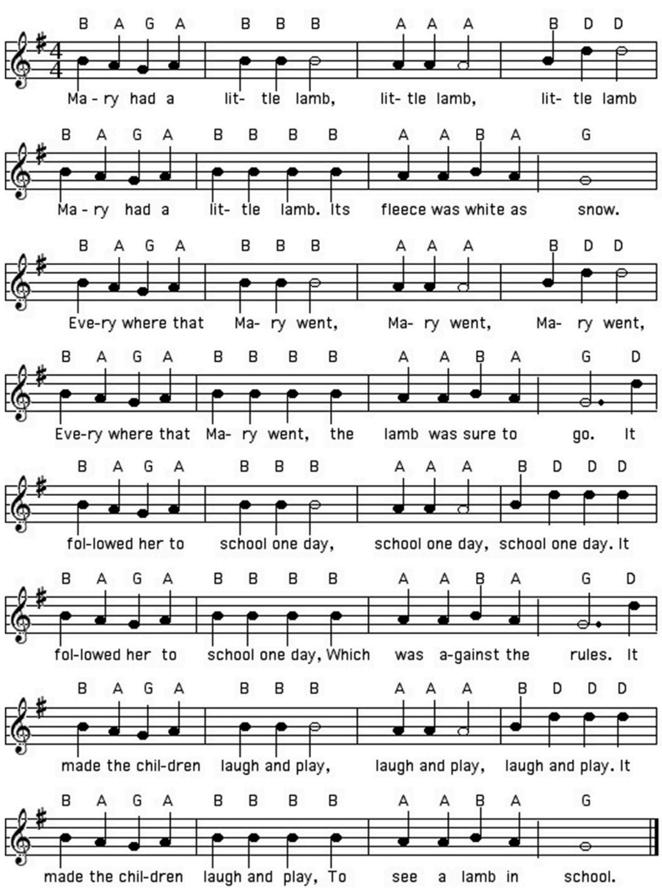
Fig. 5. Notes of the song "Mary Had a Little Lamb" (Smith et al., 2003).

sequence. However, for more complicated melodies, it is necessary to convert a larger range of notes into nucleotide sequences. Also, the method does not take into account the halftones and note combinations.

4. The new encoding method

We propose a new method for encoding of melodies, considering duration and the tone of the notes. Each note is encoded by 8 nucleotides. The first three nucleotides determine the duration of the notes and consist of G and C as follows:

ACG – whole note.

AGC – Half note.

ACC – Quarter note.

AGG – Eighth note.

CAG – Sixteenth note.

The next five nucleotides encode the tone of the notes. In this case all keys including semitones can be encoded. We have accepted the number 88 as the total number of keys. For a smaller number of octaves the corresponding keys will not be used for encoding. Sequences were selected to retain 50% GC composition in the resulting octanucleotides. All possible pentanucleotide sequences were generated using Python 3.6 (Scripts for encoding). We have selected those ones containing two C/G nucleotides and no homopolymer fragments or repeats to exclude self-complementarity (Table 3).

The proposed approach is consistent with a method for symbol encoding described earlier (Smith et al., 2003). However, it is convenient to use DNA from oligotheca representing the set of NYRN-oligonucleotides with 8 nt coding part (Garafutdinov et al., 2022). In this case each NYRN-oligonucleotide contains 1 byte of information, wherein nucleotide C is encoded as "00", T - "11", A and G (purines, R) - "0" and "1", respectively (so-called R/CT encoding). This principle can be used to translate a nucleotide sequence that stores information about a melody into a binary format. Decoding such a binary record a single "0" and "1" occurring within one byte are converted into purines A and G. Combinations "00" and "11" are converted into pyrimidines C and T respectively. The first binary symbol and the following one must be analyzed to follow the abovementioned schema. If symbols are the same they are treated as a pair and correspond to C for "00" and T for "11". If symbols differ the first one is considered as an independent and is taken as A ("0") or G ("1"). Then processing of a binary record continues. This approach slightly reduces the coding capacity and eliminates the problems associated with the presence of long homopolymer motifs. For instance, the encoding method proposed for Microvenus project (Davis, 1996) does not take into account biological and chemical properties of polynucleotide sequences. Hence, the capacity for R/CT encoding is 1.5 bits per nucleotide.

Service homopolymer sequences have been used to encode the start (AAATA) and the end (TTATT) of a chord. These homopolymer motifs differ from the nucleotide sequences that encode notes. Although homopolymer motifs increase the length of the coding sequence, they make it possible to encode all note combinations. It is proposed to use the AATAA pentanucleotide to encode a pause when none of the keys sounds. The pause also has a duration which is similar to the whole note duration.

To demonstrate the applicability of the approach, the well-known "Flohwalzer" (Flohwalzer) (dog's waltz) was converted into DNA sequence. The most popular and the simplest version of the dog's waltz is shown on Fig. 6.

"Flohwalzer" is played with the keys from 28 to 56. The melody can be represented as shown in Table 4. The first column shows the numbers of the keys (the keys played together are in brackets), and the second one shows the duration of each note.

"Flohwalzer" converted into nucleotide sequence is presented in Table 5.

The entire nucleotide sequence is as follows:
AGGAAACGAGGAACGAACCAGCAAAAATAACCGAACAACCCAA-

Table 2
Conversion of the part of the song "Mary Had a Little Lamb" into nucleotides according to Huffman algorithm.

					Ma-	ry	had	a	lit-	tle	lamb	lit-	tle	lamb	lit-	tle
m taac	u tact	* tacca	4/4 cac	(Cag	B1/4 ACG	A1/4 TGG	G1/4 AAGG	A1/4 TGG	B1/4 ACG	B1/4 ACG	B1/2 ACTT	A1/4 TGG	A1/4 TGG	A1/2 TGTT	B1/4 ACG	D1/4 CGG
lamb	**Ma-**	**ry**	**had**	**A**	**lit-**	**tle**	**lamb**	**its**	**fleece**	**was**	**white**	**as**	**snow**			
D1/2 CGTT	B1/4 ACG	A1/4 TGG	G1/4 AAGG	A1/4 TGG	B1/4 ACG	B1/4 ACG	B1/4 ACG	A1/4 ACG	A1/4 TGG	B1/4 TGG	A1/4 ACG	G1 TGG	) AAGTA	× cca	2 cct	ccg

Table 3
Encoding of keys using pentanucleotide sequences.

Key number		Nucleotide sequence	Key number		Nucleotide sequence	Key number		Nucleotide sequence	Key number		Nucleotide sequence
1	A	AACCA	23	G	GAAAG	45	F	CAGAA	67	D#	CCAAT
2	A#	ACCAA	24	G#	AGAGA	46	F#	CAAGA	68	E	CACAT
3	B	AAACC	25	A	TAGGA	47	G	CAAAG	69	F	AACCT
4	C	CCAAA	26	A#	TGGAA	48	G#	ACAGA	70	F#	ACACT
5	C#	CACAA	27	B	TAAGG	49	A	TAGCA	71	G	ACCAT
6	D	CAACA	28	C	GGTAA	50	A#	TACGA	72	G#	ACCTA
7	D#	CAAAC	29	C#	GAGTA	51	B	TAAGC	73	A	ACAGT
8	E	ACACA	30	D	GTAGT	52	C	GCATA	74	A#	ACGAT
9	F	TACCA	31	D#	TGAGA	53	C#	GACTA	75	B	CGAAT
10	F#	TCCAA	32	E	TGAAG	54	D	GAACT	76	C	CAGAT
11	G	TAACC	33	F	AAGCA	55	D#	GTAAC	77	C#	ACAGT
12	G#	CCTAA	34	F#	AGCAA	56	E	TGACA	78	D	ATCAG
13	A	CTCAA	35	G	AAAGC	57	F	TACGA	79	D#	ACGAT
14	A#	CTAAC	36	G#	GCAAA	58	F#	TCAAG	80	E	CGATA
15	B	TCACA	37	A	GACAA	59	G	TAACG	81	F	AGACT
16	C	TCAAC	38	A#	GAACA	60	G#	CGTAA	82	F#	AGCAT
17	C#	AAGGA	39	B	GAAAC	61	A	CAGTA	83	G	GACAT
18	D	AGGAA	40	C	AGACA	62	A#	CAAGT	84	G#	AATGC
19	D#	AAAGG	41	C#	AACGA	63	B	CTAAG	85	A	AAGTC
20	E	GGAAA	42	D	ACGAA	64	C	TCAGA	86	A#	AGACT
21	F	GAGAA	43	D#	AAACG	65	C#	ACACT	87	B	AGCAT
22	F#	GAAGA	44	E	CGAAA	66	D	ACCAT	88	C	ATAGC
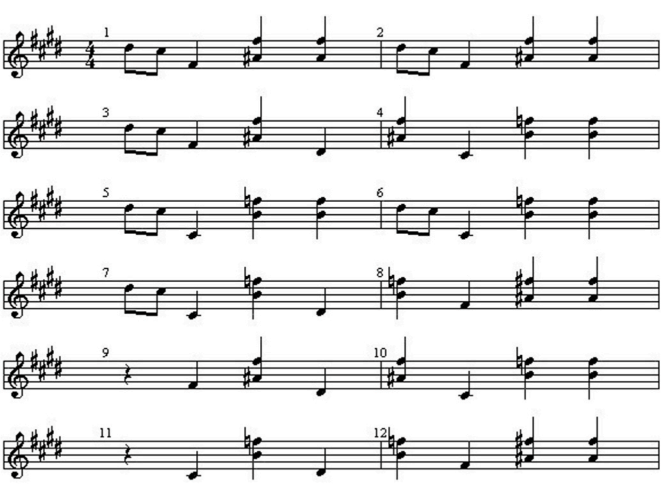
Fig. 6. "Flohwalzer" melody.

GATTATTAAATAACCGAACAACCCAA-
GATTATTAGGAAACGAGGAACGAACCAGCAAAAATAACCGAA-
CAACCCAAGATTATTAAATAACCGAACAACCCAA-
GATTATTAGGAAACGAGGAACGAACCAGCAAAAATAACCGAA-
CAACCCAAGATTATTACCTGAGAAAATAACCGAACAACCCAA-
GATTATTACCGAGTAAAATAACCGAAACACCCAGAATTATTAAA-
TAACCGAAACACCCA-
GAATTATTAGGAAACGAGGAACGAACCGAGTAAAATAACCGAAA-
CACCCAGAATTATTAAATAACCGAAACACCCA-

Table 4
Representation of the "Flohwalzer" as the numbers of the keys and the duration of the notes.

	Note sequence	Note duration
1	43-41-34- (38, 46) - (38, 46)	1/8–1/8 -1/4–1/4 - 1/4
2	43 - 41–34 - (38, 46) - (38, 46)	1/8–1/8 -1/4–1/4 - 1/4
3	43 - 41–34 - (38, 46) - 31 - (38, 46) - 29 - (39, 45) - (39, 45)	1/8–1/8 -1/4–1/4 - 1/4–1/4 -1/4–1/4 - 1/4
4	43 - 41–29 - (39, 45) - (39, 45)	1/8–1/8 -1/4–1/4 - 1/4
5	43 - 41–29 - (39, 45) - (39, 45)	1/8–1/8 -1/4–1/4 - 1/4
6	43 - 41–29 - (39, 45) - 31 - (39, 45) - 29 - (38, 46) - (38, 46)	1/8–1/8 -1/4–1/4 - 1/4–1/4 -1/4–1/4 - 1/4
7	48 - (38, 46) - 50 - (38, 46) - 53 - (39, 45) - (39, 45)	1/4–1/4 - 1/4–1/4 -1/4–1/4 - 1/4
8	53 - (39, 45) –50 - (39, 45) - 48 - (38, 46) - (38, 46)	1/4–1/4 - 1/4–1/4 -1/4–1/4 - 1/4
9	34 - (38, 46) - 31 - (38, 46) - 29 - (39, 45) - (39, 45)	1/4–1/4 - 1/4–1/4 -1/4–1/4 - 1/4
10	29 - (39, 45) - 31 - (39, 45) - 34 - (38, 46) - (38, 46)	1/4–1/4 - 1/4–1/4 -1/4–1/4 - 1/4
11	34 - (38, 46) - 31 - (38, 46) - 29 - (39, 45) - (39, 45)	1/4–1/4 - 1/4–1/4 -1/4–1/4 - 1/4
12	29 - (39, 45) - 31 - (39, 45) - 34 - (38, 46) - (38, 46)	1/4–1/4 - 1/4–1/4 -1/4–1/4 - 1/4

GAATTATTAGGAAACGAGGAACGAACCGAGTAAAATAACCGAAA-
CACCCAGAATTATTAAATAACCGAAACACCCA-
GAATTATTAGGAAACGAGGAACGAACCGAGTAAAATAACCGAAA-
CACCCAGAATTATTACCGAGTAAAATAACCGAAACACCCA-
GAATTATTACCACAGAAAATAACCGAACAACCCAA-
GATTATTACCTACGAAAATAACCGAACAACCCAA-
GATTATTACCGACTAAAATAACCGAAACACCCAGAATTATTAAA-

Table 5
Encoding of the "Flohwalzer" as octanucleotide sequences. Nucleotides in bold are responsible for the sound duration.

1	AGGAAACG AGGAACGA ACCAGCAA AAATA ACCGAACA ACCCAAGA TTATT AAATA ACCGAACA ACCCAAGA TTATT
2	**AGG**AAACG **AGG**AACGA **ACC**AGCAA AAATA **ACC**GAACA **ACC**CAAGA TTATT AAATA **ACC**GAACA **ACC**CAAGA TTATT
3	**AGG**AAACG **AGG**AACGA **ACC**AGCAA AAATA **ACC**GAACA **ACC**CAAGA TTATT **ACC**TGAGA AAATA **ACC**GAACA **ACC**CAAGA TTATT **ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT AAATA **ACC**GAAAC **ACC**CAGAA TTATT
4	**AGG**AAACG **AGG**AACGA **ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT AAATA **ACC**GAAAC **ACC**CAGAA TTATT
5	**AGG**AAACG **AGG**AACGA **ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT AAATA **ACC**GAAAC **ACC**CAGAA TTATT
6	**AGG**AAACG **AGG**AACGA **ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT **ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT
7	**ACC**ACAGA AAATA**ACC**GAACA **ACC**CAAGA TTATT **ACC**TACGA AAATA**ACC**GAACA **ACC**CAAGA TTATT **ACC**GACTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT AAATA **ACC**GAAAC **ACC**CAGAA TTATT
8	**ACC**GACTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT **ACC**TACGA AAATA **ACC**GAAAC **ACC**CAGAA TTATT **ACC**ACAGA AAATA**ACC**GAACA **ACC**CAAGA TTATT AAATA**ACC**GAACA **ACC**CAAGA TTATT
9	**ACC**AGCAA AAATA**ACC**GAACA **ACC**CAAGA TTATT **ACC**TGAGA AAATA**ACC**GAACA **ACC**CAAGA TTATT **ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT AAATA **ACC**GAAAC **ACC**CAGAA TTATT
10	**ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT **ACC**TGAGA AAATA **ACC**GAAAC **ACC**CAGAA TTATT **ACC**GAGTA AAATA**ACC**GAACA **ACC**CAAGA TTATT AAATA**ACC**GAACA **ACC**CAAGA TTATT
11	**ACC**AGCAA AAATA**ACC**GAACA **ACC**CAAGA TTATT **ACC**TGAGA AAATA**ACC**GAACA **ACC**CAAGA TTATT **ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT AAATA **ACC**GAAAC **ACC**CAGAA TTATT
12	**ACC**GAGTA AAATA **ACC**GAAAC **ACC**CAGAA TTATT **ACC**TGAGA AAATA **ACC**GAAAC **ACC**CAGAA TTATT **ACC**GAGTA AAATA**ACC**GAACA **ACC**CAAGA TTATT AAATA**ACC**GAACA **ACC**CAAGA TTATT

TAACCGAAACACCCAGAATTATTACCGACTAAAATAACCGAAA-
CACCCAGAATTATTACCTACGAAAATAACCGAAACACCCA-
GAATTATTACCACAGAAAATAACCGAACAACCCAAGATTATTAAA-
TAACCGAACAACCCAAGATTATTACCAGCAAAAATAACCGAA-
CAACCCAAGATTATTACCTGAGAAAATAACCGAACAACCCAA-
GATTATTACCGAGTAAAATAACCGAAACACCCAGAATTATTAAA-
TAACCGAAACACCCAGAATTATTACCGAGTAAAATAACCGAAA-
CACCCAGAATTATTACCTGAGAAAATAACCGAAACACCCA-
GAATTATTACCGAGTAAAATAACCGAACAACCCAAGATTATTAAA-
TAACCGAACAACCCAAGATTATTACCAGCAAAAATAACCGAA-
CAACCCAAGATTATTACCTGAGAAAATAACCGAACAACCCAA-
GATTATTACCGAGTAAAATAACCGAAACACCCAGAATTATTAAA-
TAACCGAAACACCCAGAATTATTACCGAGTAAAATAACCGAAA-
CACCCAGAATTATTACCTGAGAAAATAACCGAAACACCCA-
GAATTATTACCGAGTAAAATAACCGAACAACCCAAGATTATTAAA-
TAACCGAACAACCCAAGATTATT.

This nucleotide sequence can be presented in binary format as well:
01100000 10110000 10000000 10000000 11000000 10000000
00000001 01111011 11000110 00000100 00000000 00001011
11011110 11000001 01100001 00000001 00000001 10000001
00000000 00000010 11110111 10001100 00001000 00000000
00010111 10111101 10000010 11000010 00000010 00000011
00000010 00000000 00000101 11101111 00000111 01000011
00000010 00000000 00000101 11101111 00000101 11000011
00000010 00000000 00001001 11101111 00011000 00010000
00000000 01001111 01111011 00000101 10000100 00001011
10000110 00000100 00000000 00010011 11011110 00110000
00100000 00000000 10011110 11110110 00001011 00001000
00010111 00001100 00001000 00000000 00100111 10111100
01100000 01000000 00000001 00111101 11101100 00010110
00010000 00101110 00011000 00010000 00000000 01001111
01111000 00101110 00011000 00010000 00000000 01001111
01111000 00000010 00011000 00010000 00000000 00101111
01111000 00110001 00001100 00001000 00000000 00010111
10111100 00010001 10000110 00000100 00000000 00010011
11011110 00110000 00100000 00000000 10011110 11110000
01000110 00011000 00010000 00000000 01001111 01111000
00110001 00001100 00001000 00000000 00100111 10111100
00000001 00001100 00001000 00000000 00010111 10111100
01100000 01000000 00000000 10111101 11100000 01000000
01100000 01000000 00000000 10111101 11100000 11101000
01100000 01000000 00000000 10111101 11100000 10111000
01100000 01000000 00000001 00111101 11100011 00000010
00000000 00001001 11101111 00000101 11000011 00000010
00000000 00001001 11101111 00000111 01000011 00000010
00000000 00001001 11101111 00000101 11000011 00000010
00000000 00000101 11101111 00011000 00010000 00000000
00101111 01111000 00010000 00011000 00010000 00000000
00101111 01111000 00111010 00011000 00010000 00000000
00101111 01111000 00101110 00011000 00010000 00000000
01001111 01111000 11000000 10000000 00000010 01111011
11000001 01110000 11000000 10000000 00000010 01111011
11000001 11010000 11000000 10000000 00000010 01111011
11000001 01110000 11000000 10000000 00000001 01111011
11000110 00000100 00000000 00001011 1101111.

For "Mary had a little lamb" melody, the nucleotide sequence is as follows:
ACCGAAACACCGAACAACCGACAAACCGAACAACCGAAA-
CACCGAAACAGCGAAACACCGAACAACCGAACAAGCGAA-
CAACCGAAACACCAACGAAGCAACGAACCGAAACACCGAACAACCGA-
CAAACCGAACAACCGAAACACCGAAACACCGAAACACCGAAA-
CACCGAACAACCGAACAACCGAAACACCGAACAACGGA-
CAAACCGAAACACCGAACAACCGACAAACCGAACAACCGAAA-
CACCGAAACAGCGAAACACCGAACAACCGAACAAGCGAA-
CAACCGAAACACCAACGAAGCAACGAACCGAAACACCGAACAACCGA-
CAAACCGAACAACCGAAACACCGAAACACCGAAACACCGAAA-
CACCGAACAACCGAACAACCGAAACACCGAACAAGCGA-
CAAACCAACGAACCGAAACACCGAACAACCGACAAACCGAA-
CAACCGAAACACCGAAACAGCGAAACACCGAACAACCGAA-
CAAGCGAACAACCGAAA-
CACCAACGAACCAACGAACCAACGAACCGAAACACCGAACAACCGA-
CAAACCGAACAACCGAAACACCGAAACAGCGAAACACCGAA-
CAACCGAACAAGCGAACAACCGAAA-
CACCAACGAAGCAACGAACCGAAACACCGAACAACCGACAAACCGAA-
CAACCGAAACACCGAAACACCGAAACACCGAAACACCGAA-
CAACCGAACAACCGAAACACCGAACAAGCGA-
CAAACCAACGAACCGAAACACCGAACAACCGACAAACCGAA-
CAACCGAAACACCGAAACACCGAAACACCGAAACACCGAA-
CAACCGAACAACCGAAACACCGAACAACGGACAA.

5. Conclusions

In the current paper a new approach to encode musical notes in DNA is proposed. It is based on octanucleotide sequences (NYRN-oligonucleotides) with 50% GC composition. The first three nucleotides determine the duration and the next five encode notes tone. The sequences suitable for enzymatic manipulations and ensuring higher specific amplification were chosen, i.e., the chemical structure of oligonucleotides is taken into account. The approach provides easy melodies conversion into a binary format, where one note is encoded with one byte. Unlike other methods including Huffman's algorithm, the approach involves a wider set of sound characteristics, such as duration and semitones. Although it does not consider changing the tempo and increasing/decreasing the volume, it allows encoding of more complicated melodies containing chords rather than sequential keystrokes.

CRediT authorship contribution statement

Olga Yu Kiryanova: Conceptualization, Software, Writing – original draft. **Ravil R. Garafutdinov:** Conceptualization, Writing – original draft, Writing – review & editing. **Irek M. Gubaydullin:** Project administration, Validation. **Alexey V. Chemeris:** Methodology, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ailenberg, M., Rotstein, O.D., 2009. An improved Huffman coding method for archiving text, images, and music characters in DNA. Biotechniques 47 (3), 747–754. https://doi.org/10.2144/000113218.

Antkowiak, P.L., Lietard, J., Darestani, M.Z., Somoza, M.M., Stark, W.J., Heckel, R., Grass, R.N., 2020. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. Nat. Commun. 11 (1), 5345. https://doi.org/10.1038/s41467-020-19148-3.

Binary Code and Storing Music on Computers. URL: https://tomrocksmaths.com/2021/10/27/binary-code-and-storing-music-on-computers/.

Blawat, M., Gaedke, K., Huetter, I., Chen, X.-M., Turczyk, B., Inverso, S., Pruitt, B., Church, G., 2016a. Forward error correction for DNA data storage. Proc. Comput. Sci. 80, 1011–1022. https://doi.org/10.1016/j.procs.2016.05.398.

Blawat, M., Gaedke, K., Huetter, I., Chen, X.-M., Turczyk, B., Inverso, S., Pruitt, B., Church, G., 2016b. Forward error correction for DNA data storage. Procedia Comput. Sci. 80, 1011–1022. https://doi.org/10.1016/j.procs.2016.05.398.

Church, G.M., Gao, Y., Kosuri, S., 2012. Next-generation digital information storage in DNA. Science 337 (6102), 1628. https://doi.org/10.1126/science.1226355.

Clelland, C.T., Risca, V., Bancroft, C., 1999. Hiding messages in DNA microdots. Nature 399 (6736), 533–534. https://doi.org/10.1038/21092.

Davis, J., 1996. Microvenus. Art J. 55 (1), 70–74. https://doi.org/10.2307/777811.

DNA-Based Digital Storage//Twist Bioscience. White paper. URL: https://www.twistbioscience.com/resources/publication??products=-1&journals=-1&year=-1.

Erlich, Y., Zielinski, D., 2017. DNA Fountain enables a robust and efficient storage architecture. Science 355 (6328), 950–954. https://doi.org/10.1126/science.aaj2038.

Flohwalzer. URL: https://en.wikipedia.org/wiki/Flohwalzer(date of the application 13. January.2023).

Garafutdinov, R.R., Chemeris, D.A., Sakhabutdinova, A.R., Chemeris, A.V., Kiryanova, O.Y., Mikhaylenko, C.I., 2022. Encoding of non-biological information for its long-term storage in DNA. Biosystems 215–216, 104664.

Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B., Birney, E., 2013a. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494 (7435), 77–80. https://doi.org/10.1038/nature11875.

Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B., Birney, E., 2013b. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494 (7435), 77–80. https://doi.org/10.1038/nature11875.

Hart V. Using Binary Numbers in Music. URL: https://archive.bridgesmathart.org/2008/bridges2008-397.pdf.

Hodgson, C.P., 1990. A DNA text code. Biotechniques 9 (3), 312.

Organick, L., Ang, S.D., Chen, Y.J., Lopez, R., Yekhanin, S., Makarychev, K., Racz, M.Z., Kamath, G., Gopalan, P., Nguyen, B., Takahashi, C.N., Newman, S., Parker, H.Y., Rashtchian, C., Stewart, K., Gupta, G., Carlson, R., Mulligan, J., Carmean, D., Seelig, G., Ceze, L., Strauss, K., 2018a. Random access in large-scale DNA data storage. Nat. Biotechnol. 36 (3), 242–248. https://doi.org/10.1038/nbt.4079.

Organick, L., Ang, S.D., Chen, Y.J., Lopez, R., Yekhanin, S., Makarychev, K., Racz, M.Z., Kamath, G., Gopalan, P., Nguyen, B., Takahashi, C.N., Newman, S., Parker, H.Y., Rashtchian, C., Stewart, K., Gupta, G., Carlson, R., Mulligan, J., Carmean, D., Seelig, G., Ceze, L., Strauss, K., 2018b. Random access in large-scale DNA data storage. Nat. Biotechnol. 36 (3), 242–248. https://doi.org/10.1038/nbt.4079. Erratum: Nat. Biotechnol. 2018. V.36(7). P.660. doi: 10.1038/nbt0718-660c.

Sakhabutdinova, A.R., Mikhailenko, K.I., Garafutdinov, R.R., Kiryanova, O.Yu, Sagitova, M.A., Sagitov, A.M., Chemeris, A.V., 2019. Non-biological application of DNA molecules. Biomics 11 (3), 344–377.

Scripts for encoding URL: https://github.com/olgakiryanova/music-encoding.

Service, F.R., March 2017. DNA could store all of the world's data in one room. Science. https://doi.org/10.1126/science.aal0852.

Smith, G.C., Fiddes, C.C., Hawkins, J.P., Cox, J.P., 2003. Some possible codes for encrypting data in DNA. Biotechnol. Lett. 25 (14), 1125–1130. https://doi.org/10.1023/A:1024539608706.

Watson, J.D., Crick, F.H.C., 1953. A structure for deoxyribose nucleic acid. Nature 171 (4356), 737–738. https://doi.org/10.1038/171737a0.

Yamuna, M., Sankar, A., Ravichandran, S., Harish, V., 2013. Encryption of a binary string using music notes and graph theory. Int. J. Eng. Technol. 5 (3), 2920–2925.

Zhirnov, V., Zadegan, R.M., Sandhu, G.S., Church, G.M., Hughes, W.L., 2016. Nucleic acid memory. Nat. Mater. 15 (4), 366–370. https://doi.org/10.1038/nmat4594.